ANNALI DI STATISTICA

Anno 120 Serie IX - Vol. 11

FORZE DI LAVORO: DISEGNO DELL'INDAGINE E ANALISI STRUTTURALI

a cura di

Ugo Trivellato

ISTITUTO NAZIONALE DI STATISTICA

Roma 1991

Supplemento all'Annuario Statistico Italiano

ISSN: 0075 - 1766

Fotocomposizione: System Data Computer S.r.l.-Via A. Baldovinetti,56/64-Roma-Contratto n.35 del 28-3-1991
Stampa: Arti Grafiche Rubbettino- Soveria Mannelli (CZ) - Contratto n. 86 del 12-7-1991 - copie 1.500

INDICE Pag.

Pag.

Pag.

Presentazione

Dal 1959 l'indagine ISTAT sulle forze di lavoro rappresenta la principale fonte di informazioni sul mercato del lavoro italiano. L'impegno dell'ISTAT in questo campo è testimoniato non solo dai notevoli oneri organizzativi sostenuti, ma anche e soprattutto dalle risorse intellettuali che in periodi diversi si sono concentrate sull'argomento. Tra le ultime iniziative meritano particolare menzione l'attività della 'Commissione per un sistema informativo del lavoro' e della 'Commissione per lo studio dei campioni'.

La prima ha riunito esponenti del mondo accademico, dei principali organi di rilevazione, nonché degli operatori sul mercato del lavoro, ed ha consentito di fare passi importanti verso la creazione di un sistema integrato sulle statistiche del mercato del lavoro, mettendo a confronto informazioni provenienti da fonti diverse, e favorendo la loro armonizzazione e quindi comparabilità.

Nella Commissione per lo studio dei campioni sono stati affrontati tutti i problemi connessi con la formulazione del progetto di un generico disegno campionario, con particolare riferimento all'indagine sulle forze di lavoro, che è stata quindi esaminata con grande attenzione alla luce dei più recenti sviluppi della teoria.

Nonostante l'intensa attività scientifica di cui si è detto, l'importanza dei fenomeni riguardanti occupazione e disoccupazione, e della loro rilevazione, ha condotto alla sollecitazione di un ulteriore contributo, rivelatosi di grande rilievo e di notevole peso nel processo di analisi critica, di rinnovamento e di sviluppo dell'indagine. Tale contributo si è concretato in una iniziativa di ricerca delle Università di Padova, Firenze e Roma, alla quale l'ISTAT ha subito aderito. L'iniziativa ha preso il nome di FOLA (da FOrze di LAvoro) e si è sviluppata per un arco di cinque anni mediante un lavoro integrato di accademici e di statistici dell'ISTAT.

Il gruppo di ricerca FOLA si è dedicato all'analisi di molti aspetti relativi all'indagine sulle forze di lavoro. Il disegno campionario, la qualità delle informazioni raccolte, indagata anche mediante il confronto tra più indagini successive, l'analisi dei dati, anche con l'uso di modelli matematici sofisticati, nonché l'impiego di indagini pilota su svariati argomenti sono stati tutti esaminati con grande impegno da validi studiosi di diversa estrazione, che hanno apportato un contributo sostanziale di idee e di suggerimenti.

Circa la concreta applicazione alla rilevazione sulle forze di lavoro delle idee scaturite dall'attività del gruppo FOLA ed esposte nei capitoli che seguono, occorre far notare che alcune sono già state di fatto adottate: i suggerimenti circa le modifiche al piano di campionamento, con particolare riferimento all'utilizzo di informazioni esogene sulla struttura per classi di età della popolazione, la realizzazione di indagini pilota, l'effettuazione di analisi statistiche esplorative sui dati rilevati hanno addirittura trovato applicazione nelle more della realizzazione di questo volume o trovano eco nei programmi statistici dell'ISTAT.

Altri contributi, di più rilevante contenuto scientifico, e di più ardita concezione, devono ancora trovare pratica collocazione nell'ambito delle attività previste nei programmi di rinnovamento a medio-lungo termine: è il caso in particolare delle

analisi di qualità longitudinali e dell'utilizzo degli stimatori compositi.

Senza voler trarre conclusioni definitive sulla feconda esperienza acquisita, la cui applicazione è, come si è detto, ancora in corso, vorrei esprimere rapidi commenti su alcuni degli argomenti trattati dal gruppo di lavoro.

Inizio dal disegno di campionamento che, pur seguendo linee tradizionali, è tuttavia robusto e ben collaudato. Non sembra infatti conseguibile nel breve periodo un miglioramento della tecnica di selezione del campione significativamente maggiore di quello già recentemente ottenuto (luglio 1990), in seguito alla revisione dei criteri di stratificazione dei comuni e al riequilibrio della numerosità delle famiglie intervistate nelle singole provincie.

Il controllo della struttura della popolazione di riferimento per classe di età, già sperimentato a partire dalla rilevazione di gennaio 1991, anche se non introdotto definitivamente, non conduce ad un sostanziale aumento di stabilità e significatività delle stime, se la popolazione anagrafica impiegata per i coefficienti di espansione è tenuta costantemente aggiornata facendo ricorso ai dati più recenti sul movimento della popolazione.

Per quanto precede, i più convincenti contributi di ricerca sul piano di campionamento appaiono quelli che affrontano problematiche nuove, o, in altre parole, che suggeriscono nuove idee: lo stimatore composito, che utilizza la parte di campione in comune tra due periodi successivi, le stime entro piccole aree e le stime provvisorie, che anticipano i tempi di diffusione dei dati definitivi, sono esempi di campi di ricerca che, pur fecondamente affrontati all'interno dell'ISTAT, hanno sempre necessità di ulteriori validi contributi.

Nella rilevazione sulle forze di lavoro le famiglie incluse nel campione vengono intervistate quattro volte prima di essere abbandonate. Questa procedura, caratteristica delle indagini di tipo 'panel', viene chiamata in causa in varie occasioni: oltre alla sopra menzionata possibilità di utilizzare la parte comune del campione per stabilizzare e rendere più precise le stime dei principali aggregati (stimatori compositi), l'impiego delle osservazioni ripetute sulle stesse persone consente di ottenere stime dei flussi 'da', 'per' ed 'entro' gli stati di occupazione, ricerca di lavoro e condizione non lavorativa. Sebbene questo tipo di analisi sia già correntemente incluso nelle pubblicazioni sulle statistiche del lavoro dell'Istituto, molti utili suggerimenti sono giunti dal gruppo FOLA sui metodi per incrementare la quota di reinterviste utilizzabili recuperando quelle altrimenti scartate per via di banali errori di trascrizione degli intervistatori o di registrazione dei codificatori.

Importante è anche l'uso della ripetizione delle interviste nel controllo della qualità dei dati, specie con riferimento a certe informazioni a contenuto temporale come l'età dei rispondenti e la durata del periodo di ricerca di lavoro. I lavori compiuti dal gruppo in questo campo possono essere di grande aiuto nella preparazione e nella stesura delle regole di verifica ed imputazione dei dati incongruenti e mancanti.

Il ventaglio degli aggregati prodotti dall'indagine e regolarmente pubblicati dall'ISTAT è attualmente assai ampio, ed è articolato in un gran numero di tabelle statistiche rilevanti per quantità e qualità, soprattutto alla luce della loro capacità di soddisfare consolidati bisogni degli utenti, ai quali l'ISTAT ha fornito un'articolata risposta suggerita da molti anni di scambi di vedute con studiosi, decisori politici ed operatori economici e sociali.

Tuttavia, le analisi compiute sulle predette tabelle utilizzando strumenti statistici avanzati, pur essendo state da sempre compiute da studiosi e ricercatori (tra i quali naturalmente sono compresi anche i ricercatori dell'ISTAT), non sono state quasi mai incluse nelle pubblicazioni ufficiali dei dati. Pertanto le ricerche condotte in questo campo dal progetto FOLA costituiscono ancora di più un prezioso corpo di conoscenze in nuovi e promettenti campi quali l'analisi esplorativa dei dati e la loro modellizzazione. Tra i vari argomenti toccati reclamano attenzione le tecniche che da matrici di dati di varia origine cercano di far scaturire il significato complessivo in esse contenuto, l'identificazione di differenti tipologie di famiglie, i differenti approcci alla modellizzazione di fenomeni specifici.

Due argomenti di attualità e di estrema importanza sono poi la gestione della rete di rilevazione e le modifiche da apportare al questionario.

Sulla rete di rilevazione rimane da compiere un lavoro di riflessione ed approfondimento unito ad interventi sul campo: i miglioramenti nella rete di rilevazione costituiscono una solida garanzia sia della tempestività che della qualità ed affidabilità delle informazioni raccolte. Un notevole passo avanti in tal senso è stato quello prodotto dall'indagine sperimentale sulla 'compenetrazione delle assegnazioni agli intervistatori', condotta dall'ISTAT sotto l'egida del gruppo di lavoro. Essa è consistita, in pratica, nell'associare casualmente sottoinsiemi di famiglie di un comune a rilevatori diversi. Confrontando le risposte ottenute dai vari rilevatori è stato così possibile ricavare una stima delle differenze nei risultati addebitabili alla sola azione dei rilevatori. È questa una conoscenza preziosa per interventi tesi a migliorare il processo di rilevazione.

Le modifiche al questionario per tener conto del mutare delle condizioni occupazionali e della struttura del mercato del lavoro sono state ricorrenti da quando la rilevazione fu per la prima volta iniziata ad oggi. Su di esse c'è sempre stato un grande dibattito. Da un lato l'ISTAT, anche perché da più parti sollecitato, provvedeva ad adeguare i quesiti e la struttura dell'intervista alla mutata situazione. Dall'altro, le inevitabili discontinuità nelle serie prodotte, le difficoltà di comparare vecchi e nuovi aggregati, le incertezze nel cogliere le implicazioni dei mutamenti nelle definizioni e nelle classificazioni, inducevano gli utenti dei dati a lamentare problemi e difficoltà. Il dilemma ha trovato un parziale componimento nella pratica, pure diffusa, della 'ricostruzione' delle serie alla luce dei mutati contenuti formali o sostanziali della rilevazione. Essa tuttavia non assicura quasi mai risultati perfettamente omogenei per tutto l'arco di tempo coperto dalle serie storiche, e ha talora l'effetto di disorientare il pubblico.

Per questi motivi, prima di iniziare una consistente revisione è opportuno fare un estensivo ricorso ad indagini pilota, che consentono di stabilire, tra le altre cose: se vale davvero la pena di compiere certe modifiche; se non vi sono effetti collaterali dannosi nell'inserimento, nella soppressione o nella modificazione di certe domande; se la raccolta, il trattamento e l'elaborazione dei dati si rivela relativamente agevole. Quanto precede viene naturalmente appurato a costi molto ridotti rispetto a quelli di un'indagine corrente.

Un sottogruppo del progetto FOLA, a composizione mista Università-ISTAT, ha realizzato un'indagine sulla storia lavorativa di un campione di famiglie lombarde: i risultati di tale rilevazione, in forma di banca dati, renderanno possibile l'analisi della capacità informativa dei vari quesiti. Data l'affinità dell'argomento, conviene

menzionare qui anche l'indagine pilota sul nuovo questionario delle forze di lavoro, anche se in essa non vi è un diretto contributo del gruppo: si può affermare che l'impulso stesso a mutare i quesiti a partire dal 1992 sia una conseguenza non solo dei programmi dell'EUROSTAT ma anche dei risultati del progetto FOLA. L'indagine, i cui risultati saranno disponibili entro l'anno 1991, è tesa a verificare l'impatto delle nuove definizioni comunitarie di persone in cerca di occupazione, nonché gli effetti del raffinamento della classificazione per branca di attività economica e dell'introduzione della professione degli occupati e del tipo di titolo di studio posseduto dagli intervistati. Essa contiene inoltre un gran numero di innovazioni minori.

Sono molto grato per il contributo dato dagli ideatori e dai ricercatori del progetto. Il loro apporto trascende l'ambito specifico dell'indagine sulle forze di lavoro e investe un campo molto esteso del lavoro statistico. Il successo del progetto ci induce a rinforzare i collegamenti ISTAT-Università.

Guido Mario Rey
Presidente dell'ISTAT

Ringraziamenti

Questo volume raccoglie i risultati salienti di una ricerca su "Forze di lavoro: disegno dell'indagine e analisi strutturali", forse più nota tra statistici ed economisti del lavoro con l'acronimo con cui, per esigenze di brevità (o per conformismo alla moda delle sigle?), sin dagli inizi si prese a designarla dai componenti il gruppo di ricerca: FOLA. Forse perché si prestava, in modo sin troppo facile - e, serve dirlo?, non accidentale -, ad allusioni a lungaggini ... fiabesche, l'acronimo ha conosciuto, nella ristretta cerchia degli addetti ai lavori, una qualche fortuna. Sicché mi è capitato più d'una volta di sentirmi chiedere a che punto stava o quando, e come, finiva la FOLA. Il punto d'arrivo, per quanto di punto d'arrivo si può parlare nell'attività scientifica, è per l'appunto l'insieme dei saggi che costituiscono questo volume.

La ricerca ha preso avvio nella seconda metà del 1986, finanziata dal Ministero della Pubblica Istruzione nell'ambito dei progetti di interesse nazionale, con un'articolazione su tre unità di ricerca: presso le Università di Padova, Firenze e Roma 'La Sapienza'. Successivamente, è stata finanziata anche dall'Istat, tramite una convenzione di ricerca con il Dipartimento di Scienze Statistiche dell'Ateneo patavino. L'apporto dell'Istat è stato essenziale: non solo per le ulteriori risorse finanziarie che ha fornito, ma ancor più per le possibilità di collaborazione scientifica ed operativa che ha aperto. L'integrazione nel gruppo di ricerca di funzionari dell'Istituto ha consentito un confronto scientifico serrato, misurato in modo stringente sull'impianto dell'indagine sulle forze di lavoro e sulle problematiche che ne discendono. Il rapporto di collaborazione si è poi esteso all'Istat *tout court*, e segnatamente al Servizio Indagini sulle Famiglie, per la progettazione e lo svolgimento di alcune indagini suppletive alla rilevazione corrente sulle forze di lavoro, di cui danno parzialmente conto i capitoli conclusivi del volume. Di tutto ciò ringrazio Guido Rey, Vincenzo Siesto e Mario Agostinelli. Infine, per approfondimenti monografici in tema di disoccupazione un'ulteriore supporto è venuto da un contratto di ricerca su "La natura e la durata della disoccupazione", nell'ambito del Progetto Finalizzato CNR "Struttura ed evoluzione dell'economia italiana".

Del gruppo di ricerca hanno fatto parte, oltre agli autori dei vari capitoli del volume, il cui contributo è palese, anche Luigi Cannari, Maria Castellini, Ottorino Chillemi, Ugo Colombino, Giovanna D'Angiolini, Achille Lemmi, Guido Masarotto, Corrado Provasi e Alessandra Toti. Anche se per varie circostanze i loro nomi non compaiono nel volume conclusivo, il loro apporto è stato tutt'altro che secondario, com'è documentato dalle note e dai rapporti che hanno scritto su tematiche della ricerca. Un riconoscimento particolare va a Alessandra Toti, per aver aiutato parecchi di noi a districarsi nel guazzabuglio dei *files* di dati che ci siamo trovati a maneggiare. Non posso poi non ringraziare tutti i componenti il gruppo di ricerca, per il clima di collaborazione e di amicizia che ha sempre accompagnato il confronto scientifico, e per aver accolto di buon grado i miei stimoli al coordinamento ed i miei richiami al rispetto di scadenze. Se qualche risultato significativo è stato raggiunto in un progetto di ricerca che ha coinvolto attorno ad un tema specifico - l'indagine

sulle forze di lavoro - uno spettro di sensibilità e competenze assai ampio, parecchio si deve a questo clima.

Versioni preliminari di saggi che compaiono in questo volume sono state discusse in Seminari tenuti tra il 1986 e il 1989 nelle Università di Firenze, Padova, University College (Londra), Madison-Wisconsin, Michigan (Ann Arbor), Ottawa e UCLA (Los Angeles), nonché in occasione di relazioni o comunicazioni presentate al 'Seminario sul sistema informativo del lavoro nazionale' (Istat e Ministero del Lavoro, Roma, 11 febbraio 1987), alla Conferenza su 'Labour Force Sample Survey as an Employment Policy Instrument' (Commission des Communautés Européennes, Fontevraud, 24-26 settembre 1987), al Seminario CNR 'La struttura dell'economia italiana: evoluzione e scenari' (Venezia, 3-5 dicembre 1987), alla 34ª Riunione Scientifica della Società Italiana di Statistica (Siena, 27-30 aprile 1988), al 3rd Annual Congress of the European Economic Association (Bologna, 27-29 agosto 1988), alla Vth Annual Research Conference of the U.S. Census Bureau (Arlington, Virginia, 19-22 marzo 1989), alla 2nd Conference of the International Federation of Classification Societies (Charlottesville, Virginia, 27-30 giugno 1989), alla 47th Session of the International Statistical Institute (Parigi, 29 agosto-6 settembre 1989), alla 35ª Riunione Scientifica della Società Italiana di Statistica (Padova, 18-21 aprile 1990). Commenti e suggerimenti ricevuti sono serviti non poco per il seguito dell'attività.

Occasioni particolarmente importanti di verifica e di dibattito sono state poi offerte da due Seminari sullo specifico tema della ricerca, tenutisi a Bressanone il 23-25 settembre 1987 e il 14-16 settembre 1988, nell'ambito delle attività promosse dalla Facoltà e dal Dipartimento di Scienze Statistiche patavini. Il primo incontro ha fornito l'opportunità per la presentazione di risultati iniziali e per una proficua discussione tra l'intero gruppo dei ricercatori, allargato a un contenuto numero di esperti. Il secondo incontro, collocato ad uno stadio ormai già avanzato della ricerca e aperto ad un ampio insieme di studiosi, ha permesso un serrato confronto scientifico, arricchito anche da comunicazioni di ricercatori stranieri. Infine, v'é da ricordare che, in vista della messa a punto dell'indagine suppletiva sulla storia lavorativa, l'Istat ha organizzato a Milano, il 30 maggio 1988, un apposito Seminario di studio, che ha consentito di raccogliere commenti ed indicazioni di un qualificato gruppo di statistici, economisti e sociologi del lavoro.

Nominando qualcuno, si corre inevitabilmente il rischio di manchevolezze per omissione. È un rischio che debbo correre, per ringraziare, tra i molti che con le loro critiche e i loro suggerimenti ci hanno accompagnato nel procedere della ricerca, alcuni colleghi che lo hanno fatto con particolare attenzione: Luigi Biggeri, Bernardo Colombo, Carlo Dell'Aringa, Chris Flinn, Luigi Frey, Amato Herzel, Riccardo Leoni, Giuseppe Leti, Aldo Predetti. Ai ringraziamenti, proprio perché in parte così circostanziati, è appena ovvio accompagnare la precisazione circa la responsabilità dei soli autori per le opinioni espresse e per eventuali errori.

L'attenzione mostrata dal CINECA per alcune esigenze di spazio di memoria, connesse con la gestione di grandi *files* di dati e con l'accesso agli stessi da più sedi universitarie, è stata di notevole aiuto. Ne sono grato a Salvatore Rago e a Remo Rossi, Presidente del Consorzio. Le indagini suppletive non avrebbero potuto svolgersi, e con esiti positivi, senza l'impegno, ben superiore al dovuto, di Amedeo Ciriello, direttore dell'Ufficio Statistico Regionale dell'Istat per la Lombardia, e dei

suoi collaboratori. Di grande utilità è stata anche la collaborazione fornita dall'Ufficio Statistica della Regione Lombardia, e dal suo direttore Enrico Caperdoni.

Il personale amministrativo e tecnico del Dipartimento di Scienze Statistiche patavino ha contribuito, con capacità e cortesia, alle molteplici, essenziali attività di supporto alla ricerca. Un grazie a tutti, e in particolare a Vincenzo Porfido, che si è accollato la segreteria del gruppo di ricerca. Nella fase finale di approntamento del volume, si è aggiunto l'apporto di Daniela Serafini, che ha curato l'organizzazione della bibliografia.

I colleghi Paolo De Sandre e Giancarlo Diana si sono succeduti alla direzione del Dipartimento di Scienze Statistiche patavino nel periodo di progettazione e di avvio della ricerca, e mi hanno aiutato in molti modi: con scambi di opinioni sul merito dell'iniziativa, col convinto sostegno al progetto, con la solerte attenzione dedicata alla definizione della convenzione tra l'Ateneo e l'Istat. Se non vado errato, quello per FOLA è stato il primo contratto di ricerca stipulato dall'Istituto Nazionale di Statistica non con singoli studiosi, ma con un'Università e un suo Dipartimento. Ciò ha comportato qualche lungaggine, una sorta di inevitabile rodaggio sul fronte degli aspetti amministrativi, che è in buona misura gravato su di loro. D'altra parte, ciò è servito a tracciare uno standard per i rapporti fra Istat e Università. E il fatto che questa strada possa, oggi, essere percorsa agevolmente è il miglior riconoscimento per questa dimensione, forse non la più gratificante, del loro impegno.

L'ultimo ringraziamento, ma il più sentito, è a Silvano Bordignon e a Gianni Marliani. A Silvano e a Gianni sono infatti debitore per avermi affiancato, con acume e generosità, il primo nella direzione della ricerca, durante l'anno sabatico che ho trascorso all'University of Wisconsin-Madison, ed il secondo nelle conclusiva fatica di curatore di questo volume.

Università di Padova, settembre 1990

U.T.

Gli autori

Giorgio Alleva	Dipartimento di Studi Geoeconomici, Statistici, Storici per l'Analisi Regionale, Università di Roma 'La Sapienza'
Lorenzo Bernardi	Dipartimento di Scienze Statistiche, Università di Padova
Sergio Bolasco	Dipartimento di Sociologia, Università di Salerno
Silvano Bordignon	Dipartimento di Scienze Statistiche, Università di Padova
Daniela Cocchi	Dipartimento di Scienze Statistiche 'Paolo Fortunati', Università di Bologna
Giuliana Coccia	Servizio Studi Istat, Roma
Renato Coppi	Dipartimento di Statistica, Probabilità e Statistiche Applicate Università di Roma 'La Sapienza'
Fabio Corradi	Dipartimento di Studi Geoeconomici, Statistici, Storici per l'Analisi Regionale, Università di Roma 'La Sapienza'
M. Antonia De Marchis	Servizio Indagini sulle Famiglie Istat, Roma
Ignazio De Nicola	Servizio Indagini sulle Famiglie Istat, Roma
Ersilia Di Pietro	Servizio Indagini sulle Famiglie Istat, Roma
Luigi Fabbris	Dipartimento di Scienze Statistiche, Università di Padova
Piero D. Falorsi	Servizio Studi Istat, Roma
Giulio Ghellini	Servizio Statistica, Regione Lombardia, Milano
Andrea Giommi	Dipartimento Statistico, Università di Firenze

Antonio Giusti	Dipartimento Statistico, Università di Firenze
Silvio Lauro	Servizio Statistica, Regione Lombardia, Milano
Gianfranco Lovison	Dipartimento di Scienze Statistiche, Università di Padova
Paolo Manfroni	Servizio Indagini sulle Famiglie Istat, Roma
Gianni Marliani	Dipartimento Statistico, Università di Firenze
Fausta Ongaro	Dipartimento di Scienze Statistiche, Università di Padova
Giuliana Passamani	Istituto di Statistica e Ricerca Operativa, Università di Trento
Enrico Rettore	Dipartimento di Scienze Statistiche, Università di Padova
Aldo Russo	Servizio Studi, Istat , Roma
Ippolito Sanetti	Servizio Censimenti, Istat, Roma
Francesco Sanna	Dipartimento di Studi Geoeconomici, Statistici, Storici per l'Analisi Regionale, Università di Roma 'La Sapienza'
Isabella Santini	Dipartimento di Studi Geoeconomici, Statistici, Storici per l'Analisi Regionale, Università di Roma 'La Sapienza'
Marina Schenkel	Dipartimento di Scienze Economiche, Università di Padova
Nicola Torelli	Dipartimento di Scienze Statistiche, Università di Padova
Ugo Trivellato	Dipartimento di Scienze Statistiche, Università di Padova
Susanna Zaccarin	CoSES, Venezia
Alberto Zuliani	Dipartimento di Studi Geoeconomici, Statistici, Storici per l'Analisi Regionale, Università di Roma 'La Sapienza'

PARTE PRIMA:

IL QUADRO DELLA RICERCA

1

FOLA: SINTESI DI UNA RICERCA

Ugo Trivellato

1. *Le indagini sulle forze di lavoro: qualche considerazione introduttiva*

Le statistiche correnti sulla partecipazione al lavoro dei diversi Paesi sono naturalmente condizionate dalle specificità dei contesti nazionali: specificità che riguardano il mercato del lavoro *tout court*, le preoccupazioni politiche e le esigenze conoscitive con cui ad esso si guarda, i dispositivi di rilevazione.

Per cogliere, in via esemplificativa, marcate differenze nella situazione del mercato del lavoro, è sufficiente guardare alla variabilità territoriale dei tassi di attività e di disoccupazione, e delle disuguaglianze di tali tassi fra gruppi demografici - maschi/femmine e giovani/adulti -, in un'area tutto sommato piuttosto omogenea, quella della Comunità Europea[1].

Tali divari in parte rimandano a, e comunque sono accompagnati da, specifiche caratteristiche dell'intervento pubblico in materia di lavoro: caratteristiche che hanno profonde radici nella storia economica e sociale dei singoli Paesi. Le attività di formazione e primo inserimento nel lavoro, i sistemi di assicurazione e di previdenza obbligatoria associati all'occupazione, i servizi di avviamento al lavoro e le forme di erogazione di *unemployment benefits* [2] alle persone in cerca di lavoro, sono alcune fra le dimensioni più significative dell'intervento pubblico che mostrano ancor oggi apprezzabili

---

1 Dati correnti in materia compaiono nelle pubblicazioni periodiche dell'Eurostat. Uno scorcio in chiave comparativa riferito al 1986, attento a disparità nel livello e nell'incidenza della disoccupazione, è in Trivellato (1990). Non è forse inutile richiamarne le evidenze essenziali. Se nell'Europa dei dodici il tasso di disoccupazione è attestato attorno al 10-11%, restando ai maggiori Paesi esso è inferiore al 7% in Germania, mentre supera il 21% in Spagna. Con altrettanta chiarezza risalta come non meno diversificate territorialmente siano le disuguaglianze nella distribuzione del peso della disoccupazione fra gruppi demografici. Così, il rapporto fra il tasso di disoccupazione femminile e quello maschile, che a livello comunitario è pari a 1,4, scende a 0,9 per il Regno Unito e sale a 2,4 per l'Italia. E in maniera ancor più accentuata si riscontra una polarizzazione dei rischi di disoccupazione sui giovani: per un giovane che cerca lavoro, nell'intera Comunità la probabilità di disoccupazione è 2,7 volte quella di un adulto; a fronte di questo dato medio, stanno tuttavia situazioni marcatamente differenti, con i poli da un lato della Germania, dove il rapporto fra i tassi di disoccupazione è pari a 1,3, e dall'altro lato dell'Italia, dove l'analogo rapporto sale a 6,7.

2 Non è per un vezzo anglofono che non uso l'espressione italiana 'indennità di disoccupazione'. Essa evoca dimensioni del sussidio così esigue che rischia di essere fuorviante. Probabilmente, se si dovesse cercare un analogo italiano di *unemployment benefits*, il termine appropriato sarebbe 'salario di disoccupazione'.

diversità.

Non sorprende dunque che nei vari Paesi permangano sensibili differenze nei dispositivi di rilevazione dell'occupazione e della disoccupazione (vedi, tra gli altri, Fürst, 1988, e OECD, 1990). Esse chiamano in causa le peculiarità dei sistemi statistici nazionali, e rimandano inoltre alle diversità di indole più generale nelle caratteristiche dell'intervento pubblico in tema di lavoro, che ho appena evocato. In altre parole, le forme di regolazione sociale e di gestione amministrativa dei rapporti di occupazione da un lato e della condizione di disoccupazione dall'altro hanno un notevole impatto sulla stessa rilevazione statistica dei fenomeni. Così, tipicamente i Paesi contraddistinti da sistemi generalizzati e consolidati di sicurezza sociale e di *welfare* tendono ancor oggi a trarre l'informazione corrente, a cadenza sub-annuale, sull'occupazione e la disoccupazione soprattutto da fonti amministrative[3].

Non manca tuttavia un'eccezione, e di grande rilievo. L'eccezione è costituita dalle indagini campionarie delle forze di lavoro condotte sulle famiglie. Tali indagini hanno preso avvio nel secondo dopoguerra sulla scorta della pionieristica, e per molti versi esemplare, esperienza della *Current Population Survey* statunitense. Esse si sono progressivamente diffuse grazie anche all'opera di stimolo e di standardizzazione definitoria - e, per i Paesi in via di sviluppo, di supporto tecnico - dell'*International Labour Office* (ILO), e vengono ormai condotte correntemente in una cinquantina di Paesi (ILO, 1986). Nell'ambito della Comunità Europea, uno specifico impulso ed un più penetrante coordinamento è poi venuto dall'Eurostat, segnatamente con la versione dell'indagine, sincrona e armonizzata, adottata dal 1983 con cadenza annuale (Eurostat, 1985).

Se nei Paesi con evoluti sistemi di sicurezza sociale e di *welfare* le indagini sulle forze di lavoro hanno sovente un ruolo integrativo rispetto alla documentazione statistica corrente tratta da fonti amministrative (obbedendo ad esigenze di più articolata rilevazione delle modalità e del grado della partecipazione al lavoro e/o di armonizzazione a livello sovranazionale, esigenze che non sollecitano una frequenza particolarmente alta dell'indagine), per i Paesi dove ciò non si dà esse costituiscono la basilare fonte corrente per la misura dell'occupazione e della disoccupazione. In tal caso, tendono ad essere svolte a cadenza ravvicinata, mensile o trimestrale.

La rilevazione trimestrale delle forze di lavoro italiana (nel seguito RTFL) condivide per l'appunto queste caratteristiche e questo rilievo. Consolidatasi sin dal 1959, essa ha accompagnato lo svolgersi del dibattito sulla misurazione e l'analisi della partecipazione al lavoro nel nostro Paese: alimentandolo di evidenze empiriche e ad un tempo registrando i mutamenti di preoccupazioni conoscitive.

Il ruolo assolutamente preminente che l'indagine è venuta assumendo è stato in larga misura motivato proprio dalla sostanziale indisponibilità dei dati di origine amministrativa sugli occupati e dalla *scarsa affi*dabilità di quelli

---

3 Significativamente, nella Comunità Europea ciò avviene per tutti i Paesi, a meno di Grecia, Italia, Spagna e Portogallo.

sui disoccupati. In conseguenza di questo stato di cose, si è fatto riferimento sempre più ampiamente alla RTFL per soddisfare diversificati bisogni conoscitivi. Con qualche semplificazione, si possono riconoscere quattro essenziali esigenze cui l'indagine è oggi chiamata a rispondere: (i) quella, originaria, della tempestiva misura corrente dell'occupazione e della disoccupazione ad un livello piuttosto aggregato; (ii) quella, indiretta ma di notevole importanza, di supporto al sistema delle statistiche macroeconomiche, fornendo la base per la stima degli 'occupati presenti', cruciale per la valutazione del valore della produzione (vedi, ad es., Mamberti Pedullà, Pascarella e Abbate, 1987); (iii) quella dell'approfondimento del grado e delle modalità della partecipazione al lavoro dei singoli; (iv) infine, quella di un dettaglio territoriale via via maggiore, per il quale fornire attendibili stime correnti dell'occupazione e della disoccupazione.

Dilatazione degli obiettivi conoscitivi, evoluzione dei paradigmi interpretativi sul (e insieme trasformazioni del) mercato del lavoro, stimoli a miglioramenti metodologici hanno prodotto, già negli anni '70 e nei primi anni '80, numerose innovazioni nella RTFL[4]. D'altra parte, proprio per la sua centralità nel quadro delle rilevazioni sull'offerta di lavoro, dalla seconda metà degli anni '80 l'indagine si è venuta a trovare al cuore di un intenso impegno di ripensamento critico, ancor oggi tutt'altro che concluso, che investe tanto aspetti di metodo - come misurare i fenomeni di interesse - quanto aspetti sostanziali - che cosa misurare e secondo quali definizioni -[5].

Entro questo impegno di riflessione critica e di revisione innovativa, si colloca il progetto di ricerca "Forze di lavoro: disegno dell'indagine e analisi strutturali" (nel seguito designato con l'acronimo FOLA). Per mettere a fuoco le finalità della ricerca, è peraltro indispensabile richiamare brevemente le caratteristiche della RTFL e dar conto delle istanze, di affinamenti metodologici e di ulteriori conoscenze, con cui essa si confronta.

---

4 Richiamo soltanto le innovazioni di maggior rilievo. (i) Una prima ristrutturazione, a tutt'oggi la più importante, si ha nel 1977, sulla scorta del dibattito sull'occupazione 'sommersa' e la disoccupazione 'scoraggiata'. Vengono introdotte sostanziali revisioni nelle definizioni e nella struttura del questionario, al fine di "far rientrare esplicitamente nel campo di osservazione il lavoro a domicilio, il lavoro occasionale e marginale, il doppio lavoro, fino alle più piccole ed episodiche partecipazioni alla vita lavorativa" (Siesto, 1980, p. 57), ed inoltre di includere fra i disoccupati anche coloro che non hanno compiuto passi recenti di ricerca attiva di lavoro. (ii) Innestandosi sulla ristrutturazione del 1977, si avvia poi una revisione del programma di abbinamento dei dati individuali e della procedura di stima dei flussi, che sfocia nella pubblicazione di matrici di transizione complete, per intervalli trimestrali e annuali, a partire dal 1981 (Moriani, 1981). (iii) Sulla scorta della crescente richiesta dei poteri locali di poter trarre dall'indagine stime attendibili delle principali grandezze alla scala sub-regionale, a partire dal 1980 si avviano esperienze di ampliamento del campione, che si estendono e si consolidano sino ad attestare la dimensione campionaria della RTFL attorno alle 140.000 famiglie (a fronte delle 82.000 del disegno definito nel '77). (iv) Infine, nel 1984 la preoccupazione di rilevare in modo più articolato e affidabile modi e gradi della partecipazione al lavoro porta ad una apprezzabile modifica nella struttura del questionario, che da un formato a foglio unico per la famiglia, con una riga per persona, passa a fogli individuali.

5 Il fenomeno non è, di certo, peculiare all'indagine italiana. Soprattutto dalla fine degli anni '70, la sistematica rivisitazione dei dispositivi di rilevazione dell'occupazione e dell'offerta di lavoro ha coinvolto in maniera diffusa organismi internazionali, uffici statistici nazionali, studiosi. Il riesame ha talora riguardato l'intero sistema di informazioni sul lavoro (vedi, ad es., National Commission on Employment and Unemployment Statistics, 1979, per gli Stati Uniti e Malinvaud, 1986, per la Francia). Più spesso, si è incentrato proprio sulle indagini sulle forze di lavoro (vedi, ad es., ILO, 1983, per convenzioni definitorie; Singh e Drew, 1981, e Macredie, 1987, per la *Labour Force Survey* canadese; Butz e Plewes, 1989, per la *Current Population Survey*; Eurostat, 1985 e 1990, per l'indagine comunitaria).

2. *La rilevazione trimestrale delle forze di lavoro: disegno dell'indagine e fabbisogni conoscitivi*

2.1. *Alcuni richiami al disegno e alle caratteristiche dell'indagine*

Disegno e caratteristiche della RTFL sono analiticamente documentate in Istat (1978) e Fabbris e Bernardi (1986). Mi limito qui a scarne notazioni, funzionali alla presentazione del progetto FOLA e dei suoi principali risultati, ordinatamente su popolazione obiettivo, disegno campionario, questionario, operazioni sul campo, memorizzazione e controllo dei dati, procedure di stima degli aggregati[6].

La popolazione obiettivo della RTFL è costituita da tutti i componenti delle famiglie residenti nel Paese, comprese le persone temporaneamente emigrate all'estero. Operativamente, essa è identificata con le famiglie registrate presso l'anagrafe della popolazione. Ritardi e incompletezze nell'aggiornamento dell'anagrafe, e inoltre la regola di sostituire famiglie designate e non intervistate - per irreperibilità o rifiuto - con altre tratte da elenchi suppletivi (Istat, 1978, pp. 25-32), inducono un qualche scarto tra (teorica) popolazione obiettivo e (fattuale) popolazione da cui discende il campione di risposte valide. Dell'entità di questo scarto non si hanno evidenze. È peraltro ragionevole attendersi che, almeno per le stime *cross-section* di totali (e rapporti) a un elevato livello di aggregazione territoriale, esso sia trascurabile (vedi Fabbris e Bernardi, 1986, pp.3-6).

Il disegno campionario è 'complesso': a due stadi - il primo è lo stadio dei comuni, il secondo quello delle famiglie -, con stratificazione delle unità di primo stadio, selezione casuale dei comuni (salvi quelli con almeno 20.000 abitanti, detti 'autorappresentativi', sempre inclusi nel campione), selezione sistematica delle famiglie, rotazione delle unità campione sia di primo che di secondo stadio. La rotazione, introdotta essenzialmente per ragioni pratiche, come ragionevole compromesso fra le contrastanti esigenze di contenimento dei costi e di distribuzione dell'onere della rilevazione, è un tratto qualificante del disegno. A livello di primo stadio, i comuni non autorappresentativi sono rinnovati annualmente per un terzo, in occasione dell'indagine estiva. Nell'ambito di ogni comune campione, poi, le famiglie sono ruotate secondo uno schema 2-2-2, cioè a dire sono intervistate per due indagini consecutive, escono dal campione per le seguenti due, e vengono infine intervistate per due ultime indagini successive (vedi la Tab. 1). Di conseguenza, trascurando la parziale rotazione dei comuni e altri fenomeni di *attrition*, il 50% delle famiglie è comune in due rilevazioni successive così come in due rilevazioni svolte a distanza di un anno, ed il 25% è intervistato

---

6 Il riferimento è al disegno e alle caratteristiche della RTFL nel periodo 1984.I-1989.II, sul quale vertono i vari studi condotti nell'ambito del progetto FOLA. È da notare che si tratta di informazioni in parte obsolete. Col 1990.IV, infatti, sulla scorta anche di suggerimenti emersi dal progetto stesso, saranno sensibilmente modificate dimensione e strategia campionaria e sarà adottata una nuova procedura di stima degli aggregati (vedi Istat, 1989, e per brevi ragguagli la sez. 6.1).

Tab. 1: *Piano di rotazione delle unità di secondo stadio (famiglie) della RTFL*

Sezione (≡ gruppo di rotazione)	Sequenza delle rilevazioni			
	t.III	t.IV	t+1.I	t+1.II
A	X			
B	X	X		
C		X	X	
D			X	X
E	X			X
F	X	X		
G		X	X	
H			X	X

in quattro occasioni nell'arco di sedici mesi.

La dimensione del campione è cospicua. Già inizialmente ampio, perché definito per ottenere stime affidabili delle principali grandezze alla scala regionale, il campione è stato infatti progressivamente dilatato per rispondere alle richieste di numerose Regioni e Province di disporre di stime sufficentemente precise per domini sub-regionali. Alla fine degli anni '80, esso è così dell'ordine delle 140.000 famiglie.

Il questionario della RTFL si contraddistingue per la struttura modulare, a fogli individuali. Salvo un prospetto per la rilevazione di dati demografici per tutti i componenti la famiglia, infatti, per ciascuna persona di 14 anni o più è prevista la compilazione di un apposito foglio, a seguito dell'intervista dell'interessato (o di un *proxy*). Le definizioni accolte per individuare gli aggregati che descrivono la partecipazione al lavoro sono, di massima, conformi ai criteri suggeriti dall'ILO (Hussmanns, Merhan e Verma, 1990) e specificati per i Paesi della Comunità dall'Eurostat (1985). Le Figg. 1 e 2 evidenziano il ruolo dei diversi quesiti del questionario rispettivamente per la classificazione della popolazione negli usuali tre stati - occupato, in cerca di occupazione (o disoccupato in senso lato), non appartenente alle forze di lavoro (o inattivo) - e, con maggior dettaglio, per l'identificazione dei disoccupati. L'impianto delle definizioni risalta con sufficiente chiarezza, e non abbisogna di commenti. Piuttosto, merita di essere segnalato un paio di particolarità. In primo luogo, v'è da notare che l'intervista si apre con un quesito sulla 'condizione dichiarata', cioè a dire sulla percezione che l'intervistato ha della propria condizione. La strategia di interrogazione si discosta quindi dalla sequenza di quesiti 'oggettivi' tipica di questo genere di indagini, segnatamente per le domande iniziali sul lavoro. La condizione dichiarata non è, si badi, decisiva per l'insieme degli esiti classificatori; ha tuttavia

notevole importanza per l'identificazione degli occupati (vedi la Fig. 1), e indirettamente può avere qualche riflesso anche sugli altri due aggregati. In secondo luogo, uno dei criteri per identificare i disoccupati, quello della ricerca di lavoro, è interpretato in maniera parecchio più lasca di quanto non accada abitualmente nei Paesi sviluppati, nel senso che non viene posto alcun limite di prossimità temporale all'ultima azione di ricerca svolta (vedi la Fig. 2)[7].

Quanto alle concrete modalità di effettuazione dell'indagine, la RTFL è svolta essenzialmente per il tramite dei Comuni. Ciò comporta che il grado di governo e controllo delle operazioni sul campo da parte dell'Istat è forzatamente mediocre. Nebulosità e inconvenienti conseguenti a questo decentramento delle responsabilità operative emergono sin dalla fase di individuazione dei comuni chiamati a partecipare all'indagine: talvolta alcuni non collaborano, il che impone deformazioni 'ragionate' all'originario disegno campionario. Tali inconvenienti investono poi, in varia misura, l'estrazione del campione di famiglie, la selezione e formazione degli intervistatori, i controlli sullo svolgimento della rilevazione, i primi vagli sulla completezza e la congruenza delle informazioni rilevate (vedi Fabbris e Bernardi, 1986, pp. 18-28, e per un primo tentativo di delineare un *error profile* dell'indagine Masselli, 1989).

Previa una sommaria revisione manuale dei questionari presso il Reparto Forze di Lavoro dell'Istat, la registrazione dei dati su supporto magnetico è effettuata *in service* da una ditta specializzata, con controllo della qualità mediante doppia digitazione e prefissata frazione di errore tollerato. Infine, presso l'Istat i dati registrati sono sottoposti ad un complesso programma di controllo, che verifica il rispetto del campo di variazione di ciascuna variabile, vaglia la coerenza fra le risposte fornite a domande collegabili, procede infine all'imputazione di valori mancanti e alla sostituzione di dati errati (perché incompatibili con altri ritenuti più affidabili), secondo un protocollo deterministico o con una procedura *hot deck.*

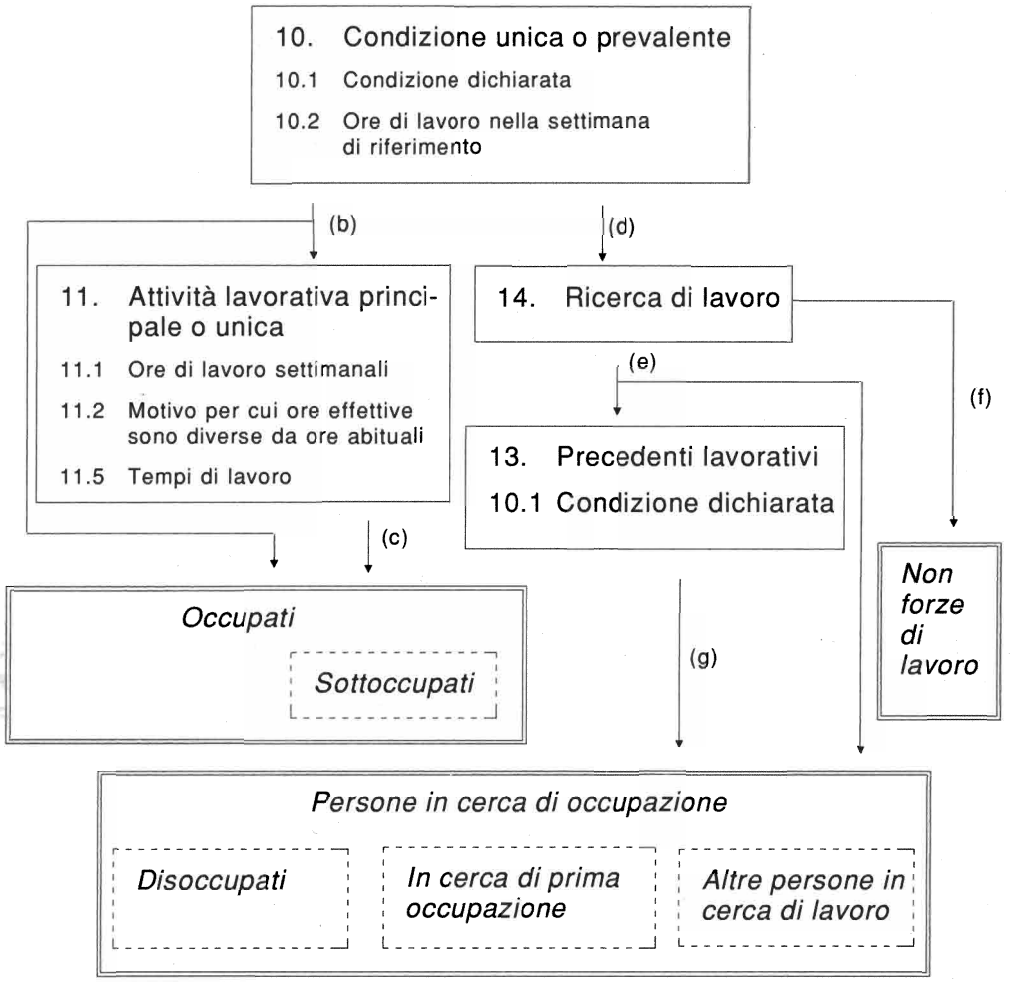
Sul *file* risultante da questo programma di controllo sono, infine, svolte le diverse elaborazioni che danno luogo alla produzione delle stime. Per le stime *cross-section* degli aggregati, di gran lunga le più note e utilizzate[8], è impiegato uno stimatore del rapporto post-stratificato per sesso: come coefficiente di espansione all'universo si utilizza, per l'appunto distintamente per maschi e femmine, il rapporto fra la popolazione residente nello strato alla data più recente e la popolazione campionata (per dettagli, vedi il cap. 2, sez. 2).

---

7 Su questo aspetto, per ora soltanto evocato, mi soffermo brevemente nella sez. 6.2.

8 Per la stima dei flussi, e di matrici di transizione, vedi Moriani (1981) e il cap. 18.

Fig. 1: *Ruolo dei blocchi di domande del questionario della RTFL nella classificazione della popolazione rispetto al lavoro* [a]



(a) I numeri rimandano alla numerazione delle domande nel questionario.

(b) Occupato: se risponde 'occupato' alla dom. 10.1 e/o 'si' alla dom. 10.2.

(c) Sottoccupato: se le ore di lavoro effettive sono inferiori a quelle abituali (dom. 11.1) e ciò dipende da carenza di domanda di lavoro (dom. 11.2), oppure se ha lavorato a tempo parziale perché non ha trovato un'occupazione a tempo pieno (dom. 11.5).

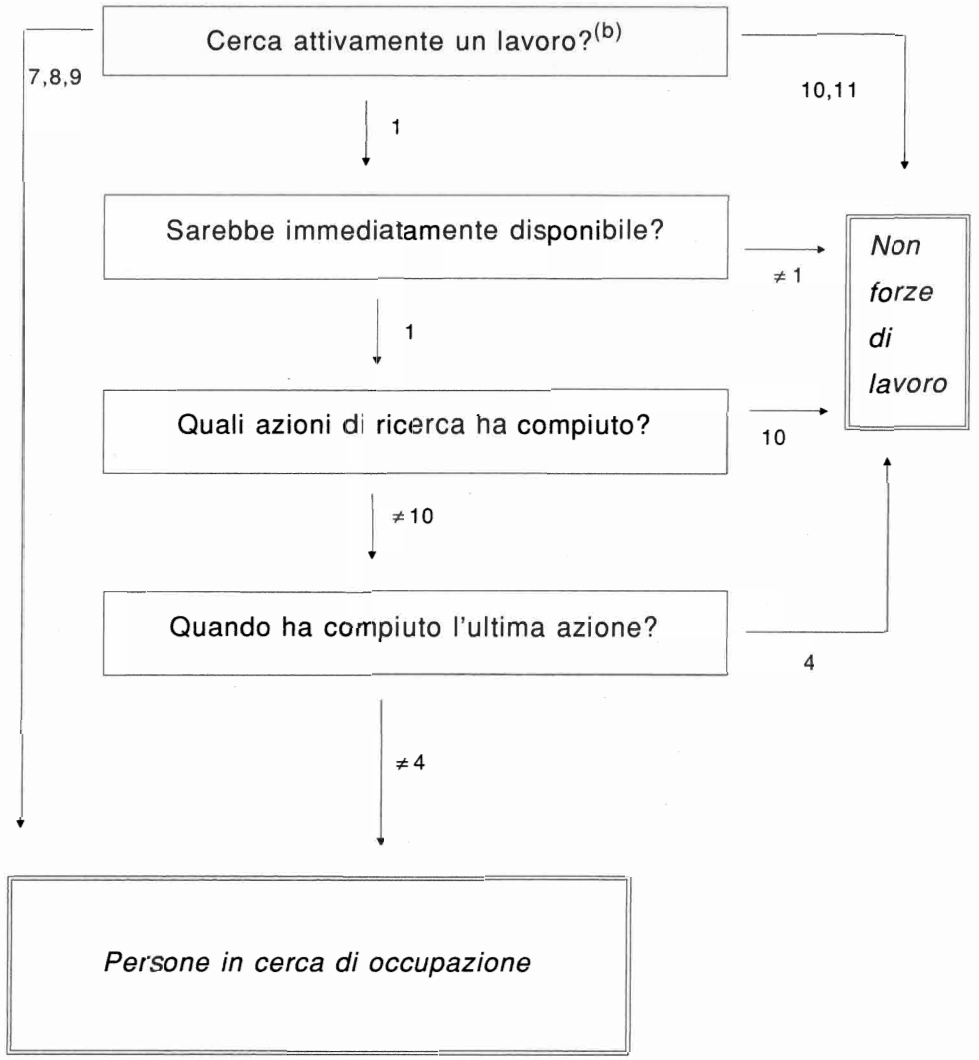
(d) ≠ (b).

(e) Per l'identificazione delle persone in cerca di occupazione, vedi la Fig. 2.

(f) ≠ (e).

(g) Disoccupato: se ha lasciato una precedente occupazione alle dipendenze per licenziamento, fine di lavoro a tempo determinato o dimissioni (dom.13). Altra persona in cerca di lavoro: se si era dichiarata in condizione non professionale (dom. 10.1). In cerca di prima occupazione: altrimenti.

Fig. 2: *Ruolo delle domande del questionario della RTFL nella definizione dell'aggregato delle persone in cerca di occupazione* [a]



(a) I numeri accanto alle frecce si riferiscono alle modalità di risposta alle domande, secondo la seguente legenda:
(i) "Cerca attivamente un lavoro?": 1=sì, alle dipendenze; 7=inizierà tra breve un lavoro alle dipendenze; 8,9=intende esercitare un lavoro in proprio; 10,11=no;
(ii) "Sarebbe immediatamente disponibile a lavorare?": 1=sì; ≠ 1=no;
(iii) "Quali azioni di ricerca ha compiuto?": ≠ 10=specifiche azioni di ricerca; 10=nessuna.
(iv) "Quando ha compiuto l'ultima azione?": 1=ultimi trenta giorni; 2=da uno a sei mesi fa; 3=oltre sei mesi fa; 4=non ha ancora iniziato.

(b) Le modalità di risposta da 2 a 6 e la 12 sono riservate agli occupati.

2.2. *Preoccupazioni metodologiche e fabbisogni conoscitivi*

Per dar conto delle riflessioni critiche sulla RTFL, e delle esigenze di affinamento e potenziamento che verso la metà degli anni '80 la investono, conviene prendere le mosse dal rapporto conclusivo della Commissione di studio dell'Istat per un sistema informativo del lavoro (Istat, 1984). Sulla scorta di un'approfondita ricognizione dello stato dell'informazione statistica in materia, confrontata criticamente con riscontrati o ragionevolmente prevedibili fabbisogni conoscitivi, tale rapporto delinea i tratti portanti di un "sistema informativo del lavoro", e in questo quadro definisce criteri ispiratori per riconsiderare il sub-sistema di rilevazioni sulle famiglie. Per quanto attiene specificamente alla RTFL, due sono i punti sui quali il rapporto ferma l'attenzione:

(a) la necessità di "dedicare notevole impegno all'affinamento dei metodi e delle tecniche di svolgimento" dell'indagine;
(b) l'opportunità di "articolare e, se necessario, differenziare gli strumenti di indagine in funzione degli obiettivi conoscitivi" (Istat, 1984, p. 55).

La sollecitazione ad affinamenti sul versante metodologico è a tutto campo. V'è da considerare, infatti, che, pur con aggiornamenti non trascurabili (peraltro, incentrati soprattutto sul questionario e sulla dimensione del campione), l'impianto dell'indagine è rimasto sostanzialmente immutato dal 1959. D'altra parte, i metodi di progettazione e conduzione di indagini campionarie, così come quelli di analisi dei dati, hanno conosciuto in questi decenni notevoli sviluppi. Di conseguenza, si fa palese l'obsolescenza di talune scelte, ad esempio in tema di disegno campionario e di procedura di stima degli aggregati. Inoltre, emerge con crescente chiarezza che il mediocre standard delle operazioni sul campo ha severe implicazioni sulla qualità dei dati. Infine, si tocca con mano che alcune delle stesse innovazioni apportate all'indagine producono effetti indesiderabili: così, la dilatazione del campione acuisce i problemi, già annosi, di qualità dei dati e di tempestività delle stime.

Citando ancora, liberamente, dal rapporto della Commissione (Istat, 1984, p. 58), l'impegno sull'affinamento dei metodi si prospetta perciò in tre direzioni:

(a) quella del disegno dell'indagine (disegno campionario, struttura del questionario e *question wording*, ecc.);
(b) quella delle modalità di svolgimento delle operazioni sul campo (sperimentazione di differenti modalità di selezione/addestramento/ supervisione dei rilevatori, misura dell'errore di risposta e del rilevatore, ecc.);
(c) quella del migliore sfruttamento delle informazioni raccolte (affinamenti nelle procedure di stima degli aggregati, ricostruzione di informazioni longitudinali basate sulle risposte fornite dagli individui nelle quattro occasioni di indagine cui partecipano, ecc.).

L'emergere di nuove preoccupazioni conoscitive è l'altro fattore che sta alla base della riconsiderazione della RTFL, e più in generale dell'insieme degli strumenti di indagine sull'offerta di lavoro. Per cogliere la direzione e la portata di queste sollecitazioni, occorre peraltro richiamare, sia pure in

estrema sintesi, le profonde trasformazioni che interessano il mercato del lavoro italiano, e il riesame dei paradigmi interpretativi che ad esse si accompagna (per maggiori dettagli, vedi, tra gli altri, Bruno, 1987). (i) Dal lato dell'offerta, si assiste ad un sensibile aumento della propensione al lavoro delle donne, e soprattutto alla crescente importanza del ruolo decisionale di individui e famiglie nel concorrere a determinare l'assetto del mercato del lavoro. La riabilitazione dell'offerta di lavoro operata, sul terreno dei paradigmi interpretativi, dalla *new home economics* è emblematica al riguardo (vedi, ad es., Colombino, 1979). (ii) Dal lato della domanda, il processo di terziarizzazione e la cosiddetta 'rivoluzione tecnologica' comportano un dinamismo ed una differenziazione crescenti nella quantità e nella qualità delle prestazioni lavorative richieste. (iii) Infine, l'aumento degli squilibri nel mercato del lavoro indotto dai fattori or ora sommariamente richiamati, e da più generali dinamiche dello sviluppo economico, induce una dilatazione delle politiche attive del lavoro, nonché nuove caratteristiche nelle stesse - più marcatamente dettate da preoccupazioni distributive ed orientate a gruppi particolari di soggetti -.

Ricomporre le implicazioni di queste trasformazioni, e dei riflessi che esse hanno sul piano degli schemi interpretativi, nella prospettiva di un'indicazione ragionata dei fabbisogni informativi è operazione complessa, anche per le caratteristiche di variabilità e di dinamismo che contraddistinguono le trasformazioni in atto. Non è forse azzardato, tuttavia, trarne alcune indicazioni di larga massima.

In primo luogo, emerge la consapevolezza di una maggiore problematicità nella definizione dei confini fra gli stati - occupato, disoccupato, inattivo -, e l'esigenza di nuovi concetti e di nuove classificazioni, capaci di cogliere con maggiore fedeltà e in modo più articolato grado e caratteristiche della partecipazione al lavoro. Tale esigenza si scontra peraltro con due ordini di questioni: da un lato l'assenza di un quadro teorico sufficientemente condiviso; dall'altro lato l'esistenza di ovvie esigenze di comparabilità, temporale e spaziale, delle misure dell'occupazione e della disoccupazione. La via d'uscita che pare profilarsi è largamente pragmatica e si muove essenzialmente lungo due direttrici: (i) per un verso, a fini di univoca misurazione dell'occupazione e della disoccupazione, soprattutto per plausibili confronti temporali e spaziali, viene accentuato il ruolo svolto da convenzioni definitorie assai circostanziate, preferibilmente condivise a livello internazionale; (ii) per un altro verso, si attribuisce notevole importanza a strategie di rilevazione che permettano flessibilità in sede di costruzione di aggregati: che consentano cioè di evidenziare sottoinsiemi di interesse di un dato aggregato, così come di costruire aggregati aggiungendo o togliendo particolari sottoinsiemi, in relazione a varianti definitorie ritenute pertinenti a fini di analisi e/o di politiche (vedi Shiskin, 1976, e ILO, 1980).

In secondo luogo, si tende a porre l'accento sulle informazioni statistiche riguardanti i *flussi:* di occupazione (fra settori e/o imprese), dalla condizione di disoccupato a quella di occupato e viceversa, dalla condizione di inattivo a quella di attivo e viceversa, ecc.. L'importanza di informazioni sui flussi della popolazione rispetto al lavoro è ormai così generalmente condivisa che

neppur serve soffermarvisi, se non per segnalare che è accentuata dalla stessa polarizzazione delle politiche attive del lavoro su persone in fasi di transizione (giovani che entrano nella vita attiva, occupati che perdono o cambiano lavoro, lavoratori anziani avviati al pre-pensionamento, ecc.).

In terzo luogo, per studiare il comportamento dei soggetti - segnatamente, individui e famiglie - sul mercato del lavoro, e in un'ottica di *policy* per rispondere a questioni circa il disegno degli interventi e la valutazione dei loro effetti, appare indispensabile disporre di informazioni statistiche a livello 'micro', possibilmente longitudinali. Il vaglio degli effetti di politiche del lavoro, rivolte a specifici gruppi di soggetti col proposito di indurre modifiche in una direzione desiderata nella loro condizione e nel loro comportamento, richiede infatti un supporto conoscitivo affatto differente da quello tradizionale, costituito da dati aggregati. A tale scopo, è necessario seguire i soggetti nel corso del tempo, cioè a dire raccogliere dati individuali longitudinali, a monte e a valle dell'intervento. E le modalità di raccolta dei dati debbono essere particolarmente curate, ed i modelli e metodi di analisi adeguatamente affinati, per poter discernere lo specifico effetto dell'intervento dai molteplici fattori individuali e sociali che concorrono a determinare lo stato del soggetto e che condizionano i successivi comportamenti (vedi, tra gli altri, Fienberg, Singer e Tanur, 1985, e Heckman e Hotz, 1989).

Muovendo da queste indicazioni sui fabbisogni informativi, ben si comprende l'enfasi posta in Istat (1984, pp. 55-56) sull'"articolazione degli strumenti di indagine". Restano peraltro cruciali interrogativi sulla fattibilità dei suggerimenti che accompagnano e specificano questo orientamento[9], e soprattutto sui modi per renderli operativi. È plausibile l'ipotesi di 'microcensimenti', a cadenza triennale o giù di lì, per disporre di informazioni disaggregate - quanto a territorio, branca, professione -? In che misura, e come, è praticabile la proposta di coordinare alla RTFL indagini saltuarie volte a investigare temi collaterali? E in ogni caso come va affrontata la revisione del disegno della RTFL, nel suo impianto complessivo e specificamente negli aspetti di contenuto e di struttura del questionario?

3. *Il progetto FOLA: motivazioni e obiettivi*

Sulle questioni, metodologiche e sostanziali, che ho brevemente richiamato, nel 1985 prendono avvio molteplici programmi di attività[10]. Tra questi, il progetto di ricerca FOLA si qualifica per l'impegno prevalente di studiosi

---

9 Le ipotesi prospettate nel rapporto conclusivo della Commissione sono, in sintesi, le seguenti: (i) 'microcensimenti' a cadenza di 3-5 anni, per disporre di informazioni disaggregate (quanto a territorio, branca, professione, ecc.); (ii) finalizzazione della RTFL alla stima corrente dell'occupazione e della disoccupazione (e degli aspetti salienti delle modalità e del grado della partecipazione al lavoro) alla scale nazionale e regionale; (iii) recupero del potenziale informativo delle fonti amministrative per documentare la dinamica congiunturale dell'occupazione e della disoccupazione a scale sub-regionali; (iv) potenziamento della pratica di coordinare alla RTFL indagini saltuarie volte a investigare temi collaterali.

10 Oltre all'attività che fa capo al Servizio Indagini sulle Famiglie, mi riferisco al 'Progetto campioni' e al 'Progetto qualità dei dati' (vedi, rispettivamente, Istat, 1985a, e Masselli, 1985).

universitari, per gli scopi essenzialmente analitici, per la selezione di uno spettro sì ampio, ma insieme chiaramente circoscritto di tematiche. Si tratta, del resto, di aspetti connessi l'un l'altro. Il fatto che la ricerca si svolga essenzialmente in sede universitaria, infatti, non può che portare, per naturale interesse degli studiosi coinvolti e per l'ovvia distanza dal contesto delle problematiche operative dell'Istat, all'accentuazione di scopi analitici, restando sullo sfondo preoccupazioni e implicazioni immediatamente progettuali. D'altronde, la stessa selezione delle tematiche risulta condizionata dal fatto che l'attività di ricerca si colloca per larghissima parte a valle dello svolgimento delle indagini, insiste cioè sulle informazioni finali che esse producono (essendo indisponibili, invece, quelle attinenti al processo di rilevazione).

Il progetto muove dalla constatazione che due tratti dell'impianto della RTFL sono sfruttati solo in parte. Ciò vale in primo luogo per il disegno campionario complesso, segnatamente per il piano di rotazione del campione e per la conseguente possibilità di creare *files* di dati longitudinali mediante abbinamento di *records* di individui che permangono in indagini successive. In secondo luogo, sono largamente inesplorate le potenzialità di analisi di dati micro: dati individuali *cross-section*, innanzitutto; inoltre, dati a livello di famiglia (o comunque informazioni sugli altri membri della famiglia collegabili a quelle dell'individuo); infine, a seguito di procedure di abbinamento, dati longitudinali.

Il progetto FOLA mira per l'appunto ad esplorare alcune opportunità di più piena utilizzazione dell'indagine consentite da questi due tratti, rispetto a due ordini di finalità:

(a) *vaglio della qualità dell'indagine e miglioramento delle stime delle variabili di interesse*, obiettivo questo collegato prevalentemente alla struttura complessa del disegno campionario e alla presenza di rotazione;

(b) *completamento della documentazione e svolgimento di analisi sulle caratteristiche strutturali e la dinamica di breve periodo del mercato del lavoro*, obiettivo questo collegato prevalentemente alla possibilità di elaborazione di dati individuali, di persone e/o famiglie, *cross-section* e longitudinali.

Avendo come riferimento queste ampie finalità, sono stati messi a fuoco dodici specifici temi. Sono questi temi che costituiscono l'ossatura del progetto, tanto dal punto di vista scientifico - degli obiettivi e dei metodi di analisi - che da quello operativo - dell'articolazione del gruppo di ricerca - (vedi Trivellato, Bernardi e Fabbris, 1987). Una sinossi dei temi specifici è nella Tab. 2. In essa si distingue, tra l'altro, fra temi (e sono la gran parte) che poggiano sulla documentazione statistica esistente - i dati pubblicati e/o quelli elementari della RTFL - e temi che hanno richiesto lo svolgimento di indagini suppletive. Per la sua ovvia importanza operativa, questo criterio di classificazione affianca utilmente quello costituito dalle finalità, e concorre a fornire una prima, sommaria guida alle tematiche affrontate.

Tab. 2: *Articolazione del progetto di ricerca FOLA in temi specifici, secondo le finalità e le informazioni richieste*[a]

Finalità	Informazioni richieste	
	Fonti esistenti	Indagini suppletive
(a) Vaglio della qualità dell'indagine e miglioramento delle stime delle variabili	(a1) Effetto del disegno di campionamento sulle stime *[2, 3]* (a2) Problemi di tempestività delle stime e di sovracampionamento *[4, 5]* (a3) Stime efficienti con campione ruotato *[6]* (a4) Studio di errori rilevabili da confronti in indagini successive *[8, 10]* (a5) Verifica di presenza di *rotation group bias* [9]	(a6) Controlli sulla qualità mediante indagini suppletive *[21]*
(b) Completamento della documentazione e analisi di caratteristiche strutturali e dinamiche del mercato del lavoro	(b1) Analisi di matrici di dati *[11, 12]* (b2) Analisi esplorative su singoli e famiglie rispetto al lavoro *[13, 14, 15]* (b3) Sviluppi nell'analisi di serie storiche del mercato del lavoro *[16, 17]* (b4) *Linkage* di unità in indagini successive e analisi di mobilità [7, 18] (b5) Analisi del comportamento rispetto al lavoro mediante modelli strutturali *[19, 20]*	(b6) Acquisizione di informazioni aggiuntive sulla storia lavorativa *[22]*

(a) I numeri in corsivo indicano i capitoli in cui il tema, o un suo particolare aspetto, è trattato.

4. *Il progetto FOLA: un filo di Arianna per leggerne i risultati*

A dar conto dei risultati del progetto è l'intero volume. Quel che mi propongo nelle pagine che seguono è, semplicemente, delinearne la trama: richiamando lo stato dell'arte da cui si muove; accennando ai diversi contributi, e soprattutto mettendo in luce le connessioni fra gli stessi; segnalando gli interrogativi che tuttora si pongono. Una sorta di filo di Arianna, dunque, che orienti alla lettura.

4.1. *Disegno campionario e stime*

Un primo insieme di contributi si struttura attorno a una tematica classica, quella delle stime e della loro variabilità, in presenza di disegno campionario complesso.

Ad aspetti di tradizionale interesse, rispettivamente l'errore di campionamento e la procedura di post-stratificazione, si riferiscono i capp. 2 e 3. Nel cap. 2, Coccia, Falorsi e Russo portano a conclusione un fecondo impegno in tema di stima della varianza campionaria e dell'effetto complessivo del disegno di campionamento (e delle sue componenti). Insieme con la compatta presentazione del metodo proposto per il calcolo approssimato di tali grandezze, essi forniscono, in via illustrativa, l'errore di campionamento percentuale e l'indice dell'effetto complessivo del disegno di campionamento per le principali stime *cross-section*, regionali e provinciali, della RTFL del 1986.IV.

Nel cap. 3, Ghellini vaglia le implicazioni della post-stratificazione per sesso sulla stima della distribuzione per età della popolazione, confrontata con quella desumibile da fonti esogene. Egli fornisce chiare evidenze di sistematica sottostima delle classi di età giovani e di sovrastima di quelle anziane, e degli effetti che, data la correlazione fra partecipazione al lavoro ed età, ne conseguono in termini di distorsione verso il basso delle stime dell'occupazione e della disoccupazione. Il suggerimento che Ghellini avanza, ed esemplifica con un'applicazione alla Lombardia, è naturale: quando si disponga - ed è per l'appunto questo il caso - di un'affidabile stima esogena della distribuzione della popolazione per età, conviene usare uno stimatore del rapporto stratificato congiuntamente per sesso ed età[11].

Con i successivi due capitoli, l'attenzione si sposta su una questione

---

11 La post-stratificazione potrebbe avvenire simultaneamente per sesso ed età, se si disponesse della distribuzione congiunta, esogena, a livello di strato. Così non è, sicché Ghellini usa un procedimento di post-stratificazione a due passi: prima per sesso, a livello di strato; poi per età, a livello regionale. Un possibile miglioramento può venire da procedure iterative, del tipo *raking ratio* (vedi, ad es., Choudhry e Hidiroglou, 1987). Dall'analisi del *pattern* della distorsione nella stima della struttura per età vengono poi altri spunti di interesse: la distorsione aumenta quanto più ci si allontana dal *benchmark* censuario, plausibilmente perché la stratificazione si fa via via più obsoleta (sicché, in prospettiva, potrebbe essere conveniente adottare criteri di stratificazione aggiornabili correntemente); il peso degli errori non campionari, in particolare l'effetto della sostituzione di famiglie, è verosimilmente tutt'altro che trascurabile (sicché il miglioramento delle modalità di conduzione dell'indagine potrebbe avere un apprezzabile impatto anche a questo riguardo).

puntuale, di notevole interesse sia metodologico che pratico: la tensione fra dilatazione della dimensione campionaria e tempestività delle stime. Anche a seguito del progressivo incremento della dimensione del campione, infatti, i tempi di raccolta ed elaborazione dei dati si sono parecchio allungati. Le stime sono ormai disponibili soltanto 100 giorni dopo la data di (teorica) conclusione della rilevazione. Questa cadenza è chiaramente insoddisfacente per tempestive analisi congiunturali della partecipazione al lavoro, così come per l'uso dei dati della RTFL per la stima corrente dei conti economici trimestrali. Le possibili linee di attacco alla questione sono due: (i) l'impiego di stimatori per piccole aree, che consentano di valutare con ragionevole margine di errore le grandezze di interesse per domini sub-regionali, con un campione di dimensione contenuta; (ii) l'elaborazione di stime precoci delle principali grandezze a livello nazionale, mantenendo nel contempo elevata la numerosità campionaria a fini di successive analisi disaggregate territorialmente.

Nel cap. 4, Fabbris, Falorsi e Russo presentano i risultati di un primo esperimento, in cui mettono a confronto l'usuale stimatore del rapporto post-stratificato applicato ad un ipotetico campione ampliato ed uno stimatore sintetico, che combina l'informazione campionaria con quella desumibile da fonti esogene, applicato al corrispondente campione di base. L'esame delle *performances* dei due stimatori, in termini di varianza e di errore quadratico medio, documenta risultati piuttosto buoni per lo stimatore sintetico.

Corradi *et al.*, d'altra parte, nel cap. 5 confrontano le *performances* di stime ottenute a diversa distanza temporale dalla (teorica) conclusione della rilevazione, sino al momento di completa disponibilità dei dati raccolti. L'obiettivo è di vagliare la possibilità di giungere ad affidabili stime precoci, tramite opportuni stimatori che utilizzano i parziali dati disponibili e/o tramite stimatori composti, che sfruttano cioè anche le informazioni (complete) relative alla precedente occasione. L'impiego di stimatori composti impone naturalmente di tener conto della struttura rotante del campione. Ciò è ad un tempo metodologicamente non banale e operativamente laborioso. Il miglioramento della precisione delle stime appare tuttavia, dall'analisi empirica condotta sulla variabile 'numero di persone in cerca di occupazione', apprezzabile.

Con il saggio di Cocchi (cap. 6), l'interesse viene portato specificamente sulle caratteristiche dinamiche del campione. Attualmente, nella determinazione delle stime delle singole variabili l'Istat non tiene conto del fatto che alcune unità campionarie erano presenti anche in rilevazioni precedenti, sicché nel campione si distinguono una frazione sovrapposta (a sua volta, distinta in sub-frazioni presenti nel campione per un diverso numero di occasioni) e una frazione non sovrapposta. Se per la frazione sovrapposta si sfruttassero le informazioni delle precedenti occasioni, sarebbe possibile ottenere stime efficienti delle variazioni da un'occasione alla successiva e di medie complessive su più occasioni. Inoltre, tenendo conto della rotazione si può migliorare anche l'efficienza delle stime degli aggregati per l'ultima occasione, sempre mediante stimatori composti che utilizzano le informazioni addizionali sulla parte sovrapposta relative alle precedenti occasioni (Pat-

terson, 1950; National Commission on Employment and Unemployment Statistics, 1979). Il tema è difficile, e soluzioni analitiche agevoli si danno per disegni di rotazione piuttosto semplici, non certo per un disegno campionario complesso quale quello della RTFL. Cocchi affronta l'argomento mescolando l'attenzione ad aspetti squisitamente metodologici - la ricerca del peso ottimo per lo stimatore composto, la determinazione della varianza dello stesso - con la preoccupazione di fornire soluzioni agevolmente praticabili. I risultati che ottiene testimoniano chiaramente i guadagni in efficienza relativa degli stimatori composti rispetto a quelli tradizionali, tanto per le stime dei livelli quanto, e ancor più, per quelle delle differenze.

4.2. *Abbinamento longitudinale e qualità dei dati*

I contributi che confluiscono nella terza parte, così come altri nel seguito, poggiano su uno zoccolo che rimanda ancora alle caratteristiche dinamiche del campione: l'abbinamento dei *records* individuali di successive indagini della RTFL. Esso costituisce, infatti, l'indispensabile pre-requisito sia per analisi della qualità dei dati che sfruttano, in varia misura, la struttura rotante del campione (capp. 8-10), sia per studi di mobilità e per modelli dinamici dell'offerta di lavoro a livello micro (capp. 18 e 19)[12].

Giusti, Marliani e Torelli hanno dedicato all'abbinamento longitudinale dei *records* della RTFL un cospicuo impegno, del quale il cap. 7 riassume approcci e risultati. La messa a punto di una procedura di abbinamento esatto basata su criteri probabilistici[13] è analiticamente ingegnosa. E quel che più conta, sono convincenti i risultati cui approda. Vi è, infatti, un apprezzabile guadagno nella quota di abbinamento rispetto a procedure euristiche, in particolare rispetto a quella utilizzata dall'Istat dal 1981. Evidenze indirette, ma concordi, inducono inoltre a ritenere che tale guadagno non sia spurio (dovuto cioè a errati abbinamenti), bensì recuperi situazioni che nelle procedure euristiche si risolvono in 'falsi negativi' (cioè in mancati abbinamenti di *records* che sono relativi alla stessa unità). La procedura proposta, infine, per le sua stessa caratteristica di variare le regole per

---

12 È da notare che, nel procedere della ricerca, il tema dell'abbinamento longitudinale di *records* individuali è venuto assumendo un ruolo ben maggiore di quanto ipotizzato inizialmente. Nel progetto originario, infatti, l'interesse al tema era circoscritto all'estensione a quattro occasioni della procedura di *linkage* approntata dall'Istat nel 1981, nell'implicito presupposto che essa fosse soddisfacente per due occasioni. Il presupposto si è però rivelato in parte erroneo: la procedura in questione non è scevra da inconvenienti; con procedure alternative si conseguono apprezzabili miglioramenti (vedi il cap. 7). Ciò ha portato ad un maggior impegno sullo specifico argomento, e - quel che più interessa - ad accentuare le interrelazioni fra la questione dell'abbinamento longitudinale e molteplici altri temi di ricerca, primi fra tutti quelli sulla qualità dei dati. Questo assestamento si riflette nell'organizzazione stessa del volume, che vede la trattazione dell'abbinamento longitudinale anticipata al cap. 7.

13 La terminologia è solo apparentemente contraddittoria. Si parla, infatti, di abbinamento 'esatto' quando l'obiettivo è di collegare informazioni pertinenti alla stessa unità, contrapponendolo all'abbinamento 'statistico' che ha per obiettivo collegare informazioni di unità simili rispetto ad un qualche criterio (Jabine e Scheuren, 1986). È del tutto naturale, poi, affrontare il problema dell'abbinamento esatto con procedure che abbiano fondamenti e usino criteri probabilistici, nell'ambito della cosiddetta *theory of exact linkage* (Fellegi e Sunter, 1969).

l'abbinamento in relazione alla configurazione degli archivi coinvolti è in grado di neutralizzare eventuali anomalie presenti nei dati, 'imparando' dai dati stessi ad assegnare un diverso potere discriminante alle variabili.

Degli errori non campionari che possono essere evidenziati sfruttando la struttura longitudinale dei dati, nel cap. 8 Giommi affronta le due manifestazioni basilari: la mancata risposta longitudinale, che si ha quando un individuo appartenente al *panel* partecipa ad almeno una, ma non a tutte le occasioni di indagine a cui avrebbe dovuto partecipare; le incompatibilità che si riscontrano fra risposte date dallo stesso individuo in tempi diversi. Le ampie e accurate analisi empiriche che presenta forniscono importanti lumi sulla qualità dei dati. Una più penetrante individuazione delle fonti di errore non campionario è peraltro preclusa dal fatto di condurre l'analisi interamente a valle del processo di raccolta e *editing* dei dati: errori di risposta, errori di registrazione, effetti del programma di controllo e imputazione, eventuali errori associati alla stessa procedura di abbinamento longitudinale si combinano, sicché è praticamente impossibile discernerne l'apporto. Significativamente, il saggio di Giommi si conclude con suggerimenti per migliorare la conduzione della RTFL, nell'ottica appunto di un più accurato controllo delle fonti di errore non campionario tramite una maggiore utilizzazione delle informazioni longitudinali.

I capp. 9 e 10 vertono su due specifiche manifestazioni, e fonti, di errore non campionario connesse alla struttura rotante del campione: il cosiddetto *rotation group bias* (cioè a dire, la distorsione nelle stime di livello di alcune variabili causata dal condizionamento che gli individui subiscono in seguito alla prolungata permanenza nel campione); le imprecisioni, dovute a effetti di memoria, nella durata della permanenza in uno stato rilevata tramite quesiti retrospettivi.

La presenza di *rotation group bias* è stata chiaramente documentata nella *Current Population Survey* per una variabile cruciale, il tasso di disoccupazione (Bailar, 1975; Shack-Marquez, 1985). Nel cap. 9, Alleva si propone di verificare se anche per alcune variabili della RTFL si riscontra un analogo fenomeno distorsivo. L'analisi non è limitata al confronto delle stime riferite alle quattro sezioni presenti nel campione da un diverso numero di occasioni. Per eliminare fattori perturbatori ed isolare il fenomeno di interesse, essa viene sviluppata anche con riguardo a gruppi di individui identificati sulla base dell'effettiva, differente anzianità nel campione (per l'appunto, tramite la procedura di abbinamento longitudinale). Alleva giunge a concludere che per la RTFL non vi è evidenza di *rotation group bias*. Quanto ciò sia dovuto all'assenza di *panel conditioning*, verosimilmente per il modesto numero di volte in cui gli individui permangono nel campione e per la distanza temporale tra le interviste, e quanto invece alla difficoltà di cogliere il fenomeno, tutto sommato sottile, in una rilevazione affetta da molteplici e apprezzabili errori non campionari, è questione che rimane aperta.

La durata della disoccupazione, rilevata nella RTFL tramite la domanda retrospettiva "Da quanti mesi è alla ricerca di occupazione?", è l'oggetto dell'analisi di accuratezza condotta, nel cap. 10, da Torelli. La letteratura sugli errori del processo di memoria suggerisce due potenziali fattori distor-

sivi: (i) l'effetto ammucchiamento (*heaping* o *digit preference*, nell'originaria terminologia anglosassone), il quale induce un'abnorme polarizzazione delle risposte su particolari durate, in buona sostanza perché il rispondente usa una scala di misura più rozza di quella adottata nel quesito; (ii) l'effetto telescopio, tipicamente in avanti, il quale porta a collocare ad una data più prossima a quella dell'indagine, rispetto alla vera, l'inizio dell'episodio di disoccupazione. Per investigare questi aspetti, Torelli utilizza sia le informazioni sulla durata riportata della disoccupazione tratte da una singola rilevazione, sia gli analoghi dati abbinati di due rilevazioni successive. Egli documenta in maniera chiara l'incidenza dei due fenomeni distorsivi e ne esplora le interrelazioni. In particolare, l'effetto *heaping*, con la tendenza a riportare durate approssimate in anni e semestri, risulta assai pronunciato, e suggerisce opportune cautele nell'utilizzazione dei dati sulla durata della disoccupazione.

4.3. *Analisi esplorative multivariate*

Con I saggi della quarta e della quinta parte, l'interesse si sposta sul secondo obiettivo del progetto FOLA: una più ampia utilizzazione delle informazioni raccolte con la RTFL, a fini di analisi delle caratteristiche strutturali e della dinamica di breve periodo della partecipazione al lavoro. Alle preoccupazioni per sviluppi metodologici si accompagnano dunque spiccati interessi sostanziali. Le analisi in chiave prevalentemente esplorativa trovano posto nella quarta parte, mentre i contributi con un taglio prevalentemente modellistico-confermativo sono raccolti nella quinta parte[14].

I primi due capitoli della quarta parte si segnalano per una marcata attenzione a questioni di metodo. Nel cap. 11, Lovison si sofferma sulle possibilità di sfruttare in modo più approfondito le informazioni rilevate con la RTFL, quelle elementari e/o le stesse tabelle pubblicate, per esplorare le relazioni fra variabili potenzialmente esplicative del comportamento sul mercato del lavoro e variabili che misurano appunto modalità e intensità di tale comportamento. Essendo la maggioranza delle variabili rilevate nominali, il contesto di riferimento è l'analisi di tabelle di frequenze, l'obiettivo la ricerca di modelli log-lineari parsimoniosi, lo scoglio con cui misurarsi il disegno campionario complesso della RTFL. Alla rassegna critica dei metodi proposti in letteratura fa seguito un vaglio empirico comparato delle *performances* di alcuni di essi, condotto su otto tabelle sulle forze di lavoro pubblicate correntemente. Lovison ne trae un suggerimento persuasivo: la pubblicazione da parte dell'Istat di un limitato insieme di informazioni aggiuntive - i *deff* di varianza - consentirebbe di migliorare di molto le possibilità di impiego dei dati pubblicati a fini di studio delle associazioni fra variabili.

Bolasco e Coppi (cap. 12) muovono da una preoccupazione metodologica

---

14 La distinzione va presa *cum granu salis*. Analisi esplorativa e analisi confermativa definiscono i due archetipi di strategie e metodi di ricerca che, nella realtà, si collocano in un *continuum*.

analoga, estrarre l'informazione essenziale da un tabella di dati multidimensionale, ma la sviluppano in una direzione affatto diversa. Trascurano il disegno campionario, e si concentrano piuttosto su metodi fattoriali a più indici e su una strategia per il loro uso sequenziale "capace di produrre un'immagine pregnante, ancorché semplificata, dell'informazione nascosta in un *array* complesso di dati". Utilizzano poi questo apparato metodologico per analisi esplorative sulla struttura e la dinamica dell'occupazione, condotte su una complessa tabella a più dimensioni.

A interrogativi più dichiaratamente sostanziali mirano a rispondere i successivi capp. 13-15. Un primo polo di interesse è costituito dalle opportunità che la RTFL offre di evidenziare tipologie familiari e di studiarne le forme di partecipazione al lavoro. Naturalmente, occorre scontare i limiti, piuttosto severi, delle tradizionali informazioni demografiche e sui rapporti di parentela acquisite con la RTFL. D'altra parte, essa presenta il non trascurabile vantaggio di essere un'indagine corrente a frequenza alta, e ben si presta quindi per tempestivi aggiornamenti di un quadro conoscitivo, pur sommario, su strutture e comportamenti familiari[15]. Nel cap. 13, Ongaro utilizza con acume i dati della RTFL per giungere, tramite un'opportuna concatenazione di tecniche di analisi esplorativa, ad una classificazione delle famiglie con giovani adulti, indicativa, per le associazioni di variabili che caratterizzano i diversi gruppi, delle stesse modalità di aggregazione familiare. Nel cap. 14, Sanna, Santini e Lauro affrontano la tematica delle connessioni fra forme familiari e caratteristiche della partecipazione al lavoro: se la complessità dell'argomento impone all'analisi di restare ad uno stadio di prima esplorazione, le evidenze empiriche che vengono fornite sono tuttavia numerose e ricche di suggestioni interpretative.

L'altro polo di interesse sostanziale è costituito da una tematica piuttosto controversa, sulla quale tornerò ancora nel seguito (vedi la sez. 6.2): l'identificazione dei disoccupati e la misura dell'*attachment* al mercato del lavoro. Rettore, Torelli e Trivellato (cap. 15) affrontano la questione conducendo un'analisi di classificazione esplorativa sullo stesso insieme di domande utilizzato dall'Istat per definire, con la classificazione *a priori* riassunta graficamente nella Fig. 2, l'aggregato dei disoccupati. L'esito saliente è l'individuazione di due distinti gruppi di non occupati che cercano lavoro: l'uno identificabile con il 'nucleo forte' della disoccupazione, l'altro con caratteristiche di *attachment* chiaramente meno pronunciate. Sulla classificazione risultante dall'analisi esplorativa essi innestano poi un esame dei flussi di mobilità trimestrale, il quale pare confermare la rilevanza della distinzione fra i due gruppi di non occupati in cerca di lavoro per cogliere, e in definitiva per predire, la dinamica di breve periodo dei comportamenti nel mercato del lavoro.

---

15 Resta invece problematico, se non addirittura proibitivo, sfruttare la struttura longitudinale della RTFL per studiare la dinamica di breve periodo della famiglia e dei suoi comportamenti. Per una lucida disamina delle difficoltà che presenta la stessa definizione e osservazione 'longitudinale' della famiglia, vedi Duncan e Hill (1985).

4.4. *Modelli di analisi delle forze di lavoro*

Dei contributi in tema di modelli del mercato del lavoro si può tracciare una sommaria articolazione guardando alle informazioni campionarie che utilizzano. Alla tipologia dei dati, infatti, sono quasi naturalmente associate classi di modelli pertinenti e, in qualche misura, gli stessi obiettivi e tagli dell'analisi. I primi due saggi (capp. 16 e 17) si caratterizzano per analizzare serie storiche di dati aggregati, ordinariamente pubblicate dall'Istat; il terzo (cap. 18) porta l'attenzione alla stima dei flussi e di matrici di transizione; gli ultimi due (capp. 18 e 19) sono accomunati dal fatto di utilizzare dati individuali.

Nel cap. 16 Bordignon conduce un approfondito studio dei problemi di destagionalizzazione delle serie sulla partecipazione al lavoro tratte dalla RTFL, incentrato sul confronto di due tra i metodi più consolidati: l' X-11-ARIMA e il metodo *model-based* MSX di Burman. Le diverse questioni di scelta del modello - moltiplicativo o additivo -, di identificazione - automatica o a cura dell'utente - del modello ARIMA, di valutazione dell'accuratezza della destagionalizzazione, di analisi della stabilità delle stime della componente stagionale in presenza di revisioni nei dati, di aggiustamento corrente (cioè, con fattori stagionali previsti per l'intero anno) ovvero concorrente (cioè, con fattori stagionali aggiornati in corso d'anno ogni volta che diventa disponibile una nuova informazione), di destagionalizzazione di serie risultanti dalla combinazione di due o più altre - per via diretta sull'aggregato o per via indiretta sulle componenti -, sono accuratamente vagliate, in particolare nelle loro implicazioni empiriche per le serie italiane. Nell'insieme, risultano confermate le buone proprietà del metodo X-11-ARIMA e vengono preziose indicazioni per un suo appropriato impiego.

L'applicazione dell'approccio VAR (*Vector Auto-Regression*) a serie storiche multivariate del mercato del lavoro e dei prezzi è l'oggetto del contributo di Passamani e Schenkel (cap. 17). Il saggio si segnala per una particolarità. A differenza di altri studi sul mercato del lavoro ispirati allo stesso approccio, l'analisi è condotta su serie dell'occupazione e della disoccupazione convenientemente disaggregate (per sesso e per settore le prime, per sesso e per le tre usuali condizioni - già occupato, in cerca di prima occupazione, altro - le seconde). Ora, l'aspetto di rilievo sta nel fatto che i risultati delle analisi multivariate dinamiche sono parecchio differenti per i diversi segmenti di forza lavoro. Essi suggeriscono che vi è un'articolazione dei meccanismi di interrelazione dinamica più marcata di quanto sinora ipotizzato, e pongono un non semplice interrogativo sul livello appropriato di disaggregazione delle serie.

Col cap. 18, Bernardi e Zaccarin recuperano i risultati dell'abbinamento longitudinale di *records* individuali ed esaminano le opportunità, e i problemi, che essi presentano a fini di stima dei flussi e di matrici di transizione. Il loro interesse è nella distorsione nella stima dei flussi indotta da errori non campionari, segnatamente da dati mancanti, e nei modelli e metodi per ovviarvi. In particolare, essi considerano la possibilità di utilizzare le osservazioni sui soggetti non trovati alla seconda occasione d'indagine (per i quali

si hanno informazioni solo alla prima) e, simmetricamente, sui soggetti non trovati alla prima (per i quali si hanno informazioni solo alla seconda), e di applicare ai dati italiani l'impianto teorico per la riallocazione delle informazioni parziali e le conseguenti procedure per la stima dei flussi. La principale difficoltà con cui si scontrano attiene all'impossibilità di identificare rigorosamente le non risposte, separandole dai mancati abbinamenti dovuti a falsi negativi e neutralizzando gli effetti perturbatori della pratica di sostituzione delle famiglie. I risultati di alcune sperimentazioni, condotte adottando sensate ipotesi semplificatrici, segnalano con chiarezza che apprezzabili miglioramenti nella stima dei flussi chiamano inevitabilmente in causa modifiche nella conduzione dell'indagine (se non altro, per documentare con precisione la sostituzione di famiglie) e indagini supplementari designate ad identificare specifiche fonti di errore.

I due capitoli conclusivi della quinta parte si segnalano, oltre che per il merito delle analisi, perché esemplificano in maniera persuasiva le notevoli opportunità di studio del comportamento dell'offerta di lavoro che vengono da campioni di dati individuali della RTFL: dati longitudinali nel caso del cap. 19; dati *cross-section* integrati con quelli di altre fonti nel caso del cap. 20.

Nel cap. 19, Torelli e Trivellato specificano e stimano un modello di durata della disoccupazione per un campione di giovani lombardi, utilizzando i dati abbinati della RTFL delle prime due indagini del 1986. Sul versante metodologico, essi fissano l'attenzione su alcuni problemi nella costruzione del modello di durata, conseguenti specificamente a incompletezze e/o inaccuratezze nei dati della RTFL. Sul versante interpretativo, l'interrogativo al quale cercano di fornire risposta verte sulla cosiddetta 'dipendenza negativa dalla durata': l'esperienza della disoccupazione agisce nel senso di ridurre le opportunità di trovare un lavoro, o, altrimenti detto, la probabilità di transitare dalla disoccupazione all'occupazione diminuisce a causa del prolungarsi dell'episodio di disoccupazione? Palesemente, la questione è di rilievo per la comprensione delle dinamiche della disoccupazione, e con implicazioni cruciali a fini di politiche. L'argomentata risposta che essi forniscono è a sostegno di una dipendenza negativa dalla durata.

A un modello dell'offerta di lavoro femminile caratterizzato dalla presa in conto di vincoli istituzionali sull'orario di lavoro, è dedicato il saggio di Rettore (cap. 20). I risvolti di interesse che esso presenta sono perlomeno due. Innanzitutto, la specificazione del modello si fa apprezzare per la realistica considerazione di peculiarità del mercato del lavoro italiano, contraddistinto da notevoli rigidità degli orari di lavoro a motivo di accordi collettivi tra le organizzazioni dei datori di lavoro e dei lavoratori. In secondo luogo, Rettore sviluppa una tecnica di stima che permette di far fronte, al meglio, al fatto che in Italia le informazioni necessarie per la stima del modello sono correntemente rilevate con due indagini diverse: con la RTFL quelle sull'orario di lavoro; con l'indagine annuale sui bilanci delle famiglie della Banca d'Italia quelle sui redditi (da lavoro e non).

4.5. *Indagini suppletive*

La sesta parte del volume dà conto delle attività di ricerca che si sono concretate in indagini suppletive, rispettivamente in tema di controlli sulla qualità dei dati (cap. 21) e di acquisizione di informazioni aggiuntive sulla storia lavorativa (cap. 22).

Preliminarmente, occorre peraltro segnalare un non trascurabile scarto fra l'originario progetto in merito al controllo della qualità dei dati tramite indagini suppletive, e gli esiti cui esso è approdato. Il progetto iniziale prevedeva, infatti, un articolato impegno di ricerca, volto ad accertare gli effetti di attori e strumenti del processo di rilevazione sulle stime delle principali grandezze (e sulla loro variabilità). Su questa base, si è giunti alla progettazione operativa di tre distinte indagini suppletive, incentrate ordinatamente su: (i) studio dell'effetto intervistatore; (ii) studio della variabilità di risposta, in particolare dell'effetto rispondente (l'interessato o un *proxy*); (iii) studio dell'effetto *question wording* (vedi Bernardi, Manfroni, Sanetti e Trivellato, 1987). Per vincoli di risorse, tuttavia, è stato possibile effettuare soltanto la prima delle tre indagini. Su uno degli obiettivi di particolare rilievo del progetto, dunque, i risultati sono sì interessanti, ma inevitabilmente parziali.

Nel cap. 21, Bernardi *et al.* illustrano disegno, svolgimento e risultati di un'indagine suppletiva su piccola scala, a carattere sperimentale, mirata a valutare l'errore dell'intervistatore sulle stime della RTFL. La tecnica impiegata è la compenetrazione delle assegnazioni degli intervistatori, che consente di stimare, con buona approssimazione, due indicatori dell'effetto intervistatore: la varianza (cioè a dire, l'incremento di varianza globale addebitabile per l'appunto agli intervistatori) e il coefficiente di correlazione intra-intervistatore (cioè a dire, una misura dello stesso effetto resa indipendente dalla scala delle variabili, quindi appropriata per confronti su variabili diverse o su sottoinsiemi del campione di dimensione differente). L'indagine ha interessato 19 comuni, e circa 3.200 famiglie, delle regioni Lombardia e Campania, ed è stata svolta contestualmente alla RTFL di ottobre 1988. L'errore dell'intervistatore risulta parecchio diversificato, tanto fra variabili che fra sottoinsiemi del campione, ma inequivocabilmente apprezzabile. È palese, quindi, l'opportunità di tenerlo in debito conto nella valutazione dell'attendibilità delle stime, così come - nella prospettiva di una sua riduzione - in sede di perfezionamenti degli strumenti di rilevazione.

Il conclusivo cap. 22, infine, presenta motivazioni, caratteristiche metodologiche e modalità di svolgimento di una impegnativa indagine suppletiva incentrata sulla storia lavorativa, condotta nel maggio 1989 su un campione di quasi 4.500 famiglie lombarde. Gli obiettivi conoscitivi sono essenzialmente due: la ricostruzione della storia lavorativa e la contestuale rilevazione di informazioni sul reddito da lavoro e sull'insieme degli altri redditi personali e familiari. Essi sono convenientemente specificati in relazione alle caratteristiche dello strumento di rilevazione: per l'appunto, un'indagine suppletiva retrospettiva. Trivellato *et al.* si soffermano in particolare sulla progettazione operativa dell'indagine: dalla predisposizione del questionario, all'effettua-

zione di un'indagine pilota, alle molteplici azioni poste in essere per assicurare elevati standards qualitativi alla rilevazione. I primi riscontri, in termini di tasso di risposta, sono promettenti, tanto più se si riflette sulla delicatezza di alcuni temi del questionario.

5. *Che insegnamenti si possono trarre? Alcune indicazioni sul fronte della ricerca*

Da questa ricognizione dei saggi raccolti nel volume risalta con chiarezza che esso è il punto di arrivo del circostanziato progetto FOLA; non certo degli interrogativi e dei propositi scientifici che ne erano alla base. Come ho qua e là messo in luce, quei propositi ed interrogativi solo in parte hanno trovato risposte soddisfacenti. E per un altro verso ne hanno stimolati di nuovi e più penetranti. Non sorprende, dunque, se alla conclusione della ricerca si accompagna, e come uno dei suoi esiti significativi, un rinnovato fervore scientifico e operativo: tanto sul fronte degli impegni dell'Istat per la revisione del disegno della RTFL, quanto su quello dello sviluppo di modelli e metodi di analisi dei dati dell'indagine.

Sul fronte della ricerca (ma, come si vedrà, spesso le problematiche di metodo si intersecano con questioni fattuali), molteplici sono i sentieri che appare promettente battere. Restando a poco più che una scarna elencazione, almeno quattro mi paiono le linee di lavoro meritevoli di rinnovata attenzione.

Innanzitutto, l'affinamento dei metodi di stima delle grandezze - livelli e variazioni -. Si tratta di sviluppare stimatori sintetici, che utilizzino al meglio informazioni esogene, e insieme di approfondire struttura e proprietà di stimatori composti. Per questa via, infatti, è ragionevole attendersi sensibili guadagni in termini di precisione delle stime e risposte soddisfacenti all'esigenza di produrre stimatori per piccole aree (per interessanti sviluppi in tal senso offerti da stimatori di tipo regressivo, vedi, ad es., Isaki, 1990).

Una seconda linea di ricerca, rilevante sia per le risultanze conoscitive sulla distorsione e la variabilità delle stime sia per i riflessi che può conseguentemente avere su revisioni nel disegno e nelle modalità di conduzione della RTFL, attiene alla valutazione degli errori non campionari (per una recente rassegna, incentrata sugli errori in indagini longitudinali, vedi Kalton, Kasprzyk e McMillen, 1989). La possibilità di identificare le diverse fonti di errore non campionario e di stimarne gli effetti è, peraltro, fortemente condizionata dalle modalità di svolgimento della rilevazione. In particolare, si richiede che siano rigorosamente documentati i diversi passi della rilevazione; che l'impatto dei diversi passi sui dati finali sia, per quanto possibile, separabile; infine, che queste informazioni siano integrate da apposite indagini supplementari. È palese, perciò, che le opportunità di ricerca in questa direzione dipendono da (e si intrecciano con) i programmi di attività dell'Istat.

In terzo luogo, meritano di essere approfondite le potenzialità conoscitive (e, specularmente, i limiti) della RTFL a fini di studio della dinamica di breve periodo della partecipazione al lavoro. La motivazione non è solo nella

riconosciuta importanza di analisi dinamiche, in chiave di flussi, per la comprensione del funzionamento del mercato del lavoro. V'è anche da considerare che i saggi sull'argomento sono fra i risultati più significativi del progetto FOLA. Tutt'altro che conclusivi, certo. Ma proprio perciò incoraggiano a un ulteriore impegno di ricerca su abbinamento longitudinale, stima dei flussi e di matrici di transizione e connessi aspetti di qualità dei dati.

Infine, è importante consolidare e generalizzare le opportunità di analisi dei microdati della RTFL. Ci si allontana in tal caso da uno specifico tema di ricerca, e ci si apre ad un ampio spettro di preoccupazioni conoscitive, di tagli analitico-interpretativi, di modelli statistici (ne sono un esempio, in questo volume, i capp. 13-15, 19 e 20). In questo quadro, la disponibilità di dati individuali della RTFL riveste forse interesse particolare per la specificazione e stima di modelli microeconomici del comportamento dell'offerta di lavoro. Non è superfluo ricordare, infatti, che lo studio empirico-quantitativo del comportamento individuale, e lo sviluppo di modelli e metodi statistici pertinenti, ha trovato proprio nel mercato del lavoro il terreno di coltura privilegiato (vedi, ad es., Heckman e Singer, 1986, e Lancaster, 1990). Ovviamente, vi è una pre-condizione perché queste opportunità possano dispiegarsi: l'approntamento e la messa a disposizione di basi di dati individuali, *cross-section* e longitudinali, della RTFL. I passi che l'Istat è chiamato a compiere in questa direzione sono molteplici, e non facili. Essi spaziano dall'affinamento delle competenze nel gestire la dimensione longitudinale della RTFL, alla soluzione dei problemi di predisposizione di *files* di uso pubblico rispettosi della tutela del segreto statistico, ad una conveniente organizzazione di basi di dati individuali orientate ad utilizzatori esterni. Anche in questo caso, dunque, si pone la questione del rapporto fra una linea di lavoro e le condizioni operative per il suo procedere.

6. *Che insegnamenti si possono trarre? Qualche riflessione in vista della revisione della RTFL*

6.1. *Alcuni punti ragionevolmente fermi*

In una diversa ottica, non è irragionevole guardare ai risultati del progetto FOLA con l'intento di trarne spunti di riflessione utili per il miglioramento del disegno e delle modalità di conduzione della RTFL. Naturalmente, le evidenze che vengono da una ricerca con finalità essenzialmente analitiche si prestano per cogliere debolezze dell'attuale impianto dell'indagine, e al più per trarre suggerimenti progettuali di larga massima, non certo per fornire compiute prescrizioni operative. Già riscontrare insufficienze è, tuttavia, imparare. E valutazioni e convincimenti di indole più generale maturati nel corso della complessiva esperienza di ricerca, se pure non tutti argomentabili in maniera stringente, possono anch'essi rivestire qualche interesse.

Tra l'altro, della RTFL è in corso un'importante revisione, raccordata a quella dell'indagine annuale comunitaria, che diverrà operativa nell'aprile

1992. Il momento è dunque opportuno per cercare di estrarre dalla rivisitazione critica condotta sull'indagine attuale ammaestramenti per affinare la progettazione della nuova.

Per dare concretezza a queste riflessioni, conviene prendere le mosse dalle linee-guida della revisione della RTFL presentate dall'Istat alcuni mesi or sono (Istat, 1989). In estrema sintesi, esse sono riassumibili come segue:

(a) mantenimento, almeno a medio termine, della cadenza trimestrale della rilevazione;
(b) razionalizzazione e snellimento del campione. Specificamente, la stratificazione delle unità di primo stadio - i comuni - avverrà sulla base della sola dimensione demografica, e sarà quindi aggiornabile annualmente. Il campione, convenientemente ridisegnato in funzione dell'obiettivo dell'efficienza di un ampio insieme di stime nazionali e regionali, si ridurrà a circa 70.000 famiglie;
(c) affinamento del metodo di stima degli aggregati, con miglioramenti nella post-stratificazione, che avverrà congiuntamente per sesso ed età a livello di regione. A medio termine, vi è inoltre il proposito di sperimentare, e possibilmente adottare, stimatori composti;
(d) revisione del questionario, indotta anche da sollecitazioni dell'Eurostat ad un maggior dettaglio e/o ampliamento nella rilevazione della partecipazione al lavoro;
(e) miglioramenti nell'insieme delle operazioni sul campo.

Si tratta di uno sforzo di rinnovamento cospicuo, sicuramente il più impegnativo nella quasi quarantennale storia della RTFL. Ed è immediato, e significativo, riconoscere come evidenze analitiche e suggerimenti progettuali emersi nella ricerca FOLA, e documentati in questo volume, abbiano trovato ampia eco.

Non indugio sui molti aspetti positivi. In particolare, le scelte prospettate in (b)-(c), destinate a divenire operative già con la rilevazione di ottobre 1990, si possono considerare punti ragionevolmente fermi[16].

6.2. *Questioni aperte*

Mi fermo invece su alcune questioni in parte ancora aperte, che a mio modo di vedere hanno implicazioni di non poco rilievo. Nell'ordine: (i) la revisione del questionario; (ii) il grado di governo e controllo dell'Istat sulle

16 Incidentalmente, noto soltanto che la decisione di continuare ad usare come *frame* per il campione l'anagrafe della popolazione è ragionevole. Il rilievo che è venuta assumendo l'immigrazione extra-comunitaria di forza lavoro e il prevedibile incremento della mobilità delle persone, segnatamente degli attivi, entro l'area comunitaria potrebbero forse far propendere per l'adozione di un campione areale. La costruzione e manutenzione di un siffatto campione è, tuttavia, operazione complicata e costosa: di dubbia utilità quando si disponga - ed è appunto il caso dell'Italia - di registri della popolazione passabilmente affidabili; in ogni caso, bisognosa di un approfondito vaglio di fattibilità e di convenienza. Almeno per il medio termine, è dunque saggio non discostarsi dal *frame* usuale, ed orientarsi piuttosto a ricorrere a fonti integrative *ad hoc* per stime sulle forze di lavoro non residenti.

operazioni sul campo; (iii) l'armonizzazione (o meno) della misura della disoccupazione ai criteri accolti a livello europeo.

La revisione del questionario è compito delicato e complesso, al quale varrà la pena di dedicare notevoli risorse. L'indagine è infatti sollecitata, in parte dallo stesso Eurostat (1990), a soddisfare ulteriori istanze conoscitive: maggiori dettagli su grado e modalità della partecipazione al lavoro per gli occupati, approfondimenti sulla disponibilità a lavorare e sulle attività di ricerca per i disoccupati, ecc.. Ora, non è facile comporre tali istanze con l'esigenza, altrettanto vitale, di mantenere snellezza e tempestività alla rilevazione. Per di più, si andrà verso la progressiva introduzione di tecnologie 'informatizzate' di gestione della RTFL, segnatamente di tecniche di intervista assistite dal calcolatore, di certo non neutre rispetto all'informazione raccolta. E, tra l'altro, è ragionevole attendersi che l''informatizzazione'[17] comporti una maggiore onerosità nell'introdurre successivi aggiustamenti, quindi una maggiore rigidità del dispositivo complessivo dell'indagine. Occorre dunque predisporre la revisione del questionario con particolare lungimiranza. Due orientamenti di indole generale possono, penso, tornare utili.

(a) Conviene orientarsi a superare la rigidità dell'attuale strumento di rilevazione. Un'opportuna flessibiltà può essere realizzata operando in varie direzioni. Da un lato, si può lasciare bianca una porzione del modello di rilevazione, da riempiere, quando necessario, con blocchi di quesiti *ad hoc*. Dall'altro lato, si possono raccordare all'indagine corrente rilevazioni suppletive saltuarie - ricorrenti e/o occasionali -, volte ad investigare temi collaterali[18].

(b) È di grande importanza un'accurata sperimentazione sulla struttura e sul *question wording* del questionario, sia per rivisitare segmenti cruciali di quello attuale (ad es., la presenza - e la collocazione all'inizio dell'intervista - del quesito sulla 'condizione dichiarata', particolarità questa tutta italiana), sia per affrontare le tematiche almeno in parte nuove. Sul piano metodologico, per migliorare le capacità di comprensione degli errori di risposta e per trarne conseguenti indicazioni per il disegno del questionario finale, notevoli vantaggi potrebbero venire dall'adozione di *cognitive laboratory techniques* seguite da indagini pilota ad ampia scala condotte su gruppi selezionati (vedi, ad es., Dippo, 1989).

In secondo luogo, v'è da osservare che tra le ragioni di difficoltà ed inadeguatezza dell'attuale indagine, responsabili tanto dei problemi di tempestività che di quelli di qualità, una - cruciale - necessita di essere messa

---

17 Oltre che alle tecniche di intervista assistite dal calcolatore (per via telefonica - CATI - o ancora faccia a faccia - CAPI -), mi riferisco a nuove tecnologie e reti di *data capture* e di trasmissione di dati, ed a procedure parzialmente o totalmente automatizzate di codifica, imputazione e controllo.

18 Entrambe queste pratiche sono correntemente adottate, con buoni risultati, nella canadese *Labour Force Survey* e nella *Current Population Survey*. A mo' di esempio, è istruttivo scorrere la sequenza dei supplementi all'indagine statunitense, mensile, in programma nel 1991: gennaio: formazione professionale; marzo: ampia indagine suppletiva annuale, nota appunto come *March Supplement*, con rilevazione del reddito, dell'esperienza di lavoro e dei movimenti migratori nell'anno solare precedente; aprile: uso di droga e programmi antidroga nei luoghi di lavoro; maggio: secondo lavoro; giungo: immigrazione-emigrazione; settembre: inabilità al lavoro di reduci di guerra; ottobre: iscrizione alla scuola; maggio/giugno/novembre: disponibilità del telefono (e accettabilità dell'intervista telefonica).

meglio a fuoco. Mi riferisco all'insoddisfacente grado di governo e controllo che l'Istat ha sullo svolgersi delle operazioni sul campo. Le evidenze fornite in proposito dalla ricerca FOLA sono indirette, indiziarie (né sarebbe potuto essere altrimenti, dato che per larghissima parte si sono utilizzate le sole informazioni finali prodotte dalla RTFL); ma concordi, e ragionevolmente preoccupanti. A mio avviso, è questo un nodo che, se non affrontato con determinazione e avviato a soluzione, rischia di compromettere qualsiasi pur pregevole progetto di ristrutturazione. Solo operando su questo terreno, infatti, si possono realizzare consistenti miglioramenti nella qualità dell'indagine, cioè a dire nel contenimento degli errori non campionari e nella tempestività dei risultati. Mettere meglio a fuoco i problemi di governo e controllo delle operazioni sul campo, cominciare ad affrontarli con maggiore determinazione, raccordare e finalizzare alla loro soluzione la stessa 'informatizzazione' delle varie fasi della rilevazione, giungere progressivamente a realizzare un efficace sistema di *survey maintenance* è perciò una tessera essenziale del mosaico di disegno della nuova indagine.

Infine, sul terreno della misura degli aggregati è verosimilmente necessario avviare una riflessione sulla convenienza a mantenere un criterio di definizione della disoccupazione parzialmente disomogeneo rispetto a quello usualmente accolto nei Paesi sviluppati. La peculiarità nostrana nella misura della disoccupazione sta nel modo di accertare una delle condizioni richieste per identificare una persona come disoccupata: la ricerca di lavoro. Tale condizione è interpretata e resa operativa senza porre alcun limite di prossimità temporale all'ultima azione di ricerca svolta (vedi ancora la Fig. 2). Questa scelta definitoria è palesemente piuttosto lontana da una più stretta interpretazione della raccomandazione dell'ILO, seguita dalla maggior parte dei Paesi sviluppati, la quale richiede che la persona abbia compiuto una specifica azione di ricerca nei trenta giorni precedenti l'intervista. È inoltre sensibilmente diversa anche dall'intermedia variante definitoria adottata per la sola Italia dall'Eurostat, che estende il criterio ILO includendo fra i disoccupati coloro i quali hanno compiuto azioni di ricerca da uno a sei mesi prima dell'intervista, purché almeno una di queste consista nella partecipazione ad un concorso pubblico o nell'iscrizione ad un ufficio pubblico di collocamento (per dettagli, vedi Rettore, Torelli e Trivellato, 1988).

Recentemente, si è aperto un vivace dibattito sulla rilevanza di queste, o analoghe, varianti definitorie per la misura della disoccupazione, e in definitiva sulla plausibiltà del criterio accolto in Italia (vedi Rettore, Torelli e Trivellato, 1988; De Nicola, 1989; Sestito, 1989; Micali, 1990). Le evidenze empiriche sono concordi nel segnalare che la scelta dell'uno o dell'altro confine definitorio fra disoccupati e inattivi ha implicazioni considerevoli sulla consistenza degli aggregati, con divergenze di diversi punti percentuali (in termini relativi, del 30-40%) nella stima del tasso di disoccupazione. Potrebbe peraltro restare il dubbio che, tutto sommato, si tratti di una questione di lana caprina. Accertato che la misura della disoccupazione è in larga parte convenzionale, si potrebbe argomentare, infatti, che dibattere sulle definizioni, e sulle stime a cui queste conducono, è un esercizio scarsamente produttivo, perlomeno se si è interessati soprattutto a confronti spaziali e/o

temporali e se i *patterns* di variabilità spaziale e temporale restano sostanzialmente inalterati al mutare delle definizioni. Ora, mentre la prima ipotesi è plausibile (salvo che i confronti spaziali siano internazionali!), la seconda è tutt'altro che scontata. I riscontri al riguardo sono scarni, e non convergenti. Mentre Sestito (1989), sulla scorta di un'analisi temporale aggregata, argomenta in favore di una scarsa sensibilità della dinamica tendenziale e ciclica della disoccupazione a varianti definitorie, nel cap. 15 Rettore, Torelli e Trivellato presentano parziali, ma chiare evidenze in senso contrario, basate su flussi di mobilità trimestrali. Per la Lombardia, essi documentano che gruppi di persone con un diverso *attachment* al mercato del lavoro, alcuni dei quali in prevalenza inclusi o esclusi dai disoccupati a seconda della definizione accolta, hanno probabilità di transizione all'occupazione (o, all'opposto, all'inattività) e probabilità di permanenza nello stato sensibilmente differenti.

Si tratta, è bene ribadirlo, di evidenze empiriche circoscritte, non conclusive. Esse si affiancano peraltro, corroborandola, alla riconsiderazione degli schemi interpretativi del mercato del lavoro italiano, segnata dalla riabilitazione del ruolo dell'offerta. Si fa infatti strada l'opinione che, sia pure in un contesto di vincoli economici e istituzionali piuttosto marcato, c'è una non trascurabile capacità dei singoli e delle famiglie di scegliere rispetto al lavoro (per una rassegna critica dei diversi punti di vista, vedi Frey, 1988). Ora, se questa opinione è condivisa ne discende che il criterio-test per qualificare una persona come disoccupato deve necessariamente essere più restrittivo di quello accolto attualmente. Che la condizione aggiuntiva di aver compiuto almeno un'azione di ricerca nell'ultimo mese sia la sola appropriata, è certo discutibile. Che occorra un accertamento più stringente, e fattuale, della volontà-disponibilità a lavorare è tuttavia, a mio avviso, indubitabile.

Da questo combinazione di evidenze empiriche e di ipotesi interpretative discendono appunto le sollecitazioni ad un riesame della definizione e della misura della disoccupazione, cui bisognerà, prima o poi, por mano. A rimarcare il carattere aperto, problematico che ha contraddistinto l'intero percorso della ricerca FOLA, non è inappropriato concludere questo quadro di sintesi formulando tali sollecitazioni in termini di interrogativi. Innanzitutto, le peculiarità del mercato del lavoro italiano sono tali da giustificare che si continui ad adottare una definizione della disoccupazione differente da quella di massima accolta nella Comunità Europea (e più in generale nell'area dell'OECD), oppure è preferibile conformarsi più pienamente alle convenzioni definitorie internazionali? E ancora, e ancor più, è conveniente restare ad una misura tutto sommato univoca della disoccupazione, o non è preferibile muovere verso la predisposizione in via corrente di un ragionevole insieme di indicatori della disoccupazione (ad esempio, sulla falsariga dei sette indicatori abitualmente pubblicati negli Stati Uniti), che affianchino quello 'ufficiale' e documentino le dimensioni del fenomeno in maniera articolata e secondo ottiche diverse?

PARTE SECONDA:

DISEGNO CAMPIONARIO E STIME

2

PRECISIONE DELLE STIME ED EFFETTO DEL DISEGNO DI CAMPIONAMENTO

Giuliana Coccia, Piero Demetrio Falorsi e Aldo Russo *

1. *Introduzione*

In un recente lavoro (Russo, Falorsi, Coccia e D'Angiolini, 1988), stimolati anche dal fondamentale studio di Verma, Scott e O'Muircheartaigh (1980), abbiamo suggerito, con riferimento all'indagine sulle forze di lavoro, una metodologia per la determinazione di una stima degli errori di campionamento, dell'effetto complessivo del disegno di campionamento e delle sue principali componenti, note con il nome di effetto stratificazione, effetto stadificazione ed effetto ponderazione.

Il presente capitolo costituisce una naturale estensione di questa linea di ricerca. Presentiamo infatti, in forma compatta, gli aspetti metodologici essenziali del calcolo dell'errore campionario e della stima dell'effetto del disegno di campionamento. Ne illustriamo poi i risultati con riguardo alla maggior parte delle stime pubblicate a livello regionale e provinciale.

Le sezioni che seguono riguardano: una breve descrizione del disegno di campionamento della rilevazione sulle forze di lavoro e dello stimatore attualmente utilizzato per l'ottenimento delle stime oggetto d'indagine (sez. 2); gli aspetti metodologici essenziali e i risultati del calcolo dell'errore campionario (sez. 3); la descrizione dell'espressione usata per la determinazione di una stima dell'effetto del disegno di campionamento e l'esposizione dei risultati ottenuti (sez. 4).

2. *Breve descrizione della strategia campionaria*

2.1. *Il disegno di campionamento*

Il campione utilizzato è a due stadi con stratificazione delle unità di primo stadio.

---

* Il capitolo è frutto della collaborazione degli autori. In particolare, P.D. Falorsi ha curato la stesura delle sezz. 1 e 2, mentre le sezz. 3 e 4 sono state redatte rispettivamente da G. Coccia e da A. Russo.

Le unità di primo stadio sono i comuni, quelle di secondo stadio le famiglie. Sono inclusi nel campione tutti i componenti delle famiglie estratte appartenenti alla popolazione oggetto di indagine.

Per la stratificazione dei comuni viene adottato un procedimento basato sui seguenti caratteri del comune: (i) provincia di appartenenza; (ii) settore statistico di appartenenza; (iii) dimensione demografica; (iv) nei comuni con popolazione inferiore a 20.000 abitanti, concatenamento della zona altimetrica e dell'attività economica prevalente.

In ciascuna regione sono definite le seguenti due aree:

- area autorappresentativa (o AR): è costituita dai comuni con popolazione uguale o superiore a 20.000 abitanti, in cui ogni comune costituisce strato a sé;
- area non autorappresentativa (o NAR): comprende i rimanenti comuni; da ogni strato viene estratto un solo comune con probabilità proporzionale al suo peso demografico.

Le famiglie da rilevare vengono estratte in modo sistematico dalle liste anagrafiche dei comuni campione.

L'indagine è ripetuta quattro volte nell'anno, nei mesi di gennaio, aprile, luglio ed ottobre (per maggiori dettagli, vedi Istat, 1978).

2.2. *Il metodo di stima*

Per l'ottenimento delle stime dell'indagine viene adottato uno stimatore del rapporto separato e post-stratificato.

Tale stimatore è espresso dalla somma di due stimatori, corrispondenti alle due aree AR e NAR.

Indicando con $X = X_{AR} + X_{NAR}$ il numero totale di individui che presentano il carattere x nella generica regione geografica, la stima del totale X è data da:

$$\hat{X} = \hat{X}_{AR} + \hat{X}_{NAR}. \qquad (1)$$

Facendo riferimento all'area AR della generica regione geografica introduciamo la seguente simbologia: h = indice di strato (h = 1,...,H); j = indice di famiglia; a = indice di sesso (a = 1,2; 1 = maschio, 2 = femmina); M_h = numero di famiglie residenti nello strato h; m_h = numero di famiglie campione nello strato h; ${}_aP_{hj}$ = numero di componenti, di sesso a, della famiglia j, residente nello strato h; ${}_aP_h$ = popolazione, di sesso a, residente nello strato h; ${}_aX_{hj}$ = numero di componenti che possiedono il carattere x, di sesso a, appartenenti alla famiglia j dello strato h; ${}_aX_h$ = numero di individui che possiedono il carattere x, di sesso a, residenti nello strato h; ${}_aX_{AR}$ = numero di persone che possiedono il carattere x, di sesso a, residenti nell'area AR; X_{AR} = numero di persone che possiedono il carattere x nella area AR.

Lo stimatore che fornisce una stima del totale X_{AR} è dato da:

$$\hat{X}_{AR} = \sum_{a=1}^{2} \sum_{h=1}^{H} \frac{\hat{X}_{ah}}{\hat{P}_{ah}} \, {}_aP_h \,, \qquad (2)$$

dove

$${}_a\hat{X}_h = \frac{M_h}{m_h} \sum_{j=1}^{m_h} {}_aX_{hj} = \sum_{j=1}^{m_h} K_h \, {}_aX_{hj} \qquad (3)$$

e

$${}_a\hat{P}_h = \frac{M_h}{m_h} \sum_{j=1}^{m_h} {}_aP_{hj} = \sum_{j=1}^{m_h} K_h \, {}_aP_{hj} \qquad (4)$$

rappresentano rispettivamente le stime corrette di ${}_aX_h$ e ${}_aP_h$, nelle quali si è posto

$$K_h = \frac{M_h}{m_h} \,. \qquad (5)$$

Relativamente all'area NAR introduciamo la seguente simbologia: t = indice di strato (t=1,...,T); i = indice di comune; s = indice di famiglia; a = indice di sesso; N_t = numero di comuni nello strato t; n_t = numero di comuni campione nello strato t ($n_t = 1$); M_{ti} = numero di famiglie nel comune i dello strato t; m_{ti} = numero di famiglie campione nel comune i dello strato t; ${}_aP_{tis}$ = numero di componenti di sesso a della famiglia s del comune i dello strato t; ${}_aP_{ti}$ = popolazione residente di sesso a, nel comune i nello strato t; ${}_aP_t$ = popolazione residente di sesso a, nello strato t; P_{ti} = popolazione residente nel comune i dello strato t; P_t = popolazione residente nello strato t; ${}_aX_{tis}$ = numero di componenti che possiedono il carattere x, di sesso a, appartenenti alla famiglia s del comune i dello strato t; ${}_aX_{ti}$ = numero di individui che possiedono il carattere x, di sesso a, residenti nel comune i dello strato t; ${}_aX_t$ = numero di individui che possiedono il carattere x, di sesso a, residenti nello strato t; ${}_aX_{NAR}$ = numero di individui che possiedono il carattere x, di sesso a, residenti nell'area NAR; X_{NAR} = numero di individui che possiedono il carattere x, nell'area NAR.

La stima del totale X_{NAR} è ottenuta mediante uno stimatore che ha la stessa struttura della (2), e cioè:

$$\hat{X}_{NAR} = \sum_{a=1}^{2} \sum_{t=1}^{T} \frac{\hat{X}_{at}}{\hat{P}_{at}} \, {}_aP_t \,, \qquad (6)$$

in cui

$$ {}_a\hat{X}_t = \frac{M_{ti}}{Z_{ti}\, m_{ti}} \sum_{s=1}^{m_{ti}} {}_aX_{tis} = \sum_{s=1}^{m_{ti}} K_{ti}\, {}_aX_{tis} \qquad (7) $$

e

$$ {}_a\hat{P}_t = \frac{M_{ti}}{Z_{ti}\, m_{ti}} \sum_{s=1}^{m_{ti}} {}_aP_{tis} = \sum_{s=1}^{m_{ti}} K_{ti}\, {}_aP_{tis}\,, \qquad (8) $$

essendo

$$ K_{ti} = \frac{M_{ti}}{Z_{ti}\, m_{ti}}\,, \qquad (9) $$

dove $Z_{ti} = P_{ti}/P_t$ rappresenta la probabilità di selezione del comune i appartenente allo strato t.

3. *Valutazione del livello di precisione delle stime*

3.1. *Aspetti metodologici*

In questa sezione vengono illustrati i livelli di precisione delle principali stime regionali e provinciali, espressi dai valori dell'errore di campionamento percentuale, che indichiamo con la seguente espressione:

$$ \hat{\varepsilon}\,(\hat{X}) = \frac{\hat{\sigma}\,(\hat{X})}{\hat{X}}\ 100\,, \qquad (10) $$

in cui il numeratore rappresenta una stima dell'errore campionario di $\hat{X}$.

A tale scopo, descriviamo gli aspetti essenziali della metodologia utilizzata per il calcolo della stima della varianza campionaria $V(\hat{X})$ (per una esposizione più organica e dettagliata rimandiamo a Russo, Falorsi, Coccia e D'Angiolini, 1988).

Tenendo presente la (1), segue immediatamente che $V(\hat{X})$ si può scrivere nella forma:

$$ V(\hat{X}) = V(\hat{X}_{AR}) + V(\hat{X}_{NAR}). \qquad (11) $$

Occupiamoci, in primo luogo, del primo addendo. E' possibile mostrare che una stima distorta, ma consistente, della varianza $V(\hat{X}_{AR})$ è fornita dall'espressione:

$$\hat{V}(\hat{X}_{AR}) = \sum_{h=1}^{H} \frac{M_h (M_h - m_h)}{m_h (m_h - 1)} \sum_{j=1}^{m_h} \left(\sum_{a=1}^{2} {}_a\hat{D}_{hj} - \frac{1}{m_h} \sum_{j=1}^{m_h} {}_a\hat{D}_{hj} \right)^2 , \qquad (12)$$

avendo posto

$${}_a\hat{D}_{hj} = {}_aX_{hj} - \frac{\hat{X}_{a\,h}}{\hat{P}_{a\,h}} {}_aP_{hj} . \qquad (13)$$

Per quanto riguarda la determinazione di una stima di $V(\hat{X}_{NAR})$, avendo un solo comune campione per strato, facciamo ricorso al metodo del collassamento degli strati, che consiste nel raggruppare a coppie gli strati. Tale metodo, com'è noto, comporta una sovrastima della varianza, dal momento che fa perdere una parte dell'effetto della stratificazione.

Indicando con g (g = 1,...,T/2) il generico 'super-strato', costituito dall'unione di due strati originari, che indichiamo con i simboli u e w, una stima approssimata della varianza $V(\hat{X}_{NAR})$ è data da:

$$\hat{V}(\hat{X}_{NAR}) = \sum_{g=1}^{T/2} \left(\sum_{a=1}^{2} {}_aD_{gu} - \sum_{a=1}^{2} {}_aD_{gw} \right)^2 , \qquad (14)$$

in cui si è posto

$${}_a\hat{D}_{gu} = {}_a\hat{X}_{gu} - \frac{\hat{X}_{a\,g}}{\hat{P}_{a\,g}} {}_a\hat{P}_{gu} \qquad (15)$$

e

$${}_a\hat{D}_{gw} = {}_a\hat{X}_{gw} - \frac{\hat{X}_{a\,g}}{\hat{P}_{a\,g}} {}_a\hat{P}_{gw} . \qquad (16)$$

3.2. *Errori di campionamento*

Per il calcolo degli errori campionari sono state utilizzate le informazioni della rilevazione eseguita nel mese di ottobre 1986.

Per quanto riguarda le regioni, abbiamo considerato le seguenti 32 stime di livello (cioè, del numero di persone che presentano una data caratteristica):

1 - occupati
2 - disoccupati
3 - in cerca di prima occupazione
4 - altre persone in cerca di lavoro

5 - totale persone in cerca di occupazione
6 - non forze di lavoro
7 - occupati in agricoltura
8 - occupati nell'industria
9 - occupati in altre attività
10 - occupati alle dipendenze in agricoltura
11 - occupati alle dipendenze nell'industria
12 - occupati alle dipendenze in altre attività
13 - totale occupati alle dipendenze
14 - occupati indipendenti agricoltura
15 - occupati indipendenti industria
16 - occupati indipendenti altre attività
17 - totale occupati indipendenti
18 - occupati indipendenti agricoltura: imprenditori, liberi professionisti, lavoratori in proprio
19 - occupati indipendenti industria: imprenditori, liberi professionisti, lavoratori in proprio
20 - occupati indipendenti altre attività: imprenditori, liberi professionisti, lavoratori in proprio
21 - totale occupati indipendenti: imprenditori, liberi professionisti, lavoratori in proprio
22 - occupati indipendenti agricoltura: coadiuvanti
23 - occupati indipendenti industria: coadiuvanti
24 - occupati indipendenti altre attività: coadiuvanti
25 - occupati dipendenti agricoltura: dirigenti ed impiegati
26 - occupati dipendenti industria: dirigenti ed impiegati
27 - occupati dipendenti altre attività: dirigenti ed impiegati
28 - totale occupati dipendenti: dirigenti ed impiegati
29 - occupati dipendenti agricoltura: operai ed assimilati
30 - occupati dipendenti industria: operai ed assimilati
31 - occupati dipendenti altre attività: operai ed assimilati
32 - totale occupati dipendenti: operai ed assimilati.

Nella Tab. 1 (che, come le seguenti, è riportata in Appendice) sono presentati, per ciascuna regione, i valori delle 32 stime e dei corrispondenti errori campionari percentuali.

Nella Tab. 2 sono riportate le distribuzioni percentuali delle stime per classi di errore percentuale.

Ai fini dell'esame dei dati raccolti in tali tabelle, è utile tener presente che, a partire dal 1977, per alcune regioni si è proceduto all'ampliamento del campione, allo scopo di poter osservare convenientemente realtà economiche e sociali sub-regionali (province, aree funzionali, bacini del lavoro, ecc.). Alla data dell'ottobre 1986, cui si riferiscono le nostre elaborazioni, il campione era sovradimensionato nelle seguenti regioni: Piemonte, Lombardia, Friuli-Venezia Giulia, Emilia-Romagna e Marche. V'è da aggiungere che l'idea, comune a tutti gli ampliamenti del campione, è che nelle aree subregionali considerate si possano ottenere con il campione ampliato stime (per i principali gruppi di popolazione) con precisione confrontabile con quella

delle analoghe stime relative alla stessa regione prima dell'ampliamento.

Giova ricordare, infine, che all'epoca di inizio della rilevazione la numerosità del campione fu determinata sotto il vincolo che l'errore massimo ammesso per la stima delle forze di lavoro a livello di regione fosse pari al 5%, al livello di confidenza del 68% .

Anche per le province il calcolo dell'errore campionario è stato effettuato per ciascuna delle 32 stime in esame. Per ragioni di spazio, tuttavia, la presentazione dei risultati è limitata alle sole 5 seguenti stime (Tab. 3):

– occupati in agricoltura
– occupati nell'industria
– occupati in altre attività
– disoccupati
– in cerca di prima occupazione.

Naturalmente, anche per l'interpretazione di questi dati va tenuto presente che nelle province appartenenti alle cinque regioni appena menzionate il campione era sovradimensionato.

Per quanto riguarda l'analisi dei risultati empirici sin qui presentati, tralasciamo un commento puntuale e ci limitiamo ad una valutazione sintetica circoscritta alle regioni.

L'esame dei dati suggerisce le seguenti considerazioni di carattere generale:

(a) anche tenendo presenti le considerazioni svolte sull'ampliamento dei campioni regionali, è immediato osservare che i valori degli errori campionari sono sensibilmente diversi da regione a regione;
(b) la Valle d'Aosta, la Liguria, l'Umbria, il Molise, la Basilicata e la Calabria sono le regioni con un numero relativamente elevato di stime i cui livelli di errore superano il 7,5% (la frazione di tali stime varia tra il 46,9% ed il 71,9%);
(c) un secondo gruppo di regioni, che comprende il Piemonte, la Lombardia, l'Emilia-Romagna e la Campania, è caratterizzato da una situazione abbastanza lusinghiera (la frazione di stime affette da un errore inferiore al 5% varia tra il 59,4% e l'81,5%);
(d) infine un terzo gruppo, costituito dalle rimanenti regioni, si colloca in una situazione intermedia (la frazione di stime con un errore campionario inferiore al 5% è dell'ordine del 40-50%).

4. *Valutazione dell'effetto complessivo del disegno di campionamento*

4.1. *Aspetti metodologici*

Come già detto, la strategia campionaria utilizzata per l'indagine sulle forze di lavoro è caratterizzata dai seguenti aspetti: (i) il disegno è a due stadi stratificato al livello di unità di primo stadio; (ii) le unità di primo stadio sono estratte con probabilità variabile, quelle di secondo stadio in modo sistematico; (iii) per l'ottenimento delle stime oggetto di indagine viene utilizzato uno stimatore del rapporto separato e post-stratificato; (iv) il cam-

pione è solo approssimativamente autoponderante.

Ciascuno di tali aspetti produce un effetto sull'errore campionario delle stime, per cui esso risulta generalmente diverso da quello ottenibile con un campione casuale semplice di pari numerosità in termini di unità finali.

Un fattore che sintetizza gli effetti delle diverse complessità del disegno di campionamento, quali quelli relativi alla stadificazione, alla stratificazione, alla post-stratificazione, al metodo del rapporto ed alla non autoponderazione del campione, è noto in letteratura con il nome di "effetto del disegno di campionamento", convenzionalmente indicato con il simbolo *deft* (Kish, 1965).

La misura di *deft* è espressa dal rapporto tra errore campionario del campione effettivamente utilizzato e quello di un ipotetico campione casuale semplice di pari numerosità in termini di unità finali, cioè a dire:

$$deft = \frac{\hat{\sigma}(\hat{X})}{\hat{\sigma}(\hat{X}^*)} , \qquad (17)$$

in cui $\hat{X}^*$ e $\hat{\sigma}(\hat{X}^*)$ indicano appunto, rispettivamente, la stima del totale X e la stima del corrispondente errore campionario che si sarebbero ottenuti se l'indagine fosse stata realizzata con un campione casuale semplice di dimensione uguale.

Avendo già definito $\hat{V}(\hat{X})$, per calcolare deft basta quindi determinare l'espressione di $\hat{\sigma}(\hat{X}^*)$.

Utilizzando un campione casuale semplice, una stima corretta di X è data da:

$$\hat{X}^* = \frac{P}{p} \sum_{k=1}^{P} X_k , \qquad (18)$$

in cui X_k è una variabile dicotomica che assume il valore 1 se il generico individuo possiede il carattere x e 0 altrimenti; P è la popolazione della generica regione e p è la dimensione dell'ipotetico campione casuale semplice, espresse rispettivamente da :

$$P = \sum_{a=1}^{2} \sum_{h=1}^{H} \sum_{j=1}^{M_h} {}_aP_{hj} + \sum_{a=1}^{2} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \sum_{s=1}^{M_{ti}} {}_aP_{tis} \qquad (19)$$

e

$$p = \sum_{a=1}^{2} \sum_{h=1}^{H} \sum_{j=1}^{m_h} {}_aP_{hj} + \sum_{a=1}^{2} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \sum_{s=1}^{m_{ti}} {}_aP_{tis} . \qquad (20)$$

La varianza di $\hat{\bar{X}}^*$ è perciò fornita dall'espressione:

$$V(\hat{\bar{X}}^*) = \frac{P(P-p)}{p} S^2 , \qquad (21)$$

dove

$$S^2 = \frac{1}{P-1} \sum_{K=1}^{P} (X_k - \bar{\bar{X}})^2 , \quad \bar{\bar{X}} = \frac{1}{P} \sum_{k=1}^{P} X_k . \qquad (22)$$

Affrontiamo ora il problema della determinazione di una stima della (21). Indichiamo con ${}_aX_{hjr}$ una variabile dicotomica che assume il valore di 1 se il generico individuo r, di sesso a appartenente alla famiglia j dello strato h dell'area AR, possiede il carattere x e 0 altrimenti. Inoltre, sia X_{tisq} una variabile dicotomica che assume il valore 1 se l'individuo q, di sesso a appartenente alla famiglia s del comune i dello strato t della area NAR, possiede il carattere x e 0 altrimenti. Mediante alcuni passaggi, che omettiamo per brevità (vedi ancora Russo, Falorsi, Coccia e D'Angiolini, 1988), si perviene alla seguente stima approssimata di $V(\hat{\bar{X}}^*)$:

$$\hat{V}(\hat{\bar{X}}^*) = \frac{P(P-p)}{p(P-1)} \left[\sum_{a=1}^{2} \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{j=1}^{m_h} \sum_{r=1}^{{}_aP_{hj}} {}_aX^2_{hjr} + \right.$$

$$\left. + \sum_{a=1}^{2} \sum_{t=1}^{T} \frac{M_{ti}}{Z_{ti}\, m_{ti}} \sum_{s=1}^{m_{ti}} \sum_{q=1}^{{}_aP_{tis}} {}_aX^2_{tisq} - \frac{\hat{\bar{X}}^2}{P} + \frac{\hat{V}(\hat{\bar{X}})}{P} \right] . \qquad (23)$$

4.2. *Effetti del disegno di campionamento*

Nelle Tabb. 4 e 5, relative rispettivamente alle regioni e alle province, sono riportati i valori di *deft* per ciascuna delle stime considerate nella precedente sezione.

L'esame dei dati della Tab. 4 mostra una situazione abbastanza omogenea sia nell'ambito di ogni regione sia tra le regioni, a meno di qualche eccezione. La maggior parte dei *deft* presenta, infatti, ordini di grandezza che variano tra 1 e 1,5. Si possono osservare, peraltro, due ulteriori gruppi di stime, entrambi costituiti da un numero relativamente basso di grandezze: il primo caratterizzato da valori di *deft* inferiori a 1, l'altro da *deft* con ordini di grandezza compresi tra 2 e 3. Infine, per alcune stime della regione Basilicata, si riscontrano valori di *deft* molto elevati: 3,79 e 3,50 rispettivamente per gli occupati indipendenti in agricoltura e per il sottoinsieme degli stessi costituito da imprenditori, liberi professionisti e lavoratori in proprio.

Sulla base delle informazioni attualmente disponibili non siamo in grado di condurre un'analisi più approfondita, volta a valutare l'importanza degli elementi che possono aver determinato le differenze osservate nei valori di *deft*, sia nell'ambito di ogni regione sia tra le regioni.

Qualche considerazione, tuttavia, può essere ragionevolmente prospettata sulla base delle seguenti argomentazioni.

Nella sez. 4.1 abbiamo illustrato, in modo schematico, la metodologia adottata per il calcolo di *deft*, sottolineando come *deft* non sia altro che un fattore il quale sintetizza le diverse complessità (stratificazione, stadificazione, ecc.) di un disegno complesso rispetto al disegno casuale semplice. Le esperienze maturate a tutt'oggi, sia nell'ambito dell'Istat, sia da parte di istituti statistici di altri Paesi, mostrano che la componente determinante di *deft* è generalmente rappresentata dall'effetto stadificazione. D'altra parte, sotto condizioni non molto restrittive, è possibile mostrare che nel campionamento a più stadi vale la relazione:

$$deft^2 = 1 + (\bar{p} - 1)\, roh\,, \qquad (24)$$

dove roh è il coefficiente di omogeneità, che misura il grado di similarità esistente tra individui entro le unità primarie, e $\bar{p}$ é il numero medio di individui intervistati per unità primaria.

Per l'indagine in esame, nell'ipotesi molto verosimile che *deft* sia determinato essenzialmente dall'effetto stadificazione, e tenendo inoltre presente la (24), le differenze osservate tra i valori di *deft* sono pertanto da attribuire, al fatto che i valori di roh non hanno un *pattern* uniforme.

Per i *deft* relativi alla province valgono considerazioni analoghe a quelle appena svolte.

Appendice: Tabelle

Tab. 1: *Stime ed errori di campionamento percentuali dei principali aggregati regionali*

Aggregati	Piemonte		Valle d'Aosta		Lombardia		Trentino - Alto Adige	
	Stima	Errore %	Stima	Errore %	Stima	Errore %	Stima	Errore %
1	1.775.216	0,76	47.810	3,02	3.645.316	0,57	357.541	1,27
2	31.082	7,90	207	0,64	58.298	5,59	8.276	10,04
3	90.379	4,82	990	0,51	143.040	3,72	7.917	10,00
4	65.611	5,68	701	4,29	107.394	4,64	7.026	14,12
5	187.073	3,54	1.899	3,04	308.732	2,76	23.220	8,12
6	1.825.322	0,87	47.837	3,46	3.577.410	0,66	340.871	1,32
7	157.981	5,16	5.231	3,93	140.052	4,56	43.376	7,94
8	744.941	1,49	14.225	6,41	1.617.806	1,19	90.228	2,87
9	872.294	1,52	28.355	3,88	1.887.457	1,16	223.937	2,07
10	22.741	9,00	2.014	6,23	45.322	6,07	12.697	10,81
11	636.708	1,65	11.525	5,93	1.401.125	1,28	74.261	3,29
12	576.715	1,76	18.857	5,85	1.325.357	1,33	159.717	2,94
13	1.236.164	0,99	32.396	4,18	2.771.804	0,74	246.674	2,14
14	135.240	5,50	3.217	9,30	94.730	5,82	30.680	9,52
15	108.233	4,20	2.700	1,32	216.681	3,93	15.967	7,24
16	295.578	2,64	9.498	3,49	562.101	2,05	64.220	5,04
17	539.052	1,94	15.415	0,85	873.511	1,64	110.867	4,39
18	98.431	5,11	2.614	9,03	74.533	5,92	20.747	8,61
19	95.360	4,22	2.331	1,79	189.294	3,86	13.535	7,55
20	247.487	2,83	7.587	1,98	460.560	2,00	48.629	4,91
21	441.278	1,97	12.532	9,90	724.388	1,64	82.911	4,18
22	36.809	9,74	603	9,74	20.197	11,20	9.933	17,60
23	12.873	11,70	369	1,63	27.386	9,28	2.432	16,05
24	48.091	5,85	1.911	4,93	101.541	4,70	15.592	9,37
25	6.150	15,41	297	3,19	10.607	11,45	1.400	24,31
26	150.819	3,33	1.864	9,64	400.195	2,52	13.814	7,82
27	347.825	2,47	11.398	7,54	783.129	1,81	82.480	3,63
28	504.795	1,93	13.558	8,10	1.193.931	1,46	97.694	3,51
29	16.591	10,58	1.718	0,30	34.715	6,91	11.297	11,64
30	485.889	2,06	9.661	7,36	1.000.931	1,56	60.446	3,74
31	228.890	2,93	7.459	0,07	542.227	2,08	77.237	3,83
32	731.369	1,59	18.838	4,89	1.577.873	1,16	148.980	2,73

Segue Tab. 1 : *Stime ed errori di campionamento percentuali dei principali aggregati regionali*

Aggregati	Veneto		Friuli-Venezia Giulia		Liguria		Emilia - Romagna	
	Stima	Errore %	Stima	Errore %	Stima	Errore %	Stima	Errore %
1	1.771.624	1,20	448.727	1,34	653.416	2,02	1.680.259	0,67
2	36.771	11,84	12.018	9,62	12.537	14,97	51.134	4,92
3	60.244	7,47	17.465	6,90	38.045	8,99	48.636	5,02
4	50.019	9,78	17.906	7,40	20.498	12,93	44.097	4,54
5	147.033	5,81	47.388	5,02	71.080	6,90	143.868	2,97
6	1.764.423	1,65	553.533	1,27	844.506	1,95	1.606.870	0,87
7	202.535	8,61	32.489	12,45	53.765	14,64	213.772	3,33
8	665.268	3,44	145.601	3,73	165.980	4,88	587.676	1,67
9	903.820	2,64	270.637	2,34	433.672	2,51	878.811	1,09
10	52.777	11,32	7.984	13,82	8.122	21,82	76.634	3,93
11	543.150	3,68	118.964	3,55	137.007	5,54	459.237	1,93
12	614.572	3,43	197.174	2,55	294.758	3,15	569.833	1,35
13	1.210.500	1,85	324.122	1,53	439.887	2,54	1.105.704	0,94
14	149.758	9,84	24.505	14,98	45.642	17,26	137.138	4,62
15	122.118	6,71	26.637	8,76	28.974	11,34	128.439	3,85
16	289.248	4,56	73.463	4,02	138.914	5,75	308.978	2,01
17	561.124	3,64	124.605	3,35	213.530	6,47	574.555	1,75
18	106.736	9,03	18.986	15,09	33.973	16,29	97.981	4,91
19	100.020	6,89	24.472	8,98	24.433	11,52	112.927	3,80
20	234.504	5,31	59.201	4,26	106.606	5,73	254.952	2,05
21	441.259	3,40	102.659	3,40	165.012	5,66	465.861	1,76
22	43.022	16,19	5.519	18,70	11.670	23,75	39.157	6,65
23	22.099	12,48	2.165	22,43	4.541	26,72	15.512	9,21
24	54.744	9,03	14.262	8,84	32.307	10,79	54.026	4,82
25	12.744	15,85	1.530	22,69	4.152	27,18	13.363	9,65
26	103.679	6,93	27.664	6,37	37.137	8,90	118.013	3,23
27	313.928	4,52	116.422	3,01	163.182	4,31	331.676	1,91
28	430.351	3,92	145.616	3,06	204.471	3,95	463.051	1,66
29	40.033	13,30	6.454	15,49	3.970	32,16	63.271	4,65
30	439.471	3,79	91.300	4,31	99.870	6,93	341.224	2,38
31	300.644	4,36	80.752	3,50	131.575	4,85	238.158	2,26
32	780.148	2,14	178.506	2,03	235.415	4,08	642.652	1,40

Segue Tab. 1 : *Stime ed errori di campionamento percentuali dei principali aggregati regionali*

Aggregati	Toscana		Umbria		Marche		Lazio	
	Stima	Errore %	Stima	Errore %	Stima	Errore %	Stima	Errore %
1	1.384.942	1,30	307.718	1,78	598.324	1,38	1.932.921	1,00
2	22.223	13,62	6.456	23,85	13.081	10,60	21.728	13,27
3	67.422	7,90	22.417	10,21	23.447	7,95	127.968	5,97
4	47.173	8,77	10.161	13,39	21.430	9,57	69.258	7,27
5	136.818	5,45	39.034	9,47	57.957	5,56	218.954	4,63
6	1.579.571	1,49	347.019	2,05	550.397	2,22	2.080.671	1,18
7	100.324	8,47	33.835	16,27	70.347	5,51	125.348	8,47
8	525.025	2,65	107.459	5,27	229.108	2,99	396.170	3,32
9	759.593	2,11	166.423	4,32	298.869	2,18	1.411.404	1,40
10	45.913	10,30	10.571	20,86	17.196	9,83	44.483	8,98
11	398.134	3,12	90.102	5,32	180.447	3,28	332.788	3,65
12	532.159	2,65	111.358	4,53	199.731	2,82	1.073.156	1,68
13	976.206	1,61	212.031	2,50	397.374	1,99	1.450.428	1,38
14	54.411	12,93	23.264	19,75	53.151	6,56	80.865	10,82
15	126.891	6,21	17.357	10,32	48.660	5,07	63.381	8,60
16	227.433	4,26	55.065	7,57	99.139	4,12	338.247	4,17
17	408.735	3,40	95.687	5,31	200.950	2,81	482.493	3,49
18	45.356	11,29	17.330	17,25	38.897	6,71	60.324	10,83
19	112.033	5,91	15.259	11,00	42.307	5,13	53.336	8,46
20	194.437	4,12	43.723	6,70	80.920	4,12	263.102	4,05
21	351.827	3,12	76.312	4,27	162.124	2,85	376.762	3,43
22	9.054	30,63	5.935	33,06	14.254	11,95	20.541	17,67
23	14.858	20,00	2.098	23,75	6.353	10,93	10.045	20,59
24	32.996	10,87	11.342	17,17	18.218	10,04	75.145	8,21
25	8.848	18,33	1.881	24,46	3.403	20,33	8.121	21,16
26	72.230	8,06	16.176	9,79	25.822	7,68	99.274	6,76
27	294.491	3,75	65.972	6,86	117.092	3,83	745.827	2,27
28	375.569	3,26	84.029	5,79	146.317	3,43	853.222	2,04
29	37.065	11,81	8.690	24,86	13.793	11,38	36.362	10,21
30	325.903	3,48	73.926	5,55	154.625	3,55	233.514	4,10
31	237.669	4,74	45.386	5,43	82.639	4,84	327.330	3,79
32	600.637	2,51	128.002	3,32	251.057	2,74	597.206	2,77

Segue Tab. 1 : *Stime ed errori di campionamento percentuali dei principali aggregati regionali*

Aggregati	Abruzzi		Molise		Campania		Puglia	
	Stima	Errore %	Stima	Errore %	Stima	Errore %	Stima	Errore %
1	476.588	1,81	116.259	3,04	1.790.662	1,08	1.269.172	1,45
2	9.790	16,42	2.012	25,94	43.126	8,50	46.848	12,10
3	25.682	9,11	8.105	14,85	254.244	4,55	112.647	6,63
4	16.793	11,81	4.438	14,94	152.662	5,87	66.632	8,78
5	52.265	6,48	14.555	13,65	450.033	3,77	226.127	5,47
6	503.815	2,10	143.784	3,75	2.098.620	1,17	1.567.223	1,47
7	79.045	9,76	28.071	9,12	268.565	4,89	276.628	5,01
8	142.780	3,65	29.408	5,69	429.627	2,59	308.036	3,46
9	254.763	2,90	58.780	4,20	1.092.470	1,87	684.508	2,56
10	10.509	14,78	3.859	12,29	101.327	6,13	171.647	7,96
11	114.237	4,16	24.959	5,89	356.649	2,95	255.781	3,82
12	171.067	2,87	40.743	5,65	765.372	2,24	464.460	3,21
13	295.812	1,90	69.560	3,34	1.223.349	1,38	891.889	1,95
14	68.536	10,84	24.212	10,34	167.238	6,71	104.981	9,49
15	28.543	8,20	4.449	15,94	72.978	6,34	52.254	7,55
16	83.696	6,08	18.038	8,46	327.098	3,79	220.048	4,78
17	180.776	4,40	46.698	7,04	567.313	3,09	377.283	4,13
18	49.305	9,78	20.847	10,63	128.817	6,91	76.663	8,35
19	24.861	8,54	4.286	15,78	61.874	5,86	47.491	8,15
20	71.227	5,48	16.133	9,12	280.122	3,39	191.334	4,39
21	145.393	3,55	41.267	7,59	470.814	2,81	315.488	3,41
22	19.231	21,55	3.364	25,50	38.421	9,92	28.318	17,92
23	3.683	30,64	163	57,42	11.103	19,22	4.763	30,79
24	12.469	14,44	1.904	23,03	46.975	10,17	28.714	13,88
25	666	49,00	811	22,34	9.712	16,99	11.790	25,16
26	18.341	9,65	2.715	18,67	49.065	7,52	36.212	11,43
27	108.711	4,28	28.887	6,12	493.731	2,87	290.584	5,42
28	127.717	4,09	32.413	5,48	552.507	2,65	338.586	5,17
29	9.843	15,08	3.047	12,54	91.616	6,79	159.857	8,41
30	95.896	4,30	22.244	5,91	307.584	3,19	219.570	4,41
31	62.356	6,11	11.856	10,45	271.642	3,64	173.876	4,27
32	168.095	3,38	37.147	4,38	670.841	2,01	553.303	3,02

Segue Tab. 1 : *Stime ed errori di campionamento percentuali dei principali aggregati regionali*

Aggregati	Basilicata		Calabria		Sicilia		Sardegna	
	Stima	Errore %	Stima	Errore %	Stima	Errore %	Stima	Errore %
1	205.624	3,62	623.617	2,18	1.464.163	1,28	499.320	1,66
2	10.404	15,48	34.318	13,80	65.930	9,27	24.995	9,92
3	21.002	9,35	69.720	10,98	146.221	5,53	61.340	8,94
4	15.416	16,03	44.755	9,70	109.795	6,29	40.996	9,47
5	46.822	6,63	148.794	8,84	321.946	4,28	127.331	5,36
6	248.220	2,94	893.894	1,78	2.167.139	1,23	676.996	1,58
7	54.680	12,11	114.688	8,74	240.357	5,67	69.409	9,18
8	59.407	6,36	129.262	5,91	308.229	3,54	118.062	4,37
9	91.537	4,24	379.668	3,29	915.576	1,77	311.849	1,98
10	16.063	14,95	76.718	11,87	142.734	7,06	21.141	13,61
11	49.848	8,81	98.451	6,62	240.101	4,06	95.656	4,51
12	66.464	4,45	288.993	4,08	628.319	2,35	217.100	2,66
13	132.375	4,43	464.162	3,25	1.011.154	1,55	333.897	2,30
14	38.617	19,98	37.970	13,42	97.623	8,56	48.268	11,79
15	9.559	11,45	30.811	11,25	68.128	7,00	22.406	12,95
16	25.073	8,19	90.674	6,26	287.258	3,95	94.750	4,17
17	73.249	12,49	159.455	5,67	453.009	3,60	165.424	4,86
18	30.986	20,71	33.583	12,92	86.613	9,27	40.957	9,83
19	8.546	9,83	29.707	11,42	62.609	7,06	21.863	13,15
20	22.085	6,93	82.469	5,98	250.755	3,90	80.655	4,23
21	61.618	11,32	145.759	5,30	399.977	3,74	143.475	4,51
22	7.631	34,27	4.387	34,89	11.010	17,03	7.311	33,28
23	1.012	36,10	1.104	50,54	5.518	27,23	543	45,21
24	2.987	23,54	8.205	22,86	36.503	10,34	14.095	10,92
25	814	38,68	4.184	28,61	9.260	17,85	2.003	26,02
26	4.528	22,42	8.876	18,83	30.762	10,82	16.692	12,96
27	40.899	6,68	220.717	4,69	442.815	2,90	125.392	3,31
28	46.240	6,26	233.777	4,79	482.836	2,70	144.087	3,06
2.	15.249	16,86	72.534	13,57	133.474	7,53	19.138	14,06
30	45.320	8,65	89.575	7,27	209.339	4,43	78.964	4,52
31	25.565	14,55	68.276	8,12	185.504	4,40	91.707	5,17
32	86.135	6,17	230.385	6,34	528.317	2,51	189.809	3,62

Tab. 2: *Distribuzione percentuale delle stime regionali per classi di errore percentuale*

Regioni	Classi di errore %				Totale
	< 2,5	2,5-5,0	5,0-7,5	> 7,5	
Piemonte	43,8	21,8	15,6	18,8	100,0
Valle d'Aosta	—	15,6	12,5	71,9	100,0
Lombardia	50,0	25,0	15,6	9,4	100,0
Trentino-Alto Adige	12,5	34,4	6,2	46,9	100,0
Veneto	12,5	34,4	18,7	34,4	100,0
Friuli-Venezia Giulia	15,6	34,4	12,5	37,5	100,0
Liguria	6,3	25,0	21,8	46,9	100,0
Emilia-Romagna	50,0	37,5	6,2	6,3	100,0
Toscana	12,5	37,5	9,4	40,6	100,0
Umbria	9,4	12,5	25,0	53,1	100,0
Marche	12,5	37,5	21,9	28,1	100,0
Lazio	21,9	31,2	9,4	37,5	100,0
Abruzzi	9,4	31,2	12,5	46,9	100,0
Molise	—	15,6	21,9	62,5	100,0
Campania	18,8	40,6	25,0	15,6	100,0
Puglia	9,4	34,4	15,6	40,6	100,0
Basilicata	—	15,6	18,8	65,6	100,0
Calabria	6,3	15,6	25,0	53,1	100,0
Sicilia	12,5	40,6	21,9	25,0	100,0
Sardegna	12,5	34,4	6,2	46,9	100,0

Tab. 3: *Errori di campionamento percentuali delle principali stime provinciali*

Province	Occupati			Disoccupati	In cerca di prima occup.
	Agr.	Ind.	Altre Att.		
Torino	17,2	2,0	2,3	11,3	6,8
Vercelli	15,2	3,9	4,2	22,6	13,8
Novara	12,6	5,4	4,8	19,5	14,3
Cuneo	10,0	4,1	4,2	21,0	14,7
Asti	10,0	8,4	4,6	19,8	16,9
Alessandria	10,9	4,0	3,4	17,7	13,6
Aosta	14,4	6,3	4,2	43,4	21,1
Varese	15,5	3,1	3,2	12,9	12,2
Como	21,4	3,5	4,2	20,7	12,9
Sondrio	15,0	6,5	4,8	18,7	18,2
Milano	15,4	2,0	1,9	9,8	5,5
Bergamo	18,3	4,3	4,4	18,1	12,5
Brescia	12,6	2,7	3,2	16,6	13,8
Pavia	9,2	6,1	4,8	18,8	10,3
Cremona	12,0	5,0	5,1	20,1	11,4
Mantova	7,1	4,1	6,0	23,1	11,6
Bolzano-Bozen	10,0	4,9	2,8	13,6	16,5
Trento	13,6	3,9	3,2	13,8	12,9
Verona	16,3	8,2	4,5	26,1	23,2
Vicenza	24,6	6,8	7,9	27,2	18,4
Belluno	37,6	8,8	3,7	25,4	20,1
Treviso	33,5	8,8	8,5	23,2	24,7
Venezia	13,0	12,5	5,9	39,7	18,2
Padova	12,1	7,4	7,1	23,3	13,7
Rovigo	21,6	8,1	5,0	15,5	12,2
Pordenone	16,6	6,4	8,5	19,0	14,7
Udine	17,9	5,9	3,3	17,2	13,2
Gorizia	40,6	8,0	2,7	21,1	13,6
Trieste	29,0	8,8	2,7	16,4	12,1
Imperia	12,5	15,9	5,5	23,5	30,0
Savona	25,7	20,5	7,9	38,5	17,6
Genova	50,8	7,3	3,1	26,7	12,3
La Spezia	16,5	11,2	9,9	31,9	21,6
Piacenza	11,3	9,8	2,8	13,2	12,3
Parma	17,6	3,9	5,7	23,5	13,2
Reggio Emilia	10,4	6,0	2,6	19,2	28,7
Modena	8,9	2,9	2,8	17,4	16,8
Bologna	8,1	3,5	2,5	11,5	11,7
Ferrara	8,5	5,1	3,1	10,5	14,5
Ravenna	9,4	5,8	4,4	11,3	9,9
Forlì	7,2	5,2	2,7	11,0	12,3
Massa Carrara	27,5	14,5	8,5	100,0	29,8
Lucca	28,6	14,8	5,9	60,2	27,5
Pistoia	27,4	9,5	6,9	50,7	16,5
Firenze	25,7	4,1	3,8	27,0	16,2
Livorno	27,8	9,8	6,7	25,3	20,8
Pisa	26,6	6,1	5,6	35,2	27,2
Arezzo	16,8	5,1	5,5	29,3	26,2
Siena	9,5	12,8	11,4	49,1	24,5
Grosseto	26,5	18,0	7,7	40,5	25,0

Segue Tab. 3: *Errori di campionamento percentuali delle principali stime provinciali*

Province	Occupati			Disoccupati	In cerca di prima occup.
	Agr.	Ind.	Altre Att.		
Perugia	19,3	6,5	5,8	35,7	12,1
Terni	22,3	8,4	4,9	19,3	19,7
Pesaro	14,1	9,6	4,3	18,0	16,5
Ancona	12,8	6,8	4,2	20,9	15,1
Macerata	9,5	3,3	4,4	24,1	13,3
Ascoli Piceno	9,7	4,6	4,8	23,4	17,4
Viterbo	16,0	10,3	6,1	27,2	24,9
Rieti	19,4	11,3	6,6	42,5	26,7
Roma	13,0	4,9	1,6	19,2	7,1
Latina	21,9	8,1	6,5	38,9	31,3
Frosinone	36,5	5,1	6,6	46,9	14,1
L'Aquila	27,3	7,3	4,6	34,2	26,1
Teramo	12,1	9,5	11,3	37,0	15,0
Pescara	46,1	6,8	4,7	34,2	16,9
Chieti	19,0	9,1	5,8	31,1	16,5
Isernia	42,8	8,7	5,1	103,8	31,6
Campobasso	8,9	7,8	6,9	26,0	14,7
Caserta	12,2	5,7	6,5	21,3	12,0
Benevento	10,9	10,5	3,7	32,8	12,5
Napoli	8,8	3,7	2,6	12,9	6,0
Avellino	11,4	6,3	4,3	22,8	16,2
Salerno	11,1	6,9	4,1	16,6	9,0
Foggia	13,5	9,8	4,0	18,3	18,2
Bari	10,7	6,5	4,1	36,4	13,2
Taranto	13,8	4,7	6,6	40,1	14,8
Brindisi	7,8	8,2	4,9	29,0	18,0
Lecce	8,2	8,2	8,0	21,7	13,9
Potenza	16,0	8,1	5,7	21,4	12,7
Matera	9,4	11,5	5,7	25,5	12,0
Cosenza	13,8	7,5	4,7	20,6	22,0
Catanzaro	14,4	5,9	3,6	24,0	8,2
Reggio Calabria	14,8	22,0	8,9	35,4	27,2
Trapani	22,9	7,7	6,2	21,3	11,8
Palermo	14,9	7,7	4,0	28,8	10,5
Messina	14,0	11,4	4,7	18,1	14,3
Agrigento	7,2	13,2	5,1	37,7	19,5
Caltanissetta	30,0	16,3	8,2	42,4	15,2
Enna	10,9	8,2	4,9	4,0	29,6
Catania	15,4	12,1	6,5	21,3	15,5
Ragusa	14,9	11,8	6,3	12,8	17,0
Siracusa	17,5	7,0	7,2	22,0	25,8
Sassari	8,9	5,3	2,8	17,3	9,8
Nuoro	16,2	10,9	5,7	21,4	29,2
Oristano	23,6	30,2	8,0	32,1	45,1
Cagliari	18,2	6,0	3,3	15,6	13,8

Tab. 4: *Effetto del disegno di campionamento sulle principali stime regionali*

Aggreggati	Piemonte	Valle d'Aosta	Lombardia	Trentino-Alto Adige	Veneto
1	1,25	1,32	1,26	1,27	1,24
2	1,32	0,90	1,20	1,16	1,37
3	1,38	0,98	1,26	1,14	1,11
4	1,38	1,38	1,37	1,51	1,31
5	1,47	1,53	1,39	1,59	1,36
6	1,46	1,52	1,46	1,25	1,72
7	1,98	1,58	1,53	2,19	2,39
8	1,34	1,23	1,48	1,16	1,83
9	1,50	1,15	1,60	1,45	1,69
10	1,29	1,89	1,15	1,57	1,60
11	1,35	1,01	1,46	1,19	1,74
12	1,36	1,32	1,48	1,65	1,74
13	1,23	1,34	1,32	1,60	1,44
14	1,95	1,65	1,60	2,20	2,32
15	1,32	1,72	1,64	1,18	1,42
16	1,40	2,12	1,41	1,69	1,53
17	1,44	2,22	1,43	2,00	1,76
18	1,53	1,47	1,44	1,62	1,79
19	1,24	1,63	1,50	1,13	1,31
20	1,37	1,67	1,24	1,42	1,59
21	1,30	1,80	1,30	1,62	1,43
22	1,78	1,07	1,43	2,29	2,03
23	1,26	0,97	1,36	1,00	1,12
24	1,22	1,68	1,34	1,49	1,29
25	1,18	0,62	1,08	1,16	1,10
26	1,24	1,30	1,45	1,18	1,36
27	1,43	1,30	1,49	1,39	1,58
28	1,38	1,54	1,53	1,48	1,63
29	1,28	2,03	1,14	1,59	1,64
30	1,44	1,13	1,47	1,21	1,59
31	1,36	1,33	1,40	1,42	1,49
32	1,41	1,10	1,43	1,47	1,25

Segue Tab. 4: *Effetto del disegno di campionamento sulle principali stime regionali*

Aggreggati	Friuli-Venezia Giulia	Liguria	Emilia-Romagna	Toscana	Umbria
1	1,49	1,28	1,24	1,07	0,94
2	1,39	1,03	1,21	1,11	1,44
3	1,21	1,08	1,20	1,12	1,16
4	1,31	1,14	1,03	1,04	1,03
5	1,46	1,15	1,23	1,12	1,44
6	1,70	1,53	1,55	1,37	1,21
7	2,98	2,25	1,71	1,49	2,29
8	2,00	1,30	1,50	1,13	1,39
9	1,81	1,17	1,25	1,13	1,48
10	1,62	1,18	1,18	1,21	1,62
11	1,70	1,32	1,50	1,14	1,27
12	1,62	1,15	1,19	1,14	1,22
13	1,34	1,20	1,26	1,01	1,00
14	3,11	2,48	1,88	1,66	2,29
15	1,91	1,22	1,51	1,23	1,03
16	1,47	1,38	1,25	1,15	1,38
17	1,64	2,01	1,55	1,27	1,31
18	2,74	1,99	1,69	1,33	1,72
19	1,88	1,13	1,40	1,09	1,03
20	1,39	1,19	1,16	1,02	1,08
21	1,49	1,51	1,38	1,07	0,93
22	1,84	1,74	1,43	1,58	1,90
23	1,37	1,23	1,23	1,32	0,83
24	1,40	1,22	1,21	1,09	1,39
25	1,16	1,06	1,20	0,95	0,82
26	1,41	1,06	1,21	1,20	0,94
27	1,42	1,12	1,24	1,16	1,38
28	1,63	1,16	1,30	1,15	1,33
29	1,63	1,20	1,27	1,25	1,74
30	1,79	1,40	1,57	1,13	1,19
31	1,35	1,12	1,22	1,30	0,89
32	1,23	1,31	1,32	1,16	0,97

Segue Tab. 4: *Effetto del disegno di campionamento sulle principali stime regionali*

Aggreggati	Marche	Lazio	Abruzzi	Molise	Campania
1	1,45	1,01	1,36	1,50	1,28
2	1,25	1,12	1,35	1,31	1,29
3	1,26	1,24	1,25	1,57	1,71
4	1,44	1,11	1,30	1,17	1,70
5	1,40	1,27	1,27	1,95	1,92
6	2,19	1,28	1,64	2,29	1,56
7	1,55	1,74	2,44	1,84	1,89
8	1,60	1,25	1,26	1,19	1,28
9	1,39	1,12	1,39	1,31	1,59
10	1,34	1,10	1,28	0,89	1,43
11	1,53	1,24	1,26	1,13	1,32
12	1,40	1,12	1,09	1,42	1,54
13	1,52	1,13	1,01	1,16	1,26
14	1,59	1,77	2,53	1,92	2,03
15	1,17	1,25	1,21	1,23	1,25
16	1,39	1,43	1,53	1,38	1,62
17	1,40	1,46	1,73	1,91	1,79
18	1,37	1,53	1,94	1,83	1,83
19	1,10	1,13	1,16	1,19	1,06
20	1,25	1,22	1,27	1,41	1,34
21	1,25	1,25	1,24	1,92	1,46
22	1,50	1,44	2,53	1,69	1,43
23	0,91	1,19	1,70	0,91	1,46
24	1,42	1,30	1,33	1,16	1,61
25	1,26	1,08	1,07	0,72	1,22
26	1,29	1,23	1,13	1,12	1,21
27	1,42	1,21	1,25	1,28	1,55
28	1,43	1,18	1,31	1,21	1,51
29	1,38	1,13	1,26	0,81	1,50
30	1,51	1,16	1,19	1,07	1,32
31	1,48	1,28	1,35	1,34	1,42
32	1,55	1,31	1,28	1,05	1,27

Segue Tab. 4: *Effetto del disegno di campionamento sulle principali stime regionali*

Aggreggati	Puglia	Basilicata	Calabria	Sicilia	Sardegna
1	1,15	1,87	1,28	1,11	1,18
2	1,53	1,47	1,60	1,44	1,35
3	1,30	1,28	1,79	1,30	1,90
4	1,31	1,89	1,28	1,28	1,64
5	1,54	1,39	2,17	1,52	1,68
6	1,38	1,79	1,39	1,47	1,45
7	1,58	2,78	1,90	1,72	2,07
8	1,15	1,51	1,36	1,22	1,31
9	1,35	1,29	1,40	1,13	1,04
10	1,95	1,80	2,09	1,64	1,67
11	1,15	1,90	1,32	1,22	1,20
12	1,34	1,13	1,47	1,20	1,12
13	1,21	1,69	1,56	1,05	1,25
14	1,81	3,79	1,63	1,63	2,21
15	1,00	1,05	1,24	1,11	1,66
16	1,34	1,22	1,20	1,32	1,11
17	1,54	3,35	1,47	1,53	1,76
18	1,36	3,50	1,48	1,66	1,69
19	1,03	0,85	1,24	1,07	1,67
20	1,14	0,97	1,09	1,21	1,04
21	1,15	2,75	1,30	1,49	1,51
22	1,77	2,82	1,43	1,09	2,38
23	1,22	1,05	1,02	1,23	0,89
24	1,36	1,19	1,29	1,21	1,10
25	1,61	1,01	1,13	1,04	0,96
26	1,27	1,41	1,11	1,15	1,40
27	1,75	1,31	1,45	1,23	1,03
28	1,82	1,31	1,53	1,20	1,02
29	1,99	1,97	2,33	1,69	1,64
30	1,22	1,78	1,37	1,24	1,09
31	1,05	2,18	1,34	1,16	1,36
32	1,67	1,81	2,00	1,16	1,41

Tab. 5: *Effetto del disegno di campionamento sulle principali stime provinciali*

Province	Occupati			Disoccupati	In cerca di prima occup.
	Agr.	Ind.	Altre Att.		
Torino	2,20	1,02	1,23	1,10	1,19
Vercelli	2,12	1,47	1,50	1,05	1,14
Novara	1,21	1,97	1,86	1,36	1,33
Cuneo	2,54	1,31	1,63	1,20	1,30
Asti	2,25	2,17	1,55	0,90	1,21
Alessandria	2,15	1,11	1,28	0,91	1,25
Aosta	1,60	1,20	1,23	0,93	1,00
Varese	1,11	1,34	1,22	0,91	1,27
Como	1,57	1,43	1,57	1,08	1,21
Sondrio	2,01	1,58	1,80	1,13	1,23
Milano	1,31	1,25	1,47	1,09	1,06
Bergamo	2,35	2,19	1,99	1,23	1,43
Brescia	1,90	1,13	1,37	1,22	1,31
Pavia	1,54	2,03	2,08	1,30	0,98
Cremona	1,66	1,55	1,73	1,42	1,07
Mantova	1,39	1,31	2,10	1,45	0,85
Bolzano-Bozen	2,17	1,26	1,42	0,90	1,14
Trento	2,18	1,18	1,50	1,29	1,14
Verona	1,83	1,50	1,10	1,09	1,25
Vicenza	2,05	1,67	1,69	1,01	0,98
Belluno	2,63	1,90	0,95	1,43	1,18
Treviso	3,45	1,71	2,00	0,95	1,03
Venezia	1,03	2,06	1,46	1,80	1,02
Padova	1,53	1,51	1,71	0,95	0,84
Rovigo	3,06	1,53	1,08	0,79	0,88
Pordenone	1,93	1,83	2,94	1,08	1,26
Udine	2,84	1,83	1,30	1,44	1,13
Gorizia	3,62	1,76	0,99	1,61	1,12
Trieste	1,32	1,65	1,20	1,10	1,18
Imperia	1,38	1,31	1,23	1,08	1,44
Savona	1,31	2,27	1,41	0,92	0,67
Genova	4,56	1,40	1,03	1,17	1,15
La Spezia	0,64	1,15	1,67	0,77	0,88
Piacenza	1,93	2,61	1,06	0,99	1,17
Parma	3,03	1,32	2,23	1,38	1,14
Reggio Emilia	1,93	1,95	0,95	1,15	1,96
Modena	1,52	1,06	1,06	1,23	1,19
Bologna	1,32	1,27	1,22	1,08	1,03
Ferrara	1,70	1,30	1,02	0,99	1,19
Ravenna	1,98	1,55	1,69	1,14	0,92
Forli	1,36	1,51	1,19	1,24	1,16
Massa Carrara	1,15	1,20	0,95	1,17	1,01
Lucca	1,38	1,71	1,09	1,60	1,42
Pistoia	1,45	1,41	1,12	0,97	0,81
Firenze	1,73	0,98	1,03	1,00	0,98
Livorno	1,13	1,16	1,16	0,96	1,20
Pisa	1,31	0,88	0,97	0,83	1,23
Arezzo	1,27	0,85	0,97	0,96	1,24
Siena	0,73	1,44	1,73	1,00	0,99
Grosseto	2,14	1,51	1,14	0,99	0,95

Segue Tab. 5: *Effetto del disegno di campionamento sulle principali stime provinciali*

Province	Occupati			Disoccupati	In cerca di prima occup.
	Agr.	Ind.	Altre Att.		
Perugia	2,22	1,42	1,64	1,73	1,09
Terni	1,82	1,20	0,94	0,70	1,39
Pesaro	1,53	2,01	1,16	1,02	1,26
Ancona	1,50	1,54	1,37	1,08	1,17
Macerata	1,67	1,11	1,55	1,56	1,08
Ascoli Piceno	1,53	1,41	1,40	1,40	1,28
Viterbo	1,79	1,21	1,28	1,17	1,51
Rieti	2,17	1,64	1,48	1,99	2,06
Roma	1,40	1,27	1,05	1,11	1,14
Latina	2,20	1,02	1,08	1,08	1,23
Frosinone	3,18	0,87	1,51	1,05	1,16
L'Aquila	2,69	1,08	1,16	1,30	1,70
Teramo	1,80	2,20	2,96	1,47	0,86
Pescara	3,21	0,90	0,97	1,49	1,10
Chieti	3,06	1,54	1,34	1,29	1,25
Isernia	4,35	1,00	0,85	1,04	1,11
Campobasso	1,57	1,36	1,79	1,34	1,47
Caserta	2,02	1,13	1,84	1,31	2,32
Benevento	3,09	1,99	1,17	1,53	1,44
Napoli	1,39	1,04	1,31	1,09	1,23
Avellino	2,41	1,33	1,40	1,14	2,00
Salerno	2,19	1,44	1,49	1,26	1,42
Foggia	2,04	1,14	0,92	1,20	1,57
Bari	1,57	1,23	1,18	1,79	1,21
Taranto	1,64	0,74	1,37	0,81	1,07
Brindisi	1,02	0,86	0,90	1,28	1,51
Lecce	1,37	1,27	1,95	1,97	1,55
Potenza	3,03	1,54	1,34	1,55	1,33
Matera	1,15	1,60	1,09	1,60	1,07
Cosenza	1,76	1,04	1,15	1,61	2,27
Catanzaro	1,93	0,94	0,91	1,87	0,91
Reggio Calabria	1,61	2,11	1,98	1,35	1,91
Trapani	2,34	0,85	1,20	1,18	0,89
Palermo	1,71	1,14	1,12	1,69	1,16
Messina	1,69	1,42	1,24	0,88	1,40
Agrigento	0,94	1,53	1,07	1,07	1,12
Caltanissetta	2,62	1,76	1,31	2,16	1,27
Enna	0,74	0,89	0,73	0,26	1,82
Catania	1,69	1,50	1,75	1,47	1,39
Ragusa	1,26	0,99	0,94	0,54	0,62
Siracusa	1,56	0,84	1,19	1,12	1,54
Sassari	1,26	1,05	1,04	1,29	1,27
Nuoro	1,85	1,38	1,24	1,20	2,21
Oristano	1,73	2,26	1,10	1,32	2,20
Cagliari	2,07	1,01	0,92	1,22	1,81

3

POST-STRATIFICAZIONE PER SESSO E DISTORSIONI DELLA STRUTTURA PER ETÀ E DELL'OFFERTA DI LAVORO

Giulio Ghellini

1. *Premessa*

L'uso di tecniche di post-stratificazione nell'elaborazione di stime campionarie ha conosciuto in questi ultimi anni un notevole sviluppo applicativo (Copeland, Peitzmeier e Hoy, 1986; Singh, Drew e Choudry, 1984), accompagnato da un rinnovato interesse metodologico sulle proprietà degli stimatori post-stratificati (Jagers, Odèn e Trulson, 1985; Jagers, 1986). La post-stratificazione è nata come tecnica di ponderazione di dati campionari volta ad attenuare effetti distorsivi nelle stime, imputabili al fenomeno delle *non*-risposte (Little, 1984; Bethelhem e Keller, 1987), ma è venuta via via assumendo la caratteristica di principale tecnica di *counter balancing* per mitigare distorsioni introdotte sia in fase di formazione del campione sia nel corso delle operazioni sul campo. Come ricorda Jagers (1986), i metodi di post-stratificazione presumono l'esistenza di una qualche relazione tra variabile oggetto di studio e altre variabili ausiliarie (ad esempio, in un'indagine d'opinione il sesso, l'età e l'istruzione). Si assume quindi che per ogni valore della variabile ausiliaria esista una distribuzione di probabilità della variabile oggetto di studio, e che la probabilità condizionata non sia disturbata dalla presenza di possibili inaccuratezze di natura campionaria e/o non- campionaria. L'idea di base è pertanto che queste inaccuratezze possano influire sulla stima della variabile ausiliaria, ma non sulla distribuzione condizionata della variabile oggetto di studio. Informazioni esterne affidabili sulla variabile ausiliaria (fornite ad esempio da indagini censuarie), se sostituite ai valori campionari, permettono di ottenere una nuova stima della variabile oggetto di studio, come media ponderata delle distribuzioni condizionate (Tremblay, 1986).

In questo capitolo ci si propone innanzitutto di vagliare empiricamente l'eventuale presenza di distorsione nella distribuzione per età della popolazione (variabile ausiliaria), come risulta dall'indagine campionaria trimestrale sulle forze di lavoro (nel seguito RTFL). Viene poi presentata una semplice proposta di post-stratificazione per età del campione della RTFL.

Com'è noto, il disegno d'indagine prevede rilevazioni trimestrali condotte su di un campione a due stadi (comuni e famiglie), con stratificazione delle

unità di primo stadio e rotazione delle unità sia di primo che di secondo stadio (vedi il cap.1, e per maggiori ragguagli Istat, 1978).

L'obiettivo principale della RTFL è la rilevazione delle caratteristiche della popolazione classificate secondo il grado di partecipazione al lavoro, ovvero la stima di aggregati come gli occupati, le persone in cerca di occupazione, le forze di lavoro, ecc. (variabili oggetto di studio). Si tratta di caratteristiche ben note in letteratura per la loro profonda connessione con il profilo demografico e territoriale di una popolazione. È quindi realistico assumere che le rispettive distribuzioni condizionate ben si adeguino ai requisiti sopra ricordati.

È pertanto di rilevante interesse verificare la rispondenza del campione all'universo rispetto ad alcune *variabili* ascrittive, come il sesso, l'età e il luogo di residenza. Una buona aderenza *(fitness)* per queste variabili, infatti, è una garanzia, sia pure indiretta, di precisione nella stima degli aggregati obiettivo dell' indagine. In caso contrario, quando si dispone di adeguate informazioni esterne, l'applicazione della post-stratificazione può permettere un miglioramento complessivo delle stime oggetto di studio.

La scelta di vagliare proprio la struttura per età stimata dalla RTFL trova ragione d'essere nelle due seguenti considerazioni, rispettivamente di ordine metodologico e sostanziale.

Le stime correntemente utilizzate sono ottenute mediante uno stimatore del rapporto separato e post-stratificato: più precisamente, la post-stratificazione avviene secondo la variabile sesso a livello di singolo strato (vedi il cap.2).

Questo metodo di stima garantisce quindi aderenza solo a due delle tre variabili sopra ricordate, sesso e luogo di residenza. Nessuna protezione è invece attivata per prevenire eventuali elementi di distorsione nella rilevazione della variabile età.

Quanto al profilo sostanziale, è noto che il grado e le modalità di partecipazione della popolazione al lavoro sono fortemente connesse con l'età dei soggetti. Si guardi, a titolo di esempio, ai differenti livelli di occupazione e/o disoccupazione presenti attualmente in Italia nelle varie fasce d'età. Sovrastime e/o sottostime sistematiche di alcune fasce d'età inducono quindi, inevitabilmente, distorsioni nelle stime delle caratteristiche lavorative della popolazione nel suo complesso.

Nel seguito del capitolo viene innanzitutto posta a confronto, distintamente per l'Italia e per la Lombardia, la distribuzione per età della popolazione ricavata dalla RTFL dell'ottobre 1981 con quella desunta, previe opportune elaborazioni per ovviare a disomogeneità definitorie, dal XII Censimento della popolazione[1]. Segue, per la sola Lombardia, un'analisi su una

---

1 Per ovviare alle diverse definizioni di popolazione esistenti tra le due rilevazioni ("domiciliata di fatto al netto delle convivenze" quella delle forze di lavoro, "residente" quella censuaria), si è proceduto in modo tale da rendere la popolazione del Censimento, sulla quale si dispone di preziose informazioni aggiuntive, il più possibile simile a quella delle forze di lavoro. Per fare ciò si è innanzitutto calcolata la "popolazione residente al netto delle convivenze" (P) per singola classe di età (i), sottraendo alla popolazione residente della classe d' età i (PR_i) i residenti nelle convivenze (PR^c_i):

serie temporale di 6 anni, dal 1981 al 1985, che consente di vagliare l'adattamento dei dati della RTFL a quelli ricavati da proiezioni per sesso ed età della popolazione lombarda[2]. È infine proposta una semplice procedura di post-stratificazione del campione secondo l'età. Su questa base, si procede al ricalcolo di alcuni tra gli aggregati tipici della partecipazione al lavoro e li si pone a confronto con le stime correntemente pubblicate della RTFL.

2. *Le verifiche sulla presenza di distorsione*

Per comprendere la natura dei vagli effettuati sulla distribuzione per età della RTFL è opportuno esplicitare i criteri guida adottati per i confronti fra le stime dall'indagine da un lato e i dati censuari (o le proiezioni demografiche) dall'altro.

Un primo criterio guida è consistito nel considerare 'vere' le strutture per età (P_i/P) del Censimento e delle proiezioni, che sono pertanto assunte come strutture di riferimento rispetto alle quali vengono misurati gli scostamenti della distribuzione stimata dalla RTFL. È ben nota la presenza di errori,

---

[segue nota]

$P_i = PR_i - PR^c_i$.
Per l' Italia la popolazione P comprende, in aggiunta a quella domiciliata di fatto (PDF), gli emigrati ancora anagraficamente residenti in Italia. Tale aggregato, non rilevato in modo diretto, da valutazioni indirette risulta tuttavia piuttosto esiguo (3-4‰.). La sua presenza non sembra quindi in grado di introdurre significative distorsioni nella distribuzione per età, anche se le persone che lo compongono hanno presumibilmente una composizione diversa dal resto dei residenti.
Per quanto riguarda le realtà territoriali sub-nazionali, la regione Lombardia nel nostro caso, oltre alla diversità descritta, bisogna tenere conto dei fenomeni di presenza/assenza rispetto al resto del Paese. Più precisamente, trascurando per semplicità gli emigrati ancora anagraficamente residenti nella regione, la popolazione domiciliata di fatto della Lombardia (PDF_L) è nella seguente relazione con quella residente al netto delle convivenze:
$PDF_L = P^L_L + (P^*_L - P^L_*)$,
dove i soprascritti indicano il luogo di residenza, i sottoscritti quello di presenza (L=Lombardia, *= altre regioni). Tuttavia i residenti di altre regioni presenti in Lombardia (P^*_L) e i residenti lombardi presenti nelle altre regioni italiane (P^L_*) sono presumibilmente molto simili quanto a struttura per età (sono ambedue "mobili per lavoro o studio"), di entità piuttosto ridotta e con saldo prossimo allo zero (nel periodo 1980-85 le entrate rappresentano mediamente il 3‰. della popolazione residente, mentre le uscite superano di poco il 2‰). Sembra pertanto accettabile l'assunto di considerare la distribuzione per età della "popolazione residente al netto delle convivenze" (P) come struttura di riferimento per valutare le eventuali distorsioni presenti nella distribuzione della popolazione delle forze di lavoro (PDF).

2 Specificamente si sono utilizzate le proiezioni demografiche del Servizio Statistica della Regione Lombardia (Blangiardo, Lauro e Semisa, 1986). Rimandando agli autori per una descrizione dettagliata del metodo di proiezione per sesso ed età della popolazione lombarda, si ricorda che per il calcolo della popolazione residente al netto delle convivenze si è ipotizzato che negli anni sucessivi al 1981 la presenza nelle convivenze di residenti della classe d'età i sia rimasta costante in termini relativi. Sono stati pertanto calcolati dei coefficienti di depurazione dei membri delle convivenze per singola classe d' età:
$C_i = (1 - PR^c_i)/P_i$.
Le popolazioni di confronto per gli anni t (t=1981,...,1985) sono state quindi calcolate nel seguente modo:
${}_tP_i = {}_tPR_i\ C_i$.
E' inoltre da ricordare che le popolazioni delle proiezioni fanno riferimento al 31 dicembre dei pertinenti anni. Per poterle meglio confrontare con quelle ottenute come media delle quattro rilevazioni trimestrali delle forze di lavoro, centrate approsimativamente a metà maggio, si sono calcolate le medie aritmetiche tra le popolazioni al 31 dicembre dell'anno t e dell'anno t-1, che risultano centrate a fine giugno circa.

spesso di natura sistematica, anche nelle rilevazioni censuarie [3]. Tuttavia, la marginalità di tali errori (perlomeno in chiave comparativa), ha indotto a guardare al Censimento come al repertorio statistico in grado di fornire la fotografia meno imprecisa della struttura per età della popolazione. Per aggiornare i confronti agli anni più recenti, poi, si sono coerentemente scelte delle proiezioni della popolazione per età basate sulle rilevazioni censuarie. L'inaffidabilità e il ritardo di aggiornamento delle anagrafi comunali hanno infatti sconsigliato di ricorrere all'alternativa rappresentata dall'utilizzo dei dati amministrativi.

L'altro criterio guida seguito è consistito nell'assumere come totale di riferimento la "popolazione domiciliata di fatto" (PDF) ottenuta con la RTFL, imponendo pertanto il seguente vincolo di uguaglianza sulle popolazioni confrontate:

$$\sum_i P_i = \sum_i PDF_i = PDF \,. \qquad (1)$$

Operativamente, ciò ha comportato la correzione proporzionale dei valori di confronto P_i per il fattore correttivo PDF/P, ottenendo la popolazione $\underline{P}_i$ come

$$\underline{P}_i = P_i \cdot (PDF/P). \qquad (2)$$

Ciò equivale ad imporre che le differenze tra i due totali P e PDF (vedi ancora la nota 1) vengano distribuite proporzionalmente tra le varie classi d'età.

A questo riguardo è opportuno ricordare che le differenze riscontrate sono assai lievi, dell'ordine del 3-4 per mille, sicché le eventuali distorsioni introdotte con tale correttivo incidono sulla singola classe d'età in modo quasi impercettibile, con riflessi di massima sulla seconda cifra decimale di una distribuzione percentuale. È inoltre interessante notare che le differenze riscontrate tra i due totali sono in accordo con evidenze indirette sull'entità delle divergenze fra le due definizioni di popolazione. A titolo di esempio, per l'Italia nel 1981 le 203 mila (4‰) persone in meno trovate nella RTFL (vedi Tab.1) non si discostano di molto dai 265 mila residenti italiani temporaneamente presenti all'estero per motivi di lavoro o studio contati dal Censimento (è questo l'aggregato che più si avvicina a quello degli esclusi dal conteggio della popolazione nella RTFL). Per la Lombardia un analogo

---

3 A questo riguardo è interessante riportare per esteso quanto Francesco Coletti scriveva nel 1911 sull'attendibilità delle tavole di mortalità e di sopravvivenza per età costruite sulla base dei dati censuari del 1901: "Vi sono quelli che non sanno oggi quanti anni abbiano per la semplice ragione che non l'hanno mai saputo con precisione. Il caso è molto frequente nelle campagne. Vi sono altri che l'hanno dimenticato. Anche questo avviene più spesso nelle campagne e più nelle età avanzate. Vi sono altri ancora che non vogliono confessare l'età cui sono giunti e dichiarano meno anni di quelli che abbiano. Vi è, infine, la tendenza mentale che sembra connaturata in noi stessi, a fissarci sulle cifre rotonde o terminanti per zero, in quelle pari o anche terminanti per cinque, ovvero richiamarci ad un anno di nascita approssimativo, che sarà un anno memorabile per qualche avvenimento (in Milano, ad esempio, coloro che affermano di essere nati nel 1848 sono grandemente in eccesso, bella e impensata manifestazione di patriottismo!) o un anno terminante per zero per numero pari o per cinque" (Coletti, 1923, pp. 10-21).

riscontro non è evidentemente possibile, ma la differenza di -6 mila (1‰) riscontrata nel 1981 tra PDF e P (vedi Tab.2), sempre negativa ma percentualmente inferiore a quella italiana, è in linea con le aspettative. Per la Lombardia, i movimenti di "domicilio di fatto" da e per le altre regioni sono infatti con saldo positivo e compensano così, almeno in parte, le assenze per domicilio in altri Paesi.

Tab.1: *Indagine sulle forze di lavoro e Censimento: strutture per età e relative differenze. Italia, 1981* (v.a. in migliaia)

Classi di età	PDF (a)	P (b)	Diff. ass. (a-b)	Diff. rel. % (a/b-1) 100
0 - 13	11.090	11.151	- 61	- 0,6
14 - 19	5.351	5.582	-231	- 4,1
20 - 24	3.978	4.08	-103	- 2,5
25 - 29	3.604	3.78	-177	- 4,7
30 - 39	7.634	7.475	159	2,1
40 - 49	7.372	7.283	89	1,2
50 - 59	7.141	7.938	203	2,9
60 - 64	2.412	2.327	85	3,7
65 - 70	3.116	3.036	80	2,6
oltre 70	4.175	4.235	- 45	- 1,1
Totale	55.873	55.873	-	-

PDF - P = - 203 (-4‰.)

MSE = 2,2; MAE = 2,6; CHI^2 = 36,9.

Tab.2: *Indagine sulle forze di lavoro e Censimento: strutture per età e relative differenze. Lombardia, 1981* (v.a. in migliaia)

Classi di età	PDF (a)	P (b)	Diff. ass. (a-b)	Diff. rel. % (a/b-1) 100
0 - 13	1.635	1.635	0	0,0
14 - 19	842	853	- 11	- 1,3
20 - 24	610	626	- 16	- 2,5
25 - 29	568	611	- 43	- 7,0
30 - 39	1.269	1.275	- 6	- 0,5
40 - 49	1.279	1.272	7	0,5
50 - 59	1.125	1.110	15	1,3
60 - 64	362	346	16	4,7
65 - 70	484	462	22	4,8
oltre 70	626	611	15	2,5
Totale	8.800	8.800	-	-

PDF - P = -6 (-1‰.)

MSE = 1,7; MAE = 2,6; CHI^2 = 6,0.

2.1. *Le verifiche al Censimento 1981*

Come già detto, si sono innanzitutto confrontati, per l'Italia e per la Lombardia, i dati relativi alla RTFL dell'ottobre 1981 (PDF) con quelli del Censimento dello stesso anno ($\underline{P}$).

Per valutare la concordanza/discordanza fra le due stime della struttura per età, si guarda alle singole differenze - assolute e relative - fra PDF_i e $\underline{P}_i$ e inoltre ad alcuni indicatori sintetici di bontà dell'accostamento (Trivellato, 1986):

(a) la radice quadrata dell'errore relativo quadratico medio:

$$MSE = \sum_i (a_i/b_i-1)^{2} \; w_i, \qquad (3)$$

dove $a_i = PDF_i$, $b_i = P_i$ e $wi = \underline{P}_i/\underline{P}$;

(b) l'errore relativo medio assoluto:

$$MAE = \sum_i \; | (a_i/b_i-1)| \; w_i; \qquad (4)$$

(c) l'usuale indice CHI^2 di *goodness of fit* della distribuzione di PDF rispetto a $\underline{P}$.

Dall'analisi delle Tabb. 1 e 2 emergono alcune evidenze degne di nota, che richiedono peraltro di essere vagliate con attenzione. Ad una prima ispezione, infatti, sembra di trovarsi di fronte a riscontri tutt'altro che chiari. A livello di indici sintetici di bontà dell'accostamento di distribuzioni relative, la radice quadrata dell'errore relativo quadratico medio fornisce lo stesso valore (2,6%) per Italia e Lombardia, mentre l'errore relativo assoluto medio è apprezzabilmente maggiore per l'Italia (2,2%) che non per la Lombardia (1,7%). Il campo di variazione delle differenze relative percentuali è d'altra parte superiore per l'aggregato territoriale più piccolo (da -7,0 a 4,8 per la Lombardia, da -4,7 a 3,7 per l'Italia), conformemente con le aspettative suggerite dalla natura campionaria della RTFL. Infine, il valore del CHI^2 segnala una differenza significativa tra le due distribuzioni solo per l'Italia, essendo forse la Lombardia favorita dal sovra-campionamento operato nella regione a partire dal 1980.

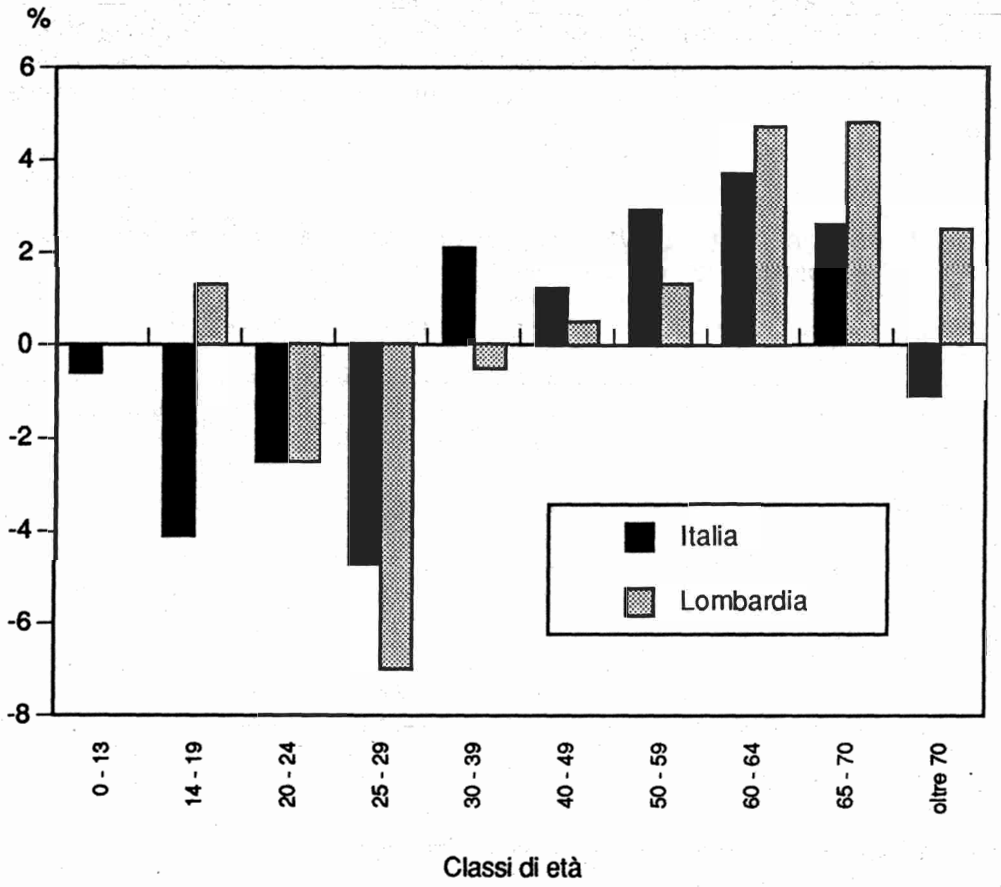
Per valutare in modo appropriato l'adattamento tra le due distribuzioni, e quindi le possibili distorsioni della struttura per età stimata dalla RTFL, è tuttavia necessario osservare l'andamento degli scarti relativi (a_i/b_i - 1) e soprattutto la loro distribuzione fra negativi e positivi, così come evidenziata dalla Fig 1. Il grafico rende manifesto come la RTFL tenda a sottostimare le classi giovanili e di conseguenza a sovrastimare quelle centrali e anziane. Ad esempio, i giovani in età 14-29 sono complessivamente sottostimati di 511 mila unità in Italia (-4,0%) e di 70 mila in Lombardia (-3,5%), mentre la classe d'età 40-70 è sovrastimata rispettivamente di 457 mila (+2,3%) e di 60 mila persone (+1,8%). Esiste sì qualche differenza tra Lombardia e Italia (la classe d'età 30-39 è sovrastimata in Italia (+2,1%) e sottostimata in

Lombardia (-0,5%), mentre per quella oltre i 70 anni accade l'opposto (-1,1%, +2,5%)), ma ciò non intacca il *pattern* complessivo della distorsione appena evidenziato.

Già questa prima verifica segnala la presenza di rilevanti fattori di distorsione nelle stime fornite dalla RTFL. Le strutture per età divergono da quelle ipotizzate vere in modo a dir poco sospetto, certamente non casuale.

Prima di passare all'ulteriore verifica, sul periodo 1981-85 per la Lombardia, è opportuno segnalare anche come la struttura degli scostamenti sia estremamente simile nelle due realtà territoriali osservate (r=0,80). È questo un ulteriore campanello d'allarme a favore dell'ipotesi di presenza di errori sistematici nella distribuzione per età rilevata con la RTFL, errori presumibilmente introdotti soprattutto in fase di rilevazione (formazione delle liste delle famiglie da intervistare, modalità di sostituzione per rifiuto o irreperibilità, ecc.). Il risultato è comunque quello di un indagine che tende a produrre un'immagine invecchiata della popolazione, con conseguenze facilmente intuibili sulle stime degli aggregati della partecipazione al lavoro.

Fig. 1: *Popolazione per classi di età: differenze relative percentuali tra RTFL e Censimento 1981. Italia e Lombardia*



2.2. *La verifica sulle proiezioni 1981-1985*

Proseguendo nella ricerca di ulteriori evidenze sulla presenza (o meno) di distorsioni nella stima della struttura per età della popolazione, si è proceduto a confrontare la serie temporale 1981-85 dei dati medi della RTFL per la Lombardia con quella desunta dalle proiezioni per età della popolazione della regione.

Il procedimento di confronto è il medesimo della sezione precedente, così come uguali sono gli indici utilizzati per valutare la bontà dell'accostamento. Nelle Tabb. 3 e 4 sono riportate le serie storiche delle differenze - assolute e relative - e dei pertinenti indicatori sintetici, che forniscono informazioni utili ad una valutazione più attenta delle distorsioni riscontrate[4] .

La stabilità nel *pattern* degli scostamenti e il progressivo peggioramento dei valori degli indicatori sintetici rendono manifesto uno stato di salute dell' indagine non soddisfacente. L'errore relativo assoluto medio (MAE) passa dal 2,3% del 1981 al 6,0% del 1985, con un valore massimo di 6,7% nel 1984 (quest'ultimo è l'anno con le peggiori *performances* in assoluto). La radice quadrata dell'errore relativo quadratico medio (MSE) sale dal 3,0% al 6,7% (7,8% nel 1984). Il campo di variazione delle differenze relative, che nel 1981 andava da -6,2 a +5,3, nel 1985 oscilla da -7,9 a +12,1 (da -9,2 a 14,3 nel 1984). Il valore dell'indice CHI^2, che per il 1981 e 1982 segnalava uno scostamento non significativo, dal 1983 in poi porta a respingere l'ipotesi di uguaglianza delle due distribuzioni con probabilità di errore inferiore all'1%.

Tab. 3: *Indagine sulle forze di lavoro e proiezioni demografiche: differenze assolute tra le strutture per età. Lombardia* (migliaia)

Classi di età	1981	1982	1983	1984	1985
0 - 13	- 6	-33	-68	-118	-106
14 - 19	-17	-27	-21	- 29	- 25
20 - 24	-26	-17	-29	- 23	- 39
25 - 29	-38	-43	-45	- 57	- 50
30 - 39	-14	-23	-53	- 67	- 43
40 - 49	16	11	8	6	23
50 - 59	17	40	74	96	69
60 - 64	18	30	41	64	55
65 - 70	22	28	58	33	33
oltre 70	28	34	55	95	83
CHI^2	7,7	13,0	27,5	53,4	40,1

4 I dati di base utilizzati sono a disposizione degli interessati presso l'autore.

Tab. 4: *Indagine sulle forze di lavoro e proiezioni demografiche: differenze relative tra le strutture per età. Lombardia* (percentuali)

Classi di età	1981	1982	1983	1984	1985
0 - 13	- 0,4	- 2,1	- 4,4	- 7,9	- 7,4
14 - 19	- 2,0	- 3,1	- 2,4	- 3,4	- 3,0
20 - 24	- 4,2	- 2,7	- 4,4	- 3,4	- 5,6
25 - 29	- 6,2	- 7,0	- 7,3	- 9,2	- 7,9
30 - 39	- 1,1	- 1,8	- 4,2	- 5,2	- 3,3
40 - 49	1,3	0,9	0,6	0,5	1,8
50 - 59	1,5	3,6	6,6	8,5	6,1
60 - 64	5,3	8,0	9,7	13,7	11,3
65 - 70	4,7	6,4	9,5	9,1	9,3
oltre 70	4,6	5,4	8,5	14,3	12,1
MAE	2,3	3,2	4,9	6,7	6,0
MSE	3,0	3,8	5,6	7,8	6,7

È forse ipotizzabile che l'assunto che la popolazione di riferimento, rappresentata dalle proiezioni, sia 'vera' si indebolisca man mano che ci si allontana dalla data del Censimento. Ma questo può influire al più nell'amplificare leggermente la dimensione delle discordanze osservate.

Inoltre, il fatto che nel 1985 (anno nel quale, con l'indagine di luglio, è stata completata per la Lombardia l'operazione di ristratificazione del campione della RTFL) si riscontri un leggero miglioramento nell'accostamento sembra suggerire l'assenza di apprezzabili errori sistematici nelle proiezioni (almeno se, com'è plausibile, tra gli effetti positivi indotti dalla nuova stratificazione vi è la capacità di produrre migliori stime della struttura per età della popolazione).

La Fig. 2 sintetizza la rilevanza delle distorsioni che affliggono la RTFL. I quarantenni sono lo spartiacque tra errori negativi e positivi e di risulta la popolazione giovanile è pesantemente sottorappresentata a vantaggio soprattutto delle classi più anziane (oltre i 60 anni).

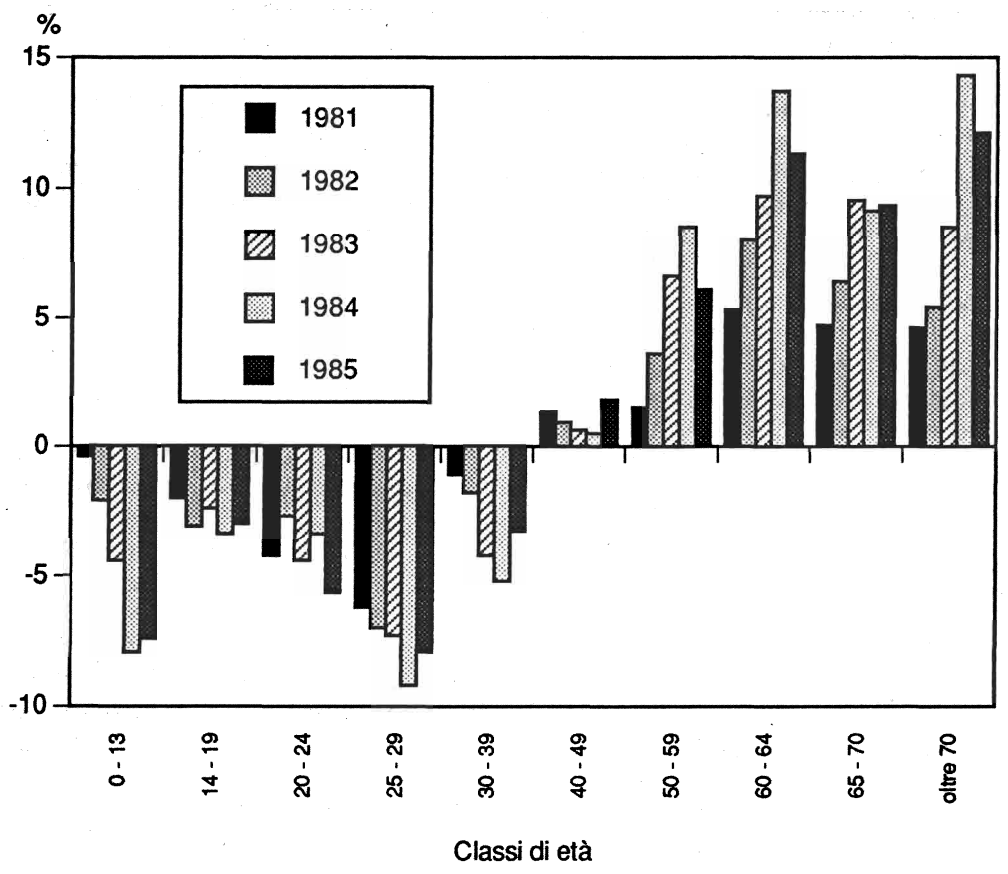
L'osservazione puntuale delle *performances* per singole classi d'età fornisce qualche prezioso elemento per l'interpretazione delle possibili cause della distorsione.

Tra le classi sottostimate, sono in particolare la 0-13 anni e la 30-39 anni a peggiorare di più, a registrare cioè il maggior incremento relativo di errore, e con un andamento piuttosto concorde. Si può ipotizzare si tratti di una sottostima di famiglie giovani (coppie in età 30-39 con 1 o 2 figli di età inferiore ai 13 anni), più difficili da rintracciare da parte degli intervistatori in quanto poco presenti presso la propria abitazione e quindi sostituite da famiglie anziane, più facili da reperire. Sono infatti proprio le classi d'età più anziane (in particolare 60-64 e oltre 70) a fare da contrappeso alle carenze riscontrate nei giovani. Ovviamente, è ragionevole il dubbio che a dar conto di queste discrepanze concorra anche l'effetto di previsioni errate della

natalità sulla distribuzione della popolazione assunta come termine di riferimento. La particolare metodologia adottata per le proiezioni tende peraltro a far respingere questa interpretazione. Almeno fino al 1984, infatti, i dati sui nati utilizzati dal modello di proiezione non sono stimati, ma introdotti invece come variabile esogena dedotta dalle rilevazioni mensili sui movimenti anagrafici.

Queste considerazioni suggeriscono, in definitiva, che gli scostamenti rilevati dipendono in larga misura da errori non campionari nella RTFL. D'altro canto, il leggero miglioramento riscontrato nel 1985, imputabile alla nuova stratificazione elaborata sulla base della rilevazione censuaria del 1981, induce a ritenere che non siano ininfluenti fattori connessi con l'obsolescenza del piano di campionamento, man mano che ci si allontana dalla data in cui viene svolta la stratificazione. Sarebbe forse opportuno pensare ad una procedura di stratificazione non vincolata a informazioni censuarie, disponibili solo a cadenza decennale, e in grado di utilizzare invece variabili rilevate annualmente, che permettano così una sorta di stratificazione aggiornabile di anno in anno.

Fig. 2: *Popolazione per classi di età: differenze relative percentuali tra RTFL e proiezioni demografiche. Lombardia, 1981-85*



3. *Una prima proposta di post-stratificazione*

L'effetto delle distorsioni riscontrate nella distribuzione per età sulle stime degli aggregati tipici della partecipazione al lavoro è ovvio, almeno nella sua direzione. Una popolazione "invecchiata" fornirà generalmente sottostime delle persone in cerca di occupazione, degli occupati, delle forze di lavoro e dei tassi di occupazione e di attività.

Per valutare quantitativamente gli effetti degli errori rilevati in Lombardia nel periodo 1981-85, si sono ricalcolati i contingenti degli occupati, delle persone in cerca di occupazione e delle non forze di lavoro (la loro somma fornisce la popolazione totale), applicando un semplice metodo di post-stratificazione per età, a livello regionale.

Definito a_{ij} l' insieme di persone stimato dalla RTFL, appartenente alla classe di età i e contraddistinto dalla modalità di partecipazione al lavoro j esaustive, (nel nostro caso j = occupato, in cerca di occupazione, non forza di lavoro), è possibile calcolare gli specifici quozienti :

$$R_{j/i} = aij / \Sigma_j a_{ij}, \text{ dove } \Sigma_j a_{ij} = PDF_i \ . \qquad (5)$$

$R_{j/i}$ è la stima rapporto della modalità di partecipazione al lavoro j, data la classe d'età i.

La post-stratificazione proposta consiste nell'applicare le distribuzioni condizionate $R_{j/i}$ alle popolazioni $\underline{P}_i$ desunte dalle proiezioni demografiche. La nuova stima dell'aggregato $a._j$ sarà quindi data dalla media, ponderata con la nuova distribuzione della variabile ausiliaria ($\underline{P}_i$), degli $R_{j/i}$:

$$\underline{a}._j = \Sigma_i R_{j/i} \underline{P}_i \ . \qquad (6)$$

La Tab. 5 riporta i risultati dell' applicazione di questo metodo ai dati lombardi precedentemente analizzati.

Appare evidente come, in accordo con le aspettative, nella RTFL si sia in presenza di una sistematica sottostima degli occupati e delle persone in cerca di occupazione, e conseguentemente delle forze di lavoro, a ovvio beneficio delle non forze di lavoro. È da notare, tra l'altro, che in questo modo si modificano sensibilmente i valori delle variazioni annuali degli aggregati, addirittura con un cambio di segno nel caso degli occupati tra il 1984 e il 1985.

La rilevanza del problema affrontato in questo capitolo, per valutazioni riferite ad aggregati della partecipazione al lavoro, emerge chiaramente da questi ultimi dati. È dunque palese la necessità di una rinnovata attenzione al problema di una buona stima della distribuzione per età della popolazione nell'ambito della RTFL.

Le modalità di intervento dovrebbero andare sia nella direzione di un ripensamento complessivo del metodo di ponderazione, con procedimenti di post-stratificazione in cui compaiano anche le classi d'età, sia in quella di una più attenta definizione del piano di campionamento, in particolare per

Tab. 5: *Confronto tra serie della RTFL e serie corrette con la post-stratificazione per età, per alcuni aggregati della partecipazione al lavoro. Lombardia* (v.a. in migliaia)

Aggregati	anni	Serie RTFL (a)	Serie corretta (b)	(a-b)	(a/b-1)
Occupati	1981	3.647	3.685	- 38	- 1,0%
	1982	3.623	3.658	- 35	- 1,0%
	1983	3.583	3.630	- 47	- 1,3%
	1984	3.568	3.618	- 50	- 1,4%
	1985	3.577	3.615	- 38	- 1,1%
In cerca di occupazione	1981	211	215	- 4	- 1,9%
	1982	242	248	- 6	- 2,4%
	1983	268	275	- 7	- 2,5%
	1984	290	300	- 10	- 3,3%
	1985	299	311	- 11	- 3,5%
Non forze di lavoro	1981	4.944	4.902	+ 42	+ 0,9%
	1982	4.942	4.901	+ 41	+ 0,8%
	1983	4.960	4.906	+ 54	+ 1,1%
	1984	4.953	4.893	+ 60	+ 1,2%
	1985	4.926	4.877	+ 49	+ 1,0%

quanto riguarda l'aggiornabilità dei criteri di stratificazione.

Ci si dovrà preoccupare di approntare anche strategie di conduzione dell'indagine che permettano di migliorare la fase di rilevazione, con particolare attenzione alle modalità di sostituzione delle famiglie: alla luce delle osservazioni fatte, questo è verosimilmente uno dei punti critici dell'indagine. L'applicazione di sistemi di post- stratificazione, del tipo di quello presentato, permette sì di ridurre l'impatto di errori introdotti in fase di rilevazione. Ma conviene ricordare anche che la loro efficienza è inversamente proporzionale alla dimensione delle distorsioni introdotte nella corso delle operazioni sul campo.

4

UTILIZZAZIONE DEGLI AMPLIAMENTI DEL CAMPIONE PER LA STIMA ENTRO PICCOLE AREE

Luigi Fabbris, Piero Demetrio Falorsi e Aldo Russo *

1. *Introduzione*

La struttura e la numerosità del campione di base dell'indagine Istat sulle forze di lavoro sono definite per ottenere stime affidabili a livello regionale e nazionale per i principali gruppi di popolazione (Istat, 1978). La dimensione del campione di base è invece insufficiente per lo studio dell'occupazione o della disoccupazione a livello locale.

Nel passato, per poter osservare convenientemente realtà economiche e sociali sub-regionali (province, USL, bacini del lavoro, ecc.), l'Istat ha fatto ricorso alla sola tecnica del sovradimensionamento del campione (Fabbris, Russo e Sanetti, 1988).

In questo capitolo, allo scopo di migliorare il livello di precisione delle stime nei domini sub-regionali, si esamina la possibilità di adottare una forma di stimatore sintetico, in luogo dello stimatore attualmente utilizzato (del rapporto post-stratificato), senza necessariamente ricorrere al sovradimensionamento del campione di base.

Lo studio è fondato sul confronto tra l'attendibilità delle stime ottenibili applicando lo stimatore del rapporto post-stratificato sul campione ampliato e quella delle stime ottenibili applicando lo stimatore sintetico sul campione di base.

Lo studio è condotto sulle USL della regione Lombardia, che costituiscono domini sub-regionali interessanti a fini di analisi delle forze di lavoro, ma non pianificati nel contesto del disegno di campionamento dell'indagine sulle forze di lavoro.

Occorre considerare, infatti, che:

(a) le USL non erano ancora create alla data di formulazione del disegno, né sono state tenute in conto nelle revisioni della stratificazione che da allora sono state realizzate;

(b) i confini di varie unità economico-sociali sub-regionali (ad esempio bacini del lavoro, ecc.) in genere non coincidono con quelli degli strati.

---

* Il capitolo è frutto della collaborazione degli autori. Per quanto riguarda la sua stesura, L. Fabbris ha redatto la sez. 1, A. Russo le sezz. 2 e 3 e P. D. Falorsi le sezz. 4 e 5.

2. *Simbologia e parametri oggetto di stima*

2.1. *Simbologia*

Con riferimento al disegno campionario attualmente adottato per l'indagine sulle forze di lavoro e relativamente ad una generica provincia, indichiamo con:

c	indice di disegno con due modalità: a=disegno ampliato; b= disegno base;
t	indice di strato;
T	insieme degli strati definiti nella provincia;
i	indice di comune;
j	indice di famiglia;
s	indice di classe demografica;
S	insieme delle classi demografiche definite secondo le modalità congiunte del sesso e delle seguenti classi di età: 14-19, 20-29, 30-39, 40-49, 50-59, 60 ed oltre;
N_{ct}	insieme dei comuni compresi nello strato t del disegno c;
M_{cti}	insieme delle famiglie residenti nel comune cti;
m_{cti}	insieme delle famiglie campione nel comune cti;
P_{cti}	popolazione residente nel comune cti;
P_{ct}	popolazione residente nello strato ct;
P_{ctijs}	numero di componenti della classe demografica s appartenenti alla famiglia ctij;
P_{cts}	popolazione residente nello strato ct ed appartenente alla classe demografica s;
P_s	popolazione residente nella provincia ed appartenente alla classe demografica s;
π_{1cti}	probabilità di selezione del comune cti;
π_{2cti}	probabilità di inclusione della generica famiglia fissato il comune cti;
K_{1cti}	inverso della probabilità di selezione π_{1cti};
K_{2cti}	inverso della probabilità di inclusione π_{2cti};
$K_{cti} = K_{1cti}\, K_{2cti}$	peso base attribuito alla generica famiglia campione del comune cti.

2.2. *Parametri oggetto di stima*

Sia X_{ctijs} il totale del generico carattere x oggetto d'indagine relativo ai P_{ctijs} componenti (ad esempio: numero di componenti occupati, appartenenti alla classe demografica s ed alla famiglia j del comune i dello strato ct).

Con riferimento ad una data USL d, il $totale_d X$ del generico carattere x

è allora dato da:

$$ {}_dX = \sum_{t \in T} \sum_{s \in S} \sum_{i \in {}_dN_{ct}} \sum_{j \in M_{cti}} X_{ctijs}, \tag{1} $$

in cui ${}_dN_{ct}$ indica l'insieme dei comuni appartenenti allo strato ct ed alla USL d.

La (1) si può poi riscrivere nelle forme equivalenti:

$$ {}_dX = \sum_{t \in T} \sum_{s \in S} {}_dX_{cts} = \sum_{t \in T} {}_dX_{ct} = \sum_{s \in S} {}_dX_s, \tag{2} $$

in cui si è posto:

$$ {}_dX_{cts} = \sum_{i \in {}_dN_{ct}} \sum_{j \in M_{cti}} X_{ctijs}, \tag{3} $$

$$ {}_dX_{ct} = \sum_{s \in S} {}_dX_{cts}, \tag{4} $$

$$ {}_dX_s = \sum_{t \in T} {}_dX_{cts}. \tag{5} $$

3. *Stimatori*

3.1. *Stimatore del rapporto post-stratificato corrispondente al disegno ampliato*

La struttura formale dello stimatore del rapporto post-stratificato applicato sul campione ampliato è definita dall'espressione:

$$ {}_d\hat{X}_{pos} = \sum_{s \in S} \sum_{t \in T} \frac{{}_d\hat{X}_{ats}}{\hat{P}_s} P_s, \tag{6} $$

in cui:

$$ \hat{P}_s = \sum_{t \in T} \sum_{j \in m_{ati}} K_{ati} P_{atijs} \tag{7} $$

e

$$ {}_d\hat{X}_{ats} = \delta_{ati} \sum_{j \in m_{ati}} K_{ati} X_{atijs}, \tag{8} $$

dove l'indice a denota che le relazioni fanno riferimento al campione ampliato; ${}_d\hat{X}_{ats}$ e $\hat{P}_s$ rappresentano rispettivamente le stime corrette di ${}_dX_{ats}$ e P_s; δ_{ati}

una variabile indicatrice che assume valore 1 se il comune ati appartiene all'insieme ${}_dN_{at}$ e valore 0 altrimenti.

Ai fini delle successive derivazioni, conviene porre la (6) nella forma:

$$ {}_d\hat{X}_{pos} = \sum_{s \in S} {}_d\hat{R}_s P_s , \qquad (9) $$

in cui

$$ d\hat{R}_s = \sum_{t \in T} \frac{d\hat{X}_{ats}}{\hat{P}_s} \qquad (10) $$

rappresenta una stima di

$$ {}_dR_s = \sum_{t \in T} \frac{{}_dX_{ats}}{P_s} . \qquad (11) $$

Stimatori del tipo (6), basati su diversi tipi di poststratificazione, sono stati studiati da Singh e Tessier (1976).

3.2. *Stimatore sintetico corrispondente al disegno base*

In letteratura sono state proposte differenti forme di stimatore sintetico, corrispondenti a modi differenti di combinare le informazioni campionarie con quelle desumibili da altre fonti (Purcell e Linacre, 1976; Gonzalez e Waksberg, 1973; Gonzalez e Hoza, 1978; Falorsi e Russo, 1987; Russo e Falorsi, 1990).

In questo studio adottiamo una versione di stimatore sintetico che, nelle sperimentazioni riportate in Russo e Falorsi (1990), ha mostrato di possedere una buona efficienza in termini di errore quadratico medio.

Lo stimatore, definito con riferimento al disegno base, dato da:

$$ {}_d\hat{X}_{sin} = \sum_{s \in S} \hat{X}_s \, {}_dW_s , \qquad (12) $$

in cui:

$$ {}_dW_s = \frac{{}_dP_s}{P_s} \qquad (13) $$

e

$$ \hat{X}_s = \sum_{t \in T} \sum_{j \in m_{bti}} K_{bti} \, X_{btijs} , \qquad (14) $$

dove l'indice b indica il riferimento al campione base.

4. *Varianza e distorsione degli stimatori proposti*

4.1. *Stimatore post-stratificato*

Lo stimatore ${}_d\hat{\bar{X}}_{pos}$ è non lineare, dunque non gode della proprietà della correttezza, ma è consistente. Il problema della determinazione della sua varianza campionaria è nel seguito affrontato nell'ottica di cercarne una soluzione approssimata per il disegno di cui si tratta. Per maggiori dettagli, rinviamo a Russo e Falorsi (1990).

Un metodo semplice per raggiungere tale scopo consiste nell'approssimare la varianza di ${}_d\hat{\bar{X}}_{pos}$ con la varianza dei termini di ordine lineare dello sviluppo in serie di Taylor di ${}_d\hat{\bar{X}}_{pos}$ (Woodruff 1971, Russo, 1988). Adottando tale metodologia, si ottiene (Russo e Falorsi 1990), che la varianza dello stimatore post-stratificato è definita dall'espressione:

$$V({}_d\hat{\bar{X}}_{pos}) \doteq \sum_{t\in T} \left(\sum_{i\in N_{at}} \pi_{1ati} (Z_{1ati} K_{1ati} - Z_{at})^2 + \sum_{i\in N_{at}} K_{ati} (M_{ati} - m_{ati}) S^2_{atiz} \right) , \qquad (15)$$

in cui:

$$Z_{atij} = \sum_{s\in S} (\delta_{ati} X_{atijs} - {}_dR_s P_{atijs}) , \qquad (16)$$

$$Z_{ati} = \sum_{j\in M_{ati}} Z_{atij} , \qquad (17)$$

$$Z_{at} = \sum_{i\in N_{at}} Z_{ati} , \qquad (18)$$

$$S^2_{atiz} = \frac{1}{M_{ati} - 1} \sum_{j\in M_{ati}} (Z_{atij} - \frac{1}{M_{ati}} Z_{ati})^2 . \qquad (19)$$

Anche lo studio della distorsione di ${}_d\hat{\bar{X}}_{pos}$ può essere condotto con un metodo fondato sulla linearizzazione, simile a quello impiegato per la determinazione della varianza. Si dimostra (Russo e Falorsi, 1990) che:

$$B({}_d\hat{\bar{X}}_{pos}) = E({}_d\hat{\bar{X}}_{pos}) - {}_d\bar{X} \doteq 0 . \qquad (20)$$

Per ottenere una distorsione non nulla dovremmo considerare lo sviluppo in serie di Taylor comprendente almeno i termini di ordine quadratico, con le relative ulteriori complessità di calcolo. Tuttavia, come mostrano alcuni studi condotti all'estero su indagine concrete, le stime ottenute mediante l'impiego di stimatori post-stratificati sono generalmente affette da livelli di distorsione trascurabili.

4.2. *Stimatore sintetico*

L'espressione esplicita della varianza dello stimatore (12) vedi Russo e Falorsi, 1990 è data:

$$V({}_d\hat{X}_{sin}) = \sum_{t \in T} \left(\sum_{i \in N_{bt}} \pi_{1bti} (Z_{bti} K_{1bti} - Z_{bt})^2 + \sum_{i \in N_{bt}} K_{bti} (M_{bti} - m_{bti}) S^2_{btiz} \right), \quad (21)$$

dove:

$$Z_{btij} = \sum_{s \in S} X_{btijs} \, {}_dW_s, \quad (22)$$

$$Z_{bti} = \sum_{j \in M_{bti}} Z_{btij}, \quad (23)$$

$$Z_{bt} = \sum_{i \in N_{bt}} Z_{bti}, \quad (24)$$

$$S^2_{btiz} = \frac{1}{M_{bti} - 1} \sum_{j \in M_{bti}} (Z_{btij} - Z_{bti})^2 . \quad (25)$$

Per la distorsione si ha inoltre:

$$B({}_d\hat{X}_{sin}) = E \left(\sum_{s \in S} \hat{X}_s \, {}_dW_s \right) - {}_dX \doteq \sum_{s \in S} (X_s \, {}_dW_s - {}_dX_s) . \quad (26)$$

5. *Analisi empirica*

Ai fini di una valutazione empirica della distorsione e della varianza degli stimatori considerati, si conduce ora una sperimentazione sui dati a livello comunale tratti dal XII Censimento generale della popolazione (ottobre 1981). Il vantaggio offerto dall'uso dei dati censuari è, palesemente, nel fatto che si opera su una popolazione (e non su un campione), sicché è possibile valutare esattamente le *performances* dei due stimatori.

L'analisi empirica è condotta con riguardo a due aggregati censuari:
– gli attivi occupati;
– gli attivi in cerca di occupazione.

Pur con qualche differenza, sono questi gli aggregati più prossimi a quelli di maggiore interesse della rilevazione trimestrale delle forze di lavoro: occupati e disoccupati.

L'itinerario seguito per svolgere l'analisi è il seguente:
(i) Tutti i comuni della regione Lombardia sono suddivisi in strati secondo due processi di stratificazione: il primo è quello del disegno base ed il

secondo è quello del disegno ampliato del campionamento per l'indagine sulle forze di lavoro.

(ii) Si procede all'identificazione della USL mediante aggregazione dei pertinenti comuni.

(iii) Per ogni USL, relativamente ai due parametri d'interesse, numero di attivi occupati ed attivi in cerca di occupazione, si calcola un opportuno insieme di indicatori.

Gli indicatori (percentuali) che nel seguito presentiamo e commentiamo sono:

(a) un indicatore del peso demografico della USL nell'ambito della provincia:

$$ {}_dP' = \frac{{}_dP}{P}\ 100\ , \tag{27} $$

dove ${}_dP$ e P indicano rispettivamente la popolazione della generica USL e quella della provincia di appartenenza;

(b) l'errore relativo dello stimatore sintetico:

$$ \varepsilon\ ({}_d\hat{X}_{sin}) = \frac{(V({}_d\hat{X}_{sin}) + B^2({}_d\hat{X}_{sin}))^{1/2}}{{}_dX}\ 100\ ; \tag{28} $$

(c) la distorsione dello stimatore sintetico rispetto al valore vero:

$$ B'\ ({}_d\hat{X}_{sin}) = \frac{|B({}_d\hat{X}_{sin})|}{{}_dX}\ 100\ ; \tag{29} $$

(d) la distorsione dello stimatore sintetico rispetto all'errore quadratico medio dello stimatore stesso:

$$ B''\ ({}_d\hat{X}_{sin}) = \frac{B^2({}_d\hat{X}_{sin})}{V({}_d\hat{X}_{sin}) + B^2({}_d\hat{X}_{sin})}\ 100\ ; \tag{30} $$

(e) l'errore relativo dello stimatore post-stratificato:

$$ \varepsilon\ ({}_d\hat{X}_{pos}) = \frac{(V({}_d\hat{X}_{pos}))^{1/2}}{{}_dX}\ 100\ ; \tag{31} $$

(f) l'efficienza dello stimatore sintetico rispetto allo stimatore post-stratificato:

$$ E({}_d\hat{X}_{sin}) = \frac{\varepsilon\ ({}_d\hat{X}_{pos})}{\varepsilon\ ({}_d\hat{X}_{sin})}\ . \tag{32} $$

I risultati salienti delle analisi empiriche effettuate sono contenuti nelle tabb. 1 e 2 (riportate in Appendice) che si riferiscono rispettivamente agli attivi occupati ed agli attivi in cerca di occupazione.

Esaminiamo dapprima la Tab. 2. Le evidenze di maggior rilievo possono essere riassunte in due ordini di osservazioni. Innanzitutto lo stimatore sintetico, applicato sul campione base risulta più efficiente dello stimatore post-stratificato impiegato con il disegno campionario ampliato. Infatti il valore medio regionale dell'efficienza del primo sul secondo è pari a 3,2, con valori che vanno da un minimo di 0,9 nella USL 3 di Pavia ad un massimo di 10,7 nella USL 23 di Milano.

A fronte di questi vantaggi dello stimatore sintetico in termini di efficienza, stanno peraltro risultati meno soddisfacenti in termini di distorsione. Lo stimatore sintetico, presenta livelli di distorsione abbastanza elevati. Gli indici $B'({}_d\hat{X}_{sin})$ e $B''({}_d\hat{X}_{sin})$ hanno infatti valori medi regionali pari rispettivamente a 18,6 e 78,6. La distorsione, inoltre, è notevolmente diversificata. L'indice $B'({}_d\hat{X}_{sin})$ varia da un valore minimo di 0,3 nella USL 1 di Sondrio ad un valore massimo pari a 91,0 nella USL 4 di Brescia. La variabilità della distorsione a livello locale evidenzia il fatto che il numero di attivi in cerca di occupazione non dipende unicamente dalla struttura della popolazione per sesso e classi di età nella singola USL.

Passiamo ora ad esaminare la Tab. 1. Anche per la stima del numero degli attivi occupati osserviamo che lo stimatore sintetico è notevolmente più efficiente dello stimatore post-stratificato. L'indice $E({}_d\hat{X}_{sin})$ presenta un valore medio regionale pari a 20,5 con un valore massimo di 106,4 nella USL 1 di Milano ed un valore minimo di 1,1 nella USL 4 di Brescia.

A differenza di quanto emerso per il numero degli attivi in cerca di occupazione, tuttavia, nella stima degli attivi occupati lo stimatore sintetico mostra una distorsione modesta; (gli indici $B'({}_d\hat{X}_{sin})$ e $B''({}_d\hat{X}_{sin})$ assumono valori medi regionali pari rispettivamente a 2,0 e 36,5) e poco variabile (l'indice $B'({}_d\hat{X}_{sin})$ presenta un campo di variazione con un valore minimo di 0,1 nella USL 4 di Varese ed un valore massimo di 8,1 nella USL 4 di Brescia). Questo fatto evidenzia che il numero di attivi occupati a livello di singola USL è ben spiegato dalla struttura della popolazione per sesso e classi di età.

I risultati di questa sperimentazione, pur non essendo sufficienti per trarre conclusioni generali, suggeriscono che per piccoli domini non pianificati, come sono le USL, le stime ottenibili applicando lo stimatore sintetico sul campione di base sono più efficienti delle stime ottenibili applicando lo stimatore del rapporto post-stratificato sul campione ampliato. In taluni casi questo guadagno in efficienza ha come contropartita un'elevata distorsione. In altri casi, invece, in particolare nella stima del numero degli attivi occupati (così come di altri aggregati che dipendano essenzialmente dalla struttura per sesso ed età della popolazione del dominio) la distorsione dello stimatore sintetico rimane modesta.

Appendice: Tabelle

Tab. 1: *Indicatori di confronto tra stimatore sintetico e stimatore post-stratificato per la stima degli attivi occupati nelle USL della regione Lombardia*

USL	$_dP'$	$\varepsilon(_d\hat{X}_{sin})$	$B'(_d\hat{X}_{sin})$	$B''(_d\hat{X}_{sin})$	$\varepsilon(_d\hat{X}_{pos})$	$E(_d\hat{X}_{sin})$
			Provincia di Varese			
1	6,4	6,3	5,6	77,8	56,5	9,0
2	7,9	2,8	0,4	1,7	41,4	14,6
3	19,4	3,0	0,9	9,9	27,7	9,2
4	5,6	2,8	0,1	0,0	76,1	26,9
5	5,5	2,9	0,9	9,3	70,6	24,1
6	22,2	5,0	4,2	70,4	43,6	8,8
7	6,1	3,0	1,0	10,6	69,8	23,1
8	17,0	3,0	1,0	11,8	53,1	17,5
9	10,0	4,3	3,2	54,4	81,4	18,9
media prov.	11,1	3,7	1,9	27,3	57,8	16,9
			Provincia di Como			
1	1,5	3,1	1,8	34,8	49,0	48,0
2	4,8	3,2	2,0	38,5	59,9	18,8
3	8,9	5,9	5,4	83,2	60,9	10,3
4	22,1	2,6	0,3	1,0	23,3	9,0
5	8,0	2,6	0,7	6,8	53,4	20,3
7	12,2	2,6	0,2	0,8	37,0	14,4
8	7,6	2,6	0,5	3,5	53,9	20,7
9	18,1	3,2	1,8	32,5	28,8	9,1
10	4,0	2,7	0,8	7,9	41,0	15,4
11	4,4	2,6	0,5	3,7	55,3	21,2
12	2,4	5,3	4,5	74,4	50,0	9,5
13	0,3	6,9	6,3	84,2	30,3	19,0
media prov.	7,7	3,5	2,0	29,3	62,0	18,4
			Provincia di Sondrio			
1	13,6	5,6	1,1	3,5	42,6	7,5
2	23,8	5,5	1,1	4,0	37,6	6,8
3	32,9	5,5	0,1	0,0	28,3	5,1
4	16,8	5,5	0,4	0,5	41,8	7,6
5	12,8	5,9	1,8	9,7	51,8	8,8
media prov.	20,0	5,6	0,9	3,5	40,4	7,2

Segue Tab. 1: *Indicatori di confronto tra stimatore sintetico e stimatore post-stratificato per la stima degli attivi occupati nelle USL della regione Lombardia*

USL	$_dP'$	$\varepsilon(_d\hat{X}_{sin})$	$B'(_d\hat{X}_{sin})$	$B''(_d\hat{X}_{sin})$	$\varepsilon(_d\hat{X}_{pos})$	$E(_d\hat{X}_{sin})$
			Provincia di Milano			
1	0,8	1,5	0,2	1,1	63,3	6,4
2	0,5	1,7	0,8	20,8	32,4	75,7
3	1,8	4,8	4,5	88,2	29,1	6,0
4	1,0	5,2	4,9	90,0	67,2	12,9
5	2,1	3,3	2,9	76,7	27,8	8,3
6	3,3	1,8	0,9	28,7	80,7	45,7
7	5,1	1,8	1,0	31,3	90,2	50,2
8	2,0	2,5	1,9	62,0	90,1	36,7
9	3,0	2,1	1,4	46,2	47,5	23,1
10	1,5	1,6	0,4	5,5	62,6	39,5
11	2,4	3,1	2,7	73,9	97,6	31,6
12	3,3	1,5	0,2	1,8	78,4	50,7
13	5,6	2,4	1,8	57,7	39,6	16,4
14	3,6	1,5	0,0	0,1	91,1	59,8
15	4,9	1,9	1,1	35,2	09,6	58,9
16	3,7	2,4	1,8	60,6	91,4	38,5
17	3,8	3,6	3,3	83,5	54,7	15,2
18	1,8	4,6	4,3	89,7	84,0	18,3
19	2,3	1,7	0,6	11,4	47,1	28,4
20	1,5	6,3	6,1	94,6	01,5	16,2
21	2,5	4,4	4,2	88,8	75,6	17,1
22	1,5	3,3	3,0	79,5	57,1	17,1
23	2,9	2,9	2,6	75,6	55,2	52,8
24	36,7	2,1	1,3	37,2	17,1	8,3
25	2,5	2,1	1,5	51,7	48,9	23,2
media prov.	4,0	2,8	2,1	51,7	77,6	4,3
			Provincia di Bergamo			
1	2,4	2,8	0,9	9,7	32,3	7,0
2	4,4	2,8	0,9	10,3	58,0	0,5
3	9,8	4,6	3,8	68,6	48,5	0,7
4	4,7	3,8	2,7	48,8	36,7	9,6
5	13,4	3,2	1,9	35,8	30,7	9,5
6	26,1	4,2	3,2	57,4	31,8	7,6
7	13,9	3,6	2,4	47,0	43,9	2,3
8	5,8	3,5	2,3	41,3	64,0	8,1
9	12,6	2,9	1,3	19,5	57,8	9,8
10	7,0	2,7	0,2	0,8	75,9	8,5
media prov.	10,0	3,4	0,1	33,9	58,0	8,4

Segue Tab. 1: *Indicatori di confronto tra stimatore sintetico e stimatore post-stratificato per la stima degli attivi occupati nelle USL della regione Lombardia*

USL	$_dP'$	$\varepsilon(_d\hat{X}_{sin})$	$B'(_d\hat{X}_{sin})$	$B''(_d\hat{X}_{sin})$	$\varepsilon(_d\hat{X}_{pos})$	$E(_d\hat{X}_{sin})$
			Provincia di Brescia			
1	6,9	2,7	1,5	31,2	74,3	8,0
2	4,6	3,0	2,0	45,7	92,2	0,9
3	5,1	2,4	1,0	16,1	50,8	1,0
4	8,5	8,4	8,1	91,8	9,0	1,1
5	9,9	3,0	1,9	40,9	21,5	7,3
6	5,4	2,2	0,1	0,2	44,7	9,9
7	9,3	2,4	0,8	12,0	30,9	2,9
8	31,6	2,7	1,3	25,6	21,2	8,0
9	4,2	4,9	4,4	80,6	97,8	0,1
10	7,6	7,2	6,9	91,5	63,4	8,8
11	7,0	3,4	2,6	58,9	69,1	0,3
media prov.	9,1	3,8	2,8	45,0	52,2	6,2
			Provincia di Pavia			
1	36,9	4,5	3,4	56,1	8,1	1,8
2	34,0	5,0	4,1	68,8	7,7	1,5
3	29,1	3,1	0,9	9,1	5,6	1,8
media prov.	33,3	4,2	2,8	44,7	7,1	1,7
			Provincia di Cremona			
1	47,5	3,8	0,5	1,6	21,9	5,7
2	11,7	5,5	4,2	56,8	44,1	8,0
3	40,8	3,8	0,7	3,0	23,1	6,0
media prov.	33,3	4,4	1,8	20,5	29,7	6,6
			Provincia di Mantova			
1	10,2	5,5	4,2	59,1	53,6	9,8
2	13,5	3,9	1,7	17,7	36,4	9,2
3	38,3	4,3	2,2	25,0	25,5	5,9
4	13,1	4,0	1,2	9,4	58,3	4,7
5	12,7	3,7	0,6	2,4	41,0	1,1
6	12,2	4,1	1,9	21,5	68,5	6,8
media prov.	16,7	4,2	2,0	22,5	47,2	1,2
media reg.	10,6	3,6	2,0	36,5	59,0	0,5

Tab. 2: *Indicatori di confronto tra stimatore sintetico e stimatore post-stratificato per la stima degli attivi in cerca di occupazione nelle USL della regione Lombardia*

USL	$_dP'$	$\varepsilon(_d\hat{X}_{sin})$	$B'(_d\hat{X}_{sin})$	$B''(_d\hat{X}_{sin})$	$\varepsilon(_d\hat{X}_{pos})$	$E(_d\hat{X}_{sin})$
			Provincia di Varese			
1	6,4	19,3	0,6	0,1	75,6	3,9
2	7,9	19,3	6,2	10,4	59,2	3,1
3	19,4	20,3	3,2	2,4	37,6	1,9
4	5,6	20,6	11,1	29,0	90,5	4,4
5	5,5	27,4	15,6	32,5	97,9	3,6
6	22,2	20,8	4,4	4,4	52,5	2,5
7	6,1	19,6	0,9	0,2	93,1	4,7
8	17,0	23,3	16,8	51,9	62,0	2,7
9	10,0	41,4	32,4	61,3	92,7	2,2
media prov.	11,1	23,6	10,2	21,3	73,5	3,2
			Provincia di Como			
1	1,5	42,0	40,2	91,4	71,4	4,1
2	4,8	26,2	20,4	60,7	84,4	3,2
3	8,9	22,7	6,1	7,3	78,9	3,5
4	22,1	25,6	19,4	57,8	32,6	1,3
5	8,0	21,4	11,1	26,6	68,7	3,2
6	5,8	24,5	9,5	15,2	82,5	3,4
7	12,2	37,0	26,3	50,5	60,7	1,6
8	7,6	34,5	23,3	45,7	73,9	2,1
9	18,1	37,3	26,7	51,1	42,0	1,1
10	4,0	33,8	22,4	44,1	90,2	2,7
11	4,4	21,0	1,4	0,4	83,9	4,0
12	2,4	28,0	23,1	68,0	88,4	3,2
13	0,3	47,1	45,7	94,3	25,2	4,8
media prov.	7,7	30,8	21,2	47,2	91,0	2,9
			Provincia di Sondrio			
1	13,6	34,4	0,3	0,0	67,8	2,0
2	23,8	51,1	26,6	27,0	62,6	1,2
3	32,9	50,6	26,0	26,3	62,2	1,2
4	16,8	33,0	6,2	3,5	61,4	1,9
5	12,8	48,1	44,0	83,7	72,4	1,5
media prov.	20,0	43,4	28,1	28,1	65,3	1,6

Segue Tab. 2: *Indicatori di confronto tra stimatore sintetico e stimatore post-stratificato per la stima degli attivi in cerca di occupazione nelle USL della regione Lombardia*

USL	$_dP'$	$\varepsilon(_d\hat{X}_{sin})$	$B'(_d\hat{X}_{sin})$	$B''(_d\hat{X}_{sin})$	$\varepsilon(_d\hat{X}_{pos})$	$E(_d\hat{X}_{sin})$
			Provincia di Milano			
1	0,8	65,4	63,2	93,4	92,2	2,9
2	0,5	40,5	37,9	87,7	56,6	3,9
3	1,8	44,2	41,7	89,1	58,9	1,3
4	1,0	30,7	27,8	81,7	96,1	3,1
5	2,1	43,2	40,7	88,8	53,0	1,2
6	3,3	35,0	32,2	84,9	02,6	2,9
7	5,1	31,2	28,3	82,2	69,2	2,2
8	2,0	53,1	50,8	91,5	06,6	2,0
9	3,0	55,5	53,2	92,0	69,1	1,2
10	1,5	24,6	21,2	74,4	78,9	3,2
11	2,4	34,4	31,7	84,6	07,0	3,1
12	3,3	10,7	4,0	14,1	90,9	8,5
13	5,6	17,1	12,6	54,3	39,5	2,3
14	3,6	27,1	26,1	92,2	93,8	3,5
15	4,9	22,9	21,3	87,6	03,9	4,5
16	3,7	14,4	11,2	59,7	95,8	6,7
17	3,8	26,1	22,6	76,6	63,0	2,4
18	1,8	12,0	5,7	19,4	00,5	8,3
19	2,3	12,8	8,6	46,2	63,3	4,9
20	1,5	24,3	20,4	73,8	15,3	4,8
21	2,5	10,5	1,2	2,0	80,0	7,6
22	1,5	43,7	41,8	89,0	80,8	1,9
23	2,9	13,1	9,3	49,5	40,1	10,7
24	36,7	13,0	9,4	47,9	19,0	1,5
25	2,5	17,6	15,6	75,3	55,8	3,2
media prov.	4,0	28,9	25,6	69,5	89,3	3,9
			Provincia di Bergamo			
1	2,4	39,4	29,5	55,9	64,7	4,2
2	4,4	22,2	13,7	38,4	83,6	3,8
3	9,8	21,8	12,8	34,6	62,8	2,9
4	4,7	34,4	31,5	83,9	66,8	1,9
5	13,4	27,9	15,4	30,4	45,2	1,6
6	26,1	19,9	5,5	7,8	37,3	1,9
7	13,9	27,8	15,1	29,8	66,2	2,4
8	5,8	19,8	3,8	3,6	80,2	4,1
9	12,6	20,2	0,3	0,0	73,9	3,7
10	7,0	47,5	38,5	65,4	06,5	2,2
media prov.	10,0	28,1	16,6	35,0	78,7	2,9

Segue Tab. 2: *Indicatori di confronto tra stimatore sintetico e stimatore post-stratificato per la stima degli attivi in cerca di occupazione nelle USL della regione Lombardia*

USL	$_dP'$	$\varepsilon(_d\hat{X}_{sin})$	$B'(_d\hat{X}_{sin})$	$B''(_d\hat{X}_{sin})$	$\varepsilon(_d\hat{X}_{pos})$	$E(_d\hat{X}_{sin})$
			Provincia di Brescia			
1	6,9	30,4	23,0	57,2	87,2	2,9
2	4,6	17,7	5,0	7,9	03,6	5,9
3	5,1	16,0	4,1	6,7	63,7	4,0
4	8,5	35,6	33,9	91,0	35,0	1,0
5	9,9	16,7	2,4	2,1	42,5	2,5
6	5,4	16,7	2,5	2,3	64,0	3,8
7	9,3	23,4	19,4	69,1	40,7	1,7
8	31,6	18,1	6,0	10,8	24,7	1,4
9	4,2	21,6	11,8	29,9	28,7	6,0
10	7,6	29,2	21,6	54,7	87,3	3,0
11	7,0	31,4	24,2	59,3	85,4	2,7
media prov.	9,1	23,4	14,0	35,5	69,3	3,2
			Provincia di Pavia			
1	36,9	24,2	4,8	3,8	28,2	1,2
2	34,0	25,4	17,1	45,2	24,9	1,0
3	29,1	37,2	24,2	42,4	33,5	0,9
media prov.	33,3	28,9	15,4	30,5	28,8	1,0
			Provincia di Cremona			
1	47,5	34,1	2,1	0,4	37,9	1,1
2	11,7	56,9	34,9	37,6	86,9	1,5
3	40,8	31,6	8,5	7,2	42,1	1,3
media prov.	33,3	40,9	15,2	15,0	55,6	1,3
			Provincia di Mantova			
1	10,2	37,7	11,3	9,0	99,2	2,6
2	13,5	50,5	28,6	32,1	81,2	1,6
3	38,3	31,2	4,8	2,3	40,6	1,3
4	13,1	31,6	2,7	0,7	84,8	2,7
5	12,7	30,8	8,5	7,6	67,3	2,2
6	12,2	31,2	4,9	2,5	98,6	3,2
media prov.	16,7	35,5	10,2	9,0	78,6	2,3
media reg.	10,6	29,6	18,6	42,5	78,6	3,0

5

PROPOSTE IN TEMA DI STIME TEMPESTIVE DEI DISOCCUPATI

Fabio Corradi, Luigi Fabbris, Ippolito Sanetti e Alberto Zuliani *

1. *Premessa*

Nel 1959, quando per l'indagine corrente sulle forze di lavoro fu adottata la cadenza trimestrale, l'Istat intese assegnare ad essa carattere di osservatorio permanente dell'evoluzione a breve termine del mercato del lavoro.

Le modifiche apportate all'indagine nel corso degli anni, comprendenti la ristrutturazione e l'ampliamento del questionario avvenuti nel 1977 e nel 1984 ed il sovracampionamento introdotto per ottenere stime più analitiche, l'hanno resa sempre più vasta e complessa (si è passati da 1500 comuni e 80.000 famiglie a circa, rispettivamente, 2.300 e 140.000), dilatando i tempi di raccolta ed elaborazione dei dati e, in definitiva, di diffusione dei risultati.

In effetti, nonostante che nel tempo sia migliorato il rapporto dei rilevatori con le famiglie, sia aumentata la disponibilità degli amministratori locali e l'efficienza degli uffici provinciali di statistica, coadiuvati dagli uffici regionali dell'Istat e, infine, la procedura di elaborazione sia stata completamente automatizzata, i risultati, sotto forma sia di dati individuali che di tavole statistiche di sintesi, divengono disponibili circa tre mesi dopo la data prevista per la rilevazione. Più precisamente:

(a) i dati individuali sono disponibili su supporto magnetico non prima di 90 giorni e soltanto per particolari sub-campioni;
(b) i dati analitici relativi alle regioni e province per le quali è stato effettuato il sovracampionamento sono anch' essi predisposti non prima di 90 giorni;
(c) la pubblicazione monografica trimestrale, contenente i risultati analitici della rilevazione, a livello nazionale e regionale, è approntata dopo 100 giorni.

Per questo, da tempo, sia all'interno dell'Istat che tra gli utilizzatori dei risultati dell'indagine, è avvertita l'esigenza di disporre di stime precoci relativamente alle principali variabili dell'occupazione e della disoccupazione.

---

* Il capitolo è frutto della collaborazione degli autori. Quando alla sua stesura, I. Sanetti ha curato la sez. 1, F. Corradi ha curato le sezz. 3 e 4, mentre le sezz. 2 e 5 sono dovute congiuntamente a tutti gli autori.

Occorre considerare, infine, che uno dei motivi per i quali si accumula il ritardo nella disponibilità dei risultati risiede nella circostanza che gli enti periferici che effettuano la raccolta dei dati restituiscono i questionari all'Istat a distanza più o meno lunga rispetto alla data prevista quale termine per la rilevazione.

Diviene allora importante valutare quale livello di informazioni sia necessario per poter procedere alla determinazione e quindi alla diffusione di stime precoci. Un criterio statistico di carattere generale è legato al raggiungimento di un prefissato livello di significatività di tali stime, per cui la possibilità di diffonderle viene in definitiva a dipendere dalla disponibilità ad abbassare la loro attendibilità[1].

2. *Obiettivi dello studio*

In questo capitolo si analizzano le *performances* di stime ottenute a diversa distanza fra il termine della rilevazione e il momento di completa disponibilità dell'informazione raccolta, con l'intento di determinare se sia possibile procedere all'effettuazione di stime precoci per le principali grandezze concernenti l'offerta di lavoro, per il complesso del territorio nazionale e, eventualmente, per le diverse ripartizioni territoriali.

Per la sua rilevanza relativamente ad interventi di politica economica e del lavoro, si fissa l' attenzione sulla stima di livello " numero di persone in cerca d'occupazione " .

L'analisi riguarda i dati aggregati a livello comunale relativi alle quattro rilevazioni del 1988, poiché è relativamente ad essi che l'Istat ha reso disponibile l'informazione relativa alla settimana di ricezione dei questionari a partire dal momento di chiusura della rilevazione. Nella Tab. 1 sono riportate, per le quattro occasioni del 1988, le percentuali di residenti nei comuni appartenenti al campione secondo la settimana durante la quale i rispettivi questionari sono pervenuti all'Istat. Va aggiunto che i dati sono stati forniti dall'Istat sotto forma di stime riferite allo strato di appartenenza dei comuni campionati [2].

Tali stime si considerano come osservazioni in tempi successivi della variabile " numero di persone in cerca d'occupazione " , osservazione che al trascorrere dal tempo della chiusura della rilevazione, e quindi con la disponibilità di un campione sempre più ampio di informazioni, diventano sempre più accurate fino al momento in cui l'Istat considera chiusa la raccolta dei questionari.

---

1 Herzel, commentando una precedente versione del lavoro, ha suggerito scherzosamente che la qualità attesa delle stime è funzione del quadrato del tempo intercorso dalla rilevazione dei dati. Suo peraltro, è il suggerimento di utilizzare stimatori basati sulle differenze, che viene approfondito nella successiva sez.4.

2 L'espansione all'universo rappresentato dallo strato è stata effettuata con la metodologia standard utilizzata dall'Istat per l'indagine sulle forze di lavoro.

Tab. 1 : *Percentuali cumulate di residenti appartenenti agli strati secondo la settimana dalla conclusione della rilevazione, durante la quale i comuni campione hanno restituito i questionari all' Istat, per area geografica. Anno 1988*

Aree geografiche	Settimane 1	2	3	4	5	6	7	8(*)
				Rilevazione di gennaio				
Nord Ovest	19,9	31,9	45,0	88,7	98,7	99,0	99,3	99,4
Nord est	12,9	61,3	63,6	97,7	100,0	100,0	100,0	100,0
Centro	39,9	64,9	97,6	98,4	98,8	99,8	100,0	100,0
Sud	19,9	25,9	29,2	30,5	58,8	96,6	98,0	99,5
Isole	32,7	47,3	58,4	95,9	98,4	99,0	99,5	100,0
Italia	24,0	43,8	56,1	78,6	89,0	98,7	99,1	99,6
				Rilevazione di aprile				
Nord Ovest	27,2	48,4	77,0	77,5	80,2	92,7	100,0	100,0
Nord est	21,6	36,2	66,4	99,4	99,9	100,0	100,0	100,0
Centro	32,9	90,3	94,4	98,4	100,0	100,0	100,0	100,0
Sud	16,6	27,3	30,1	31,5	31,5	62,9	98,7	99,0
Isole	31,9	65,9	73,3	74,9	99,1	99,3	99,3	100,0
Italia	25,3	51,0	66,3	73,7	85,5	97,7	99,7	99,9
				Rilevazione di luglio				
Nord Ovest	5,6	23,3	51,0	60,5	62,6	63,3	74,5	99,1
Nord est	21,1	45,4	48,0	48,1	96,2	98,2	99,9	100,0
Centro	3,9	30,9	60,9	63,6	69,6	70,5	71,0	72,4
Sud	8,4	18,7	29,9	32,2	34,7	62,8	63,1	89,9
Isole	4,8	29,7	48,6	79,8	83,8	86,2	88,3	99,2
Italia	5,1	24,0	46,4	54,1	56,9	73,2	76,9	91,9
				Rilevazione di ottobre				
Nord Ovest	5,4	18,5	39,6	63,9	77,4	92,6	98,1	99,7
Nord est	3,6	23,7	48,6	60,3	68,2	95,8	95,8	95,8
Centro	20,0	41,5	60,3	66,9	96,0	97,8	98,2	99,5
Sud	11,7	18,2	25,1	38,2	66,1	66,7	98,5	99,7
Isole	13,2	35,5	51,4	54,9	67,7	98,8	99,1	100,0
Italia	10,4	25,8	43,0	56,4	75,3	88,5	98,0	98,8

(*) Settimana alla quale l'Istat ha considerato chiusa la raccolta di questionari.

Considerazioni di merito relativamente alle esigenze dell' Istituto di statistica, e in definitiva dei decisori pubblici, hanno portato a valutare stime disaggregate per le cinque ripartizioni geografiche comunemente usate:

Nord Ovest : Piemonte, Valle d'Aosta, Lombardia, Liguria.
Nord Est : Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna;
Centro : Toscana, Umbria, Marche, Lazio;
Sud : Abruzzi, Molise, Campania, Puglia, Basilicata, Calabria;
Isole : Sicilia, Sardegna.

In questo ambito, il problema della stima precoce del numero di persone in cerca d'occupazione (o di altri aggregati concernenti le forze di lavoro) presenta alcune peculiarietà:

(a) i dati accumulati nelle successive settimane sono influenzati dall'ordine d'arrivo dei questionari all'Istat, il quale dipende dalla speditezza con la quale operano i singoli comuni e dalla circostanza che le province, collettori delle informazioni comunali, talora attendono che tutti i comuni dell'area di competenza abbiano restituito i questionari compilati prima di trasmetterli all'ISTAT;
(b) i problemi derivanti dalla rotazione del campione di famiglie possono essere ignorati, perché si impiegano dati aggregati al livello del comune. Anche la rotazione annuale dei comuni può essere ignorata, se le stime si basano sulle sole informazioni raccolte nell'occasione in cui le stime stesse vengono prodotte.

Prima di procedere, introduciamo alcune notazioni di carattere generale. La frequenza di persone in cerca d'occupazione verrà indicata con y, il numero di residenti con x; la sovrapposizione di un accento circonflesso (^) indica una stima, mentre una barra sovrapposta (-) nella medesima posizione indica la media aritmetica della variabile pertinente. Con il deponente destro si indica, in generale, la quantità di informazione a cui si fa riferimento. In particolare i deponenti p o q indicano se si fa riferimento, rispettivamente, alla parte di campione già pervenuta o non ancora pervenuta all'Istat. I deponenti 1 o 0, pure collocati a destra, indicano l'occasione per la quale vengono prodotte stime precoci e, rispettivamente, quella immediatamente precedente. La presenza di più di un deponente delimita la quantità di informazione risultante dal prodotto logico dei deponenti implicati, mentre l'assenza di deponente indica che l'informazione corrispondente è quella relativa all'universo. Notazioni particolari saranno definite allorché impiegate.

3. *Stima del "numero di persone in cerca d'occupazione" mediante lo stimatore rapporto*

Per un aggregato quale il "numero di persone in cerca di occupazione", l' usuale metodo di stima si basa sullo stimatore di un quoziente, ben noto in letteratura (Cochran, 1953, pp. 154-184). Esso è definito quando a ciascuna unità campionaria sono associati una quantità nota x, denominata variabile accessoria, ed un'altra, y, non conosciuta prima che il campiona-

mento sia avvenuto. Nel caso in esame, esse sono rispettivamente la popolazione residente ed il numero di persone in cerca di occupazione negli strati. Lo stimatore basato sul rapporto è allora definito facendo riferimento all'insieme s di unità campionate come:

$$\hat{y}_p = x(y_p/x_p). \qquad (1)$$

Due sono gli approcci principali per effettuare inferenza relativamente ad esso (Thomsen e Tesfu, 1988,pp. 370-374). Il primo desume la varianza dello stimatore dalla sua distribuzione campionaria, dato il piano di campionamento. Ad esempio, per un campione casuale semplice e per una numerosità campionaria sufficientemente elevata, lo stimatore di un rapporto appare leggermente distorto e con varianza approssimativamente pari a:

$$V = (N/n)(N-n) \sum_{i \in p} (y_i-(\bar{y}/\bar{x})x_i)^2/(N-1), \qquad (2)$$

dove N ed n sono rispettivamente il numero di unità statistiche nella popolazione e nel campione.

Il secondo approccio distingue due componenti nella stima, quella desumibile in base alle unità del campione già pervenute e quella corrispondente all'informazione pertinente alle unità complementari del campione. Trattandosi di ammontari avremo:

$$\hat{y} = \hat{y}_p + \hat{y}_q. \qquad (3)$$

Si definisce altresì un modello stocastico di relazione fra le singole determinazioni della variabile d'interesse e quella/e accessoria/e. Un semplice modello lineare che viene generalmente ìutilizzato (Cochran 1977, pp. 158), e che riconduce allo stimatore rapporto, tranne per aspetti concernenti il metodo di stima, è il seguente:

$$y_i = \beta x_i + e_i \qquad (i \in p \text{ e } x > 0)$$

$$E(e_i|x_i) = 0;\ E(e^2_i|x_i) = \sigma^2(x_i);\ E(e_i e_j) = 0,\ i \neq j. \qquad (4)$$

Da questa formulazione modellistica, discende che la stima relativa alla variabile d'interesse, per la componente $\hat{y}_q$, viene effettuata tramite la relazione stabilita dal modello, a sua volta stimata attraverso il metodo dei minimi quadrati ponderati, applicato ai dati resi disponibili dal campione (y_i, x_i; $i \varepsilon p$). Vari stimatori della varianza di $\hat{y}_p$ sono stati proposti in letteratura e studiati in relazione alle caratteristiche che essi assumono allorché il modello (4) non risulti correttamente specificato (Royall e Cumberland, 1981). La stima più largamente accettata è la seguente:

$$V= (N/n^2)(N-n)(\bar{x}_q\bar{x}/\bar{x}_p^2) \sum_{i \in p} (y_i-(\bar{y}_p/\bar{x}_p)x_i)^2/(1-(x_i/n\bar{x}_p)). \qquad (5)$$

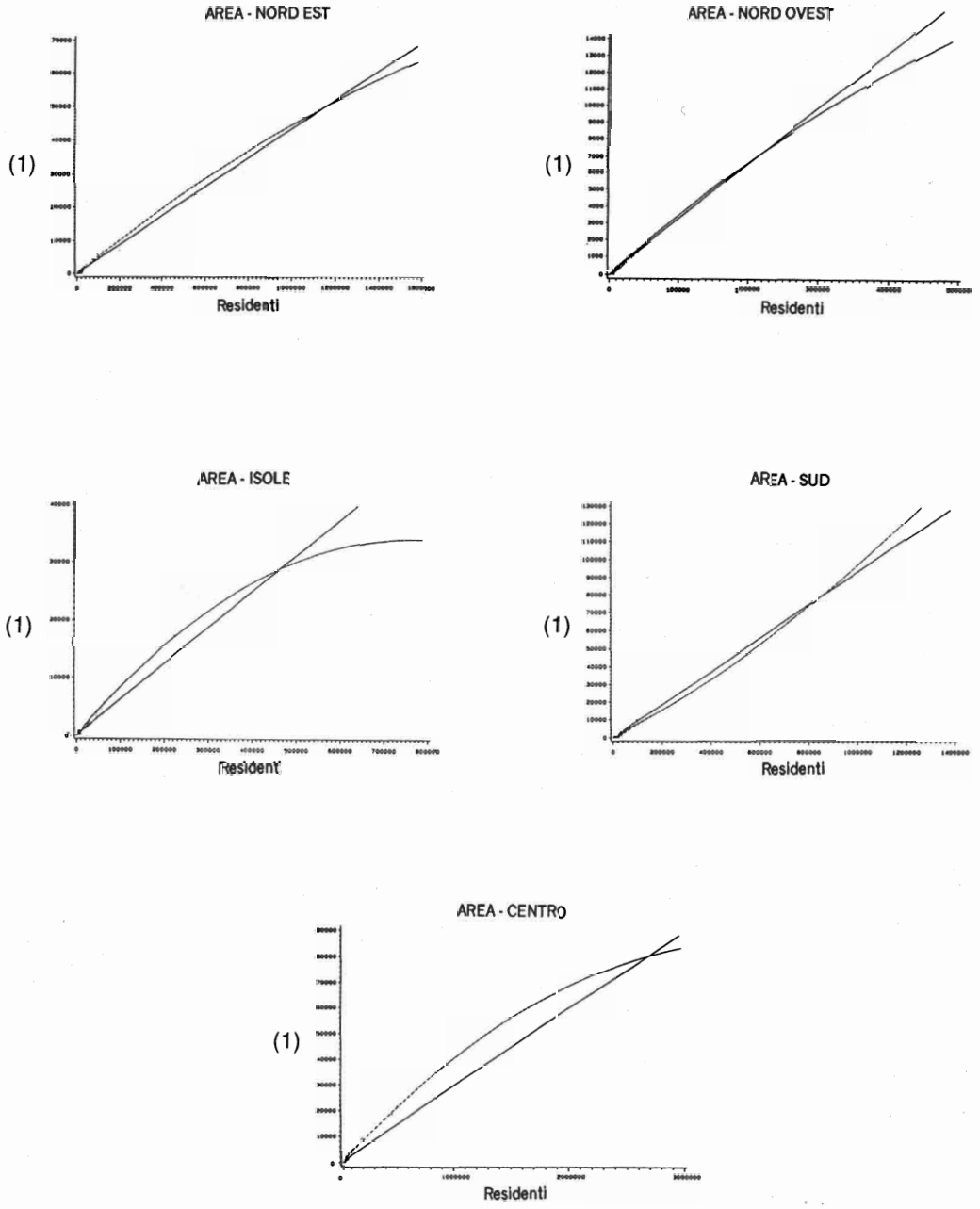
Valori bassi della varianza si ottengono per valori elevati di $\bar{x}_p$, cosicché stime intervallari più ampie si ottengono allorché le informazioni sulle quali ci si basa siano corrispondenti a bassi valori della variabile accessoria. Per una prima esplorazione circa l'adeguatezza dei modelli (1) e (4) per il caso che ci interessa, nella Fig. 1 vengono presentati, distintamente per aree geografiche, i dati relativi alla rilevazione di Gennaio 1988 (le altre occasioni investigate non mostrano, in proposito, apprezzabili differenze), insieme alla rappresentazione delle equazioni di regressione relative al modello proporzionale e, a fini di confronto, a quello con componente aggiuntiva quadratica. Osservando gli andamenti delle due equazioni di regressione, si nota che essi si differenziano essenzialmente a causa delle (poche) osservazioni concernenti le grandi città, le quali risultano, perciò, particolarmente influenti per la stima dei parametri del modello quadratico. Sebbene quest'ultimo segua con maggiore precisione i dati, esso non appare particolarmente conveniente ai nostri fini in quanto, qualora le grandi città fossero presenti nel campione, la stima perderebbe in larga misura le caratteristiche sostanziali di precocità. Infatti, risulterebbe rappresentata gran parte della popolazione residente e si potrebbero ottenere stime quasi definitive. Qualora, invece, esse fossero assenti, la stima della componente quadratica risulterebbe poco robusta.

Per ottenere stime dell'ammontare della variabile "numero di persone in cerca d'occupazione" si è applicato il modello (1), in base ai dati pervenuti nelle prime otto settimane dalla conclusione della rilevazione. Nella Fig. 2 sono riportate le stime intervallari, al 95%, utilizzando la varianza stimata secondo la (2), relativamente alle occasioni esaminate (Gennaio, Aprile, Luglio, Ottobre 1988), per l'Italia nel suo complesso. Nella Tab. 2, per una migliore valutazione dei risultati ottenuti, sono compendiati gli scarti relativi percentuali fra le stime precoci ed il numero di persone in cerca d'occupazione stimato definitivamente, in maniera distinta per area geografica e per ciascuna delle occasioni.

Considerando i risultati per l'intero territorio nazionale, si nota che gli intervalli di confidenza proposti comprendono quasi sempre, fin dalle prime settimane, il valore della stima finale. Va però osservato che la rilevante ampiezza di questi intervalli, specialmente nella fase iniziale, comporta che le prime stime siano del tutto inutilizzabili. Ciò risulta evidente tenendo conto del fatto che, fra le successive occasioni di rilevazione analizzate, si è avuta una variazione relativa massima pari a -3% (fra le occasioni di gennaio ed aprile 1988). In questa situazione, un intervallo di confidenza al 95%, con un'ampiezza pari al 5% del numero di persone in cerca d'occupazione, come approssimativamente risulta per le stime effettuate nelle prime settimane dalla conclusione della rilevazione, non consente alcuna attendibile induzione.

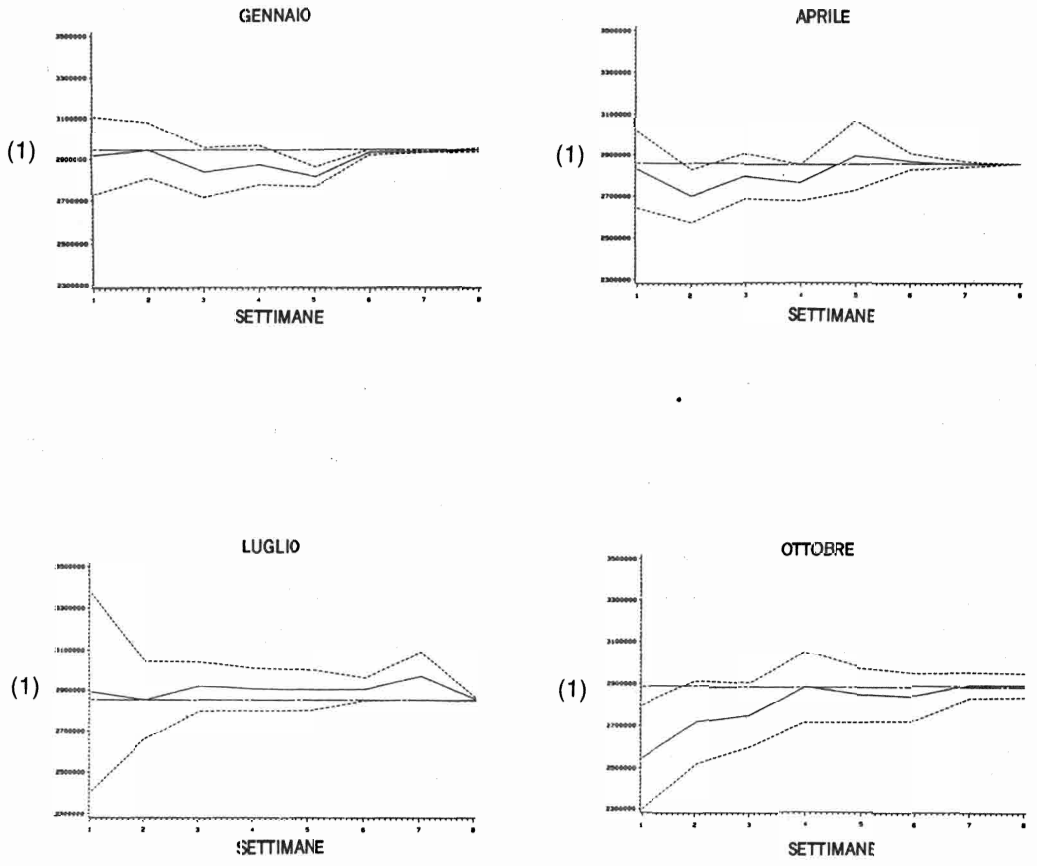
Può essere interessante considerare se le stime precoci colgano la tendenza all'aumento o alla diminuzione del numero di persone in cerca d'occupazione fra due successive rilevazioni. Tale tendenza può essere valutata, per le occasioni di aprile, luglio e ottobre, tenendo conto della riga continua orizzontale per due successive sezioni della Fig. 2. Tale linea

Fig. 1: *Residenti e persone in cerca di occupazione (DIS) per area geografica. Valori relativi agli strati ed equazioni di regressione per i modelli lineare e quadrico, con intercetta nulla. Rilevazione di gennaio 1988.*



(1) In cerca di occupazione

Fig. 2: *Stime precoci intervellari del numero delle persone in cerca d'occupazione, ottenute mediante lo stimatore rapporto ed effettuate nelle successive settimane dalla conclusione della rilevazione per le occasioni di gennaio, aprile, luglio e ottobre 1988 (1- α = 0,95)*
linee tratteggiate = intervalli di confidenza;
linea intera = stima puntuale;
linea parallela all'asse delle ascisse = stima finale.



(1) In cerca di occupazione

Tab. 2 : *Scarti relativi percentuali fra la stima precoce del numero di persone in cerca d'occupazione ottenuta con lo stimatore del rapporto e la stima definitiva per grande ripartizione geografica. Anno 1988.*

Aree geografiche	Settimane 1	2	3	4	5	6	7
			Rilevazione di gennaio				
Nord Ovest	- 6,2	-10,2	-13,5	1,8	0,5	0,3	0,1
Nord est	- 9,0	3,7	2,8	- 0,7	0,0	0,0	0,0
Centro	14,9	9,5	- 0,7	- 0,1	- 0,1	0,0	0,0
Sud	- 8,2	- 4,5	- 6,6	- 7,1	-11,4	- 1,0	- 0,8
Isole	11,2	8,7	6,1	0,5	- 0,5	- 0,5	- 0,5
Italia	- 1,0	- 0,1	- 3,6	- 2,4	- 4,3	- 0,4	- 0,3
			Rilevazione di aprile				
Nord Ovest	-11,5	-11,1	4,1	3,7	2,7	3,2	0,0
Nord est	0,1	- 1,3	2,8	0,2	0,1	0,0	0,0
Centro	11,3	- 1,9	- 1,3	- 0,7	0,0	0,0	0,0
Sud	- 6,7	- 8,5	- 6,8	- 8,7	2,3	0,4	0,4
Isole	9,2	0,2	- 0,1	- 0,7	0,4	0,4	0,4
Italia	- 1,0	- 5,5	- 2,0	- 3,1	1,4	0,3	- 0,1
			Rilevazione di luglio				
Nord Ovest	0,7	- 3,2	3,3	1,4	0,7	0,1	5,8
Nord est	-32,7	- 1,1	0,1	1,5	1,2	1,3	0,5
Centro	16,8	0,7	0,9	1,0	0,1	- 0,1	- 0,4
Sud	5,7	4,3	3,0	3,0	3,6	8,4	8,3
Isole	- 3,6	- 6,4	2,2	0,2	0,1	- 0,6	0,1
Italia	1,1	- 0,2	2,2	1,7	1,6	3,1	3,9
			Rilevazione di ottobre				
Nord Ovest	- 8,6	- 2,3	- 8,3	- 8,2	5,2	2,5	0,1
Nord est	-16,2	-10,8	- 4,8	- 0,2	- 0,5	0,0	0,0
Centro	- 0,9	13,0	5,8	4,1	- 0,6	0,1	- 0,3
Sud	-24,1	-17,3	-11,5	1,1	- 5,9	- 6,1	0,1
Isole	2,9	- 0,1	2,1	0,1	0,5	- 0,1	0,0
Italia	-11,9	- 6,0	-4,9	- 0,2	- 1,4	- 1,9	- 0,1

rappresenta infatti il numero di persone in cerca d'occupazione stimato definitivamente al termine della precedente occasione di rilevazione.
La tendenza è stata colta nei casi in cui la variazione occorsa è risultata ampia. Ad esempio, nell'occasione di aprile, allorché si è registrata una diminuzione del 3% del numero di persone in cerca d'occupazione rispetto al

dato di gennaio, tale tendenza era colta già dalla stima precoce ottenuta dopo una sola settimana dalla conclusione della rilevazione e poi da tutte le successive. Al contrario, per la rilevazione di luglio che ha fatto registrare una diminuzione pari ad appena lo 0,13% rispetto al dato del trimestre precedente, la tendenza non viene colta dalle stime precoci, pur essendo tutte quante assai prossime al valore finale. Per l'occasione di ottobre, che ha fatto registrare un aumento nel numero di persone in cerca d'occupazione pari all'1,1%, il comportamento delle stime precoci non risulta in nitido.

Considerando ora gli scarti delle stime precoci dalla stima finale riportati nella Tab. 2, congiuntamente con le percentuali di copertura della popolazione residente corrispondenti alle informazioni via via pervenute, riportate nella Tab. 1, si nota che differenze inferiori al 2% in valore assoluto rispetto alla stima finale si ottengono quando la popolazione residente negli strati di appartenenza dei comuni, per i quali l'informazione risulta disponibile, rappresenta circa l'85% di quella finale. Tale tasso di copertura si raggiunge, nelle diverse occasioni considerate e a seconda delle aree geografiche di riferimento, alla quarta o quinta settimana dalla conclusione della rilevazione. E' questo un risultato apprezzabile, considerato l'attuale ritardo nella diffusione dei risultati da parte dell'Istat e la semplicità del modello adottato per la stima precoce. Esso non può tuttavia essere considerato pienamente soddisfacente, avendo ridotto soltanto del 15% circa la numerosità programmata per arrivare ad un risultato sufficientemente accurato (uno scarto di ± 2% dalla stima finale può essere empiricamente fissato come soglia, tenendo conto di quanto già detto circa la variabilità del fenomeno sotto esame).

Per migliorare i risultati, sempre nell' ambito di modelli che sfruttano l'informazione raccolta con una sola rilevazione, i campioni di comuni acquisiti settimanalmente sono stati post-stratificati in base a tre ulteriori variabili accessorie: ampiezza demografica (secondo cinque classi di residenti: 1-5.000, 5.001-10.000, 10.001-50.000, 50.001-200.000, oltre 200.000); attività economica prevalente (agricoltura, industria, servizi); altitudine (montagna, collina, pianura). Palesemente, si tratta di variabili note *a priori*, quindi per l'appunto accessorie, considerato che costituiscono criteri di stratificazione del campione di comuni. L'obiettivo generalmente perseguito utilizzando questa tecnica (Holt e Smith, 1979,pp. 33-35) è quello di produrre dei modelli locali per sottoinsiemi di dati che presentino maggiore omogeneità, e quindi minore variabilità, rispetto alla variabile d'interesse: in questo modo, le stime dovrebbero risultare maggiomente efficienti.

Le prove effettuate hanno però messo in luce che:

(a) una eccessiva segmentazione in gruppi di post-stratificazione rende vano il tentativo di miglioramento della stima, se non si hanno almeno circa 30 comuni campione per sottoinsieme;

(b) l'ausilio della post-stratificazione non riduce in modo apprezzabile la variabilità della stima.

Risultati assai modesti sono stati ottenuti post-stratificando secondo la ampiezza demografica del comune, per la scarsità di punti campione relativi

alle città più grandi. I migliori risultati, in quest'ambito di prove, sono stati ottenuti post-stratificando secondo l'attività economica prevalente. Essi sono presentati nella Tab. 3.

Tab. 3 : *Scarti relativi percentuali fra la stima precoce del numero di persone in cerca d'occupazione ottenuta con lo stimatore del rapporto post stratificando secondo la attività economica prevalente e la stima definitiva, per grande ripartizione geografica. Anno 1988.*

Aree geografiche	Settimane 1	2	3	4	5	6	7
			Rilevazione di gennaio				
Nord Ovest	- 4,9	-15,9	-24,6	0,6	0,8	0,6	0,3
Nord est	19,7	14,2	12,5	- 1,0	0,0	0,0	0,0
Centro	35,9	15,5	- 1,0	- 0,1	- 0,1	0,0	0,0
Sud	-25,4	-15,0	-20,9	-23,0	-24,1	- 1,3	- 1,2
Isole	-10,7	8,4	2,7	4,4	- 0,7	- 0,7	- 0,7
Italia	- 1,1	- 0,2	- 3,7	- 2,7	- 4,8	- 0,4	- 0,3
			Rilevazione di aprile				
Nord Ovest	21,3	1,2	15,5	14,8	11,8	6,7	0,0
Nord est	-18,0	-35,2	9,7	0,5	0,2	0,0	0,0
Centro	46,6	0,5	1,0	- 1,0	0,0	0,0	0,0
Sud	-19,6	-16,7	-12,6	-19,5	- 8,1	- 0,6	- 1,3
Isole	11,7	- 2,8	- 3,6	- 4,4	0,8	0,5	0,5
Italia	- 1,1	- 6,3	- 2,2	- 3,4	1,7	0,2	- 0,1
			Rilevazione di luglio				
Nord Ovest	16,3	-12,6	7,3	10,8	11,4	7,3	18,1
Nord est	- 79,1	-28,5	- 0,4	1,0	0,7	2,6	1,0
Centro 85,4	16,2	9,1	9,8	0,8	0,2	- 0,6	
Sud	-14,7	-32,6	3,9	2,8	19,4	15,5	15,6
Isole	-41,3	-32,7	3,9	2,8	1,1	- 0,3	- 0,4
Italia	- 5,6	- 1,4	1,6	1,5	1,4	3,5	4,1
			Rilevazione di ottobre				
Nord Ovest	- 8,6	- 1,4	-13,9	- 2,0	17,2	6,9	- 0,4
Nord est	-32,6	-55,1	-30,8	1,8	3,2	0,0	0,0
Centro	60,6	69,4	17,8	15,7	- 0,2	- 0,8	- 0,3
Sud	- 72,7	-53,4	-32,5	9,0	- 4,3	- 4,6	1,6
Isole	15,2	10,3	7,0	3,0	4,2	- 0,4	0,0
Italia	- 9,5	- 5,6	- 4,0	0,5	- 1,5	- 2,2	- 0,1

Si nota che la post-stratificazione ha un effetto decisamente distorcente sulle stime precoci effettuate nelle prime settimane specie se disaggregate per aree geografiche, mentre apporta un lieve miglioramento relativamente alle stime per l'Italia nel complesso (ma solo perché si compensano gli errori che si presentano a livello delle sub-aree). Inoltre, si registra un miglioramento e per le stime effettuate con la disponibilità di almeno il 50% dell'informazione finale (valutata in termini di popolazione residente), cioé, a seconda delle occasioni, dopo tre- quattro settimane dal termine della rilevazione (e sempre qualora, in ciascuna delle partizioni ottenute post-stratificando, il numero di punti campione risulti sufficiente)[3].

4. *Stime ottenute utilizzando stimatori comprendenti informazioni relative alla rilevazione precedente*

In questo sezione, si dà conto dei risultati ottenuti utilizzando stimatori compositi (Wolter, 1979), cioé, in questo caso, combinazioni lineari di stimatori basati su informazioni relative a due successive rilevazioni. Si è scelto di operare su due sole rilevazioni, anziché tentare un approccio basato su serie storiche (di lunghezza sufficiente a consentire la stima di modelli regressivi e/o autoregressivi), perché si ritiene opportuno privilegiare l'informazione che si rende via via disponibile, raccolta relativamente alla occasione per la quale si vogliono produrre stime. L'informazione può essere utilizzata per aggiornare la stima basata sull'informazione completa, di cui si dispone con riferimento all'occasione precedente. Il procedimento che, in definitiva, si adotta rende poco importante la conoscenza del processo che governa la serie e risulta abbastanza semplice. Descriviamolo in dettaglio.

Si parte dalla stima finale del numero di persone in cerca d'occupazione relativa alla precedente occasione di rilevazione (tempo 0) e la si assume come stima iniziale per l'occasione attuale. Di mano in mano che le informazioni derivanti dalla nuova rilevazione si rendono disponibili, vengono utilizzate per stimare la variazione del numero di persone in cerca d'occupazione intervenuta fra le due occasioni contigue. Tale stima viene utilizzata per aggiornare quella iniziale ed ottenere stime precoci per l'occasione attuale, con una sensibile riduzione della varianza ad esse relativa. Infatti, poiché la stima effettuata al tempo 0 è nota con certezza al tempo 1, la variabilità della stima precoce sarà da attribuirsi unicamente alla variazione intervenuta, la cui varianza può essere espressa come:

$$\mathrm{var}(\hat{y}_1-\hat{y}_0) = \mathrm{var}(\hat{y}_1) + \mathrm{var}(\hat{y}_0) - 2\mathrm{cov}(\hat{y}_1,\hat{y}_0).$$

---

3 Altri tentativi di post-stratificazione sono stati effettuati raggruppando i comuni in base ad alcune altre variabili quali il tasso di disoccupazione nell'agricoltura, nell'industria e nei servizi. Ciò senza ottenere miglioramenti apprezzabili delle stime, in quanto i gruppi che via via si formavano, in particolare nelle prime settimane dalla conclusione della rilevazione, non realizzavano gradi di copertura, in termini di punti campione, adeguati per garantire una varianza sufficientemente contenuta.

Nella ragionevole ipotesi che la varianza della variabile d'interesse resti pressoché costante fra due occasioni contigue e pari, diciamo, a $var(\hat{y})$, l'efficienza sarà funzione del coefficiente di correlazione. Avremo pertanto che:

$$var(\hat{y}_1-\hat{y}_0) = 2var(\hat{y}) - 2var(\hat{y})r(\hat{y}_1,\hat{y}_0) = 2var(\hat{y})(1-r(\hat{y}_1,\hat{y}_0)),$$

che tende ad annullarsi se la correlazione tra due occasioni è prossima ad 1.

Prima di procedere alla proposta finale di stimatore, si sono dovuti affrontare due problemi .

Il primo sorge per effetto della rotazione del campione: per i comuni nuovi entrati non è evidentemente possibile procedere alla stima della variazione del numero di persone in cerca d'occupazione. Naturalmente, il problema si pone solamente per le occasioni in cui viene effettuata la rotazione. Nel caso in esame, tale circostanza si è verificata con l'occasione di luglio ed ha riguardato circa un terzo dei comuni con meno di 20.000 abitanti. Alcune possibili soluzioni sono le seguenti :

(a) assumere come stima del numero di persone in cerca d'occupazione, per gli strati rappresentati dalle unità campionarie di primo stadio subentrate nell'occasione 1, quella finale definita al tempo 0 per le unità sostituite. Conseguentemente, per la frazione del campione ruotata la stima della variazione risulta nulla. In definitiva lo stimatore vale:

$$\hat{y}_{1,p} = x_0[(1-P)(y_{0,nr}/x_{0,nr} + y_{1,nr,p}/x_{1,nr,p}-y_{0,nr,p}/x_{0,nr,p}) + Py_{0,r}/x_{0,r}], \qquad (6)$$

dove:

– P indica la frazione di residenti appartenenti agli strati dai comuni ruotati,

– r ed nr indicano, rispettivamente, che si tratta di informazioni relative alla parte di campione ruotata e non ruotata fra le due successive occasioni;

(b) estendere agli strati rappresentati dalle unità campionarie di primo stadio subentrate nell'occasione 1, la stima della variazione del numero di persone in cerca d'occupazione definita per la parte di campione non ruotato. In questo caso, lo stimatore risulta essere:

$$\hat{y}_{1,p} = x_0(y_0/x_0 + y_{1,nr,p}/x_{1,nr,p} - y_{0,nr,p}/x_{0,nr,p}); \qquad (7)$$

(c) stimare la variazione del numero di persone in cerca d'occupazione, per gli strati rappresentati dalle unità campionarie di primo stadio subentrate nell'occasione 1, con riferimento alla stima finale definita al tempo 0 per le unità sostituite, nell'ipotesi che l'appartenenza al medesimo strato ne consenta la scambiabilità. In questo caso, lo stimatore risulta:

$$\hat{y}_{1,p} = x_0[(1-P)(y_{0,nr}/x_{0,nr} + y_{1,nr,p}/x_{1,nr,p} - y_{0,nr,p}/x_{0,nr,p}) + P(y_{0,r}/x_{0,r} + y_{1,r,p}/x_{1,r,p} - y_{0,r,p}/x_{0,r,p})]. \qquad (8)$$

Nella Tab.4 sono riportate le stime finali dei tassi (percentuali) di persone in cerca d'occupazione rispetto alla popolazione residente per le due occa-

Tab. 4: *Rapporti percentuali fra il numero di persone in cerca d'occupazione e la popolazione residente, calcolati per le occasioni di aprile e luglio per la parte ruotata e non ruotata del campione, per area geografica. Anno 1988.*

Aree geografiche	Parte ruotata			Parte non ruotata		
	aprile	luglio	differenza	aprile	luglio	differenza
Nord Ovest	2,358	2,382	0,023	2,018	2,879	0,816
Nord Est	3,017	2,669	-0,348	2,773	2,780	0,007
Centro	4,726	4,420	-0,306	4,166	5,552	1,385
Sud	7,533	7,156	-0,377	8,435	10,273	1,837
Isole	7,628	8,457	0,828	7,700	9,456	1,765
Italia	4,632	4,521	-0,110	4,033	5,008	0,975

sioni contigue di aprile e luglio 1988, nonché la loro differenza: oltre alla consueta disaggregazione per aree geografiche, è fornita anche l'informazione separatamente per gli strati rappresentati dalla parte di campione ruotata e da quella che ha mantenuto la stessa composizione relativamente alle unità di primo stadio, e ciò per i soli strati soggetti al meccanismo della rotazione dei comuni. Anche se l'esplorazione è limitata ad una sola occasione di rotazione, d'altronde l'unica disponibile, le differenze riscontrabili fra i risultati ottenuti per le parti di campione ruotato e non ruotato appaiono apprezzabili. In molti casi, infatti si ottiene un segno diverso per la variazione tra aprile e luglio. In definitiva, tra le possibili soluzioni appena illustrate appare prudente adottare l'ipotesi (a), anche in considerazione del fatto che i comuni interessati alla rotazione hanno un numero di residenti poco elevato cosicchè la variazione del numero di persone in cerca d'occupazione ad essi imputabile non risulterebbe comunque particolarmente influente sulla stima complessiva.

Il secondo problema riguarda il peso da assegnare alla variazione del numero di persone in cerca d'occupazione ottenuta in base all'informazione che via via si rende disponibile per l'occasione 1, nella determinazione della stima precoce $\hat{y}_{1,p}$. Sembra infatti ragionevole tenere conto del grado di rappresentatività dell'informazione, riferita al tempo 1, che si rende via via disponibile rispetto a quella definitiva. Un possibile peso può essere costruito tenendo conto della somiglianza fra due distribuzioni relative cumulate secondo l'ampiezza demografica degli strati: quella del numero di residenti negli strati rappresentati dalla frazione di comuni per i quali l'informazione risulta acquisita al momento della stima, e quella del numero di residenti in tutti gli strati. Alla stima della variazione si può assegnare un peso pari a:

$$w_p = 1 - D_p = 1 - \text{Sup}\,|F(x) - F_p(x)|, \qquad (9)$$

dove:
$F(x)$ è il numero dei residenti negli strati con popolazione $\leq x$, rapportato al numero complessivo di residenti;
$F_p(x)$ è il numero dei residenti negli strati con popolazione $\leq x$, per i quali l'informazione risulta disponibile al momento della stima, rapportato al numero complessivo di residenti negli stessi strati.

Il peso w_p può assumere valori compresi fra zero ed uno, raggiungendo i due estremi rispettivamente nei casi di massima discordanza fra $F_p(x)$ e $F(x)$ e di loro coincidenza.

In definitiva, lo stimatore risulta:

$$\hat{y}_{1,p} = x_0[(1-P)(y_{0,nr}/x_{0,nr} + w_p(y_{1,nr,p}/x_{1,nr,p} - y_{0,nr,p}/x_{0,nr,p})) + Py_{0,r}/x_{0,r}]. \qquad (10)$$

Nella Tab. 5 sono riportati gli scarti delle stime precoci del numero di

Tab. 5: *Scarti relativi percentuali fra la stima tempestiva del numero di persone in cerca d'occupazione ottenuta con lo stimatore composito (10) e la stima definitiva, per area geografica. Anno 1988.*

Aree geografiche	Settimane 1	2	3	4	5	6	7	8
			Rilevazione di aprile					
Nord Ovest	2,2	1,1	1,6	1,4	0,8	1,7	0,0	0,0
Nord est	2,6	0,1	0,1	0,2	0,1	0,0	0,0	0,0
Centro	- 6,0	- 0,7	- 0,4	- 0,1	0,0	0,0	0,0	0,0
Sud	1,6	- 0,1	- 1,4	- 2,1	- 0,3	- 0,4	- 0,3	0,0
Isole	- 7,7	0,4	2,1	1,9	0,2	0,1	0,1	0,0
Italia	- 1,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0
			Rilevazione di luglio					
Nord Ovest	6,9	5,2	1,7	2,1	2,1	2,1	2,8	0,9
Nord est	4,1	3,6	0,5	2,3	2,3	1,8	1,3	1,4
Centro	15,9	1,9	0,4	1,2	1,3	1,2	0,8	0,9
Sud	- 4,0	1,2	2,4	2,2	3,9	0,6	0,7	- 0,7
Isole	-12,7	- 6,3	2,0	- 2,4	- 1,4	- 1,6	- 2,1	- 0,8
Italia	1,0	0,0	1,7	1,1	1,9	0,6	0,5	0,0
			Rilevazione di ottobre					
Nord Ovest	- 6,7	- 0,4	0,2	- 2,5	1,3	0,0	- 0,4	0,0
Nord est	1,7	4,4	5,3	2,3	0,8	0,2	0,2	0,2
Centro	- 1,3	0,7	0,4	- 3,0	1,0	1,5	0,3	0,3
Sud	-22,8	- 9,1	- 7,4	- 2,9	1,1	1,0	- 0,3	0,0
Isole	- 3,5	- 7,1	- 3,9	- 5,1	- 3,3	0,0	0,0	0,0
Italia	-10,6	- 4,5	-3,0	- 2,8	0,2	0,6	- 0,2	0,1

persone in cerca di occupazione ottenute utilizzando lo stimatore (10) da quelle finali, distintamente per aree geografiche.

Rispetto ai risultati ottenuti con i precedenti stimatori, si registra un consistente miglioramento. Per la rilevazione di aprile, la più favorevole fra quelle analizzate, la stima precoce alla prima settimana, ottenuta con appena il 25% dell'informazione finale, dà un errore di previsione di appena l'1% rispetto alla stima definitiva al livello dell'intero territorio nazionale, cogliendo subito la tendenza alla diminuzione. Nel caso più problematico, relativo alla rilevazione di ottobre, si giunge ad un'approssimazione dello 0,2%, con il 75% circa d'informazione finale.

5. *Conclusioni*

Alcune valutazioni conclusive si possono trarre confrontando i risultati delle stime precoci ottenute tramite lo stimatore (1) e lo stimatore (10), in termini di scarti relativi delle stime stesse rispetto a quelle finali, già presentati nelle Tabb. 3 e 5. Tale confronto può essere effettuato per le rilevazioni di aprile, luglio e ottobre. Si registra un pressoché generale miglioramento dell'accuratezza delle stime ottenute con lo stimatore composito (10), circostanza questa particolarmente evidente per le stime effettuate nell' area Sud, la meno facilmente prevedibile in tutte le occasioni considerate, e per l'occasione di ottobre. L'indicazione che se ne trae, quindi, è senz'altro in favore dell' uso contemporaneo dell'informazione raccolta in due occasioni contigue, anche se approfondimenti sono necessari in particolare per ottenere stime intervallari. Sviluppi in questa direzione, basati essenzialmente sul filtro di Kalman, già sono stati tentati in contesti differenti (Bordignon e Trivellato, 1989) e potranno essere oggetto di ulteriori approfondimenti con riferimento al tema qui trattato.

6

INDAGINE SULLE FORZE DI LAVORO E STIMATORI PER CAMPIONI RUOTATI

*Daniela Cocchi**

1. *Lo schema di rotazione nell'indagine italiana sulle forze di lavoro*

1.1. *Descrizione del procedimento di rotazione*

La rilevazione trimestrale sulle forze di lavoro, (nel seguito RTFL) utilizza, nelle diverse occasioni di indagine, un campione parzialmente ruotato. In ciascuno strato, cioè, metà delle famiglie intervistate al tempo *t* è stata intervistata anche al tempo t-1. Lo schema di rotazione può essere descritto più compiutamente come segue. Ciascuna famiglia estratta a costituire il campione rimane coinvolta nell'indagine in un arco di 6 rilevazioni, per un totale di 15 mesi. L'intervista ha luogo per le prime due occasioni; nel corso delle due indagini successive il contatto viene sospeso; viene poi ripreso per le ultime due rilevazioni, in modo che la prima e la terza intervista, come anche la seconda e la quarta, si svolgano rispettivamente a distanza di un anno. Secondo un linguaggio ormai consolidato, si tratta quindi di uno schema di rotazione del tipo 2-2-2 (vedi il cap.1).

Quando viene realizzata la rotazione, le famiglie campionate in ciascuno strato in una qualunque occasione sono attribuite a 4 'sezioni', ciascuna contenente lo stesso numero di famiglie. A ciascuna sezione corrispondono le famiglie intervistate per la prima, seconda, terza e quarta volta, tenuto ovviamente conto del periodo di interruzione dell'intervista. Maggiori dettagli sulla predisposizione delle sezioni, oltre che sulle differenze, a questo riguardo, tra ciò che accade rispettivamente per gli strati autorappresentativi (AR) e non autorappresentativi (NAR), sono in Cocchi e Castellini (1988).

1.2. *Impiego attuale della rotazione e il problema della qualità dei dati*

I vantaggi che derivano dalla scelta di un campione parzialmente ruotato

---

* Ringrazio S. Corsi e G. Guagnano per l'assistenza prestata nella fase di calcolo. Ringrazio inoltre M. Castellini e P.D. Falorsi per le utili discussioni.

sono, per il momento, esclusivamente organizzativi: superate le difficoltà iniziali di costruzione del campione, in ciascun comune estratto soltanto una parte del campione (il 50%) deve essere sostituita ad ogni trimestre. Lo schema di rotazione garantisce il 50% di sovrapposizione anche per la distanza di un anno, mentre a 9 e a 15 mesi il tasso di sovrapposizione è del 25% (salvi i fenomeni di *attrition*).

La sovrapposizione parziale del campione comporta inoltre che le stime delle variazioni da un periodo all'altro sono più affidabili di quelle ottenute per mezzo di campioni completamente rinnovati in ciascuna occasione (vedi, ad esempio, Duncan e Kalton, 1987), anche se ciò non è per ora documentato in modo preciso. Infatti, la valutazione quantitativa del miglioramento dell'efficienza delle stime dovuto alla rotazione presuppone la costruzione di stimatori delle caratteristiche di interesse che tengano conto della rotazione del campione e si basa sul confronto delle misure di variabilità tra essi e gli stimatori fino ad ora impiegati, che trascurano l'aspetto di rotazione.

I vantaggi dell'impiego delle informazioni sulla rotazione (che obbliga peraltro a risalire ai dati individuali o almeno a organizzare la costruzione delle stime a partire dalle sezioni) possono essere rilevanti, soprattutto nell'ipotesi di dimensioni campionarie più ridotte delle attuali, per ottenere stime attendibili a livello di sub-aree o per ridurre la variabilità delle stime meno affidabili.

1.3. *La modifica necessaria per tenere conto della rotazione*

Per ora, nelle singole occasioni, all'interno di uno strato il conteggio degli individui che possiedono una particolare caratteristica avviene senza tenere conto della sezione di riferimento, in altre parole marginalizzando rispetto alla variabile 'codice di sezione'. Dal punto di vista operativo, tuttavia, non si deve fare molto per iniziare a sfruttare l'aspetto della rotazione.

Nel presente studio si incentra l'attenzione sull'impatto dello schema di rotazione sulla costruzione delle stime e si propone uno stimatore alternativo a quello correntemente usato. A tale scopo, sono stati predisposti programmi che trattano i dati individuali tenendo sempre conto della sezione di appartenenza.

La base di tutti gli sviluppi che tengono conto della rotazione consiste infatti nella costruzione, per ogni occasione e per ogni strato h = 1, ..., H, di 4 stimatori del rapporto, denotati con il suffisso R, separatamente per i due sessi, a = 1,2, del tipo correntemente utilizzato dall'Istat per la produzione di stime (vedi il cap. 2). In tal modo la sezione di appartenenza, vale a dire l'occasione di intervista, è identificabile ed è quindi possibile mettere in relazione le coppie di stimatori ottenuti sulla base delle stesse famiglie intervistate a distanze temporali prefissate. Lo stimatore per la p-ma sezione può essere definito come:

$$ {}_at_{p,R}(y)_h = \frac{{}_at_p(y)_h}{{}_at_p(z)_h} \; {}_at(\zeta)_h \;, \; p = 1, \ldots, 4 \;, \qquad (1) $$

dove

$$ {}_at(\zeta)_h = \sum_{\lambda=1}^{{}_aN_h} 1 \qquad (2) $$

rappresenta la popolazione, per sesso, dello strato h, mentre

$$ {}_at_p(y)_h = \sum_{i=1}^{{}_an_{h,p}} {}_ay_i \qquad (3) $$

corrisponde, per ciascun sesso, al conteggio degli individui che, estratti nella p-esima sezione dello strato campionato, possiedono la caratteristica di interesse, denotata in modo dicotomico: y={0,1} . Infine,

$$ {}_at_p(z)_h = \sum_{i=1}^{{}_an_{h,p}} 1 \qquad (4) $$

rappresenta il totale degli intervistati, separatamente per i due sessi, nella p-esima sezione dello strato h-esimo.

La relazione tra gli stimatori di sezione e lo stimatore correntemente utilizzato dall'Istat è illustrata formalmente nelle (2.1)-(2.9) di Cocchi (1990) e si basa sull'ipotesi che ogni sezione sia un campione casuale della popolazione dello strato, di numerosità pari a 1/4 del numero di unità complessivamente campionate. Solamente in tal caso, infatti,

$$ {}_at_R(y)_h = \frac{1}{4} \sum_{p=1}^{4} {}_at_{p,R}(y)_h. \qquad (5) $$

Per descrivere la costruzione degli stimatori di sezione nell'ambito dello schema di rotazione italiano, nella Tab. 1 è presentata la successione degli stimatori di sezione (1) che possono essere calcolati nell'arco di 7 rilevazioni.

Tab. 1: *Stimatori di sezione in base alla rotazione 2-2-2* (a)

t-6	t-5	t-4	t-3	t-2	t-1	t
$t_4(y)_{t-6}$	$t_1(y)_{t-5}$	$t_2(y)_{t-4}$			$t_3(y)_{t-1}$	$t_4(y)_t$
$t_3(y)_{t-6}$	$t_4(y)_{t-5}$	$t_1(y)_{t-4}$	$t_2(y)_{t-3}$			$t_3(y)_t$
	$t_3(y)_{t-5}$	$t_4(y)_{t-4}$	$t_1(y)_{t-3}$	$t_2(y)_{t-2}$		
		$t_3(y)_{t-4}$	$t_4(y)_{t-3}$	$t_1(y)_{t-2}$	$t_2(y)_{t-1}$	
$t_2(y)_{t-6}$			$t_3(y)_{t-3}$	$t_4(y)_{t-2}$	$t_1(y)_{t-1}$	$t_2(y)_t$
$t_1(y)_{t-6}$	$t_2(y)_{t-5}$			$t_3(y)_{t-2}$	$t_4(y)_{t-1}$	$t_1(y)_t$

(a) Per semplicità, sono omessi i suffissi che denotano sesso e strato.

In essa, i suffissi da 1 a 4 indicano le 4 sezioni, intendendo con ciò che una sezione venga intervistata per la prima, seconda, terza o quarta volta. I suffissi t-i, i = 0,..., 6, indicano invece le occasioni temporali. Le sezioni possono quindi essere riconosciute, in senso longitudinale, secondo lo schema 2-2-2, seguendo la successione dei suffissi che vanno da 1 a 4 (il suffisso 1 successivo a un 4 denota un rinnovo di sezione). Considerando uno strato e facendo riferimento a due generiche occasioni consecutive, t e t-1, le due sezioni che compaiono per la prima e per la terza volta nell'indagine al tempo t-1 si sovrappongono alle sezioni che compaiono per la seconda e per la quarta volta nell'indagine al tempo t.

Il fatto che, nella pratica della elaborazione dei risultati dell'indagine, non si sia tenuto finora conto della rotazione fa sorgere alcune riserve sulla risultante qualità dei dati. In particolare, si può temere che le famiglie assegnate ad una sezione non vi rimangano per i periodi stabiliti e che avvengano perciò sostituzioni non controllate, con la conseguenza che gli abbinamenti dei dati individuali nel tempo possano risultare inferiori a quanto teoricamente implicato dal piano di rotazione e talora fallaci (vedi il cap.7).

2. *La versione per la RTFL dello stimatore composto correntemente usato nelle indagini degli istituti centrali di statistica*

La rotazione parziale del campione è molto diffusa nelle indagini periodiche svolte dagli uffici di statistica di altri Paesi. Il numero di occasioni in cui la famiglia rimane nel campione, nonché il periodo di abbandono temporaneo dell'indagine, sono molto variabili, e sono fortemente legati alla cadenza dell'indagine (trimestrale o mensile). Ad esempio, la CPS (Current Population Survey) mensile degli Stati Uniti adotta uno schema del tipo 4-8-4 (vedi U.S. Bureau of Census, 1978), mentre nella rilevazione mensile canadese sulle forze di lavoro ogni sezione rimane nel campione per sei mesi consecutivi prima di essere sostituita (vedi Statistics Canada, 1977).

Le proprietà di uno schema di rotazione sono legate alle caratteristiche dello stimatore adottato, la cui struttura consegue da particolari assunzioni. In generale, uno stimatore che tenga conto della parziale sovrapposizione nel tempo delle unità campionate viene proposto come una media ponderata di componenti riguardanti, rispettivamente, la parte del campione che resta nell'indagine dalla rilevazione precedente e la parte che non si sovrappone ad essa. Dal punto di vista della facilità di impiego è inoltre auspicabile che, al susseguirsi delle occasioni di indagine, il valore della stima sia facilmente aggiornabile sulla base della valutazione ottenuta all'occasione precedente e di calcoli effettuati sul solo campione corrente.

Lo stimatore 'composto' utilizzato nella CPS statunitense e sperimentato in Canada ha una forma riconducibile alle caratteristiche menzionate. Sulla base di un particolare peso k, su cui si discuterà in seguito, per sesso e per strato si può costruire (vedi Hansen , Hurwitz, Nisselson e Steinberg, 1955) l' espressione:

$$t_c(y)_t = k[t_c(y)_{t-1} + t_s(y)_t - t_s(y)_{t-1}] + (1-k)t(y)_t \,, \qquad (6)$$

dove $t(y)_t$ è la stima del totale di interesse sulla base dell'intero campione al tempo t, cioè la (5), ed i suffissi c ed s denotano rispettivamente i termini 'composto' e 'relativo alle unità sovrapposte'. Quindi, con $t_c(y)_{t-1}$ viene denotato il calcolo della stessa (6) al periodo precedente.

In particolare, per lo schema italiano è sufficiente, con riferimento a due periodi consecutivi, riprendere gli stimatori di sezione (1) descritti nella Tab. 1 per ottenere le restanti componenti della (6), nell'ipotesi che ogni sezione comprenda esattamente 1/4 degli individui campionati:

$$t_s(y)_t = (t_{4,R}(y)_t + t_{2,R}(y)_t)/2 \,, \qquad (7)$$

$$t_s(y)_{t-1} = (t_{3,R}(y)_{t-1} + t_{1,R}(y)_{t-1})/2 \,. \qquad (8)$$

Gli stimatori composti (6) possono poi essere sommati per sesso e per strato, analogamente a quanto avviene per gli stimatori del rapporto correntemente impiegati dall'Istat.

La (6) costituisce una giustapposizione ragionevole di elementi desumibili da un campione organizzato in sezioni, ed è facilmente utilizzabile in forma ricorsiva. Per comprenderne meglio la struttura, si può notare che essa consiste in una rielaborazione dell'espressione

$$t_c(y)_t = t^*(y)_t + k\ t^*(y)_{t-1} + k^2\ t^*(y)_{t-2} + \ldots + k^t\ t^*(y)_0 = \sum_{i=0}^{t} k^i\ t^*(y)_{t-i}, \qquad (9)$$

in cui viene isolata la somma degli elementi relativi alla parte sovrapposta

$$t^*(y)_t = k[t_s(y)_{t-1}] + (1-k)t(y)_t \qquad (10)$$

e dove $t^*(y)_0 = t_c(y)_0$ è una stima iniziale per il periodo che precede il primo per il quale si voglia calcolare una stima composta. La varianza della (9) si può quindi esprimere come:

$$V[t_c(y)_t] = \sum_{i=0}^{t} k^{2i}\, V[t^*(y)_{t-1}] + 2 \sum_{j=1}^{t} \sum_{i=0}^{t} k^{2i+j}\ C[t^*(y)_{t-i-1}, t^*(y)_{t-i-j}] \,, \qquad (11)$$

e facendo tendere t all'infinito, utilizzando la somma di una serie, si ottiene l' espressione

$$V[t_c(y)_t] = \frac{1}{1-k^2}\left\{V[t^*(y)_t] + 2\sum_{j=1}^{t} k^j\, C[t^*(y)_t, t^*(y)_{t-j}]\right\}. \qquad (12)$$

Per poter calcolare la (12) nel contesto delle diverse rilevazioni, si dovrà specificare la struttura nel tempo delle varianze e delle covarianze. Al fine di mantenere semplicità nei calcoli, si formulano solitamente le seguenti ipotesi:

(1) identica varianza degli stimatori di sezione nel tempo e qualunque sia l'occasione di intervista, cioè:

$$V[t_p(y)_t] = V[t_p(y)_{t-j}] = \sigma^2, \quad p = 1,\ldots,4 \quad j = 0,\ldots,t, \tag{13}$$

(2) costanza della covarianza tra stimatori riferiti alle stesse sezioni, ad intervalli temporali prefissati, cioè:

$$C[t_p(y)_t, t_h(y)_{t-j}] = \rho_j \sigma^2, \quad p,h = 1,\ldots,4 \quad j = 1,\ldots,t, \tag{14a}$$

se c'è sovrapposizione di unità tra t e t-j, e

$$C[t_p(y)_t, t_h(y)_{t-j}] = 0, \tag{14b}$$

altrimenti.

Il calcolo della varianza della (6) prendendo le mosse dalla (9) fa sì che ulteriori sviluppi della (12) possano essere effettuati soltanto seguendo lo schema di rotazione di volta in volta adottato. Nel caso italiano, in base alle (14) potranno essere diversi da zero i coefficienti ρ_1, ρ_3, ρ_4 e ρ_5, in quanto ρ_2 non è calcolabile, come non lo sono le correlazioni tra periodi distanti oltre 15 mesi. Infatti, riprendendo la Tab. 1 e ponendosi al generico tempo t, per il calcolo della covariazione a un trimestre di distanza si utilizzeranno le sezioni che compaiono per la quarta e seconda volta al tempo t e per la terza e prima volta al tempo t-1. Allo stesso modo, per la covariazione a un anno di distanza serviranno le sezioni che compaiono per la quarta e terza volta al tempo t e per la seconda e prima volta al tempo t-4. Per i due importanti intervalli temporali pari al trimestre e all'anno, per ogni periodo t, la struttura di rotazione offre quindi due replicazioni della realizzazione dello stimatore di sezione, nell'ipotesi che non vi siano distorsioni dovute alla rotazione campionaria. Per la RTFL, d'altra parte, la cadenza trimestrale della rilevazione e il basso numero di interviste che viene fatto a una stessa famiglia inducono a ritenere che il fenomeno della distorsione imputabile alla permanenza delle famiglie nel campione, che si riscontra ad esempio nella CPS statunitense, non sia rilevante (vedi il cap. 9). Continuando l'esame degli stimatori di sezione si ha inoltre una sola possibilità di abbinamento sia per la distanza tra interviste pari a nove mesi, con riferimento alla sezione che compare per la terza volta al tempo t e per la seconda al tempo t-3, sia per quanto riguarda la distanza di 15 mesi, per la sezione che compare per la quarta volta al tempo t e per la prima al tempo t-5.

La varianza della (6), a partire dalla (12), sulla base delle sole (13) e (14), è pari a:

$$t_c(y)_t = \frac{\sigma^2}{8(1-k^2)}\ 2+2k^2+4k^3-\rho_1(2k+4k^2+2k^3)+$$

$$+\rho_3(k^3-3k^5-2k^6)+\rho_4(2k^5+4k^6+2k^7)-\rho_5(k^5+2k^6+k^7) \qquad (15)$$

mentre la varianza di una variazione tra due trimestri,

$$d_c(y)_t = t_c(y)_t - t_c(y)_{t-1}, \qquad (16)$$

risulta:

$$V[d_c(y)_t] = \frac{\sigma^2}{2(1-k^2)}\ 4(1-k^4) + \rho_1(-2-4k+4k^3+2k^4) +$$

$$+\rho_3(2k^7-k^6-4k^5+2k^4+2k^3-k^2) + \rho_4(-2k^8+4k^6-2k^4) + \rho_5(k^8-2k^6+k^4) \qquad (17)$$

Le dimostrazioni, riportate nell'Appendice, sono state elaborate, in una prima stesura, nella tesi di laurea di Castellini. Esse discendono facilmente ricordando che le varianze della (7) e della (8) sono, in virtù della (13):

$$V[t_s(y)_t] = V[t_s(y)_{t-1}] = \frac{\sigma^2}{2},$$

mentre le covarianze corrispondenti, per le (14) e tenendo sempre presente gli abbinamenti della Tab. 1, si calcolano in modo simile alla

$$C[t_s(y)_t, t_s(y)_{t-1}] = C[(t_4(y)_t+t_2(y)_t)/2, (t_3(y)_{t-1} + t_1(y)_{t-1})/2] = \rho_1 \frac{\sigma^2}{2}.$$

Vale la pena di segnalare la contraddizione tra l'apparente semplicità dello stimatore (6) e la portata delle ipotesi che vengono tacitamente accettate.

La (15) e la (17) mostrano come eventuali riduzioni della variabilità delle stime ottenute impiegando stimatori composti derivano dal gioco tra la struttura di correlazione ed il peso k. Infatti, le varianze degli stimatori di tipo (5) correntemente usati, se rimangono valide le ipotesi (13) e (14), possono essere espresse come

$$V[t(y)_t] = \frac{\sigma^2}{4} \qquad (18)$$

$$V[d(y)_t] = V[t(y)_t - t(y)_{t-1}] = \frac{\sigma^2(1-0,5\rho_1)}{2}, \qquad (19)$$

e perciò, a seconda della struttura di correlazione valutata per ciascuna

variabile, si può ricercare il peso k che massimizza l'efficienza dello stimatore composto rispetto a quello tradizionale.

3. *Altri stimatori che tengono conto della rotazione parziale del campione*

La proposta (6) costituisce la più diffusa utilizzazione della sovrapposizione parziale delle unità nelle indagini periodiche. Ricordiamo che gli sviluppi di calcolo richiesti dalla struttura di rotazione della RTFL non sono certo i più complessi. Infatti, gli stimatori che essa comporta sono 'ad un livello', nel senso che i dati rilevati al tempo t riguardano solo quell'occasione e non momenti precedenti, mentre stimatori che tengono conto di più livelli sono stati considerati, ad esempio, da Eckler (1955), Manoussakis (1977) e Wolter (1979).

Non mancano però in letteratura sviluppi legati alla (6), che mirano a ulteriori miglioramenti dell'efficienza delle stime tenendo conto del contributo della covariazione *tra altri* elementi a disposizione, ad esempio sezioni consecutive nel tempo, e quindi formate da famiglie diverse, che in tempi diversi hanno eguale anzianità nel campione. Un esempio in tal senso sono i cosiddetti stimatori AK, proposti da Gurney e Daly (1965) ed ulteriormente sviluppati da Huang ed Ernst (1981) per gli USA, da Kumar e Lee (1985) per il Canada e da Russo (1990) per l'Italia, ricercando, assieme al peso k, un ulteriore peso A sulla base del quale costruire una media ponderata simile alla (6). Nel caso degli Stati Uniti, ad esempio, viene dato un ulteriore peso ai gruppi che sono nel campione per la prima e quinta volta, vale a dire ai casi rispettivamente dell'intervista esente da distorsione da rotazione e della *ripresa* del contatto dopo un anno di interruzione.

La proposta della (6) non è stata accompagnata soltanto da applicazioni pratiche. Uno studio teorico di Rao e Graham (1964) mostra, ad esempio, che tale stimatore presenta alcune interessanti caratteristiche di ottimalità, come la correttezza nel senso del disegno. Inoltre, per tale stimatore, considerando congiuntamente la dimensione della popolazione, quella del campione, il tasso di sovrapposizione ed il numero di occasioni di permanenza in un ciclo di campionamento, si possono determinare le relazioni ottimali tra questi elementi.

La struttura di media ponderata, con pesi ottimi, di componenti legate alla parte sovrapposta e di componenti legate alla parte non sovrapposta è anche la forma del miglior stimatore lineare corretto, ottenuto come funzione lineare degli stimatori di sezione, per una qualunque funzione lineare delle caratteristiche oggetto di interesse. Senza dover infatti tener conto dei dettagli del disegno campionario specifico, che nel caso dello stimatore (6) impediscono di procedere oltre la scrittura della (12), la stima dei livelli o delle variazioni o di totali su un periodo diventa un problema di analisi multivariata che può essere risolto sulla base del teorema di Gauss-Markov.

Questa soluzione è stata proposta e discussa, in svariate forme, da Patterson (1950), Eckler (1955), Gurney e Daly (1965), Wolter (1979). In particolare, i primi contributi hanno preso in esame la particolare struttura

di correlazione esponenziale, in cui cioè $\rho_j = \rho_1^j$, e hanno considerato variabili quantitative misurate a livello individuale. Successivamente sono state sviluppate estensioni per medie e totali, quindi per proporzioni e conteggi, e infine è stato abbandonato il vincolo della struttura esponenziale. Nelle diverse proposte, la scrittura dello stimatore varia, a seconda dell'interpretazione più naturale allo svolgimento della soluzione; in qualche caso la struttura fondamentale di media ponderata rimane in ombra. Le peculiarità del disegno di rotazione determinano la struttura delle matrici di varianze e covarianze che entrano nella soluzione.

Gli stimatori lineari a varianza minima non vengono di solito impiegati nelle grandi indagini correnti, perchè la dimensione della soluzione aumenta all'aumentare dei periodi considerati. In un caso come quello italiano, tali stimatori potrebbero però assumere un interesse particolare nell'ambito dei cosiddetti comuni di tipo B, che ruotano a loro volta nel campione, rimanendo nell'indagine per 12 rilevazioni, dando luogo a 48 stime elementari di sezione. Naturalmente, un impiego di *routine* di questa soluzione esatta acquisterebbe rilevanza con l'aumento dell'incidenza dei comuni di tipo B nella RTFL.

Non è tuttavia immediato stabilire la relazione tra gli stimatori lineari a varianza minima e gli stimatori composti come la (6), che costituiscono una soluzione sub-ottimale. Un tentativo è stato fatto da Wolter (1979) riguardo ad una indagine specifica, mentre nel caso italiano l'esperienza non è ancora stata effettuata.

Seguendo l'impostazione della ricerca dello stimatore lineare a varianza minima, non si ammette però alcuna relazione tra il totale al tempo t e quello al tempo t-k. Secondo la tradizione della teoria del campionamento da popolazioni finite, infatti, il totale di popolazione non è aleatorio. I problemi che sorgono nell'affrontare il problema in questo modo sono stati messi in luce da Smith (1978) e ulteriormente sviluppati da Tam (1987).

I contributi teorici più recenti, piuttosto che con la proposta di stimatori composti sub-ottimali, tentano di risolvere i problemi sollevati dalla ricerca del miglior stimatore lineare corretto a varianza minima sfruttando l'aspetto di serie temporali delle indagini ripetute nel tempo (vedi Blight e Scott, 1973; Jones, 1980; Scott e Smith, 1974; Scott, Smith e Jones, 1977; Smith, 1978 e Tam, 1987).

4. *La valutazione della struttura di correlazione e la ricerca del peso ottimo*

Sia nel caso dell'adozione dello stimatore (6), sia in quello della ricerca dello stimatore lineare a varianza minima, la valutazione dei coefficienti di correlazione tra stimatori di sezione costituisce un elemento essenziale della soluzione del problema. Nel caso della ricerca dello stimatore lineare a varianza minima con il teorema di Gauss-Markov, le correlazioni tra le stime calcolate nelle diverse occasioni *entr*ano infatti nella matrice delle varianze e covarianze. D'altra parte, nel caso si adotti lo stimatore sub-ottimo (6), la valutazione della struttura di correlazione dà, per ogni variabile, la possibilità

di scegliere il peso k che massimizza l'efficienza relativa dello stimatore composto rispetto a quello tradizionale.

Uno studio svolto in modo mirato alla determinazione del peso ottimale per le singole variabili permetterà di ottenere risultati che presentano una variabilità teorica più contenuta di quanto avverrebbe con un unico peso.

Tuttavia, il calcolo del peso ottimale per ciascuna variabile non può essere l'obiettivo da perseguire nella pubblicazione corrente di stime sulle forze di lavoro da parte di un istituto nazionale di statistica. Infatti, l'utilizzazione di pesi diversi per variabili diverse nell'intento di ricercare risultati ottimali mal si combina con l'esigenza di coerenza tra i dati pubblicati. I valori delle stime debbono essere compatibili con il sistema di relazioni contabili tra le variabili, per dar luogo a totali generali partendo da somme parziali, e ciò non è garantito se si analizza ciascuna variabile in modo indipendente dalle altre. In questa situazione, la proposta di uno stimatore composto implicherà la scelta di un unico peso per tutte le variabili, a scapito della ricerca della minimizzazione della varianza della stima distintamente per ciascuna variabile.

Prima di passare alla stima delle correlazioni con i dati della RTFL, sono state effettuate diverse simulazioni, sintetizzate in parte in Cocchi e Castellini (1988), per saggiare l' efficienza della (15) e della (17) al variare di k, sulla base di strutture di correlazione decrescenti. Come era da attendersi, tali simulazioni hanno mostrato, a parità di valori delle correlazioni, maggiori vantaggi nel caso di stima delle variazioni piuttosto che dei livelli delle variabili. Per lo schema di rotazione italiano si è potuto notare come i valori del peso k che danno luogo al maggiore aumento di precisione dello stimatore composto rispetto a quello ordinario siano abbastanza prossimi a 0,5. Il valore 0,5 è inoltre quello che si usa anche nella rilevazione statunitense, la quale, seppur con la differenza della cadenza mensile anziché trimestrale, ha caratteristiche dello schema di rotazione e del tasso di sovrapposizione simili al caso italiano.

Data la natura esplorativa di questo studio, vale la pena di stimare la successione delle correlazioni per alcune variabili rilevanti ed ottenere i pesi ottimi, valutando poi la perdita di efficienza nei diversi casi, quando si opti per un unico peso.

Inoltre, poichè si può ragionevolmente sospettare un peggioramento della qualità dei dati all'aumentare della distanza temporale tra le coppie di elementi da abbinare, si potrà saggiare l'efficienza delle stime ottenute usando solo una parte della struttura di correlazione, quella più affidabile perchè relativa alle distanze temporali più brevi. Assumerà un particolare interesse la valutazione della correlazione ad un periodo di distanza, che si pone come componente della variabilità di qualunque stimatore composto.

La scelta del metodo di valutazione delle correlazioni si collega alla distinzione tra dati individuali e dati aggregati, secondo quanto è stato messo in evidenza da Smith (1978). La letteratura meno recente sui campioni ruotati, ed in particolare le chiarissime esposizioni che sono nei manuali più noti, ad esempio Cochran (1977) o Yates (1981), fanno riferimento a variabili quantitative individuali. Invece, il risultato fondamentale che si ottiene nella

pratica dei sondaggi è basato su stime aggregate di periodo, o al massimo sulle replicazioni delle stime di periodo calcolate nelle sezioni ruotate.

Si deve notare che l'uso del coefficiente di correlazione tra stime di sezione, anche se formalmente corretto, specie per il caso italiano si presta a qualche critica: in primo luogo perché il calcolo dei coefficienti di correlazione tra stime di sezione, se non si dispone di una successione di rilevazioni sufficientemente lunga, rischia di basarsi su un numero di casi esiguo (in particolare quando si dispone di indagini trimestrali); secondariamente perchè, con le (14), si forza un andamento costante su quantità che possono essere variabili nel tempo. Lo stimatore composto, a prima vista svincolato da stringenti assunzioni distributive, viene in realtà ad essere fortemente condizionato dalle ipotesi avanzate, nonostante dal punto di vista formale manchi un modello sulle variabili. In Cocchi (1990) si tenta di saggiare la validità dell'assunto di invarianza nel tempo.

Gli stimatori della RTFL sono del tipo (1) o (5), e quindi non lineari. Per la valutazione della loro varianza, si usa sovente un' approssimazione in serie di Taylor, nella versione proposta da Woodruff (1971), come è fatto anche nel cap. 2, distinguendo tra strati AR e NAR. In questo modo, le varianze dello stimatore vengono valutate in modo indipendente per ogni periodo. La determinazione della struttura di correlazione tra stimatori di sezione nasconde invece un vincolo più restrittivo: sfruttando le ipotesi espresse con la (13) e la (14), ci si propone di valutare misure di variabilità tenendo conto dei dati ottenuti nel corso del tempo. Così, il numero delle replicazioni della stima (1) che possono entrare nei calcoli può essere denotato come: $J=J_1J_2J_3$, in cui:

J_1: numero di replicazioni, al tempo t, di stime di sezione abbinabili a quelle di un'altra rilevazione, nel caso italiano una o due. Infatti, nell'occasione t-esima, come si è detto, all'interno dello strato le stime di sezione da considerare sono due nel caso delle distanze temporali trimestrali o annuali, ed una nel caso delle distanze di nove e quindici mesi;

J_2: numero di replicazioni nel tempo utilizzabili, che dipende dal numero di rilevazioni disponibili. Per la valutazione delle correlazioni si considera infatti una successione di indagini sulla base delle restrittive ipotesi di stazionarietà (13) e (14). Per questo motivo, disponendo dei dati di più indagini, all'interno di uno stesso strato si possono considerare le replicazioni nel tempo delle stime di sezione calcolate sulle famiglie che rispondono alle diverse occasioni di intervista;

J_3: eventuali replicazioni di comuni all'interno dello strato, che nel piano di campionamento della RTFL non sono disponibili quando non vi sia sovracampionamento, e che nel caso di una sola stima per strato possono essere ottenute abbinando strati che possono essere considerati simili.

L'approssimazione in serie di Taylor secondo Woodruff (1971) permette anche di calcolare con facilità un'approssimazione della varianza di una differenza tra periodi, perchè scinde la varianza da calcolare nella somma di varianze di elementi indipendenti. Nel caso della differenza tra il tempo t ed il tempo t-i, ad esempio, l'elemento alla base dei calcoli, per sesso e per strato, è:

$$ {}_a t(y)_{h,t,j} - \frac{{}_a t(y)_{h,t}}{{}_a t(z)_{h,t}} {}_a t(z)_{h,t,j} - \left({}_a t(y)_{h,t-i,j} - \frac{{}_a t(y)_{h,t-i}}{{}_a t(z)_{h,t-i}} {}_a t(z)_{h,t-i,j} \right), \qquad (24) $$

dove con j=1, ..., J si indicano le replicazioni che possono essere considerate. A seconda che gli strati siano AR o NAR, si useranno le espressioni adeguate per le stime delle varianze (vedi il cap. 2). A partire dalle varianze in t, t-i e dalla corrispondente varianza della differenza, si potrà ricavare una valutazione delle correlazioni alla distanza temporale i come

$$ r_i = \frac{V[t(y)_t] + V[t(y)_{t-i}] - V[t(y)_t - t(y)_{t-i}]}{2(V[t(y)_t])^{1/2}(V[t(y)_{t-i}])^{1/2}}. \qquad (25) $$

Questo procedimento approssimato, simile a quello proposto da Lee (1985), considera le replicazioni all'interno dello strato nel modo appena descritto, tenendo conto della possibilità di replicazioni nel tempo.

Il calcolo delle varianze necessarie per la (25) viene quindi effettuato su insiemi diversi di dati a seconda dell'ordine del ritardo considerato. Ne consegue che, a meno di avere successioni di rilevazioni molto lunghe, si rischia di ottenere stime diverse delle varianze di livello per gli stessi periodi.

5. *Un esempio su una particolare regione*

5.1. *Descrizione dei dati utilizzati e delle fasi di calcolo*

Per valutare empiricamente la proposta di stimatori composti sono stati analizzati i dati individuali della regione Veneto per una successione di 6 indagini, dal 1985.I al 1986.II (il periodo di 15 mesi consente di tentare anche la valutazione della correlazione a 5 periodi di distanza). Su tale base è stato impostato un procedimento di trattamento dei dati che, a differenza di quanto accade attualmente, viene effettuato subordinatamente alle modalità della variabile 'sezione', vale a dire occasione di intervista.

L'eventualità dell'adozione di stimatori composti da parte dell'Istat imporrebbe, in realtà, la ricerca di pesi ottimi a livello nazionale. Riteniamo tuttavia che l'ambito regionale si mostri adeguato (vedi anche Falorsi, 1990).

Come è stato menzionato nell'introduzione, la necessità di trattare i dati individuali per il calcolo delle (1) ha sollevato il problema della scelta tra la riutilizzazione di procedure di gestione dei dati correntemente usate dall'Istat e l'organizzazione di una base di dati sulle forze di lavoro finalizzata alla soluzione dei problemi da affrontare.

Si è scelta questa seconda strada, che permette tra l'altro di verificare passo per passo la qualità delle elaborazioni fatte e di integrare informazioni che a lungo sono appartenute ad archivi diversi e tra loro non collegati. La mancanza del codice comunale Istat nel *file* di dati sulle forze di lavoro non consentiva d'altra parte il legame esplicito con le informazioni ausiliarie

necessarie al calcolo della (1), come la popolazione dello strato.

Ancora, il fatto che nel *file* corrente dei dati dalla RTFL si disponga del solo codice di strato dei comuni campionati è causa di ambiguità quando strati diversi, all'interno della stessa provincia, sono identificati da uno stesso codice, come avviene spesso per gli strati AR. Il problema può essere risolto solo risalendo alla prima cifra del codice progressivo di comune (1 per il capoluogo e 3 per i comuni oltre i 20.000 abitanti). Inoltre, soltanto la conoscenza dell'universo degli strati Istat permette di controllare la rotazione dei comuni di tipo B, e quindi di effettuare in modo appropriato gli abbinamenti delle sole famiglie che appartengono alle stesse sezioni.

Per risolvere le difficoltà appena menzionate è stato messo a punto presso il CINECA un sistema di integrazione tra le banche dati Istat-CINECA e i dati dalla RTFL, che permette di compiere tutte le verifiche sulla sostituzione dei comuni e sul tasso di campionamento. In particolare è possibile associare in modo esplicito tutte le componenti degli stimatori di tipo (1) o (5) sulla base dei valori della (2) senza servirsi di coefficienti di riporto equivalenti, nella (1), al rapporto (2)/(3). Ciò avviene utilizzando congiuntamente FOCUS, per la realizzazione e la gestione della base di dati, e SAS per le elaborazioni successive (vedi Cocchi, 1990). La costruzione del sistema di corrispondenze descritto permette anche, qualora lo si desideri, di introdurre con relativa facilità metodi per correggere le anomalie nella post-stratificazione documentate nel cap. 3, consentendo altresì di considerare eventuali miglioramenti anche nell'ottica della stima per piccole aree.

Il procedimento messo a punto può essere descritto nel modo seguente. In primo luogo, per mezzo del sistema di gestione di basi di dati FOCUS, il *file* di dati della RTFL, per ogni rilevazione, viene associato ad una struttura di decodifica, dando luogo a due *files* di dati i quali corrispondono rispettivamente al conteggio degli individui che possiedono le caratteristiche di interesse e al totale degli individui campionati. Ciò avviene rispettando il massimo di disaggregazione, mantenendo la possibilità di marginalizzare per sesso, sezione di rotazione, comune, ecc., in base ad una gerarchia comunque scelta e considerando le successive occasioni temporali come modalità di una variabile 'tempo'. Viene contemporaneamente creato un *file* di dati che contiene la descrizione degli strati definiti nella regione e che permette di riconoscere i comuni appartenenti allo stesso strato, gli strati AR, gli strati NAR con un solo comune campionato, gli strati NAR con più comuni campionati, il numero di famiglie e la popolazione per sesso al 31 dicembre dei diversi anni.

In questo modo si possono costruire la (2) e la (4) e, per ciascuna variabile, procedere al conteggio della (3).

Da questo punto in poi si utilizzano programmi SAS elaborati per:

(a) costruire le stime di sezione (1) all'interno dello strato; nel caso in cui ci siano più comuni all'interno dello strato, la (1) non è altro che la media degli stimatori di strato determinati su base comunale, ponderati con l'incidenza della popolazione dei comuni nello strato;

(b) ricalcolare le stime per strato correntemente ricavate dall'Istat (vedi il cap. 2);

(c) verificare la sostanziale coincidenza tra tali stime e la media aritmetica (5) delle stime di sezione;
(d) selezionare gli elementi che per le distanze temporali prefissate verranno utilizzati per la valutazione delle correlazioni;
(e) stimare la struttura di correlazione;
(f) determinare il k ottimale per la stima di livelli e di variazioni, sia tenendo conto di tutte le correlazioni stimate, sia eliminando una ad una quelle relative alle maggiori distanze temporali, valutando di volta in volta la perdita in efficienza;
(g) calcolare lo stimatore composto, sia per i livelli sia per le variazioni, per i rispettivi k ottimi e per un peso eguale per tutte le variabili, ad esempio k =0,5.

5.2. *Sulla stima della struttura di correlazione*

La stima della struttura di correlazione è uno dei passi fondamentali del procedimento appena riassunto.

Inizialmente, la procedura di calcolo delle correlazioni per un ambito regionale è stata effettuata senza tenere conto del piano di campionamento, considerando perciò gli stimatori di sezione (1) come realizzazioni della stessa variabile casuale. Ogni risultato di strato (più precisamente, ogni risultato di sezione e di comune all'interno dello strato, per tenere presente la possibilità di 'sovracampionamento') è stato quindi considerato come proveniente da un campionamento casuale semplice.

Questa semplificazione ha permesso di predisporre abbastanza velocemente la successione delle fasi di calcolo, ed anche di impostare una verifica, seppure rudimentale, della stazionarietà nel tempo che viene ipotizzata nella definizione della soluzione qui adottata (vedi Cocchi, 1990). Infatti, considerando come replicazioni le stime comunali, si sono potuti distinguere fin dall'inizio, per la distanza di un anno, gli abbinamenti tra stime di sezione riguardanti intervistati alla prima e terza occasione dagli abbinamenti tra stime che coinvolgevano intervistati per la seconda e quarta volta. Per la distanza di un trimestre si è invece distinto tra le stime riguardanti individui intervistati per la prima e seconda volta da quelli intervistati per la terza e quarta. Soltanto dopo aver verificato che non c'erano anomalie rilevanti nell'andamento di questi correlogrammi nel tempo, si è proceduto al calcolo di un solo valore, per ogni periodo, della correlazione a un trimestre e ad un anno. Dopo la verifica sui correlogrammi per le distanze di un trimestre, nove mesi e un anno, si è calcolato un unico valore per ciascuna delle quattro distanze temporali sulla base di tutti i dati disponibili. Per quanto riguarda la correlazione a 15 mesi, lavorando in questo modo si hanno semplicemente tante replicazioni quanti sono gli strati (o i comuni considerati distinti).

Poichè si è lavorato su 6 rilevazioni, il fatto di calcolare una sola stima delle correlazioni per strato significa che, a meno di sovracampionamento, in ciascuno strato per valutare ρ_1 si dispone di una coppia di vettori di

dimensione 10, per ρ_4 di una coppia di vettori di dimensione 4, per ρ_3 di una coppia di vettori di dimensione 3 e per ρ_5 di due soli valori. Infatti, l' ipotesi di stazionarietà temporale consente di considerare anche le replicazioni nel tempo come realizzazioni della stessa variabile casuale. Si è di fatto mantenuta, però, la possibilità di calcolare stime di sezione secondo la (1) per ciascun comune all'interno dello strato, per mantenere distinte le replicazioni di comuni all'interno dello strato, che si presentano regolarmente quando c'è sovracampionamento. Nel caso si abbiano più comuni per strato, o si costruiscano 'superstrati' abbinando strati che vengono considerati affini, aumenta J_3, il numero delle replicazioni delle stime di sezione per strato.

Seguendo questo procedimento, sulla base della stime di sezione disponibili, la successione delle stime delle correlazioni per alcune variabili fondamentali ha mostrato un andamento costante, e non decrescente come c'è da attendersi. Ciò, oltre che all'esiguità dei dati, potrebbe essere parzialmente attribuito alla non buona qualità dei dati stessi in senso longitudinale. Tuttavia vale anche la pena di ricordare che il periodo di riferimento di 6 trimestri (gennaio 1985-aprile 1986) è molto breve, e non permette di cogliere le dinamiche cicliche che tipicamente caratterizzano il sistema economico. Ciò costituisce un ulteriore motivo di inerzia nei valori dei coefficienti di correlazione per le diverse distanze temporali.

Conforta tuttavia ricordare la sostanziale coincidenza dei valori di $\hat{\rho}_1$, la correlazione più affidabile, con i risultati ottenuti calcolando stime a partire dai valori dicotomici individuali all'interno degli strati effettuate recentemente dall'Istat. Inoltre, la maniera con cui sono state effettuate le prove permetterà di passare con relativa facilità a metodi alternativi di stima delle varianze, e conseguentemente delle correlazioni.

5.3. *I risultati*

Nella Tab. 2 riportiamo il valore di $\hat{\rho}_1$, per il totale e separatamente per i due sessi, per alcune variabili. Non siamo al corrente di studi per l'argomento effettuati per sesso, ma i risultati per il totale sono paragonabili a quelli ottenuti in altri Paesi, con correlazioni alte per le variabili legate all'occupazione e più basse per le variabili di disoccupazione, mentre gli andamenti sono relativamente difformi per sesso.

Si può notare infatti che, di massima, le correlazioni assumono valori più bassi in corrispondenza delle variabili che danno maggiori difficoltà di valutazione e che comunque sono caratterizzate da un maggior numero di entrate e/o uscite. Tale *pattern* è analogo a quello riscontrato, ad esempio, in Canada (vedi Kumar e Lee, 1985). Gli andamenti sono peraltro opposti quando si analizzano separatamente i dati dei due sessi.

Nelle Tabb. 3 e 4 sono riportati i valori dell'efficienza relativa rispettivamente dello stimatore di livello (6) e dello stimatore per le differenze (16), rispetto agli stimatori tradizionali. In ciascuna colonna compaiono:

Tab. 2: *Stime dei* ρ_1 *ottenute considerando gli stimatori di sezione come provenienti da un campionamento casuale semplice*

	forze di lavoro	occupati	occ. in agricoltura	in cerca di lavoro	disoccupati
Totale	0,962	0,959	0,655	0,660	0,390
Maschi	0,963	0,959	0,632	0,408	0,117
Femmine	0,839	0,828	0,476	0,508	0,429

(a) i valori dell'efficienza relativa, in particolare i rapporti (18)/(15) nella Tab. 3 e (19)/(17) nella Tab. 4, calcolati per ogni variabile sulla base del k ottimale e dei valori $\hat{\rho}_j$ costanti ed uguali a quelli della Tab. 2,
(b) tra parentesi tonde il k ottimale,
(c) tra parentesi quadre la differenza tra l'efficienza valutata con il k ottimo e quella valutata con k=0,5, rispettivamente nel caso di: (I) considerazione di tutte e 5 le correlazioni, (II) eliminazione di $\hat{\rho}_5$, (III) considerazione dei soli $\hat{\rho}_4$ e $\hat{\rho}_1$ (a parità di altre condizioni, si tratta di stime valutate su un numero doppio di dati rispetto a $\hat{\rho}_3$ e $\hat{\rho}_5$) (IV) eliminazione di $\hat{\rho}_5$ e $\hat{\rho}_4$, (V) mantenimento del solo $\hat{\rho}_1$.

Le due tabelle mostrano quindi i risultati ottenuti eliminando successivamente le correlazioni meno affidabili. Scorrendo le colonne, l'abbandono progressivo delle correlazioni corrispondenti alle distanze più lontane equivale a postulare l'assenza delle correlazioni eliminate.

Di solito, negli studi di questo tipo per molte delle caratteristiche considerate si ottengono vantaggi del 18-21% nell'efficienza relativa per gli stimatori del tipo (6). I miglioramenti di entità superiore ottenuti dipendono dalla sostanziale invarianza delle correlazioni nel tempo.

Tab. 3: *Efficienza della stima (6) rispetto alla stima tradizionale e k ottimo corrispondente*

	forze di lavoro	occupati	occ. agr.	disoccupati
I	1,497 (0,50) [0]	1,496 (0,50) [0]	1,140 (0,30) [*]	1,061 (0,20) [*]
II	1,729 (0,60) [0,05]	1,724 (0,60) [0,05]	1,152 (0,30) [0,12]	1,063 (0,20) [*]
III	1,438 (0,45) [0,03]	1,438 (0,45) [0,04]	1,138 (0,30) [*]	1,061 (0,20) [*]
IV	1,476 (0,45) [0,02]	1,476 (0,45) [0,02]	1,139 (0,30) [*]	1,061 (0,20) [*]
V	2,116 (0,70) [0,35]	2,108 (0,70) [0,34]	1,155 (0,30) [0,10]	1,063 (0,20) [*]

* indica che, impiegando k=0,5, si ha un peggioramento dell'efficienza

Tab. 4: *Efficienza delle stime (16) rispetto alla stima tradizionale e k ottimo corrispondente*

	forze di lavoro	occupati	occ. agr.	disoccupati
I	2,217 (0,50) [0]	2,214 (0,50) [0]	1,330 (0,40) [0,03]	1,138 (0,30) [0,06]
II	2,632 (0,55) [0,01]	2,623 (0,55) [0,01]	1,367 (0,45) [0,01]	1,148 (0,35) [0,05]
III	2,103 (0,45) [0,09]	2,100 (0,45) [0,09]	1,315 (0,40) [0,07]	1,137 (0,30) [0,09]
IV	2,232 (0,50) [0]	1,475 (0,45) [0,02]	1,139 (0,30) [0,17]	1,138 (0,30) [0,01]
V	6,246 (0,90) [3,25]	6,187 (0,90) [3,19]	1,414 (0,55) [0]	1,151 (0,35) [0,02]

* indica che, impiegando k=0,5, si ha un peggioramento dell'efficienza.

Possiamo anche notare la conferma dei migliori risultati per le variazioni piuttosto che per i livelli. L'elevato aumento in efficienza constatato quando si mantiene solo ρ_1 è attribuibile al valore molto elevato di questo coefficiente, che permette di ridurre sensibilmente la varianza della stima.

A parità di struttura di correlazione otteniamo pesi più elevati per lo stimatore composto delle differenze, per le quali abbiamo in genere risultati migliori e maggiore attesa di ottenere miglioramenti apprezzabili sulla base del peso comune k=0,5. Ciò non sembra avvenire nel caso dei livelli, per il quale, almeno in base alla struttura di correlazione valutata, si dovrebbero cercare i pesi ottimi per ciascuna variabile invece di ricorrere ad un peso comune, oppure si dovrebbe adottare un peso minore di 0,5, per ridurre il numero di differenze negative. Un problema consiste senz'altro nella scarsa rilevanza dei miglioramenti dell' impiego di uno stimatore composto per la delicata variabile disoccupati. Infine, qualunque sia il numero di correlazioni ritenute, la perdita di efficienza nella scelta di un unico peso non appare rilevante, se comparata all'ordine di grandezza dell' efficienza raggiunta.

Questo studio, seppure limitato, ha permesso di mettere in luce alcune caratteristiche, tutto sommato positive, dell'introduzione della rotazione nel calcolo delle stime. Vale la pena di sottolineare tuttavia che, affinchè queste caratteristiche risultino apprezzabili, è necessario che i periodi di riferimento siano sufficientemente lunghi, in modo da permettere che il meccanismo della rotazione si eserciti in modo completo ed il suo effetto possa essere misurato più compiutamente.

Appendice

Derivazione della (15). Si parte dalla (12):

$$V[t_c(y)_t] = \frac{1}{1-k^2}\left\{V[t^*(y)_t] + 2\sum_{j=1}^{t} k^j\, C[t^*(y)_t, t^*(y)_{t-j}]\right\}. \qquad (A.1)$$

Per il primo addendo,

$$V[t^*(y)_t] = V\left\{k[t_s(y)_{t,(t-1)} - t_s(y)_{t-1,(t)}] + (1-k)t(y)_t\right\},$$

poiché

$$C[t_s(y)_{t,(t-1)}, t_s(y)_{t-1,(t)}] = \rho_1\sigma^2/2,$$
$$C[t_s(y)_{t,(t-1)}, t_s(y)_t] = \sigma^2/4,$$
$$C[t_s(y)_{t-1,(t)}, t(y)_t] = \rho_1\sigma^2/4,$$

sostituendo e sommando si ottiene:

$$V[t^*(y)_t] = (\sigma^2/4)\left\{1+3k^2-2\rho 1k\,(1+k)\right\}. \qquad (A.2)$$

Calcoliamo ora i termini della sommatoria, cioè le

$$C[t^*(y)_t, t^*(y)_{t-1}] \qquad \text{per} \qquad i = 1,2,\ldots .$$

Sia i=1. Si avrà:

$$C[t^*(y)_t, t^*(y)_{t-1}] = C\left\{k[t_s(y)_{t,(t-1)} - t_s(y)_{t-1,(t)}] + (1-k)t(y)_t,\right.$$
$$\left. k[t_s(y)_{t-1,(t-2)} - t_s(y)_{t-2,(t-1)}] + (1-k)t(y)_{t-1}\right\}$$
$$= \frac{\sigma^2}{4}\left[-k + k^2 + \frac{\rho_1}{2}(1-k^2)\right], \qquad (A.3)$$

poiché

$$C[t_s(y)_{t,(t-1)}, t(y)_{t-1}] = \rho_1\sigma^2/4,$$
$$C[t_s(y)_{t-1,(t)}, t(y)_{t-1}] = \sigma^2/4,$$
$$C[t(y)_t, t(y)_{[t-1]}] = \rho_1\sigma^2/8,)$$

e le restanti covarianze sono nulle.
Sia i=2. Si avrà:

$$C[t^*(y)_t, t^*(y)_{t-2}] = 0,$$

poiché è sempre nulla la frazione di sovrapposizione a distanza di due intervalli.
Sia i=3. Si avrà:

$$C[t^*(y)_t, t^*(y)_{t-3}] = C\{k[t_s(y)_{t,(t-1)} - t_s(y)_{t-1,(t)}] + (1-k)t(y)_t,\ k[t_s(y)_{t-3,(t-4)} - t_s(y)_{t-4,(t-3)}] + (1-k)t(y)_{t-3}\}$$
$$= \frac{\sigma^2}{4}\left[\frac{\rho_3}{4}(1-k^2) - \frac{k\rho_4}{2}(1-k)\right]. \qquad (A.4)$$

Infatti:

$C[t(y)_t, t_s(y)_{t-3,(t-4)}] = \rho_3\sigma^2/8,$
$C[t(y)_t, t_s(y)_{t-4,(t-3)}] = \rho_4\sigma^2/8,$
$C[t(y)_t, t_s(y)_{t-3}] = \rho_3\sigma^2/16$

e le restanti covarianze sono nulle.
Sia i=4. Si avrà:

$$C[t^*(y)_t, t^*(y)_{t-4}] = C\{k[t_s(y)_{t,(t-1)} - t_s(y)_{t-1,(t)}] + (1-k)t(y)_t,\ k[t_s(y)_{t-4,(t-5)} - t_s(y)_{t-5,(t-4)}] + (1-k)t(y)_{t-4}\}$$
$$= \frac{\sigma^2}{4}\left\{-\frac{k\rho_3}{2}(k+1) + \frac{\rho_4}{2}(3k^2+1) - \frac{k\rho_5}{2}(k+1)\right\}, \qquad (A.5)$$

essendo:

$C[t_s(y)_{t,(t-1)}, t_s(y)_{t-4,(t-5)}] = \rho_4\sigma^2/4,$
$C[t_s(y)_{t,(t-1)}, t_s(y)_{t-5,(t-4)}] = \rho_5\sigma^2/4,$
$C[t_s(y)_{t,(t-1)}, t(y)_{t-4}] = \rho_4\sigma^2/8,$
$C[t_s(y)_{t-1,(t)}, t_s(y)_{t-4,(t-5)}] = \rho_3\sigma^2/4,$
$C[t_s(y)_{t-1,(t)}, t_s(y)_{t-5,(t-4)}] = \rho_4\sigma^2/4,$
$C[t_s(y)_{t-1,(t)}, t(y)_{t-4}] = \rho_3\sigma^2/8,$
$C[t(y)_t, t_s(y)_{t-4,(t-5)}] = \rho_4\sigma^2/8,$
$C[t(y)_t, t_s(y)_{t-5,(t-4)}] = \rho_5\sigma^2/8,$
$C[t(y)_t, t(y)_{t-4}] = \rho_4\sigma^2/8.$

Sia i=5. Si avrà:

$$C[t^*(y)_t, t^*(y)_{t-5}] = C\{k[t_s(y)_{t,(t-1)} - t_s(y)_{t-1,(t)}] + (1-k)t(y)_t,\ k[t_s(y)_{t-5,(t-6)} - t_s(y)_{t-6,(t-5)}] + (1-k)t(y)_{t-5}\}$$
$$= \frac{\sigma^2}{4}\left\{\frac{k\rho_4}{2}(k-1) + \frac{\rho_5}{4}(1-k)^2\right\}, \qquad (A.6)$$

dato che:

$C[t_s(y)_{t,(t-1)}, t(y)_{t-5}] = \rho_5\sigma^2/8,$
$C[t_s(y)_{t-1,(t)}, t(y)_{t-5}] = \rho_4\sigma^2/8,$
$C[t(y)_t, t(y)_{t-5}] = \rho_5\sigma^2/16$

e le restanti covarianze sono nulle.
Inoltre,

$C[t^*(y)_t, t^*(y)_{t-i}] = 0$, per ogni $i > 5$.

Sostituendo le (A.2)-(A.6) nella (A.1) con gli opportuni pesi, si ottiene finalmente la (15).

Calcolo della (17). Si parte dalla:

$$V[d_c(y)_t] = V[t_c(y)_t - t_c(y)_{t-1}] = 2\{V[t_c(y)_t] - C[t_c(y)_t, t_c(y)_{t-1}]\}$$
$$= \frac{2}{1-k^2}\{(k-1)C[t^*(y)_t, t_c(y)_{t-1}] + (1-k)C[t^*(y)_t, t_c(y)_t]\}$$
$$= \frac{2}{1-k^2}\{(1-k)V[t^*(y)_t] - (1-k)^2 C[t^*(y)_t, t_c(y)_{t-1}]\} .$$

Poichè:

$$C[t^*(y)_t, t_c(y)_{t-1}] = C[t^*(y)_t, kt_c(y)_{t-2} + t^*(y)_{t-1}]$$
$$= kC[t^*(y)_t, t_c(y)_{t-2}] + C[t^*(y)_t, t\,(y)_{t-1}]$$
$$= \sum_{i=1}^{t-1} k^{i-1} C[t^*(y)_t, t^*(y)_{t-i}],$$

sarà anche:

$$V[d_c(y)_t] = \frac{2}{1-k^2}(1-k)\left\{V[t^*(y)_t] - \frac{(1-k)^2}{k}\sum_{i=1}^{t-1} k^i\, C[t^*(y)_t, t^*(y)_{t-i}]\right\} . \qquad \text{(A.7)}$$

Sommando opportunamente le (A.2)-(A.6) si ottiene la (17).

PARTE TERZA:

ABBINAMENTO LONGITUDINALE E QUALITA' DEI DATI

7

PROCEDURE PER L'ABBINAMENTO DEI DATI INDIVIDUALI DELLE FORZE DI LAVORO

Antonio Giusti, Gianni Marliani e Nicola Torelli

1. *Introduzione*

La rilevazione trimestrale delle forze di lavoro (RTFL), condotta dall'Istat, è stata progettata essenzialmente come un'indagine *cross-section*, con l'obiettivo di ottenere, a cadenza trimestrale, una misura e una descrizione delle caratteristiche dell'occupazione e della disoccupazione. Tuttavia, il particolare disegno campionario adottato, che prevede una rotazione parziale degli intervistati in quattro successive occasioni di indagine, consente di ottenere dati di tipo longitudinale, sfruttando le informazioni fornite dagli individui che permangono nel campione per più rilevazioni.

Finora, l'uso di questi dati è stato limitato alla costruzione di matrici di transizione tra stati occupazionali, ad intervalli trimestrali ed annuali, che l'Istat pubblica regolarmente dal 1979 (vedi Moriani, 1981) e sulle quali è possibile condurre alcune parziali analisi di mobilità (Pollastri, 1976a e 1976b; Zaccarin e Bernardi, 1984; vedi anche il cap. 18).

In realtà, la possibilità di creare dati longitudinali, seppure riferiti ad una durata relativamente modesta (massimo 15 mesi), apre la strada a tutta una serie di utilizzazioni che, da un lato, coinvolgono interessanti analisi sulla dinamica di breve periodo del mercato del lavoro e, dall'altro, consentono di controllare la qualità dei dati dell'indagine stessa (U.S. Department of Labor, 1980; Duncan e Kalton, 1987).

Presupposto indispensabile per queste applicazioni è il collegamento *(linkage)* degli archivi riferiti a successive indagini trimestrali, che si ottiene abbinando opportunamente le informazioni individuali fornite nelle diverse rilevazioni.

Com'è noto, il problema del collegamento di informazioni provenienti da fonti diverse può essere affrontato attraverso tecniche di abbinamento 'esatto', se il principale obiettivo è quello di collegare le informazioni relative alla stessa unità (Jabine e Scheuren, 1986), o di abbinamento 'statistico', se si cerca di collegare le informazioni di unità simili rispetto ad un qualche criterio (Paass, 1984).

Ovviamente, le due tecniche differiscono sensibilmente per ciò che concerne le implicazioni sulla *privacy* dei rispondenti, l'affidabilità dei risultati,

le possibilità di uso dei dati che ne derivano. E la scelta di utilizzare l'una o l'altra è dettata prevalentemente dalle finalità del collegamento (Cox e Boruch, 1988).

All'abbinamento esatto si può ricorrere solo quando le basi di dati da collegare si riferiscono almeno in parte allo stesso insieme di unità e quando negli archivi sono presenti variabili, quali, ad esempio, nome e cognome, data di nascita, sesso, che consentono l'identificazione delle unità stesse.

Una applicazione di particolare interesse delle tecniche di abbinamento esatto riguarda la costruzione di informazioni longitudinali attraverso il collegamento dei dati osservati su un campione di unità, che rimane totalmente o parzialmente immutato per più rilevazioni successive (indagini *panel* o con campioni ruotati).

E' evidente che, se le variabili di identificazione sono registrate senza errore e non esiste la possibilità che più unità presentino la stessa modalità per tali variabili ('omonimie'), il problema dell'abbinamento esatto si risolve nella definizione di un algoritmo informatico efficiente, che permetta di ricercare i *records* corrispondenti negli archivi posti a confronto. Nella maggior parte dei casi, tuttavia, non è così. Le variabili di identificazione possono infatti non corrispondere a causa di errori di registrazione nei dati o perché gli archivi hanno un diverso riferimento temporale. E, d'altra parte, la corrispondenza in alcune variabili non garantisce che le unità siano necessariamente le stesse, potendosi avere, ad esempio, individui diversi che hanno lo stesso nome, la stessa data di nascita, lo stesso sesso. In questa situazione, confrontando un *record* di un archivio con quello di un altro, non è sempre possibile stabilire con certezza se essi si riferiscono o meno alla stessa unità ed è necessario introdurre un criterio decisionale che permetta di risolvere le situazioni ambigue. La formalizzazione teorica del problema risale agli anni '60 e la si ritrova in Newcombe *et al.* (1959), Tepping (1968), Fellegi e Sunter (1969). Questi ultimi, in particolare, propongono una elegante soluzione, nota come 'teoria dell'abbinamento esatto', basata su valutazioni di carattere probabilistico all'interno di un approccio di tipo decisionale.

Nel presente capitolo viene presentata una procedura di abbinamento per i dati individuali della RTFL basata su valutazioni probabilistiche. Tale procedura viene sperimentata sui dati individuali del Veneto e della Lombardia.

Il capitolo è organizzato come segue. La sez. 2 riassume brevemente gli aspetti teorici e i problemi operativi legati ad una procedura di abbinamento esatto di informazioni provenienti da fonti diverse. Nella sez. 3 si esaminano le caratteristiche della RTFL, in relazione alle possibilità che essa offre di creare informazioni longitudinali ed ai problemi specifici che pone, si fa un breve richiamo alle precedenti esperienze di abbinamento dei dati della rtfl e si presenta, infine, una procedura di abbinamento basata su criteri probabilistici. I risultati della procedura sono brevemente illustrati nella sez. 4 e sottoposti ad alcune verifiche empiriche nella sez. 5. La sez. 6, infine, contiene alcune considerazioni conclusive.

2. *L'abbinamento esatto di informazioni provenienti da fonti diverse: metodi e problemi*

Al fine di delineare i termini del problema dell'abbinamento esatto, si considerino, nel caso più semplice, due archivi di informazioni, A e B, contenenti dati che, per una certa quota, si riferiscono al medesimo insieme di unità. In ogni *record* individuale siano presenti variabili di identificazione, per le quali non sia possibile escludere la presenza di errori e omonimie.

La regola attraverso la quale decidere se abbinare o meno due *records* individuali può essere così schematizzata:

(a) si individua un insieme di n variabili di confronto, che, in genere, sono, *a priori*, o non soggette a variabilità individuale (ad esempio, la data di nascita o il sesso) o estremamente poco variabili (ad esempio, l'indirizzo);

(b) si considerano tutte le possibili coppie di *records* (a,b) ($a \in A$ e $b \in B$), dette coppie di confronto, e si associa ad ognuna di esse un vettore $\mathbf{x}$, la cui generica componente x_i esprime il risultato del confronto per l'i-esima variabile. La codifica del confronto può essere articolata nel modo che si ritiene più conveniente. Nel caso più semplice, x_i può essere ricondotta ad una variabile dicotomica che assume il valore 0 se c'è accordo e 1 se c'è disaccordo, in base ad una opportuna definizione di accordo/disaccordo;

(c) si associa ad $\mathbf{x}$ un valore sintetico $T(\mathbf{x})$, usualmente denominato 'peso totale', che è in relazione diretta con la propensione a ritenere che quel confronto riguardi due *records* relativi alla stessa unità;

(d) si sceglie un valore soglia k che discrimina tra la decisione di abbinamento, se $T(\mathbf{x}) \geq k$, e di non abbinamento, se $T(\mathbf{x}) < k$[1].

Naturalmente, considerata la possibilità di errori e omonimie nelle variabili di confronto, una regola siffatta può condurre a decisioni non corrette, ovvero si può pervenire: (i) ad un errato abbinamento ('falso positivo'), quando si decide di abbinare ed i *records* si riferiscono ad unità differenti; (ii) ad un mancato abbinamento ('falso negativo'), quando si decide di non abbinare ed i *records* sono relativi alla stessa unità. Il rapporto tra i due tipi di errore dipende - oltre che dalla distribuzione degli errori e delle omonimie nei due archivi - dal numero e dal tipo delle variabili di confronto che si scelgono, dalla rigidità con cui si definisce l'accordo o disaccordo nel confronto e dal valore soglia che si adotta.

Il punto cruciale della regola di decisione sopra indicata è la definizione del peso totale, $T(\mathbf{x})$, da associare al risultato di ciascun confronto. Tale valore può essere determinato sulla base di criteri *ad hoc*, oppure può essere desunto da opportune valutazioni probabilistiche. Questa seconda strada, aperta dal lavoro di Fellegi e Sunter (1969), appare decisamente più rigorosa.

---

1 In taluni casi, può essere opportuno adottare un insieme di decisioni più ampio e, di conseguenza, scegliere un conveniente insieme di valori soglia. Nell'impostazione di Fellegi e Sunter (1969), ad esempio, le decisioni finali sono tre. Si considera, infatti, anche la possibilità che alcuni confronti, denominati 'abbinamenti possibili', conducano ad una non decisione, rinviando eventualmente ad una successiva ispezione manuale. Tepping (1968) fornisce, invece, un esempio con cinque decisioni terminali.

Infatti, nonostante essa presenti difficoltà operative non indifferenti, consente di pervenire a procedure di abbinamento che, in relazione ad un prefissato obiettivo, hanno criteri di ottimalità.

Si richiamano qui gli elementi essenziali della 'teoria dell'abbinamento esatto', rinviando per maggiori dettagli al lavoro originale di Fellegi e Sunter, alla utile rilettura in termini di teoria dell'informazione fatta da Kirkendall (1985) e all'ottima rassegna di Jabine e Scheuren (1986).

Il vettore di confronto $\mathbf{x}$ è considerato come la determinazione di una variabile casuale (discreta), che ha funzione di probabilità $p(\mathbf{x}|H_0)$ e $p(\mathbf{x}|H_1)$ condizionatamente alle due ipotesi H_0: "i due *records* si riferiscono alla stessa unità" e H_1: "i due *records* sono relativi ad unità differenti".

Il rapporto $T(\mathbf{x}) = P(\mathbf{x}|H_0)/P(\mathbf{x}|H_1)$ è una statistica sufficiente per discriminare tra H_0 e H_1. In generale, $T(\mathbf{x})$ assume valori crescenti tanto maggiore è la probabilità che i due *records* si riferiscano alla stessa unità. Si perviene così ad una regola di decisione, che, in corrispondenza di valori elevati di $T(\mathbf{x})$, o di una sua conveniente trasformazione[2], porta all'abbinamento mentre, per valori bassi, conduce a non abbinare.

Formalmente, la situazione descritta presenta strette analogie con il problema della verifica di ipotesi statistiche semplici. Ciò rende comprensibile come la scelta del valore soglia debba essere affidata ad un criterio di ottimalità in relazione ad un dato obiettivo, quale potrebbe essere, ad esempio, rendere minima la probabilità di mancato abbinamento, una volta fissata la probabilità di errato abbinamento[3].

Seppure possa essere derivata teoricamente, la definizione di una regola ottima si scontra con problemi operativi non indifferenti. Vale la pena di richiamare alcuni tra gli ostacoli più rilevanti che si pongono nelle applicazioni pratiche.

Risulta spesso economicamente inopportuno, se non addirittura improponibile, il confronto di tutte le possibili coppie di *records*. In questa situazione, la scelta che si rivela spesso più ragionevole è quella di ricorrere al 'bloccaggio'. Tale pratica consiste nel restringere il confronto alle sole coppie (a,b) che presentano lo stesso valore in una o più variabili (variabili di blocco), riducendo così il carico computazionale. Non considerando i confronti tra *records* appartenenti a blocchi diversi, si tende a ridurre la probabilità di falsi positivi, mentre risulta più elevata la probabilità di mancati abbinamenti (Kelley, 1985). In linea teorica, nel calcolare i pesi, occorrerebbe tener conto del bloccaggio utilizzato ma, generalmente, questo aspetto viene trascurato nelle applicazioni pratiche.

In generale, le funzioni di probabilità $p(\mathbf{x}|H_0)$ e $p(\mathbf{x}|H_1)$ sono ignote. È

---

2 Nella pratica, è usuale ricorrere al logaritmo in base 2 di $T(\mathbf{x})$, che può essere interpretato come guadagno di informazione che si ha, nel decidere in favore di H_0, in seguito all'osservazione di $\mathbf{x}$, rispetto alle valutazioni di probabilità su H_0 e H_1 disponibili a priori (Kirkendall, 1985, p.196).

3 Nel sistema a tre decisioni di Fellegi e Sunter (vedi la nota 1) si arriva ad una regola 'ottima' che, fissata la probabilità di errore di primo e secondo tipo (falsi negativi e falsi positivi) minimizza il numero dei confronti che cadono nella zona di incertezza (abbinamenti possibili). Una impostazione leggermente differente si ha invece in Tepping (1968), che affronta il problema in un'ottica più propriamente decisionale, facendo esplicito riferimento ad una funzione di costo connessa con i differenti tipi di errore.

pertanto necessario procedere ad una loro stima per la quale, a meno di non disporre di sostanziali informazioni *a priori*, sono possibili due metodi. Il primo, proposto da Fellegi e Sunter, parte dall'assunto di indipendenza tra gli elementi del vettore **x**, condizionatamente alle due ipotesi e stima i pesi dalle distribuzioni marginali dei due archivi da collegare. In particolare, data l'ipotesi di indipendenza, il rapporto $T(\mathbf{x})$ può essere scomposto come $\Pi_i P(x_i|H_0)/P(x_i|H_1)$ e ciascuna componente, che può essere interpretata come il contributo fornito al peso totale dal risultato del confronto dell'i-esima variabile, viene stimata utilizzando la distribuzione marginale della variabile nei due archivi, unitamente ad alcune informazioni, supposte note, sulla probabilità di errori di rilevazione, trascrizione e registrazione nelle variabili. Il secondo metodo, suggerito da Tepping (1968), consiste nell'identificare preliminarmente, con buona approssimazione, l'insieme dei confronti relativi agli stessi individui e calcolare quindi, sulla base di questa prima valutazione, i rapporti di probabilità. In tal caso, è possibile ottenere direttamente la stima di $T(\mathbf{x})$, senza ricorrere all'ipotesi di indipendenza, che spesso può risultare eccessivamente restrittiva.

La scelta del valore soglia costituisce un'ulteriore problema. Per una soluzione ottimale, sarebbe necessario conoscere la distribuzione del vettore **x** condizionatamente alle due ipotesi H_0 e H_1. Nella pratica, è usuale determinarlo in modo empirico, ricorrendo all'ispezione della distribuzione dei pesi totali (Howe e Lindsay, 1981), unitamente a considerazioni che riguardano la propensione al rischio di falso positivo rispetto al rischio di falso negativo (Newcombe *et al.*, 1983; Newcombe, 1988).

3. *L'abbinamento dei dati di successive indagini della Rilevazione Trimestrale delle Forze di Lavoro*

3.1. *Alcune peculiarità dei dati della RTFL*

Com'è descritto nel cap. 1, il disegno della RTFL prevede la reintervista di parte delle unità campionarie in quattro diverse occasioni di indagine nell'arco di quindici mesi, secondo uno schema in base al quale, teoricamente, il 50% del campione di famiglie è comune da un trimestre al successivo e da un trimestre allo stesso trimestre dell'anno successivo, mentre il 25% del campione può essere seguito per quattro occasioni di indagine in un arco di quindici mesi.

Ogni famiglia mantiene nelle quattro occasioni di indagine un codice, che contraddistingue ciascuno dei *records* individuali dei singoli componenti il nucleo familiare. Attraverso il codice, è quindi possibile individuare le famiglie che sono presenti in indagini successive e, all'interno di queste, procedere all'abbinamento dei *records* individuali.

Mancando nel *record* un identificatore dell'individuo all'interno della famiglia, è necessario affiancare al codice familiare, che costituisce la chiave principale dell'abbinamento, una serie di altre variabili, confrontando le quali

sia possibile decidere se due *records* sono da riferire alla stessa persona. In teoria, possono essere usate, a questo fine, tutte quelle variabili individuali per le quali: (i) non è logicamente ammissibile una variazione di stato tra le due indagini, come il sesso o l'anno di nascita; (ii) la variazione è ammissibile entro limiti di tolleranza prefissati, come l'istruzione e lo stato civile; (iii) la variazione è ammissibile, ma può ritenersi rara nell'arco di tempo considerato, come la relazione con il capofamiglia.

In questa situazione, il corretto abbinamento degli individui può essere ostacolato: (i) da errori di trascrizione e registrazione del codice di identificazione della famiglia; (ii) da errori di rilevazione, trascrizione o registrazione nelle variabili di confronto; (iii) dalla possibilità che, a parità di codice familiare, esistano più individui che presentano modalità identiche nelle variabili di confronto.

Così posto, il problema non presenta particolarità di rilievo rispetto a quello generale dell'abbinamento di informazioni da due archivi, sinteticamente illustrato nella sez. 2. Tuttavia, non appena si esaminano le caratteristiche peculiari della RTFL, è possibile cogliere alcune specificità non trascurabili.

Innanzitutto, la conduzione dell'indagine ed il suo impianto non propriamente longitudinale possono generare ulteriori ostacoli al corretto abbinamento. Ci riferiamo, in particolare: (i) al fatto che alle famiglie che entrano nel campione in sostituzione di quelle non più reperibili viene attribuito lo stesso numero di codice di queste ultime; (ii) all'adozione da parte dell'Istat di un programma automatico di correzione ed imputazione basato sulla verifica di coerenza tra dati relativi ad una singola occasione (è possibile che tale programma modifichi le modalità di una variabile, generando problemi di incoerenza nel confronto tra due occasioni successive; vedi in proposito il cap. 8).

Una seconda considerazione concerne gli accorgimenti necessari per adattare la procedura di abbinamento a seconda delle occasioni di indagine che si vogliono collegare. Lo schema di rotazione, infatti, consente di abbinare gli individui a distanza di tre, dodici o quindici mesi. E' chiaro che si dovranno adottare opportune modifiche nella definizione dell'accordo e disaccordo in quelle variabili, come ad esempio il titolo di studio, che possono subire variazioni in un arco di tempo più lungo.

Un'ultima considerazione riguarda, infine, l'abbinamento di informazioni individuali per quattro occasioni di indagine. Di massima, è possibile ottenere l'abbinamento semplicemente iterando la procedura per due occasioni, con l'unica avvertenza di modificare opportunamente la definizione dell'accordo e disaccordo tra le variabili di confronto, in relazione alla distanza tra le due occasioni di volta in volta considerate. Per economia di esposizione, non ci soffermiamo in questa sede sulla strategia approntata per utilizzare convenientemente la procedura per due occasioni, presentata nella sez. 3.3, ai fini di un agevole collegamento tra quattro occasioni.

3.2. *Procedure di abbinamento dei dati della RTFL basate su criteri euristici*

Come ricordato nell'introduzione, l'Istat costruisce correntemente matrici di transizione tra stati occupazionali. A questo fine, utilizza una procedura per l'abbinamento dei dati di successive occasioni di indagine della RTFL, che viene qui richiamata in modo estremamente sintetico (per maggiori dettagli, vedi Moriani, 1981).

La procedura utilizza il codice familiare come variabile di blocco ed opera su due archivi ordinati per tale codice, riducendo così notevolmente il carico computazionale[4]. Le variabili di confronto utilizzate per l'identificazione degli individui sono: l'età in anni compiuti, il sesso, la relazione con il capofamiglia e il grado di istruzione. Il criterio seguito consiste nell'abbinare solo i *records* individuali che presentano accordo in tutte le variabili.

Reinterpretando tale procedura secondo la logica illustrata nella sez. 2, si ha che essa associa ad ogni confronto un vettore $\mathbf{x} = \{x_i\}$ (i=1,..,4), con $x_i=0$ se c'è accordo nell'i-esima variabile e $x_i=1$ se non c'è accordo. A ciascun elemento del vettore viene attribuito un peso w_{0i} pari a 0 se $x_i=0$ e w_{1i} minore di 0 se $x_i=1$. Il modo di codificare l'accordo e il disaccordo e il sistema dei pesi è riassunto schematicamente nella Tab. 1.

Il valore sulla base del quale si decide dell'abbinamento è

$$T(\mathbf{x}) = \Sigma_i (1-x_i)w_{0i} + \Sigma_i x_i w_{1i} ,$$

che, essendo $w_{0i}=0$ per ogni i, si riduce a

$$T(\mathbf{x}) = \Sigma_i i\ x_i w_{1i} .$$

La regola di abbinamento è:

- se $T(\mathbf{x}) = 0$ si abbina,
- se $T(\mathbf{x}) < 0$ non si abbina.

A prima vista, il criterio appare estremamente rigido e viene giustificato dalla preoccupazione di contenere al massimo il numero di falsi positivi. Nella costruzione di matrici di flusso, infatti, si ritiene generalmente che un errato abbinamento sia più dannoso di un mancato abbinamento, poiché il primo comporta, di solito, una sovrastima della mobilità, mentre il secondo si risolverebbe, in definitiva, solo in una riduzione della numerosità campionaria. In realtà, ciò è vero se i falsi negativi si distribuiscono in modo casuale. Quando la procedura di abbinamento è troppo rigida, però, essa rischia di

---

4 Questa scelta, anche se praticamente imposta dalla dimensione degli archivi su cui si opera, non è ovviamente senza conseguenze. Eliminando dall'insieme dei possibili confronti tutti quelli relativi ad individui che appartengono a blocchi diversi, si esclude la possibilità di abbinare lo stesso individuo che alle due occasioni presenta, per errore, un codice familiare diverso (secondo Masselli, 1988, gli errori nei codici familiari sono relativamente frequenti).

Tab. 1: *Insieme dei confronti, codifica e pesi utilizzati nella procedura di abbinamento dell'Istat*

elemento del vettore	variabile di confronto	valore assunto	condizione [(a)]	peso [(b)]
x_1	età	0	0≤età(2)-età(1)≤1	$w_{01} = 0$
		1	altrimenti	$w_{11} = -p$
x_2	sesso	0	sesso(1)=sesso(2)	$w_{02} = 0$
		1	altrimenti	$w_{12} = -p$
x_3	relazione col capofamiglia	0	relazione(1)=relazione(2)	$w_{03} = 0$
		1	altrimenti	$w_{13} = -p$
x_4	istruzione	0	istruzione(1)=istruzione(2)	$w_{04} = 0$
		1	altrimenti	$w_{14} = -p$

(a) (1) e (2) indicano che la modalità della variabile è quella osservata alla prima o alla seconda occasione.
(b) p è un qualsiasi valore positivo.

divenire selettiva, nel qual caso anche i mancati abbinamenti possono produrre una distorsione nella stima della mobilità[5].

A ben guardare, poi, anche la protezione contro il rischio dei falsi positivi non appare sufficiente. E' vero, infatti, che si abbina solo in caso di completo accordo, ma si usano poche variabili di confronto e si definisce l'accordo/disaccordo in modo piuttosto impreciso (si usa l'età in anni compiuti anziché l'anno di nascita; non si tiene conto che il grado di istruzione può modificarsi tra le due occasioni di indagine). Infine, non pare convincente la scelta di attribuire alle differenti variabili la stessa importanza ai fini della decisione.

Partendo da queste considerazioni, abbiamo inizialmente proposta una procedura alternativa, denominata MATCH, che, pur rimanendo nell'ottica di un sistema di pesi *ad hoc*, cercava di ovviare ad alcuni degli inconvenienti prima segnalati. Richiamiamo qui gli aspetti salienti e i principali risultati di tale procedura, rinviando per maggiori dettagli a Giusti, Marliani e Torelli (1987).

Così come la procedura adottata dall'Istat, MATCH utilizza il codice

5 Basta pensare, ad esempio, che, non ammettendo l'abbinamento di *records* individuali che presentano differenze nella relazione col capofamiglia, si esclude dal campione longitudinale una quota di persone per le quali il cambiamento può essere legato ad un evento particolare (morte del precedente capofamiglia, scissione del nucleo familiare, ecc.) che ha buone probabilità di essere associato anche a fenomeni di mobilità occupazionale. Alcune verifiche (vedi Giusti, Marliani e Torelli, 1987) sembrano confermare la selettività della procedura Istat: la distribuzione dei non abbinati rispetto ad alcuni caratteri strutturali (età, sesso, titolo di studio, condizione professionale) risulta abbastanza diversa da quella corrispondente degli abbinati.

familiare come fattore di bloccaggio, ma si differenzia dalla prima soprattutto per: (i) il maggior numero di variabili di confronto utilizzate; (ii) una definizione più accurata dell'accordo/disaccordo tra variabili; (iii) l'introduzione di un sistema di pesi che assegna un ruolo differenziato alle diverse variabili nel determinare la decisione; (iv) l'adozione di una regola di decisione che prevede la possibilità di abbinare anche individui per i quali si rileva disaccordo in alcune variabili e quindi, implicitamente, l'abbandono dell'ipotesi, alquanto irrealistica, dell'assenza di errori di rilevazione, trascrizione e registrazione nelle variabili di confronto.

Il modo di codificare l'accordo e il disaccordo e il sistema dei pesi della procedura MATCH è riassunto schematicamente nella Tab. 2.

Il valore sulla base del quale si decide dell'abbinamento è

$$T(\mathbf{x}) = \sum_i (1-x_i)w_{0i} + \sum_i x_i w_{1i}.$$

La regola di abbinamento è:

- se $T(\mathbf{x}) \geq 0$ si abbina,
- se $T(\mathbf{x}) < 0$ non si abbina.

Si osservi, che mentre il peso associato all'accordo è lo stesso per tutte le variabili, il disaccordo tra le differenti variabili assume un'importanza diversa nel discriminare tra la decisione di abbinamento e non. In particolare, un disaccordo nel mese di nascita o nell'anno di nascita o nel sesso conduce ad un abbinamento solo se le altre sette variabili risultano in accordo. Un disaccordo nella relazione con il capofamiglia o nel livello di istruzione o nello stato civile può invece condurre ad un abbinamento anche se accompagnato da un disaccordo nella condizione professionale o nei precedenti lavorativi.

Sul piano dei risultati, le differenze tra le due procedure possono essere riassunte nel modo seguente.

(a) La procedura MATCH consente di abbinare, tra due occasioni di indagine, un numero di individui notevolmente maggiore. Tale capacità emerge in modo ancora più netto nell'abbinamento tra quattro occasioni di indagine, per il quale si ha un incremento nella quota di abbinati prossimo al 40%.

(b) Numerose verifiche effettuate consentono di affermare che, nonostante il ragguardevole guadagno nella quota di abbinamenti, la procedura MATCH rischia ancora meno della più restrittiva procedura correntemente adottata dall'Istat sul fronte dei falsi positivi.

(c) Una eccessiva rigidità dei controlli si rivela esiziale in situazioni di errori 'anomali', che talvolta si riscontrano nei dati[6].

---

6 Nell'indagine del primo trimestre del 1985, la variabile relazione con il capofamiglia presenta, per motivi che non ci è dato conoscere, una distribuzione palesemente distorta. Ciò determina un mediocre risultato di entrambe le procedure nei collegamenti che coinvolgono il 1985.I, ma, nel caso della procedura adottata dall'Istat, la caduta nella quota di abbinati è decisamente più consistente.

Tab. 2: *Insieme dei confronti, codifica e pesi utilizzati nella procedura MATCH*

elem. del vett.	variabile di confronto	valore assunto	condizione[a]	peso
x_1	mese di nascita	0	mese(1)=mese(2)	w_{01}= 0,375
		1	altrimenti	w_{11}=-2,625
x_2	anno di nascita	0	anno(1)=anno(2)	w_{02}= 0,375
		1	altrimenti	w_{12}=-2,625
x_3	sesso	0	sesso(1)=sesso(2)	w_{03}= 0,375
		1	altrimenti	w_{13}=-2,625
x_4	relazione col capofamiglia	0	relazione(1)=relazione(2)	w_{04}= 0,375
		1	altrimenti	w_{14}=-1,750
x_5	istruzione (per abbinamenti a distanza di un trimestre)	0	istruzione(1)=istruzione(2) [b]	w_{05}= 0,375
		1	altrimenti	w_{15}=-1,750
x_5	istruzione (per abbinamenti a distanza di un anno)	0	0≤istruzione(2)-istruzione(1)≤1	w_{05}= 0,375
		1	altrimenti	w_{15}=-1,750
x_6	stato civile	1	stato(1)≠celibe e stato(2)=celibe	w_{06}=-1,750
		0	altrimenti	w_{16}= 0,375
x_7	condizione professionale	1	cond(1)=occup. e cond(2) in cerca 1ª occ.	w_{07}=-0,875
		0	altrimenti	w_{17}=0,375
x_8	precedenti lavorativi	1	non ha precedenti(2) e ha precedenti(1)	w_{08}=-0,875
		0	altrimenti	w_{18}= 0,375

(a) (1) e (2) indicano che la modalità della variabile è quella osservata alla prima o alla seconda occasione.

(b) Questo modo di codificare l'accordo è stato adottato quando la seconda delle due rilevazioni trimestrali non è quella di luglio (nel qual caso è ammissibile il passaggio da un titolo di studio a quello immediatamente superiore). In ogni caso, invece, si codifica come accordo il passaggio da diploma di scuola media superiore a laurea (che può avvenire in qualunque trimestre).

3.3. *Una procedura basata su criteri probabilistici*

L'aspetto meno convincente delle procedure illustrate nella sez. precedente è certamente la scelta del sistema di pesi, basata su valutazioni euristiche, non rigorosamente giustificabili. Ci è sembrato quindi opportuno arrivare alla definizione di una procedura nella quale il sistema dei pesi derivasse da una stima del rapporto di verosimiglianza $T(\mathbf{x}) = P(\mathbf{x}|H_0)/P(\mathbf{x}|H_1)$.

Disponendo già della procedura MATCH, che appare in grado di identificare, con margini di errore trascurabili, l'insieme dei confronti relativi agli stessi individui, la strada che è sembrata preferibile è quella, suggerita da Tepping, di utilizzare questo insieme di confronti per stimare direttamente i due termini del rapporto.

Sostanzialmente, quindi, la procedura approntata, denominata LINK, si sviluppa in due passi, il primo dei quali consiste in un abbinamento preliminare effettuato con MATCH. Dai risultati di tale abbinamento si ottiene una stima dei pesi da utilizzare nel passo successivo.

Fatta salva la differente valutazione dei pesi, lo schema generale della procedura è identico a quello di MATCH. Si usa ancora il codice familiare come fattore di bloccaggio e le variabili utilizzate per il confronto e le definizioni di accordo/disaccordo sono quelle riportate nella Tab. 2.

Al fine di illustrare come i risultati del primo passo sono utilizzati per ottenere la stima finale dei pesi, conviene partire da due considerazioni.

Consideriamo innanzitutto l'ipotesi di indipendenza tra le componenti del vettore **x**, alla quale spesso si fa ricorso per la stima del rapporto di verosimiglianza. Come sottolineato da Fellegi e Sunter (1969, pp. 1189-1190), ciò che si assume è che le componenti di **x** siano condizionalmente indipendenti con riferimento alle due ipotesi H_0 (il numeratore di $T(\mathbf{x})$) ed H_1 (il denominatore di $T(\mathbf{x})$). Le implicazioni sostanziali dell'ipotesi di indipendenza sono piuttosto differenti nei due casi. Riguardo al numeratore, conviene anzitutto osservare come $P(\mathbf{x}|H_0)$ dipenda esclusivamente dalla probabilità di errori di rilevazione, trascrizione e registrazione nelle variabili di confronto. Se le variabili fossero osservate senza errore, infatti, nel confrontare i *records* relativi ad uno stesso individuo (e cioè subordinatamente all'ipotesi H_0) si registrerebbe sempre un accordo e, quindi, si avrebbe $P(x_i=1|H_0)=0$[7]. In presenza di errori, ciò non accade e il valore di $P(x_i=1|H_0)$ dipende dalla probabilità di errore di osservazione nella variabile i-esima. In definitiva, assumere l'indipendenza tra le componenti del vettore **x**, subordinatamente ad H_0 equivale a ritenere indipendenti gli errori tra le diverse variabili di

---

7 Ovviamente, ciò presuppone che accordo e disaccordo siano stati codificati in modo che, in assenza di errori, non sia possibile osservare disaccordo quando i due *records* si riferiscono alla stessa unità. Se, invece, si codifica come disaccordo anche una differenza che per lo stesso individuo non è impossibile, ma soltanto rara (come avviene nel nostro caso per la relazione con il capofamiglia), $P(x_i=1|H_0)$ risulterà positiva, seppure piccola, anche in assenza di errori.

confronto. Nella misura in cui la generazione degli errori avviene prevalentemente nella fase di trascrizione e registrazione delle informazioni, l'assunzione può ritenersi ragionevole[8]. Riguardo al denominatore, invece, l'ipotesi di indipendenza appare decisamente più restrittiva, almeno per alcune delle variabili di confronto considerate. Prendiamo, ad esempio, il sesso e la relazione con il capofamiglia. Se si confrontano due individui diversi (quindi, condizionatamente ad H_1) e si osserva, poniamo, un accordo sul sesso, è molto probabile che questo si accompagni ad un accordo anche nella relazione con il capofamiglia, data la forte prevalenza della modalità 'capofamiglia' tra i maschi e 'coniuge' tra le femmine.

Una seconda osservazione è legata al fatto che, generalmente, nel determinare i pesi da attribuire ai vari confronti, non si tiene conto del bloccaggio. Ad ogni confronto viene assegnato un peso medio, desunto da informazioni rilevate sull'intero archivio, come se il blocco, identificato, nel nostro caso, dallo stesso codice familiare, fosse un campione casuale dell'archivio stesso. Anche in questa circostanza, le implicazioni sono diverse a seconda che si consideri la stima del numeratore e del denominatore. Per il numeratore, che come abbiamo detto dipende dalla probabilità di errore di osservazione nell'i-esima variabile, attribuire ai confronti dei differenti blocchi uno stesso peso significa ritenere che la probabilità di errore sia la stessa in ogni famiglia. Anche in questo caso, se gli errori sono prevalentemente di trascrizione e/o registrazione l'ipotesi appare, tutto sommato, sostenibile. Per il denominatore, che, sotto l'assunto di indipendenza, viene generalmente valutato sulla base delle distribuzioni di frequenza marginali delle variabili di confronto nei due archivi, un'analoga ipotesi equivale a sostenere che le variabili di confronto hanno la stessa distribuzione marginale nelle diverse famiglie. Per certe variabili l'ipotesi è palesemente inverosimile. Ad esempio, la distribuzione della relazione con il capofamiglia in famiglie di due persone è certamente diversa da quella analoga in famiglie di quattro persone. Di conseguenza, il peso da assegnare ad un accordo/disaccordo casuale andrebbe valutato, se non separatamente per ogni blocco, almeno tenendo conto di alcune appropriate caratteristiche familiari quali, ad esempio, la numerosità dei componenti.

Le due osservazioni precedenti motivano la strategia adottata nell'utilizzare i risultati di MATCH nel secondo passo della procedura LINK. L'idea base da cui si è partiti è quella di tenere in esplicita considerazione il fattore di bloccaggio nella stima del peso totale da assegnare a ciascun vettore confronto, subordinando le valutazioni di probabilità di **x** ai soli confronti fatti tra individui che hanno lo stesso codice familiare. Ciò significa che il peso totale è ancora definito come rapporto tra le verosimiglianze di **x**, dati H_0 e H_1; solo che adesso H_0 e H_1 assumono un significato leggermente diverso, indicando, rispettivamente, che gli individui confrontati sono gli stessi o sono diversi, ma hanno lo stesso codice familiare in entrambi i casi.

---

8 L'assunzione appare più discutibile se si tratta, invece, di errori di risposta. Si pensi, ad esempio, all'effetto del *proxy respondent*.

Alla luce delle osservazioni precedenti, l'ipotesi di indipendenza e la mancata considerazione del fattore di bloccaggio appaiono ragionevoli condizionatamente ad H_0. Perciò, il numeratore del rapporto viene scomposto in

$$P(\mathbf{x}|H_0) = \Pi_i P(x_i|H_0),$$

e ciascuna componente $P(x_i|H0)$ è stimata separatamente come frequenza relativa del risultato x_i nell'insieme degli abbinati individuato al primo passo dalla procedura MATCH.

Apprezzabili modifiche sono invece introdotte per la stima di $P(\mathbf{x}|H_1)$, al fine di riflettere l'impatto del bloccaggio e di evitare il ricorso all'ipotesi di indipendenza. Il significato di tali modifiche può essere chiarito dalla seguente scomposizione:

$$\begin{aligned} P(\mathbf{x}|H_1) &= [P(\mathbf{x}H_1F) + P(\mathbf{x}H_1F\hat{})]/P(H_1) \\ &= [P(\mathbf{x}|H_1F)P(H_1|F)P(F) + P(\mathbf{x}|H_1F\hat{})P(H_1|F\hat{})P(F\hat{})]/P(H_1) , \end{aligned}$$

dove F e F^ indicano, rispettivamente, che il confronto si riferisce ad individui appartenenti o non appartenenti alla stessa famiglia.

Tale scomposizione permette di considerare esplicitamente la possibilità che i confronti effettuati all'interno di ogni blocco possano riferirsi anche a famiglie diverse, a causa delle sostituzioni e della presenza di errori nel codice famiglia.

Evidentemente, in assenza di sostituzioni e/o errori, il secondo addendo in parentesi quadra sarebbe nullo (risultando $P(F\hat{}) = 0$). Quando ciò non avviene, pare verosimile che la distribuzione di **x**, per i confronti che avvengono tra individui diversi della stessa famiglia, differisca dall'analoga distribuzione relativa ai confronti all'interno di famiglie diverse e conviene, quindi, valutare in modo differenziato i due addendi.

Il modo in cui si è pervenuti ad una stima delle varie componenti di $P(\mathbf{x}|H_1)$ è il seguente:

- $P(\mathbf{x}|H_1F)$: è la probabilità di osservare **x** confrontando individui diversi della stessa famiglia. Tale probabilità è stimata, separatamente per famiglie di differente numerosità, come rapporto tra il numero di confronti che in MATCH hanno condotto ad un non abbinamento e il numero totale dei confronti fatti. In questo modo, si evita il ricorso all'ipotesi di indipendenza e si introduce, nella stima del peso, un primo elemento che tiene conto delle eventuali differenze tra le distribuzioni di **x** condizionate alle diverse caratteristiche del blocco nel quale di volta in volta si opera.
- $P(H_1|F)$: se la famiglia è la stessa nelle due occasioni di indagine, la probabilità che un confronto non si riferisca allo stesso individuo dipende, evidentemente, dal numero totale dei confronti che si possono fare all'interno della famiglia e dal numero di individui che sono realmente presenti in entrambe le occasioni. Indicando con n_A ed n_B la numerosità della famiglia alle due occasioni ed ipotizzando che, trattandosi della stessa famiglia, il $\min(n_A,n_B)$ sia il numero di individui presenti entrambe le volte, una stima di tale quantità può essere ottenuta come $[n_An_B$-

$\min(n_A,n_B)]/n_An_B=1-1/\max(n_A,n_B)$. In questo modo il peso attribuito ai vari confronti varia al variare della numerosità della famiglia. Si introduce, cioè, nella valutazione, un ulteriore elemento legato alle caratteristiche del blocco.

- P(F): è la probabilità che, dato il bloccaggio, il confronto si riferisca ad individui della stessa famiglia. Tale probabilità è legata a tutti quei fattori, come le sostituzioni di famiglie, gli errori nei codici, i cambiamenti di liste non previsti, che si risolvono nell'attribuzione dello stesso codice familiare a due famiglie diverse nelle due occasioni. Da valutazioni già riportate in Giusti, Marliani e Torelli (1987) e da informazioni contenute in Masselli (1988), si può ritenere che questi fattori agiscano in circa l'8% dei casi, sicché si è assunto $P(F^\wedge)=0,08$ e, quindi, $P(F)=0,92$[9].
- $P(\mathbf{x}|H_1F^\wedge)$: è la probabilità di osservare **x** quando il confronto si riferisce ad individui diversi che appartengono a famiglie diverse e può essere ragio>evolmente interpretata come la probabilità di osservare quel confronto estraendo a caso un individuo da ciascuno dei due archivi. In mancanza di informazioni sui confronti tra individui di famiglie diverse, si ricorre all'ipotesi di indipendenza e si stima separatamente ciascuna componente $P(x_i|H_1F^\wedge)$ come $\sum_i \sum_j {}_Af_{hi}\ {}_Bf_{ji}$, dove ${}_Af_{hi}$ e ${}_Bf_{ji}$ indicano la frequenza relativa della i-esima variabile di confronto nei due archivi A e B e la doppia sommatoria si estende alternativamente a tutte le coppie di modalità (h,j) che sono codificate come accordo se ($x_i=0$) o disaccordo se ($x_i=1$)[10].
- $P(H_1|F^\wedge)$: quando il confronto avviene tra famiglie diverse c'è la pratica certezza che si riferisca ad individui diversi. Si assume quindi che questa probabilità sia pari ad uno.
- $P(H_1)$: rappresenta la probabilità che un confronto preso a caso, tra quelli effettuati tenendo conto del bloccaggio, non si riferisca allo stesso individuo. Essa può essere ragionevolmente stimata, sulla base dei risultati ottenuti da MATCH, come rapporto tra il numero dei confronti che non hanno dato luogo ad abbinamento ed il numero totale dei confronti fatti.

Valutate nel modo descritto le diverse componenti, si determina il peso totale $T(\mathbf{x})$ e se ne calcola il $\log_2$. La procedura prevede, infine, che si individui un valore soglia k e adotta la seguente regola di decisione:

- se $\log_2 T(\mathbf{x}) \geq K$ si abbina,
- se $\log_2 T(\mathbf{x}) < K$ non si abbina.

---

9 La scelta dei due valori 0,92 e 0,08, anche se basata su valutazioni empiriche, può apparire non sufficientemente motivata, soprattutto considerando che i due valori vengono mantenuti fissi nei vari abbinamenti, mentre il numero di errori nei codici familiari e/o sostituzioni può cambiare da occasione ad occasione. Si osservi però che i due valori agiscono come pesi nella media ponderata delle due componenti del denominatore, delle quali la prima è certamente dominante. Finché il rapporto tra i due pesi è dell'ordine di 10 a 1, come nel nostro caso, piccole variazioni, necessarie se si volesse tener conto della diversa probabilità di errore alle differenti indagini, non dovrebbero comportare una sostanziale modifica nel peso totale.

10 Se si definisce accordo solo la situazione nella quale la i-esima variabile presenta la stessa modalità nei due *records*, la stima è data da $P(x_i=0|H_1) = \Sigma_h\ {}_Af_{hi}\ {}_Bf_{hi}$ e $P(x_i=1|H_1) = 1-P(x_i=0|H_1)$.

La scelta del valore soglia è stata effettuata empiricamente (sulla scorta delle indicazioni di Howe e Lindsay, 1981) ispezionando la distribuzione empirica dei pesi e tenendo conto che il fattore di bloccaggio basato sul codice familiare costituisce già una ragionevole protezione contro il rischio di falsi positivi, e consente quindi di utilizzare un valore soglia abbastanza basso. Nella nostra applicazione ai dati del Veneto e Lombardia abbiamo posto k=0. Ciò equivale ad abbinare quando la probabilità di osservare un vettore di confronto **x**, sotto l'ipotesi che si tratti dello stesso individuo, è superiore a quella dello stesso risultato nell'ipotesi che gli individui confrontati siano diversi.

La procedura non consente abbinamenti multipli. Se più individui del file A risultano abbinabili con lo stesso individuo del file B o viceversa, si abbinano i due al cui confronto è associato il peso maggiore e, nel caso che più confronti presentino lo stesso peso (cosa estremamente improbabile dato il fattore di bloccaggio), si abbinano i due relativi al primo confronto effettuato.

Possono risultare utili alcuni chiarimenti circa la logica di funzionamento di LINK, al fine di evidenziarne gli aspetti più convincenti.

Avendo introdotto un peso differenziato in relazione alla dimensione familiare, si considera esplicitamente una fonte di variabilità nei pesi legata alle caratteristiche del blocco. In particolare, a parità di **x**, il criterio di abbinamento risulta più selettivo al crescere della dimensione familiare. Si consideri, ad esempio, il confronto per famiglie composte da un solo individuo. In tal caso il denominatore del peso totale $T(\mathbf{x})$ è rappresentato dalla sola componente $P(\mathbf{x}H_1F^\wedge)/P(H_1)$ (la prima componente si annulla, essendo $P(H_1|F)=0$). In altri termini, per famiglie di una sola persona, si tende ad abbinare con maggior facilità (soprattutto se, com'è nel nostro caso, $P(F^\wedge)$ è piccolo rispetto a $P(H_1)$). Addirittura, se non fossero presenti errori nei codici familiari, gli individui appartenenti a famiglie di una sola persona risulterebbero sempre abbinati, com'è ragionevole, anche quando vi fosse disaccordo in tutte le variabili di confronto. Al crescere della numerosità familiare, invece, la prima componente del denominatore assume via via più importanza e, di conseguenza, a parità di **x**, il peso tende a ridursi.

Nonostante questa caratteristica, intuitivamente attraente, non vanno sottaciute alcune limitazioni legate soprattutto alla stima di $P(\mathbf{x}|H_1F)$, che andrebbe riconsiderata con maggiore attenzione. Tuttavia, un suo miglioramento è legato alla possibilità di identificare in MATCH i confronti fatti in famiglie che sono realmente le stesse nelle due occasioni. Tale strada non è percorribile, se non con larghi margini di approssimazione e a meno di ricorrere ad una ispezione manuale; operazione, quest'ultima, antieconomica, se non impraticabile, per larghe basi di dati.

4. *Un breve esame dei risultati*

In questa sezione si presentano brevemente i risultati ottenuti con LINK per l'abbinamento di due occasioni di indagine della RTFL.

Gli abbinamenti riguardano, distintamente, le due regioni Veneto e Lom-

bardia. Le occasioni di indagine considerate sono 1985.I, 1985.II, 1986.I e 1986.II. A partire da queste è possibile effettuare due abbinamenti a distanza di un trimestre (85.I-II e 86.I-II) e due a distanza di un anno (85.I-86.I e 85.II-86.II).

Prima di passare ad esaminare i risultati ottenuti in termini di numero di abbinamenti, è utile proporre alcune considerazioni sul significato dei pesi prodotti dalla procedura e sulla giustificazione del valore soglia adottato.

Ovviamente non è possibile presentare i pesi effettivamente utilizzati da LINK poiché essi variano a seconda della configurazione del vettore **x** e della dimensione familiare. A solo titolo indicativo, nella Tab. 3 si riporta, per i soli abbinamenti della Lombardia, il peso teorico che LINK avrebbe attribuito a ciascuna variabile, nell'ipotesi che questa fosse stata la sola utilizzata come termine di confronto per decidere dell'abbinamento, distinguendo tra famiglie di due e quattro persone. L'ispezione dei dati consente alcune prime osservazioni.

Come era ragionevole attendersi, le variabili che hanno un maggior potere discriminante, sia in accordo che in disaccordo, sono il mese e soprattutto l'anno di nascita. La probabilità di osservare un accordo casuale in tali variabili è, infatti, relativamente più bassa, dato l'alto numero di modalità, e questo riduce il denominatore di T(**x**). Inoltre, l'errore di osservazione risulta, per le stesse variabili, tra i più contenuti e ciò contribuisce, in caso di disaccordo, a ridurre il numeratore.

Se si prende invece, ad esempio, il sesso, si osserva che un disaccordo può avere un peso determinante nel decidere di non abbinare, mentre, per contro, l'elevata probabilità che due individui estratti a caso abbiano lo stesso sesso, fa sì che un accordo comporti un contributo modesto a favore della decisione di abbinamento. Allo stesso modo, l'elevata probabilità di un accordo casuale, legata al criterio con cui l'accordo viene definito, giustifica il basso peso attribuito all'accordo in variabili come lo stato civile, la condizione professionale e le precedenti esperienze lavorative, che hanno invece un'alta capacità discriminante in caso di disaccordo.

L'aspetto più rilevante, comunque, risiede nella capacità della procedura proposta di produrre pesi che si adattano all'eventuale presenza di errori anomali nei dati. Al riguardo, si può osservare il peso della variabile relazione con il capofamiglia nei diversi collegamenti effettuati. Quando sono coinvolti i dati dell'indagine del primo trimestre del 1985, l'importanza del disaccordo in questa variabile si riduce sensibilmente. Ciò è da attribuire alla elevata quota di errori nella relazione col capofamiglia, che si riscontrano nei dati del 1985.I (vedi la precedente nota 6). In altri termini, valutando il peso sulla base delle distribuzioni dei due archivi, la procedura è in grado di 'imparare' dai dati stessi ad assegnare un diverso potere discriminante alle variabili.

Un'ulteriore osservazione sui dati della Tab. 3 riguarda la diversa importanza attribuita da LINK a ciascuna variabile, in relazione alla differente dimensione familiare. Generalmente, il peso di un accordo decresce con l'aumentare della dimensione familiare (è più alta la probabilità di un accordo casuale), mentre tende ad aumentare, in valore assoluto, quello di un disaccordo. Vi sono variabili, come, ad esempio, il sesso, per le quali il peso di

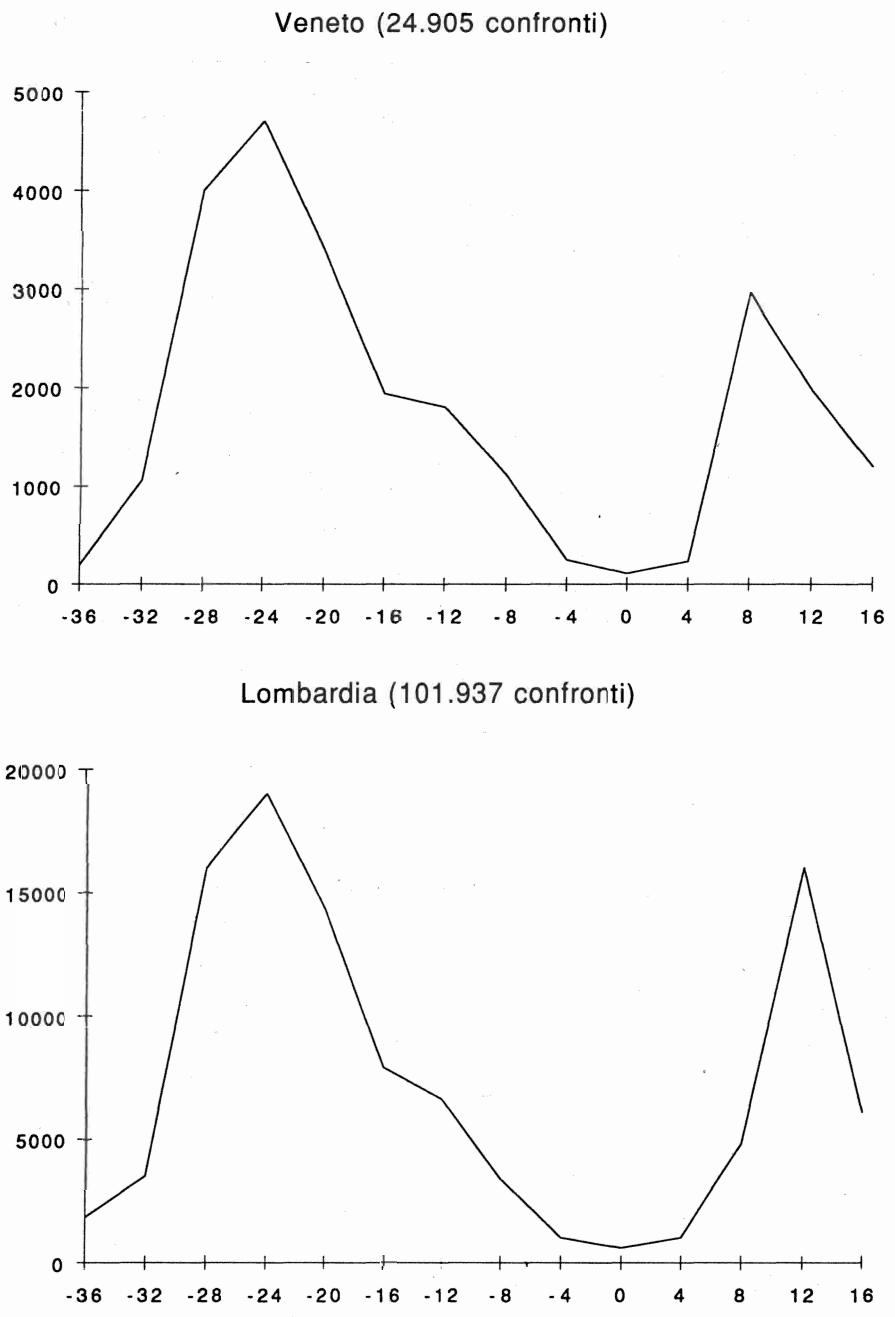
un accordo osservato in famiglie di due persone è più che doppio rispetto a quello analogo osservato in famiglie di quattro persone.

Infine, qualche osservazione può essere fatta anche con riferimento alla scelta del valore soglia. In buona parte essa deriva dalle considerazioni, già svolte nella precedente sezione, a proposito della protezione che il fattore di bloccaggio adottato garantisce nei confronti dei falsi positivi. Tale opzione non è, ovviamente, definitiva. Una sua revisione può derivare, oltre che dalla diversa propensione al rischio di un errato abbinamento - legata eventualmente a particolari esigenze di analisi dei dati longitudinali -, dall'esame della distribuzione empirica dei pesi. Nella Fig. 1 sono riportate, come esempio, le rappresentazioni delle distribuzioni dei pesi totali di tutti i confronti fatti con LINK per i dati del Veneto e della Lombardia 86.I-II. Tale distribuzione può essere interpretata come miscuglio tra le due distribuzioni $p(\mathbf{x}|H_1)$ e $p(\mathbf{x}|H_0)$. Se non si hanno particolari motivi per proteggersi maggiormente contro uno dei due errori, è ragionevole collocare il valore soglia in corrispondenza del minimo nella porzione positiva della curva (Coulter, 1985), che nel nostro caso risulta prossimo allo zero per tutti gli abbinamenti effettuati.

Tab. 3: *Pesi teorici che la procedura LINK avrebbe assegnato all'accordo e al disaccordo in ciascuna variabile di confronto se questa fosse stata la sola utilizzata nell'abbinamento (abbinamenti della Lombardia)*

variabile di confronto	risultato del confronto	85.I-II		86.I-II		85.I-86.I		85.II-86.II	
		fam. 2 pers.	fam. 4 pers.	fam. 2 pers.	fam. 4 pers.	fam. 2 pers.	fam. 4 pers.	fam. 2 pers.	fam. 4 pers.
Mese nascita	accordo	3,71	2,46	3,88	2,55	2,97	2,19	3,13	2,32
	disacc.	-5,97	-6,49	-5,44	-5,97	-5,29	-5,81	-4,84	-5,36
Anno nascita	accordo	4,64	4,40	4,51	4,85	3,94	3,86	4,28	4,37
	disacc.	-5,83	-6,35	-5,23	-5,75	-5,08	-5,59	-4,48	-5,00
Sesso	accordo	2,32	1,13	2,58	1,15	1,66	,96	1,71	1,01
	disacc.	-6,03	-6,57	-5,90	-6,45	-5,65	-6,19	-5,69	-6,23
Rel. capof.	accordo	2,66	1,41	3,57	1,78	1,69	1,03	2,23	1,44
	disacc.	-1,42	-2,04	-5,92	-6,46	-1,25	-1,81	-5,18	-5,71
Istruzione	accordo	1,95	1,25	2,02	1,25	1,57	1,04	1,56	1,04
	disacc.	-3,02	-3,62	-2,74	-3,36	-3,64	-4,19	-3,03	-3,59
Stato civile	accordo	0,94	0,40	0,94	0,39	0,92	0,39	0,93	0,39
	disacc.	-6,66	-7,51	-6,87	-7,78	-7,44	-8,07	-7,28	-7,91
Condiz. prof.	accordo	1,42	0,70	1,49	0,71	1,18	0,63	1,20	0,65
	disacc.	-9,44	-10,16	-9,14	-9,70	-9,01	-9,64	-8,11	-8,72
Preced. lavor.	accordo	1,50	0,73	1,57	0,75	1,24	0,66	1,25	0,67
	disacc.	-5,67	-6,32	-5,31	-5,97	-5,26	-5,84	-4,78	-5,37

Fig. 1: *Distribuzione empirica dei pesi assegnati dalla procedura LINK a tutti i confronti effettuati negli abbinamenti 86.I-II del Veneto e della Lombardia*



Veniamo adesso all'esame delle Tabb. 4 e 5 che riassumo i risultati di LINK, per i diversi abbinamenti, ponendoli a confronto con quelli ottenuti al primo passo utilizzando la procedura MATCH. Ci limitiamo ad alcune sintetiche considerazioni.

(a) LINK produce una quota di abbinati lievemente superiore a quella, già elevata, ottenuta con MATCH. L'incremento non è certo rilevante (mediamente mezzo punto percentuale nei collegamenti 86.I-II e 85.II-86.II), ma la cosa pare da attribuire prevalentemente al buon risultato già raggiunto con MATCH. Infatti, laddove esistono margini di recupero, come negli abbinamenti che coinvolgono il primo trimestre 1985, il guadagno è nettamente più consistente (circa 1,5 punti percentuali). In questo caso, la flessibilità del sistema dei pesi giuoca un ruolo rilevante. La minor severità di MATCH, rispetto alla procedura utilizzata dall'Istat, consente

Tab. 4: *Abbinati dalle procedure MATCH e LINK secondo il risultato dell'abbinamento all'interno del blocco familiare*

Procedura	Veneto				Lombardia			
	85.I 85.II	86.I 86.II	85.I 86.I	85.II 86.II	85.I 85.II	86.I 86.II	85.I 86.I	85.II 86.II
Presenti 1° occas.	7.540	7.457	6.371	6.235	33.234	33.426	25.387	25.859
Presenti 2° occas.	7.516	7.427	6.161	6.142	33.289	33.149	27.182	27.068
Abbinati in complesso								
MATCH	6.227	6.272	4.781	4.922	25.550	27.615	16.384	17.167
(% su abbinabili) [(a)]	*83,8*	*84,4*	*79,1*	*80,1*	*76,9*	*83,3*	*64,5*	*66,4*
LINK	6.395	6.301	4.954	4.916	26.070	27.741	17.147	17.525
(% su abbinabili)	*85,1*	*84,8*	*80,4*	*80,0*	*78,3*	*83,7*	*67,5*	*67,8*
Abbinati con codici familiari per i quali si abbinano tutti i componenti								
MATCH	5.445	5.757	4.079	4.205	23.028	25.450	14.014	14.084
(% su abbinabili)	*72,4*	*77,5*	*66,2*	*68,5*	*69,3*	*76,8*	*55,2*	*54,5*
LINK	5.864	5.816	4.294	4.198	24.252	25.715	15.060	15.083
(% su abbinabili)	*78,0*	*78,3*	*69,7*	*68,3*	*72,9*	*77,6*	*59,3*	*58,4*
Abbinati con codici familiari per i quali si abbinano solo alcuni componenti								
MATCH	782	515	792	717	2.522	2.165	2.370	3.083
(% su abbinabili)	*10,4*	*6,9*	*12,9*	*11,8*	*7,6*	*6,5*	*9,3*	*11,9*
LINK	531	485	660	718	1.818	2.026	2.087	2.442
(% su abbinabili)	*7,1*	*6,5*	*10,7*	*11,7*	*5,4*	*6,1*	*8,2*	*9,4*

(a) La percentuale è riferita al più piccolo numero di presenti ad una delle due occasioni, ovvero al numero di abbinamenti teoricamente possibili.

Tab. 5: *Non abbinati dalle procedure MATCH e LINK secondo il risultato dell'abbinamento all'interno del blocco familiare*

Procedura		Veneto				Lombardia			
		85.I 85.II	86.I 86.II	85.I 86.I	85.II 86.II	85.I 85.II	86.I 86.II	85.I 86.I	85.II 86.II
	Non abbinati con codici familiari per i quali si abbina almeno un componente								
MATCH:	tipo 10 [(a)]	370	232	369	365	1.322	992	1.541	1.540
	tipo 01	388	246	317	264	1.408	1.042	1.007	1.202
LINK:	tipo 10	226	220	289	341	955	932	1.541	1.540
	tipo 01	245	235	232	251	1.033	990	1.270	1.454
	Non abbinati con codici familiari, presenti ad entrambe le occasioni, per i quali non si abbina alcun individuo								
MATCH:	tipo 10	497	578	690	624	3.440	2.437	4.946	4.193
	tipo 01	474	608	665	624	3.359	2.451	5.057	4.238
LINK:	tipo 10	473	561	687	654	3.287	2.371	4.114	3.578
	tipo 01	449	590	667	643	3.214	2.377	4.031	3.628
	Non abbinati perché il codice familiare è presente ad una sola occasione								
	tipo 10	446	375	441	324	2.922	2.382	2.516	2.959
	tipo 01	427	301	308	332	2.972	2.041	4.734	4.461

(a) 1 identifica la presenza e 0 l'assenza. Il tipo 10 si riferisce, quindi, ad un individuo che è presente solo alla prima occasione, mentre il tipo 01 ad un individuo che è presente solo alla seconda.

già un notevole recupero su una situazione compromessa da forti errori nei dati; ma la rigidità del sistema dei pesi non permette di andare oltre certi limiti. Tra l'altro, il fatto che si ottengano risultati numericamente vicini anche partendo da un abbinamento preliminare meno efficace, fa ritenere che la procedura LINK non sia eccessivamente condizionata dai risultati ottenuti al primo passo.

(b) E' confortante osservare che la maggior quota di abbinamenti si ottiene da un considerevole aumento di famiglie abbinate completamente (famiglie nelle quali si abbinano tutti i componenti). Quando ciò avviene, è verosimile pensare che il guadagno non sia imputabile ad un aumento di errati abbinamenti ma debba, invece, essere attribuito al recupero di situazioni che in MATCH si risolvevano in falsi negativi.

(c) Per tutte le procedure, compresa quella più semplice adottata dall'Istat, per i cui risultati si rinvia a Giusti, Marliani e Torelli (1987), la quota di abbinati diminuisce passando da due occasioni a distanza di un trimestre a due occasioni a distanza di un anno. Questa differenza trova certamente una sua giustificazione se si considera che, in arco di tempo più lungo, aumenta la possibilità che si verifichino eventi (ingressi, uscite, irreperibilità delle famiglie e conseguente sostituzione) che provocano una riduzione dei presenti ad entrambe le indagini. Colpisce però la dimensione

del fenomeno, soprattutto se si guarda ai dati della Lombardia, che vede ridursi di 15 punti percentuali la quota di abbinati. Evidentemente, in questo caso, altri fattori entrano in giuoco nel determinare il non abbinamento; e di ciò si ha una conferma dai dati della Tab.5.

(d) Tali dati mostrano, infatti, come il numero teorico di abbinabili, utilizzato per il calcolo della percentuale, sovrastimi decisamente il numero di individui effettivamente presenti alle due occasioni. In generale, la gran parte dei non abbinamenti è, per così dire, di tipo familiare, cioè coinvolge tutti gli individui che presentano lo stesso codice famiglia. Evidentemente, le sostituzione e gli errori nei codici comportano o che si mettano a confronto famiglie diverse, generando una situazione di interi blocchi famiglia che risultano presenti alle due occasioni ma per i quali non si abbina alcun individuo, o che alcuni codici familiari risultino presenti ad una sola delle due occasioni.

5. *Alcune verifiche sui risultati dell'abbinamento tra due occasioni di indagine*

Al di là della valutazione numerica delle *performances* in termini di numero di abbinamenti effettuati, la verifica dei risultati di una procedura di abbinamento esatto dovrebbe essere basata, almeno in linea teorica, sull'esame degli errori ad essa connessi, vale a dire sulla individuazione ed analisi dei falsi positivi e negativi. Purtroppo, a meno di non ricorrere a simulazioni, individuare gli errori è tutt'altro che agevole. Si può cercare, tuttavia, di identificare una serie di situazioni in un certo senso sospette, nelle quali, cioè, il rischio di errato abbinamento o mancato abbinamento sembra più elevato. In questa sezione si presentano in modo sintetico i risultati di alcune verifiche condotte in tal senso sugli abbinamenti effettuati con LINK. Verifiche analoghe, per il confronto tra i risultati della procedura MATCH rispetto a quella dell'Istat, sono in Giusti, Marliani e Torelli (1987), al quale si rinvia per maggiori dettagli sui criteri cui le verifiche stesse sono ispirate.

5.1. *Abbinati e falsi positivi*

Come abbiamo visto, la procedura LINK, così come MATCH, ammette l'abbinamento anche in presenza di disaccordo tra alcune variabili. Ne consegue che, nei *files* longitudinali riferiti a ciascun individuo, possono essere presenti informazioni che risultano incompatibili o incoerenti, se confrontate alle due occasioni[11].

Nell'ottica di individuare situazioni di sospetti falsi positivi, una prima

---

11 La logica della procedura Istat, invece, è quella di produrre abbinati 'perfetti' dal punto di vista delle incompatibilità. In realtà, essa garantisce l'assenza di incompatibilità limitatamente alle sole variabili sesso e relazione con il capofamiglia. Infatti, se si controllano gli abbinati Istat sulla base delle più precise

verifica si può condurre esaminando il numero di incompatibilità presenti nei *records* degli abbinati. La Tab. 6 riporta, appunto, le incompatibilità, definite secondo lo schema della Tab. 2, rilevate per gli abbinati tra due occasioni di indagine con le procedura LINK.

Il numero di abbinati per i quali si registra completo accordo nelle otto variabili di confronto rappresenta la grande maggioranza degli abbinamenti effettuati (intorno all'85%). Ciò naturalmente non garantisce che in tali abbinamenti non siano presenti falsi positivi, ma certo questi non dovrebbero essere troppo probabili.

Consistente è anche la quota di individui che vengono abbinati in presenza di un disaccordo. C'è da dire però che, pagando questo prezzo, si riescono riescono a recuperare situazioni altrimenti compromesse, come quella, già citata, degli errori nella relazione con il capofamiglia del primo trimestre 1985. Dall'altro canto, anche nel caso di un solo disaccordo il rischio di falso positivo non dovrebbe essere molto elevato. Conviene ricordare, infatti, che gli individui da abbinare vengono scelti in famiglie che, alle due indagini, presentano lo stesso codice di identificazione. Nella maggior parte dei casi, questo significa che si tratta della stessa famiglia; in tale ipotesi, un abbinamento tra due individui che presentano un accordo in sette

Tab. 6: *Distribuzione degli abbinati secondo il numero di disaccordi osservati nelle variabili di confronto*

Numero di disaccordi	85.I-II		86.I-II		85.I-86.I		85.II-86.II	
	N.	%	N.	%	N.	%	N.	%
				Veneto				
0	4.360	68,2	5.299	84,1	3.400	68,6	4.287	87,2
1	1.823	28,5	956	15,2	1.420	28,7	598	12,2
2	211	3,3	46	0,7	133	2,7	31	0,6
3	1	..	-	-	1	..	-	-
				Lombardia				
0	18.530	71,1	24.030	86,7	11.437	66,7	14.667	83,7
1	6.872	26,4	3.521	12,6	4.839	28,2	2.470	14,1
2	665	2,5	189	0,7	830	4,9	379	2,2
3	3	..	1	..	41	0,2	9	..

[segue nota]

definizioni di accordo/disaccordo adottate in LINK, possono risultare incompatibili la data di nascita e il titolo di studio, così come, ovviamente, possono essere incompatibili le informazioni relative a stato civile, condizione professionale e precedenti esperienze lavorative, che l'Istat non utilizza come variabili di confronto.

variabili su otto e per i quali non esiste, in quella famiglia, la possibilità di trovare un abbinamento migliore (accordo completo), è da considerare quasi certamente un abbinamento corretto. Se, invece, le due famiglie sono diverse, o per un errore nel codice o perché vi è stata una sostituzione, e quindi i confronti vengono effettuati tra individui diversi, è molto difficile che si osservi un solo disaccordo nelle otto variabili considerate.

Un sospetto crescente si può avanzare nei riguardi degli abbinamenti che avvengono in presenza di più di un disaccordo. Con la solita esclusione degli abbinamenti interessati dal 1985.I, questi rappresentano tuttavia una quota sostanzialmente trascurabile.

Per identificare situazioni sospette quanto a falsi positivi, una seconda verifica può essere fatta analizzando il risultato degli abbinamenti effettuati all'interno di uno stesso codice familiare. In particolare, è opportuno controllare se, nell'ambito della famiglia, si è abbinato, tra gli altri, un elemento *pivot*, come il capofamiglia o il coniuge, oppure se l'abbinamento riguarda solo altri componenti, come, ad esempio, i figli. Si deve osservare infatti che, quando si pongono a confronto individui diversi, per effetto di errori nei codici o sostituzioni di famiglie, il rischio di falso positivo è maggiore per le coppie di giovanissimi poiché, nel confrontare due bambini, anche diversi, la relazione con il capofamiglia, il titolo di studio, lo stato civile, la condizione professionale e i precedenti lavorativi risultano forzatamente in accordo. Le uniche variabili che mantengono un potere discriminante sono, quindi, mese ed anno di nascita e sesso e non è improbabile registrare, per puro caso, due accordi (ad esempio, sesso e mese di nascita) in queste tre variabili.

In base a questa considerazione, si può ammettere che se valgono contemporaneamente le due condizioni:

(a) nella famiglia si abbina meno della metà degli individui rispetto al numero dei teoricamente abbinabili (il più piccolo numero tra i componenti alla prima e alla seconda occasione)

(b) tra gli abbinati non è presente almeno il capofamiglia e/o il coniuge,

allora è ragionevole sospettare che gli abbinati di quella famiglia siano in realtà dei falsi positivi.

I dati della Tab. 7 forniscono a riguardo risultati confortanti. Il numero di abbinati da LINK in famiglie nelle quali non si abbina né il capofamiglia né il coniuge è assai ridotto. Tra questi, la situazione più sospetta, relativa a famiglie nelle quali si abbina non più del 50% dei componenti (indicata in tabella con A1) raggiunge, nel peggiore dei casi, lo 0,9% degli abbinati.

Un'ulteriore verifica, che può fornire qualche indicazione sul rischio di falso positivo, è legata all'esame della distribuzione empirica dei pesi totali.

Come si ricorderà, non tutti i confronti per i quali risulta $T(\mathbf{x})> 0$ si risolvono in un abbinamento. Avendo escluso la possibilità di abbinamenti multipli, infatti, può accadere che due individui, il cui confronto dà luogo ad un peso totale superiore alla soglia, non vengano abbinati perché, per uno dei due, si trova un abbinamento migliore (peso più alto). La proporzione di questi casi *può essere* assunta come valutazione approssimata di $P(T(\mathbf{x})> 0|H_1)$ ed

Tab. 7: *Distribuzione degli abbinati secondo il numero di abbinamenti all'interno della famiglia e l'abbinamento o meno del capofamiglia e/o del coniuge.*

Tipo di abbinamento [(a)]	85.I-II N.	85.I-II %	86.I-II N.	86.I-II %	85.I-86.I N.	85.I-86.I %	85.II-86.II N.	85.II-86.II %
			Veneto					
A1	18	0,3	12	0,2	18	0,3	12	0,2
A2	28	0,4	11	0,2	29	0,6	11	0,2
A3	22	0,3	36	0,6	29	0,6	44	0,9
A4	6.327	99,0	6.242	99,0	4.878	98,5	4.849	98,7
Totale	6.395	100,0	6.301	100,0	4.954	100,0	4.916	100,0
			Lombardia					
A1	58	0,2	48	0,2	175	0,9	138	0,8
A2	88	0,3	36	0,1	121	0,7	37	0,2
A3	145	0,6	185	0,7	247	1,5	303	1,7
A4	25.779	98,9	27.472	99,0	16.604	96,9	17.084	97,3
Totale	26.070	100,0	27.741	100,0	17.147	100,0	17.525	100,0

(a) A1: Abbinati in famiglie nelle quali non si abbina né capofamiglia né coniuge e che presentano un numero di abbinati minore o uguale al 50% dei teoricamente abbinabili.

A2: Abbinati in famiglie nelle quali non si abbina né capofamiglia né coniuge e che presentano un numero di abbinati maggiore al 50% dei teoricamente abbinabili.

A3: Abbinati in famiglie nelle quali si abbina il capofamiglia e/o il coniuge e che presentano un numero di abbinati minore o uguale al 50% dei teoricamente abbinabili.

A4: Abbinati in famiglie nelle quali si abbina il capofamiglia e/o il coniuge e che presentano un numero di abbinati maggiore al 50% dei teoricamente abbinabili.

Tab. 8: *Numero di confronti con peso positivo e inferiore a 12 per risultato dell'abbinamento e classe di peso.*

Occasioni di indagine collegate	classe di peso	Veneto Abbinati	Veneto Non abbinati	Veneto % di non abbinati	Lombardia Abbinati	Lombardia Non abbinati	Lombardia % di non abbinati
85.I-II	0 - 3	191	53	21,7	594	255	30.0
	3 - 6	67	9	11,8	397	66	14,2
	6 - 9	631	16	2,5	2.988	52	1,7
	9 -12	2.276	15	0,6	11.370	50	0,4
86.I-II	0 - 3	57	55	49,1	276	241	46,7
	3 - 6	149	14	8,6	424	51	10,7
	6 - 9	533	10	1,8	4.046	64	1,6
	9 -12	2.453	20	0,8	8.483	41	0,5
85.I-86.I	0 - 3	198	48	19,5	313	219	41,2
	3 - 6	63	9	12,5	358	44	10,9
	6 - 9	425	9	2,1	1.841	29	1,6
	9 -12	2.874	10	0,3	5.646	30	0,5
85.II-86.II	0 - 3	47	42	47,2	232	218	48,4
	3 - 6	53	8	13,1	411	40	8,9
	6 - 9	675	14	2,0	944	24	2,5
	9 -12	1.646	12	0,7	6.668	35	0,5

è in grado di fornire un'idea del rischio di falso positivo connesso ad una procedura e alla scelta di un determinato valore soglia.

La Tab. 8 contiene il numero di confronti con peso maggiore della soglia scelta, distinguendo tra quelli che hanno condotto all'abbinamento e quelli 'potenzialmente abbinabili', che sono, invece, stati scartati. I dati forniscono un'ulteriore testimonianza della validità della procedura adottata: la quota di non abbinati (espressione del rischio) cade drasticamente all'aumentare del peso e si mostra relativamente consistente solo nei dintorni della soglia; tuttavia, anche in questi casi, il problema riguarda un numero di confronti assai contenuto.

5.2. *Non abbinati e falsi negativi*

L'altro versante sul quale occorre valutare una procedura di abbinamento è quello dei falsi negativi. In altri termini, ci si può chiedere se sussistono margini di guadagno sulla quota dei non abbinati.

A tale riguardo, sarebbe necessario distinguere l'insieme dei non abbinati nei due sottoinsiemi dei falsi negativi e dei non compresenti. I primi riguardano direttamente il modo di operare della procedura di abbinamento. Come si è detto, si tratta di *records* che, pur essendo relativi al medesimo soggetto, non danno luogo ad abbinamento a causa di errori nelle variabili di confronto. Tra i secondi, invece, rientrano tutti quei *records* relativi a coloro che sono entrati a far parte del campione solo successivamente alla prima occasione di indagine (nati e nuove acquisizioni dei nuclei familiari inclusi nel campione), come anche coloro che, presenti una prima volta, sono poi usciti (per morte, migrazione, scissione del nucleo familiare). Naturalmente, la non compresenza può verificarsi anche a livello di un'intera famiglia, nel caso in cui questa venga sostituita.

E' ovvio che una corretta valutazione della *performance* di una procedura di abbinamento deve tener conto soltanto dei falsi negativi, escludendo i presenti ad una sola occasione, che dovrebbero essere eliminati dal computo degli abbinabili.

L'identificazione dei due insiemi (falsi negativi e non compresenti) è tuttavia problematica. Sul piano dell'abbinamento, infatti sia il falso negativo che il non compresente producono lo stesso risultato, generando *records* della forma 10 (individuo presente solo alla prima occasione) oppure 01 (individuo presente solo alla seconda occasione). Sempre nell'ottica di individuare situazioni sospette, conviene distinguere due casi:

(a) il non abbinamento è a livello individuale, cioè si verifica all'interno di una famiglia (codice familiare) in cui sono stati abbinati alcuni individui;
(b) il non abbinamento riguarda l'intera famiglia, cioè si riscontra un intero codice familiare per il quale si osservano esclusivamente *records* della forma 10 e/o 01.

A livello individuale, la presenza di un falso negativo comporta sempre, nell'insieme dei *records* riferiti ad una famiglia, una situazione di ingressi e uscite fittizie che si bilanciano numericamente: uno stesso individuo presente

alle due occasioni e non abbinato a causa di errori nelle variabili di confronto dà, infatti, luogo a due *records*, uno di tipo 10 e uno di tipo 01. Tuttavia, l'esistenza nella famiglia di un numero di ingressi e uscite bilanciati non permette di concludere con certezza che si è in presenza di falsi negativi. Possono infatti verificarsi situazioni di bilanciamento anche per il solo movimento naturale di ingresso e uscita e, d'altra parte, vi può essere uno sbilanciamento anche in presenza di falsi negativi, quando questi si verificano in famiglie che presentano anche ingressi o uscite per cause naturali.

Una valutazione al riguardo non è quindi agevole. Si può tuttavia convenire che le situazioni più sospette siano quelle in cui valgono contemporaneamente le seguenti condizioni:

(a) si abbinano due o più persone, almeno una delle quali è il capofamiglia o il coniuge (ciò dovrebbe fornire sufficienti garanzie che il codice familiare si riferisca realmente alla stessa famiglia in entrambe le occasioni);
(b) si hanno ingressi bilanciati da uno stesso numero di uscite (il caso di movimenti naturali di questo tipo non dovrebbe essere frequente).

In questo caso, il numero di ingressi e uscite bilanciate può essere visto come il numero di sospetti falsi negativi nella famiglia.

La Tab. 9 riporta il numero totale di situazioni di questo tipo riscontrate negli abbinamenti ottenuti con LINK. Come si vede, anche su questo fronte i dati sono confortanti, risultando il numero di sospetti falsi negativi a livello individuale piuttosto contenuto.

Mentre i falsi negativi a livello individuale sono da imputare alla esistenza, nelle variabili di confronto, di errori di rilevazione, trascrizione e registrazione, che vanno oltre i limiti di tolleranza previsti dalla procedura di abbinamento, gran parte dei falsi negativi a livello familiare sono invece da attribuire a errori nei codici identificativi della famiglia. Anche in questo caso, distinguere i falsi negativi non è agevole.

Al riguardo, è di un certo interesse richiamare i risultati di alcune verifiche, presentate in Giusti, Marliani e Torelli (1987), dalle quali risulta che i non abbinamenti di tipo familiare sono concentrati per la gran parte in un numero

Tab. 9: *Sospetti falsi negativi a livello individuale* [(a)]

Regione	85.I-II	86.I-II	85.I-86.I	85.II-86.II
Veneto	54	75	46	72
Lombardia	185	244	122	197

(a) I sospetti falsi negativi sono individuati dal numero di ingressi (records di tipo 10) e uscite (records di tipo 01) che si bilanciano, in famiglie nelle quali si abbinano due o più componenti, uno dei quali almeno è il capofamiglia o il coniuge

limitato di comuni e sembrano da imputare più al meccanismo di conduzione dell'indagine e agli errori nei codici di identificazione della famiglia che ad una maggiore o minore rigidità della procedura di abbinamento.

Quest'ultima notazione merita qualche ulteriore riflessione, in quanto potrebbe mettere in dubbio l'opportunità di considerare solo i confronti tra individui con lo stesso codice familiare. Tuttavia, alcune verifiche condotte utilizzando un bloccaggio meno restrittivo (il codice del comune) hanno consentito un recupero nella quota di abbinati decisamente modesto e ciò al prezzo di oneri computazionali non indifferenti (vedi Giusti Marliani e Torelli, 1987).

In definitiva, quindi, pur essendo l'errore nei codici familiari il maggior responsabile dei mancati abbinamenti, la scelta del bloccaggio per codice familiare continua a sembrare la più plausibile. Un suo abbandono non consentirebbe, infatti, un incremento di abbinati sufficiente a giustificare il maggior costo e, soprattutto, verrebbe meno un elemento che, almeno a giudicare dalle verifiche riportate nella sez. 5.1, pare costituire una notevole garanzia contro il rischio dei falsi positivi.

6. *Alcune note conclusive*

Riassumendo quanto emerso, si possono avanzare alcune considerazioni conclusive, distinguendo tra gli aspetti teorici e quelli operativi dell'abbinamento dei dati della RTFL.

Sul piano teorico, la procedura LINK appare certamente più soddisfacente di quelle euristiche, tipo Istat o MATCH. Il modo in cui essa assegna alle variabili di confronto una diversa importanza, ai fini della decisione di abbinamento, risponde ad un criterio oggettivo, che si mostra in grado di 'assorbire' eventuali anomalie presenti nei dati. Il fatto poi che i pesi che ne risultano, nonostante alcune limitazioni derivanti dalle ipotesi necessarie alla loro stima, siano conformi alle aspettative e intuitivamente convincenti, costituisce un ulteriore elemento in suo favore.

Per una valutazione più precisa occorrerebbe forse sperimentare un più articolato sistema di codifica dell'accordo/disaccordo nelle diverse variabili e cercare vie alternative per la stima di alcune componenti del peso. Si ha tuttavia la sensazione che i margini per ulteriori miglioramenti siano sostanzialmente ridotti. Occorre infatti tener presente che, da un certo punto di vista, i dati della RTFL possono considerarsi dati facili da abbinare, soprattutto per la presenza del codice familiare nei *records* individuali. E' pur vero che sono gli errori in questo campo a costituire il maggior ostacolo all'abbinamento, ma è altrettanto vero che il codice familiare costituisce un naturale fattore di bloccaggio, che si rivela assai efficace. Su questi dati, anche una procedura che usi pesi *ad hoc*, come MATCH, raggiunge risultati parecchio soddisfacenti.

Questa osservazione ci porta su un piano operativo ed è chiaro che, in questa ottica, non si può fare a meno di chiedersi se valga la pena di adottare una procedura più onerosa, quando questa garantisce guadagni sostanzial-

mente modesti. Un miglioramento del sistema di codifica ed un maggior controllo dello svolgimento dell'indagine assicurerebbe, infatti, margini di guadagno nell'abbinamento certamente superiori.

Ci sembra però che, anche sul piano operativo, le procedure probabilistiche mantengano una loro validità. Anche nel caso che per l'abbinamento dei dati della RTFL si preferisca continuare ad utilizzare una procedura deterministica, tipo MATCH, che risulta certamente meno onerosa, rimarrebbe sempre il problema della scelta dei pesi. A questo proposito, appare naturale giungere alla definizione di un sistema di pesi attraverso l'uso preliminare di procedure probabilistiche, tipo LINK, che sono in grado di assegnare alle variabili di confronto una differente importanza in relazione al potere discriminante che risulta dai dati.

Inoltre, al fine di limitare gli inconvenienti connessi all'uso di pesi costanti nel tempo, la procedura probabilistica potrebbe essere utilizzata correntemente, su un insieme ridotto dei dati dell'intera indagine RTFL, allo scopo di tenere sotto controllo la funzionalità del sistema dei pesi, che sarebbe sottoposto a revisione o a cadenze periodiche o non appena il risultato del controllo inducesse un qualche sospetto sulla presenza di eventuali anomalie nei dati.

8

UN'ANALISI DELLA QUALITA' DEI DATI BASATA SUL CONFRONTO DEI 'RECORDS' INDIVIDUALI IN PIU' OCCASIONI

Andrea Giommi

1. *Introduzione*

Il problema dell'individuazione e, quando possibile, della correzione degli errori non campionari assume connotazioni di particolare interesse in indagini ripetute con campioni a struttura longitudinale o a parziale rotazione delle unità, come la rilevazione trimestrale delle forze di lavoro (RTFL), che prevede la reintervista di parte delle unità campionarie in quattro occasioni di indagine nell'arco di quindici mesi (vedi il cap. 1).

Nelle indagini con disegno *cross-section*, tali errori si manifestano attraverso la mancanza di dati (assenza di risposta) e la presenza di risposte incoerenti all'interno di un questionario (ad esempio, donna in servizio di leva) o tra questionari relativi ad individui appartenenti a gruppi ben identificati, come il nucleo familiare (un esempio è dato dalla presenza in una famiglia di più di un capofamiglia). Nelle indagini del tipo della RTFL, grazie alla possibilità di costruire dati a struttura longitudinale, possono anche essere evidenziati errori che si presentano come incompatibilità tra informazioni date dallo stesso individuo in tempi diversi. Anche l'assenza di risposta, pur manifestandosi come nei dati *cross-section*, assume in quelli longitudinali connotazioni particolari e dà luogo a problemi del tutto peculiari anche per quanto riguarda le possibilità di trattamento.

Nelle indagini longitudinali si manifesta spesso un altro fenomeno distorsivo normalmente indicato con la terminologia inglese *rotation group bias*. Si tratta sostanzialmente di una distorsione rilevabile nelle stime di livello di alcune variabili di indagine, causata da forme di condizionamento che gli individui subirebbero in seguito alla prolungata permanenza nel campione. L'individuazione di tale fenomeno non è tuttavia legata alla possibilità di ricostruire storie individuali o familiari dei partecipanti alle varie occasioni di indagine.

Lo scopo principale di questo capitolo è di fornire alcuni elementi per una valutazione della qualità dei dati nella RTFL attraverso l'individuazione e l'esame della mancata risposta e delle incompatibilità presenti nei dati longitudinali, ricostruibili mediante una procedura di abbinamento esatto di *records* individuali. I dati utilizzati sono quelli delle regioni Veneto e Lom-

bardia rilevati nelle prime due occasioni di indagine (gennaio e aprile) del 1985 e del 1986. Le strutture longitudinali studiate sono i possibili abbinamenti di due occasioni di indagine a distanza di tre mesi, l'abbinamento di due occasioni a distanza di un anno ed infine il collegamento delle quattro occasioni di indagine. Tutti i dati longitudinali studiati in questo capitolo sono stati ottenuti con la procedura di abbinamento esatto LINK, discussa dettagliatamente nel cap. 7.

La sezione che segue è dedicata al problema dell'assenza di risposta. Nelle sezz. 3 e 4 si esamina il problema delle incompatibilità di risposta nei *records* longitudinali (individuali) relativi, rispettivamente, a due e a quattro occasioni di indagine. Il capitolo si conclude con alcune osservazioni sulla qualità dei dati quale risulta dalle analisi svolte e con un suggerimento per migliorarla, soprattutto nell'ottica di un più ampio sfruttamento delle informazioni longitudinali. Alla verifica dell'esistenza di un *rotation group bias* è dedicato il successivo cap. 9.

2. *Assenza di risposta*

Dell'assenza di risposta si considerano comunemente due tipologie: (i) la mancata risposta ad uno o più quesiti del questionario (*item nonresponse*); (ii) la mancata risposta a tutto il questionario, cioè la mancata partecipazione di un'unità campionaria all'indagine (*unit nonresponse*).

Per la RTFL, così come per ogni altra indagine longitudinale, è necessario introdurre una terza tipologia che possiamo denominare mancata risposta parziale o mancata risposta longitudinale (*panel nonresponse*). Questa si verifica quando un individuo appartenente al *panel* partecipa (risponde più o meno compiutamente al questionario) ad almeno una ma non a tutte le occasioni di indagine alle quali dovrebbe partecipare.

Il maggior problema legato alla mancata risposta è quello della scelta di un'opportuna metodologia statistica per il suo trattamento; normalmente si tratta di scegliere fra metodi di ponderazione e metodi di imputazione[1]. Nella pratica delle indagini questi ultimi sono comunemente utilizzati per i casi di *item nonresponse* ed i metodi di ponderazione per quelli di *unit nonresponse*. Per la non risposta longitudinale la scelta della metodologia non è ovvia (Kalton, 1986; Kalton e Lepkowski, 1985; Kalton e Miller, 1986). Rinviando l'approfondimento di questo aspetto ai lavori citati, ci limitiamo ad osservare che la non risposta longitudinale può essere vista come un insieme di *item nonresponses* nell'ambito del *record* longitudinale di un individuo, e, in questa ottica, sembrerebbe opportuno un intervento di imputazione. Ma può anche essere vista come *unit nonresponse* nell'ambito delle strutture *cross-section* che compongono quella longitudinale, ed in questo senso sarebbe

---

1 Talvolta si può procedere a recuperi 'sul campo', ma ciò è raramente praticabile in indagini di grosse dimensione come la RTFL, per motivi di costo e di tempo. Un'ulteriore possibilità è quella di ignorare la non risposta, ma è facile mostrare che, almeno per quanto riguarda le stime univariate di livello, questo corrisponde ad un particolare tipo di ponderazione e di imputazione (Kalton, 1983).

più appropriato un trattamento per ponderazione.

Anche per il caso di *item nonresponse* alla singola occasione di indagine, vi sono maggiori possibilità di trattamento rispetto a quelle che si hanno nelle indagini *cross-section*, potendo spesso disporre, per l'imputazione di un dato mancante, dello stesso dato rilevato in un'altra occasione di indagine.

Tornando alla non risposta longitudinale, si deve osservare che a monte del problema del trattamento, in ogni caso, devono essere affrontati altri due problemi non banali relativi alla sua rilevazione e all'individuazione delle cause che l'hanno originata.

Riguardo al primo punto occorre ricordare che la non risposta *panel* può essere rilevata sia per *records* individuali che familiari.

L'unità di rilevazione nell'indagine è rappresentata dalla famiglia. Le famiglie non trovate o che non collaborano sono sostituite con famiglie che dovrebbero avere stessa struttura e dimensione. Alla famiglia che sostituisce viene dato lo stesso codice (numero progressivo nell'area di rilevazione) di quella sostituita. Inoltre una famiglia sostituita dovrebbe essere contattata o cercata di nuovo alla successiva occasione cui deve partecipare.

Dal punto di vista della struttura longitudinale dei dati familiari, la sostituzione equivale ad una mancata risposta. Nella fase della sua ricostruzione mediante il programma di abbinamento - che procede confrontando *records* individuali all'interno di famiglie con identico codice nelle due occasioni che si collegano - la sostituzione dà luogo a mancati abbinamenti individuali o, raramente, a falsi positivi. I mancati abbinamenti dovrebbero costituire la norma in quanto, anche se una famiglia è sostituita con un'altra identica relativamente ad una serie di caratteri propri dell'aggregato (quali, ad esempio, numerosità complessiva, numero di figli in età scolare, di anziani ecc.), è molto improbabile che, scendendo a livello individuale possano trovarsi ancora, nei caratteri di confronto, identità in numero sufficiente da consentire l'erroneo abbinamento (falso positivo).

Tuttavia, se da una parte possiamo avere la ragionevole certezza che la quasi totalità delle sostituzioni daranno luogo a mancati abbinamenti individuali, non possiamo certo pensare che questi ultimi abbiano origine soltanto, o anche in massima parte, dalle sostituzioni familiari. Altre possibili cause sono infatti:

- la mancanza effettiva di un membro in una famiglia presente alle occasioni che si collegano;
- un falso negativo familiare (mancato abbinamento per errore nel codice famiglia);
- un falso negativo individuale (mancato abbinamento per errori nei caratteri di confronto).

Di queste cause, i falsi negativi, sono in qualche modo riconducibili alla qualità dei dati e/o alle caratteristiche del programma di abbinamento, mentre sostituzioni familiari e mancanza effettiva di singoli sono legate al comportamento dei partecipanti all'indagine. I dati a nostra disposizione non permettono di attribuire con certezza un caso di non abbinamento ad una delle cause alternative elencate. Pertanto non è possibile parlare di non risposta individuale in senso proprio, ma soltanto dare a questa espressione il signi-

ficato assai più ampio di mancato abbinamento.

L'Istat ha recentemente annunciato un'innovazione rappresentata dall'introduzione di un apposito codice di avvenuta sostituzione familiare: un'iniziativa senz'altro opportuna che tuttavia non necessariamente elimina i problemi di identificazione della non risposta (vedi la. sez. 2.2).

Questi problemi di identificazione implicano a loro volta l'impossibilità di associare ai partecipanti all'indagine le tipologie di non risposta longitudinale normalmente individuabili nel collegamento di più occasioni di indagine. Chiariremo tra breve questa affermazione dopo aver introdotto e illustrato tali tipologie, considerando inizialmente l'abbinamento di due occasioni di indagine.

2.1. *Non riposta longitudinale su due occasioni*

Ciascuna tipologia, che chiameremo sinteticamente *pattern*, è rappresentabile mediante una sequenza di 1 (risposta/abbinamento) e di 0 (non risposta/non abbinamento). Collegando due occasioni, vi è un unico *pattern* che indica abbinamento, 11, e due che indicano non abbinamento, 10 e 01. Per chiarire il motivo per cui non è possibile far corrispondere singoli individui ai *patterns* suindicati, si consideri un ipotetico collegamento di due occasioni a ciascuna delle quali abbiano partecipato, rispondendo sempre, gli stessi 1.000 individui. In assenza di errori si dovrebbero rilevare 1.000 *patterns* uguali a 11. Supponiamo invece che, per errori nelle variabili identificative, si abbinino solo 700 soggetti. Con il programma di abbinamento da noi utilizzato questa situazione darà luogo a 700 *records* con *pattern* 11, corrispondenti ad altrettanti individui e ben 600 *records* per i 300 individui restanti: 300 con *pattern* 10 e ancora 300 con *pattern* 01, per un totale di 1.300 *records* contro i 1.000 partecipanti effettivi al complesso delle due occasioni.

I dati della Tab. 1 soffrono di questo inconveniente. I totali risultanti dalla somma delle frequenze assolute dei tre *patterns*, per i quattro collegamenti effettuati, si riferiscono soltanto a casi originati dal programma di abbinamento. Il numero di individui presenti ad ogni occasione di indagine può comunque essere ricavato. I presenti alla prima occasione risultano dalla somma dei casi associati ai *patterns* con 1 in prima posizione e i presenti alla seconda, dalla somma dei casi associati ai *patterns* con 1 in seconda posizione. Il tutto a meno di falsi positivi, ovviamente non individuabili, tra i *patterns* 11. Considerando i dati del Veneto per il 1985.I-II, risultano presenti 7.560 individui alla I occasione, 7.516 alla seconda e 6.395 ad ambedue. Non si può invece determinare con esattezza il numero di individui effettivamente mancanti ad ogni occasione di indagine e, di conseguenza, neanche quello dei presenti ad una sola delle occasioni.

Nel commentare i dati delle tabelle di questa sezione faremo comunque uso di termini come 'caduta di partecipazione' o simili, che richiamano con immediatezza il concetto di non risposta in senso stretto, ma che, per quanto abbiamo fin qui osservato, dovranno essere presi in un'accezione ben più ampia.

Per la Lombardia si osservano cadute del 21,6 e 17% tra le prime due occasioni rispettivamente dell'85 e dell'86, e del 32,5 e del 32,2% per le due occasioni collegabili ad un anno di distanza; per il Veneto del 15,4 e 15,5% a tre mesi e del 22,2 e 21,2% ad un anno. Appare del tutto normale che in entrambe le regioni l'aumento delle perdite individuali sia altamente correlato con il tempo che intercorre con le occasioni di indagine. Ma, mentre per il Veneto l'incremento delle perdite non supera il 43%, per la Lombardia si arriva quasi al 100%. Una differenza tanto marcata non può ragionevolmente essere attribuita ad un comportamento differenziale dei partecipanti all'indagine, né ad una maggiore quota di errori nella fase della raccolta dei dati. L'unica possibilità sembra essere una riestrazione del campione in uno o più comuni, una procedura che ci risulta venga attuata tra la seconda e la terza occasione di indagine nei comuni che, per variazioni di dimensioni demografiche e/o di attività prevalente, devono cambiare strato di appartenenza. Non disponendo del codice di attività prevalente, siamo indotti a pensare che la differenza tra le percentuali delle due regioni dipenda interamente o quasi dalla mancata enucleazione dei comuni passati di strato, prima del collegamento.

Tab. 1: *Distribuzione delle tipologie di risposta nei possibili collegamenti di due occasioni di indagine*

Patterns	Veneto				Lombardia		
	V.A.	%	(a) %		V.A.	%	(a) %
				85.I-II			
11	6.397	73,7	84,6		26.070	64,5	78,4
10	1.165	13,4	15,4		7.164	17,8	21,6
01	1.121	12,9			7.219	17,7	
				86.I-II			
11	6.301	73,4	84,5		27.741	71,4	83,0
10	1.156	13,5	15,5		5.685	14,6	17,0
01	1.126	13,1			5.408	14,0	
				85.I-86.I			
11	4.954	65,4	77,8		17.147	48,4	67,5
10	1.417	18,7	22,2		8.240	23,3	32,5
01	1.207	15,9			10.035	28,3	
				85.II-86.II			
11	4.916	65,9	78,8		17.525	49,5	67,8
10	1.319	17,7	21,2		8.334	23,5	32,2
01	1.226	16,4			9.543	27,0	

(a) Presenti alla prima occasione.

Dal punto di vista della qualità dei dati, quindi, nessun problema sorge se tali riestrazioni sono programmate e risultano dalla documentazione in possesso dell'Istat.

Oltre che dalla distanza tra le occasioni di indagine la quota di risposta può anche dipendere dal tempo di permanenza nell'indagine. Nel leggere la Tab. 1 occorre tenere presente che, in ogni occasione di indagine, si seguono longitudinalmente individui appartenenti a due delle quattro sezioni in cui si suddivide il campione. Nel collegamento a distanza di tre mesi, il campione della prima occasione è formato dalle sezioni dei partecipanti per la prima e per la terza volta; quello della seconda occasione, dalle sezioni dei partecipanti per la seconda e per la quarta volta. Se il collegamento è a distanza di un anno, nella prima occasione posta a confronto troviamo le sezioni dei partecipanti per la prima e per la seconda volta, e nella seconda occasione le sezioni dei partecipanti per la terza e per la quarta volta. Di conseguenza, le percentuali di caduta su due occasioni della Tab. 1 mediano il comportamento di due gruppi di individui che, essendo entrati nell'indagine in tempi diversi possono anche esprimere comportamenti diversi. Dai dati a nostra disposizione (qui non riportati) si sono rilevate differenze modeste tra le percentuali rilevate sulle diverse sezioni. Per le prime due occasioni dell'85 nel Veneto, ad esempio, la caduta tra le prime due partecipazioni è risultata del 15,5% contro il 14,8% tra le ultime due. Situazione analoga si ha per le stesse occasioni di indagine della Lombardia con il 22,3% di caduta tra le prime due partecipazioni e il 20,8% tra le ultime due. In definitiva, a parità di distanza tra occasioni collegate si osserva sistematicamente una caduta un po' più elevata tra le prime due partecipazioni che non tra le ultime due.

2.2. *Assenza di risposta e codici di famiglia*

Per coppie di occasioni di indagine l'assenza di risposta è stata analizzata anche a livello familiare.

Il programma di abbinamento pone a confronto *records* individuali soltanto a parità di codice familiare. Un errore in questo codice, rendendo impossibile un confronto corretto di *records*, provoca nella maggior parte dei casi falsi negativi, corrispondenti a *patterns* 10 o 01. In ogni altro caso e ciò dovrebbe accadere meno frequentemente, genera falsi positivi (*pattern* 11), abbinando *records* simili ma di individui diversi (vedi il cap. 7).

L'analisi dei codici familiari ci ha consentito di classificare i casi di mancato abbinamento (corrispondenti all'insieme dei *patterns* 10 e 01) in tre categorie (Tab. 2):

(a) mancati abbinamenti in famiglie (leggi codici familiari) con almeno un *pattern* 11;
(b) mancati abbinamenti in famiglie senza alcun *pattern* 11;
(c) mancati abbinamenti per impossibilità di effettuare l'abbinamento a causa della presenza del codice familiare in una sola delle due occasioni di indagine.

Anche in questo caso la dizione 'mancati abbinamenti' esprime un effetto

del quale non si è in grado di discriminare le cause. La classificazione introdotta permette di affinare l'analisi dell'assenza di risposta, ma non ancora di attribuire univocamente agli individui i *patterns* osservati. Ad esempio, la classe (b) sembrerebbe, ad un primo esame, originata da sostituzioni familiari. Se così fosse tutti i *patterns* 10 e 01 che vi cadono corrisponderebbero a individui. Purtroppo la realtà è più complessa. In fase di rilevazione, più difficilmente in fase di registrazione dei dati, si può verificare uno 'slittamento' dei codici familiari: ad un gruppo di famiglie può essere attribuito anziché il giusto codice, lo stesso aumentato o diminuito di una costante. Per effetto di questo errore, che può avere origini svariate ma sostanzialmente riconducibili alla complessità della codifica in rapporto al grado di addestramento dei rilevatori (Masselli, 1988), al momento del collegamento tra due occasioni si pongono a confronto codici familiari uguali cui corrispondono famiglie diverse. L'effetto più probabile è che per questi codici non si verifichi alcun abbinamento, proprio come avviene quando c'è una sostituzione; ma in questo caso i *patterns* 10 e 01 non corrispondono più ad uscite ed entrate, bensì tutti, o quasi, a falsi negativi familiari[2].

Che errori come quello appena descritto si verifichino, e non pochi, è fuori di dubbio: la classe (c) dei *patterns* non abbinati, per la presenza del codice familiare in una sola delle due occasioni che si collegano, dovrebbe derivare anche da questi. La parte iniziale o finale di una lista di codifiche che ha subito uno slittamento perde la corrispondenza numerica con le altre liste ad essa collegabili. Un'altra possibilità è data dall'ingresso e dall'uscita di codici per errore del rilevatore. La Tab. 2 mostra che quantitativamente questo fenomeno è assai rilevante. Considerando tutti gli abbinamenti del Veneto e quelli della Lombardia a tre mesi di distanza e complessivamente i *patterns* 10 e 01 della classe (c), cioè sommando in tabella le percentuali relative ai due *patterns*, si osservano quote di non abbinati che vanno dal 7,9 al 14,5%. Nei due abbinamenti ad un anno della Lombardia si arriva al 20,5 e al 21%. Ma è molto probabile che in questi ultimi casi il fenomeno sia stato amplificato dalla citata riestrazione del campione in alcuni comuni, nel senso che questa operazione potrebbe aver comportato mutamenti dimensionali dei campioni riestratti con conseguente perdita o aggiunta di codici familiari.

Anche i *patterns* 10 e 01 della classe (a) non hanno origine univoca. Il fatto che provengano dal confronto di famiglie nelle quali si verifica almeno un abbinamento non esclude la possibilità che siano dei falsi negativi, derivanti da errori nelle variabili di collegamento. E' tuttavia in questa classe che troviamo le effettive entrate ed uscite individuali in e da famiglie che permangono nell'indagine. Un fenomeno la cui ridotta rilevanza quantitativa

---

2 Questi particolari falsi negativi sarebbero eliminati da un programma di abbinamento che operasse confronti superando il vincolo del codice familiare, ad esempio, confrontando tra loro tutti gli individui all'interno di ogni sezione. In questo caso tuttavia, data la dimensione dell'indagine e la frequenza con cui si effettua questa operazione, il tempo di elaborazione per realizzare l'abbinamento aumenterebbe enormemente, con costi inaccettabili in rapporto ai possibili benefici rappresentati dal recupero di un numero relativamente modesto di abbinati.

Tab. 2: *Distribuzione delle tipologie di risposta/non risposta a livello di codici familiari*

Patterns		85.I-86.I		85.II-86.II		85.I-II		86.I-II	
		V.A.	%	V.A.	%	V.A.	%	V.A.	%
					Veneto				
11		4.954	65,4	4.916	65,8	6.395	73,8	6.301	73,4
10	(a)	289	3,8	341	4,6	226	2,6	220	2,6
	(b)	687	9,1	654	8,8	473	5,5	561	6,5
	(c)	441	5,8	324	4,4	446	5,2	375	4,4
01	(a)	232	3,1	251	3,4	245	2,8	235	2,7
	(b)	667	8,8	643	8,6	449	5,2	590	6,9
	(c)	308	4,1	332	4,5	427	4,9	301	3,5
					Lombardia				
11		17.147	48,4	17.525	49,5	26.070	64,4	27.741	71,4
10	(a)	1.610	4,5	1.797	5,1	955	2,4	932	2,4
	(b)	4.114	11,6	3.578	10,1	3.287	8,1	2.371	6,1
	(c)	2.516	7,1	2.959	8,4	2.922	7,2	2.382	6,1
01	(a)	1.270	3,6	1.454	4,1	1.033	2,6	990	2,6
	(b)	4.031	11,4	3.628	10,2	3.214	8,0	2.377	6,1
	(c)	4.734	13,4	4.461	12,6	2.972	7,3	2.041	5,3

(a) mancati abbinamenti in famiglie presenti alle due occasioni

(b) mancati abbinamenti in famiglie prive di abbinamenti ma presenti alle due occasioni

(c) mancati abbinamenti per presenza del codice familiare in una sola occasione

è confermata dalla frequenza percentuale dei *patterns* della classe (a), che non arrivano al 3%, salvo per le due situazioni particolari della Lombardia.

In conclusione, la classificazione dei *patterns* di risposta della Tab. 2 non ci permette di discriminare le entrate e le uscite effettive, che nell'indagine sono rappresentate per la massima parte dalle sostituzioni familiari, da quelle apparenti, che derivano dai falsi negativi. Ma ci consente di affermare che, più che dalla maggiore o minore rigidità della procedura di abbinamento rispetto agli errori nelle variabili identificative, i falsi negativi sono causati da errori nei codici di famiglia o di livello superiore. E' un fatto che invita a riflettere sull'eventualità che l'introduzione del codice di avvenuta sostituzione familiare lasci aperto più di un problema, almeno quando si voglia disporre dell'informazione longitudinale dell'indagine.

2.3. *Non risposta su quattro occasioni*

Passando al collegamento di quattro occasioni di indagine, occorre fare una premessa riguardo alle diverse possibilità che si offrono per effettuare questa operazione.

L'abbinamento scaturisce dal confronto di coppie di occasioni. Quattro occasioni possono essere confrontate a due a due in diversi modi.

Ad esempio, la prima occasione può essere presa come punto di riferimento per tutti i confronti, ed è questo il criterio adottato in questo capitolo, oppure un'altra occasione può essere presa come base 'fissa' dei confronti. Mutuando per l'analogia termini propri della teoria dei numeri indici, in contrapposizione a confronti con base fissa, si può pensare ad una serie di confronti a 'base mobile', cioè effettuati tra ogni occasione e la precedente. Questa premessa è giustificata dal fatto che i risultati numerici dell'abbinamento dipendono in una certa misura dal criterio che si sceglie.

Collegando quattro occasioni, sono teoricamente rilevabili 16 *patterns* che si riducono a 15 escludendo il *pattern* 0000, corrispondente alla totale assenza di informazione e, di conseguenza, ad una situazione non rilevabile.

L'elenco dei *patterns*, unitamente alla frequenza con cui li abbiamo rilevati nelle quattro occasioni disponibili per il Veneto e la Lombardia, è riportato nella parte superiore della Tab. 3.

I *patterns* possono essere raggruppati secondo vari criteri. Little e David (1982) considerano accanto alla situazione di completa risposta, le situazioni di 'logorio' (*attrition*), 'rientro' (*reentry*) e 'ingresso ritardato' (*later entry*). Il logorio si verifica quando un individuo che ha partecipato alla prima occasione dell'indagine l'abbandona definitivamente in una qualunque delle successive occasioni cui sarebbe chiamato a partecipare. Su quattro occasioni ciò è espresso dai *patterns* 1110, 1100, 1000. Il rientro è la situazione in cui, dopo la partecipazione iniziale, un individuo abbandona l'indagine per una o più occasioni per tornare poi a parteciparvi per le rimanenti. I *patterns* possibili sono 1101, 1011, 1001. Infine, l'ingresso ritardato si verifica quando l'individuo partecipa all'indagine iniziando da un'occasione successiva alla prima: 0111, 0011, 0001. Restano fuori da questa classificazione situazioni che potremmo definire 'miste' in quanto risultanti da combinazioni delle precedenti. Ad esempio, 1010 può essere vista come una situazione di rientro seguita da logorio; 0101 come ingresso ritardato più rientro, e così via.

L'interesse della classificazione sta nel suo obiettivo di individuare situazioni tipiche, cui corrispondono gruppi di soggetti relativamente omogenei in rapporto al modo di partecipare all'indagine e, verosimilmente, anche ai caratteri sotto inchiesta. La sua corretta adozione è tuttavia legata alla possibilità di associare ad ogni partecipante all'indagine un unico *pattern* di risposta. Poiché con i dati della RTFL questo non è possibile, si è ritenuta più opportuna una classificazione basata sul numero di occasioni per le quali è disponibile l'informazione.

I 15 *patterns* da noi rilevati non sono espressione del comportamento dei partecipanti all'indagine, ma soltanto del tipo di informazione disponibile per i *records* 'prodotti' dal programma di abbinamento. Il *pattern* 1100 rappresenta un *record* individuale con informazioni soltanto per le prime due occasioni. Purtroppo, come abbiamo già avuto modo di osservare, l'individuo cui è associato questo *record* può essere lo stesso cui è associato un altro *record* contraddistinto dal *pattern* 0011 o anche 0010 o 0001. In generale tutti i *patterns* che hanno i simboli 1 in posizione diversa possono teorica-

Tab. 3: *Distribuzione delle tipologie di risposta non risposta individuale per quattro occasioni di indagine*

Patterns		Veneto		Lombardia	
		V.A.	%	V.A.	%
		A - Tutti i casi			
Risposta totale	1111	2.223	49,5	7.143	29,5
Risposta a tre occasioni	1110	111	2,5	501	2,1
	1101	173	3,9	301	1,2
	1011	168	3,7	653	2,7
	0111	62	1,4	407	1,7
Risposta a due occasioni	1100	306	6,8	1.753	7;2
	1010	11	0,2	141	0,6
	1001	7	0,2	40	0,2
	0110	10	0,2	185	0,8
	0101	-	-	157	0,6
	0011	47	1,0	1.150	4,7
Risposta a una occasione	1000	186	4,1	2.177	9,0
	0100	283	6,3	2.476	10,2
	0010	451	10,0	3.471	14,3
	0001	454	10,1	3.687	15,2
Totale		*4.492*	*100,0*	*24.242*	*100,0*
		B - Presenti alla I occasione			
Risposta totale	1111	2.223	69,8	7.143	56,2
Risposta a tre occasioni	1110	111	3,5	501	3,9
	1101	173	5,4	301	2,4
	1011	168	5,3	653	5,1
Risposta a due occasioni	1100	306	9,6	1.753	13,8
	1010	11	0,3	141	1,1
	1001	7	0,2	40	0,3
Risposta a una occasione	1000	186	5,8	2.177	17,1
Totale		*3.185*	*100,0*	*12.709*	*100,0*

mente far capo allo stesso soggetto.

Come abbiamo fatto in precedenza per la Tab. 1, possiamo associare univocamente il *pattern* all'individuo solo riducendo opportunamente l'insieme dei *records* prodotti dal programma di collegamento. Per il collegamento di quattro occasioni è possibile ridurre l'insieme dei *patterns* a quelli associati ai presenti ad una qualsiasi di queste. Nella Tab. 3, si è scelto di rappresentare i presenti alla prima occasione di indagine. In questo insieme è ristabilita la corrispondenza univoca tra *pattern* e individui, ma il loro totale (3.185 per il Veneto e 12.709 per la Lombardia) è inferiore al numero dei partecipanti effettivi all'indagine (nella sezione collegabile su quattro occasioni). E' infatti altamente improbabile che nell'arco di tempo di 15 mesi la parte di campione collegabile per quattro occasioni non abbia fatto registrare 'ingressi ritardati'.

Un potenziale indicatore della qualità dell'indagine è rappresentato dalla frequenza percentuale degli individui con *pattern* 1111 su tutti i partecipanti alla prima occasione o, alternativamente, dal suo complemento a 100, vale a dire dalla percentuale di perdita complessiva di individui nell'arco di quattro occasioni (sempre ovviamente sull'insieme dei partecipanti alla prima occasione e ricordando il significato ampio dei termini caduta, perdita, ecc.).

E' difficile esprimere una valutazione sul 30% circa di perdita di individui nel *follow-up* su quattro occasioni nel Veneto (a fronte del 69,8% di completa risposta, Tab. 3/B); né è possibile confrontare questa percentuale con quelle, abbastanza inferiori, rilevate negli Stati Uniti per indagini analoghe (Kalton, 1986; Kalton e Lepkowski, 1985) a causa delle differenze nella cadenza delle interviste e nelle regole di *follow-up* familiare.

Il confronto più significativo è ancora quello tra dati della Lombardia e del Veneto. Alla luce di quanto abbiamo già osservato per il collegamento di due occasioni a tre mesi non sorprende la differenza di caduta complessiva tra le due regioni (oltre il 43% per la Lombardia contro il 30% del Veneto) che risulta giustificata dall'ipotesi di riestrazione del campione in alcuni comuni della Lombardia.

Per quanto riguarda il comportamento dei rispondenti in rapporto al tempo di permanenza dell'indagine, nel collegamento di quattro occasioni si seguono nel tempo individui appartenenti ad una sola sezione di indagine. I criteri adottati per i confronti tra occasioni di indagine (vedi la sez. 4) ci permettono di confrontare soltanto le percentuali di caduta tra le prime due partecipazioni all'indagine, rilevate nel collegamento a due occasioni, con quella rilevabile nel collegamento a quattro, che si ottiene rapportando la frequenza dei *patterns* che iniziano con 10 a quella di tutti i *patterns*. I valori che si osservano sono 11,7% per il Veneto e 23,6% per la Lombardia. Le differenze con i valori visti in precedenza, per il collegamento a tre mesi, sono dovute al fatto che, prima di procedere al collegamento di quattro occasioni, sono stati enucleati i *records* dei comuni che sarebbero usciti dall'indagine per la rotazione annuale. Questa operazione non viene invece eseguita prima dell'abbinamento delle prime due occasioni di ogni anno. Si deve osservare che mentre per la Lombardia, dove le cadute sono mediamente più alte, l'uscita di questi comuni non provoca in pratica alcuna differenza, per il

Veneto si riscontra un miglioramento, cioè una diminuzione della caduta di oltre 4 punti percentuali. Ciò indicherebbe che i comuni che, nell'anno da noi considerato, stavano terminando il loro ciclo di indagine presentavano percentuali di caduta notevolmente più alte degli altri, stimabili intorno al 20%.

2.4. *Fattori di distorsione introdotti dall'assenza di risposta*

Come abbiamo già detto, non affrontiamo in questo studio il problema del trattamento della non risposta longitudinale. Vogliamo però metterne in evidenza l'importanza, soprattutto nell'ottica di utilizzare nel modo più completo e corretto i dati longitudinali ricostruibili nella RTFL.

Nella Tab. 4 abbiamo riportato la distribuzione di una serie di caratteri (sesso, stato civile, grado di istruzione, condizione professionale, età) rilevati alla prima occasione di indagine separatamente sull'insieme dei presenti in quattro occasioni (*pattern* 1111) e sull'insieme dei presenti alla prima occasione con gli altri possibili *patterns* (1110, 1100, 1000, 1101, 1011, 1010). Si può osservare che l'insieme dei presenti a tutte e quattro le occasioni ha caratteristiche mediamente diverse da quello dei mancanti ad almeno una occasione. In sintesi, gli uomini sono in proporzione maggiore delle donne nel gruppo dei sempre presenti; lo sono ancora i coniugati rispetto alle altre categorie dello stato civile; quelli con licenza elementare rispetto agli altri. Come condizione professionale prevalgono, nel gruppo dei sempre presenti, le casalinghe, gli studenti, i ritirati dal lavoro. Considerando l'età, constatiamo la prevalenza, in questo gruppo, di individui nella classe 14-20 anni (gli studenti) e 36-60 anni (coloro che, mediamente, hanno situazione lavorativa stabilizzata); oltre i 60 anni la proporzione è maggiore nel gruppo che presenta uscite, presumibilmente per il maggior peso della mortalità.

Anche se le differenze sono modeste soprattutto per la Lombardia - ma per questa regione ciò dovrebbe essere conseguente alla riestrazione del campione - i dati fanno comunque riflettere sulle possibili distorsioni che si introducono limitando le analisi ai rispondenti che danno luogo ad una struttura longitudinale completa, soprattutto se questa è fortemente 'selezionata', come è appunto quella risultante dal metodo di abbinamento utilizzato dall'Istat per la costruzione delle matrici di transizione (vedi il cap. 7).

2.5. *Non risposta a singoli quesiti del questionario*

Come abbiamo osservato in precedenza, i programmi di pulitura dei dati utilizzati dall'Istat per eliminare incongruenze ad ogni singola occasione, operano anche delle imputazioni per non risposta ai vari *items* del questionario. Il criterio è essenzialmente di tipo logico: un dato mancante viene imputato con un valore logicamente deducibile dalle risposte disponibili nello stesso questionario. Se ciò non è possibile l'imputazione avviene secondo

Tab. 4: *Distribuzione di alcuni caratteri alla I occasione di indagine per i presenti a 4 occasioni e i presenti alla I occasione ma non a tutte le altre*

Caratteri	Veneto		Lombardia	
	Sempre presenti	Non sempre presenti	Sempre presenti	Non sempre presenti
Sesso				
Maschi	49,44	46,36	48,68	47,25
Femminile	50,56	53,64	51,32	52,75
Stato civile				
Celibe/nubile	29,19	30,94	28,92	28,83
Coniugato/a	61,99	58,11	60,96	60,16
Vedovo	7,85	9,43	8,85	10,03
Separato ecc.	0,97	1,51	1,27	0,97
Titolo di studio				
Analfabeta	0,40	0,52	0,21	0,25
Nessun titolo	19,48	24,01	15,26	17,05
Licenza elementare	40,94	33,58	38,58	40,84
Scuola media inferiore	27,17	26,30	28,53	27,00
Scuola media superiore	9,67	12,47	14,43	12,81
Laurea	2,34	3,12	2,98	2,05
Condizione professionale				
Occupato	44,62	44,78	46,64	46,36
In cerca di nuova occup.	1,72	2,64	1,52	1,21
In cerca di prima occup.	2,10	1,26	1,98	2,09
In servizio di leva	0,65	0,75	0,77	0,47
Casalinga	22,90	21,13	19,32	19,79
Studente	8,92	7,67	9,89	7,95
Inabile	1,72	2,52	1,09	1,72
Ritirato	14,09	12,70	17,42	18,50
Altro	3,12	6,42	1,37	1,92
Età				
fino a 13	16,33	17,36	15,08	16,56
14 - 20	11,56	9,36	11,96	10,28
21 -25	7,69	9,04	7,42	7,20
26 - 30	5,80	8,94	6,06	7,35
31 - 35	7,24	8,32	6,83	7,17
36 - 40	7,47	7,17	7,55	6,99
41 - 50	13,54	11,23	15,22	14,27
51 - 60	13,72	9,67	12,99	12,27
più di 60	16,64	18,92	16,90	17,61

criteri prestabiliti[3]. Per la costruzione di dati longitudinali, l'imputazione potrebbe essere effettuata in maniera diversa, utilizzando cioè le informazioni che provengono anche dalla stessa struttura longitudinale dei dati. In altri termini si tratterebbe, prima, di effettuare l'abbinamento sui dati grezzi (non ancora sottoposti al piano di compatibilità) e successivamente di imputare e controllare le incongruenze tra ed interne alle occasioni per eliminare quelle logicamente correggibili. Inoltre sarebbe interessante verificare l'effetto di metodi di imputazione diversi operanti sui dati *cross-section* (*hot-deck*, imputazioni multiple, ecc.). Queste verifiche potrebbero essere limitate, per la verità, alla sola condizione professionale, l'unica variabile con il titolo di studio[4] che nei dati grezzi raggiunge una quota di non risposta tra l'1% e il 2%. Non è impossibile, a nostro avviso, che, cambiando il metodo di imputazione di questo 1%, si osservino variazioni rilevanti, ad esempio, sulla stima del tasso di disoccupazione. Per tutte le altre variabili l'effetto dovrebbe essere irrilevante.

Il discorso cambia se si considera l'effetto complessivo del piano di compatibilità. Vedremo nella sezione che segue che tale programma può creare delle incompatibilità tra occasioni successive, riducendo la quota di abbinati privi di incompatibilità nelle variabili di abbinamento.

3. *Incompatibilità fra due occasioni di indagine*

Consideriamo ora il problema delle incompatibilità che emergono dal confronto di due occasioni di indagine. Nella sezione successiva lo stesso problema viene brevemente affrontato in rapporto all'abbinamento sull'arco di quattro occasioni. Come già detto, tutti gli abbinamenti sono stati realizzati mediante il programma LINK, e, pertanto, considereremo incompatibili situazioni trattate come tali dallo stesso programma.

Queste situazioni, che conviene riassumere brevemente, sono:

(a) una qualsiasi differenza nel mese o nell'anno di nascita;
(b) una qualsiasi differenza nel sesso;
(c) una qualsiasi differenza nella relazione col capofamiglia;
(d) una variazione nel grado di istruzione, diversa dal conseguimento della laurea, da un'occasione alla successiva per i collegamenti a distanza di tre mesi. E' invece ammissibile l'avanzamento di un grado negli abbinamenti ad un anno di distanza;
(e) lo stato di celibe/nubile per tutti coloro che in un'occasione precedente si trovassero in uno stato diverso;
(f) l'essere in cerca di prima occupazione avendo dichiarato in un'occasione

---

3 Nei dati sottoposti al piano di compatibilità abbiamo trovato piccole quote di non risposta. Non escludiamo che ciò possa essere dovuto ad una imperfetta enucleazione da parte nostra di coloro che non dovevano rispondere ai quesiti presi di volta in volta in esame. In ogni caso tali quote sono, per ogni domanda, inferiori allo 0,2% delle risposte e quindi senz'altro trascurabili.

4 Fa eccezione, sia per i dati grezzi sia per quelli sottoposti al piano di compatibilità , la domanda su 'chi ha risposto ai quesiti', che presenta comunque quote di mancata risposta superiori al 50%.

precedente di essere occupato (condizione professionale, quesito 10.1); (g) la dichiarazione di non aver mai lavorato in passato avendo dichiarato in occasione precedente il contrario (precedenti lavorativi, quesito 13.1).

Le Tabb. 5-9 riportano le incompatibilità rilevabili nelle variabili sopra citate dopo l'abbinamento di coppie di occasioni di indagine. Le Tabb. 5,6 e 9 riguardano i possibili abbinamenti a 3 mesi effettuabili con i dati del Veneto e della Lombardia a nostra disposizione e cioè: 85.I-II, 86.I-II ed ancora 86.I-II con dati grezzi. Le Tabb. 7 e 8 contengono dati relativi ad abbinamenti di *records* a distanza di un anno, rispettivamente 85.I-86.I e 85.II-86.II.

Non è stato possibile in nessuno degli abbinamenti considerati approfondire l'esame delle incompatibilità analizzandole in funzione del 'tipo di rispondente', ovverosia del componente della famiglia che ha risposto nelle due occasioni. Questa variabile, infatti, presenta quote di non risposta intorno al 50%. E' difficile da imputare, ed anche facendolo, il tipo di imputazione risulta avere un peso eccessivo sugli effetti osservabili (cfr. A. Giommi, A. Giusti e N. Torelli, 1987).

Le indicazioni che emergono dall'analisi delle tabelle sono sostanzialmente conformi a quelle già emerse per i dati del Veneto in Giommi, Giusti e Torelli (1987), in cui gli abbinamenti delle occasioni di indagine erano stati effettuati con un diverso programma di abbinamento. E ciò, sia perché il miglioramento del metodo di abbinamento ha un effetto assai relativo sulle quote di incompatibilità osservabili, sia perché i dati della Lombardia presentano distribuzioni del tutto analoghe a quelli del Veneto.

Dal punto di vista della qualità dei dati dobbiamo ancora segnalare che le percentuali di incompatibilità su anno e mese di nascita, sesso, e stato civile sono piuttosto contenute: mediamente si aggirano intorno all'1%, risultando un po' superiori negli abbinamenti a un anno di distanza, soprattutto per la Lombardia. Anche la relazione con il capofamiglia rientra tra queste variabili, con l'eccezione della I occasione dell'85. La quota di incongruenze eccezionalmente elevata che avevamo già rilevata a suo tempo per il Veneto nell'abbinamento delle prime due occasioni di indagine dell'85 (22,7%), si ripresenta anche in Lombardia per lo stesso periodo (19%) e in ambedue le regioni negli abbinamenti ad un anno che partono dalla prima occasione dell'85. Un'evidenza che non siamo tuttora in grado di spiegare se non ipotizzando che, nel gennaio 85, non fosse operante o messo a punto il programma di correzione dei dati a livello di codice familiare, previsto dal piano di compatibilità.

A parte la relazione con il capofamiglia, le uniche due variabili che sistematicamente presentano quote di incompatibilità di un certo rilievo sono il titolo di studio e i precedenti lavorativi.

Per il titolo di studio la percentuale di incompatibilità è sorprendentemente alta, sempre superiore al 7% negli abbinamenti a due mesi, e al 5% in quelli ad un anno. Poiché non crediamo che per questa variabile vi sia minore attenzione rispetto alle altre da parte del rilevatore, dobbiamo pensare che per il titolo di studio vi possa essere una maggiore propensione a fornire indicazioni errate sul proprio o sul conto altrui, cui si sommano problemi di

Tab. 5: *Incompatibilità rilevabili nel collegamento di 'records' individuali su due occasioni di indagine: 85.I-II*

Caratteri	Veneto			Lombardia		
	Totale risposte	Incompatibilità V.A.	%	Totale risposte	Incompatibilità V.A.	%
Anno di nascita	6.395	40	0,6	26.070	251	1,0
Mese di nascita	6.395	97	1,5	26.070	295	1,1
Sesso	6.395	35	0,5	26.070	271	1,0
Relaz. col capofamiglia	6.395	1.453	22,7	26.070	4.949	19,0
Titolo di studio	6.395	489	7,6	26.070	1.958	7,5
Stato civile	5.291	29	0,5	21.875	107	0,5
Condizione professionale	5.291	6	0,1	21.875	31	0,1
Precedenti lavorativi	2.914	99	3,4	11.786	349	3,0

Tab. 6: *Incompatibilità rilevabili nel collegamento di 'records' individuali su due occasioni di indagine: 86 .I-II*

Caratteri	Veneto			Lombardia		
	Totale risposte	Incompatibilità V.A.	%	Totale risposte	Incompatibilità V.A.	%
Anno di nascita	6.301	83	1,3	27.741	396	1,4
Mese di nascita	6.301	102	1,6	27.741	388	1,4
Sesso	6.301	66	1,0	27.741	267	1,0
Relaz. col capofamiglia	6.301	96	1,5	27.741	270	1,0
Titolo di studio	6.301	573	9,1	27.741	2.134	7,7
Stato civile	5.315	18	0,3	23.455	74	0,3
Condizione professionale	5.315	8	0,2	23.455	26	0,1
Precedenti lavorativi	2.987	103	3,5	12.260	352	2,9

Tab. 7: *Incompatibilità rilevabili nel collegamento di 'records' individuali su due occasioni di indagine: 85.I-86.I*

Caratteri	Veneto			Lombardia		
	Totale risposte	Incompatibilità V.A.	%	Totale risposte	Incompatibilità V.A.	%
Anno di nascita	4.954	49	1,0	17.147	577	3,4
Mese di nascita	4.954	77	1,6	17.147	368	2,1
Sesso	4.954	33	0,7	17.147	241	1,4
Relaz. col capofamiglia	4.954	1.157	23,4	17.147	3.962	23,1
Titolo di studio	4.954	259	5,2	17.147	1.049	6,1
Stato civile	4.085	12	0,3	14.481	71	0,5
Condizione professionale	4.085	4	0,1	14.481	21	0,1
Precedenti lavorativi	2.274	98	4,3	7.876	333	4,2

Tab. 8: *Incompatibilità rilevabili nel collegamento di 'records' individuali su due occasioni di indagine: 85.II - 86.II*

Caratteri	Veneto			Lombardia		
	Totale risposte	Incompatibilità V.A.	%	Totale risposte	Incompatibilità V.A.	%
Anno di nascita	4.916	61	1,2	17.525	601	3,4
Mese di nascita	4.916	71	1,4	17.525	443	2,5
Sesso	4.916	46	0,9	17.525	263	1,5
Relaz. col capofamiglia	4.916	88	1,6	17.525	291	1,7
Titolo di studio	4.916	281	5,7	17.525	1.254	7,2
Stato civile	4.100	18	0,4	14.846	49	0,3
Condizione professionale	4.100	11	0,3	14.846	30	0,2
Precedenti lavorativi	2.242	95	4,2	7.884	323	4,1

memoria e di conoscenza della risposta, soprattutto nel fornirla per i familiari (sull'importanza dell'effetto intervistatore per la variabile titolo di studio, vedi peraltro il cap. 21).

Non sorprende invece la percentuale di incompatibilità nei precedenti lavorativi, mediamente superiore al 3%, poiché il relativo quesito è tipicamente di quelli che implicano uno sforzo di memoria, che, come è noto, fa aumentare il rischio di errata risposta.

Per tutte le altre variabili la quota di risposte incompatibili è assai limitata. Lo è, in particolare, per la condizione lavorativa. Ma a questo proposito bisogna osservare che, se la risposta per questa variabile fosse data in modo del tutto casuale, la quota di incompatibilità, per il modo in cui l'abbiamo definita, non sarebbe di molto superiore all'1%.

3.1. *Abbinamento di 'records' non sottoposti al piano di compatibilità*

Nella Tab. 9 sono riportate le incompatibilità rilevate dall'abbinamento di dati grezzi del Veneto e della Lombardia. Gli abbinamenti si riferiscono alle due prime occasioni dell'86[5].

Si osservi che le quote di incompatibilità, sia per il Veneto che per la Lombardia, sono sostanzialmente le stesse che si sono rilevate sui dati sottoposti al piano di compatibilità (Tab. 6). Soltanto per il titolo di studio i dati grezzi presentano quote superiori di circa un punto, un punto e mezzo percentuale.

La notevole differenza che emerge dal confronto delle due tabelle riguarda il numero totale di abbinati. Per il Veneto, operando sui dati non corretti si abbinano 201 individui in più, con un incremento pari al 3% circa sui dati corretti; per la Lombardia 2.029 persone in più, che rappresentano oltre il 7% degli abbinati corretti. Si deve inoltre osservare che anche il numero degli abbinati che non presentano alcun errore nelle variabili di collegamento è risultato superiore sui dati non corretti (Tab. 10): di poco nel Veneto (5.302 contro 5.298), ma quasi del 4,5% in Lombardia (25.107 contro 24.025). L'indicazione che emerge da queste evidenze è chiara. Il piano di compatibilità, mentre elimina incongruenze interne alla singola occasione di indagine, ne introduce tra occasioni successive o comunque collegabili, andando a forzare o imputare anche variabili identificative. Il maggior numero di abbinamenti privi di incompatibilità è presumibilmente dovuto alle imputazioni. Il programma LINK non considera incompatibile la mancata risposta ad una variabile di abbinamento, quando si verifica per ambedue le occasioni poste a confronto. Se l'imputazione per le due occasioni è difforme, può generare incompatibilità e anche mancato abbinamento. Non pensiamo tuttavia che, complessivamente, la maggior quota di abbinamento dei dati non corretti

---

5 Dei dati grezzi avevamo a disposizione soltanto le prime due occasioni dell'86. Sarebbe stato indubbiamente più interessante controllare l'abbinamento delle prime due occasioni dell'85 per verificare la quota di incompatibilità nella relazione con il capofamiglia.

Tab. 9: *Incompatibilità rilevabili nel collegamento di 'records' individuali su due occasioni di indagine: 86.I - II. Dati non sottoposti al "Piano di compatibilità".*

Caratteri	Veneto			Lombardia		
	Totale risposte	Incompatibilità V.A.	%	Totale risposte	Incompatibilità V.A.	%
Anno di nascita	6.502	93	1,43	29.770	427	1,43
Mese di nascita	6.502	104	1,60	29.770	400	1,34
Sesso	6.502	76	1,17	29.770	418	1,40
Relaz. col capofamiglia	6.502	127	1,95	29.770	400	1,34
Titolo di studio	6.502	727	11,18	29.770	2.742	9,21
Stato civile	5.517	32	0,58	25.177	129	0,05
Condizione professionale	5.517	4	0,07	25.177	21	0,01
Precedenti lavorativi	3.137	107	3,41	13.263	378	2,85

Tab. 10: *'Records' privi di incompatibilità negli abbinamenti effettuati*

Abbinamenti	Veneto			Lombardia		
	Totale abbinati	Abbinamenti corretti V.A.	%	Totale abbinati	Abbinamenti corretti V.A.	%
85.I - II	6.365	4.360	68,2	26.070	18.530	71,1
86.I - II	6.301	5.298	84,1	27.741	24.025	86,6
85.I - 86.I	4.954	3.400	68.6	17.147	11.430	66,7
85.II - 86.II	4.916	4.286	87,2	17.525	14.668	83,7
86.I - II dati grezzi	6.502	5.302	81,5	29.970	25.107	84,3

derivi completamente dall'assenza di risposta, dato che il programma di abbinamento controlla otto variabili, sei delle quali presentano, nei dati grezzi, percentuali di non risposta inferiori all'1% e le altre due tra l'1 e il 2%.

3.2. *Distorsioni dovute alla selezione dei 'records' abbinati*

Nelle precedenti sezioni abbiamo osservato come nella ricostruzione dell'informazione longitudinale mediante l'abbinamento esatto tra occasioni di indagine sia opportuno non selezionare eccessivamente l'insieme degli abbinati per evitare di introdurvi fattori di distorsione.

Una possibile causa di eccessiva selezione è l'assenza di risposta longitudinale, dipendente solo in parte dal metodo di abbinamento da noi adottato. Un metodo con minore tolleranza di errore, come quello adottato dall'Istat per la costruzione delle matrici di transizione, dà luogo ad un minor numero di abbinamenti e quindi ad una quota maggiore di non risposta longitudinale.

Gli abbinamenti ammessi dalla procedura Istat sono assimilabili - anche se non pienamente corrispondenti - a quelli da noi indicati come perfetti nelle Tabb. 11 e 12, vale a dire all'insieme dei *records* privi di incompatibilità sul totale dei *records* abbinati dal programma LINK. Le tabelle riportano, separatamente per questi *records* e per il totale degli abbinati, le distribuzioni di frequenza di alcuni caratteri (sesso, stato civile, titolo di studio, età e condizione professionale), rilevate alla prima occasione dell'85, su abbinati a tre mesi e ad un anno di distanza.

Per tutte le variabili considerate si osservano, nell'insieme degli abbinati perfetti, modalità sovrarappresentate o sottorappresentate rispetto al totale degli abbinati; alcune di queste sono rilevanti per l'analisi statistico-economica conseguente alla costruzione delle matrici di transizione. Ad esempio, gli uomini sono più rappresentati delle donne nei *records* perfetti, così come lo sono gli occupati; le casalinghe, al contrario, vi sono sottorappresentate. Queste differenze tra abbinati perfetti e non, sono in molti casi abbastanza modeste, potremmo dire poco significative, ma ci sembra importante il fatto che si ripetono sistematicamente per le stesse modalità, nei due abbinamenti considerati, sia per il Veneto che per la Lombardia.

4. *Incompatibilità sull'arco di quattro occasioni*

Un primo esame delle incompatibilità è stato effettuato anche sulla struttura longitudinale più estesa, rappresentata dall'abbinamento dei *records* individuali su quattro occasioni di indagine.

I risultati, riportati nella Tab. 13, sono stati ottenuti esaminando sequenzialmente le strutture longitudinali e rilevando eventuali incompatibilità nei tre passaggi dalla I alla II occasione dell'85, dalla II dell'85 alla I dell'86 e dalla I alla II dello stesso anno. In seguito, per semplicità, indicheremo la

Tab.11: *Distribuzione di alcuni caratteri alla I occasione di indagine dell'85, per abbinati privi di errori nelle variabili di collegamento e totale abbinati. Abbinamento 85.I-II*

Caratteri	Veneto		Lombardia	
	Perfetti	Totale records	Perfetti	Totale records
Sesso				
Maschi	52,43	48,44	52,74	48,67
Femminile	47,57	51,56	47,53	51,33
Stato civile				
Celibe/nubile	26,59	29,15	27,62	29,33
Coniugato/a	62,92	61,55	60,67	60,26
Vedovo/a	9,42	8,43	10,51	9,33
Separato/a	1,07	0,87	1,20	1,08
Titolo di studio				
Analfabeta	0,48	0,44	0,16	0,21
Nessun titolo	24,01	20,38	19,00	16,24
Licenza elementare	38,97	39,66	42,00	41,34
Scuola media inferiore	24,52	27,01	25,63	27,90
Scuola media superiore	9,50	10,29	10,71	11,85
Laurea	2,52	2,22	2,50	2,47
Età				
fino a 13	23,05	17,26	20,63	16,09
14 - 20	8,76	11,04	9,49	11,19
21 -25	5,92	7,36	6,22	7,24
26 - 30	6,19	6,74	5,77	6,28
31 - 35	7,39	7,60	6,58	6,77
36 - 40	7,41	7,63	7,38	7,53
41 - 50	12,82	13,18	14,21	14,62
51 - 60	11,47	11,71	12,60	12,85
più di 60	17,00	17,49	17,12	17,42
Condizione professionale				
Occupato	48,12	44,82	48,62	46,09
In cerca di nuova occup.	1,67	1,70	1,13	1,34
In cerca di prima occup.	2,00	2,26	1,89	2,09
In servizio di leva	0,51	0,62	0,54	0,62
Casalinga	18,57	22,33	17,76	19,91
Studente	8,08	8,63	7,63	8,63
Inabile	1,85	1,74	1,44	1,46
Ritirato	16,85	14,57	19,44	18,33
Altro	2,35	3,33	1,55	1,53

Tab.12: *Distribuzione di alcuni caratteri alla I occasione di indagine dell'85, per abbinati privi di errori nelle variabili di collegamento e totale abbinati. Abbinamento 85.I-86.I*

Caratteri	Veneto		Lombardia	
	Perfetti	Totale records	Perfetti	Totale records
Sesso				
Maschi	54,10	48,87	53,06	48,12
Femminile	45,90	59,80	46,94	51,83
Stato civile				
Celibe/nubile	27,77	29,28	25,95	28,84
Coniugato/a	62,35	61,76	61,92	60,60
Vedovo/a	8,76	8,16	10,77	9,33
Separato/a	1,12	0,80	1,36	1,23
Titolo di studio				
Analfabeta	0,42	0,48	0,13	0,22
Nessun titolo	23,47	20,86	17,88	16,08
Licenza elementare	39,83	40,33	39,47	38,91
Scuola media inferiore	23,69	26,36	25,92	28,18
Scuola media superiore	10,36	10,01	13,28	13,66
Laurea	2,23	1,95	3,23	2,96
Età				
fino a 13	22,24	17,54	19,37	15,56
14 - 20	9,07	11,22	9,05	11,58
21 -25	6,04	7,03	5,69	6,60
26 - 30	5,46	6,02	5,58	6,11
31 - 35	7,26	7,49	6,51	6,65
36 - 40	7,23	7,45	7,27	7,15
41 - 50	14,07	13,93	15,08	15,31
51 - 60	12,15	12,12	13,42	13,32
più di 60	16,53	17,20	18,03	17,73
Condizione professionale				
Occupato	47,96	44,20	48,99	45,61
In cerca di nuova occup.	2,33	1,98	1,16	1,38
In cerca di prima occup.	2,37	2,25	1,67	1,97
In servizio di leva	0,67	0,76	0,54	0,64
Casalinga	18,95	22,43	16,55	19,81
Studente	7,36	8,44	7,96	9,61
Inabile	2,12	1,96	1,35	1,33
Ritirato	16,25	14,97	20,35	18,59
Altro	1,99	3,01	1,43	1,67

Tab.13: *Distribuzione percentuale del numero di incompatibilità rilevabili nel collegamento di quattro occasioni e loro distribuzione nel passaggio tra successive occasioni di indagine*

Caratteri	Veneto				Lombardia			
	A - Totale occasioni							
	numero di incompatib.				numero di incompatib.			
	0	1	2	3	0	1	2	3
Anno di nascita	98,1	0,9	1,0	-	97,2	1,7	1,0	0,1
Mese di nascita	97,3	1,3	1,4	-	97,1	1,9	1,0	-
Sesso	98,4	1,0	0,5	-	97,8	1,2	1,0	-
Relaz. col capofamiglia	76,6	21,9	1,3	0,2	76,1	22,6	1,1	0,2
Titolo di studio	85,3	10,5	3,3	0,9	85,4	10,3	3,9	0,4
Stato civile	98,3	1,7	-	-	98,7	1,3	-	-
Condizione professionale	99,4	0,6	-	-	99,6	0,4	-	-
Precedenti lavorativi	88,8	10,8	0,4	-	91,4	8,3	0,3	-
	B - Tra occasioni successive							
	incomp. tra occasioni				incomp. tra occasioni			
	I--II	II--III	III--IV		I--II	II--III	III--IV	
Anno di nascita	16,2	44,1	39,7		24,5	40,1	35,4	
Mese di nascita	30,5	33,7	35,8		23,0	33,7	43,3	
Sesso	20,0	32,0	48,0		34,7	30,2	35,1	
Relaz. col capofamiglia	89,4	5,6	5,0		90,1	5,8	4,1	
Titolo di studio	34,5	26,7	38,8		37,9	25,5	36,6	
Stato civile	48,4	25,8	25,8		33,7	42,5	23,8	
Condizione professionale	16,6	41,7	41,7		23,1	53,8	23,1	
Precedenti lavorativi	37,4	32,2	30,4		30,2	44,6	25,2	

sequenza di occasioni con I, II, III e IV.

Il criterio scelto per il confronto tra occasioni ci consente di rilevare, per ogni variabile considerata, da 0 a 3 incompatibilità, che tuttavia non corrispondono necessariamente ad altrettanti errori. Due esempi serviranno a chiarire il criterio adottato. Si supponga che per la variabile sesso un record presenti: M F M F, vale a dire maschio alla I e III occasione e femmina alla II e IV. E' ovvio che complessivamente non possono esservi che due errori. C'è però anche un'incompatibilità ad ogni passaggio tra un'occasione e la successiva. Il criterio che abbiamo adottato segnalerà quindi tre incompatibilità. Se invece un *record* presenta M F M M, l'errore è molto probabilmente uno (è estremamente improbabile, che il *record* possa riferirsi ad una donna), mentre i passaggi non compatibili sono due: MF dalla I alla II e FM dalla II alla III; quindi, con il nostro criterio, due incompatibilità.

Possiamo allora osservare che, da un punto di vista quantitativo, la situazione delle incompatibilità su quattro occasioni rispecchia quella già esaminata su due. Da un punto di vista qualitativo, dato il criterio di costruzione della tabella, siamo indotti a considerare le doppie variazioni come delle rettifiche, almeno nella maggior parte dei casi. Per il sesso, ad esempio, che ha due sole modalità, deve trattarsi necessariamente di rettifiche. Lo stesso vale per i precedenti lavorativi e varrebbe per lo stato civile, dato il modo in cui sono configurati i casi di incompatibilità, se la variabile presentasse variazioni doppie. Per le altre variabili, in via teorica, sono possibili passaggi definiti incompatibili su modalità sempre diverse.

Le percentuali di maggiore rilievo nella Tab. 13 riguardano la relazione con il capofamiglia, a causa delle anomalie discusse nei precedenti paragrafi, e le due variabili più ricche di incompatibilità negli abbinamenti a due mesi: titolo di studio e precedenti lavorativi.

Nella seconda parte della Tab. 13 abbiamo riportato la percentuale di incompatibilità per ciascuna variabile, nei tre successivi passaggi tra le occasioni collegate. Le situazioni si presentano abbastanza diversificate e non ci sembra emergano sistematicità degne di rilievo. Fa eccezione, ovviamente, la relazione con il capofamiglia, le cui incompatibilità sono concentrate nella I occasione e quindi compaiono, nel 90% circa dei casi, nel passaggio da questa alla successiva[6].

5. *Note conclusive*

In questo capitolo si è proceduto all'esame dei dati longitudinali della RTFL (ottenuti dall'abbinamento delle prime due occasioni di indagine dell'85 e dell'86 nel Veneto e in Lombardia), tentando di mettere in rilievo gli elementi utili ad una valutazione della loro qualità, soprattutto in rapporto alla individuazione di mancate risposte e di incompatibilità di risposta.

Purtroppo i dati a nostra disposizione se da un lato consentono analisi anche approfondite degli aspetti quantitativi dei fenomeni studiati, dall'altro costituiscono il limite maggiore per un'interpretazione convincente ed argomentata dei fenomeni stessi. Più di una volta non siamo stati in grado di interpretare compiutamente i dati osservati. L'esempio tipico è rappresentato dalla quota rilevante di incompatibilità nella variabile relazione con il capofamiglia. Abbiamo ipotizzato un qualche problema nel piano di compatibilità Istat, che solo l'accesso a documentazione interna all'Istituto potrebbe confermare o smentire.

Un discorso analogo può essere fatto per la non risposta, particolarmente per la non risposta longitudinale. Nonostante l'uso di un programma di abbinamento piuttosto sofisticato, non siamo in grado di esprimere veri e propri giudizi sugli aspetti qualitativi dell'indagine legati all'individuazione

---

6 Nella lettura delle distribuzioni della Tab. 13/B occorre tenere presente che, sebbene si sia fatto uguale a 100 il totale dei passaggi incompatibili e non quello dei *records*, alcune percentuali derivano dal rapporto di numeri molto piccoli; ciò avviene in particolare per sesso, stato civile e condizione professionale.

dei dati mancanti.

Gli elementi raccolti sembrano confermare quanto era già emerso anche in ricerche svolte dall'Istat (Masselli, 1988), e cioè che la non risposta attribuibile al comportamento dei partecipanti all'indagine è una piccola parte di quella che, nei nostri dati, leggiamo sotto forma di mancati abbinamenti. La maggior parte di questi, ma è difficile quantificarli esattamente, sono da ascrivere ad errori nelle codifiche familiari.

Nell'ottica di una utilizzazione corretta dell'informazione longitudinale dell'indagine, il trattamento della non risposta assume notevole importanza. La scelta della metodologia da adottare è legata ad una corretta interpretazione delle situazioni che si osservano (vedi il cap. 18). Questa è tanto più difficile quanto maggiore è la libertà di interpretazione delle regole di conduzione dell'indagine. Ci riferiamo in particolare alle regole dettate per il *follow-up* e la sostituzione delle famiglie, sulla cui uniformità di accezione è ragionevole nutrire qualche dubbio, alimentato da colloqui avuti da una parte con funzionari Istat, dall'altra con i responsabili dell'indagine in alcuni Comuni.

Restando nell'ottica dell'informazione longitudinale, riteniamo che i nostri risultati possano ancora indurre a qualche riflessione sull'opportunità di studiare e, in un prossimo futuro, realizzare ulteriori programmi di imputazione e correzione dei dati, che integrino il piano di compatibilità dell'indagine. Se è fuori discussione l'importanza di quello attuale al fine della tempestiva pubblicazione dei risultati trimestrali, non vediamo perché lo stesso piano dovrebbe condizionare i dati longitudinali, la struttura dei quali offre possibilità di trattamento aggiuntive e più adeguate di quelle attualmente operanti.

9

SULLA PRESENZA DI DISTORSIONE NELLE STIME INDOTTA DALLA ROTAZIONE CAMPIONARIA

Giorgio Alleva

1. *Termini del problema ed obiettivi dello studio*

Quando un'indagine statistica viene effettuata periodicamente usando lo stesso, o più o meno lo stesso, questionario, ci possono essere dei vantaggi nella rotazione del campione; cioè può essere utile disporre di un regolare programma nel quale nuove unità sono estratte per rimpiazzare vecchie unità che sono state nel campione per un certo numero di interviste.

Il vantaggio della rotazione è rappresentato dal fatto che la stima del livello delle variabili oggetto di studio può essere migliorata utilizzando informazioni di periodi passati. Per ottenere questo miglioramento è necessaria una parziale intersezione tra campioni precedenti e campione corrente (Ekler, 1955; Fellegi, 1963; Rao e Graham, 1964; Castellini, 1982). Se lo scopo dell'indagine è la stima delle variazioni dei fenomeni oggetto dello studio, teoricamente sarebbe meglio mantenere il campione identico. Anche in questo caso è comunque vantaggiosa la rotazione campionaria, ovvero l'eliminazione di una parte del campione e la sua sostituzione con altre unità, ad intervalli predeterminati. La ragione principale di ciò è rappresentata dal voler evitare le distorsioni nelle stime indotte dalla prolungata presenza nel campione di unità statistiche (Williams e Mallow, 1970). Tale fenomeno è denominato generalmente *sample fatigue*.

Esperienze di altri Paesi hanno dimostrato come l'osservazione ripetuta sulle stesse unità introduca un particolare fattore di distorsione nelle stime di livello delle variabili studiate, in funzione del numero di volte in cui le unità sono state oggetto di indagine (Bailar, 1975; Ghangurde, 1982; Shack-Marquez, 1985). In particolare è stato osservato che per molte variabili, le stime costruite in sezioni del campione presenti nello stesso da un diverso numero di occasioni di indagine non hanno lo stesso valor medio.

Gli obiettivi alla base di questo studio sono i seguenti: (i) verificare l'esistenza di una distorsione nella stima delle diverse variabili oggetto dello studio indotta dal numero di volte in cui le unità (gli individui intervistati) sono oggetto di indagine; (ii) identificare i possibili fattori all'origine della eventuale distorsione nelle stime.

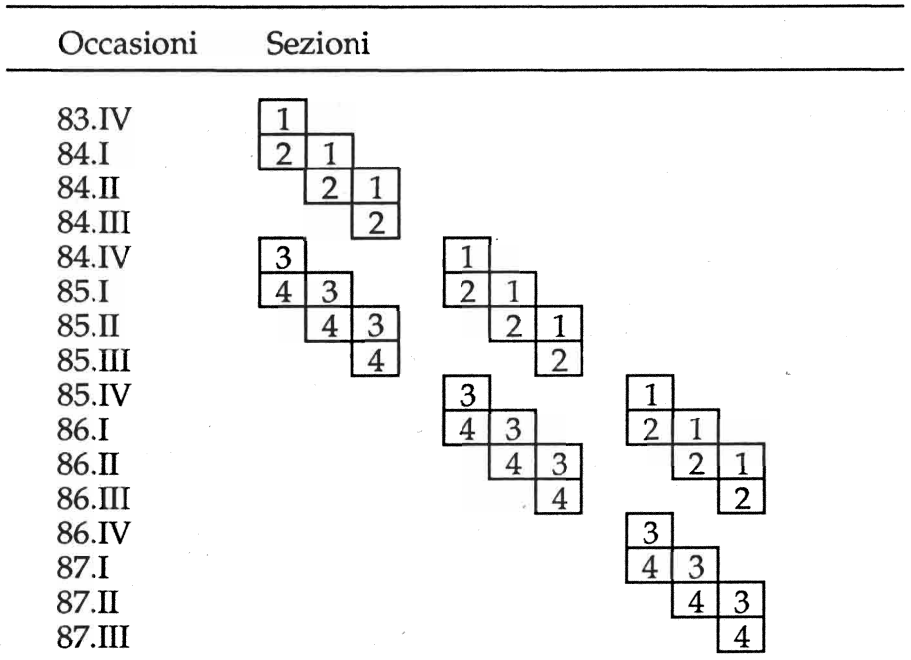
Allo scopo di verificare l'esistenza del fattore di distorsione, per le diverse

occasioni e per ognuno dei caratteri presi in considerazione, si intendono confrontare le stime di livello relative alle quattro sezioni di famiglie oggetto di indagine per un diverso numero di volte.

Si ricorda che in ognuna delle occasioni dell'indagine il campione è formato da quattro distinte sezioni di famiglie, rispettivamente presenti per la prima, la seconda, la terza e la quarta volta nell'indagine (vedi il cap. 1 e la Fig. 1). Le famiglie appartenenti a ciascuna sezione vengono intervistate per due occasioni consecutive, poi eliminate dal campione per altrettante occasioni e, quindi, intervistate nuovamente per altre due indagini consecutive.

Occorre ricordare che per un corretto confronto tra sezioni, ciascuna di esse deve essere composta da individui intervistati per lo stesso numero di volte. E' pertanto indispensabile eliminare da ciascuna sezione sia le famiglie di riserva, inserite per sostituire famiglie non reperibili o che non hanno voluto essere nuovamente intervistate, sia individui intervistati per un numero di volte diverso da quello delle famiglie di riferimento.

Fig. 1: *Schema del piano di rotazione della rilevazione delle forze di lavoro (il numero della sezione identifica il numero di volte da cui risulta presente)*



Ad esempio, se si vuole effettuare uno studio per l'occasione di indagine relativa al primo trimestre del 1985 (indicata nel seguito come occasione 1985.I), occorre effettuare un analisi che prende in considerazione anche le occasioni 1983.IV, 1984.I, 1984.II e 1984.IV. Infatti, osservando la Fig. 1 nel senso delle colonne (le sezioni), mentre non occorre effettuare controlli sugli individui presenti nell'indagine 1985.I per la prima volta, occorre assicurarsi che quelli considerati presenti per la seconda volta siano effettivamente stati presenti nell'indagine precedente (1984.IV), che quelli considerati presenti per la terza volta siano stati effettivamente intervistati nelle precedenti occasioni 1984.I e 1984.II, e infine che gli individui indicati come presenti per la quarta ed ultima volta nel 1985.I siano stati effettivamente intervistati nelle occasioni 1983.IV, 1984.I e 1984.IV. A questo scopo, si è utilizzata la procedura di abbinamento LINK descritta nel cap. 7.

2. *Piano di lavoro*

La verifica dell'esistenza del fattore di distorsione nelle stime di livello delle variabili per ciascuna occasione di indagine può essere effettuata attraverso due differenti approcci, ai quali corrispondono le due fasi attraverso le quali si è sviluppato il presente studio.

In una *prima fase*, allo scopo di individuare quali siano i caratteri sui quali ricadono i maggiori indizi della presenza della distorsione, il problema è stato affrontato in termini descrittivi-esplorativi, attraverso la costruzione e il confronto di indicatori associati a gruppi di individui presenti nel campione da un diverso numero di volte. Tali indicatori risultano differenti a seconda della natura delle variabili stesse: in particolare, si distinguono i tre casi delle variabili quantitative, delle mutabili dicotomiche e delle mutabili politomiche sconnesse. Se la dimensione e la variabilità dei valori di tali indici consente di ottenere indizi della presenza della distorsione, l'andamento dei valori in funzione del numero di volte in cui gli individui sono presenti nel campione permette di ottenere indizi sulle cause della distorsione. La metodologia utilizzata e i principali risultati sono presentati nelle sezz. 3.1 e 3.2.

Nella *seconda fase*, allo scopo di valutare se i caratteri risultati indiziati di distorsione nella precedente analisi siano o meno affetti dalla distorsione, è stata verificata l'ipotesi di dipendenza tra le modalità di risposta e il numero di volte in cui gli individui sono presenti nel campione. Pertanto, se il primo approccio, quello di natura esplorativa, si base sulla statistica descrittiva, il secondo, di natura valutativa-confermativa, si basa sulla statistica inferenziale. La metodologia utilizzata e i principali risultati di tale analisi sono illustrati nelle sezz. 4.1 e 4.2.

Come verrà ampiamente descritto nel seguito (sez. 3), la prima fase dello studio, quella di tipo esplorativo, è stata effettuata su un rilevante numero di caratteri ed ha condotto al risultato di focalizzare la ricerca e la verifica dell'eventuale distorsione sulla classe degli individui che si sono dichiarati in cerca di occupazione. Tale verifica, di tipo inferenziale, effettuata nella

seconda fase (sez. 4), ha condotto alla conclusione di dover escludere la presenza di distorsione nella stima dell'aggregato delle persone in cerca di occupazione indotta dalla rotazione campionaria.

L'analisi inferenziale è stata effettuata sfruttando al massimo le informazioni disponibili e con il rigore necessario alla valutazione del problema. In primo luogo, mentre l'analisi esplorativa è stata condotta solo sui questionari relativi alla regione Veneto, quella inferenziale ha preso in considerazione il complesso dei dati disponibili, relativi al Veneto e alla Lombardia. In secondo luogo, il confronto tra le risposte fornite dagli stessi individui in occasioni diverse è stato effettuato previa la citata operazione di *matching* (vedi il cap. 7) soltanto nell'analisi inferenziale. (Nell'analisi esplorativa la già ridotta dimensione del campione utilizzato ne ha sconsigliato una ulteriore diminuzione ad opera di tale procedura.)

Un ulteriore affinamento è stato infine introdotto nella definizione degli individui in cerca di occupazione. Infatti, nell'analisi esplorativa si è presa in considerazione semplicemente la dichiarazione fornita dagli intervistati alla domanda 10.1 del questionario circa la condizione professionale. Così sono stati considerati "in cerca di occupazione" coloro che hanno risposto alla seconda e alla terza modalità, rispettivamente associate agli individui che sono in cerca di "nuova" e "prima" occupazione. Di contro, nell'analisi inferenziale, la definizione adottata risulta quella corrente dell'Istat (al riguardo, vedi il cap. 1).

3. *Individuazione dei caratteri sui quali ricadono i maggiori indizi della presenza della distorsione nelle stime indotta dalla rotazione campionaria: analisi esplorativa*

3.1. *Metodologia*

Come già detto, obiettivo di tale fase della ricerca è il confronto, in ciascuna occasione e per ciascuna variabile, delle stime di livello relative a sezioni di famiglie presenti nel campione da un numero diverso di volte. Gli indicatori attraverso i quali viene istituito tale confronto sono pertanto calcolati disgiuntamente per le diverse occasioni di indagine e sono diversi a seconda che si intenda valutare la distorsione nelle stime di variabili quantitative, di mutabili dicotomiche o di mutabili politomiche sconnesse.

Un primo indicatore, costituito da un rapporto tra valori medi, risulta adeguato per la valutazione delle stime di variabili quantitative. Avendo riguardo all'i-esima occasione di indagine e alla j-esima variabile oggetto di studio, si può costruire un indice pari al rapporto tra il valor medio della stima di livello della j-esima variabile nella k-esima sezione e il valor medio della stima di livello della medesima variabile calcolato per il complesso delle quattro sezioni:

$$\frac{E({}_iY_{j,k})}{E({}_iY_j)} \quad (1)$$

In caso di presenza della distorsione, la stima della variabile studiata è dipendente dal numero di volte in cui le famiglie sono intervistate (ovvero dalle sezioni). Valori significativamente inferiori o superiori all'unità saranno indizio della distorsione, indicando rispettivamente stime minori o maggiori per quelle determinate sezioni presenti un diverso numero di volte nel campione.

Ai fini dell'utilizzazione di tali indici, si noti che le variabili quantitative rappresentano soltanto una minoranza delle informazioni desumibili dall'indagine (età, ore di lavoro, durata della ricerca, e poche altre).

Nel caso di mutabili dicotomiche, l'indice proposto è rappresentato da un rapporto tra frequenze relative (indicate con n). In particolare, si tratta, per ciascuna i-esima occasione di indagine, del rapporto tra la quota degli individui contraddistinti da una delle due modalità (j_1 e j_2) della j-esima variabile nella k-esima sezione e la quota degli individui che sono caratterizzati dalla medesima modalità nel complesso delle sezioni:

$$\frac{({}_in_{j1,k}/{}_in_k)}{({}_in_{j1}/{}_in)} \quad (2)$$

Come per l'indice proposto per le variabili quantitative, anche tali indici tenderanno ad essere pari ad uno nel caso di assenza della distorsione.

Pertanto, l'osservazione dei quattro valori, corrispondenti alle quattro sezioni, degli indici così costruiti consentirà di valutare, nella specifica occasione di indagine e per la variabile in esame:

(a) la loro dispersione intorno al valore medio (pari all'unità), ovvero l'entità della distorsione;

(b) il 'tipo' della distorsione, ovvero il suo verso (crescente o decrescente al variare del numero di volte in cui una sezione compare nell'indagine) e il suo 'punto di rottura' (dopo quante indagini la distorsione si presenti superiore ad una certa intensità).

Se il confronto tra gli indici relativi alla medesima variabile, costruiti nelle diverse occasioni disponibili, consentirà di valutare la sistematicità o meno, nel tempo, della distorsione, quello tra gli indici relativi alle diverse variabili consentirà invece di graduare le variabili a seconda dell'entità, sistematicità e 'tipo' della distorsione.

Nel caso di mutabili politomiche sconnesse, molto numerose nel questionario, il confronto tra le risposte fornite da individui appartenenti a diverse sezioni deve essere improntato prendendo in considerazione il complesso delle modalità in cui si manifesta il carattere in esame. Per ciascuna variabile e in ciascuna occasione si dispone in particolare di una tabella a doppia entrata le cui righe rappresentano le diverse modalità della variabile e le colonne le diverse sezioni. Elemento generico di tale tavola risulta ${}_in_{jj,k}$, ovvero il numero di individui che sono contraddistinti nell'i-esima occasione

e nella k-esima sezione dalla modalità jj della variabile j. I margini di riga $_in_{jj}$ rappresentano pertanto il numero di individui che nel complesso del campione hanno risposto con la modalità jj, e quelli di colonna $_in_k$ il totale di individui appartenenti alla k-esima sezione.

Avendo riguardo alle singole modalità, la valutazione della distorsione potrà essere effettuata analogamente a quanto proposto per le mutabili dicotomiche. Infatti, l'interesse nelle risposte ad una domanda caratterizzata da diverse modalità è spesso concentrato soltanto su alcune di esse. Ad esempio, prendendo in considerazione una delle variabili cruciali del questionario, quella relativa alla condizione professionale (la domanda 10.1 del questionario), nonostante le modalità di risposta siano 9, l'interesse nella valutazione della distorsione è riposto essenzialmente negli intervistati che si dichiarano in cerca di occupazione (ovvero nella seconda e nella terza modalità di risposta, rispettivamente associate agli individui che sono in cerca di "nuova" e "prima" occupazione).

Inoltre, sono state trattate come mutabili dicotomiche, le modalità di risposta che si riferiscono a domande che ammettono risposte plurime.

Se la dimensione e la variabilità dei valori di tali indici consente di ottenere indizi della presenza della distorsione, l'andamento dei valori in funzione del numero di volte in cui gli individui sono presenti nel campione permette di ottenere indizi sulle cause della distorsione. L'analisi circa il 'tipo' della distorsione, può essere effettuata a partire dalla seguente osservazione. Se si considera il rango dei 4 valori dell'indice associato ad un determinato carattere, in corrispondenza del numero di volte nelle quali gli individui sono intervistati (che corrisponde alle quattro sezioni), è possibile classificare la distribuzione dei ranghi in una delle 4 ! = 24 diverse permutazioni.

Allo scopo di associare a ciascuno dei possibili andamenti degli indici una interpretazione del 'tipo' di distorsione, è possibile suddividere le 24 permutazioni in diverse tipologie. Una prima suddivisione può essere effettuata tra andamenti crescenti ed andamenti decrescenti. Inoltre è possibile suddividere gli andamenti crescenti (decrescenti) in quattro classi (Fig. 2).

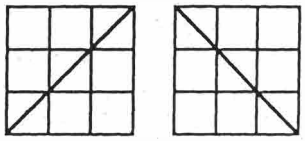
Prima classe di distorsione. E' costituita dal caso di andamento costantemente crescente (decrescente). Nel caso costantemente crescente, nella successione delle occasioni di interviste, un sempre maggior numero di individui tende a riconoscersi in una certa categoria (modalità di risposta).

Seconda classe di distorsione. E' costituita dagli andamenti crescenti (decrescenti) per entrambe le due coppie di occasioni consecutive. Si potrebbe trattare della distorsione indotta dalla cosiddetta *sample fatigue*. Gli individui, infatti, sono intervistati in due occasioni consecutive, eliminati dal campione per altrettante occasioni e poi nuovamente intervistati per due occasioni consecutive. Pertanto, a fronte di una tendenza a riconoscersi nella successione di interviste in misura crescente (decrescente) in determinate categorie, il lasso temporale tra le due coppie di interviste sembrerebbe indurre un certo 'recupero'. I cinque diversi andamenti che costituiscono tale classe si differenziano nell'intensità della distorsione (rappresentata dalla tendenza di fondo di andamento crescente/decrescente) e nella capacità del 'recupero'.

Fig. 2: *Classi di distorsione*

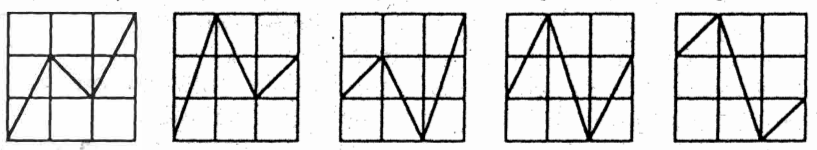
Prima classe di distorsione

Tipo 1: 1,2;3,4 Tipo 2: 4,3;2,1

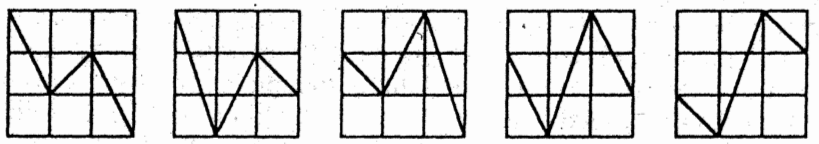


Seconda classe di distorsione

Tipo 3: 1,3;2,4 Tipo 4: 1,4;2,3 Tipo 5: 2,3;1,4 Tipo 6: 2,4;1,3 Tipo 7: 3,4;1,2

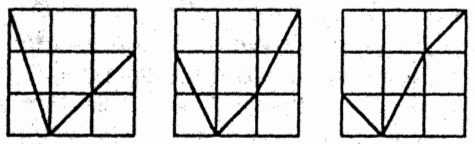


Tipo 8: 4,2;3,1 Tipo 9: 4,1;3,2 Tipo 10: 3,2;4,1 Tipo 11: 3,1;4,2 Tipo 12: 2,1;4,3

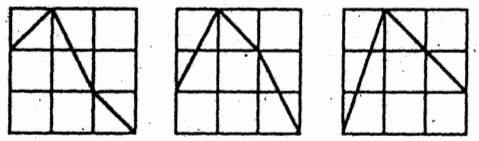


Terza classe di distorsione

Tipo 13: 4,1;2,3 Tipo 14: 3,1;2,4 Tipo 15: 2,1;3,4

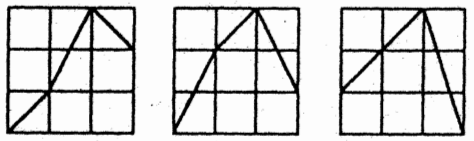


Tipo 16: 3,4;2,1 Tipo: 17: 2,4;3,1 Tipo: 18: 1,4;3,2

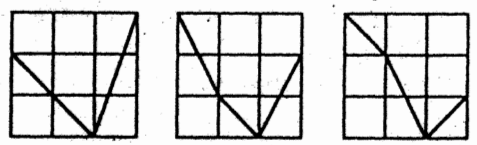


Quarta classe di distorsione

Tipo 19: 1,2;4,3 Tipo 20: 1,3;4,2 Tipo 21: 2,3;4,1



Tipo 22: 3,2;1,4 Tipo 23: 4,2;1,3 Tipo 24: 4,3;1,2



Terza classe di distorsione. E' costituita dagli andamenti crescenti (decrescenti) a partire dalla seconda volta in cui gli individui sono presenti nel campione. Si tratta dei casi in cui, a fronte di una tendenza a riconoscersi nella successione di interviste in misura crescente (decrescente) in determinate categorie, nella prima intervista gli individui tendono a non riconoscersi in quella determinata categoria. La causa alla base di questo tipo di distorsione potrebbe essere legata alla difficoltà, oggettiva o soggettiva, nel riconoscimento della modalità di risposta. L'intensità della distorsione (rappresentata dalla tendenza di fondo di andamento crescente/decrescente) e l' 'effetto prima intervista' differenziano i tre diversi andamenti di tale classe di distorsione.

Quarta classe di distorsione. E' formata dagli andamenti crescenti (decrescenti) fino alla terza volta in cui gli individui sono presenti nel campione. Si tratta dei casi in cui, a fronte di una tendenza a riconoscersi nella successione di interviste in misura crescente (decrescente) in determinate categorie, nell'ultima intervista gli individui tendono a non riconoscersi in quella determinata categoria. La causa alla base di questo tipo di distorsione potrebbe essere legata a diverse cause: (i) caduta nell'attenzione a fornire una risposta coerente; (ii) difficoltà degli intervistati nell'ammettere il perseverare della modalità di risposta fornita nelle precedenti occasioni, in specie se si tratta di riconoscere una situazione di disagio economico; (iii) nel caso che gli intervistati sappiano che si tratta dell'ultima intervista, tendenza a fornire informazioni non più relative alla situazione attuale, bensì a quella prevista (o auspicata). I tre diversi andamenti che costituiscono tale classe si differenziano nell'intensità della distorsione (rappresentata dalla tendenza di fondo di andamento crescente/decrescente) e nell' 'effetto ultima risposta'.

La valutazione del 'tipo' di distorsione, effettuata in una prima fase della ricerca sui caratteri maggiormente indiziati, non è qui riportata. Come già messo in evidenza, la fase successiva di ricerca, quella tipo inferenziale (sez. 4), ha escluso la presenza di distorsione nelle stime indotta dalla rotazione campionaria. Pertanto le precedenti osservazioni vanno intese soltanto come una proposta di classificazione di distorsione delle stime secondo le diverse possibili cause alla sua base.

3.2. *Primi risultati dell'analisi*

L'analisi sulla presenza di distorsione nelle stime si è basata in questa prima fase sui dati relativi alla regione Veneto, in 4 distinte occasioni; in particolare si tratta delle prime due indagini del 1985 e del 1986 (indicate nel seguito con 1985.I, 1985.II, 1986.I e 1986.II).

Come già riportato in precedenza, in questa prima fase non è stato previamente operato l'abbinamento dei dati individuali, necessario per un corretto confronto tra le risposte fornite dalle varie sezioni di individui presenti nel campione. I confronti sono stati condotti su medie, o frequenze relative, riferite *tout court* alle sezioni.

Occorre sottolineare immediatamente che i risultati di tale fase esplora-

tiva hanno condotto a focalizzare la ricerca della distorsione sulla stima del numero di persone in cerca di occupazione.

I risultati delle prime analisi effettuate sono riportati nelle Tabb. 1 e 2. Mentre la Tab. 1 riguarda l'analisi del complesso dei caratteri considerati, la Tab. 2 riporta i risultati dell'analisi specifica condotta sulla posizione professionale (e sulla disoccupazione in particolare).

Nella Tab. 1 sono riportati, per l'occasione di indagine 1986.I:

(a) per ciascuna variabile quantitativa considerata, l'indice indicato dall'espressione (1), moltiplicato per 100;

(b) per ciascuna modalità di mutabile dicotomica o politomica sconnessa considerata, l'indice indicato dall'espressione (2). Anche in questo caso l'indice è stato moltiplicato per 100.

L'osservazione dei quattro valori degli indici, corrispondenti alle quattro sezioni di famiglie, consente di valutare l'entità e il tipo della distorsione per ciascuna delle variabili quantitative e per ciascuna modalità delle mutabili. L'entità della distorsione può essere sintetizzata da una misura della dispersione dell'indice intorno alla sua media (che risulta pari a 100). A tale proposito, per ciascuna modalità, si riportano i valori dello scostamento quadratico medio dei 4 valori degli indici (indicato con s.q.m.).

Allo scopo di tener conto dell'importanza relativa di ciascuna modalità di risposta, si riporta la frequenza media delle singole risposte, nelle 4 sezioni. Poiché le risposte si concentrano spesso soltanto in alcune modalità, la frequenza media consente di valutare la significatività dell'entità della distorsione.

Infine, si noti che l'analisi è stata effettuata anche per una serie di caratteristiche ascrittive degli intervistati, o assimilabili a tali. Si tratta in particolare del sesso, del titolo di studio, dell'età e della dimensione demografica del comune di residenza. Quest'ultimo carattere è stato considerato nelle due modalità "grandi comuni" (si tratta dei capoluoghi di provincia e dei centri con oltre 20.000 abitanti) e "piccoli comuni" (centri con meno di 20.000 abitanti). L'analisi trova giustificazione nel fatto che si intende valutare la capacità interpretativa degli indici proposti nel caso di caratteri presumibilmente non affetti da distorsione.

Prendendo in considerazione la misura dello scarto quadratico medio dei valori degli indici proposti, si può osservare quanto segue.

(a) La distorsione risulta nulla in corrispondenza dei caratteri acrittivi e delle informazioni plausibilmente di natura meno incerta (dimensione demografica del comune di residenza, sesso, titolo di studio, occupazione permanente, occupazione a tempo pieno, casalinga, studente, ecc.). Infatti, le differenze tra i valori dell'indice risultano estremamente ridotte, il che induce a ritenere che si tratti di differenze di natura casuale e non del risultato del particolare fattore di distorsione in esame. Si consideri ad esempio la modalità "maschi" della mutabile sesso degli intervistati. I valori dell'indice variano in corrispondenza del numero di volte che gli individui sono intervistati da 99,40 a 98,99, a 100,24, a 101,34. La minima variabilità di tali valori (s.q.m. pari a 1,0) è indizio di assenza della distorsione. Stessa considerazione può essere effettuata anche per le

Tab. 1: *Distribuzione degli indici di distorsione secondo la sezione e le diverse modalità di risposta. Veneto, 1986.I.* (s.q.m.: scarto quadratico medio degli indici di distorsione; frequenza media: numero medio di risposte per sezione)

Mutabile - variabile	Sezione (n. volte nel campione)				s.q.m.	frequenza media
	1	2	3	4		
10.1 Qual'è attualmente la sua condizione						
Occupato	98,33	99,13	101,25	101,21	1,5	1396,0
In cerca di nuova occupazione	75,73	104,03	106,07	113,20	16,5	51,8
In cerca di prima occupazione	70,19	95,54	110,28	122,73	22,7	72,0
In servizio di leva	116,35	132,93	65,35	86,56	30,1	19,5
Casalinga	104,22	99,40	101,72	94,79	4,0	662,0
Studente	105,69	103,23	90,38	101,05	6,7	282,0
Inabile al lavoro	89,28	96,72	111,19	102,25	9,2	67,0
Persona ritirata dal lavoro	103,08	98,52	98,35	100,18	2,2	527,3
Altra condizione (benestante, anziano e simili)	101,66	116,79	93,82	87,94	12,5	70,0
11.1 Ore effettive svolte	99,53	101,10	98,22	101,15	1,4	1433,0
11.1 Ore svolte abitualmente	100,94	99,26	99,13	95,55	2,3	1433,0
11.3 Posizione nella professione						
Operaio, subalterno o assimilati	99,18	98,36	101,77	100,60	1,5	626,0
Lavoratore in proprio	101,23	100,52	99,40	98,93	1,0	309,8
Impiegato o intermedio	95,50	103,32	101,52	99,41	3,4	304,3
Coadiuvante	114,91	94,54	96,26	95,11	9,8	85,5
Apprendista	117,06	86,22	86,75	111,07	16,1	31,3
Dirigente	98,61	109,21	84,04	108,53	11,7	26,5
Libero professionista	97,42	116,26	97,76	88,58	11,6	25,8
Imprenditore	108,82	98,42	84,89	108,68	11,3	18,3
Lavorante a domicilio per conto d'imprese	36,36	121,48	117,87	120,73	42,2	5,8
11.5 Tempi di lavoro						
— L'attività è esercitata a tempo pieno	100,08	99,68	100,58	99,65	0,4	1353,5
L'attività è esercitata a tempo parziale per uno dei seguenti motivi:						
— Non desidera un lavoro a tempo pieno	104,52	103,49	86,07	106,52	9,5	27,0
— Non ha potuto trovare un lavoro a tempo pieno	103,40	107,30	87,45	102,37	8,7	23,3
— Altri motivi	86,88	103,68	115,69	92,73	12,7	19,3
— Malattia o invalidità	57,01	90,72	88,02	162,28	44,8	5,5
— Frequenta corsi scolastici o di formazione professionale	139,36	133,05	21,52	110,19	53,9	4,5
11.6 Carattere permanente o meno dell'occupazione						
— Ha un'occupazione permanente	99,82	100,62	99,93	99,63	0,4	1388,5
Ha un'occupazione temporanea perché						
— Non ha potuto trovare un lavoro permanente	110,27	78,96	93,63	117,70	17,3	22,8

segue Tab. 1: *Distribuzione degli indici di distorsione secondo la sezione e le diverse modalità di risposta. Veneto, 1986.I.* (s.q.m.: scarto quadratico medio degli indici di distorsione; frequenza media: numero medio di risposte per sezione)

Mutabile - variabile	Sezione (n. volte nel campione)				s.q.m.	frequenza media
	1	2	3	4		
— Altri motivi	104,52	87,32	72,62	136,36	27,3	8,0
— Il contratto di lavoro riguarda un periodo di formazione	89,59	57,02	138,32	113,34	34,8	7,0
— Non desidera un lavoro permanente	108,39	103,49	129,10	58,77	29,6	6,8
12 Eventuale seconda attività lavorativa						
Svolge soltanto un'attività	100,14	99,91	99,82	100,13	0,2	1414,3
Possiede anche una seconda attività che:						
— ha svolto nella settimana di riferimento	64,32	112,58	119,17	101,71	24,6	9,8
— ha svolto o prevede di svolgere nel corso dell'anno	116,14	99,79	107,58	77,13	16,7	9,0
13.1 Hai mai lavorato in passato?						
Sì	100,78	98,23	102,80	98,21	2,2	963,5
No	99,01	102,28	96,41	102,30	2,8	751,0
13.2 Da quanti mesi ha lasciato l'ultima occupazione?	100,37	98,40	100,74	100,46	1,1	962,8
13.3 Motivo per cui ha lasciato l'occupazione						
Fine di un lavoro a tempo determinato	100,86	94,96	113,49	90,53	10,0	46,0
Pensionamento per raggiunti limiti di età o per altri motivi	100,30	95,20	98,50	106,65	4,8	41,0
Licenziamento	81,35	103,56	95,69	118,99	15,7	35,0
Ritiro dal lavoro per motivi di salute o maternità	133,48	94,11	109,72	63,26	29,4	19,8
Dimissioni o cessazioni di attività in proprio	97,34	131,54	78,80	89,70	22,9	16,3
Pensionamento anticipato per motivi economici	120,51	70,81	75,05	138,82	33,2	5,3
Servizio di leva	..	148,70	..	249,87	122,7	1,3
13.4.1 Posizione nella professione (ultima occupazione)						
Operaio, subalterno o assimilati	97,85	94,55	100,90	107,01	5,3	103,8
Impiegato o intermedio	116,20	94,34	104,53	86,23	12,9	25,8
Lavoratore in proprio	106,86	116,79	87,23	88,31	14,5	16,0
Coadiuvante	40,07	140,14	124,62	88,31	45,2	8,0
Apprendista	85,49	130,80	119,63	60,56	32,4	5,0
Dirigente	189,98	41,52	44,31	134,57	70,7	2,3
Lavorante a domicilio per conto d'imprese	256,47	74,74	79,76	..	105,9	1,3
Imprenditore	..	186,86	..	201,85	115,6	1,0
Libero professionista	..	186,86	..	201,85	115,6	0,5

segue Tab. 1: *Distribuzione degli indici di distorsione secondo la sezione e le diverse modalità di risposta. Veneto, 1986.I.* (s.q.m.: scarto quadratico medio degli indici di distorsione; frequenza media: numero medio di risposte per sezione)

Mutabile - variabile	Sezione (n. volte nel campione)				s.q.m.	frequenza media
	1	2	3	4		
14.1 Cerca attivamente lavoro?						
No, non ha possibilità o interesse a lavorare	102,79	99,76	98,73	98,83	1,9	1526,0
No, ha già un lavoro e non ne cerca un altro	98,46	100,45	101,70	99,31	1,4	1364,8
Si, cerca un lavoro alle dipendenze	90,30	100,53	102,74	106,04	6,8	150,8
No, ma potrebbe lavorare a particolari condizioni	110,19	111,99	85,94	92,49	12,9	36,5
Si, ha già un'occupazione ma aspira a un lavoro migliore	97,06	89,92	99,94	112,92	9,6	25,5
Intende esercitare un lavoro in proprio ma non ha i mezzi	80,36	88,05	91,66	139,26	26,8	19,3
Si, ha già un'occupazione ma è temporanea	72,80	93,83	103,79	128,49	23,2	8,5
Si, ha già un'occupazione ma cerca un altro lavoro per altri motovo	71,12	137,52	40,56	150,65	52,7	7,3
Si, ha già un'occupazione ma teme di perderla	79,33	46,01	120,64	152,75	46,9	6,5
Inizierà tra breve un lavoro alle dipendenze	68,75	..	196,04	132,39	84,8	1,5
Si, ha già un'occupazione ma cerca un secondo lavoro	206,26	..	196,04	..	115,5	0,5
Inizierà un lavoro in proprio ed ha già predisposto i mezzi	..	199,40	..	198,58	115,5	0,5
14.2 Come dovrebbe essere l'occupazione ricercata						
Dipendente esclusivamente a tempo pieno	87,95	106,16	112,78	92,35	11,6	79,5
Dipendente preferibilmente a tempo pieno	100,50	100,09	98,35	101,03	1,2	54,5
Senza preferenza	104,53	94,81	87,41	112,30	10,9	41,3
Indipendente	80,04	83,17	94,50	134,19	25,4	24,8
Dipendente preferibilmente a tempo parziale	182,40	89,50	93,22	53,20	52,5	11,5
Dipendente esclusivamente a tempo parziale	133,18	129,39	77,96	69,92	32,5	8,8
14.4 Da quanti mesi è alla ricerca dell'occupazione	90,83	98,80	99,73	108,11	7,1	220,3
14.6 Quali azioni concrete di ricerca ha compiuto						
Visita personale a possibili datori di lavoro	117,30	93,72	98,77	93,68	11,1	26,3

segue Tab. 1: *Distribuzione degli indici di distorsione secondo la sezione e le diverse modalità di risposta. Veneto, 1986.I.* (s.q.m.: scarto quadratico medio degli indici di distorsione; frequenza media: numero medio di risposte per sezione)

Mutabile - variabile	Sezione (n. volte nel campione)				s.q.m.	frequenza media
	1	2	3	4		
Iscrizione presso ufficio pubblico di collocamento senza indennità	85,09	96,03	102,98	112,32	11,6	93,3
Segnalazione a datori di lavoro da parte di amici o di conoscenti	118,42	73,80	144,04	68,31	35,9	4,0
Invio a datori di lavoro di domande scritte di assunzione	102,37	103,41	99,28	95,71	3,5	118,0
Iscrizione presso ufficio pubblico di collocamento con indennità	93,89	91,01	108,43	105,03	8,5	83,3
Azioni concrete di ricerca non ancora iniziate	104,52	101,64	88,36	106,04	8,1	53,3
Risposta a offerte di lavoro pubblicate sui giornali	104,52	105,33	95,58	95,78	5,3	53,3
Partecipazione a concorsi per l'assunzione nel settore pubblico	66,10	146,46	89,33	93,20	34,4	10,8
Inserzioni sui giornali per richiesta di lavoro	118,42	116,85	120,03	51,23	33,1	16,0
Iscrizione presso un ufficio privato di collocamento	150,72	89,46	81,48	88,32	31,6	16,5
Altre azioni di ricerca	43,06	107,35	69,84	165,60	55,0	2,8
14.8 Per quale motivo non cerca lavoro						
Motivi personali o familiari	98,52	99,39	102,54	99,53	1,8	535,8
Ritiro dal lavoro per età	98,78	99,00	103,86	98,34	2,6	453,3
Motivi di studio	102,39	103,41	92,15	102,10	5,3	270,0
Motivo di salute, invalidità o altro impedimento fisico	99,67	97,76	104,89	97,65	3,4	204,0
Assenza di bisogno	104,59	93,54	96,88	105,06	5,7	56,5
E' considerato troppo giovane o troppo vecchio dai datori di lavoro	95,39	104,46	66,35	134,15	27,9	21,0
Servizio di leva	119,11	129,36	59,18	92,45	31,4	18,5
Vana ricerca di un lavoro in passato	44,52	88,63	44,23	223,58	84,6	2,3
Non sa	320,51	79,77	..	..	151,6	1,3
Convinzione di non disporre di sufficiente preparazione professionale	200,32	199,43	..	..	115,5	1,0
1 Dimensione demografica dei comuni di resistenza						
Grandi comuni	100,68	99,226	100,31	99,802	0,6	1211,8
Piccoli comuni	99,573	100,48	99,807	100,12	0,4	1935,8
3 Sesso						
Maschi	99,402	98,992	100,24	101,34	1,0	1514,8
Femmine	100,56	100,94	99,778	98,758	1,0	1632,8

segue Tab. 1: *Distribuzione degli indici di distorsione secondo la sezione e le diverse modalità di risposta. Veneto, 1986.I.* (s.q.m.: scarto quadratico medio degli indici di distorsione; frequenza media: numero medio di risposte per sezione)

Mutabile - variabile	Sezione (n. volte nel campione)				s.q.m.	frequenza media
	1	2	3	4		
4 Età						
14 - 19 anni	100,34	100,46	96,717	102,51	2,4	965,0
20 - 24 anni	99,407	99,604	103,39	97,565	2,4	292,0
25 - 29 anni	103,47	108,28	99,531	88,921	8,2	254,8
30 - 39 anni	100,14	99,217	99,92	100,72	0,6	548,0
40 - 49 anni	94,629	101,2	97,628	106,37	5,1	478,0
50 - 59 anni	98,028	98,934	105,87	97,073	4,0	468,8
Oltre 59 anni	103,19	97,206	100,97	98,721	2,6	747,5
8 Titolo di studio						
Analfabeta	101,02	56,971	120,03	121,58	30,1	12,3
Nessun titolo	109,85	90,931	100,58	98,967	7,8	307,0
Licenza elementare	96,508	99,604	102,78	100,94	2,6	1312,3
Licenza scuola media inferiore	100,22	102,8	95,447	101,61	3,2	1046,5
Diploma scuola media superiore	102,86	101,49	100,54	95,206	3,4	389,0
Laurea	102,49	104,03	105,94	87,573	8,4	80,5

quattro variabili quantitative considerate (ore di lavoro effettive ed abituali, mesi di non occupazione e mesi di ricerca di occupazione).

(b) La distorsione sembrerebbe massima in corrispondenza delle modalità contraddistinte da un numero molto basso di risposte. Tale risultato, d'altronde presumibile, suggerisce l'opportunità di determinare una soglia nella frequenza media delle risposte, allo scopo di eliminare dall'analisi le modalità contraddistinte da valori non significativi della distorsione.

(c) Prendendo in considerazione le modalità di risposta con frequenza media di sezione superiore alle 50 unità, l'entità della distorsione sembrerebbe massima proprio in corrispondenza delle due modalità della mutabile condizione professionale attuale: "in cerca di prima occupazione" (s.q.m. pari a 22,7) e "in cerca di nuova occupazione" (s.q.m. pari a 16,5). Avendo riguardo alla modalità "in cerca di prima occupazione", i valori dell'indice variano in corrispondenza del numero di volte che gli individui sono intervistati da 70,19 a 95,54, a 110,28, a 122,73. Per quanto riguarda la modalità "in cerca di nuova occupazione", i valori dell'indice variano da 75,73 a 104,03, a 106,07, a 113,20. La variabilità di tali valori e la presenza di un verso, nel caso crescente, nel loro andamento, suggeriscono la presenza della distorsione. Ciò conferma l'opportunità di svolgere una specifica analisi su tali informazioni, anche allo scopo di verificare se trovano conferma per l'indagine italiana i risultati di analisi condotte su indagini sulle forze lavoro in altri Paesi.

Allo scopo di valutare più accuratamente l'eventuale sistematicità della distorsione nelle modalità attinenti alla "prima" e alla "nuova" ricerca di lavoro, il confronto tra la distorsione nelle risposte fornite dagli occupati e da coloro in cerca di occupazione è istituito in tutte e quattro le occasioni di indagine (1985.I, 1985.II, 1986.I e 1986.II). Inoltre, per poter disporre di una numerosità delle risposte tale da rendere significativo il confronto tra classi di intervistati, si è considerato anche il totale delle informazioni per le quattro occasioni di indagine (in tal modo il numero medio di risposte per sezione di coloro che si dichiarano in cerca di occupazione diventa pari a 481,3). Così, le risposte fornite dagli individui presenti, ad esempio, per la prima volta nel campione sono state ottenute come somma di quelle dei presenti per la prima volta in ciascuna delle 4 indagini. I risultati salienti sono nella Tab. 2.

Tab. 2: *Distribuzione degli indici di distorsione secondo la sezione e classi di intervistati secondo la posizione nella professione. Veneto: diverse occasione di indagine*

	Sezione (n. volte nel campione)				s.q.m.	frequenza media
	1	2	3	4		
* Occupati						
1985.I	101,87	97,49	100,75	99,90	1,9	1375,5
1985.II	101,19	98,63	100,86	99,33	1,2	2025,8
1986.I	98,33	99,13	101,25	101,21	1,5	1396,0
1986.II	102,98	96,96	100,78	99,29	2,5	1377,5
totale 4 occasioni	101,16	98,17	100,89	99,79	1,4	6174,8
* in cerca di occupazione						
1985.I	94,81	111,83	103,20	90,13	9,6	126,5
1985.II	94,90	111,41	102,78	90,76	9,1	126,5
1986.I	72,50	99,09	108,52	118,75	19,9	123,8
1986.II	98,97	87,81	92,91	119,40	13,9	104,5
totale 4 occasioni	90,09	103,35	102,28	104,14	6,6	481,3
di cui in cerca di nuova occupazione						
1985.I	111,94	116,89	82,13	88,90	17,0	56,3
1985.II	112,04	116,45	81,79	89,52	16,9	56,3
1986.I	75,73	104,03	106,07	113,20	16,5	51,8
1986.II	105,71	91,11	93,70	109,07	8,8	47,0
totale 4 occasioni	101,90	108,01	90,56	99,59	7,2	211,3
di cui in cerca di prima occupazione						
1985.I	81,09	107,78	120,08	91,11	17,3	70,3
1985.II	81,17	107,37	119,59	91,75	16,9	70,3
1986.I	70,19	95,54	110,28	122,73	22,7	72,0
1986.II	93,46	85,11	92,26	127,85	19,2	57,5
totale 4 occasioni	80,85	99,71	111,45	107,69	13,6	270,0

Le risposte fornite dagli occupati non sembrano affette da distorsione. Questa evidenza emerge in tutte le occasioni di indagine, così come per le varie classi degli intervistati. Infatti, prendendo in considerazione il sesso, la dimensione demografica del comune di residenza, il titolo di studio e l'età degli intervistati "occupati" nelle varie occasioni, il valore della dispersione degli indici appare assai ridotto.

Di contro, in tutte le occasioni e per le varie classi di intervistati considerate, le risposte di coloro che hanno dichiarato di essere in cerca di occupazione risultano indiziate di distorsione. Tuttavia, considerando il totale degli intervistati, la distorsione sembrerebbe assumere diversa entità e verso nelle quattro occasioni considerate. Questa risulta maggiore nelle due occasioni del 1986 (s.q.m. pari a 19,9 e 13,9) rispetto alle due del 1985 (s.q.m. pari a 9,6 e 9,1).

Tali differenze nei risultati alle varie occasioni inducono a sospendere il giudizio sulla presenza di distorsione e ad affidarne la verifica ad un test specifico, allo scopo di valutare se tali differenze nelle risposte degli intervistati siano di natura casuale o sistematica.

4. *Valutazione della presenza della distorsione nelle stime degli individui in cerca di occupazione indotta dalla rotazione campionaria: analisi inferenziale*

Allo scopo di verificare l'esistenza di distorsione nelle risposte fornite da individui presenti per un diverso numero di volte nel campione è stato considerato l'indice Chi Quadrato[1]. Questo è stato calcolato su tabelle a doppia entrata, nelle quali sulle righe compare la modalità occupato o in cerca di occupazione (due modalità) e sulle colonne il numero di volte nel quale gli individui sono presenti nel campione. Ciò ha permesso l'applicazione del test necessario alla verifica dell'ipotesi di indipendenza tra le modalità di risposta relative alla condizione professionale e il numero di volte in cui gli individui sono presenti nel campione.

Allo scopo di verificare l'esistenza della distorsione non soltanto sul complesso degli individui ma su diverse specifiche classi di intervistati, le risposte fornite dagli occupati e da coloro che si sono dichiarati in cerca di occupazione sono state analizzate anche secondo il sesso, l'età, il titolo di studio, lo stato civile.

Si ricorda che a differenza della prima fase (sez. 3), sono stati considerati i dati sia della regione Veneto che della regione Lombardia. Inoltre, come già detto, l'analisi è stata condotta rigorosamente su individui abbinati il pertinente numero di volte, non già su sezioni di rotazione. L'operazione di *matching* è stata effettuata prendendo in considerazione gli individui presenti

---

1 Indicando con n_{ij} e $\bar{n}_{ij}$, rispettivamente le frequenze effettive e quelle teoriche associate agli intervistati in i-esima condizione professionale e appartenti alla j-esima sezione del campione, il Chi Quadrato risulta $\Sigma_i \Sigma_j \frac{(n_{ij} - \bar{n}_{ij}^2)}{\bar{n}_{ij}}$.

nelle prime due occasioni del 1985 e nelle prime due occasioni del 1986. Pertanto, i controlli che si sono potuti effettuare sono quelli evidenziati nella Fig. 3.

Si noti che, per una parte degli individui, si è potuto effettuare l'abbinamento completo su 4 occasioni, ovvero estendere il controllo agli individui presenti per la 4ª volta nell'occasione 1986.II che erano presenti per la 3ª volta nell'occasione 1986.I, per la 2ª nell'occasione 1985.II e per la 1ª volta nell'occasione 1985.I.

Molto più numerosi sono naturalmente i confronti che è stato possibile condurre tra sezioni di individui abbinati in diverse coppie di occasioni di indagine.

Per le due coppie di occasioni contigue, quella 1985.I-II e quella 1986.I-II il confronto è stato realizzato tra due coppie di sezioni di individui: quelli presenti per la 1ª e 2ª volta e per quelli presenti per la 3ª e 4ª volta.

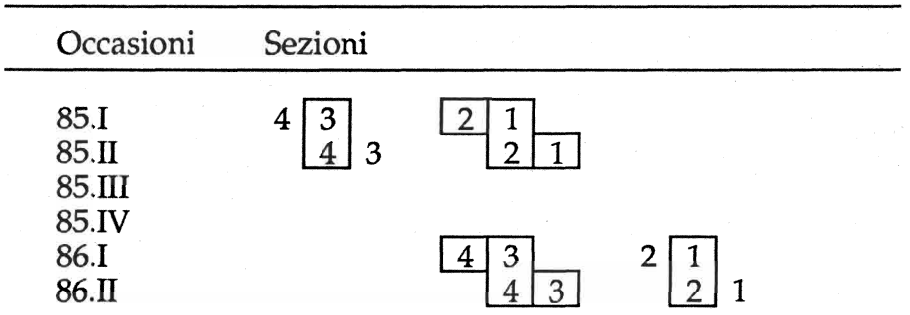
Anche per le due coppie di occasioni a distanza di un anno, quella 1985.I-1986.I e quella 1985.II-1986.II, il confronto è stato realizzato tra due coppie di sezioni di individui: quelli presenti per la 2ª e 4ª volta e per quelli presenti per la 1ª e 3ª volta.

Prendendo in considerazione la coppia di indagini 1985.II-1986.I, il confronto è stato realizzato tra una coppia di sezioni di individui: quelli presenti per la 2ª e 3ª volta. Si tratta di un controllo compreso in quello completo sulla sequenza di 4 interviste.

Infine, prendendo in considerazione la coppia di indagini più distanti, quella 1985.I-1986.II, il confronto è stato realizzato tra una coppia di sezioni di individui: quelli presenti per la 1ª e 4ª volta. Anche tale controllo è compreso in quello completo.

Vi è da precisare che il *matching* effettuato assicura soltanto che gli individui siano stati intervistati in un numero di volte corrispondente alla sezione di appartenenza. Infatti, osservando la Fig. 3 nel senso delle colonne

Fig. 3: *Controlli sulle presenza di distorsione da rotazione effettuati a seguito di abbinamento di dati individuali. Lombardia e Veneto, 1985.I e 1986.I e II*



(le sezioni), se per gli individui presenti per la prima volta nel campione non sussistono problemi, per quelli inseriti nella sezione 2 occorre effettuare il controllo che siano stati presenti nell'occasione precedente. Il *matching* effettuato sulle quattro occasioni 1985.I e II e 1986.I e II ha consentito di operare tale controllo per le occasioni 1985.II e 1986.II, ma non per il 1985.I e 1986.I. (Infatti, non si sono prese in considerazione le due occasioni 1984.IV e 1985.IV che sarebbero state necessarie a questo scopo.) Analogamente, per gli individui inseriti nella sezione 3 occorre effettuare il controllo che siano stati presenti in due occasioni precedenti. Il *matching* effettuato ha consentito di operare tale controllo solo per l'occasione 1986.I. Tuttavia, si noti che nell'occasione 1986.II tale *matching* è stato operato parzialmente, soltanto su una delle due occasioni precedenti (la prima). Per gli individui inseriti nella sezione 4, infine, occorre effettuare il controllo che siano stati presenti in tre occasioni precedenti. Il *matching* effettuato ha consentito di operare tale controllo solo per l'occasione 1986.II. Nelle occasioni 1985.II e 1986.II il *matching* è stato operato parzialmente, soltanto su una delle tre occasioni precedenti (rispettivamente la terza e la seconda).

Allo scopo di effettuare il confronto tra le risposte fornite da intervistati presenti nell'indagine da un numero diverso di volte, si è deciso di considerare in ciascuna occasione le sezioni di famiglie su cui fosse stato operato almeno un controllo a monte o a valle dell'occasione stessa. Nell'ipotesi considerata i confronti effettuati nelle diverse occasioni risultano:
1985.I: 1^a-2^a-3^a;
1985.II: 1^a-2^a-4^a;
1986.I: 1^a-3^a-4^a;
1986.II: 2^a-3^a-4^a [2].

Si noti che in nessuna occasione è stato possibile considerare un confronto tra tutte e quattro le sezioni presenti nel campione. Infatti, in ogni occasione è stato possibile considerare una terna di sezioni. Complessivamente, risultano presenti, una volta, tutte le quattro possibili terne di sezioni.

Si noti ancora che prendendo in considerazione coppie di sezioni, è stato possibile instaurare i seguenti confronti:
1985.I: 1^a-2^a; 1^a-3^a; 2^a-3^a;
1985.II: 1^a-2^a; 1^a-4^a; 2^a-4^a;
1986.I: 1^a-3^a; 1^a-4^a; 3^a-4^a;
1986.II: 2^a-3^a; 2^a-4^a; 3^a-4^a.

In sintesi si tratta di tratta di 12 confronti, due per ciascuna delle sei possibili coppie di sezioni (1-2; 1-3; 1-4; 2-3; 2-4; 3-4).

Al livello di significatività del 5%, ovvero ammettendo tale probabilità di errore nel test, l'ipotesi di base di indipendenza tra condizione professionale e numero di volte in cui gli intervistati sono presenti nel campione non può essere rigettata. In particolare la stima della condizione professionale non risulta distorta dalla rotazione campionaria né per il complesso degli individui,

---

2 Qualora si fossero considerate soltanto le sezioni su cui fosse stato operato l'intero controllo, il *matching* operato avrebbe portato a considerare soltanto i confronti: 1985.II: 1^a-2^a; 1986.I: 1^a-3^a; 1986.II: 1^a-2^a.

né per alcuna delle specifiche classi di individui prese in considerazione.

Lo scarso numero di volte in cui gli individui permangono nel campione (solo quattro) e la distanza temporale tra le interviste (pari a 3 mesi tra le prime due e le ultime due occasioni e pari a 9 mesi tra la seconda e la terza occasione) sembrano pertanto modalità di conduzione dell'indagine che assicurano non si verifichino distorsioni nelle stime indotte dalla prolungata permanenza degli stessi individui nel campione.

10

LA DURATA RIPORTATA DELLA DISOCCUPAZIONE: UN'ANALISI DI ACCURATEZZA

Nicola Torelli

1. *Adeguatezza e accuratezza della durata riportata della disoccupazione: riflessioni introduttive*

Uno dei principali obiettivi della rilevazione trimestrale delle forze di lavoro (nel seguito RTFL) è quello di documentare in modo completo ed articolato il livello e la dinamica della disoccupazione in Italia. A tale scopo, hanno notevole rilievo le informazioni sulla durata della ricerca di lavoro. Esse consentono di arricchire il quadro delle conoscenze sul fenomeno, sia in chiave descrittiva sia in chiave di modelli di comportamento dinamico, a livello micro, dell'offerta di lavoro.

Nella RTFL, come d'altronde nelle principali indagini sulle forze di lavoro svolte nei Paesi sviluppati, vengono correntemente raccolti dati sulla durata della ricerca di occupazione[1]. Tipicamente, tali informazioni vengono rilevate con un quesito retrospettivo. In particolare, nella rilevazione italiana ad ogni individuo del campione che, alla data dell'indagine, risulti disoccupato viene chiesto: "Da quanti mesi è alla ricerca di occupazione?"[2].

Sia che si vogliano impiegare le informazioni risultanti per la descrizione del fenomeno della disoccupazione, sia che ci si proponga di utilizzarle nel contesto di modelli di durata, si pongono riguardo ad esse problemi di (i) adeguatezza e di (ii) accuratezza.

La prima delle due questioni enunciate, l'*adeguatezza*, ha a che vedere con la capacità delle durate rilevate, supposte esenti da errori di misura, di

---

1 Va notato che con la RTFL viene rilevata anche un'altra informazione retrospettiva sulla durata: si tratta, per i non occupati già occupati, del periodo trascorso dalla conclusione dell'ultima occupazione. Palesemente, l'informazione è rilevata per un aggregato diverso da quello al quale si richiede di riportare la durata della ricerca di lavoro. Inoltre, tale informazione ha un rilievo differente (e, in prima approssimazione, meno chiaro) per l'analisi della dinamica della disoccupazione e della transizione all'occupazione, perché riguarda genericamente un periodo di non lavoro, che può quindi comprendere anche sub-periodi in cui non è stata svolta alcuna ricerca di occupazione (o addirittura coincidere *tout court* con un periodo di inattività).

2 Si impone una precisazione. Il quesito sulla durata della ricerca di lavoro è in realtà posto a tutti coloro che hanno dichiarato di essere alla ricerca (quindi, ad esempio, anche agli occupati in cerca di un diverso o ulteriore lavoro). Nel seguito, restringiamo tuttavia l'attenzione sui soli disoccupati, sicché nel contesto di questo studio è equivalente parlare di 'disoccupazione' o di 'ricerca di lavoro'.

rappresentare in maniera appropriata le durate degli episodi completi di disoccupazione. La questione è legata essenzialmente al disegno dell'indagine sulle forze di lavoro, disegno che comporta due fenomeni di inadeguatezza: la censura a destra e il cosiddetto *length bias*. La censura a destra è l'ovvia risultante del fatto che l'informazione riportata riguarda la durata dell'episodio ancora in corso (il pertinente quesito è, per l'appunto, posto a chi è alla ricerca di lavoro), e non già la durata completa dell'episodio di disoccupazione. Inoltre, il fatto che la durata della disoccupazione sia rilevata soltanto per gli episodi in corso alla data dell'indagine implica che verosimilmente i dati sono affetti da *length bias*: il fenomeno si manifesta perché, detto semplicemente, nella maggior parte delle situazioni episodi di disoccupazione più lunghi hanno maggiore probabilità di essere interrotti, e quindi osservati, da un piano di osservazione puntuale.

Il problema della censura a destra può essere in parte alleviato sfruttando il disegno campionario con rotazione della RTFL (vedi il cap. 1), che consente, per una parte del campione, di osservare le risposte date dagli stessi soggetti in successive occasioni di indagine e quindi di osservare anche episodi di disoccupazione completi (relativi cioè a coloro che, essendo alla ricerca di lavoro ad una occasione di indagine, alla successiva dichiarano di essere occupati o di avere smesso di cercare lavoro). In tal caso, tuttavia, la durata dell'episodio di disoccupazione completo è nota con approssimazione, perché della sua conclusione si sa soltanto che è avvenuta nel periodo compreso fra le due rilevazioni[3]. I problemi connessi con la stima di modelli di durata della disoccupazione muovendo da un campione di dati affetti da inadeguatezze, cioè a dire censurati a destra (o approssimati) e *length biased*, hanno trovato una prima trattazione sistematica in Salant (1977) e sono affrontati nel cap. 19.

La seconda questione enunciata, l'*accuratezza*, ha a che vedere con imprecisioni ed errori di misura della durata riportata, connesse essenzialmente alla modalità con cui la si rileva, modalità costituita, come si è detto, da un quesito retrospettivo. Il tema presenta almeno due ramificazioni. Un primo aspetto riguarda la grossolanità della durata riportata, in termini di unità di tempo con cui la si misura (nella RTFL, lo si è appena ricordato, è il mese): si tratta essenzialmente di un problema di grado, cioè a dire di precisione dell'unità di misura rispetto al campo di variazione dei dati. Il secondo, e più importante, aspetto attiene all'accuratezza della durata riportata, in risposta ad un quesito retrospettivo: l'informazione è verosimilmente affetta da errori connessi al processo di memoria.

Recuperando la terminologia classica (Sudman e Brandburn, 1973) ed estendendola convenientemente, le inaccuratezze nei dati su eventi (o durate) rilevati mediante quesiti retrospettivi si possono ricondurre:

(a) all'effetto memoria, che conduce a non riportare uno o più eventi poichè il rispondente li ha dimenticati;

---

3 Per una più generale trattazione delle diverse questioni derivanti da uno schema di osservazione caratterizzato dalla combinazione di quesiti retrospettivi e di un campione con rotazione, vedi Trivellato e Torelli (1989).

(b) all'effetto telescopio, che ha luogo quando c'è memoria dell'evento, ma si ha difficoltà a collocare lo stesso nel tempo. Usualmente, si parla di effetto telescopio in avanti quando l'evento viene collocato in una data più prossima a quella dell'indagine, rispetto a quella vera, e di effetto telescopio all'indietro se l'evento è collocato in un istante più remoto;
(c) all'effetto *heaping*, o ammucchiamento, che induce una abnorme polarizzazione delle risposte su alcune particolari date (nel caso di collocazione di eventi) ovvero su alcune particolari durate (nel caso di tempi di permanenza in uno stato)[4].

È appunto sull'accuratezza della durata riportata della disoccupazione, e sugli errori del processo di memoria che possono intaccarla, che si incentra l'attenzione in questo capitolo. L'analisi, conviene precisarlo subito, si avvale dei soli dati sulla durata riportata tratti dalla RTFL. Essa consente quindi valutazioni interne di accuratezza, e non i più stringenti giudizi traibili da studi di validazione basati sul confronto con fonti esterne altamente attendibili (per esempi in tal senso, vedi Mathiowetz e Duncan, 1988, e Duncan e Hill, 1989). Come si vedrà, le evidenze che emergono sono tuttavia di non poco interesse, e illuminanti circa alcuni nitidi *patterns* di inaccuratezza presenti nei dati. Ancora, v'è da aggiungere che l'analisi è condotta secondo un'ottica essenzialmente descrittivo-esplorativa. Questa scelta ha una semplice motivazione. Modelli statistici degli effetti del processo di memoria si ritrovano certo in letteratura (vedi, tra gli altri, Sikkel, 1985, e von Dosselaar, van den Hurk e Israël, 1989), ma sono fortemente stilizzati, e soprattutto orientati alla trattazione del ricordo del numero di eventi. Quando l'informazione di interesse è la durata riportata di un episodio, le indicazioni della psicologia cognitiva e della metodologia della ricerca sulle caratteristiche del processo di memoria si fanno più sfumate ed incerte (Bernard, Killworth, Kronenfeld e Sailer, 1984), e tali modelli richiedono verosimilmente apprezzabili modificazioni ed affinamenti, che non possono non basarsi sui riscontri di attente analisi descrittivo-esplorative.

Per l'analisi empirica dell'accuratezza si utilizzano, in modo combinato, le informazioni sulla durata riportata tratte da una singola rilevazione e gli analoghi dati abbinati relativi a coloro che partecipano a due rilevazioni successive distanziate di un trimestre. In sostanza, i dati rilevati in una singola occasione si prestano per fornire evidenze in merito alla presenza, ed alle caratteristiche, dell'effetto *heaping*. Sfruttando opportunamente le risposte date dagli stessi soggetti in due successive occasioni, e in particolare guardando al cambiamento nella durata riportata della ricerca di lavoro da una indagine alla successiva, si possono avere indicazioni sulla coerenza delle risposte, e indirettamente sull'effetto telescopio. Qualche ulteriore chiarimento sui fenomeni di inaccuratezza può infine venire ponendo in relazione le evidenze fornite dalle due precedenti linee di analisi.

---

4 Palesemente, l'effetto *heaping* può essere interpretato come una particolare manifestazione dell'effetto telescopio. Come si vedrà nel seguito, esso assume peraltro un rilievo particolare nei dati sulla durata della disoccupazione della RTFL. E' quindi conveniente considerarlo come un distinto fenomeno di *recall bias*.

Il capitolo è organizzato come segue. Dopo un breve esame delle caratteristiche del campione di dati utilizzati, volto essenzialmente a vagliare l'incidenza di eventuali altre fonti di errore connesse al disegno della RTFL (sez. 2), si analizza l'effetto *heaping* nella durata riportata della disoccupazione, sfruttando le informazioni che vengono da una singola indagine (sez. 3). Successivamente, utilizzando i dati riguardanti gli individui abbinati che risultano essere continuativamente in cerca di lavoro in due successive occasioni di indagine, si esamina la coerenza delle risposte fornite riguardo al medesimo episodio di disoccupazione, e la si mette in relazione con quanto emerso dalla precedente analisi sull'effetto *heaping* (sez. 4). La conclusiva sez. 5 contiene brevi considerazioni di sintesi e prospetta alcune linee di ricerca per approfondire lo studio dell'accuratezza di risposte a quesiti retrospettivi, in vista soprattutto di un oculato impiego dei dati sulla durata riportata della disoccupazione per analisi descrittive e in modelli di analisi dinamica dell'offerta di lavoro.

2. *Dati e informazioni usate*

I dati utilizzati sono quelli della RTFL per le regioni Lombardia e Veneto nelle due occasioni di indagine 1986.I e 1986.II, relativi alle persone disoccupate alla prima occasione. Nelle due regioni, il campione di disoccupati dell'indagine di gennaio 1986 consta di 2.930 individui. Per gran parte delle analisi, il campione di interesse è peraltro costituito dai disoccupati abbinati che nelle due occasioni di indagine risultano continuativamente alla ricerca di lavoro, campione che comprende 746 individui. E' infatti per questi ultimi che è ragionevole procedere al confronto delle risposte fornite in due successive rilevazioni. Ed è quindi ancora per questo sottoinsieme che conviene procedere all'analisi delle risposte fornite alla prima occasione, per poter poi porre in relazione le evidenze di inaccuratezza che emergono dalle due linee di analisi.

L'identificazione del campione di interesse avviene in due passi:

(a) in primo luogo, tramite la procedura di abbinamento dei dati individuali illustrata nel cap. 7, si seleziona il campione dei disoccupati (a gennaio 1986) che partecipano anche all'indagine di aprile 1986;
(b) in secondo luogo, nell'ambito di questi si individuano i disoccupati continuativamente alla ricerca di lavoro. Sono considerati tali coloro che, alla seconda occasione d'indagine: (b1) sono disoccupati; (b2) riportano una durata della ricerca di lavoro pari a quattro mesi o più[5].

Va notato che le informazioni di cui si dispone non sempre consentono di classificare con certezza gli individui come continuativamente alla ricerca.

---

5 Il restante sottoinsieme di disoccupati abbinati è evidentemente costituito da individui con episodio di disoccupazione completo. Ciò si da per quanti transitano all'occupazione o all'inattività nell'intervallo fra le due indagini, cioé a dire: coloro che hanno dichiarato di essere occupati o inattivi alla seconda occasione; coloro che, pur essendo disoccupati alla seconda occasione, dichiarano di essere in cerca di lavoro da meno di quattro mesi.

Due specifici fattori connessi al quesito retrospettivo sulla durata della ricerca possono, infatti, interferire nella corretta identificazione dell'aggregato di interesse. (i) Come si è detto, vengono esclusi dal campione dei disoccupati continuativamente alla ricerca coloro che alla seconda occasione riportano una durata della ricerca inferiore ai quattro mesi. Ora, ciò può essere dovuto all'uscita e al successivo rientro nella condizione di disoccupazione entro il trimestre che separa le due indagini, e in tal caso l'esclusione dall'aggregato di interesse è appropriata. Alternativamente, si può essere in presenza di una errata risposta alla seconda occasione, dovuta ad esempio ad un effetto telescopio in avanti, e in tal caso l'esclusione potrebbe risultare erronea. La distinzione tra le due situazioni non è possibile sulla base dei dati disponibili, e non si può quindi escludere l'evenienza di qualche errore del tipo indicato. (ii) Sono inoltre esclusi dal campione di interesse i disoccupati con durata della ricerca, alla prima occasione, superiore ai 95 mesi. È questa la conseguenza del fatto che la codifica prevista per la durata riportata è su due cifre, quindi con un massimo di 99 mesi, sicchè per questi disoccupati non sarebbe possibile l'analisi della coerenza fra le risposte fornite alle due indagini. Si tratta, peraltro, di un numero esiguo di casi: episodi di disoccupazione cronica sono, fortunatamente, molto rari.

Nel complesso, queste limitazioni nell'identificazione del campione di interesse appaiono di scarso rilievo. Piuttosto, occorre interrogarsi sulla ragionevolezza e la generalità delle conclusioni che si possono trarre dall'analisi di un campione così circoscritto. L'interrogativo ha perlomeno due risvolti. Che livello di generalità si può attribuire ad evidenze relative ai disoccupati di due sole regioni, e limitate al periodo gennnaio-aprile 1986? È ragionevole ritenere che i *patterns* di inaccuratezza nella durata riportata della disoccupazione dipendano dal processo di memoria (e dalla modalità di interrogazione), oppure essi sono inquinati da altri fattori connessi al disegno dell'indagine e ai criteri di selezione dell'aggregato di interesse?

L'insieme dei dati è palesemente circoscritto quanto a dimensione territoriale e temporale. Sommarie analisi su dati di altre regioni e relative ad altre occasioni di indagine hanno mostrato, tuttavia, come i fenomeni di inaccuratezza siano presenti secondo modalità sostanzialmente invariate. Più che all'incidenza ed alle caratteristiche con cui si manifesta la disoccupazione, essi possono quindi essere attribuiti agli effetti del processo di memoria a fronte di quesiti retrospettivi.Qualche considerazione più circostanziata va spesa in merito alla possibilità che l'insieme dei disoccupati abbinati continuativamente alla ricerca di lavoro presenti caratteristiche marcatamente diverse dall'aggregato più ampio da cui proviene - il totale dei disoccupati alla prima occasione -. In sostanza, si tratta di accertare che il campione di interesse non risenta, o comunque risenta in misura trascurabile, di fenomeni di selezione. Un'eventuale distorsione associata alla selezione può essere ricondotta a tre possibili cause: (i) il disegno campionario con rotazione; (ii) la procedura di abbinamento; (iii) errori di classificazione nella condizione rispetto al lavoro. Vediamo di valutare brevemente, per ora in via congetturale e indiretta, il possibile impatto di ciascuna di queste cause.

Innanzitutto, è ragionevole attendersi che l'iniziale limitazione al sottoin-

sieme di disoccupati di due delle quattro sezioni di rotazione, quelle per le quali è prevista la reintervista nel trimestre successivo, non conduca a distorsioni apprezzabili. Infatti, per ogni singola rilevazione ciascuna sezione è un sottocampione casuale del campione complessivo di famiglie. Qualche dubbio potrebbe eventualmente sorgere in presenza di un effetto di *rotation group bias*, ma i risultati del cap. 9 portano ad escludere questa eventualità: per i dati della RTFL tale effetto risulta trascurabile.

Più problematico è escludere che vi sia distorsione da selezione dinamica, conseguente all'abbinamento. Tale fenomeno si riscontra quando i soggetti che permangono nel campione di abbinati nelle due occasioni di indagine differiscono sistematicamente dai soggetti non abbinati. Il mancato abbinamento è dovuto essenzialmente a due motivi: l'incapacità della RTFL, per le caratteristiche stesse del suo impianto, di seguire gli individui che si staccano dalla famiglia o le famiglie che cambiano residenza[6]; un *pattern* di selezione nella procedura di abbinamento. Ora, i riscontri del cap. 7 paiono escludere che la procedura di abbinamento sia affetta da apprezzabili fenomeni selettivi. Verosimilmente, è quindi il primo motivo ad essere responsabile di una selezione non casuale del campione. Essa dovrebbe comunque essere di dimensioni contenute, per la mobilità relativamente poco marcata che caratterizza ormai la popolazione italiana e per il breve intervallo temporale fra le due occasioni di indagine.

Infine, la possibilità di distinguere correttamente episodi di disoccupazione ininterrotti e completi è condizionata dalla capacità di classificare senza errore ogni individuo entro uno degli stati - occupato, disoccupato, inattivo -. Gli errori di classificazione possono risultare da errori di rilevazione e/o registrazione e/o codifica e/o imputazione, dalla presenza di *rotation group bias*, da errori della procedura di abbinamento (segnatamente, da errati abbinamenti, ai quali è facilmente associato un cambiamento spurio). Tipicamente, essi conducono alla rilevazione di cambiamenti di stato in realtà mai avvenuti, quindi ad una distorsione verso il basso delle misure di durata della disoccupazione: in altre parole, sulla base del confronto di risposte fornite a due successive indagini, un episodio di disoccupazione può apparire completo quando in realtà esso è ancora in corso. Con riguardo ai dati della *Current Population Survey* statunitense, Poterba e Summers (1986) presentano interessanti evidenze empiriche in merito all'effetto di errori di classificazione sulla stima della mobilità fra stati occupazionali. Sfortunatamente, per i dati della RTFL poco è noto circa la presenza e l'impatto dei fattori che possono indurre errori di classificazione nello stato occupazionale. Non esistono, infatti, complete e rigorose analisi su tali temi, e la stessa possibilità di svolgerle è preclusa dalla mediocre documentazione sullo svolgimento dell'indagine, che non consente di discernere il ruolo dei diversi fattori nel

---

6 Com'è noto, le unità di ultimo stadio della RTFL sono le famiglie anagrafiche, operativamente individuate con l'indirizzo di residenza. Per la parte del campione soggetta a reintervista, non è previsto alcun tentativo per seguire individui o famiglie che si muovono: gli individui, semplicemente, si perdono; le famiglie sono sostituite con altre, tratte da una lista supplementare.

processo di produzione dei dati finali. I parziali risultati presentati nei capp. 7, 8 e 9 permettono peraltro di escludere almeno alcune delle potenziali cause di distorsione citate (in particolare, come già accennato non c'è evidenza di *rotation group bias*).

3. *Distribuzione di durata ed effetto ammucchiamento*

Prime indicazioni sull'accuratezza si possono avere esaminando la distribuzione di frequenza delle durate di ricerca di lavoro per i disoccupati, ottenute ad una singola rilevazione.

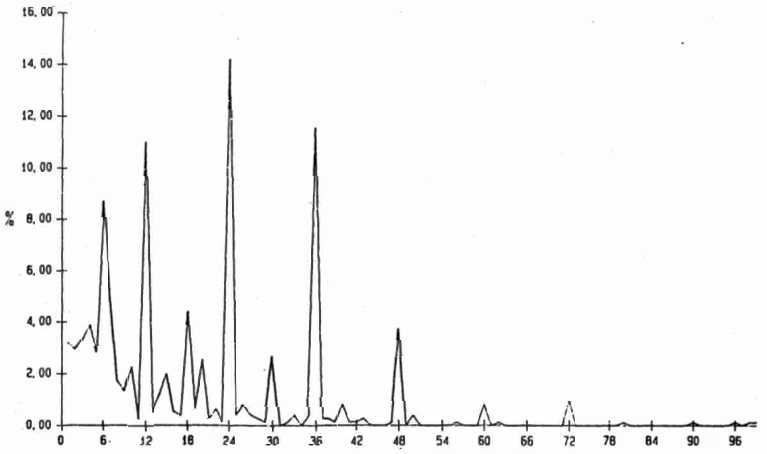
La Fig. 1 riporta il grafico della distribuzione, per i dati della RTFL di gennaio 1986, relativo all'insieme dei disoccupati abbinati nelle due indagini 1986.I-II che risultano continuativamente alla ricerca. E' immediato osservare come vi sia una forte polarizzazione delle risposte su durate pari a 12 mesi e ai suoi multipli, e inoltre, seppure in modo meno marcato, sulle durate di 6 mesi, 18(=6+12) mesi, e via dicendo. La concentrazione delle durate riportate su alcuni valori caratteristici è palesemente abnorme. Le attese, infatti, sarebbero per una distribuzione delle durate essenzialmente regolare, coerentemente con i tipici andamenti osservati per dati della stessa natura rilevati con differenti modalità (ad es., con registrazioni di tipo amministrativo). I picchi della distribuzione lasciano dunque spazio a ragionevoli sospetti riguardo all'accuratezza delle misure di durata.

Occorre considerare, tuttavia, che le cause di un siffatto fenomeno di ammucchiamento possono essere molteplici. Vi possono concorrere, infatti:

(a) l'effetto *heaping* propriamente inteso, cioè la tendenza del processo di memoria ad arrotondare l'effettiva durata ad una cifra intera (l'anno o il semestre, nel caso della RTFL);
(b) altri fattori di distorsione connessi al disegno dell'indagine ed ai criteri di selezione dell'aggregato;
(c) fattori, che possiamo sinteticamente designare col termine di 'ammucchiamento strutturale', i quali determinano una reale polarizzazione delle durate su alcuni valori. Ciò vale, ad esempio, quando vi sia una marcata stagionalità negli ingressi nel mercato del lavoro. Un caso ancor più emblematico è quello di politiche di *welfare* le quali prevedano l'erogazione di un'indennità di disoccupazione per un numero massimo di mesi: conseguentemente, in accordo con l'atteso comportamento individuale, gran numero di transizioni all'occupazione si concentra alla scadenza del periodo coperto dall'indennità (vedi, ad es., Han e Hausman, 1990).

Per i dati della RTFL, peraltro, la causa dominante, se non l'unica, appare essere la tendenza ad arrotondare la durata vera, riportandola approssimata in anni e semestri, come conseguenza di inaccuratezze del processo di memoria a fronte di un quesito retrospettivo. Per l'Italia, e in particolare per le due regioni considerate, non vi sono infatti ragioni per attendersi apprezzabili fenomeni di ammucchiamento strutturale: l'indennità di disoccupazione è, o comunque nel 1986 era, insignificante; il *pattern* di ammucchiamento documentato dalla Fig. 1 è in ogni caso incoerente con la stagionalità, del

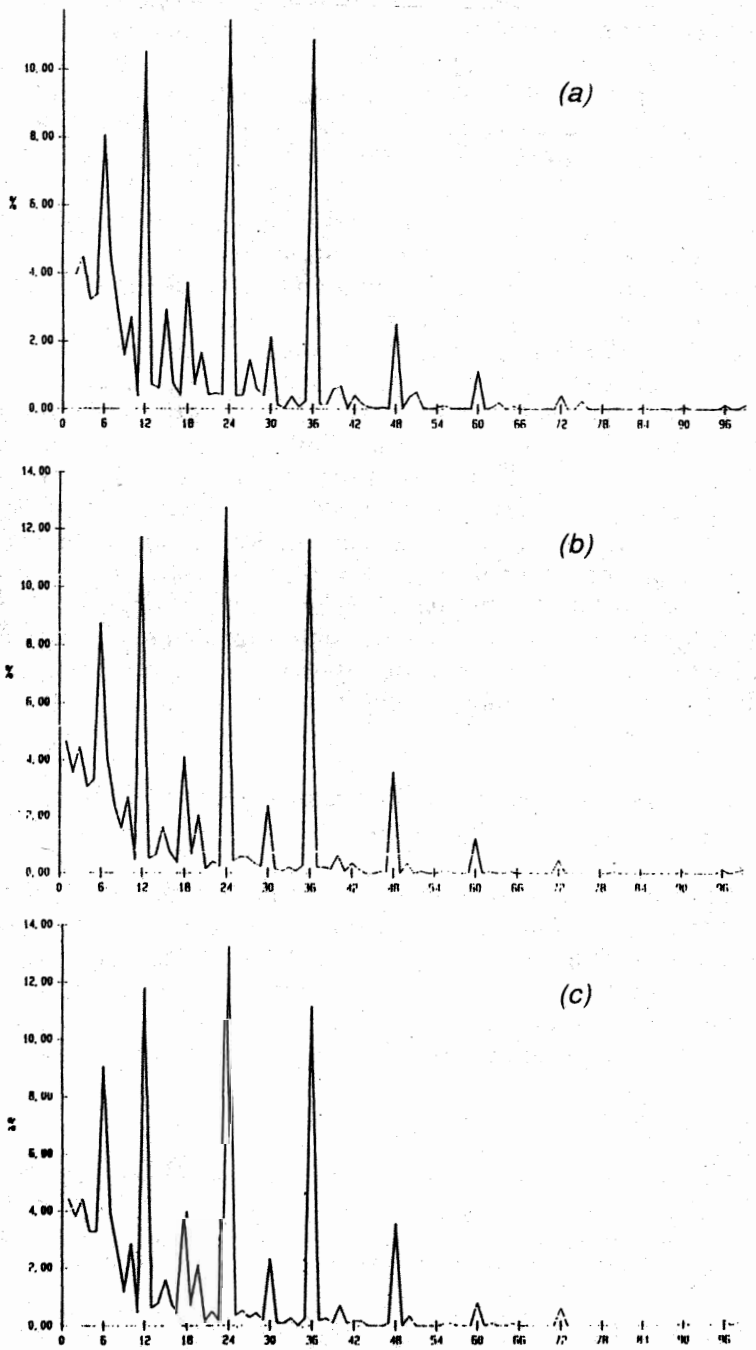
Fig. 1: *Distribuzione dei disoccupati continuativamente alla ricerca di lavoro in due successive indagini secondo la durata della ricerca riportata alla prima occasione. Lombardia e Veneto, 1986. I-II (N=746)*



resto non vistosa, nell'inizio della ricerca di lavoro (vedi il cap. 16). D'altra parte, a conferma delle argomentazioni *a priori* sulla scarsa rilevanza di possibili fattori di distorsione del tipo (b) presentate nella sez. precedente, la Fig. 2 mostra come l'effetto *heaping* si manifesti con caratteristiche del tutto analoghe negli aggregati più ampi costituiti rispettivamente dall'insieme dei disoccupati alla prima occasione di indagine, dall'insieme dei disoccupati appartenenti alle sezioni di rotazione per cui è prevista la reintervista nel trimestre successivo, dall'insieme dei disoccupati abbinati nelle due indagini.

È pertanto ragionevole concludere che l'abnorme polarizzazione delle durate riportate su anni e semestri risultante dalla Fig. 1 chiama in causa essenzialmente l'effetto *heaping*. A ulteriore conferma, si può osservare come la tendenza ad arrotondare risulti più marcata per periodi di ricerca più lunghi. Per durate superiori a 24 mesi le risposte concentrate su un multiplo di 6 sono pari al 69%, mentre scendono al 45% per le durate fino a 24 mesi. Ciò è coerente con l'evidenza, comune negli studi su problemi di memoria in indagini retrospettive, che la tendenza a ricordare un evento declina con la distanza dell'evento dal tempo dell'intervista (Sudman e Brandburn, 1973; Baddeley, 1979; Sikkel, 1985). Sempre in questa chiave, è interessante notare ancora come la quota di dati ammucchiati sia maggiore sulle durate corrispondenti ad anni (multipli di 12) rispetto a quelle corrispondenti ad anni e mezzo (cioè a 6+k12, con k=1,2,3...), e come per tempi di disoccupazione molto lunghi le durate corrispondenti agli anni assorbano la quasi totalità delle risposte. Chiaramente, l'approssimazione nelle risposte si fa più marcata quanto più lontano è l'evento-origine dell'episodio di disoc-

Fig. 2: *Distribuzione dei disoccupati secondo la durata riportata dalla ricerca:*
(a) Lombardia e Veneto, 1986.I, intero campione (N=2.930)
(b) Lombardia e Veneto, 1986.I, sezioni del campione di cui è prevista la reintervista nell'indagine successiva (N=1.412)
(c) Lombardia e Veneto, 1986.I-II, abbinati nelle due indagini, durata riportata alla prima occasione (N=1.128)



cupazione in corso[7].

Naturalmente, *pattern* e intensità con cui l'effetto *heaping* si manifesta dipendono anche dalle modalità con cui avviene l'interrogazione retrospettiva. Ad esempio, è plausibile attendersi che una diversa formulazione del quesito sulla durata della ricerca - non già "Da quanti mesi è alla ricerca di occupazione?", ma "In che anno e mese ha iniziato la ricerca di occupazione?" - comporti un profilo parzialmente diverso dell'effetto *heaping* (nell'esempio in questione, verosimilmente associato a risposte polarizzate su gennaio). Più in generale, è ragionevole attendersi che abbiano particolare rilievo tre caratteristiche del disegno di interrogazione retrospettiva:

(a) il fatto che l'interrogazione sia centrata su episodi ovvero sia basata sul calendario (come nell'indagine suppletiva sulla storia lavorativa presentata nel cap. 22);
(b) il fatto che si richieda di collocare nel tempo eventi (come nell'esempio appena fatto) ovvero di ricordare durate di episodi;
(c) la lunghezza del periodo su cui si estende l'interrogazione, e in particolare il fatto che il periodo di ricordo sia illimitato ovvero limitato.

L'intreccio fra effetti del processo di memoria e disegno dell'indagine retrospettiva è tema in larga parte ancora inesplorato. È comunque degno di nota che, con specifico riguardo alla durata della disoccupazione e/o dell'occupazione, le evidenze di svariati tipi di indagini, tanto del tipo RTFL - cioè centrate sull'episodio, con quesito sulla durata e con periodo di ricordo illimitato -, quanto genuinamente longitudinali e con più sofisticate caratteristiche di interrogazione - cioè basate sul calendario, con quesiti sugli eventi e con periodo di ricordo limitato - (vedi Martini, 1987, e Hill, 1988, sulla statunitense *Survey on Income and Program Participation*), documentino un effetto *heaping* sempre vistoso, anche se differenziato nel profilo. Si è dunque in presenza di un fenomeno di indole generale, che chiama in causa l'influenza del processo di memoria e che merita di essere approfondito.

4. *Coerenza delle risposte a due successive occasioni di indagine, effetto telescopio ed effetto ammucchiamento*

Per i disoccupati continuativamente alla ricerca di lavoro, a partire dal confronto delle risposte fornite in due successive occasioni di indagine è possibile condurre un'analisi della 'coerenza' delle durate riportate. In questo contesto per 'coerenza' si intende la relazione, di accordo o disaccordo, che c'è fra il cambiamento nella durata in mesi riportata nelle due occasioni e il numero di mesi che separa le indagini. Ad esempio, un individuo continua-

---

7 Evidenze aggiuntive sul fatto che la durata riportata della ricerca sia dominata, e con modalità sostanzialmente invariate, dall'effetto *heaping* vengono da altre analisi, qui non riportate per ragioni di spazio (vedi Cielo, 1989, e materiali disponibili presso l'autore). Esse hanno avuto per oggetto: (i) per un verso i non disoccupati con attività di ricerca di lavoro: segnatamente, gli occupati in cerca di un nuovo o ulteriore lavoro e le persone classificate fra non forze di lavoro che pure dichiarano di cercare lavoro; (ii) per un altro verso i diversi sottoinsiemi di disoccupati - già occupati, in cerca di prima occupazione, altri in cerca di lavoro -.

tivamente alla ricerca di lavoro in due successive indagini distanziate di 3 mesi, che alla prima intervista dichiari una durata di ricerca di 5 mesi e che nella seconda intervista dichiari 12 mesi di ricerca, riporta un cambiamento nella durata non coerente.

Seguendo Bowers e Horvath (1984), per ciascun disoccupato continuativamente alla ricerca si può calcolare l'errore nel cambiamento della durata riportata in due indagini successive (ECD), definito come segue:

$$ECD = Durata_t - Durata_{t-1} - 3,$$

dove t-1 e t identificano le due successive occasioni di indagine, e 3 è il numero di mesi che separa le due indagini. Chiaramente, si possono distinguere tre casi: (i) per risposte coerenti, ECD assume il valore 0; (ii) un valore di ECD negativo segnala un errore di sottostima nel cambiamento nella durata riportata; (iii) all'opposto, un valore di ECD positivo indica una sovrastima del cambiamento.

Valutazioni sintetiche su direzione e intensità dell'incoerenza nel riportare il cambiamento possono poi venire dalla distribuzione di ECD, e inoltre dalla considerazione congiunta dell'errore medio e dell'errore medio assoluto (se errore medio e errore medio assoluto sono uguali in modulo, gli errori sono ovviamente tutti nella stessa direzione).

La Tab. 1 riporta la distribuzione in classi di ECD, condizionata alla durata riportata della disoccupazione alla prima occasione. Risalta con nitidezza una relazione negativa fra la durata riportata all'indagine iniziale e la percentuale di risposte non coerenti, che nel complesso non raggiunge il 40%. In particolare, all'aumentare della durata crescono in modo evidente le risposte con ECD negativo, mentre si ha un andamento opposto per le risposte con ECD positivo. Le incoerenze nel senso della sovrastima del cambiamento sono la maggioranza per le durate iniziali brevi, mentre l'indice ECD risulta negativo per ben il 75% delle incoerenze associate a durate elevate alla prima occasione.

Tali risultati sono largamente in accordo con quelli riportati da Bowers e Horvath (1984), che conducono un'analoga analisi sui dati della *Current Population Survey* statunitense. Essi sono, d'altra parte, conformi alle attese suggerite da precedenti studi sugli effetti del processo di memoria per dati retrospettivi. Si ha una conferma che al crescere della distanza temporale fra l'evento-origine e il momento dell'intervista risulta più problematico riportare accuratamente la durata dell'episodio. Inoltre, per coloro che sperimentano un episodio di disoccupazione piuttosto lungo si manifesta un chiaro effetto telescopio in avanti, che porta a collocare l'evento 'inizio della ricerca di lavoro' ad una data progressivamente più prossima a quella dell'indagine.

Ulteriori, convincenti evidenze nell'analisi della coerenza emergono poi quando si pongono in connessione i riscontri sulla distribuzione di ECD con la presenza dell'effetto *heaping*. Guardando alla Tab. 2, dove è riportata la distribuzione di ECD, si hanno chiari segnali del fatto che la tendenza a riportare durate arrotondate sui semestri e gli anni possa essere responsabile

Tab. 1 : *Distribuzione dei disoccupati continuativamente alla ricerca di lavoro in due successive indagini, secondo la coerenza delle risposte sulla durata della ricerca e secondo la durata della ricerca riportata alla prima occassione. Lombardia e Veneto, 1986. I-II* [(a)]

Durata della disoccupazione alla prima indagine	Disoccupati continuativamente alla ricerca					
	Distribuzione delle risposte				Errore medio	Errore medio assoluto
	incoerenti negative	coerenti	incoerenti positive	Totale		
meno di 3 mesi	3 *4,23*	38 *53,52*	30 *42,25*	71	2,79	2,87
da 4 a 6 mesi	20 *17,39*	57 *49,57*	38 *33,04*	115	1,52	2,18
da 7 mesi a un anno	51 *32,08*	64 *40,25*	44 *27,67*	159	1,81	3,90
da 1 a 2 anni	68 *33,01*	78 *37,86*	68 *29,13*	206	-0,29	4,81
oltre i due anni	104 *53,33*	49 *25,13*	42 *21,54*	195	-5,67	10,18
Totale	246 *32,97*	286 *38,34*	214 *28,69*	746	-0,67	5,43

(a) Valori percentuali (per riga) in corsivo.

di una quota non trascurabile delle incoerenze osservate. A tal fine, torna utile fissare l'attenzione sulle due seguenti situazioni:

(a) vi è un arrotondamento alla stessa durata in entrambe le occasioni, il che comporta un valore di ECD pari a -3;
(b) nelle due occasioni la durata viene arrotondata su valori diversi: se l'arrotondamento alla seconda occasione di indagine avviene su durate più lunghe di un semestre o di un anno, ciò implica valori di ECD pari a 3 o 9.

Ora, è interessante notare che questi tre valori di ECD assorbono una quota rilevante del totale delle risposte (circa il 18%), che equivale a quasi un terzo delle risposte incoerenti.

L'esame della Tab. 3 permette di approfondire ed estendere le considerazioni appena abbozzate. In essa sono riportate, per coloro per i quali si registrano valori di ECD diversi da 0, le durate dichiarate nelle due successive occasioni, mettendo in evidenza esclusivamente i valori su cui si concentra l'effetto *heaping* e accorpando le risposte diverse in un unica classe. Oltre il 37% delle incoerenze risulta da coppie di risposte con entrambi i valori riportati pari a 6 mesi o multipli, e delle restanti una quota rilevante registra tali valori per almeno una delle due durate. La frazione di incoerenze asso-

Tab. 2: *Distribuzione dell'errore nel riportare il cambiamento della durata (ECD) in due successive occasioni di indagine. Lombardia e Veneto, 1986.I-II*

ECD (in mesi)	N.	%
< -15	39	5,2
-15	16	2,1
-14	2	0,3
-13	3	0,4
-12	3	0,4
-11	10	1,3
-10	3	0,4
-9	14	1,9
-8	5	0,7
-7	3	0,4
-6	4	0,5
-5	8	1,1
-4	1	0,1
-3	83	11,1
-2	14	1,9
-1	38	5,1
0	286	38,3
1	62	8,3
2	13	1,7
3	25	3,4
4	5	0,7
5	8	1,1
6	10	1,3
7	8	1,1
8	0	0,0
9	26	3,5
10	2	0,3
11	8	1,1
12	1	0,1
13	6	0,8
14	0	0,0
15	7	0,9
> 15	33	4,4
Totale	746	100,0

ciata a durate riportate diverse da 6 mesi o multipli in entrambe le occasioni non raggiunge il 21%.

Una più puntuale conferma delle connessioni fra effetto *heaping* e incoerenze nel cambiamento della durata riportata si ha considerando che, in presenza di incoerenze associate a coppie di durate 'arrotondate', vi è una marcata tendenza a riportare la stessa durata nelle due occasioni (sono i casi sulla diagonale principale della Tab. 3: ben 71 su 173). Questo fenomeno, che porta ad un valore di ECD negativo, specificamente pari a -3, avviene con maggiore frequenza per le durate riportate più lunghe. L'intreccio

Tab. 3: *Distribuzione dei disoccupati continuativamente alla ricerca di lavoro con risposte incoerenti, secondo la durata della ricerca riportata nelle due indagini. Lombardia e Veneto, 1986.I-II* [(a)]

Durata della ricerca alla I occasione	Durata della ricerca alla II occasione										
	pari a 6 mesi e multipli									diversi	Totale
	6	12	18	24	30	36	42	48	>48		
pari a 6 mesi e multipli											
6	2	6	1	3	-	1	-	-	-	19	32
12	1	15	5	5	-	2	-	1	1	17	47
18	-	4	2	4	1	2	-	-	-	8	21
24	-	3	2	13	3	12	-	3	1	35	72
30	-	1	-	3	-	1	-	1	-	4	10
36	-	3	-	8	2	30	-	4	4	3	74
42	-	-	-	-	-	-	-	-	-	-	-
48	-	2	1	1	-	2	-	4	3	6	19
48	-	-	1	2	-	2	-	-	5	3	13
diversi	8	28	2	19	3	9	1	3	3	96	172
Totale	11	62	14	58	9	61	1	16	17	211	460

(a) Dati al netto delle risposte coerenti.

fra effetto *heaping* e effetto telescopio in avanti è palese[8].

Una ricognizione più sistematica dei fattori che contribuiscono a dar conto di incoerenze nella durata riportata può, infine, essere tentata attraverso analisi di regressione di ECD su convenienti insiemi di potenziali variabili esplicative. La Tab. 4 presenta i risultati di alcune specificazioni di regressione, contraddistinte dall'inclusione di un numero via via maggiore di regressori. Più precisamente:

(a) la prima specificazione attiene alla regressione di ECD su una sola va-

8 L'eccezione, peraltro anch'essa conforme alle attese, si ha per le durate molto brevi - inferiori ai 3 mesi - alla prima occasione. Per queste, la risposta non è arrotondata per definizione, mentre l'effetto *heaping* emerge facilmente alla seconda occasione d'indagine, con un arrotondamento tipicamente ai 6 mesi. Tra l'altro, ciò da conto della forte prevalenza di incoerenze con ECD positivo in corrispondenza di brevi durate iniziali, documentata nella Tab. 1.

Tab. 4: *Risultati di analisi di regressione dell'errore nel riportare il cambiamento della durata (ECD), rispetto a possibili determinanti. Lombardia e Veneto, 1986.I-II* [(a)]

Variabili indipendenti		Specificazioni		
		(A)	(B)	(C)
Intercetta		4,25	3,43	2,94
Durata		-0,26** *(10,80)*	-0,28** *(10,78)*	-0,29** *(10,48)*
Heapet	(=1 se risposta su 6 o multipli in entrambi i trimestri)		3,46** *(3,03)*	3,33** *(2,92)*
Heap1t	(=1 se risposta su 6 o multipli alla I occ.)		1,51 *(1,53)*	1,26 *(1,27)*
Heap2t	(=1 se risposta su 6 o multipli alla II occ.)		3,58** *(2,89)*	3,21** *(2,64)*
Selfself	(=1 se risponde l'interessato ad entrambe le occ.)		-0,97 *(0,84)*	-0,36 *(0,29)*
Selfprox	(=1 se risponde l'interessato alla I occ.)		-3,29** *(2,45)*	-3,09** *(2,29)*
Proxself	(=1 se risponde l'interessato alla II occ.)		-1,38 *(0,83)*	-1,11 *(0,70)*
Occ	(=1 se precedentemente occupato)			-0,32 *(0,35)*
Sex	(=1 se femmina)			1,41* *(1,77)*
Eta1	(=1 se età tra 25 e 50)			0,21 *(0,25)*
Eta2	(=1 se età superiore a 50)			-2,68 *(1,22)*
Dipla	(=1 se diplomato o laureato)			-0,31 *(0,38)*
R^2		0,136	0,157	0,164
$\bar{R}^2$		0,134	0,149	0,150

(a) Statistica t in parentesi. **: significativo al livello del 5%. *: significativo al livello del 10% (test a due code, $H_o: \beta_j = 0$)

riabile, la durata riportata alla prima occasione;

(b) la seconda equazione di regressione include fra i regressori anche un insieme di *dummies*, volte a cogliere l'influenza dell'effetto *heaping* (in entrambe le occasioni o solo alla prima o solo alla seconda) e dell'identità del rispondente - l'interessato o un'interposta persona -;

(c) infine, la terza equazione di regressione comprende ulteriormente variabili *dummy* relative a talune caratteristiche dei disoccupati - sesso, età, livello di istruzione, precedenti lavorativi -.

Il risultato più significativo di queste semplici analisi di regressione è la presenza di una forte relazione negativa fra la durata della ricerca di lavoro riportata alla prima occasione e i valori di ECD. Tale relazione, si noti, è molto stabile, ed è solo marginalmente influenzata dall'introduzione di ulteriori variabili esplicative, a conferma di un progressivo effetto telescopio in avanti. Delle restanti variabili, solo alcune, segnatamente le *dummies* per l'effetto ammucchiamento, mostrano coefficienti significativamente diversi da zero.

5. *Osservazioni conclusive*

La questione da cui ha preso le mosse questo studio, circa l'accuratezza dei dati di durata della ricerca di lavoro nella RTFL e quindi circa il loro potenziale informativo per analisi 'fini' sulla caratteristiche e la dinamica della disoccupazione, trova nelle analisi presentate elementi che inducono a qualche cautela. Ciò va tenuto a mente quando si vogliano usare dati sulla durata riportata, vuoi a fini descrittivi vuoi per la stima di modelli di durata, anche se non pare irragionevole poterne trarre importanti indicazioni.

Allo scopo di rendere chiare le implicazioni delle evidenze emerse per analisi descrittive, può essere utile guardare alla Tab. 5. In essa è riportata, per il totale dei disoccupati in Lombardia e Veneto nell'indagine 1986.I e per i tre sottoinsiemi in cui vengono usualmente distinti - già occupati, in cerca di prima occupazione, altri in cerca di lavoro -, la distribuzione secondo la durata di ricerca in opportune classi. L'interesse è nel raffronto fra due criteri di definizione delle classi di durata molto prossimi l'un l'altro, ma che conducono a risultati vistosamente diverso in presenza di un marcato fenomeno di *heaping*. In un caso le classi di durata sono quelle correntemente adottate nelle pubblicazioni dell'Istat, cioè con estremo superiore incluso; nell'altro caso si impiegano analoghe classi, considerandole però aperte a destra (il valore corrispondente all'estremo superiore di una classe è quindi incluso nella classe successiva).

Com'è evidente dall'ispezione della Tab. 5, il fatto che gli estremi delle classi coincidano con valori per cui si è documentato l'effetto *heaping* fa sì che fra le due distribuzioni vi siano differenze sensibili. Occorre quindi essere particolarmente cauti nella lettura dei risultati. Ad esempio, facendo riferimento ad un'ulteriore aggregazione delle classi e distinguendo semplicemente fra brevi e lunghe durate, col il confine posto sui 12 mesi, la frazione di soggetti con lunga durata della disoccupazione passa dal 50 al 61% a

Tab. 5: *Distribuzione dei disoccupati secondo la durata riportata della ricerca, utilizzando due diversi criteri di chiusura delle classi di durata. Lombardia e Veneto, 1986.I* [(a)]

Durata della ricerca (mesi)	Disoccupati già occupati	Persone in cerca di prima occupazione	Altre persone in cerca di lavoro	Totale
Classi con limite destro incluso				
0 - 3	180	87	98	365
	24,33	*6,78*	*10,79*	*12,46*
3 - 6	131	201	97	429
	17,70	*15,78*	*10,68*	*14,64*
6 - 12	166	318	179	663
	22,43	*24,80*	*19,71*	*22,63*
12 - 24	185	338	188	711
	25,00	*26,37*	*20,71*	*24,26*
oltre 24	78	338	346	762
	10,54	*26,37*	*38,11*	*26,01*
Classi con limite destro escluso				
0 - 3	127	42	65	234
	17,16	*3,28*	*7,16*	*7,99*
3 - 6	131	118	75	324
	17,70	*9,20*	*8,26*	*11,06*
6 - 12	144	318	129	591
	19,46	*24,81*	*14,21*	*20,17*
12 - 24	177	319	187	683
	23,92	*24,88*	*20,59*	*23,31*
24 e oltre	161	485	452	1.098
	21,76	*37,83*	*49,78*	*37,47*
Totale	740	1.282	908	2.930

(a) Percentuali (per colonna) in corsivo.

seconda che si adotti la prima o la seconda convenzione per la cesura.

Quanto all'uso di dati sulla durata riportata della disoccupazione in modelli stocastici di comportamento individuale, si tratta essenzialmente di vagliare se e in che misura le inaccuratezze documentate inducano distor-

sione nella stima dei parametri di interesse. A tale riguardo, qualche evidenza confortante è in Torelli e Trivellato (1989) e nel cap. 19, dove viene impiegata una semplice procedura *ad hoc* per valutare l'impatto dell'effetto *heaping* sui valori delle stime dei parametri di un modello di durata della disoccupazione giovanile. Nell'ambito di ragionevoli correzioni per l'*heaping*, e condizionatamente ai dati analizzati ed al modello di durata impiegato, non risulta che le inaccuratezze dovute all'effetto ammucchiamento abbiano un impatto apprezzabile sulle stime dei parametri. È appena ovvio osservare che si tratta, tuttavia, di primi circoscritti riscontri.

L'analisi dell'impatto di errori connessi al processo di memoria sulle stime dei parametri di modelli di durata è argomento che, anche alla luce di quanto mostrato in questo sede, è meritevole di approfondimenti e verosimilmente troverà ampio spazio nel procedere degli studi statistici su modelli per dati longitudinali. Una prospettiva di ricerca di particolare interesse è legata alla possibilità di specificare, accanto al modello di durata, un opportuno modello di misura. Tuttavia, è facile prevedere che tale strada possa riservare non poche difficoltà nella fase di specificazione di convincenti modelli di misura per dati di durata, come è prevedibile che in una successiva fase di stima si possano presentare problemi computazionali di una certa complessità.

PARTE QUARTA:

ANALISI ESPLORATIVE

11

LA SELEZIONE DI MODELLI LOG-LINEARI IN PRESENZA DI DISEGNO CAMPIONARIO COMPLESSO: UN'ESPERIENZA SUI DATI DELLE FORZE DI LAVORO

Gianfranco Lovison

1. *Introduzione*

Per inquadrare il contributo che questo capitolo intende fornire, è opportuno richiamare la distinzione introdotta da McCarthy (1966), e ormai largamente accettata, fra uso analitico e uso enumerativo dei dati provenienti da indagini con disegni campionari complessi, come è appunto l'indagine sulle forze di lavoro.

Per *uso enumerativo* si intende l'uso tradizionale dei dati per stimare parametri della popolazione finita oggetto d'indagine, quali medie, totali, percentuali, ecc. Per *uso analitico* si intende l'impiego dei dati per analizzare le relazioni fra variabili rilevate con l'indagine: tale uso può concretizzarsi a vari livelli di complessità, dalla stima di semplici misure di relazione (quali coefficienti di correlazione o misure di associazione fra coppie di variabili) alla analisi di nessi causali complessi mediante modelli statistico-probabilistici.

Tradizionalmente, l'utilizzatore dei dati sulle forze di lavoro pubblicati dall'Istat si è interessato alle stime di livello di vari fenomeni, quali la disoccupazione, l'occupazione, la ricerca di lavoro, ecc., possibilmente disaggregate secondo le modalità di variabili ritenute in qualche senso esplicative di detti fenomeni, quali la regione, l'età, il sesso, l'istruzione, ecc.. Ciò spiega l'attenzione rivolta dall'Istat nelle varie fasi (disegno dell'indagine, stima, pubblicazione dei risultati) alle proprietà statistiche e alla fruibilità pratica di tali stime puntuali di parametri della popolazione di riferimento. Nella terminologia prima ricordata, si può dire che sia la progettazione dell'indagine, sia la scelta degli stimatori, sia i criteri di presentazione dei dati sia, infine, il loro impiego da parte degli utenti, sono di tipo enumerativo.

Tuttavia, già oggi non è infrequente anche un impiego analitico dei dati pubblicati sulle forze di lavoro, ed è lecito attendersi che tale impiego si diffonderà in futuro. La spinta principale in questa direzione viene dal giustificato desiderio da parte dell'utilizzatore di sfruttare in modo più approfondito le informazioni che l'indagine rileva e che le tabelle pubblicate mettono a disposizione, per esplorare le relazioni, o, se si preferisce, i nessi

causali, fra variabili potenzialmente esplicative dei comportamenti di individui e/o famiglie sul mercato del lavoro e variabili che misurano intensità e modalità di tali comportamenti.

Che questa tendenza dall'enumerativo all'analitico sia in atto sembra confermato da un lato dal crescente numero di esperienze straniere su indagini analoghe (si vedano ad esempio i lavori di Kumar e Rao (1984) sulla *Labour Force Survey* canadese e di Aitkin e Healey (1985) sulla *Labour Force Survey* della CEE), sia la presenza stessa di un rilevante settore 'analitico' nel progetto di ricerca in cui questo studio si inserisce.

Il passaggio dall'uso enumerativo a quello analitico, in particolare quando quest'ultimo implica l'utilizzo di modelli statistico-probabilistici e delle relative procedure inferenziali, non è però privo di insidie. La prima tentazione di fronte alla grande abbondanza di dati multivariati forniti da indagini campionarie su larga scala come quella sulle forze di lavoro ed alla parallela disponibilità di *software* statistico è quella di inserire i dati nel calcolatore e interpretare i risultati forniti dalle numerose procedure analitiche disponibili. Ciò significa però, implicitamente o meno, ignorare il disegno campionario complesso e le procedure di stima che hanno dato origine alle stime pubblicate, dato che grandissima parte dei metodi implementati nei *packages* più diffusi poggia sulle ipotesi classiche di campionamento casuale semplice.

Sui pericoli insiti in questo utilizzo di metodi inferenziali classici su dati da campionamenti complessi, in particolare se comprendenti più stadi e una qualche forma di *clustering*, vi è ormai una vasta letteratura disponibile. Tuttavia, va riconosciuto che l'utilizzatore che non voglia ignorare il disegno campionario complesso e desideri utilizzare metodi di analisi appropriati per questa situazione non standard, si trova di fronte a problemi non semplici.

Questo capitolo tenta appunto di indicare alcune direzioni di lavoro per affrontare questi problemi. Ci si concentrerà sui problemi che emergono nell'analisi di relazioni fra variabili categoriali, dato che la maggioranza delle variabili rilevate dall'indagine sulle forze di lavoro sono di questa natura e i dati sono conseguentemente pubblicati sotto forma di tabelle di frequenze. In particolare, per 'analisi di relazioni' fra variabili categoriali si intenderà in questa sede la ricerca di modelli log-lineari parsimoniosi, capaci di adattarsi bene ai dati mettendo in evidenza la struttura di interdipendenza presente fra le variabili. Sebbene questo non sia probabilmente l'approccio più diffuso fra gli utilizzatori 'analitici' dei dati sulle forze di lavoro, esso offre alcuni vantaggi: da un lato, perchè sta in qualche modo a mezza strada fra un approccio esplorativo e uno strettamente inferenziale, e sembra quindi particolarmente appropriato alle prime fasi di analisi, soprattutto su dati pubblicati; dall'altro, perchè dà una cornice sufficientemente generale dal punto di vista metodologico, entro cui possono essere inseriti molti altri problemi applicativi, quali la verifica di singole ipotesi di indipendenza o omogeneità, anche in tabelle a più di due entrate, la costruzione di misure sintetiche di associazione fra variabili, e così via.

La struttura del capitolo è la seguente. La sez. 2 è dedicata ad un richiamo dei metodi proposti in letteratura per l'analisi di tabelle di frequenze da campionamenti complessi, accompagnato da alcune considerazioni compa-

rative fra tali metodi. La sez. 3 illustra una verifica empirica di alcuni di questi metodi , condotta su otto tabelle sulle forze di lavoro correntemente pubblicate dall'Istat. La sez. 4 contiene alcune indicazioni per l'utilizzatore e qualche suggerimento per il produttore di dati, aventi lo scopo di migliorare le opportunità di impiego analitico dei dati rilevati con l'indagine.

2. *Approcci proposti in letteratura*

La letteratura sull'analisi di tabelle di contingenza da campionamenti complessi è ormai abbastanza ricca, ma scarsi sono stati finora i tentativi di sistemazione organica. Pur senza avere la pretesa di proporre qui una tale rassegna, si ritiene opportuno presentare alcuni richiami sui metodi proposti in letteratura per l'inferenza su modelli per tabelle di frequenze da campionamenti complessi. Il tentativo è quello di introdurre una prima riunificazione della notazione usata dai vari autori (alquanto varia, e quindi fonte di difficoltà nell'individuazione dei nessi fra i vari approcci) e di mettere in evidenza la logica dei vari approcci piuttosto che le derivazioni formali. I metodi presentati nelle sezz. 2.2 e 2.3 sono trattati con un qualche maggior dettaglio, essendo quelli sottoposti a verifica empirica nella sezione successiva; una illustrazione più precisa si può comunque trovare in Lovison e Falorsi (1990).

2.1 *Notazione*

Siano A_h, h=1,...,H, variabili categoriali, cioè qualitative o quantitative discretizzate, caratterizzate da I_h modalità ciascuna.

Il supporto dell'analisi è la tabella di contingenza multipla costituita dalle $I = \prod_{h=1}^{H} I_h$ celle ottenibili combinando fattorialmente le modalità delle H variabili. Per convenienza, si userà spesso, nel seguito, la notazione vettorializzata in cui le celle sono ordinate lessicograficamente e il generico indice i=1,...,I, corrisponde alla generica combinazione di indici $(i_1,i_2,...,i_H)$.

Si assumerà che oggetto di indagine sia π, I-vettore delle proporzioni nella popolazione finita di riferimento (ovvero: I-vettore di probabilità nella super-popolazione infinita di riferimento, se si assume un punto di vista *model-based*).

Ovviamente, vale il vincolo $\pi'\mathbf{1} = 1$, dove $\mathbf{1}$ è l'I-vettore unitario. Tutto ciò che viene qui esposto per l'inferenza su π vale ovviamente, *mutatis mutandis*, per μ, I-vettore di frequenze (o di frequenze attese, in un'ottica *model-based*). Per non appensatire la presentazione, il parallelismo fra i due casi verrà d'ora in poi abbandonato.

L'informazione campionaria su π è contenuta in un campione di dimensione n. Lo schema di campionamento impiegato per acquisire tale campione è indicato con s. Il caso particolare di campionamento casuale semplice,

cioè il caso standard di schema multinomiale o prodotto multinomiale usato nell'inferenza classica su tabelle di contingenza, è indicato con srs.

L'inferenza su π è basata su **p**, stimatore s-consistente (cioè: consistente rispetto al disegno s) di π; naturalmente anche per **p** vale il vincolo **p'1** = 1.

Per ciò che riguarda gli operatori campionari (valori attesi, varianze, covarianze, ecc.), essi vengono contrassegnati con i deponenti s e srs ogniqualvolta si voglia mettere in evidenza il disegno campionario rispetto al quale vengono calcolati, mentre la quantità campionaria su cui operano viene indicata fra parentesi. In particolare: $\mathbf{V}_s(\mathbf{p})$ e $\mathbf{V}_{srs}(\mathbf{p})$ sono le matrici IxI di varianze/covarianze di **p** rispetto al disegno effettivamente impiegato per rilevare i dati (s) e se il disegno impiegato fosse quello casuale semplice (srs). Le rispettive stime sono denotate con $\hat{\mathbf{V}}_s(\mathbf{p})$ e $\hat{\mathbf{V}}_{srs}(\mathbf{p})$. In base a risultati noti per il caso multinomiale, si può scrivere:

$$\mathbf{V}_{srs}(\mathbf{p}) = n^{-1}(\Delta_\pi - \pi\pi'), \text{ dove } \Delta_\pi = \text{diag}(\pi) . \qquad (1)$$

Una stima consistente di $\mathbf{V}_{srs}(\mathbf{p})$ è data da:

$$\hat{\mathbf{V}}_{srs}(\mathbf{p}) = n^{-1}(\mathbf{D}_p - \mathbf{pp}'), \text{ dove } \mathbf{D}_p = \text{diag}(\mathbf{p}). \qquad (2)$$

Una delle quantità più usate per valutare l'impatto del disegno s rispetto ad un disegno srs nel caso univariato, è come noto, l'effetto di disegno o *deff* (Kish, 1965, p.258). Nel contesto dell'analisi di tabelle di contingenza, data la natura multivariata del parametro π, si ha una matrice IxI di *deff*, che nella notazione qui impiegata si può indicare come:

$\mathbf{D} = \mathbf{V}_s(\mathbf{p})/\mathbf{V}_{srs}(\mathbf{p})$, dove / denota il quoziente elemento per elemento.

La matrice **D** è quindi una matrice simmetrica, i cui elementi sono:

$$d_{ij} = \begin{cases} \dfrac{var_s(p_i)}{var_{srs}(p_i)} & \text{(cioè gli usuali } deff\text{, che converrà qui chiamare } deff \text{ di varianza)} \quad \text{per } i=j, \\ \dfrac{cov_s(p_i, p_j)}{cov_{srs}(p_i, p_j)} & (deff \text{ di covarianza}) \quad \text{per } i \neq j. \end{cases}$$

La matrice **D** può essere stimata dalla sua controparte campionaria:

$$\hat{\mathbf{D}} = \hat{\mathbf{V}}_s(\mathbf{p})/\hat{\mathbf{V}}_{srs}(\mathbf{p}).$$

Nel seguito si farà uso soltanto degli elementi sulla diagonale principale di $\hat{\mathbf{D}}$, cioè dei *deff* stimati di varianza:

$$\hat{d}_{ii} = \frac{\hat{var}_s(p_i)}{p_i(1-p_i)/n} .$$

Le relazioni fra le variabili A_h possono essere studiate mediante un modello log-lineare $\mathbf{M}_\vartheta$:

$$\mathbf{M}_\vartheta : \phi = u(\vartheta)\mathbf{1} + \mathbf{X}\vartheta \qquad (3)$$

con: ϕ = log(π), dove log(.) è l'operatore logaritmo naturale elemento per elemento;
X matrice del disegno Ixr nota, di rango r ≤ I-1; se r = I-1 si ottiene il modello saturato $\mathbf{X}_{max}$; **X'1** = **0**, dove **1** è l'I-vettore unitario e **0** è l'r-vettore nullo;
ϑ r-vettore di parametri ignoti che caratterizzano il modello $\mathbf{M}_\vartheta$;
$u(\vartheta)$ = log{1/[**1**'exp(**X**ϑ)]} fattore di normalizzazione che assicura il rispetto del vincolo π'**1**=1; exp(.) è l'operatore esponenziale elemento per elemento.

Alternativamente, il modello $\mathbf{M}_\vartheta$ può essere specificato in termini di equazioni di vincolo:

$$\mathbf{M}_\vartheta : \mathbf{C}'\phi = \mathbf{0}, \qquad (4)$$

con: **C** qualsiasi matrice Ix(I-r-1) di rango pieno, tale che **C'X** = **O** e **C'1** = **0**, dove **O** è la matrice (I-r-1)xI nulla e **0** è l'(I-r-1)-vettore nullo.

In particolare una scelta conveniente per **C** nell'ambito della usuale parametrizzazione log-lineare è rappresentata dalla matrice ortocomplementare di **X**, che 'accostata' a quest'ultima dà la matrice $\mathbf{X}_{max}$ del modello saturato, ossia:

$\mathbf{X}_{max}$ = [**X** | **C**], con: **C'X** = **O**, **X'1** = **0**, **C'1** = **0**.

Avendo così specificato **C**, **C'**ϑ rappresenta il vettore dei contrasti log-lineari postulati nulli se è valido il modello (4). È questa la scelta di **C** che adotteremo in seguito.

Tutte le quantità che si riferiscono al modello non saturato (4) sono caratterizzate dal deponente ϑ, cioè dal parametro specifico del modello $\mathbf{M}_\vartheta$. Così, π_ϑ è l'I-vettore di proporzioni (o di probabilità) sotto il modello $\mathbf{M}_\vartheta$. Nella stima di π_ϑ si incontra una prima difficoltà derivante dalla violazione dell'assunto classico di multinomialità, poichè non si è in generale in grado di scrivere la funzione di verosimiglianza per i campionamenti complessi utilizzati in pratica. Di conseguenza, non è possibile ottenere stime di massima verosimiglianza di π_ϑ . L'approccio suggerito da Imrey, Koch e Stokes (1982) per aggirare l'ostacolo, e utilizzato in questo capitolo , è il seguente.

Nel caso standard di campionamento multinomiale (srs), il sistema di equazioni di verosimiglianza risulta, come noto:

$$\mathbf{X}'\hat{\pi}_{\vartheta} = \mathbf{X}'(\mathbf{n}/n)$$

dove **n** è l'I-vettore di frequenze campionarie osservate (non ponderate).

Sostituendo **p** a **n**/n, si ottiene:

$$\mathbf{X}'\mathbf{p}_{\vartheta} = \mathbf{X}'\mathbf{p} ,$$

la cui soluzione mediante un qualsiasi algoritmo iterativo (IPFP, Fisher *scoring method*, ecc.), fornisce stime di 'pseudo massima verosimiglianza' $\mathbf{p}_{\vartheta}$ di π_{ϑ}. La s-consistenza di **p** assicura quella delle stime $\mathbf{p}_{\vartheta}$, sotto condizioni di regolarità standard.

2.2. *L'approccio dei minimi quadrati ponderati*

Supponiamo che l'analista abbia a disposizione gli stimatori s-consistenti **p** e $\hat{\mathbf{V}}_s(\mathbf{p})$. Facendo ricorso ad una versione del Teorema del Limite Centrale appropriata per il disegno s (per una rassegna , vedi Rao e Scott, 1981) ed appellandosi alle grandi numerosità campionarie tipiche di indagini svolte con campionamenti complessi, si può assumere che $\sqrt{n}\mathbf{p}$ sia asintoticamente normale multivariato:

$$\sqrt{n}\mathbf{p} \xrightarrow{d} N_I (\sqrt{n}\pi, n\mathbf{V}_s (\mathbf{p})).$$

Sia: $\mathbf{f} = \log(\mathbf{p})$. Allora: $\mathbf{C}'\mathbf{f}$ è una stima s-consistente di $\mathbf{C}'\phi$, il vettore di contrasti log-lineari che si postula nullo nella formulazione (4) del modello (3); $\hat{\mathbf{V}}_s (\mathbf{C}'\mathbf{f}) = \mathbf{C}'\Delta_p^{-1} \hat{\mathbf{V}}_s (\mathbf{p}) \Delta_p^{-1} \mathbf{C}$ è una stima s-consistente di $\mathbf{V}_s(\mathbf{C}'\mathbf{f})$, matrice di varianze/covarianze di $\mathbf{C}'\mathbf{f}$ rispetto al disegno s.

Applicando l'approccio di Grizzle, Starmer e Koch (1969) è allora naturale verificare l'adattamento del modello (4) mediante la statistica-test:

$$X^2_w = \mathbf{f}'\mathbf{C}[\mathbf{C}'\Delta_p^{-1}\hat{\mathbf{V}}_s (\mathbf{p}) \Delta_p^{-1}\mathbf{C}]^{-1}\mathbf{C}'\mathbf{f} . \qquad (5)$$

Se l'assunto di normalità asintotica di $\sqrt{n}\mathbf{p}$ regge, esso permette di considerare anche $\mathbf{C}'\mathbf{f}$ come un vettore asintoticamente normale. Di conseguenza X^2_w ha la struttura di una statistica-test di tipo-Wald e ad essa sono applicabili i classici risultati distributivi:

$$X^2_w \xrightarrow{d} \chi^2_v , \text{ con } v = \text{rango}(\mathbf{C}) = I\text{-}r\text{-}1.$$

I più convinti assertori di questo approccio sono Koch e i suoi collaboratori; un'esposizione generale si può trovare in Koch, Freeman e Freeman

(1975). Altri contributi in questo ambito sono in Schuster e Downing (1976), Tomberlin (1979), Imrey, Sobel e Francis (1980).

2.3. *L'approccio di aggiustamento di statistiche-test standard*

I metodi che possono essere raggruppati in questo filone muovono dalla convinzione che l'analista di dati qualitativi tenda, erroneamente, a utilizzare le statistiche-test standard che gli sono familiari anche in presenza di dati da campionamenti complessi, per i quali non valgono i classici risultati distributivi basati sullo schema multinomiale. Di qui, l'interesse per lo studio della distribuzione indotta, per tali statistiche standard, dai campionamenti non casuali semplici e la proposta di possibili aggiustamenti che rendano legittimo il confronto con il χ^2 tabulare.

I lavori fondamentali su questa linea di ricerca sono di Rao e Scott (1981,1984) e di Fellegi (1980). Altri contributi sono in Holt, Scott, Ewings (1980), Bedrick (1983), Rao e Scott (1985). Una rassegna è in Rao (1985).

2.3.1. *Aggiustamenti basati sui 'deff ' generalizzati*

Il risultato fondamentale è dovuto a Rao e Scott (1981,1984). Siano:

$$X^2 = n \sum_{i=1}^{I} (p_i - p_{\vartheta i})^2 / p_{\vartheta i} \qquad (6)$$

$$G^2 = 2n \sum_{i=1}^{I} p_i \log (p_i / p_{\vartheta i}) \qquad (7)$$

le usuali statistiche-test X^2 di Pearson e G^2 rapporto di verosimiglianze, calcolate utilizzando stimatori s-consistenti $\mathbf{p}$ e $\mathbf{p}_\vartheta$. Sotto l'assunto di asintotica normalità di $\sqrt{n}\mathbf{p}$, se cioè: $\sqrt{n}\mathbf{p} \xrightarrow{d} N_I (\sqrt{n}\pi, n\mathbf{V}_s (\mathbf{p}))$ le due statistiche-test (6) e (7) sono asintoticamente equivalenti e la loro distribuzione asintotica è :

$$G^2 , X^2 \xrightarrow{d} \sum_{j=1}^{I-r-1} \delta_j W_j , \qquad (8)$$

dove: W_j , j = 1,...,I-r-1, sono v.c. χ^2 indipendenti;
$\delta_1 , \delta_2 , \ldots , \delta_j , \ldots , \delta_{I-r-1}$ sono gli autovalori ordinati della matrice
$\mathbf{A} = [\mathbf{V}_{srs} (\mathbf{C'f})]^{-1} [\mathbf{V}_s (\mathbf{C'f})]$.

La matrice **A** gioca un ruolo fondamentale in questo approccio. Essa può essere anche riscritta, per mettere in evidenza il ruolo delle matrici di va-

rianze/covarianze di **p** rispetto al disegno s e rispetto ad un campionamento casuale semplice srs, come segue:

$$\mathbf{A} = [\mathbf{C}'\Delta_\pi^{-1}\, \mathbf{V}_{srs}(\mathbf{p})\Delta_\pi^{-1}\, \mathbf{C}]^{-1}\, [\mathbf{C}'\Delta_\pi^{-1}\, \mathbf{V}_s(\mathbf{p})\Delta_\pi^{-1}\, \mathbf{C}]\,,$$

ed è denominata da Rao e Scott 'matrice degli effetti generalizzati del disegno', in quanto può essere vista come l'estensione multivariata del concetto di *deff* univariato. Ad esempio δ_1, il più grande autovalore di **A**, è il massimo tra i *deff* di tutte le possibili combinazioni lineari degli elementi di **C'f**. Se si ha a disposizione una stima consistente $\hat{\mathbf{V}}_s(\mathbf{p})$, e facendo ricorso alla stima $\hat{\mathbf{V}}_{srs}(\mathbf{p})$ suggerita in (2), si può stimare consistentemente **A** mediante:

$$\hat{\mathbf{A}} = [\mathbf{C}'\Delta_p^{-1}\, \hat{\mathbf{V}}_{srs}(\mathbf{p})\, \Delta_p^{-1}\, \mathbf{C}]^{-1}\, [\mathbf{C}'\Delta_p^{-1}\, \hat{\mathbf{V}}_s(\mathbf{p})\, \Delta_p^{-1}\, \mathbf{C}].$$

Impiegando gli autovalori di $\hat{\mathbf{A}}$, $\hat{\delta}_j$, j=1 ,...,I–r–1, è allora possibile migliorare l'approssimazione a χ^2_{I-r-1} di X^2 (o di G^2) mediante le seguenti statistiche-test modificate:
(a) con correzione del primo ordine:

$$X_c^2 = (\hat{\delta}_\cdot)^{-1}X^2 \text{ , dove } \hat{\delta}_\cdot = (\, I-r-1\,)^{-1}\, tr(\hat{\mathbf{A}}) \text{ è uno stimatore consistente della media degli autovalori } \delta_j; \qquad (9)$$

(b) con correzione del secondo ordine:

$$X_s^2 = [\hat{\delta}_\cdot\, (1+\hat{a}^2)\,]^{-1}X^2 \text{ , dove } \hat{a}^2 = \frac{\Sigma(\hat{\delta}_j - \hat{\delta}_\cdot)^2}{(I-r-1)\,\hat{\delta}_\cdot^2} \text{ è uno stimatore consistente del quadrato del coefficiente di variazione dei } \hat{\delta}_j. \qquad (10)$$

Quest'ultima correzione è basata sulla approssimazione di Satterthwaite (1963) alla distribuzione (8).

Altre due statistiche-test modificate proposte in letteratura si basano sull'aggiustamento di X^2 non con funzioni dei *deff* generalizzati, ma con un loro estremo superiore, esatto o approssimato. L'aggiustamento che ne risulta è ovviamente molto conservativo, ma ha il pregio di richiedere molte meno informazioni sul disegno e di non dipendere dallo specifico modello indagato.

Il risultato chiave è fornito da Rao e Scott (1985), che hanno dimostrato che un estremo superiore esatto (*Exact Upper bound*) per $\delta_\cdot$ è dato da:

$$\delta_\cdot \le \sum_{i=1}^{I-r-1} \lambda_i/(\, I-r-1\,) \text{ , dove } \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_{i-1} \text{ sono gli autovalori non nulli della matrice } \mathbf{B} = n\Delta_\pi^{-1}\, \mathbf{V}_s(\mathbf{p}).$$

L'implicazione pratica di questo risultato è che, se l'analista ha a dispo-

sizione gli I-1 autovalori non nulli stimati $\hat{\lambda}_i$ di $\hat{\mathbf{B}} = n\Delta_p^{-1}\mathbf{V}_s(\mathbf{p})$, può stimare l'estremo superiore esatto:

$$_e\hat{u}_{\delta.} = \sum_{i=1}^{I-r-1} \hat{\lambda}_i / (I-r-1)$$

e correggere con esso la statistica X^2 grezza , costruendo la statistica modificata:

$$X^2_{EU} = (\,_e\hat{u}_{\delta.})^{-1} X^2 \,. \qquad (11)$$

Infine, se l'analista non ha a disposizione gli autovalori $\hat{\lambda}_i$, i=1,...,I-1, ma solo la loro media $\hat{\lambda}. = \sum_{i=1}^{I-1} \hat{\lambda}_i / (I-1)$, un estremo superiore approssimato (*approximate upper bound*) per $\delta.$ si può ancora ricavare osservando che:

$$(I-r-1)\delta. \leq \sum_{i=1}^{I-r-1} \lambda_i \leq \sum_{i=1}^{I-1} \lambda_i = (I-1)\lambda. = tr(\mathbf{B}) \,.$$

Quindi, avendo a disposizione $\hat{\lambda}.$, si può stimare l'estremo superiore approssimato con:

$$_a\hat{u}_{\delta.} = \frac{(I-1)\,\hat{\lambda}.}{(I-r-1)}$$

e correggere con questa quantità la statistica X^2 , ottenendo:

$$X^2_{AU} = (\,_a\hat{u}_{\delta.})^{-1} X^2 \,. \qquad (12)$$

Si noti che quest'ultima correzione è possibile anche avendo a disposizione solo $\mathbf{p}$ e i *deff* di varianza di $\mathbf{p}$, $\hat{d}_{ii}$, i=1,...,I, sfruttando la relazione: $tr(\hat{\mathbf{B}}) = \sum_{i=1}^{I} (1-p_i)\,\hat{d}_{ii}$, e stimando quindi l'estremo superiore con:

$$_a\hat{u}_{\delta.} = (I-r-1)^{-1} \sum_{i=1}^{I} (1-p_i)\,\hat{d}_{ii} \,.$$

2.3.2. *Un aggiustamento basato sui soli 'deff' di varianza*

Il calcolo dei *deff* (di varianza) $\hat{d}_{ii}$ delle stime delle proporzioni (o delle frequenze) congiunte rappresenta ormai una pratica diffusa nella fase di elaborazione dei dati rilevati in indagini campionarie con disegni complessi. Per alcune variabili rilevate con l'Indagine sulle forze di lavoro, ad esempio, tali quantità sono pubblicate nel cap. 2 di questo stesso volume. E' quindi

desiderabile sfruttare queste informazioni sull'impatto del disegno di campionamento non solo a fini di valutazione della qualità delle stime pubblicate, ma anche per la correzione di statistiche-test o di altre quantità impiegate a fini analitici.

Una correzione di X^2 basata sui *deff* di varianza $\hat{d}_{ii}$ è stata proposta, su basi intuitive, da Fellegi (1980):

$$X_F^2 = (\hat{d}.)^{-1} X^2 \text{ , dove } \hat{d}. = \sum_{i=1}^{I} \hat{d}_{ii} / I \text{ è il deff medio nella tabella.} \qquad (13)$$

Se si esclude la maggiorazione già vista nel paragrafo precedente, non vi sono relazioni analitiche note fra i d_{ii} e i δ_j, e quindi fra $\hat{d}$. e $\hat{\delta}$., utilizzabili per una valutazione comparativa dei meriti della statistica-test X_F^2 rispetto a X_C^2 o X_S^2 . La giustificazione teorica per l'uso di X_F^2 consiste nel fatto che, riscrivendo:

$$X_F^2 = \sum_{i=1}^{I} (n/\hat{d}.) (p_i - p_{\vartheta i})^2 / p_{\vartheta i} ,$$

risulta chiaro come essa derivi dalla correzione degli scarti ponderati $(p_i - p_{\vartheta i})^2 / p_{\vartheta i}$ con la quantità $(n/\hat{d}.)$ denominata da alcuni autori *effective sample size*, poichè rappresenta la numerosità di un campione casuale semplice che ha la stessa precisione (media) di stima del campione di dimensione n effettivamente usato.

2.4. *L'approccio basato su tecniche di replicazione*

L'idea fondamentale delle tecniche di replicazione (BRR - replicazioni ripetute bilanciate - in varie versioni e *jackknife*) è che la variabilità di una statistica basata su un intero campione può essere stimata in termini di sottocampioni (repliche) che siano selezionati in modo da riprodurre, eccetto che per la dimensione, il disegno complesso dell'intero campione. L'applicazione di questo principio alla verifica dell'adattamento di modelli moltiplicativi o log-lineari a tabelle di contingenza ha portato alla costruzione di statistiche-test alternative a quelle standard appropriate al caso srs.

I primi autori ad aver lavorato in questa direzione sono stati Chapman (1966), Nathan(1973) e Fellegi(1980), che hanno proposto stimatori e statistiche-test per il vettore $\mathbf{u} = \pi - \pi_\vartheta$, basati sulla tecnica delle BRR di McCarthy (1966,1969).

Pur essendo interessanti, anche per il loro carattere pionieristico in questo filone, questi contributi si sono limitati al caso di indipendenza in tabelle a due entrate e non hanno fornito *software* facilmente accessibile per la loro implementazione. Per questi motivi, essi risultano di scarsa utilità pratica, e non verranno ulteriormente trattati .

Più utile da questo punto di vista risulta il contributo di Fay (1979, 1982, 1984, 1985), che ha applicato il metodo del *jackknife* al problema della costruzione di una statistica-test utilizzabile su dati provenienti da disegni campionari che permettono l'uso di metodi di replicazione. L'idea di partenza è quella di applicare il *jackknife* non a $\mathbf{p}$, o a $\hat{\mathbf{u}}$, come pure si può fare per arrivare ad una stima di $\mathbf{V}_s(\mathbf{p})$ o di $\mathbf{V}_s(\hat{\mathbf{u}})$, ma direttamente all'usuale statistica X^2 per valutarne la variabilità. Schematicamente, la procedura proposta da Fay è la seguente.

Lo stimatore campionario $\mathbf{p}$ si può pensare, se il disegno campionario permette l'individuazione di K repliche almeno approssimativamente indipendenti, come la somma di K stime $\mathbf{z}^{(k)}$, indipendenti e identicamente distribuite, dove $\mathbf{z}^{(k)}$ è calcolata sulla k-esima replica. Inoltre, da $\mathbf{p}$ e da ciascuna $\mathbf{z}^{(k)}$ si possono ottenere, ad esempio mediante *Iterative Proportional Fitting,* le stime $\mathbf{p}_\vartheta$ e $\mathbf{z}_\vartheta^{(k)}$ delle proporzioni attese sotto il modello $\mathbf{M}_\vartheta$. Ciò rende possibile il calcolo della statistica X^2 'globale', cioè di (6), impiegando $\mathbf{p}$ e $\mathbf{p}_\theta$, e della statistica X^2 ottenuta escludendo la k-esima replica dal calcolo:

$$X^2(-k) = (n-n^{(k)}) \sum_{i=1}^{I} [\,(p_i - z_i^{(k)}) - (p_{\vartheta i} - z_{\vartheta i}^{(k)})\,]^2 / (p_{\vartheta i} - z_{\vartheta i}^{(k)}) .$$

Vi sono ovviamente K repliche $X^2(-k)$ e, corrispondentemente, K 'pseudo-valori' $B_k = X^2 - X^2(-k)$ su cui si basa la procedura *jackknife*. Intuitivamente: se la media dei B_k è 'vicina' a zero, ciò significa che il valore assunto da X^2 è compatibile con la sua variabilità casuale. Al contrario, se gli pseudo-valori sono 'significativamente positivi', ciò indica che X^2 divergerebbe al crescere di n, cioè è il modello $\mathbf{M}_\vartheta$ che non si adatta ai dati. La statistica proposta da Fay per quantificare la significatività statistica dell'andamento degli pseudo-valori è:

$$X_J = \frac{[X^2]^{1/2} - [K_{n+}]^{1/2}}{n^{1/2}\,[\sum_{k=1}^{K}(B_k - \sum_{h=1}^{K} B_h/n)^2]^{1/2} \,/\, (n-1)^{1/2}[8X^2]^{1/2}}, \qquad (14)$$

$$\text{dove: } K_{n+} = \begin{cases} K_n = X^2 - \sum_{h=1}^{K} B_h & \text{se } K_n > 0, \\ 0 & \text{altrimenti.} \end{cases}$$

Fay (1984,1985) dà anche risultati distributivi riguardanti la statistica X_J. In particolare, nel caso in cui l'effetto di *clustering* sia costante per tutte le proporzioni o frequenze stimate in una tabella, egli dimostra che tale distribuzione è una trasformata monotona della distribuzione χ^2 e fornisce una tavola di valori critici per $\alpha = 0{,}05$ e $\alpha = 0{,}01$.

2.5. *L'approccio 'model-based'*

Come in altri settori dell'inferenza su popolazioni finite, anche nel caso

dell'analisi delle relazioni fra variabili categoriali vi sono stati tentativi di fondare le procedure inferenziali su modelli di super-popolazione (approccio *model-based*) piuttosto che sulle distribuzioni indotte dal disegno di campionamento effettivamente impiegato (approccio *design-based*).

In questo specifico contesto, si possono considerare come appartenenti all'approccio *design-based* i metodi descritti in 2.2, 2.3 e 2.4, mentre si possono far risalire all'approccio *model-based* i contributi di Nathan (1969), Altham (1976), Cohen (1976), Brier(1980), Tomberlin (1980), Chambless e Boyle (1985). Nel seguito si descriverà brevemente il modello proposto da Brier che, pur essendo limitato ad uno schema particolare di campionamento, rappresenta una generalizzazione rispetto ad altri contributi in questo filone ed è comunque utile per spiegare come l'approccio *model-based* possa essere impiegato in questo contesto.

Brier considera un disegno di campionamento a due stadi da una super-popolazione divisa in R *clusters* di dimensione infinita. Da tale super-popolazione vengono estratti con campionamento casuale semplice r *clusters,* da ognuno dei quali vengono estratte, sempre con campionamento casuale semplice, un numero costante $n_q = t$, $q = 1,..,r$, di unità finali. Queste ultime vengono classificate secondo le I combinazioni di modalità delle H variabili categoriali in esame e disposte nell'I-vettore $\mathbf{n}_q$, $q = 1,...,r$. Dato lo schema multinomiale impiegato entro ciascun *cluster,* esse sono distribuite come v.c. Multinomiali:

$$\mathbf{n}_q|\pi_q \sim M_I(t,\pi_q)$$

di parametri t e π_q, vettore di probabilità specifico del q-esimo *cluster.*

Qui si inserisce il modello di super-popolazione, secondo il quale si assume che i vettori di probabilità π_q siano generati da una v.c. *Dirichlet,* con funzione di densità:

$$f(\pi_q \mid \pi,\nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^{I} \Gamma(\nu\pi_i)} \prod_{i=1}^{I} \Gamma(\pi_{qi})^{\nu\pi_i-1} \quad , \ \nu > 0 \ .$$

La distribuzione marginale del vettore $\mathbf{n}_q$ di frequenze risulta pertanto Dirichlet-Multinomiale (Mosimann, 1962):

$$\mathbf{n}_q \mid \pi \sim DM_I(t,\pi,\nu).$$

Dato che per questo disegno campionario lo stimatore di π è:

$$\mathbf{p} = (rt)^{-1} \sum_{q=1}^{r} n_q$$

si ha: $E_s(\mathbf{p}) = \pi$,

$$\mathbf{V}_s(\mathbf{p}) = \frac{(v+t)}{(rt)^2(v+1)} \mathbf{V}_{srs}(\mathbf{p}) .$$

Come si vede, con questo approccio è possibile esplicitare la distribuzione di **p**, o almeno i suoi primi due momenti. Di conseguenza, diventano utilizzabili i principi dell'inferenza classica, quale quello di verosimiglianza o dei minimi quadrati, per costruire stimatori, statistiche-test, intervalli di confidenza, ecc.. Alternativamente, Brier suggerisce di usare questi risultati *model-based* per correggere la statistica-test X^2 di Pearson:

$$X^2_B = \frac{(\hat{v}+1)}{(\hat{v}+t)} X^2 \text{ , dove } \hat{v} \text{ è uno stimatore consistente del parametro } v \text{ della Dirichlet-Multinomiale.} \qquad (15)$$

2.6. *Considerazioni comparative*

Non esistono in letteratura lavori generali di confronto fra i quattro approcci presentati nelle sezioni precedenti. E' comunque possibile formulare alcune considerazioni comparative alla luce di tre criteri di valutazione:

(a) utilizzabilità, sia in termini di quantità di informazioni necessarie per impiegare ogni metodo sia in termini di generalità e flessibilità del metodo stesso;
(b) *performances* statistiche, dal punto di vista della convergenza alla distribuzione asintotica, della potenza, ecc.;
(c) disponibilità di *software* per l'implementazione.

Per ciò che riguarda il punto (a), una prima distinzione va fatta fra metodi che richiedono l'accesso dell'utilizzatore ai dati individuali (quali quelli illustrati in 2.2, 2.4 e 2.5) e metodi applicabili avendo accesso ai soli dati pubblicati e ad informazioni aggiuntive sul disegno campionario, quali sono le statistiche-test modificate illustrate in 2.3. Fra queste ultime, vanno in particolare menzionate X^2_F e X^2_{AU} (per le quali è sufficiente una sola informazione sul disegno, rispettivamente $\hat{d}.$ e $\hat{\lambda}.$) e X^2_{EU} che richiede la conoscenza degli I-1 autovalori $\hat{\lambda}_i$. Più complesso è il discorso riguardante X^2_C e X^2_S. In teoria, queste statistiche-test richiedono la disponibilità di un numero ridotto di informazioni sul disegno, ma la quantità di queste informazioni dipende dai modelli che l'utilizzatore intende analizzare. Poichè il numero di tali modelli può essere anche molto elevato già per tabelle a tre o quattro dimensioni, è piuttosto irrealistico che l'Istat possa fornire queste informazioni per tutte le tabelle pubblicate. In pratica, quindi, anche queste due statistiche-test sono applicabili solo da parte di utilizzatori che abbiano accesso ai dati individuali dell'indagine.

Dal punto di vista della generalità e flessibilità, i metodi delle sottosezioni 2.2 e 2.3 e l'approccio basato sull'impiego del *jackknife* sembrano i più convincenti. I primi richiedono sostanzialmente la stimabilità della matrice

di varianze/covarianze $\mathbf{V}_s(\mathbf{p})$; sebbene tale stima possa essere complicata da ricavare in modo esatto per disegni molto complessi, è usualmente abbastanza agevole ottenerla in via approssimata, o trascurando alcune complessità del disegno che si reputano secondarie o mediante tecniche di replicazione (per una buona discussione su questo punto, vedi Imrey, Sobel e Francis, 1980). La grande flessibilità è poi uno dei vantaggi dell'approccio del *jackknife* sviluppato da Fay: in sostanza, purchè il disegno consenta la formazione di repliche plausibilmente indipendenti, esso è utilizzabile per qualunque disegno e qualunque modello.

Una considerazione specifica va infine fatta, proprio dal punto di vista dell'applicabilità, per l'approccio *model-based*. Tale approccio è, dal punto di vista teorico e a giudizio di chi scrive, il più elegante e completo, in quanto consente di tradurre in termini probabilistici l'effetto delle operazioni che vengono compiute nella realizzazione pratica di un dato disegno di campionamento. Purtroppo, sembra però ancora lontana una sua applicabilità generale, dato che schemi di campionamento realistici implicano meccanismi di generazione dei dati pressochè intrattabili da un punto di vista *model-based*. Si noti ad esempio che, come dimostrato da Rao e Scott (1981), il modello di Brier, pur impiegando una verosimiglianza Dirichlet-Multinomiale non del tutto banale, copre solo il caso abbastanza irrealistico di campionamenti a *clusters* con effetto di *clustering* costante per tutte le proporzioni stimate. Sembra quindi esservi necessità di maggiore lavoro di ricerca prima che questo approccio possa dare strumenti inferenziali realmente applicabili a disegni complessi del tipo di quelli impiegati nella pratica del campionamento da popolazioni finite.

Alcune indicazioni comparative sul punto (b) sono reperibili in lavori parziali di confronto (Rao, 1985; Thomas e Rao, 1984). La statistica X^2_W tende ad avere mediocri *performances* in termini di convergenza al χ^2 in presenza di campioni di limitata numerosità o di 0-campionari. Inoltre, essendo basata sull'inversione di $\hat{\mathbf{V}}_s(\mathbf{C'f})$, tende ad essere numericamente molto più instabile delle concorrenti, in particolare per tabelle grandi.

Nel caso particolare di effetto di *clustering* costante per tutte le stime in $\mathbf{p}$, le statistiche-test X^2_C, X_J e X^2_B sono asintoticamente equivalenti e convergono al χ^2 tabulare. In pratica, queste tre statistiche sono del tutto soddisfacenti ogni qualvolta gli autovalori $\hat{\delta}_j$ di $\hat{\mathbf{A}}$ non sono molto variabili ovvero, condizione più facilmente verificabile in concreto, quando i *deff* di varianza $\hat{d}_{ii}$ sono abbastanza simili tra loro. In questi casi, X^2_C si impone per la semplicità di calcolo. Nel caso più realistico in cui l'impatto del disegno sia diverso per le diverse proporzioni in $\mathbf{p}$, il che implica una forte variabilità dei $\hat{\delta}_j$ e dei $\hat{d}_{ii}$, X^2_C e X^2_B tendono ad essere anche gravemente anticonservative, e occorre fare ricorso a X_J o a X^2_S. Fra queste due, nessuna è uniformemente superiore all'altra; X^2_S sembra comunque preferibile dal punto di vista del calcolo. Il difetto maggiore di X_J è infatti di carattere computazionale: il suo calcolo richiede la ri-stima del modello M_θ per KxL repliche (dove L è il numero eventuale di strati) e tale procedura può ovviamente essere molto

dispendiosa.

Passando a considerare le statistiche-test modificate utilizzando stime dell'estremo superiore di $\hat{\delta}_.$, è ovvio che X^2_{EU} e X^2_{AU} tendono ad essere conservative rispetto a x^2_C, poichè correggono X^2 con una quantità maggiore di (o tutt'al più uguale a) quella usata da X^2_C . La tendenza a sovracorreggere X^2 è però almeno in parte compensata da vantaggi nella fase di calcolo e di pubblicazione. La matrice $\hat{\mathbf{B}}$, infatti, a differenza di $\hat{\mathbf{A}}$, non dipende da uno specifico modello e quindi le funzioni-test modificate X^2_{EU} e X^2_{AU} richiedono, come già accennato in precedenza, il calcolo e la pubblicazione di una quantità di informazioni aggiuntive sul disegno limitata e, soprattutto, svincolata dagli specifici modelli che l'analista intende esaminare.

Venendo infine al punto (c), va osservato che, per quel che riguarda i metodi *model-based*, non esiste *software* specifico per la loro applicazione. Per ciò che concerne invece il metodo del *jackknife,* Fay (1982) ha predisposto un *package*, chiamato CPLX, per implementare il suo metodo. Discorso più articolato va fatto per le statistiche X^2_W , X^2_C e X^2_S , che richiedono la stima completa di $\hat{\mathbf{V}}_s(\mathbf{p})$. A questo proposito va osservato che sebbene siano numerosi i *packages* generali per la stima di varianze di stime campionarie (per una interessante rassegna comparativa vedi Kaplan, Francis e Sedransk, 1979), pochi comprendono opzioni per la stima delle covarianze, che sono ovviamente necessarie per la stima completa di $\hat{\mathbf{V}}_s(\mathbf{p})$. Le possibilità per l'utente sono qui o di scrivere programmi di stima specifici per il proprio disegno campionario o di ricorrere ad uno dei pochi prodotti che contengono opzioni a questo scopo, scontando la necessità di operare approssimazioni più o meno pesanti per poter analizzare i propri dati. Fra questi prodotti, vanno ricordati OSIRIS IV (Lepkowski, 1982) e SURREGR (Shah e La Vange, 1982). Per le potenzialità di utilizzo di OSIRIS IV sui dati delle forze di lavoro, si può vedere Lunardi (1989).

3. *Una verifica empirica sui dati delle forze di lavoro*

In questa sezione si presenta una verifica empirica di alcuni dei metodi illustrati nella sezione precedente. Su otto tabelle correntemente pubblicate dall'Istat nei volumi *Rilevazione delle forze di lavoro* della Collana d'Informazioni è stata condotta la ricerca di un modello log-lineare parsimonioso che si adatti in modo soddisfacente ai dati. Tale ricerca è stata realizzata utilizzando come criterio la statistica X^2 di Pearson non corretta e sei delle statistiche-test illustrate in precedenza: X^2_W , X^2_C , X^2_S , X^2_F , X^2_{EU} e X^2_{AU} . Per maggiori dettagli su questa verifica, e in particolare sulla procedura di selezione adottata, si rimanda ancora una volta a Lovison e Falorsi (1990).

La domanda centrale è se la ricerca di un 'buon modello' conduce allo stesso modello finale sia facendosi guidare nella scelta da X^2 non corretta, e quindi non tenendo conto del disegno campionario che ha generato i dati, sia impiegando come criterio statistiche-test che incorporano informazioni

più o meno accurate sul disegno.

Le otto tabelle esaminate sono:

(1) Occupati, secondo il sesso (SEX), la posizione nella professione (POS), il ramo di attività economica (ATT);
(2) Occupati, secondo il sesso , la posizione nella professione, il ramo di attività economica, l'età (ETA);
(3) Occupati, secondo la condizione dichiarata (DIC), la posizione nella professione, il ramo di attività economica;
(4) Occupati, secondo la condizione dichiarata, il sesso, la posizione nella professione, il ramo di attività economica;
(5) Occupati, secondo il sesso, il ramo di attività economica, il tipo di attività (TAT), il tipo di orario (TOR);
(6) Occupati, secondo il tipo di attività, il sesso , la posizione nella professione, il ramo di attività economica;
(7) Forze di lavoro, secondo il sesso , la condizione sul mercato del lavoro (CON), l'età;
(8) Forze di lavoro, secondo il sesso, l'età , la condizione sul mercato del lavoro, l'istruzione (IST).

I dati riguardano il Veneto, rilevazione di Ottobre 1986.

La Tab. 1 riporta sinteticamente i risultati delle otto procedure di scelta. I modelli sono rappresentati con la notazione abbreviata proposta da Fienberg (1980), che impiega una proprietà dei modelli log-lineari gerarchici, in forza della quale ogni modello di questa classe è univocamente identificato dagli insiemi generatori, cioè dalle interazioni più alte, di ogni ordine, presenti nel modello stesso. I casi in cui le statistiche-test che tengono conto dell'impatto del disegno conducono alla scelta dello stesso modello scelto in base ad X^2 non corretto sono segnalati dal simbolo =. Nei casi in cui due o più modelli di analoga complessità parametrica si adattano bene ai dati si è preferito, coerentemente con l'ispirazione esplorativa dell'intero esercizio, non tentare di discriminare ulteriormente per arrivare alla scelta di un unico modello e riportare quindi tutti i modelli classificati come accettabili dalla procedura di selezione.

L'ispezione della Tab. 1 suggerisce alcune considerazioni di fondo:

(a) nella maggioranza dei casi tenere o non tenere conto del disegno comporta conclusioni diverse in termini di scelta di modelli;
(b) l'impatto del disegno influenza maggiormente tale scelta in tabelle a quattro che in tabelle a tre entrate. Ciò è dovuto al fatto che gli aggiustamenti apportati per tener conto del disegno hanno conseguenze operative solo su interazioni 'deboli', caratterizzate da significatività marginali. Con le grandi numerosità campionarie qui disponibili (6.273 per le tabelle sugli occupati, 9.086 per quelle sulle forze di lavoro) le interazioni di primo ordine risultano quasi sempre altamente significative, cosicchè il tener conto o meno del disegno comporta valutazioni diversificate di significatività soprattutto per ciò che riguarda le interazioni di secondo e terzo ordine che caratterizzano le tabelle a molte entrate;
(c) l'impatto del disegno, e in particolare la componente di *clustering*, fa assumere valori patologicamente alti alla statistica X^2 grezza, che con-

Tab. 1 : *Modelli finali risultanti da procedure di selezione condotte usando come criterio la statistica X^2 grezza e sei statistiche-test modificate per tenere conto dell' impatto del disegno di campionamento, su otto tabelle dall' indagine sulle forze di lavoro: Veneto, ottobre 1986.*

Tabella	Variabili	X^2	modello scelto in base a X^2_W	X^2_C	X^2_S	X^2_F	X^2_{EU}	X^2_{AU}
1	[1]SEX [2]POS [3]ATT	[12][13][23]	=	=	=	=	[13][23]	[13][23]
2	[1]SEX [2]POS [3]ATT [4]ETA'	[124][134][234]	=	=	[134][234][12] [134][234][13]	[134][234][12]	[234][12][13][14] [134][12][23][24] [124][13][23][34]	[234][12][13][14] [134][12][23][24] [124][13][23][34]
3	[1]DIC [2]POS [3]ATT	[13][23]	=	=	=	=	=	=
4	[1]DIC [2]SEX [3]POS [4]ATT	[124][13][23][34] [134][12][23][24] [234][12][13][14]	[12][13][14][23][24][34]	[12][14][23][24][34]	[12][14][23][24][34]	[12][14][23][24][34]	[12][14][24][34]	[12][14][24][34]
5	[1]SEX [2]ATT [3]TOR [4]TAT	[123][134][24]	=	=	=	[123][14][24][34]	[12][13][14][24][34] [12][14][23][24][34]	[12][14][24][34]
6	[1]TAT [2]SEX [3]POS [4]ATT	[124][134][23]	[124][13][23][34]	[124][13][23][24]	[124][13][23][34]	[124][13][23][34]	[12][14][23][24][34]	[12][14][23][24][34]
7	[1]SEX [2]COM [3]ETA'	[123]	=	=	=	=	[12][13][23]	[12][13][23]
8	[1]SEX [2]ETA' [3]COM [4]IST	[123][234][14]	[123][14][24][34]	[123][14][24][34]	[123][14][24][34]	[123][14][24][34]	[12][13][23][24][34]	[12][13][23][24][34] [12][13][14][23][24] [12][13][14][23][24]

seguentemente conduce alla scelta di modelli finali sovra-parametrizzati;

(d) pur con notevoli differenze nei valori numerici che assumono (si vedano le tab. 1-8 in Lovison e Falorsi, 1990) le statistiche-test X^2_W, X^2_C, X^2_S, e X^2_F conducono quasi sempre a selezionare il medesimo modello, risultando così, almeno per ciò che riguarda queste tabelle e questo particolare esercizio inferenziale, operativamente analoghe;

(e) è confermata la tendenza conservativa , già prevista a livello teorico, delle statistiche-test X^2_{EU}, X^2_{AU}, che conducono a sotto-parametrizzare il modello finale, escludendo interazioni probabilmente non eleminabili per una soddisfacente descrizione della tabella esaminata.

4. *Considerazioni conclusive*

E' appena il caso di dire che è necessario usare grande cautela nel trarre indicazioni generali da una verifica empirica come quella sinteticamente illustrata nella sezione precedente, in cui la valutazione dell'impatto del disegno campionario sulla scelta di un 'buon modello' deriva non da risultati analitici di carattere generale, ma dall'accumulazione di analisi condotte su tabelle diverse per dimensione, tipo di variabili, sotto-popolazione di riferimento. Tuttavia, è possibile prendere spunto da questa applicazione per avanzare almeno alcuni suggerimenti per l'utilizzatore e per il produttore dei dati sulle forze di lavoro.

Il messaggio fondamentale per l'utilizzatore analitico è che è indispensabile correggere la statistica X^2 grezza nel cercare modelli parsimoniosi per i dati sulle forze di lavoro: l'impatto di campionamento fa assumere valori patologicamente alti ad X^2, rendendo praticamente impossibile la semplificazione del modello. Il fatto che tale impatto si faccia sentire di più su interazioni 'deboli' (tipicamente quelle di ordine superiore al primo) che su interazioni 'forti' (tipicamente quelle di primo ordine) e diminuisca di peso al crescere della numerosità campionaria non deve far sottovalutare la rilevanza del problema. Dal punto di vista conoscitivo sono spesso le interazioni 'deboli', non immediatamente evidenti all'ispezione dei dati, quelle più interessanti e che legittimano il ricorso ad un approccio modellistico. Ancora, sono proprio sotto-gruppi selezionati, e quindi caratterizzati da una modesta presenza nel campione, a rivestire spesso un particolare interesse per l'analista (si pensi, nel caso delle forze di lavoro, ai disoccupati o ai giovani in cerca di prima occupazione).

L'esigenza di tenere conto del disegno comporta, come si è visto nella sez. 2, la necessità di avere informazioni aggiuntive sul disegno stesso. Ciò pone ovviamente l'utilizzatore in conflitto con il produttore dei dati, per il quale ogni aggiunta di informazioni rappresenta un allungamento dei tempi e un appesantimanto delle pubblicazioni. Le indicazioni per bilanciare queste opposte esigenze vanno diversificate per i casi in cui l'Istat fornisce all'utilizzatore i dati originali e per ciò che riguarda invece i dati correntemente pubblicati.

Nel caso in cui l'Istat fornisce all'utilizzatore i dati originali, è essenziale che nel *record* di ogni unità finale compaiano tutte le informazioni sul disegno di campionamento (strato originale, ed eventualmente collassato, di appartenenza; *cluster* di appartenenza; coefficienti di riporto all'universo; ecc.). Solo così l'utente è nelle condizioni di usare i metodi suggeriti dalla letteratura e il relativo *software*.

Per ciò che riguarda i dati pubblicati, trovare un punto di equilibrio fra le contrapposte esigenze dell'utilizzatore e del produttore è certamente più difficile. Sfortunatamente, la scorciatoia rappresentata dalla pubblicazione, per ogni tabella, di $\hat{\lambda}_.$ o, al più, dei $\hat{\lambda}_i$, i=1,...,I-1, per rendere possibile il calcolo delle statistiche modificate X^2_{EU} e X^2_{AU}, non è percorribile. Tali statistiche-test, infatti, non sembrano soddisfacenti come criteri-guida per la scelta di modelli, poiché, essendo eccessivamente conservative, tendono ad eliminare dal modello interazioni 'deboli' ma rilevanti per la comprensione della struttura di interdipendenza delle variabili esaminate.

Un possibile compromesso sembra essere invece fornito dalle ottime *performances* della statistica modificata X^2_F. Questa statistica, così come la statistica X^2_C per il sottoinsieme di modelli che ammettono stime MV in forma esplicita nel caso multinomiale (vedi Rao e Scott, 1984), può essere facilmente calcolata avendo a disposizione i *deff* di varianza $\hat{d}_{ii}$. La pubblicazione di tali *deff* con cadenza annuale, ad esempio in occasione della pubblicazione delle tabelle medie sulle quattro rilevazioni trimestrali, non dovrebbe essere proibitiva nè dal punto di vista computazionale nè dal punto di vista del volume delle pubblicazioni (si tratta di I informazioni per ogni tabella con I combinazioni di modalità) e avrebbe il vantaggio di fornire un fondamentale contributo sia all'utilizzatore analitico, per la correzione delle sue statistiche-test, sia all'utilizzatore enumerativo, per la valutazione dell'accuratezza delle stime pubblicate.

12

IL RUOLO DEI METODI DI ANALISI DEI DATI 'MULTIWAY' NELLO STUDIO DELLA STRUTTURA E DELLA DINAMICA DELL'OCCUPAZIONE

Sergio Bolasco e Renato Coppi

1. *Introduzione*

Scopo di questo capitolo è di mostrare l'utilità di costruire ed analizzare appropriati *array* di dati a più indici ricavabili da un archivio originario di dati, al fine di estrarre l'informazione essenziale che questo offre riguardo ad un complesso fenomeno in esame.

In questo quadro, riprendiamo nostri contributi all'analisi di tavole di dati sulle forze di lavoro (vedi Bolasco e Coppi, 1987 e 1988). Il lettore, pur dovendo riferirsi ai due lavori citati per approfondimenti e dettagli, trova in questo capitolo le coordinate essenziali, sia metodologiche che applicative, della strategia di analisi suggerita. Tali coordinate sono qui proposte in un quadro di sintesi che, rispetto ai precedenti lavori, dà anche conto di ulteriori analisi empiriche.

I dati statistici sulle forze di lavoro sono in generale caratterizzati da una pluralità di tagli osservativi. Da un lato, le numerose variabili di classificazione (ad es., sesso, condizione lavorativa, posizione nella professione, ecc.); d'altro lato le dimensioni spaziali e temporali del fenomeno (ad es., variabilità da regione a regione, dinamica temporale, ecc.). Allo scopo di pervenire ad una visione complessiva è spesso utile comporre diversi tagli osservativi in un'unico *array* di dati da sottoporre ad analisi. Ovviamente, la scelta dei tagli osservativi e la maniera di comporli sono elementi cruciali ai fini dei risultati che è ragionevole attendersi. Ad ognuna di tali scelte corrisponde un determinato *array* di dati, in genere di struttura complessa.

Sul piano metodologico, nei suoi sviluppi più recenti, l'analisi statistica multivariata considera esplicitamente questa complessità (vedi, ad es., Law, Snyder, Hattie e Mc Donald, 1984; Coppi e Bolasco, 1989). Alla classica matrice bidimensionale a 2 modi e 2 indici $X = \{x_{ij}\}$ che incrocia individui e variabili (x_{ij} = valore assunto dalla variabile j-ma sull'individuo i-mo), viene sostituito l'*array* a più modi, più indici e più dimensioni. Seguendo le definizioni di Carroll e Arabie (1980), per 'modo' si intende una particolare classe di entità considerata nel volume di informazioni disponibili (ad es., il modo individui, il modo variabili, il modo luoghi, il modo tempi, ecc.).

Per 'indice', reso in inglese con la parola *way*, si intende il criterio di

classificazione più fine dei dati, oggetto di analisi (per un esempio, vedi oltre l'applicazione). Gli indici sono il risultato del prodotto cartesiano di un certo numero di modi, alcuni dei quali possono essere ripetuti (ad es., dati i modi A e B, si può considerare l'*array* a due modi e tre indici AxBxB). Peraltro un *array* con un determinato numero di modi e indici può essere trattato, a seconda dei metodi, con un differente numero di dimensioni. Per 'dimensione' intendiamo il supporto fisico dell'*array* su cui si possono disporre uno o più modi: ad esempio, un *array* a 4 dimensioni è composto di elementi x_{ijkl} caratterizzati dalle 4 coordinate i,j,k,l, ciascuna delle quali si riferisce ad un supporto costituito da uno o più modi. Pertanto, si può avere un *array* a 4 modi, 5 indici e 3 dimensioni.

2. *Strategie di analisi basate su 'array' individui x variabili x occasioni*

2.1. *L' 'array' individui x variabili x occasioni*

Un tipo di *array* di dati la cui costruzione si rivela molto utile nelle indagini socio-economiche è quello a 3 modi, 3 indici e 3 dimensioni: $\mathbf{X} = \{x_{ijk}\}$ (i = 1,I; j= 1,J; k = 1,K), dove l'indice i si riferisce agli individui (in senso lato)[1], l'indice j alle variabili (in senso lato)[2] e l'indice k alle occasioni (se queste esprimono dei tempi si tratta di occasioni ordinate). La costruzione di questo tipo di *array* è indicata quando si desideri analizzare simultaneamente la struttura e la dinamica di un fenomeno complesso.

In questa sede facciamo riferimento a tale tipo di *array*, che sarà costruito a partire da una tavola statistica che fornisce per ogni anno (T), dal 1978 al 1986, e per le varie regioni italiane (R) la ripartizione degli occupati secondo i caratteri: sesso (S), posizione professionale (CLP) e settore di attività economica (AE). In base a questa tavola è possibile costruire numerosi *array* di dati con diversi modi ed indici. Ciascun *array*, frutto di una scelta ponderata, costituisce l'oggetto di una particolare analisi rispondente ad un problema specifico.

Indichiamo con $\mathbf{X}_k$ la matrice individui (righe) x variabili (colonne) nell'occasione k (strati) e con $\mathbf{Y}_k$ una matrice delle stesse dimensioni, ottenuta mediante trasformazioni apportate sugli elementi di $\mathbf{X}_k$ (ad es., mediante la definizione di 'fibre' ottenute dividendo ogni elemento x_{ijk} per il totale di riga $x_{i.k}$). L'analisi può vertere sull'insieme delle matrici originali $\mathbf{X}= \{\mathbf{X}_k\}$ $k \in K$ oppure sull'insieme delle trasformate $\mathbf{Y}= \{\mathbf{Y}_K\}$ $k \in K$.

Per quanto riguarda l'analisi, è utile distinguere una fase esplorativo-descrittiva da una più propriamente induttiva volta alla costruzione di un modello interpretativo. Nella prima fase l'elemento centrale è la misura e la

---

1 Non sempre per individui occorre intendere le unità di osservazione. Spesso accade di considerare come unità di analisi dei gruppi di individui (espressi in termini di uno o più criteri di classificazione).

2 Con il termine variabili a volte si intendono anche le singole modalità di caratteri qualitativi

rappresentazione delle diverse componenti di variabilità del fenomeno in esame. In questo caso, occorre manipolare l'*array* iniziale di dati in diverse maniere prima di pervenire ad una sintesi utile per passare alla seconda fase. Con riferimento ai dati oggetto della nostra indagine, consistenti essenzialmente in una tabella di contingenza a 5 entrate, è possibile, innanzitutto, valutare la variabilità spiegata dai singoli caratteri e dalle loro interazioni, tramite l'applicazione di un modello loglineare non saturato (ad esempio, con le interazioni ìdoppie ed alcune interazioni triple).

2.2. *L'approccio fattoriale*

In un secondo momento, la costruzione di specifici *array* a 3 indici consente di arricchire l'analisi esplorativa della variabilità in particolari direzioni. Nell'approccio fattoriale da noi adottato in Bolasco e Coppi (1987), l'enfasi è stata posta sullo studio della dinamica temporale della struttura occupazionale per regioni. In tal caso è il tempo a giocare il ruolo di occasione. In concreto, l'*array* considerato è stato costruito componendo due modi (attributi e luoghi: [(CLPxS)+R] sulla dimensione righe , mentre una parte del modo attributi (AE) è disposta sulle colonne, e il modo tempi costituisce i vari strati. L'elemento generico x_{ijk} definisce quindi il numero di occupati che, nell'anno k, sono caratterizzati dall'appartenenza al tipo i (residente in una certa regione, di un certo sesso, in una certa posizione professionale) e alla variabile j (cioè a un dato settore di attività economica).

Qualunque sia la scelta dell'*array* a 3 indici, secondo l'approccio fattoriale gli obiettivi fondamentali dell'analisi sono i seguenti:

(a) *Confronto 'globale'* tra occasioni. Si tratta di comparare la struttura complessiva delle matrici $\mathbf{X}_k$, o $\mathbf{Y}_k$, al variare di k. Ciò si traduce nel trovare delle dimensioni fattoriali lungo le quali rappresentare simultaneamente le diverse matrici. Nel contesto della classica scelta di piani fattoriali per la rappresentazione, ogni occasione diviene un vettore bidimensionale. Un problema importante è quello di interpretare gli assi fattoriali e, di conseguenza, la 'distanza' tra occasioni.

(b) *Analisi strutturale 'media'*. Si tratta di determinare le relazioni di fondo tra individui e tra variabili, a prescindere dall'influenza esercitata dalle singole occasioni. Ciò comporta la determinazione di spazi fattoriali 'medi', che consentano la rappresentazione (simultanea o meno) degli individui e delle variabili mediamente considerati.

(c) *Analisi strutturale 'fine'* (secondo le occasioni). Si tratta di esaminare in dettaglio i cambiamenti di struttura delle matrici $\mathbf{X}_k$, o $\mathbf{Y}_k$, al variare di k. Occorre dunque rappresentare le 'traiettorie' seguite da tali 'strutture'(siano esse definite da uno o più elementi 'omologhi'), nelle diverse occasioni, su degli spazi fattoriali. Per motivi di comparabilità e di interpretazione,è generalmente ritenuto utile effettuare tali rappresentazioni sugli spazi fattoriali medi.

Affinchè si possano osservare variazioni fra 'istanti' o fra 'soggetti' occorre che esista una struttura comune alle diverse matrici (tutti i metodi qui

considerati, infatti, misurano l'evoluzione essenzialmente in termini di statica comparata). Solo in condizioni di relativa stabilità saranno determinabili dei fattori o componenti comuni, grazie ai quali sia le rappresentazioni 'medie' sia quelle strutturali 'fini' in spazi comuni, assumono interesse e significato. Pertanto, il modo che scandisce le varie occasioni non dovrà corrispondere alla maggiore fonte di variabilità dell'*array*.

In Bolasco e Coppi (1987) abbiamo presentato i risultati conseguiti mediante tale approccio, con riferimento all'utilizzazione in parallelo di tre specifiche tecniche: i) l'Analisi Fattoriale Multipla (AFM: Escofier e Pages, 1988); ii) STATIS (Lavit, 1988); iii) L'Analisi di Matrici di Relazione (AMR: Coppi, 1986).

Riprenderemo alcuni di questi risultati nella sez. 3, in un quadro di sintesi. Peraltro, desideriamo sottolineare subito che un tale studio costituisce solo il primo passo di una strategia di analisi orientata alla costruzione di un modello interpretativo.

2.3. *L'uso sequenziale di tecniche 'multiway'*

In Bolasco e Coppi (1988), abbiamo prospettato un primo approfondimento in questa direzione. In particolare, abbiamo trattato dell'uso sequenziale di un insieme di tecniche *multiway*, come un 'apparato rivelatore' capace di produrre un'immagine pregnante, ancorchè semplificata, dell'informazione nascosta in un *array* complesso di dati.

In tale ottica abbiamo preso in considerazione due altri tipi di *array* a 3 indici, fondati sulla fibra costituita dal vettore dei rapporti di composizione degli occupati per settore di attività economica (AE), vista come modo variabili[3]. Specificamente, con riferimento ad un anno (1986) sono stati così costruiti i seguenti due *array*: i) l'*array*, che abbiamo denominato A, in cui il modo unità è definito da CLPxS e il modo occasioni da R; ii) l'*array*, che abbiamo denominato B, in cui il modo unita è definito da RxS e il modo occasioni da CLP[4].

Nell'analisi di tali *array*, all'applicazione in chiave esplorativa dei metodi STATIS e AFM è stata affiancata l'utilizzazione del metodo TUCKALS2, basato sul modello di decomposizione di Tucker (vedi Kroonenberg, 1983). Quest'ultimo modello ha una valenza interpretativa in termini di struttura latente e consente, pertanto, di passare alla fase terminale della strategia d'analisi, consistente nell'identificazione di un modello interpretativo del fenomeno occupazionale.

---

3 Si osservi che una volta costruito un certo tipo di fibre è possibile assumere solo due *slice*. Nel nostro caso, considerando le fibre secondo la variabile AE, per ciascun anno le occasioni compatibili con un simile *array* sono le regioni (*slice*: CLP x AE) oppure le posizioni professionali (*slice*: R x AE). In tal modo la struttura della variabile AE resta intatta. (Se considerassimo, invece, come *slice* i settori (CLP x R), essi avrebbero, per componenti, elementi di AE appartenenti a fibre diverse.) Questa situazione crea una asimmetria nell'analisi, che non è insita nei dati grezzi, ma scaturisce da scelte del ricercatore.

4 Palesemente questi due *array* possono riferirsi a singoli anni del periodo considerato, oppure a valori medi calcolati su tale arco temporale.

Alcuni elementi interpretativi sono già stati forniti in Bolasco e Coppi (1988). Tuttavia, essendo quell'analisi limitata ad un anno, nella sez. 3 di questo capitolo essa viene convenientemente estesa e i principali risultati già ottenuti sono corroborati, e arricchiti, dalla considerazione di *array* di dati relativi ad altri anni.

Dal punto di vista metodologico, è appena il caso di sottolineare che la strategia di analisi ora delineata non costituisce certamente un paradigma valido per qualsiasi indagine complessa. Le tecniche *multiway* utilizzabili possono, infatti, essere differenti. Permane tuttavia, crediamo, la centralità di un arco investigativo che si muova dal polo esplorativo a quello interpretativo attraverso una concatenazione di manipolazioni (*preprocessing*) ed analisi di dati, fondate sull'estrinseco carattere *multiway* delle informazioni disponibili.

3. *Applicazione*

3.1. *Il contesto*

Come già detto, si è considerata una tavola proveniente dall'indagine Istat sulle forze di lavoro, relativa agli occupati dichiarati. La tavola riguarda le medie annuali delle rilevazioni trimestrali (T), e classifica gli occupati dichiarati, a livello regionale (R), secondo il sesso (S), la posizione professionale (CLP) e il settore d'attività economica {AE}. Questa tavola, corren-

Tab. 1: *Modalità delle variabili di classificazione R, S, CLP e AE e loro legenda* [(a)]

Regione (R):
L = Lombardia
V = Veneto
Z = Lazio
C = Campania
P = Puglia

Sesso (S)
M = Maschi
F = Femmine

Posizione professionale (CLP)
L = Imprenditori e liberi professinisti
A = Lavoratori autonomi
C = Coadiuvanti
I = Impiegati e dirigenti
O = Operai e assimilati

Attività economica (AE):
AGR = Agricoltura
ENE = Industria energetica
MIN = Industria trasformazione dei minerali e industria chimica
MTL = Industria trasformazione dei metalli e industria meccanica
MAN = Industria manifatturiera
CST = Industria delle costruzioni e impianti
COM = Commercio, alberghi e pubblici esercizi
TRA = Trasporti e comunicazioni
CRE = Credito, assicurazioni, servizi alle imprese
PAM = Pubblica amministrazione, assistenza e previdenza
ASE = Altri servizi e Attività sociali varie

(a) Le sigle delle unità "tipo" sono una combinazione delle modalità dei caratteri corrispondenti: ad es. VML indica il tipo "imprenditori maschi del Veneto".

Tab. 2 - *Occupati dichiarati per sesso, condizione professionale e settore di attività economica. Lombardia, 1978*

	AGR	ENE	MIN	MTL	MAN	CST	COM	TRA	CRE	PAM	ASE
LML	2753	94	2304	6068	8625	10118	5844	1028	3131	1620	25929
LMA	77385	1540	8041	21715	58348	40835	166493	19.19	2525	1169	23.23
LMC	14463	23	657	2243	6699	3791	25394	1984	510	546	2157
LMI	4892	12486	44281	88834	88422	28064	61447	29814	70.354	50566	71317
LMO	35176	17369	92743	333210	306856	182831	146929	68486	5587	31975	63810
LFL	931	0	195	355	1350	267	1580	63	399	1021	4617
LFA	12761	0	763	1809	17474	386	56830	658	507	822	1808
LFC	14166	0	766	2236	7257	692	55077	536	309	151	3804
LFI	2324	3632	19884	33968	73867	8381	52139	14973	28599	32557	127832
LFO	6265	681	21731	634082	74135	4980	48192	4123	1760	14290	108357

temente non pubblicata, presenta l'ulteriore vantaggio di avere i dati espressi in unità, non in migliaia.

Per varie ragioni, nelle successive analisi empiriche si prendono in considerazione soltanto cinque regioni: Lombardia, Veneto, Lazio, Campania e Puglia. Esse rappresentano peraltro diverse, tipiche combinazioni di occupati nei tre rami di attività economica, e indirettamente differenti contesti economici (vedi Bolasco, 1983). Sono inoltre sufficientemente rappresentative della situazione italiana nel suo complesso (comprendono circa il 50% del totale occupati). I modi della tavola sono tre: il modo attributi (CLP,S,AE), il modo luoghi (R) e il modo tempi (T). Le modalità delle variabili di classificazione, e la simbologia utilizzata nel seguito per denotarle, sono nella Tab. 1. A mo' di esempio, la Tab.2 presenta una parte della tavola, relativa alla Lombardia, anno 1978.

3.2. *Studio delle fonti di variabilità: i risultati dell'applicazione di un modello loglineare*

In coerenza con quanto esposto nella sez. 2, il primo obiettivo che caratterizza la nostra strategia consiste nel valutare l'importanza relativa delle diverse fonti di variabilità dei dati che compaiono nella tavola originale. Per avere un quadro generale sulla situazione della variabilità, procediamo all'applicazione di un modello loglineare sull'*array* originario, considerato nella sua essenza di tabella di contingenza a 5 entrate.

Tale applicazione risulta utile per valutare in particolare l'entità dei contributi degli effetti semplici, doppi e di alcuni effetti tripli dei vari elementi costitutivi dell'*array*. Nella scelta dello specifico modello non saturato ci siamo avvalsi di risultati già ottenuti in Bolasco e Coppi (1987), tramite l'applicazione di tecniche *multiway* (nel caso in cui non esistessero indica-

zioni preliminari, occorrerebbe vagliare un ampio insieme di possibili modelli). E' stato considerato il seguente modello:

$$\log z_{ijklm} = AE + CLP + S + R + T + RxS + RxCLP + RxAE + SxCLP + SxAE + CLPxAE + SxCLPxAE + RxSxCLP,$$

dove abbiamo indicato con z_{ijklm} il valore atteso della frequenza nella generica cella della tabella e con le sigle dei caratteri gli effetti semplici doppi e tripli ad essi associati. L'adattamento di questo modello alla tavola osservata non risulta buono, in termini di significatività. Bisogna notare, tuttavia, che i dati in esame si riferiscono ad una intera popolazione, per giunta di numerosità assai elevata. In tal caso, per ottenere un buon grado di adattamento è in genere necessario ricorrere ad interazioni di ordine superiore, se non, addirittura, al modello saturato. Peraltro, trattandosi di dati di popolazione, le misure di attendibilità espresse in termini probabilistici hanno scarsa importanza pratica (ciò vale anche per i test di significatività concernenti i singoli parametri del modello e per il calcolo degli errori standard delle rispettive stime). Nel commentare brevemente i risultati per il modello in esame ci limiteremo, dunque, ai valori stimati dei parametri. L'ordine di grandezza di tali stime, valutato in senso relativo, permette in effetti di individuare quei caratteri che, presi singolarmente o a coppie o a terne, maggiormente influenzano la variabilità delle frequenze della tabella di contingenza considerata. Per quanto concerne l'effetto dei singoli caratteri, la maggiore variabilità è assorbita da CLP, seguita da AE e S. Scarsa variabilità si riscontra rispetto agli anni e alle regioni. Ciò risulta chiaramente dalla Tab. 3, dove riportiamo, i valori minimo e massimo degli effetti semplici relativamente a ciascun carattere e degli effetti interattivi doppi e tripli considerati nel modello.

Ancora con riferimento alla Tab. 3, si nota che, tra le interazioni doppie e triple, le maggiori componenti di variabilità sono attribuibili a CLPxAE, RxAE, RxSxCLP. Anche in tal caso, dunque, i caratteri CLP e AE sono quelli

Tab. 3 - *Valori minimo e massimo degli effetti semplici e d'interazione stimati con il modello loglineare*

Caratteri	Effetto		Caratteri	Effetto	
	min	max		min	max
AE	-1,16	1,78	RxAE	-1,04	1,00
CLP	-1,73	1,43	SxCLP	-0,59	0,40
S	-0,65	0,65	SxAE	-0,52	0,85
R	-0,09	0,77	CLPxAE	-2,78	1,75
T	-0,34	0,21	SxCLPxAE	-0,14	0,10
RxS	-0,09	0,03	SxCLPxR	-0,55	0,47
RxCLP	-0,19	0,24			

che assumono maggiore rilevanza.

Quest'analisi preliminare, anche se utilizzata in modo schematico, fornisce interessanti indicazioni ai fini delle successive investigazioni. Essa offre, tra l'altro, un chiaro supporto alla scelta di AE come fibra fondamentale degli *array* a tre indici costruibili a partire dalla tavola originale dei dati, e di CLP come elemento essenziale di un'altra delle rimanenti due dimensioni.

3.3. *Analisi descrittiva ed esplorativa dei dati con metodi fattoriali 'multiway'*

L'obiettivo caratteristico di questa seconda fase consiste nell'individuazione delle strutture empiriche più significative che emergono dai dati in esame.

Per questa fase, come già detto, i dati della tabella di contingenza sono stati trasformati in fibre del tipo profili (rapporti di composizione sull'attività economica). In Bolasco e Coppi (1987) questi profili sono stati standardizzati: in tal modo si considerano tutti i tipi con lo stesso peso, e tutti i settori con lo stesso peso. In Bolasco e Coppi (1988) si è scelto invece di non standardizzare le colonne (lasciando quindi che ogni settore di attività economica mantenga la propria variabilità) e di fissare un anno, il 1986, per studiare la situazione occupazionale sia delle varie regioni (*array* A) sia delle varie posizioni professionali (*array* B). Si è trascurato il settore ENE, in quanto generante un insieme di zero strutturali[5].

Percorriamo ora i passi salienti dell'analisi esplorativa dei dati.

3.3.1. *Confronto globale*

(A) *Struttura intertemporale*

Esaminata attraverso il metodo STATIS, l'entità delle variazioni annuali della struttura occupazionale è limitata e non descrive sui piani fattoriali una traiettoria significativa. L'unica differenziazione di un certo rilievo è l'andamento degli anni 84-86 rispetto ai precedenti. Tale andamento si osserva nel piano F1-F2 con l'opposizione rispetto al primo asse dei punti rappresentanti le matrici 84, 85, 86 in rapporto alle altre (vedi Bolasco e Coppi, 1987, fig. 3a). Combinando F2-F3 si può osservare l'omogeneità del triennio 84-86, la notevole differenza del 78 rispetto al 79 e 80, e quella dell'81 rispetto all'82 e 83.

Andamenti analoghi si ottengono applicando AFM. La principale analogia risiede nella contrapposizione degli anni 84-86 rispetto ai restanti in termini di F2. Peraltro, questa relativa invarianza interstrutturale costituisce una buona proprietà per l'applicazione delle tecniche AFM e STATIS.

Con l'applicazione di AMR, si evince ancora che gli anni non differiscono

---

5 La posizione di ENE nelle configurazioni ottenute per le attività economiche nei vari anni è nota da Bolasco e Coppi (1987).

comunque di molto, nella loro struttura, in nessuna delle dimensioni individuate tramite l'*unfolding*. Si rammenta che la 'struttura' di ciascun anno é, qui, fornita dall'insieme delle associazioni tra ogni tipo e ogni settore di attività economica. Ad un esame più dettagliato, si può tuttavia notare la distinzione degli anni in 3 gruppi 78, 79-83, 84-86. Da qui la possibilità di considerare 3 soli anni, uno per ciascun gruppo, nel caso si voglia verificare, in diversi periodi temporali, l'esistenza di una determinata configurazione strutturale di altri tipi di *array*. Per lo studio delle variazioni interregionali nel seguito sono stati scelti appunto il 78, l'82 ed l'86.

(B) *Struttura interregionale e interprofessionale*

La struttura interregionale e quella interprofessionale sono state analizzate in Bolasco e Coppi (1988) mediante il metodo STATIS, con riferimento rispettivamente agli *array* A e B.

Tale metodo offre come misura delle distanze fra strutture regionali o professionali la matrice di correlazione fra occasioni.

Si trova che le tavole per regioni (come quelle per anni) sono molto simili: la correlazione fra occasioni, infatti, non scende mai al di sotto di 0,9. Invece le posizioni professionali, viste come occasioni, risultano abbastanza diverse, al punto da porre in dubbio la possibilità di svolgere correttamente un'analisi media (che in effetti nel seguito tralasceremo).

3.3.2. *Analisi strutturale 'media'*

(A) *Configurazione temporale media*

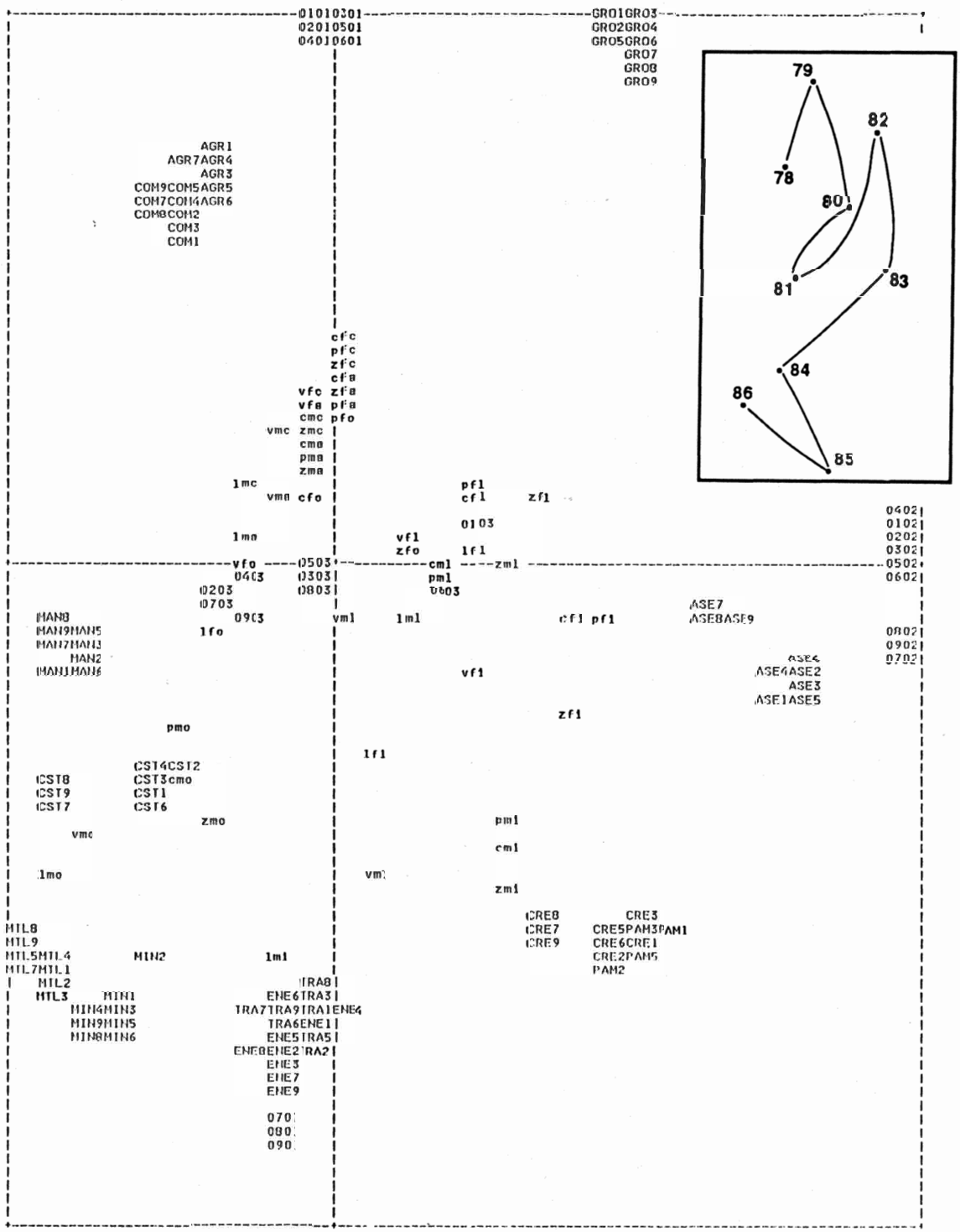
I risultati ottenuti con STATIS e con AFM sono in pratica sovrapponibili, poiché i coefficienti di ponderazione delle singole matrici $\mathbf{X}_k$ utilizzati nei due metodi, sono assai simili. Ne consegue, ad esempio, che le coordinate delle variabili, in STATIS, coincidono con quelle dell'AFM. Il loro posizionamento in AFM riproduce sul primo piano principale la fondamentale differenziazione dei settori di attività in rapporto alla posizione professionale, all'interno delle regioni (vedi Fig 1).

Peraltro, per quanto concerne le configurazioni regionali un'altra interessante informazione si ottiene dalla ricostruzione delle 'forme regioni' (vedi Bolasco e Coppi, 1987, fig. 6b). Tali forme, espresse in termini di posizioni professionali distinte per sesso, risultano simili nelle varie regioni, convalidando sotto altri aspetti la debole variabilità interregionale.

(B) *Configurazione regionale media*

Scegliendo per l'analisi dell'*array* A il modello di decomposizone di Tucker (metodo TUCKALS2) a tre fattori per ciascuno dei due modi, otteniamo delle dimensioni latenti (fattori) che forniscono, in sintesi, le seguenti informazioni. Per il modo unità (CLPxS): (i) il primo asse costituisce una buona misura della posizione nella professione, graduando progressivamente da destra verso sinistra i coadiuvanti, i lavoratori autonomi, gli operai, gli impiegati e dirigenti, i liberi professionisti e imprenditori; (ii) il secondo asse

Fig. 1: *A.F.M. - Piano principale: prime due variabili generali (x:F2;y:F1)(B)*



discrimina il lavoro indipendente dal lavoro dipendente, associando prevalentemente al primo il sesso femminile e al secondo quello maschile; in altre parole, permette di osservare la posizione delle donne rispetto agli uomini per ciascuna posizione professionale; (iii) il terzo asse contrappone gli operai alle restanti categorie, esprimendo così una polarizzazione fra addetti preposti alla produzione industriale strettamente intesa e altri addetti.

Per le componenti del modo variabili (AE): (i) il primo fattore gradua i settori d'attività da quelli con minor livello di complessità e diffusi nel territorio, fondamentalmente legati alla sussistenza della popolazione (AGR e COM), via via attraverso settori più complessi come l'industria fino a quelli evoluti e di tipo post-industriale (ASE); (ii) il secondo fattore oppone i settori maggiormente strutturati in termini di stabilità dell'occupazione, come molti di quelli industriali e la Pubblica Amministrazione, ai settori poco strutturati (ASE, COM e AGR), per i quali il cosiddetto "sistema di garanzie" è relativamente debole; (iii) il terzo fattore è chiaramente un asse di terziarizzazione, poichè polarizza le attività economiche volte alla produzione di beni (AGR e la totalità dei settori industriali) contro quelle di produzione di servizi (la totalità dei settori del terziario).

La configurazione fattoriale identificata con il metodo TUCKALS2 permette di valutare la struttura associativa del modo unità e del modo variabili tramite la visualizzazione congiunta degli elementi dei due modi nei cosiddetti *joint plots* (vedi Bolasco e Coppi, 1988, appendice). In effetti, con riferimento ai primi due fattori, dalla Fig.2 si traggono agevolmente i principali elementi di interazione fra i due modi. In particolare, si osservano le seguenti corrispondenze: Altri Servizi (ASE) con imprenditori e liberi professionisti (LF, LM) e impiegate (IF); Commercio (COM) e Agricoltura (AGR) con lavoratori autonomi (AF, AM) e coadiuvanti (CM, CF), con prevalenza sulle donne; Pubblica Amministrazione (PAM) con Impiegati (IM, IF), con prevalenza sugli uomini.

3.3.3. *Analisi delle 'traiettorie' secondo le occasioni*

L'interesse prevalente di questa analisi riguarda la struttura del modo unità, che può essere studiata in dettaglio tramite le traiettorie dei suoi elementi al variare delle occasioni. In questa sede ci limiteremo all'illustrazione delle configurazioni di punti omologhi secondo il tempo e le regioni.

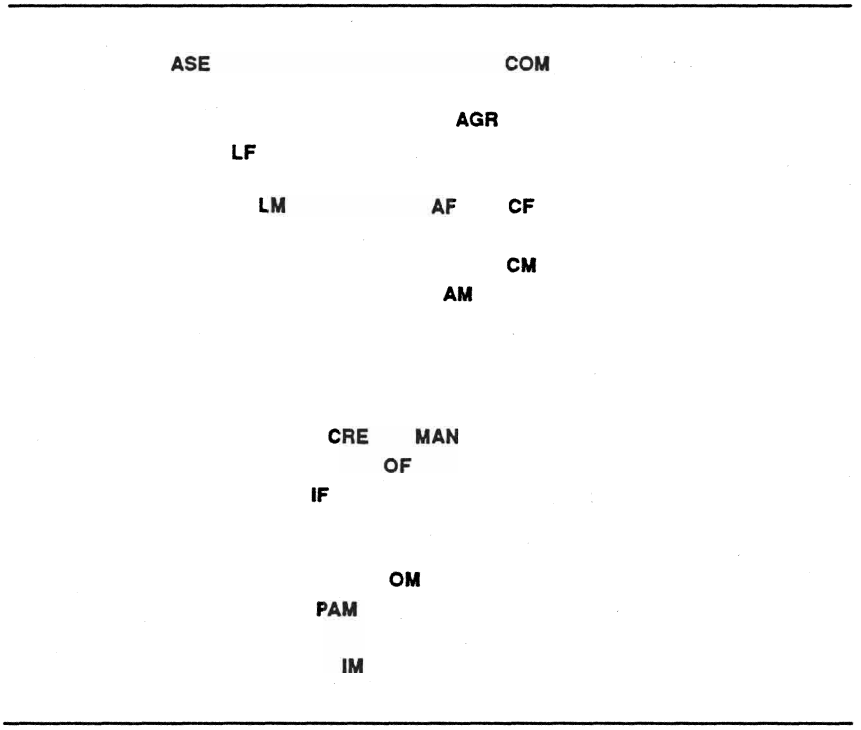
(A) *Dinamica temporale*

L'evoluzione temporale delle figure professionali all'interno delle regioni è scarsamente informativa, soprattutto a causa delle deboli variazioni annuali degli elementi costituenti le fibre dell'*array* in esame.

(B) *Differenziazione regionale*

Dal metodo TUCKALS2, applicato all'*array* A, si trae essenzialmente quanto segue (per dettagli, vedi Bolasco e Coppi, 1988, figg. 4a e 4b):

Fig. 2: *Metodo TUCKALS2: 'joint plot' del modo 2 (AE) e del modo 1 (CLP) sul piano delle prime due componenti (fattori) (x:F1;y:F2)*



(B1) Innanzitutto, esiste una diversa variabilità fra lavoratori dipendenti e indipendenti: eterogenei i primi, omogenei i secondi (la sola variabilità apprezzabile, fra questi ultimi, è quella dei lavoratori autonomi e dei coadiuvanti sul terzo asse). Si deduce che solo le condizioni lavorative eterogenee potranno descrivere traiettorie con un significato interpretabile.

(B2) Le microconfigurazioni per posizione professionale sono stabili (stesso ordinamento delle regioni) e somiglianti nella distinzione per sesso, sia per gli impiegati che per gli operai (la sola eccezione è quella delle operaie pugliesi, che precedono quelle lombarde sul terzo fattore). In particolare le traiettorie degli operai sono opposte, in termini di ordinamenti regionali, rispetto a quelle degli impiegati, coerentemente con i significati che tali traiettorie assumono sugli assi fattoriali, come vedremo più avanti.

(B3) Le traiettorie degli operai (sia M che F) sono opposte a quelle degli impiegati in termini del primo fattore. Poiché, come già detto, tale fattore costituisce una graduatoria dello status professionale, la traiettoria testimonia ad esempio come l'operaio lombardo possieda uno status migliore di quello pugliese, mentre è l'impiegato pugliese a trovarsi in posizione relativamente migliore dell'omologo lombardo. Interessanti sono anche le diverse distanze esistenti fra le regioni in ciascuna traiettoria.

(B4) Sempre limitandosi agli operai e impiegati, la graduatoria regionale rispetta il significato del terzo fattore (grado di terziarizzazione crescente verso il basso): gli operai più terziarizzati risultano quelli del Lazio, mentre fra gli impiegati ciò vale maggiormente per quelli campani e pugliesi (a netta distanza dagli omologhi del Veneto). Per questi ultimi, si nota che la loro posizione comprova la relativa maggiore percentuale di impiegati nei settori del terziario (PAM, CRE, COM, ASE), mentre per quanto riguarda gli impiegati in Lombardia la posizione rivela la loro relativa maggiore percentuale nei settori industriali.

3.4. *Verso un modello interpretativo*

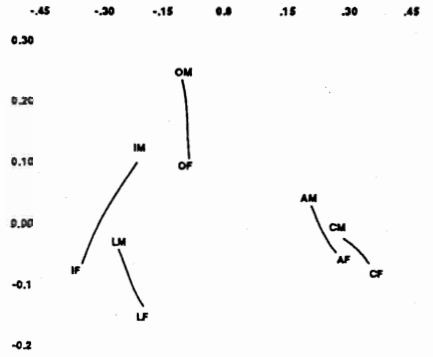
Alla luce della strategia di analisi tratteggiata nella sez.2, è possibile ora definire alcuni primi elementi costitutivi di un modello interpretativo della situazione occupazionale. Lo spunto è offerto dai risultati ottenuti tramite l'applicazione del metodo TUCKALS2 all'*array* A. Osserviamo, infatti, la stretta correlazione di significati fra gli assi corrispondenti ai due modi. Questo fatto si sostanzia nella struttura della cosiddetta *core matrix*, di cui, in Tab. 4, si riporta l'informazione in termini di variabilità spiegata. La forma prevalentemente diagonale di tale matrice denota un'assenza di interdipendenza fra i successivi fattori e una identità di fatto fra assi corrispondenti nei due modi.

Tenendo conto delle configurazioni medie già commentate nella sez. 3.3.2, siamo in grado di formulare l'ipotesi che il fenomeno in studio possa essere letto secondo una struttura tri-fattoriale costituita rispettivamente dallo status professionale, dalla 'sicurezza' del rapporto di lavoro e dalla relazione diretta/indiretta con i processi produttivi.

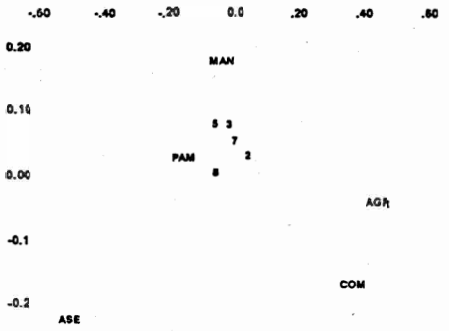
In particolare, in questo abbozzo di modello interpretativo: (i) lo status professionale assorbe la maggior quantità di variabilità totale; (ii) i settori di maggior peso sono da un lato AGR e COM e dall'altro gli altri servizi (ASE); (iii) le categorie con maggior variabilità sono i lavoratori dipendenti (operai e impiegati); (iv) esistono alcune forti interazioni fra gli altri servizi e l'imprenditorialità (LM, LF), fra i settori AGR e COM e il complesso del lavoro autonomo (con maggior peso dell'occupazione femminile), fra il settore manifatturiero e la fascia operaia (in particolare femminile).

In Bolasco e Coppi (1988, p.14) notavamo come i valori di tali interazioni potessero costituire utili elementi di valutazione sintetica della situazione occupazionale, qualora, ripetendo l'analisi su differenti anni o insiemi di

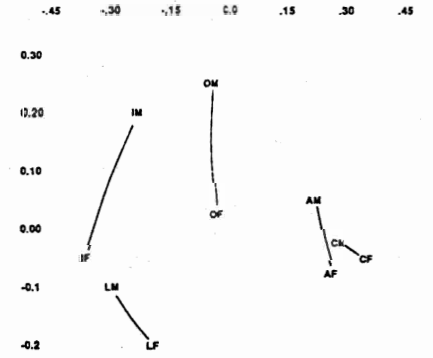
Fig. 3: *Piani delle prime due componenti relative ai modi 1 e 2, per gli anni 1978, 1982 e 1986.*



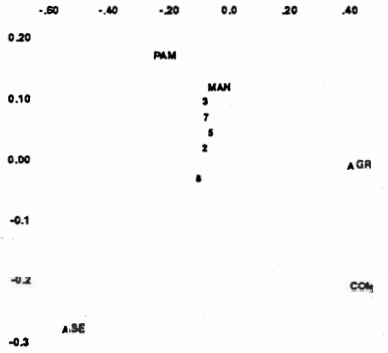
3A: MODO 1, ANNO 1978



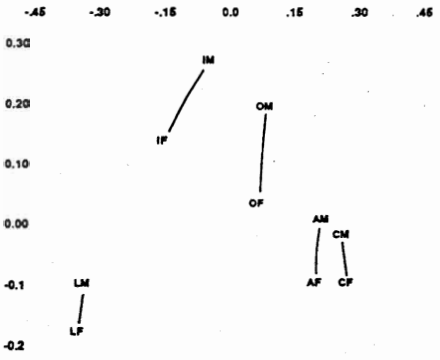
3D: MODO 2, ANNO 1978



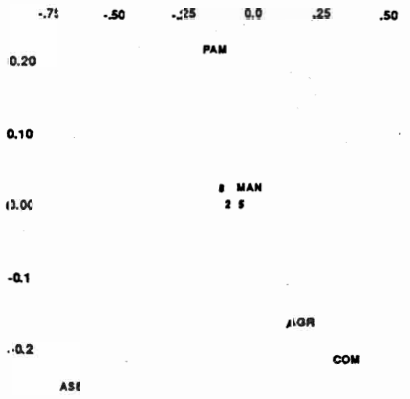
3B: MODO 1, ANNO 1982



3E: MODO 2, ANNO 1982



3C: MODO 1, ANNO 1986



3F: MODO 2, ANNO 1986

Tab. 4: *'Core-matrix' e variabilità spiegata secondo la combinazione delle componenti dei due modi*

	1	2	3	variabilità spiegata (ss)		
1	0,6331	0,0003	0,0048	ss(*core matrix*)	68.995,649	0,88735
2	0,0033	0,1721	0,0067	ss(elem. diagon.)	66.664,271	0,85737
3	0,0113	0,0035	0,0522	ss(el. extradiag.)	2.331,378	0,02998

regioni, si volessero cogliere le principali variazioni. In effetti, abbiamo ripetuto la medesima analisi per gli anni 1978 e 1982 (scelti in base a quanto detto nella sez. 3.3.1). I risultati salienti, per quanto riguarda gli elementi dei due modi secondo le prime due componenti del modello, sono nella Fig. 3. Essi possono essere così sintetizzati: (i) il modello risulta stabile, con un grado di adattamento crescente negli anni; (ii) il significato delle componenti è costante; (iii) le differenze fra i sessi e fra i settori appaiono in diminuzione nel corso del tempo.

4. *Conclusioni*

La conclusione essenziale è che dall'insieme delle analisi emergono elementi rilevanti per la definizione di un potenziale modello interpretativo dell'occupazione. Infatti, le dimensioni latenti che abbiamo identificato sembrano una buona base per la rappresentazione della struttura dell'occupazione, anche nella sua dimensione evolutiva.

Le ulteriori analisi empiriche presentate in questo capitolo confermano e arricchiscono le osservazioni da noi già prospettate in Bolasco e Coppi (1988). In particolare ci sembra opportuno richiamare, ribadendola, una considerazione.

Il fatto che le dimensioni latenti identificate appaiono essere fattori determinanti anche in un'analisi intertemporale, suggerisce che esse dovrebbero essere tenute sotto controllo nelle fasi di ridisegno di alcuni aspetti dell'indagine: sia del campionamento, sia del controllo delle stime, sia dell'affinamento delle modalità di classificazione vuoi per attività economica vuoi per condizione professionale.

13

FORME DI AGGREGAZIONE DEGLI INDIVIDUI SU BASE FAMILIARE: UN'ANALISI ESPLORATIVA

Fausta Ongaro

1. *Premessa*

Le modificazioni osservate recentemente nei tempi di accadimento e nella sequenza degli eventi che caratterizzano le prime fasi della vita adulta (acquisizione di autonomia abitativa ed economica; assunzione di responsabilità coniugali, riproduttive, ecc.) hanno contribuito, da un lato, a promuovere (assieme ad altri fenomeni quali la divorzialità) lo sviluppo di un più vivace interesse per lo studio delle famiglie e, dall'altro, a far emergere come argomento autonomo di ricerca il tema della transizione del giovane allo stato adulto (Kiernan, 1986; Young, 1987; Bianchi, 1987; Kobrin e Waite, 1987).

La diffusione presso la popolazione giovanile di forme familiari un tempo rare o inesistenti (convivenza al fuori del matrimonio, famiglie anucleari), che si sono poste in concorrenza con la famiglia nucleare[1] costituita al momento del primo matrimonio, ha complicato l'evoluzione del percorso familiare e posto le basi per la nascita di nuove esigenze conoscitive sulle interazioni che sussistono tra tipologie familiari ed eventi o stati che direttamente (perchè riferibili all'esperienza del giovane) o indirettamente (perchè relativi ad altri componenti la famiglia) coinvolgono l'individuo in questa fase dell'esistenza.

Nel nostro Paese l'interesse per le tematiche familiari ha trovato riscontro, in primo luogo, nella predisposizione di nuove indagini in cui la famiglia, da oggetto di rilevazione utile prevalentemente a fini campionari, diventa a pieno titolo unità di osservazione. L'indagine Istat del 1983 sulle Strutture e i comportamenti familiari (Istat, 1985b) e le successive indagini Multiscopo sulle famiglie sono un esempio di tale attenzione non solo per i contenuti che affrontano, ma anche per la cura che esprimono nella definizione del concetto di famiglia e nella rilevazione di quelle variabili che, assieme o singolarmente, favoriscono una classificazione tipologica (per nuclei) capace di esprimere al meglio i legami tra i componenti e la presumibile organizza-

1 Si intende per nucleo una coppia di individui con o senza figli o un genitore con figli che non costituiscono a loro volta un nucleo.

zione interna delle unità.

Se, tuttavia, cerchiamo nel panorama italiano dati utili per lo studio mirato della dinamica familiare giovanile i risultati sono meno incoraggianti. È vero che le giovani generazioni del nostro Paese sembrano abbracciare nuove forme di vita familiare con molta più cautela dei loro coetanei del centro-nord Europa e del nord America (Istat 1986; Golini, 1987; De Sandre, 1988). Dati relativi all'inizio degli anni '80 mostrano che in Italia la convivenza è ancora poco diffusa (su 100 donne in età 20-24 anni, quelle conviventi sono 42 in Danimarca e in Italia solo 1), la costituzione di famiglie composte da individui senza relazioni parental-coniugali è pressochè inesistente (in Irlanda su 100 individui in età 20-24 non meno di 10 vivono in gruppi anucleari) e l'abitudine ad andare a vivere da soli è ancora confinata a gruppi selezionati di popolazione (contro il 31% di donne sole in età 20-24 della Danimarca, l'Italia presenta solo l'1% di donne nella stessa condizione). Le modificazioni dei comportamenti, legate soprattutto a variazioni nella cadenza degli eventi (in particolare aumento dell'età al primo matrimonio), alterano solo i tempi di permanenza dei giovani nelle diverse forme familiari e lasciano che il matrimonio sia ancora la causa principale di uscita dalla famiglia d'origine.

All'interno di modelli di comportamento tradizionale sono ancora numerosi, tuttavia, i quesiti cui dare risposta.

Che responsabilità hanno alcuni eventi (superamento di un'età critica, acquisizione di indipendenza economico-lavorativa) nel passaggio del giovane alla vita di coppia o alla costituzione di una famiglia unipersonale? Il tempo trascorso in una certa condizione influisce sul rischio di sperimentare tale transizione? Quanto incidono altre caratteristiche dell'individuo (sesso, livello di istruzione), la vitalità o meno di entrambi i genitori e il livello di indipendenza economica di genitori e fratelli nel delineare i percorsi familiari del giovane e le durate di permanenza in ciascuno stato? Quali circostanze favoriscono la coabitazione di una nuova coppia con la famiglia d'origine di uno dei componenti? E quali (nascita di un figlio, superamento di una certa durata temporale, acquisizione di sicurezza economica), invece, determinano la cessazione della stessa coabitazione? Che peso hanno i fattori congiunturali quali il mercato del lavoro o delle abitazioni in questi processi?

Ogni tentativo di fornire risposte esaurienti anche ad uno solo di tali interrogativi non può che passare attraverso l'impiego di dati miranti a cogliere le numerose variabili in gioco nella loro evoluzione temporale. Fonti di natura trasversale non centrate sull'osservazione della popolazione giovanile, quali quelle ancora disponibili, possono solo fornire indicazioni di massima e parziali sull'andamento dei fenomeni.

Obiettivo di questo studio è di mettere a punto, senza alcuna ambizione di ricerca eziologica e di dettaglio, una prima serie di riflessioni su un tema ancora relativamente sconosciuto.

Nei limiti di dati provenienti da un'indagine tradizionale sulle famiglie, si vogliono individuare, con tecniche di analisi multivariata, gruppi omogenei di famiglie che, per le associazioni di variabili che esprimono, offrano indicazioni sulle modalità di aggregazione familiare proprie delle unità con giovani adulti. Ciò non consente solo di classificare gli aggregati secondo

criteri che, essendo sensibili ai fenomeni che interessano questa fase della vita, esprimono in forma più mirata e sintetica le condizioni familiari della popolazione giovanile (si osservi che la classificazione per nuclei nella versione che evidenzia le coppie con figli ed i nuclei monogenitore lascia spazio per ampi margini di ambiguità in questo sottogruppo di famiglie); con le precauzioni del caso, i risultati possono essere utilizzati anche per mettere a punto alcune ipotesi sulle dinamiche e sui comportamenti che sono all'origine delle medesime condizioni di vita familiare.

Oggetto di osservazione sono le famiglie del Veneto campionate per l'indagine sulle forze di lavoro del primo trimestre 1986[2] . Considerando che fino ai 19 anni le opportunità di sperimentare forme familiari diverse da quella d'origine sono piuttosto rare e che la maggior parte delle transizioni avvengono nel decennio successivo, dal campione originario di 5.179 unità sono state estratte 1.607 famiglie con almeno un individuo in età 20-29 anni (che nel seguito saranno più sinteticamente denominati giovani) e su queste è stata condotta l'analisi esplorativa.

2. *Le variabili*

L'insieme delle variabili attribuibili all'unità familiare è sensibilmente condizionato dalla natura trasversale dell'osservazione e dal fatto che la famiglia è utilizzata in forma prevalentemente strumentale a fini di formazione del campione. Informazioni a livello familiare sono ottenibili solo con l'elaborazione di dati sui singoli componenti e limitatamente a contenuti demografici di tipo tradizionale (età, sesso, stato civile, ecc.). Su questa base è, tuttavia, possibile individuare un gruppo di variabili categoriali relative ai giovani, alla composizione del resto della famiglia e all'intero aggregato familiare.

I dati sui giovani riguardano il numero, l'età, il sesso, lo stato civile, il livello di istruzione e la condizione professionale.

Stato civile. Considerando che i casi di convivenza e di interruzione di matrimonio sono poco diffusi, e che, peraltro, la presenza di almeno un giovane coniugato può essere indicativa dell'esistenza di un nuovo nucleo, l'interesse per questa variabile è stato limitato a due sole modalità: coniugato e non coniugato (celibe-nubile, separato, divorziato, vedovo).

Numero. Il numero di giovani di un aggregato assume importanza soprattutto in vista dell'individuazione delle situazioni di prolungata permanenza in famiglia d'origine di un unico figlio non coniugato.

Età. Se è indubbio che l'età gioca un ruolo rilevante nell'accompagnare il percorso evolutivo del giovane verso l'autonomia (psicologica, economica, abitativa, ecc.), è anche vero che analisi trasversali di questo tipo possono essere incapaci di cogliere le sfumature e le interazioni di tale impatto. Il suo impiego è stato quindi predisposto in modo da evidenziare eventuali

---

2 Per una descrizione delle caratteristiche dell'indagine, vedi il cap. 1 e più diffusamente Fabbris e Bernardi (1986).

associazioni con altre variabili solo all'interno di ampie tipologie familiari. A tale scopo l'età è stata considerata separatamente per coniugati e non coniugati: nel primo caso, con l'obiettivo di cogliere l'anzianità della coppia attraverso quella della donna, essa esprime l'età dell'individuo più giovane; nel secondo caso, con l'intento di far emergere situazioni di prolungata coabitazione con i(l) genitori (e), essa indica, invece, quella del giovane più anziano.

Sesso. Utilizzato in combinazione con il numero di giovani non coniugati, dovrebbe contribuire a precisare se certe forme di permanenza in famiglia d'origine si associano ad un particolare sesso dei giovani.

Condizione professionale. La variabile, in quanto *proxy* di autosufficienza economica, può essere strettamente correlata con la forma familiare sperimentata (uno stato di disoccupazione può essere la causa di una forzata permanenza in famiglia d'origine o costringere i fratelli coresidenti ad entrare precocemente nel mondo del lavoro). La sua importanza è sottolineata dall'utilizzo di un insieme di caratteri che rilevano la presenza di occupazione e quella delle più comuni forme di non occupazione (studente, in cerca di occupazione, ecc.). Per i giovani occupati, infine, nell'ipotesi che il lavoro indipendente configuri una maggiore incertezza economica, si distingue tra occupazione dipendente e indipendente.

Livello di istruzione. Espresso come titolo di studio più elevato tra gli individui di 20-29 anni, l'informazione ha significato solo per quanti hanno completato la formazione scolastica. Il suo scopo è quello di evidenziare in che misura differenze socio-culturali delle giovani generazioni interagiscono con la tipologia familiare sperimentata.

Le variabili sul resto della famiglia riguardano i figli, i genitori ed i fratelli dei giovani.

Figli. L'informazione assume importanza in quanto indica se almeno uno dei giovani dell'aggregato ha avviato il processo riproduttivo. Mancando di indicazioni precise sui rapporti di parentela con i giovani[3], la presenza di 'figli' è stata rilevata considerando tali tutti gli appartenenti alla famiglia in età inferiore di almeno 16 anni rispetto al giovane più anziano.

Genitori. Il dato sulla loro presenza è indicativo del fatto che i giovani abbiano avviato o meno il processo di autonomia abitativa rispetto alla famiglia d'origine. Analogamente che per i 'figli', sono stati considerati 'genitori' i membri della famiglia la cui differenza d'età con il giovane più anziano supera i 15 anni. Ulteriori dettagli sul numero, sesso (limitatamente al caso di un unico 'genitore') ed età degli stessi (in presenza di più di un 'genitore' l'anzianità dell'eventuale coppia è sintetizzata dall'età dell'individuo più giovane) possono offrire elementi di riflessione aggiuntivi sulle forme di aggregazione familiare che li coinvolgono.

---

3 La rilevazione della relazione di parentela è riferita al capofamiglia e prevede cinque modalità: coniuge, figlio, ascendente, altro parente, altro non parente.

Fratelli. L'introduzione di questa variabile ha senso soprattutto con riferimento allo studio delle tipologie che esprimono la permanenza del giovane in famiglia d'origine e, più in particolare, per capire se la presenza di altri individui in concorrenza con i giovani nella spartizione delle risorse familiari gioca un qualche ruolo nelle associazioni tra i caratteri. A questo scopo, sono stati considerati solo i 'fratelli' non coniugati di età inferiore ai 20 anni, individuandoli tra quelli che, rispetto al giovane più anziano, hanno una differenza d'età inferiore ai 16 anni. Per i 'fratelli' in età 14-19 anni si è proceduto, inoltre, a rilevare l'esistenza di un'occupazione.

I dati sull'intera famiglia riguardano il numero dei componenti e l'ampiezza demografica del comune di residenza.

Numero complessivo dei componenti la famiglia. La variabile risulta utile per discriminare tra famiglie unipersonali e altre e, all'interno di queste, tra aggregati di numerosità piccola (2-3 individui) e medio-grande (4 o più).

Ampiezza demografica del comune di residenza. Il carattere dovrebbe permettere di cogliere eventuali associazioni tra tipologia familiare ed ambiente culturale ed organizzativo in senso lato.

3. *La metodologia*

L'analisi esplorativa è stata condotta con il duplice obiettivo di evitare l'emergere di strutture banali e di ridurre al minimo il rischio di instabilità dei risultati che spesso accompagna questo tipo di approcci.

La consapevolezza che l'associazione tra le variabili espressa dai dati può non essere estranea al complesso delle scelte di metodo ha suggerito, in primo luogo, una riflessione organica sulla suddivisione delle variabili in attive e supplementari[4] e sulla loro precisazione a livello di modalità. Le Tabb. 1/A-B, che sintetizzano il prodotto di questo impegno, sono il risultato dell'incontro tra le considerazioni di merito svolte nella sezione precedente e decisioni di ordine tecnico orientate soprattutto a:

- limitare, per quanto possibile, tra le variabili attive, quelle suscettibili di associazioni strutturali (o banali), quelle con modalità rare e quelle che, alla luce di elaborazioni preliminari, indebolivano la struttura senza, peraltro, essere essenziali per la comprensione del fenomeno: questo spiega, ad esempio, l'incrocio tra la variabile età e stato civile dei giovani e l'inserimento tra le variabili supplementari delle informazioni sui 'fratelli', sul titolo di studio e sulle caratteristiche del comune di residenza;
- utilizzare contestualmente in forma flessibile l'insieme delle variabili supplementari in modo che ad esso non siano riconducibili solo elementi di contorno delle famiglie, ma anche quei caratteri che, in base al criterio

4 Sono considerate variabili attive le variabili per le quali sembra sussistere una struttura interessante di relazioni evidenziabile mediante un'analisi multidimensionale. Esse consentono di definire una struttura di riferimento alla quale poi correlare altre informazioni o valutazioni parziali (variabili supplementari).

precedente, non potevano essere inseriti *in toto* o nella forma desiderata tra le variabili attive[5] .

Tab. 1/A: *Variabili sui giovani impiegate nell'analisi esplorativa, loro codifica e loro utilizzo nel modello*

Variabili	Modalità	Codifica	Utilizzo [(a)]
- numero di coniugati per età del più giovane	0	C1	A
	1+ ≤ 24 anni	C2	A
	1+ ≥ 25 anni	C3	A
- numero di non coniugati per età del più anziano	0	NC1	A
	1 ≤ 24 anni	NC2	A
	1 ≥ 25 anni	NC3	A
	2+	NC4	A
- numero di non coniugati per sesso	0	NCS1	S
	1 M	NCS2	S
	1 F	NCS3	S
	2+	NCS4	S
- età del più anziano	≤ 22	E1	S
	23-26	E2	S
	≥ 27	E3	S
- numero di occupati per tipo di occupaz.	0	O1	A
	1+ almeno 1 ind.[(b)]	O2	A
	1+ tutti dipend.[(c)]	O3	A
- numero di indiv. in cerca di occupazione	0	D1	S
	1+	D2	S
- numero di studenti	0	S1	S
	1+	S2	S
- numero di inoccupati [(d)]	0	I1	S
	1+	I2	S
- titolo di studio più alto	≤ media inf.	T1	S
	medio super.	T2	S
	laurea	T3	S

(a) A=variabile attiva; S=variabile supplementare.
(b) Comprende: imprenditore, libero professionista, lavoratore in proprio, coadiuvante.
(c) Comprende: dirigente, impiegato, operaio, lavoratore conto terzi, apprendista.
(d) Comprende: casalinghe, milit. di leva, pensionati, inabili, benestanti.

5 Tra le alternative per contenere la formazione di strutture poco significative è stata considerata anche la possibilità di effettuare analisi asplorative differenziate per sottogruppi di famiglie. Ragioni di merito avevano anche prospettato come interessante la suddivisione delle unità in famiglie con almeno un giovane coniugato e famiglie composte solo da giovani non coniugati, ma la relativa bassa numerosità e la scarsa mutabilità della variabile (attiva) sul numero di 'genitori' che avrebbero caratterizzato il primo insieme hanno sconsigliato tale scelta.

Tab. 1/B: *Variabili sulla famiglia del giovane impiegate per l'analisi esplorativa, loro codifica e loro utilizzo nel modello*

Variabili	Modalità	Codifica	Utilizzo (a)
Variab. sui 'genitori'			
— numero	0	NG1	A
	1	NG2	A
	2+	NG3	A
— numero per età del 'genitore' più anziano	0	EG1	S
	1 ≤ 54	EG2	S
	1 ≥ 55	EG3	S
	2+ ≤ 54	EG4	S
	2+ ≥ 55	EG5	S
— numero per sesso	0	SG1	S
	1 M	SG2	S
	1 F	SG3	S
	2+	SG4	S
Variab. sui 'fratelli'			
— numero non coniugati in età 14-19 per condiz.professionale	0	FA1	S
	1+ alm.1 occup.	FA2	S
	1+ non occupati	FA3	S
— numero in età ≤ 13	0	FB1	S
	1+	FB2	S
Variab. sui 'figli'			
— numero	0	NF1	A
	1+	NF2	A
Variab. su intera fam.			
—numero di componenti	1	N1	A
	2	N2	A
	3	N3	A
	4+	N4	A
— caratteristiche del comune di residenza	capol. provincia	R1	S
	com.> 20.000 ab.	R2	S
	com. ≤ 20.000 ab.	R3	S

(a) A=variabili attive; S=variabili supplementari.

L'analisi delle associazioni è stata svolta con una concatenazione di tecniche esplorative tendente a contenere l'arbitrarietà delle scelte ed ottenere, così, una classificazione finale sufficientemente robusta (Griguolo e Palermo, 1984).

Tralasciando di entrare nel merito delle singole tecniche, peraltro ampiamente trattate in letteratura (Lebart, Morineau e Tabard, 1977), la proce-

dura prevede nell'ordine: un'analisi delle corrispondenze multiple, un'analisi dei *clusters* non gerarchica e un'analisi dei *clusters* gerarchica [6].

(a) Analisi delle corrispondenze multiple. Disponendo di variabili qualitative la fase è preliminare ad ogni trattamento successivo di tipo classificatorio. L'analisi è stata svolta sulle 1607 famiglie del campione con le 6 variabili attive e le 11 variabili supplementari descritte nelle Tabb. 1/A-B.

(b) Analisi dei *clusters* non gerarchica in via esplorativa. Ciò ha comportato: (b1) il recupero dei 10 fattori che in base all'analisi delle corrispondenze multiple spiegano il 97,7% dell'inerzia totale; (b2) la conduzione su di essi di un numero elevato (30) di partizioni esplorative secondo un algoritmo non gerarchico (scelta casuale dei centri) fissando a tre il numero di classi di ciascuna partizione; (b3) l'incrocio delle 6 (migliori) partizioni per le quali la quota d'inerzia tra le classi è risultata più elevata; (b4) la formazione conseguente di 21 classi stabili ciascuna delle quali risulta composta dagli oggetti (famiglie) sempre associati nelle partizioni esplorative.

(c) Analisi dei *clusters* gerarchica. L'analisi è stata condotta sulle 21 classi stabili ottenute al passo precedente: ogni classe è rappresentata dal vettore delle medie che i 10 fattori assumono in essa e pesata con la quota di famiglie che la compongono.

4. *I risultati*

La Fig.1, che rappresenta il dendogramma ottenuto applicando l'analisi dei *clusters* gerarchica alle 21 classi stabili e le variazioni d'inerzia spiegata dovute alle modificazioni dei livelli di aggregazione delle classi, fornisce gli elementi necessari per individuare un numero di gruppi omogenei utile ai nostri fini. Assumendo che il criterio per definire la somiglianza tra i gruppi sia la quota d'inerzia perduta a seguito dell'aggregazione di due classi (due gruppi sono tanto più simili quanto minore è la riduzione d'inerzia), si puo' osservare che un taglio dell'albero gerarchico al livello III (i livelli I e II corrispondono alla formazione rispettivamente di due e cinque gruppi) consente una suddivisione delle famiglie in un numero relativamente elevato (sette) di classi da permettere un'analisi sufficientemente articolata delle aggregazioni tra variabili.

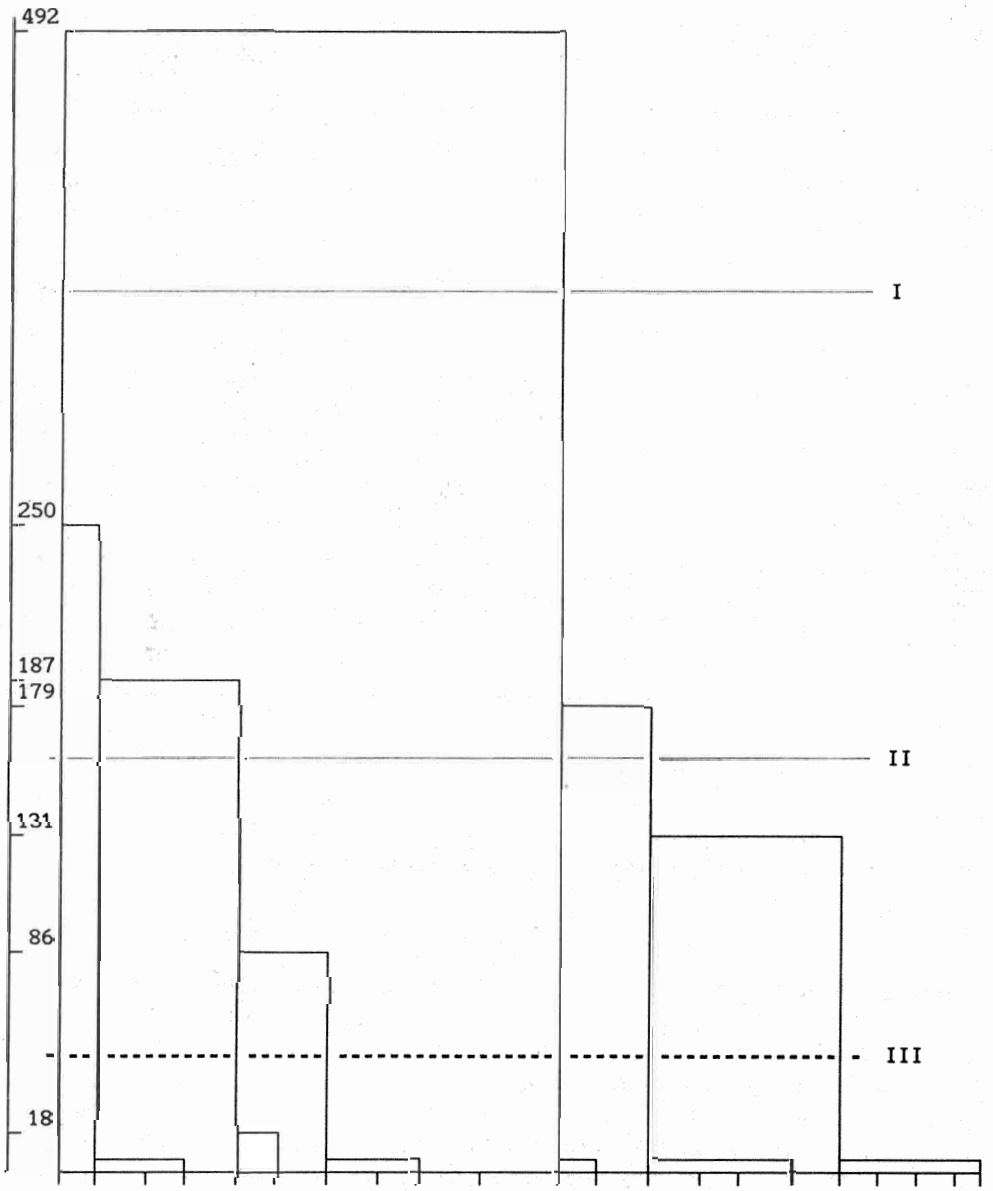
La Tab. 2, che descrive i profili dei sette gruppi ottenuti rispetto alle variabili attive e le quote d'inerzia (totale, intraclasse, interclasse) assorbite da tale taglio dell'albero, consente di riconoscere quattro ampie categorie di famiglie.

1. Un gruppo composto esclusivamente da giovani che vivono soli (A). Sono questi individui per la quasi totalità (96%) non coniugati, occupati nell'82% dei casi e di età tendenzialmente superiore ai 24 anni (oltre il 60%).

---

6 Per l'elaborazione dei dati è stato utilizzato il package ADDAEST descritto in Griguolo e Vettoretto (1983).

Fig. 1: *Albero gerarchico delle aggregazioni tra le 21 classi stabili, inerzia perduta (x 1000) a seguito delle variazioni dei livelli di aggregazione e tagli del dendogramma suggeriti dalle variazioni d'inerzia.*



Tab. 2: *Profilo dei sette gruppi omogenei rispetto alle variabili attive e quote d'inerzia assorbite dalla classificazione*

Variabili	A	B	C	D	E	F	G	Totale
N1	1,00	-	-	-	-	-	-	0,03
N2	-	0,99	-	-	-	0,02	0,49	0,13
N3	-	-	0,61	0,38	-	0,59	0,43	0,36
N4	-	0,01	0,39	0,62	1,00	0,39	0,08	0,49
C1	0,96	0,06	-	0,98	0,99	0,98	0,89	0,66
C2	-	0,41	0,22	0,02	-	0,01	0,05	0,10
C3	0,04	0,53	0,78	-	0,01	0,01	0,06	0,24
NC1	0,04	0,93	0,99	0,01	-	-	0,09	0,32
NC2	0,33	0,02	0,01	0,99	-	-	0,46	0,35
NC3	0,62	0,03	-	-	-	1,00	0,34	0,15
NC4	-	0,02	-	-	1,00	-	0,11	0,17
NG1	1,00	0,99	0,89	-	0,01	0,02	-	0,33
NG2	-	0,01	0,04	-	0,10	-	1,00	0,11
NG3	-	-	0,06	1,00	0,89	0,98	-	0,56
NF1	1,00	0,98	0,03	0,99	0,97	0,96	1,00	0,75
NF2	-	0,02	0,97	0,01	0,03	0,04	-	0,25
O1	0,18	0,08	0,33	0,37	0,16	0,25	0,36	0,28
O2	0,18	0,13	0,20	0,10	0,19	0,13	0,14	0,15
O3	0,64	0,79	0,47	0,53	0,65	0,62	0,50	0,57
Numero unità	45	135	379	481	260	167	140	1.607
Inerzia intraclasse (a)	0,02	0,07	0,26	0,16	0,08	0,06	0,12	0,77

(a) Considerando che l'inerzia totale è pari a 2,12, l'inerzia interclasse ne rappresenta il 64%.

2. Un gruppo composto da famiglie di due o più componenti con una generalizzata (non meno del 94%) presenza di giovani coniugati (B,C). Tale insieme puo' essere ulteriormente suddiviso in due sottogruppi.

2.1. Famiglie di due componenti (99%) con giovani coniugati (94%) e senza 'figli' (98%) (B). L'individuo più giovane della presumibile coppia (generalmente la donna) ha un'età di poco sbilanciata verso la classe 25-29 (ciò è vero nel 56% delle famiglie con coniugati) e nel 99% dei casi non esistono 'genitori'.

2.2. Famiglie di tre o più componenti (100%) con 'figli' (97%) e con almeno un giovane coniugato (100%) di età superiore ai 24 anni (78% delle unità) (C). Diversamente dal gruppo precedente, qui è presente anche un 11% di

casi con (uno o due) 'genitori'.

3. Un gruppo composto da famiglie caratterizzate dalla presenza di due 'genitori' (89% come minimo) con (nel 99% dei casi) giovani non coniugati (D,E,F). Tra queste possiamo distinguere diversi sottogruppi.

3.1. Famiglie con entrambi i 'genitori' (100%) ed un solo individuo non coniugato in età 20-24 anni (99%) (D). L'età del giovane, la presenza di due 'genitori' e la dimensione familiare complessiva (per il 62% di 4 o più componenti) fanno ritenere che, pur nella sua eterogeneità (qui possono confluire sia famiglie al termine della fase di contrazione, che aggregati alle soglie della stessa fase), questa tipologia esprima prevalentemente unità giovani sotto il profilo del ciclo evolutivo, in cui l'unico individuo di 20-24 anni è il più anziano di un insieme di fratelli che non ha ancora avviato il processo di allontanamento dalla famiglia d'origine. In questo gruppo oltre la metà dei giovani (63%) risulta occupato ed il 16% di questi presenta un'occupazione indipendente.

3.2. Famiglie con un'estesa maggioranza (89%) di due 'genitori' (nel 10% dei casi il 'genitore' è uno solo) e due o più individui non coniugati in età 20-29 anni (100%) (E). La numerosità e lo stato civile dei giovani fanno di queste famiglie una sorta di luogo di concentrazione di giovani adulti che si trattengono in famiglia d'origine. In questa tipologia la proporzione di famiglie con almeno un individuo occupato sale all'84% (e di queste il 23% registra almeno un individuo impiegato in attività indipendenti).

3.3. Famiglie con due 'genitori' (98%) ed un solo individuo non coniugato in età 25-29 anni (100%) (F). L' esistenza di entrambi i 'genitori', l'età elevata del giovane e la dimensione familiare (per il 59% si tratta di unità con tre componenti) fanno ritenere che in questo gruppo ricadano principalmente unità che sperimentano uno stadio evolutivo relativamente avanzato in cui l'unico figlio (rimasto) si attarda a vivere con i genitori. Il 75% di questi giovani risulta occupato e il 17% di essi svolge un lavoro indipendente.

4. Famiglie di un solo 'genitore' (100%) (G). Il gruppo è composto da aggregati di numerosità contenuta (due, tre componenti), in cui l'unico genitore ha al suo fianco soprattutto (91%) giovani non coniugati (generalmente in numero di uno) e, solo in misura minore (11%), giovani coniugati. In entrambi i casi l'età degli individui è variabile un po' su tutto l'arco del decennio considerato. Non essendo presenti (100%) 'figli' si può supporre che i giovani coniugati di questa tipologia non abbiano ancora avviato il processo riproduttivo. Nel 64% delle famiglie esiste almeno un giovane occupato e nel 22% di queste troviamo almeno un occupato con attività indipendente.

La Tab. 3, che descrive i gruppi rispetto alle variabili supplementari, aiuta a precisare ulteriormente le caratteristiche dei giovani e degli eventuali 'genitori' o 'fratelli'.

In particolare si può osservare che:

- gli individui che appartengono al gruppo A sono prevalentemente maschi (58%), in età maggiore di 22 anni (86%, per il 44% in età 23-26 e per il 42% in età 27-29) e, per il 53%, con istruzione medio-superiore;

Tab.3: *Profilo dei sette gruppi omogenei rispetto alle variabili supplementari*

Variabili	A	B	C	D	E	F	G	Totale
NCS1	0,04	0,93	0,99	0,01	-	-	0,09	0,32
NCS2	0,58	0,01	0,01	0,55	-	0,62	0,54	0,29
NCS3	0,38	0,04	-	0,44	-	0,38	0,26	0,21
NCS4	-	0,02	-	-	1,00	-	0,11	0,17
E1	0,13	0,04	0,01	0,71	0,19	-	0,34	0,28
E2	0,44	0,39	0,30	0,29	0,52	0,50	0,34	0,37
E3	0,42	0,57	0,69	-	0,29	0,50	0,32	0,35
D1	0,93	0,94	0,96	0,84	0,76	0,92	0,82	0,87
D2	0,07	0,06	0,04	0,16	0,24	0,08	0,18	0,13
S1	0,91	0,99	1,00	0,83	0,77	0,92	0,85	0,89
S2	0,09	0,01	-	0,17	0,23	0,08	0,15	0,11
I1	0,98	0,83	0,48	0,93	0,86	0,90	0,89	0,80
I2	0,02	0,17	0,52	0,07	0,14	0,10	0,11	0,20
T1	0,44	0,63	0,80	0,60	0,48	0,60	0,66	0,63
T2	0,53	0,34	0,20	0,40	0,48	0,33	0,31	0,35
T3	0,02	0,03	-	-	0,04	0,07	0,03	0,02
EG1	1,00	0,99	0,89	-	0,01	0,02	-	0,33
EG2	-	-	-	-	0,05	-	0,42	0,04
EG3	-	0,01	0,04	-	0,05	-	0,58	0,07
EG4	-	-	0,02	0,78	0,67	0,30	-	0,38
EG5	-	-	0,04	0,22	0,21	0,68	-	0,18
SG1	1,00	0,99	0,89	-	0,01	0,02	-	0,33
SG2	-	0,01	0,01	-	0,02	-	0,14	0,02
SG3	-	-	0,03	-	0,08	-	0,86	0,09
SG4	-	-	0,06	1,00	0,89	0,98	-	0,56
FA1	1,00	1,00	0,99	0,49	0,67	0,80	0,79	0,75
FA2	-	-	-	0,15	0,08	0,07	0,08	0,07
FA3	-	-	0,01	0,36	0,25	0,13	0,13	0,17
FB1	1,00	0,99	1,00	0,87	0,91	0,97	0,97	0,94
FB2	-	0,01	-	0,13	0,09	0,03	0,03	0,06
R1	0,38	0,20	0,14	0,32	0,28	0,20	0,27	0,24
R2	0,09	0,06	0,14	0,12	0,11	0,13	0,07	0,12
R3	0,53	0,74	0,72	0,56	0,61	0,67	0,66	0,64

– a conferma dello stadio evolutivo relativamente arretrato delle famiglie del gruppo D, il 71% dei giovani ha meno di 23 anni, l'età del 'genitore' più giovane è per il 78% dei casi inferiore ai 55 anni e nel 51% delle famiglie sono presenti 'fratelli' con 14-19 anni ('fratelli' con meno di 14 anni sono invece presenti nel 13% delle unità); tra i giovani non occupati, infine, le categorie maggiormente rappresentate sono gli studenti (17%)

e quanti sono in cerca di occupazione (16%);

- in E la quota di famiglie con almeno un giovane non occupato è relativamente elevata (24% con almeno un individuo in cerca di occupazione; 23% con almeno uno studente; 14% con almeno un inoccupato): considerando che dalle variabili attive risulta che gli aggregati con nessun giovane occupato sono solo il 16%, c'è da presumere che in questo gruppo sia piuttosto diffuso tra i giovani della stessa famiglia un regime di occupazione misto;
- a conferma dello stadio evolutivo avanzato delle famiglie del gruppo F, il 50% dei giovani ha un'età variabile tra i 27 e i 29 anni, l'età del 'genitore' più giovane è superiore ai 54 anni nel 70% dei casi e i 'fratelli' con meno di 20 anni sono relativamente assenti (nel 20% delle unità ci sono 'fratelli' di 14-19 anni e solo nel 3% 'fratelli' con meno di 14 anni); si osserva, inoltre, che nel 62% dei casi il giovane presente è di sesso maschile e che il 25% di non occupati si ripartisce pressochè equamente tra studenti, persone in cerca di occupazione e inoccupati; — nonostante la netta prevalenza di maschi (almeno quando il giovane è uno solo ed è celibe) e un basso livello di istruzione (nel 66% delle unità i giovani hanno terminato da almeno sei anni la formazione scolastica), in G una quota relativamente alta (18%) di famiglie registra almeno un giovane in cerca di occupazione;
- rispetto al profilo generale, nei capoluoghi di provincia trovano diffusione soprattutto i gruppi A e D (per contro è scarsamente rappresentata la tipologia C) mentre i gruppi B e C sono presenti con maggiore frequenza nei comuni con meno di 20.000 abitanti (dove invece il gruppo A è poco rappresentato).

La classificazione incrociata delle famiglie secondo la tipologia *ex post* appena individuata e una tipologia costruita a priori in base al criterio della presenza o meno di nuclei nell'aggregato familiare[7] consente di apprezzare ulteriormente i risultati ottenuti.

Dai dati della Tab. 4 risulta confermato che la tipologia *ex post* è in grado di esprimere ampie categorie familiari, tanto più capaci di catturare la realtà quanto piu questa si esprime in forma semplice (le famiglie complesse non emergono) e tradizionale (la ridotta mutabilità delle distribuzioni condizionate secondo tipologia *ex post* è dovuta anche alla rarità delle famiglie anucleari e delle coppie di giovani non coniugati). Essa è, comunque, preferibile alla tipologia *ex ante* che, non riuscendo a discriminare all'interno delle famiglie mononucleari (in particolare coppie con figli e nuclei monogenitore) tra quelle costituite dai giovani e quelle attribuibili ai loro genitori, trascura aspetti importanti degli aggregati che contano tra i loro componenti individui in età 20-29 anni.

---

7 Per una descrizione del metodo adottato per la costruzione della tipologia per nuclei partendo da dati individuali dei componenti la famiglia, vedi Ongaro (1990).

Tab. 4: *Famiglie con giovani secondo la classificazione ottenuta 'ex post' ed una classificazione per nuclei costruita 'ex ante'*

tipologia *ex ante*	tipologia *ex post* A	B	C	D	E	F	G	Tot.	%
fam.anucl.									
- soli	45	-	-	-	-	-	-	45	2,8
- gruppi	-	6	-	-	1	2	3	12	0,7
fam.mono-nucleari									
-coppie	-	124	-	-	-	1	-	125	7,8
-coppie+fi.	-	1	335	418	202	142	1	1.099	68,4
-monogenit.	-	4	1	-	25	2	123	155	9,7
fam.com-plesse (a)	-	-	43	63	32	20	13	171	10,6
Tot.	45	135	379	481	260	167	140	1.607	100,0
%	2,8	8,4	23,6	29,9	16,2	10,4	8,7	100,0	

(a) Per famiglie complesse si intendono quelle con un nucleo allargato ad individui aggregati e le famiglie polinucleari.

L'intersezione delle due classificazioni offre, invece, interessanti elementi informativi aggiuntivi sulla natura delle famiglie con giovani.

Considerando i gruppi D, E ed F come rappresentativi di situazioni in cui i giovani vivono ancora la condizione di figlio all'interno della famiglia dei genitori (si ricordi che tra queste famiglie le unità con giovani coniugati non superano il 2% e che sono pressoché inesistenti i 'figli'), si può concludere che la maggior parte delle famiglie sono espressione della presenza di un forte legame dei giovani con la famiglia d'origine. Ciò è vero, in particolare, per le 1099 coppie con figli che, almeno per il 69%, sono costituite dai genitori dei giovani. Ma, meno ovviamente, ciò è vero anche per le famiglie complesse la cui articolazione interna è raramente analizzata a causa delle pesanti (e non sempre affidabili nei risultati) elaborazioni informatiche che richiederebbe: la Tab. 4 mostra allora che nel campione considerato almeno il 67% delle famiglie complesse ha origini che non dipendono dalla costituzione di un nucleo da parte dei giovani dell'aggregato.

5. *Ulteriori riflessioni*

Ogni tentativo di lettura dei gruppi in chiave di comportamenti familiari giovanili va affrontato con molta cautela.

I problemi non derivano solo dall'utilizzo di dati provenienti da un'osservazione trasversale che non consente di isolare gli effetti di generazione,

tanto più insidiosi quanto più si lavora con stati soggetti a *turn-over* nell'arco di tempo considerato (ad esempio, se cambia da una generazione all'altra l'età di inizio/fine a cui i giovani entrano/escono dallo stato di famiglia unipersonale può essere fuorviante ragionare sulle proporzioni di soli per età).

La prudenza è auspicabile soprattutto a causa della difficoltà di controllare tutte quelle forme di eterogeneità tra i gruppi che possono avere effetti sulla direzione e sull'intensità di specifiche associazioni. L'incertezza non dipende solo dall'assenza di informazioni su fenomeni rilevanti per la comprensione dei risultati (vitalità dei genitori dei giovani che sperimentano una famiglia unipersonale; numero totale di fratelli viventi; durata del matrimonio delle coppie di giovani; ecc.). Problemi nascono anche quando i gruppi sono caratterizzati da variabili che, essendo correlate strutturalmente con quelle oggetto di studio (per esempio, con il crescere dell'età del giovane aumenta la probabilità di trovare giovani occupati, con livello di istruzione più elevato o con genitori più anziani; con il crescere del numero di giovani in famiglia aumenta, a parità di altre condizioni, la probabilità di trovare almeno un giovane occupato in lavori indipendenti) non consentono di valutare le differenze di associazioni al netto di fattori perturbanti.

Le associazioni di variabili espresse dai gruppi suggeriscono, tuttavia, alcune riflessioni che, pur nella loro generalità, permettono di gettare un primo sguardo sul complesso e ancora relativamente sconosciuto tema dei comportamenti familiari in età 20-29 anni.

1. Il matrimonio porta alla costituzione da parte dei giovani di nuclei autonomi rispetto alla famiglia d'origine (B e C) e ciò è tanto più vero quanto più la coppia è senza figli (B). Genitori e figli coniugati formano semmai un unico aggregato quando questi ultimi hanno avviato il processo riproduttivo (C) e, nel caso questo non fosse ancora iniziato, la coabitazione ha luogo solo in presenza di un unico genitore (G). La natura dei dati disponibili non consente di fornire spiegazioni del fenomeno; tuttavia, già in questa sede è possibile avanzare alcune ipotesi, non necessariamente alternative, che non trovano smentite nei risultati ottenuti:

- la coabitazione di un nuovo nucleo con i genitori avviene preferibilmente in concomitanza con particolari 'debolezze' dei genitori quali l' età avanzata (C) e/o la vedovanza (G);
- la nascita di un figlio, alterando l'organizzazione ed i bisogni della coppia giovane, facilita il ripristino di più stretti contatti (compresa la coabitazione) con la famiglia d'origine di uno dei due componenti (C);
- le condizioni socio-economiche e culturali che si associano alla presenza di coppie precocemente feconde stanno alla base anche di un'eventuale coabitazione della coppia con la famiglia d'origine (C).

2. Le uscite dalla famiglia d'origine per cause diverse dal matrimonio (A), non sono tanto associate ad un'età avanzata (nonostante che la proporzione di soli aumenti al crescere dell'età il salto di densità più rilevante avviene tra la classe 20-23 e la classe 23-26) o ad un sesso particolare (tra i non coniugati i maschi sono tuttavia il 58%), quanto all'acquisizione di un'indipendenza economico-lavorativa. Confrontando la proporzione di occupati del

gruppo A con quella di D ed F (che hanno entrambi un solo giovane in età variabile nell'arco 20-29) si osserva infatti che tale quota (82%) è decisamente più alta che negli altri gruppi dove gli stessi valori oscillano da un minimo di 63% ad un massimo di 75%. La condizione di solo, infine, sembra privilegiare relativamente giovani con istruzione medio-superiore. Le spiegazioni, anche in questo caso, possono essere molteplici, non ultima quella che tiene conto di eventuali selezioni dovute al tipo di occupazione (dipendente-indipendente; in posizione elevata o bassa) sperimentata da questi giovani. E' possibile, tuttavia, osservare che tale risultato si giustifica anche nell'ipotesi che la famiglia unipersonale rappresenti un'esperienza temporanea che tende a concludersi con il matrimonio e che non inizia immediatamente dopo la conclusione degli studi: in tal caso, nella classe 20-29 quanti hanno un basso livello di istruzione dovrebbero aver in gran parte già contratto matrimonio e quanti sono invece in possesso di un diploma di laurea, o hanno raggiunto un'età sufficientemente elevata da transitare direttamente allo stato di coniugato oppure stanno ancora vivendo in famiglia d'origine in attesa di costituire *unità* a se stante.

3. Se l'indipendenza economica è requisito necessario per la costituzione di una nuova famiglia, essa non è, tuttavia, sufficiente a determinarla. Oltre il 70% degli aggregati in cui i giovani vivono con i genitori ha almeno un individuo tra 20 e 29 anni occupato. Considerando che ciò è vero anche per il gruppo F, dove i giovani hanno tutti un'età superiore ai 24 anni e la maggioranza di essi ha già da tempo concluso il ciclo scolastico, si può escludere che la presenza di un'occupazione, anche quando è associata ad un prolungato periodo di indipendenza economica (nell'ipotesi che non si manifestino effetti selettivi rispetto alla storia lavorativa), sia sufficiente a far uscire comunque dalla famiglia d'origine.

4. La relativamente elevata quota di casi con giovani in cerca di occupazione (18%) e, più in generale, non occupati (36%) - non giustificabile né con l'età, né con la presenza di studenti, ne' con una consistente quota di unità con casalinghe coniugate a mariti di età superiore ai 29 anni - fa presumere che le famiglie di G, oltre al disagio dovuto all'assenza di uno dei genitori, sperimentino anche quello di una maggiore debolezza economica.

5. Poco si può aggiungere a quanto si è già detto sull'interazione esistente tra indipendenza economico-lavorativa dei giovani e tipologia familiare sperimentata. Le caratteristiche dei gruppi, fortemente condizionate da variabili di natura demografica[8] , l'impossibilità di controllare elementi importanti per la spiegazione di alcuni risultati (ad esempio, il numero complessivo di figli della famiglia d'origine) e la doppia valenza che può assumere il concetto di giovane in cerca di occupazione (può trattarsi di un disoccupato o di un

---

8 Rispetto alle variabili sulla condizione occupativa dei giovani, i gruppi B e C non sono tra loro confrontabili perchè l'età mediamente più anziana delle coppie di C esclude più spesso dalla considerazione le condizioni occupazionali dei coniugi (presumibilmente maschi) in età superiore ai 29 anni; i gruppi D ed F, pur contando in famiglia un unico giovane, sono, invece, eterogenei per quanto concerne età e sesso dello stesso.

individuo in cerca di prima occupazione, e in questo caso differenze nei gruppi rispetto al tempo trascorso dai giovani dopo la conclusione della formazione scolastica diventano rilevanti) rendono difficile districarsi nell'intreccio delle variabili in gioco. Abbandonato ogni tentativo di approfondire aspetti particolari del fenomeno (del tipo: la condizione occupazionale dei giovani che restano in famiglia d'origine interagisce con quella degli altri fratelli coetanei o di età inferiore ai 20 anni o con la probabilità degli stessi di uscire dalla famiglia d'origine?), si osserva, tuttavia, che alcuni risultati offrono spunti per ulteriori elementi di riflessione.

- I valori sistematicamente più elevati che altrove delle proporzioni di giovani in cerca di occupazione di D, E, F[9] e G suggeriscono che condizioni di dipendenza economica si associano più frequentemente a una permanenza da non coniugato in famiglia d'origine.
- L'elevata proporzione di famiglie con almeno un giovane occupato in lavori indipendenti osservabile in C (rapportando le quote di O2 e di O2+O3 si ottiene livelli del 30% contro valori massimi degli altri gruppi del 22%), non imputabile esclusivamente a cause strutturali[10] , fa sospettare che la presenza di un'occupazione indipendente non sia, in generale, occasione per prolungare, da non coniugato, la permanenza in famiglia d'origine. I dati, per la verità, non ci dicono se tale occupazione è stata assunta prima o dopo il matrimonio; tuttavia, considerando che in C è più frequente che l'unico giovane eventualmente catturato sia la donna, si può avanzare l'ipotesi che, per ragioni ancora da precisare (per es. fattori culturali), il lavoro autonomo per la donna sia un elemento che favorisce l'assunzione congiunta di responsabilità coniugali e riproduttive.
- Collegando quanto appena osservato con quanto che è stato rilevato per le famiglie unipersonali, non si può escludere che in questa fascia d'età il lavoro indipendente, contrariamente a ciò che si era ritenuto inizialmente, sia un fattore di allontanamento del giovane dalla famiglia d'origine

Le riflessioni ed i quesiti disseminati in questa ultima parte del capitolo precisano ulteriormente limiti e potenzialità di questo tipo di fonti e di analisi. Le conclusioni cui si è pervenuti possono essere arricchite con analisi esplorative su altre aree geografiche e differenti periodi temporali al fine di evidenziare diversità ed analogie di comportamento delle popolazioni.

L'aggiunta di qualche variabile (per esempio, il tipo di professione) o la conduzione di esplorazioni in sottogruppi sufficientemente numerosi di fami-

---

9 Il dato di F (8%) potrebbe essere spiegato dalla più bassa incidenza del fenomeno nella popolazione di età 25-29 e in particolare dalla più contenuta presenza di individui in cerca di prima occupazione. Ciò non esclude che l'età non abbia comunque effetti sul rischio di allontanarsi dalla famiglia d'origine (per es. per matrimonio) e attraverso questo su una modifica della condizione professionale o più specificamente su una conversione dello stato non occupazionale (per es. donne coniugate che si dichiarano casalinghe).

10 Il dato è più alto anche rispetto ad E che conta sempre due giovani. Rispetto a B, inoltre, l'affermazione è ancora più vera se si considera la relativamente maggiore femminilizzazione delle coppie di C (il coniuge che può sfuggire perchè di età superiore ai 29 anni è più probabile sia di sesso maschile) e la maggior presenza (oltre il 50%) di famiglie con inoccupati (presumibilmente casalinghe).

glie possono far emergere più chiaramente l'esistenza o meno di associazioni (positive o negative) che i gruppi attualmente formatisi lasciano magari un po' in ombra. Elaborazioni su altre fonti di natura trasversale che, a differenza dell'indagine sulle forze di lavoro non si basano sulla famiglia anagrafica (censimenti, recenti indagini Istat sulle famiglie), possono cogliere sia forme familiari meno tradizionali che una maggiore articolazione dei comportamenti negli aggregati più numerosi.

Un sostanziale passo avanti nell'approfondimento di queste tematiche può, tuttavia, essere fatto solo disponendo di fonti mirate di natura longitudinale. La conoscenza dei tempi di accadimento dei diversi eventi che interessano il giovane (fine scolarità, inizio e fine di eventuali occupazioni, inizio di matrimonio e/o convivenza, ecc.) ed i suoi parenti più prossimi coabitanti e non e dello sfondo socio-culturale ed economico in cui questi si inseriscono contribuirebbe a gettare una luce determinante su come viene vissuta questa fase delle vita delle persone sotto il profilo familiare. In tale contesto potrebbero, tra l'altro, trovare applicazioni interessanti anche le tecniche esplorative di analisi multivariata.

14

FORME FAMILIARI E CARATTERISTICHE DELL'OCCUPAZIONE

Francesco Sanna, Isabella Santini e Silvano Lauro

1. *Premessa*

Individuare regole, tendenze ed omogeneità familiari di approccio al mercato del lavoro, significa essenzialmente analizzare la famiglia secondo un duplice punto di vista:

- uno, che si può definire *interno*, si prefigge come obiettivo l'analisi dei legami demografici ed economici esistenti all'interno del nucleo familiare;
- un secondo, che si può definire *esterno*, parte dalla considerazione che la famiglia svolge, in veste di intermediaria tra propri componenti e sistema socio-economico, l'importante funzione di mediazione e rielaborazione degli stimoli provenienti dall'esterno, e quindi anche dal mercato del lavoro.

La fisionomia interna ed esterna dei nuclei è generata dalla connessione di numerosi fattori (storici, sociali, economici) che insieme concorrono a delineare distinti profili familiari, molteplici forme aggregative ed organizzative delle quali si ha, ancora, un quadro frammentario sia in termini di evoluzione spaziale che temporale.

Approfondire alcuni fenomeni connessi con le strategie interne ed esterne più comunemente adottate dai nuclei familiari, eventuali legami e nessi causali tra scelte individuali e strategie familiari, tendenziali omogeneità di approccio della famiglia al mercato del lavoro è lo scopo che ci si è prefissi. L'approfondimento di queste tematiche avviene tramite l'analisi delle informazioni desumibili dall'indagine Istat sulle forze di lavoro, limitatamente alle regioni Veneto e Lombardia, sui dati del I trimestre 1986.

2. *Le strategie organizzative interne familiari*

L'unità familiare è stata da sempre intesa come sede di soddisfazione dei bisogni, affettivi, sociali, economici, e come luogo in cui, secondo strategie ottimali, viene organizzata l'attività per il procacciamento delle risorse necessarie per soddisfare tali bisogni. Da ciò sono progressivamente derivati nella famiglia notevoli elementi di rigidità (una precisa divisione dei ruoli tra chi lavora in casa e chi produce per il mercato; la tendenza a far sopravvivere i nuclei di tipo esteso al fine di rendere più immediata ed esauriente la

soddisfazione dei bisogni dei singoli), che hanno favorito, tra l'altro, il costituirsi di nuclei con una configurazione assimilabile a quella delle 'imprese familiari' (Del Boca e Turvani , 1979a e 1979b).

Tuttavia, le progressive trasformazioni economiche e sociali, ed in particolare lo sviluppo della società industriale e la tendenziale urbanizzazione del territorio (trasformazioni che hanno offerto gradualmente alla famiglia la possibilità di soddisfare gran parte dei propri bisogni all'esterno) hanno contribuito, nel l'ultimo decennio, a rendere maggiormente flessibile e a semplificare la struttura dei nuclei. Ne sono testimonianza le stesse statistiche ufficiali (Censimenti generali della popolazione, 1971 e 1981) dalle quali emergono alcune precise tendenze demografiche e socio-economiche a livello nazionale, quali: un progressivo restringimento della dimensione del nucleo familiare; per i giovani, l'acquisizione di una più precoce autonomia dai genitori; una tendenza verso la separazione dei nuovi nuclei familiari da quelli di origine; un sensibile aumento dell'occupazione femminile .

L'analisi delle informazioni desumibili dall'indagine sulle forze di lavoro (in particolare il confronto territoriale tra le due regioni prese in esame) offre importanti spunti di riflessione.

(a) Congiuntamente alle accennate tendenze demografiche e socio-economiche , si rilevano mutamenti piuttosto significativi nel modo in cui la famiglia si organizza. Se è, infatti, vero che la famiglia da un lato è ancora legata a schemi organizzativi classici (è il caso, ad esempio, dei nuclei familiari in cui, accanto alla figura dell'occupato, si ritrova costantemente quella della casalinga), dall'altro èanche vero che essa tende gradualmente a proiettarsi verso forme non tradizionali (è tipico l'esempio di famiglie di soli occupati) anche se con *trends* non omogenei a livello territoriale, proprio in virtù della persistenza in Italia di realtà regionali molto diverse tra loro (Tab. 1).

(b) In questo graduale proiettarsi verso forme non tradizionali e verso una maggiore flessibilità organizzativa, la famiglia tende ad indebolirsi, in quanto, probabilmente, il sistema sociale non è ancora in grado o non trova, comunque, conveniente farsi carico di indispensabili funzioni sociali ed economiche, in sostituzione del nucleo familiare. La famiglia tende, quindi, ad abbandonare forme organizzative classiche solo quando ha conseguito una propria indipendenza economica ed è, nello stesso tempo, in grado di garantire un sostegno socio-assistenziale ai suoi componenti.

Infatti, se è pur vero che la tipologia 'solo occupati ' presenta connotati piuttosto interessanti riguardo all'incidenza degli isolati (il 22,24 % in Veneto e il 27,32 % in Lombardia), al sesso e al titolo di studio del capofamiglia (molte le donne capofamiglia; prevalenza di un livello di istruzione medio-superiore), quando ci si sofferma ad analizzare l'età del capofamiglia secondo la tipologia demografica familiare, essa risulta inferiore alla media generale solo in corrispondenza delle tipologie di coppia e di coppia con figli (Tab. 2). Ciò può in parte dipendere dal fatto che, salvo casi particolari (separazione, vedovanza,) è, in genere, ancora difficile, che un giovane, anche se autonomo dal punto di vista economico,

Tab. 1 : *Caratterizzazione delle famiglie con almeno un occupato*

Famiglie	Veneto (%)	Lombardia (%)
Solo occupati	25,53	33,92
Presenza di almeno una casalinga	55,91	43,64
Presenza di almeno un ritirato	20,57	18,89

Tab. 2 : *Eta' media del capofamiglia nelle famiglie di solo occupati*

Tipologia familiare	Veneto	Lombardia
Isolati	43,16	41,26
Gruppi	41,38	43,18
Coppia	35,60	37,05
Coppia con figli	36,81	37,67
Nucleo monogenitore	45,58	43,07
Nucleo esteso	45,66	42,74
Complesso	38,65	38,93

Tab. 3 : *Distribuzione percentuale delle famiglie con almeno un occupato, secondo la tipologia demografica e la tipologia familiare (condizione unica o prevalente). Composizione % per colonna*

Regione e Tipologia demografica	Solo occupati	Presenza casalinga	Presenza ritirato lavoro	Presenza studente,leva, in cerca occ.	Totale
VENETO					
Gruppi	2,25	0,42	1,72	0,15	1,09
Coppia	22,11	8,35	6,22	2,04	11,62
Coppia con figli	67,53	72,98	51,01	80,93	68,89
Nucleo monogenitore	6,05	4,68	12,76	5,66	7,75
Nucleo esteso	2,06	13,57	28,29	11,22	10,65
Totale	100,00	100,00	100,00	100,00	100,00
LOMBARDIA					
Gruppi	1,83	0,49	2,76	0,46	1,38
Coppia	23,74	8,61	7,66	1,51	12,66
Coppia con figli	65,94	79,00	56,58	84,15	70,97
Nucleo monogenitore	7,36	5,66	17,15	7,45	9,80
Nucleo esteso	1,13	6,24	15,85	6,43	5,19
Totale	100,00	100,00	100,00	100,00	100,00

Tab. 4 : *Percentuale di famiglie con almeno un occupato nelle quali almeno un componente svolge un lavoro a tempo parziale*

Tipologia familiare (Condiz. unica o prevalente)	Veneto	Lombardia
Solo occupati	9,54	6,41
Altra tipologia in cui è assente la casalinga	11,19	7,23
Presenza di almeno una casalinga	4,29	2,60
Totale	7,04	4,93

lasci la famiglia di origine. Lo troverà conveniente solo quando formerà un nuovo nucleo e, in questo caso, il coniuge costituirà un punto di riferimento non trascurabile nell'economia familiare.
Si aggiunga, inoltre, che un'influenza significativa sulla scelta della strategia organizzativa familiare più opportuna la esercitano i figli, in quanto è, in genere, proprio la loro presenza a spingere il coniuge di sesso femminile a modificare il suo status da occupata a casalinga (Tab.3).

(c) Vale la pena, infine, osservare che anche le strategie organizzative di tipo non tradizionale soffrono di alcune rigidità, in quanto poco si differenziano dalle forme classiche nella sostanza dei ruoli assunti dai componenti. Come si rileva, infatti, dalla Tab. 4 è proprio nelle famiglie in cui è assente la figura della casalinga che si registra una maggiore presenza di componenti con un lavoro a tempo parziale, uno stato occupazionale indispensabile quando il nucleo deve far fronte con le proprie forze sia alle esigenze economiche che a quelle socio-assistenziali dei singoli. Il maggior peso che questo fenomeno assume nelle famiglie in cui, oltre all'assenza della casalinga, vi è almeno un componente in condizione non professionale (studente, in servizio di leva, in cerca di occupazione, ritirato dal lavoro) indica quanto questa necessità si consolidi nel momento in cui al componente occupato venga richiesto non solo di partecipare attivamente alla conduzione della casa ma anche di occuparsi dei bisogni dei giovani e degli anziani. In questo modo si conferma, anche, il fatto che queste due ultime figure costituiscono dei veri e propri componenti a carico.

In definitiva, le classiche configurazioni occupazionali familiari, pur non nascondendo quelle rigidità cui sopra si accennava, riescono ancora bene ad adattarsi alla realtà odierna. Infatti, oltre a procurare un vantaggio diretto all'economia globale del nucleo, attivano, ancora, delle convenienze non trascurabili per i singoli componenti. Quando tali strategie organizzative non risultano direttamente attuabili, la famiglia opta per una forma di tipo sostitutivo, quale un lavoro *part-time* per almeno un componente occupato, l'aggregazione in gruppi di isolati (soprattutto in presenza di un anziano) o, in casi particolari, la dipendenza dal nucleo di origine (costituendo nuclei monogenitori o estesi, come si rileva dalla Tab.3) attuando, in entrambi questi ultimi due casi, reciproche convenienze per i componenti la famiglia: per le coppie o gli individui soli, occupati, protezione e sicurezza familiare; per i genitori o altri parenti un sostegno di tipo economico e/o assistenziale.

3. *Le strategie professionali della famiglia: alcuni criteri di valutazione e principali tendenze in atto*

Da tutto ciò sembra evidenziarsi una tendenza alla semplificazione della configurazione interna della famiglia, con una maggiore flessibilità ed intercambiabilità dei ruoli al suo interno, pur permanendo alcune rigidità tipiche delle forme organizzative classiche. Parallelamente, emergono forme nuove di approccio al mercato del lavoro, nuove figure professionali nell'agricoltura,

nell'industria e nel terziario, e le attività economiche di tipo tradizionale acquistano una nuova fisionomia. La famiglia, attraverso le scelte professionali dei suoi componenti, segue questo processo evolutivo del mercato del lavoro adattandosi ad esso gradualmente, pur se in modo non uniforme a livello territoriale. A questo proposito può risultare interessante individuare come questo processo abbia inciso sulle scelte professionali dei singoli, non tanto nelle famiglie con un solo occupato, dove le conseguenze del processo sono piuttosto scontate (esse si identificano, principalmente, in un graduale adattamento alla reale offerta di posti di lavoro), quanto in quelle con più di un occupato, dove la scelta professionale del singolo è sempre in qualche modo condizionata da quella degli altri componenti, e dove una strategia professionale piuttosto che un'altra può direttamente incidere sui risultati economici del nucleo familiare (si pensi sempre ai vantaggi che possono derivare in una famiglia da un'attività imprenditoriale comune). Dall'indagine sulle forze di lavoro risulta che, in Veneto, nel 49,82 % delle famiglie con almeno due occupati (pari al 30,95 % del totale delle famiglie della regione), i componenti dichiarano di svolgere la medesima professione; tale percentuale supera il 50 % in Lombardia, dove queste famiglie rappresentano il 31,57 % del totale (Tab. 5). In Veneto, nel 60,03 % delle famiglie con almeno due occupati i componenti dichiarano di svolgere la professione nel medesimo ramo di attività; l'omologa percentuale per la Lombardia è del 57,85 % (Tab. 6). Per avere un quadro sufficientemente esaustivo del grado di omogeneità-disomogeneità delle professioni in un nucleo familiare, è necessario analizzare congiuntamente settore di attività e professione svolta e tener presente che il concetto di omogeneità professionale in una famiglia è, tra l'altro, funzione della configurazione economica del territorio in cui essa risiede. Allo scopo di individuare, quindi, associazioni e connessioni tra profili professionali e settori di attività e validi criteri di interpretazione del concetto di omogeneità-disomogeneità professionale-settoriale in un nucleo familiare (pur con i limiti imposti dallo scarso dettaglio delle modalità proposte dal questionario dell'indagine ISTAT) si sono sottoposte le due tavole di contingenza 'tipologia familiare posizione nella professione' e 'tipologia familiare ramo di attività economica' all'analisi delle corrispondenze semplici. Dall'analisi di dettaglio degli autovalori e della percentuale di variabilità spiegata da ciascun asse fattoriale (Tab. 7) risultano significativamente interpretabili solo i primi tre assi.

Dalla lettura congiunta del secondo e terzo asse fattoriale si individuano zone di omogeneità professionale-settoriale familiare. L'asse 2 oppone, infatti, il settore di attività primario a quello terziario, evidenziando al tempo stesso delle zone intermedie di disomogeneità. L'asse 3, invece, contrappone posizioni professionali omogenee a tipologie con un alto grado di disomogeneità. Sul piano generato dal secondo e terzo asse fattoriale è possibile, quindi, individuare quattro zone ben distinte (Figg. 1 e 2 e Tab. 8)[1]:

---

1 Il primo asse fattoriale costituisce, invece, un indice del grado di 'dipendenza-indipendenza' delle professioni in un nucleo familiare.

Tab. 5 : *Distribuzione percentuale delle famiglie con almeno due occupati per 'tipologia familiare: posizione nella professione'*

Posizione/i	Veneto	Lombardia
Solo imprenditori	-	0,14
Solo professionisti	0,17	0,42
Solo lavoratori in proprio	5,52	3,06
Solo coadiuvanti	0,10	0,11
Solo dirigenti	0,28	0,14
Solo impiegati	11,39	18,01
Solo operai	32,36	29,05
Professioni miste	50,18	49,07
Totale	100,00	100,00

Tab. 6 : *Distribuzione percentuale delle famiglie con almeno due occupati per 'tipologia familiare: ramo di attività economica'*

Ramo/i	Veneto	Lombardia
Solo agricoltura	4,12	1,29
Solo industria	28,14	33,00
Solo servizi	25,84	22,04
Solo Pubblica Amministrazione	1,93	1,52
Rami misti	39,97	42,15
Totale	100,00	100,00

Tab. 7 : *Analisi delle corrispondenze semplici : percentuale di variabilità spiegata da ciascun asse fattoriale*

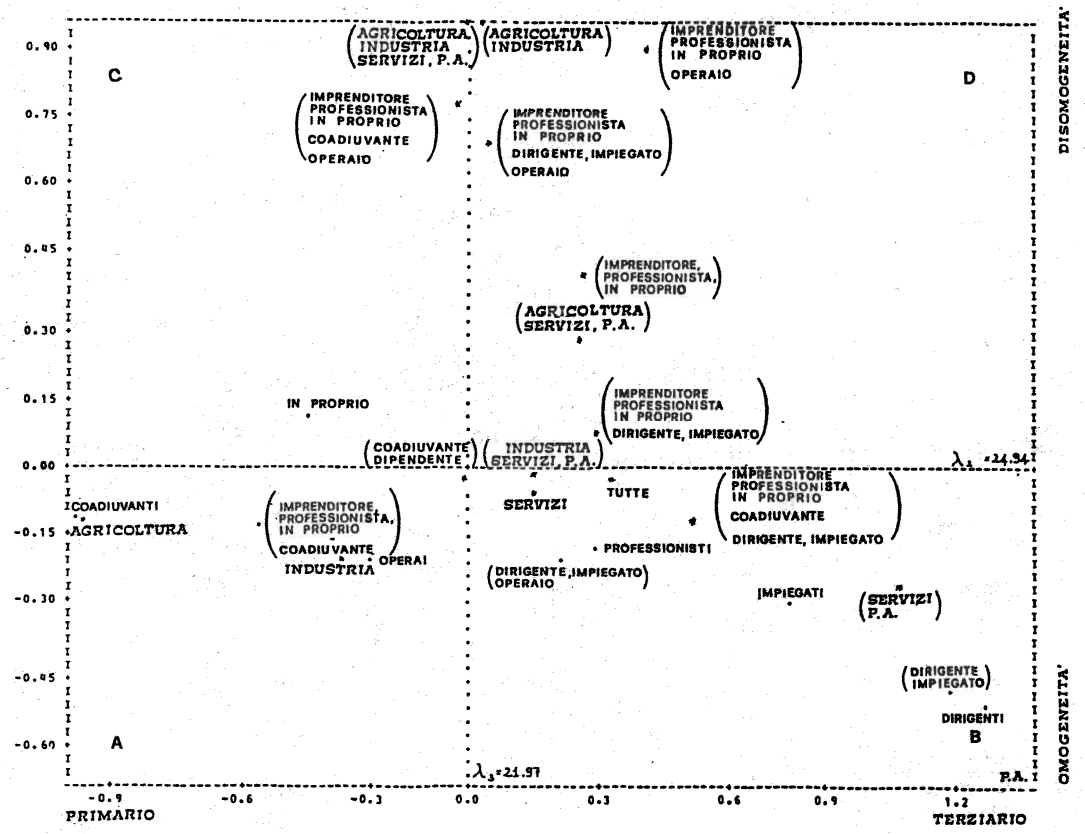
Regione	Asse							
	1	2	3	4	5	6	7	8
Veneto	41,67	24,94	21,97	4,42	3,75	1,59	1,27	0,39
Lombardia	47,36	27,71	12,73	4,24	3,58	3,32	0,69	0,37

Tab. 8: *Distribuzione percentuale delle famiglie per tipologia omogenea professione-settore (A - D)*

Regione	Tipologia				
	A	B	C	D	Totale
Veneto	42,94	26,04	7,04	23,98	100,00
Lombardia	36,53	38,22	3,16	12,09	100,00

Fig. 1: *Regione Veneto: proiezione sul piano fattoriale (assi 2 e 3) dei punti-modalità: 'tipologia familiare: posizione nella professione' e 'tipologia familiare: ramo di attività'*

A. area omogenea agricoltura industria;
B. area omogenea terziario ;
C. area disomogenea agricoltura-industria ;
D. area disomogenea (agricoltura-industria) e servizi.

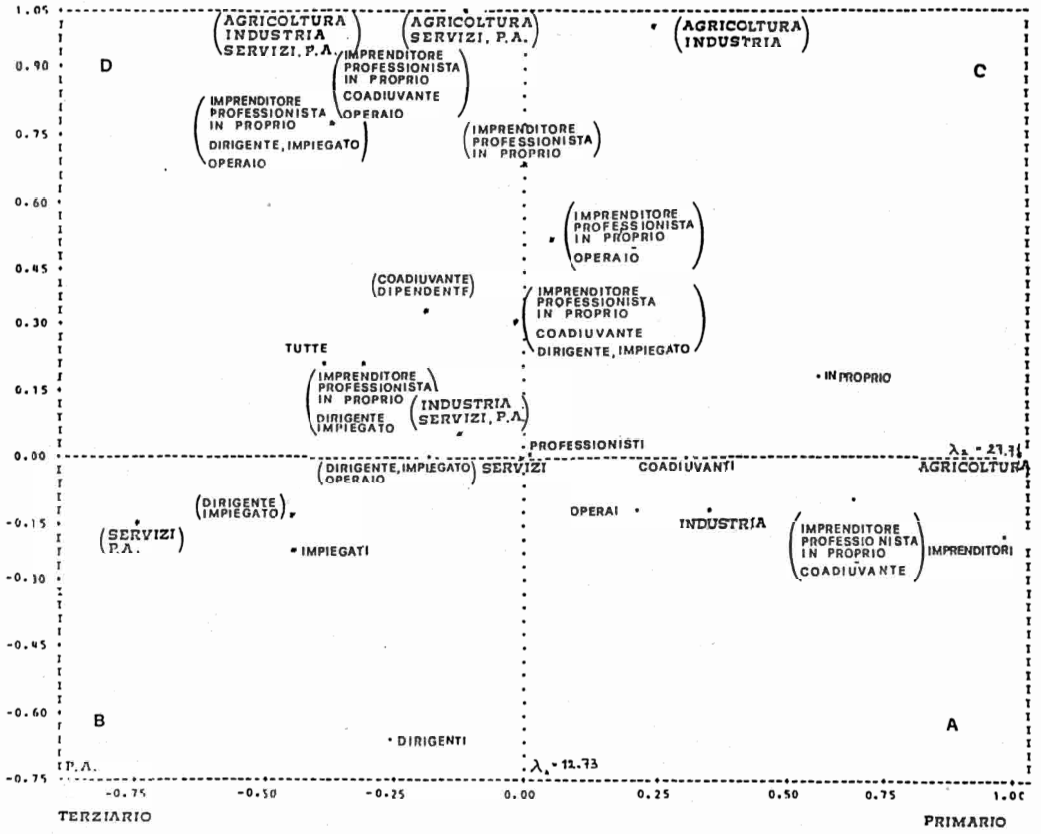


Sia in Veneto che in Lombardia prevale una tendenza dei componenti il nucleo familiare ad adottare strategie professionali omogenee (tipologie A e B; 68,98 % delle famiglie in Veneto e 74,75 % in Lombardia), anche se queste è ovvio che si differenzino in funzione della configurazione economica del territorio in cui la famiglia risiede.

Emerge, inoltre, molto chiaramente la fase di transizione che la famiglia, in Veneto, sta attraversando per quanto riguarda le strategie adottate dai componenti in tema di collocazione professionale: un cammino graduale da agricoltura a terziario, a stadio più avanzato in Lombardia (tipologia D). In questa fase di transizione, vengono gradualmente meno delle forme di omo-

Fig. 2: *Regione Lombardia: proiezione sul piano fattoriale (assi 2 e 3) dei punti modalità: 'tipologia familiare: posizione nella professione' e 'tipologia familiare: ramo di attività'*

A. area omogenea agricoltura industria;
B. area omogenea terziario ;
C. area disomogenea agricoltura-industria ;
D. area disomogenea (agricoltura-industria) e servizi.



geneità professionali familiari (nei nuclei estesi nel settore agricolo), ma altre se ne creano (nelle coppie nel terziario), anche se assumono connotati ben diversi dalle prime: tra l'altro non concorrono a delineare nuove imprese familiari. Non a caso, infatti, le forme imprenditoriali familiari sono ancora prevalentemente di tipo agricolo e industriale.

La prevalente disomogeneità che si riscontra in Lombardia rispetto ai settori agricoltura-industria (tipologia C) va interpretata con attenzione. Infatti, non si tratta tanto di una forma di disomogeneità quanto di una tendenza dei due settori a coincidere; questa considerazione è ampiamente confermata dai risultati esposti nel seguito.

4. *Caratterizzazione socio-professionale della famiglia*

La lettura degli atteggiamenti e dei comportamenti delle famiglie in merito alle strategie organizzative interne ed esterne da esse adottate, può essere arricchita notevolmente dall'analisi delle corrispondenze multiple, che consente di estrarre l'essenziale delle informazioni contenute nella tavola dei dati di riferimento e di fornirne una rappresentazione che si presti ad una più immediata interpretazione. Essa permette, infatti, di sintetizzare e valutare simultaneamente il comportamento degli indicatori di riferimento, di individuare il grado e il verso delle interrelazioni esistenti tra il complesso delle variabili e tra variabili e individui (nel caso in esame, le famiglie), il tutto con la minor perdita di informazioni possibile. L'analisi è stata condotta, sia per la regione Veneto che per la regione Lombardia, a partire da 21 variabili attive[2], che si possono raggruppare secondo due tematiche (vedi la Tab. 9 per un maggior dettaglio delle caratteristiche delle tavole di dati di riferimento):

(i) Caratteristiche demo-socio-economiche del capofamiglia (variabili originarie, desunte cioè direttamente dal questionario dell' indagine);

(ii) Caratteristiche demo-socio-economiche della famiglia (variabili derivate, ottenute per semplice aggregazione di informazioni elementari del questionario o conseguenti alla definizione di tipologie[3] delle famiglie)[4].

Se l'analisi viene estesa a tutte le famiglie, il primo asse fattoriale, da solo, assorbe il 10% della variabilità totale del fenomeno, contrapponendo, in sostanza, le famiglie senza occupati a quelle con almeno un occupato e celando altri aspetti caratteristici, e forse più rilevanti, della fisionomia socio-professionale delle famiglie. D'altra parte, questo risultato era atteso, dato che sia in Veneto che in Lombardia le famiglie senza occupati sono il 30% circa del totale. Si è quindi ritenuto opportuno, per meglio approfondire il fenomeno in esame, limitare l'analisi alle famiglie con almeno un occupato.

Un primo strumento di interpretazione e lettura dei risultati ottenuti è costituito dalle proiezioni sui piani fattoriali dei punti modalità attivi, le quali permettono abbastanza agevolmente di desumere le corrispondenze esisten-

---

2 Variabili di tipo A (capofamiglia): sesso, età, titolo di studio, stato civile, condizione professionale, posizione nella professione, branca di attività economica, tempi di lavoro, carattere dell'occupazione, seconda attività, ore di lavoro nella settimana. Variabili di tipo B (famiglia): tipologia lavorativa, ramo di attività, carattere dell'occupazione, tempi di lavoro, numero di attività svolte, posizione nella professione, presenza/assenza di non occupati che svolgono attività, bilancio formativo familiare, tipologia demografica 1 (componenti per età), tipologia demografica 2(tipo di nuclei). L'elenco completo di tutte le modalità attive, contenente la distribuzione semplice di ciascuna variabile, distintamente per le due regioni e per le famiglie con almeno uno e con almeno due componentioccupati, è in Sanna, Santini e Lauro (1988).

3 Per una dettagliata decrizione delle tipologie qui impiegate, vedi Ongaro (1987).

4 Le variabili di tipo A tendono a migliorare sensibilmente la descrizione dei profili tipici familiari. Si sarebbero potute aggiungere variabili attive riguardanti le caratteristiche demografiche e socio-economiche del coniuge. E' ormai convinzione diffusa, infatti, che la posizione sociale della famiglia non derivi più esclusivamente da quella del suo capo, bensì dai profili e dalle attitudini di entrambi i coniugi. In quest'ottica, si dovrebbero rivedere anche le 'vere' posizioni di ciascun componente all'interno degli altri nuclei-gruppi, nuclei estesi, ecc.. Così operando si introdurrebbero però notevoli margini di soggettività e di incertezza interpretativa, con benefici marginali rispetto agli obiettivi primari della ricerca, che ha privilegiato, in questa fase, l'analisi generale rispetto ad una più particolareggiata. Per una più dettagliata disamina di queste problematiche, vedi Santini, Sanna e Lauro (1989).

Tab. 9 *Analisi delle corrispondenze multiple: caratteristiche delle tavole di dati*

	N. famiglie	%	% Inerzia
Totale famiglie			
Veneto	1.494.938	100,00	13,49
Lombardia	3.314.884	100,00	13,39
Famiglie con almeno un occupato			
Veneto	1.040.891	69,63	10,73
Lombardia	2.263.734	68,29	10,61
Famiglie con almeno due occupati			
Veneto	462.743	30,95	11,71
Lombardia	1.046.457	31,57	10,75

ti tra una o più modalità e simiglianze in profilo tra famiglie, e in definitiva di costituire approssimativamente un'ossatura interpretativa del meccanismo di formazione di raggruppamenti tipici di famiglie.

La percentuale di inerzia spiegata dai primi due assi fattoriali supera il 10% (Tab.9): un buon risultato, se si pensa alle dimensioni dello spazio vettoriale di origine (140 coordinate) e alla configurazione della tavola dei dati, espressa in forma disgiuntiva completa.

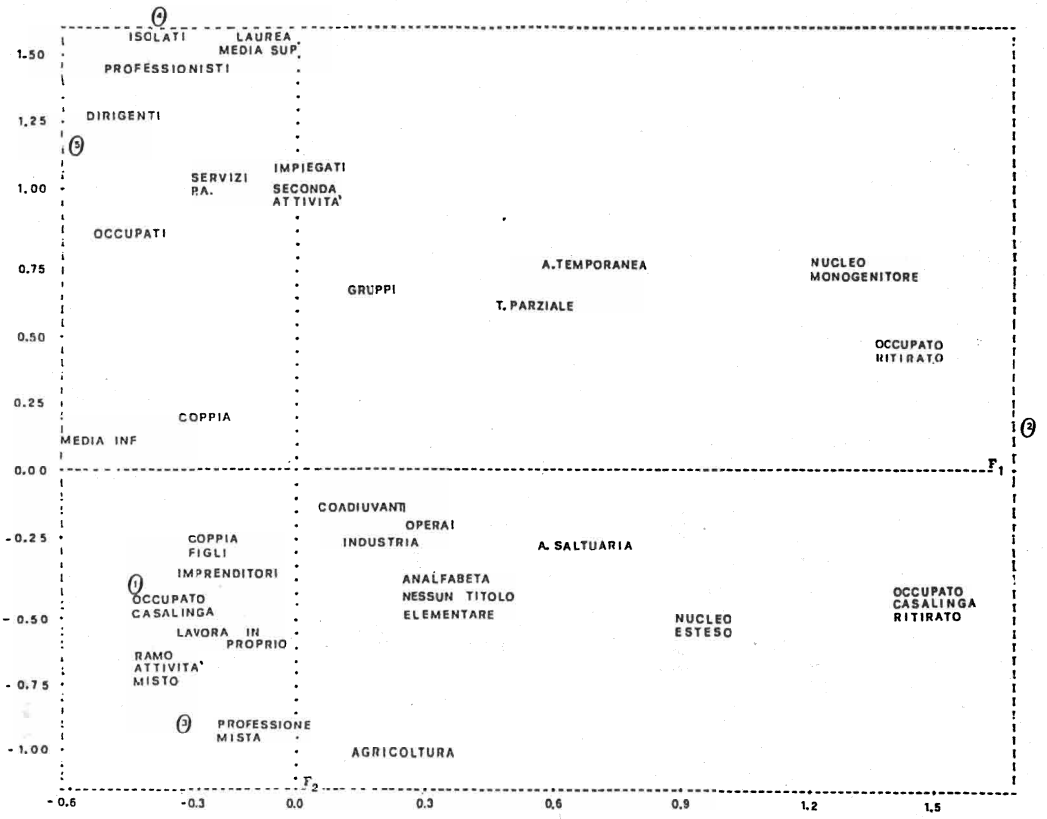
Per evitare un'eccessiva dispersione nell'interpretazione dei risultati, l'analisi è stata quindi limitata al piano principale, (proposto dalla Fig. 3 per il solo Veneto, in quanto non emergono sostanziali differenze dall'analisi comparata dei risultati delle due regioni). Sul piano principale sono riportati i punti modalità attivi più caratteristici, i quali consentono di cogliere le corrispondenze esistenti tra modalità e di affermare l'esistenza di un nesso significativo tra il primo asse fattoriale e l'insieme delle variabili che, pur secondo aspetti differenti, esprimono il grado di occupazione all'interno del nucleo familiare (condizione prevalente, età, grado di istruzione, stato civile, ecc.). Infatti, emerge chiaramente come al decrescere dei valori del primo asse fattoriale risulta più rara la presenza di ritirati dal lavoro, diminuisce l'età del capofamiglia, ecc.. Il secondo asse fattoriale, invece, meglio descrive la configurazione 'interna' ed 'esterna' del nucleo, sia da un punto di vista demografico che socio-economico-strutturale. A parità di grado di occupazione, quindi, le famiglie tendono a spaccarsi in due grandi tipologie:

(a) una matrice classica, con una configurazione organizzativa equilibrata (coppie con figli, costante presenza della casalinga, ecc..) e un approccio al mercato del lavoro di tipo tradizionale (nell'agricoltura o nell'industria, a tempo pieno, ecc.);

(b) una seconda, di costituzione più recente, meno tradizionale, ma nella quale è presente un forte stato di precarietà (più marcato in Lombardia che in Veneto). In questa tipologia muta, naturalmente, anche l'approccio al mercato del lavoro soprattutto dei nuclei isolati, giovani, maggiormente proiettati verso le nuove professioni tipiche del terziario avanzato.

Tra tali due grandi tipologie esistono delle zone intermedie che evidenziano il cammino lento e graduale attraverso il quale le forme familiari si trasformano. Nelle professioni si sta tendendo ad una sostanziale omoge-

Fig. 3: *Regione Veneto: proiezione sul piano principale dei punti-modalità attivi (caratteristiche della famiglia). Famiglie con almeno un occupato*



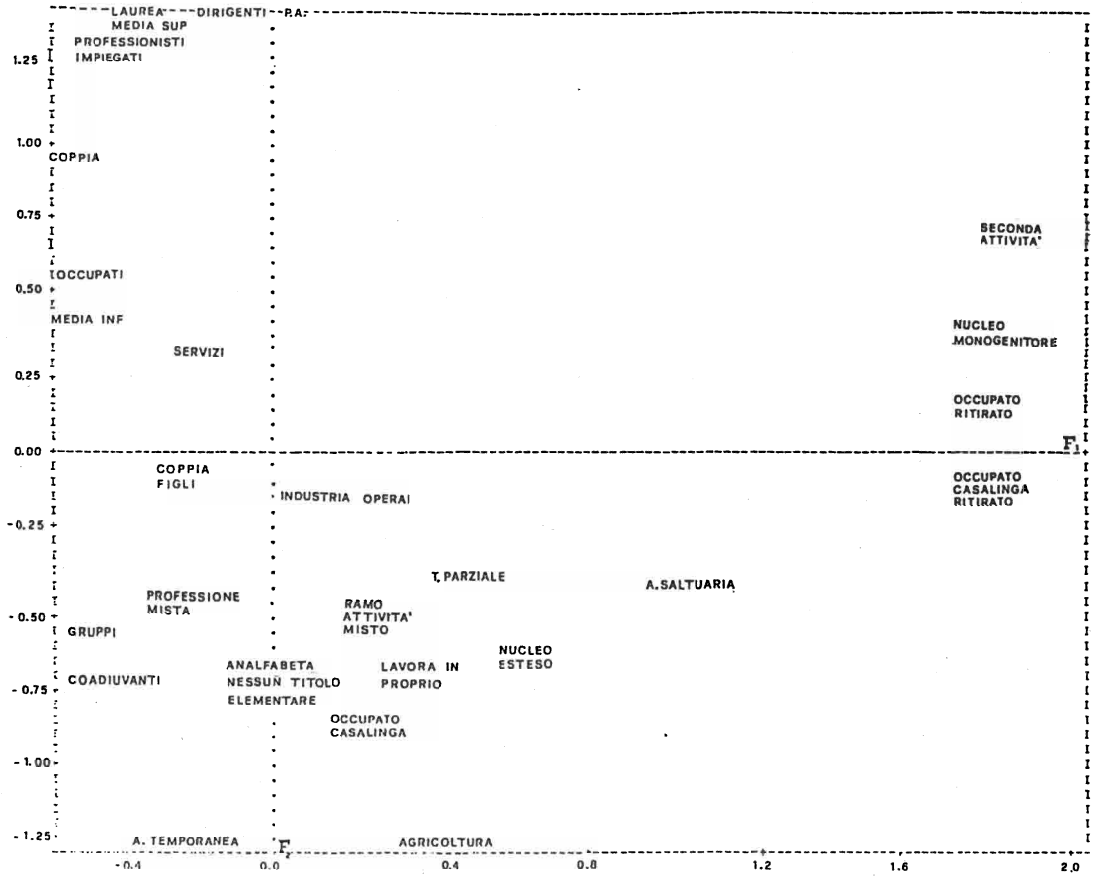
neità di tipo terziario, secondo le linee evolutive della configurazione economica delle due regioni.

I risultati confermano quanto già accennato in precedenza, e cioè che il costituirsi di forme organizzative non tradizionali è fortemente condizionato dalla configurazione demografica del nucleo e da alcune specifiche caratteristiche dello stesso, quale, ad esempio, la presenza dei figli. L'emergere di forme non tradizionali è ancora dettato da motivazioni che in parte esulano dalla volontà dell'individuo (si veda a tale proposito la posizione sul piano fattoriale dei separati).

Differenze sensibili emergono dall'analisi comparata dei risultati delle due regioni rispetto a fenomeni specifici:

– in Veneto ancora ben marcata è la presenza di una realtà agricola piuttosto consolidata, che, invece, in Lombardia tende a confondersi con quella di tipo industriale, con tutti gli effetti socio-economici e demografici che ne derivano (quali la presenza, nella prima, di famiglie 'estese di

Fig. 4: *Regione Veneto: proiezione sul piano principale dei punti-modalità attivi (caratteristiche della famiglia). Famiglie con almeno due occupati*



tipo complesso' che da sole costituiscono delle vere e proprie imprese familiari);

– in Lombardia emerge, in modo piuttosto marcato, il processo di semplificazione e di innovazione delle forme organizzative e professionali familiari, che si desume da una più netta distinzione, nell'ambito delle forme nuove, tra nuclei di nuovissima costituzione e nuclei più consolidati. I primi manifestano un modo nuovo di approccio al mercato del lavoro, professionale nel terziario avanzato, ma complessivamente più instabile rispetto a forme nuove più consolidate (dirigenti e impiegati nella Pubblica Amministrazione e nel settore creditizio). In Veneto, questo fenomeno risulta ancora piuttosto confuso.

Il quadro generale non muta sostanzialmente se l'analisi viene condotta rispetto alle sole famiglie con almeno due occupati (Fig.4).

Infatti, ancora una volta il primo asse fattoriale è interpretabile come indice sintetico del grado di occupazione del nucleo familiare mentre il

secondo asse esplicita meglio, da un punto di vista demografico e socio-economico, la configurazione 'interna' ed 'esterna' del nucleo,definendo anche in questo caso due distinte tipologie familiari, una a matrice classica e una con connotati meno tradizionali. I risultati offrono delle importanti conferme su alcune tendenze già in parte emerse. In particolare:

- la famiglia tende ad abbandonare forme organizzative classiche a favore di forme meno tradizionali solo quando può garantire ai suoi componenti adeguati benefici economici ed assistenziali. Come infatti emerge dall'analisi comparata dei risultati ottenuti (famiglie con almeno un occupato e famiglie con almeno due occupati) i giovani solo se costretti abbandonano la famiglia di origine per creare da soli un nuovo nucleo e, se ciò avviene, generano inevitabilmente elementi di debolezza che si riscontrano soprattutto nel particolare tipo di lavoro svolto (temporaneo, seconda occupazione, etc.). Nelle coppie e nelle coppie con figli, invece, questi elementi di debolezza tendono a scomparire, segno quindi dell'importanza del ruolo svolto dal coniuge nell'economia familiare, sia se occupato, apportando un'altra fonte di reddito, sia se in condizione non professionale, dedicando il proprio tempo alla cura della casa e dei figli;
- ancora una volta, in Lombardia, si trova coincidenza tra realtà agricola ed industriale e un progressivo affacciarsi dell'imprenditoria familiare nel terziario, fenomeni questi ancora piuttosto confusi in Veneto;
- una conferma, rispetto a quanto già in precedenza anticipato, si ha anche riguardo alle strategie professionali comunemente adottate dai nuclei familiari. Il secondo asse sintetizza adeguatamente le tendenze regionali in atto: un progressivo spostamento della famiglia verso attività omogenee nel terziario più di tipo impiegatizio che di tipo imprenditoriale; d'altro canto vanno anche scomparendo progressivamente forme imprenditoriali familiari nell'agricoltura, con un tendenziale adattamento delle stesse alla mutata configurazione del mercato del lavoro.

5. *Raggruppamenti tipologici delle famiglie*

In questa parte, operando sui medesimi dati già analizzati nelle sezioni precedenti, si tenta di giungere a raggruppamenti omogenei delle famiglie delle due regioni in esame, in relazione a variabili socio-demografiche (del capofamiglia e dell'intera famiglia) e professionali.

Il fine principale di queste analisi è di carattere esplorativo (mancano, infatti, almeno per quanto concerne la realtà italiana, schemi e/o modelli di riferimento). *A posteriori*, peraltro, si tenta di capire se ed in quale misura i raggruppamenti cui si è pervenuti possano configurarsi come reali tipologie familiari in relazione alle caratteristiche occupazionali delle famiglie considerate.

Il metodo utilizzato è quello della *cluster analysis*[5] (privo, perciò, di

5 Le elaborazioni sono state effettuate utilizzando il programma SPAD - (vedi Lebart, Morineau,1985).

specifiche ipotesi distribuzionali delle osservazioni). I risultati presentati nel seguito sono quelli relativi ai raggruppamenti apparsi di maggiore e più chiaro significato, tra l'altro con notevoli similitudini tra le due regioni, pur sensibilmente diverse sia sotto il profilo socio-economico che sotto quello più strettamente demografico.

Mentre indicatori frequentemente impiegati in numerose analisi socio-economiche (quali, ad esempio, il reddito per abitante, la percentuale di addetti all'agricoltura, i consumi di energia elettrica, la densità telefonica, etc.) forniscono una immagine abbastanza differente delle due regioni, le analisi qui presentate hanno invece dato risultati per molti versi assai simili. Vero è che le due regioni sono limitrofe, con frequenti ed intensi scambi (e perciò con ampie possibilità di esportazione di modelli demografici, sociali, di comportamento), ma le analogie riscontrate vanno spesso ben al di là di quanto ci si poteva inizialmente attendere.

Le notevoli similitudini emerse dalle analisi sulle due regioni consentono una lettura parallela dei principali risultati. Tale parallelismo deve già considerarsi, in sè, un importante risultato, stanti le differenze tra le due realtà indagate, cui si è appena accennato.

5.1. *Famiglie con almeno un occupato*

Le analisi condotte sulle famiglie con almeno un occupato hanno portato alla individuazione di 5 gruppi in Veneto e 6 in Lombardia. Tali gruppi, in sintesi, si caratterizzano come segue.

Il gruppo 1 (*operai*) si caratterizza, in entrambe le regioni, per la presenza di famiglie con capofamiglia operaio, un'età media del capofamiglia più bassa di quella media generale, la quasi totale assenza di addetti all'agricoltura, lo scarso numero di famiglie in cui sono presenti anziani. Numericamente, è il gruppo più consistente: comprende infatti il 37% delle famiglie nel Veneto ed oltre il 41% in Lombardia.

Il capofamiglia, maschio, è piuttosto giovane (età modale 30-39 anni), è sposato ed ha almeno un figlio (in genere di meno di 10 anni). Nella sua famiglia si individua un solo nucleo; sono in genere assenti gli anziani (o, comunque, i ritirati dal lavoro).

Il bilancio formativo familiare è modesto: entrambi i coniugi non hanno proseguito gli studi oltre l'obbligo, e così pure (spesso) i figli, se ultraquattordicenni.

Il capofamiglia, come detto, ha una qualifica operaia; la sua occupazione è a tempo pieno ed ha carattere permanente; il settore in cui lavora è industriale in senso stretto (manifatturiero, di trasformazione) o è quello delle costruzioni; raramente svolge (o, almeno, dichiara di svolgere) una seconda attività.

La moglie, se lavora (il che accade più spesso in Lombardia che in Veneto), ha una occupazione con qualifica operaia, come il marito.

La caratteristica tipologica familiare di coppia (abbastanza giovane) con figli fa sì che in tali famiglie spesso non vi siano altri occupati oltre il

capofamiglia e, eventualmente, la moglie.

La configurazione demografica tipo è quella della coppia con figli.

Il gruppo 2 (*ritirati dal lavoro*) si caratterizza (in modo molto netto) per la presenza di famiglie con capofamiglia ritirato dal lavoro, quindi di età elevata (mediamente oltre i 60 anni). Sono nuclei anziani, in cui raramente sono presenti componenti in età inferiore a 14 anni. Le famiglie caratterizzate da nuclei estesi (o da 2 o più nuclei) si presentano in sovrafrequenza in questo gruppo, che comprende il 18% delle famiglie nel Veneto e poco più del 15% in Lombardia.

Il capofamiglia, non di rado di sesso femminile, è in genere ultrasessantenne, spesso vedovo/a.

La configurazione demografica tipo è quella del nucleo monogenitore (con figlio, adulto, in genere di meno di 40 anni, occupato). Allorchè sono presenti entrambi i coniugi, il più delle volte sono entrambi ritirati dal lavoro.

Il bilancio formativo familiare è assai modesto: il capofamiglia è molte volte privo di qualsiasi titolo di studio; anche l'adulto/i presente/i difficilmente ha/hanno un titolo di studio che va oltre l'obbligo.

La tipologia lavorativa familiare è mista, vi è cioè almeno un occupato adulto, accanto ad uno o più ritirati dal lavoro.

In queste famiglie vi è, in genere, un solo componente occupato, nell'industria, con qualifica operaia.

E' da notare che questo gruppo è, tra tutti quelli individuati nelle diverse analisi, quello che più e meglio si differenzia da ogni altro. Occorre altresì rilevare che tale netta tipologizzazione ha connotati assai più spiccatamente demografici che occupazionali. Infatti analizzando, il gruppo in un'ottica non familiare, ma focalizzando l'attenzione sull'unico (in genere) adulto che lavora per famiglia, le sue caratteristiche occupazionali non si discostano apprezzabilmente da quelle degli adulti occupati delle famiglie del gruppo 1 (operai), per quel che concerne le famiglie venete. In Lombardia, non emergono connotazioni di rilievo né rispetto al livello culturale né rispetto al ramo ed alla posizione nella professione del (o dei) componente/i occupato/i.

Il gruppo 3 (*lavoratori in proprio*) presenta una spiccata connotazione legata al settore di attività. Si caratterizza infatti per la presenza di famiglie agricole, con capofamiglia non più giovane (mediamente ha poco più di 50 anni), lavoratore indipendente agricolo. Tra gli altri componenti occupati, spesso qualcuno lo è nel settore terziario (ma non il coniuge, che lavora pure in agricoltura, come lavoratore in proprio o come coadiuvante).

Il numero di ore settimanalmente lavorate dalla famiglia in complesso è sensibilmente più elevato di quello medio (75 ore contro 60). Ciò dipende in parte dal più alto numero di occupati per famiglia che caratterizza le famiglie di questo gruppo, in parte dal maggior impegno orario settimanale di ciascun occupato.

Il gruppo comprende il 20% delle famiglie in Veneto (regione tuttora a più spiccata vocazione agricola), circa il 9% delle famiglie in Lombardia.

Spesso si tratta di famiglie abbastanza numerose.

Del gruppo fanno parte, con caratteristiche occupazionali simili, anche famiglie il cui capofamiglia è lavoratore in proprio nel commercio; egli associa

a sé, nella conduzione dell'attività, il coniuge ed uno o più altri componenti della famiglia.

Il bilancio formativo familiare è basso: capofamiglia e coniuge in genere non hanno più della licenza elementare, in relazione anche alla loro età, in genere più avanzata (classe di età modale 50-59 anni per entrambi) di quella degli altri gruppi di famiglie con capofamiglia occupato.

Il basso livello di istruzione dà conto della qualifica operaia del coniuge, se questi svolge attività lavorativa in settori extraagricoli. I figli, se presenti, non hanno proseguito gli studi oltre l'obbligo.

Nel gruppo sono relativamente frequenti le famiglie con 3 occupati: capofamiglia e coniuge in agricoltura, figlio nella industria, generalmente con qualifica operaia.

Questo gruppo è quello che più di ogni altro si caratterizza sotto l'aspetto delle variabili occupazionali di settore, soprattutto in Veneto. Ciò può essere ricollegato alla vocazione agricola che tuttora caratterizza vaste aree (geografiche, ma anche sociali) del Veneto.

Peraltro, si riscontra anche qui, come in altre zone del Paese, la tendenza dei figli a non restare in agricoltura, ma a cercare lavoro nell'industria. Potrebbe trattarsi, in definitiva, di un gruppo in via di estinzione: laddove il processo di industrializzazione è più avanzato (Lombardia) questo gruppo, pur presente, è assai meno numeroso e, soprattutto, presenta contorni meno nitidi.

Il gruppo 4 (*isolati*) nel Veneto è il meno numeroso tra quelli individuati (comprende il 7% delle famiglie) e si caratterizza per la presenza di isolati (spesso di sesso femminile), senza figli.

Il capofamiglia (in genere unico componente della famiglia) è celibe o nubile, con sovrafrequenza di donne (spesso separate). Nel caso di famiglie con più di un componente, la tipologia demografica è quella del 'gruppo di parenti'; quindi, in ogni caso, una famiglia senza nuclei.

Il capofamiglia è occupato a tempo pieno, prevalentemente nei servizi, a vario livello (come testimonia l'ampio ventaglio dei titoli di studio, che vanno dalla licenza elementare alla laurea); la qualifica lavorativa, di conseguenza, è talora operaia, talora impiegatizia.

Si tratta quindi di un gruppo con una precisa caratterizzazione demografica (le famiglie senza nuclei), ma privo di tratti ben delineati sotto l'aspetto occupazionale. Considerazione in sè non marginale, ma di scarso rilievo in un contesto in cui si mira a studiare i comportamenti familiari rispetto al mercato del lavoro: questo gruppo comprende infatti individui isolati o, comunque, gruppi senza nuclei. Le decisioni rispetto al lavoro sono quindi decisioni certamente 'individuali', svincolate cioè (nè potrebbe essere altrimenti, visto il contesto) da considerazioni di tipo 'familiare'.

In Lombardia, invece, pur in presenza di una decisa sovrafrequenza di famiglie senza nuclei (soprattutto isolati), nel gruppo (che comprende poco meno dell'8% delle famiglie) confluiscono tipologie demografiche abbastanza varie. Gli isolati sono occupati, di medio livello culturale, piuttosto giovani, e fra di essi hanno un peso non trascurabile le donne separate. La condizione sociale delle famiglie di questo gruppo è piuttosto varia; le qualifiche diri-

genziali e le libere professioni sono comunque scarsamente presenti.

Il gruppo 5 (*impiegati, dirigenti, liberi professionisti*) comprende circa il 18% delle famiglie in Veneto e si differenzia dagli altri per il più elevato livello culturale dei membri delle famiglie che lo compongono (i giovani sono quasi tutti studenti) e per la prevalenza di occupati nel terziario e nella Pubblica Amministrazione.

Il capofamiglia (maschio, tra i 30 e i 50 anni) è sposato con figli, di cui almeno uno di meno di 14 anni.

La configurazione demografica tipo è quella della coppia con figli.

Il bilancio formativo familiare è decisamente alto: il capofamiglia è diplomato o laureato, e così pure il coniuge; i figli (se presenti) continuano a studiare oltre l'obbligo.

Talora il capofamiglia è il solo occupato, nei settori dei servizi, del credito e delle assicurazioni, della Pubblica Amministrazione (se lavoratore dipendente), con qualifica impiegatizia o dirigenziale; più spesso anche la moglie lavora, nei medesimi settori.

Del gruppo fanno parte anche le famiglie il cui capofamiglia è libero professionista. Anche in questo caso la moglie spesso lavora, nei settori suindicati.

Ciò che caratterizza questo gruppo rispetto ad altri (in particolare rispetto al gruppo 1), è il più elevato livello culturale delle famiglie che ad esso appartengono (ed al quale sono correlate le diversità nei settori e nelle qualifiche occupazionali).

In Lombardia, il gruppo, per così dire, si spacca in due: da un lato gli impiegati (gruppo 5a), cui ben si attagliano le considerazioni appena svolte per l'omologo gruppo del Veneto, e che comprende oltre il 21% delle famiglie della regione; dall'altro i dirigenti e liberi professionisti (gruppo 5b): è il gruppo meno numeroso (5% delle famiglie) e si caratterizza per l'elevato bilancio formativo familiare: entrambi i coniugi sono laureati o, quanto meno, uno laureato ed uno diplomato.

Il capofamiglia, in genere laureato, ha mediamente 45 anni ed è coniugato con figli; anche la moglie in genere lavora (con qualifica impiegatizia).

La composizione demografica tipo è quella della coppia con figli; nel gruppo sono però pure presenti, con discreta frequenza, gli isolati.

Caratteristica delle famiglie di questo gruppo è l'elevato impegno lavorativo settimanale del capofamiglia: oltre 45 ore, in media.

Nella Tab. 10 sono sintetizzate le caratteristiche dei gruppi appena descritti.

È da notare, altresì, che tra i diversi gruppi individuati permangono taluni elementi di omogeneità, in particolare quanto a:

(a) numero medio di componenti (di ogni età) per famiglia, che si mantiene tra 3,5 e 3,7 per quasi tutti i gruppi, con la sola rilevante eccezione del quarto (quello degli isolati), nel quale scende a 1,2;

(b) numero medio di figli per famiglia[6], che è assai simile nei vari gruppi

---

6 Si ricorda al riguardo che si tratta dei figli facenti parte della famiiglia al momento dell'intervista, non di tutti i figli che una coppia (o una donna) ha avuto fino al momento dell'intervista.

Tab. 10: *'Cluster analysis ' : famiglie con almeno un occupato*

Gruppo	Veneto			Lombardia		
	Famiglie N.	%	Inerzia intragruppo	Famiglie N.	%	Inerzia intragruppo
1	384.678	37,0	0,10659	932.322	41,2	0,10490
2	185.236	17,8	0,11896	349.463	15,4	0,07936
3	209.542	20,1	0,06994	177.857	7,9	0,05070
4	72.146	6,9	0,06967	198.547	8,7	0,04561
5 (5a)	189.293	18,2	0,09558	492.935	21,8	0,07053
(5b)				112.646	5,0	0,03570
Tot.	1.040.895	100,0		2.263.770	100,0	
Inerzia intragr.		0,73808			0,77848	
			= 0,61567			= 0,66806
Inerzia totale		1,19882			1,16528	

(intorno a 1,4 - 1,5), tranne che nel quarto, per il quale vale 0,1;

(c) scarsa presenza di ascendenti e/o altri parenti: gli uni e gli altri si ritrovano con una maggiore (ma comunque bassa) frequenza nelle famiglie agricole e, i secondi, anche in quelle con capofamiglia ritirato dal lavoro.

Le principali caratteristiche demografiche ed occupazionali delle famiglie appartenenti ai diversi gruppi sono riportate nella Tab. 11.

Alcune peculiarità emerse da altre analisi effettuate meritano almeno un cenno.

La prima (rilevante dal punto di vista demografico e, in prospettiva, per i riflessi sul mercato del lavoro) riguarda, per il Veneto, il gruppo degli operai, che, in un'analisi a 7 gruppi, si 'spacca' in relazione alla differente età del

Tab. 11: *Caratteristiche demografiche ed occupazionali delle famiglie con almeno un occupato.*

Gruppo	Eta' media del capo-famiglia		N. medio di occupati per famiglia		N. medio di non occupati (>14anni) per famiglia	
	Veneto	Lombardia	Veneto	Lombardia	Veneto	Lombardia
1	40,4	43,1	1,57	1,68	1,08	1,07
2	63,1	61,8	1,39	1,31	1,95	1,77
3	50,0	50,7	1,86	1,75	1,34	0,98
4	44,1	39,3	1,04	1,03	0,14	0,04
5 (5a)	40,9	40,8	1,52	1,64	1,09	0,91
(5b)		45,4		1,43		0,88
Complesso	46,7	45,8	1,55	1,55	1,22	1,03

capofamiglia. Da un lato si raggruppano le famiglie meno giovani (quelle con capofamiglia ultraquarantenne), che presentano un bilancio formativo familiare più basso ed una dimensione familiare mediamente più elevata (di conseguenza, non sono infrequenti le famiglie con più di due occupati). Dall'altro, quelle più giovani (capofamiglia sotto i 40 anni), che si distinguono, principalmente, per un bilancio formativo familiare meno basso (il capofamiglia ha completato la scuola dell'obbligo) ed una dimensione familiare (forse non finale, però) più ridotta.

Un secondo spunto di riflessione deriva dalla enucleazione, anche nel Veneto, di un gruppo delle famiglie dei dirigenti e liberi professionisti (già classificate prevalentemente nel gruppo 5). La sua caratteristica principale è la marcata sovrafrequenza di addetti nel settore dei servizi (dirigente il capofamiglia, dirigente o impiegata la moglie). Un'altra peculiarità del gruppo (transitoria, però, in quanto legata all'età dei figli) è rappresentata dalla compresenza di famiglie con rapporto non occupati/occupati nullo (famiglie con figlio/i di meno di 14 anni) e famiglie con rapporto pari a due o tre (famiglie con figli ultraquattordicenni, studenti).

Per la Lombardia, è interessante notare che, riducendo a cinque i gruppi, scompare il gruppo delle famiglie agricole, in larga parte riassorbito nel gruppo 1 (operai), a riprova dell'accorciamento di distanze (sotto tutti i punti di vista: dallo stile di vita, al livello culturale, alle scelte occupazionali, ecc.) tra le famiglie operaie e quelle contadine in un contesto di agricoltura avanzata e largamente industrializzata (e di una piccola e piccolissima industria distribuita sul territorio), quale quello lombardo.

5.2. *Famiglie con almeno due occupati*

Tra le famiglie del Veneto con almeno un occupato, il 36% circa annovera al suo interno due o più occupati. In Lombardia, le famiglie con almeno due occupati sono oltre il 46% delle famiglie in cui almeno un componente lavora. Su questi sottoinsiemi di famiglie sono state ripetute le analisi precedentemente illustrate relativamente a tutte le famiglie con almeno un occupato, al fine di individuarne eventuali peculiarità (soprattutto in termini di similarità delle scelte occupazionali).

Scompare, evidentemente, in queste analisi il gruppo degli isolati (gruppo 4), ed i gruppi che si formano ricalcano sostanzialmente le caratteristiche dei gruppi individuati considerando il complesso delle famiglie. Per quanto concerne il Veneto, si trovano ancora una volta i gruppi:

- dei ritirati dal lavoro (gruppo 2; 12% delle famiglie). Si tratta di nuclei estesi, con capofamiglia appunto ritirato dal lavoro (di età media oltre i 60 anni), assai spesso di sesso femminile, vedova. A lavorare sono due (o più) dei figli rimasti in famiglia, o un figlio/a e la nuora/genero;
- dei lavoratori in proprio dell'agricoltura e del commercio (gruppo 3; 16% delle famiglie). In questo gruppo è frequente trovare più di due componenti occupati, tutti nel medesimo settore di attività (in pratica, nella

stessa azienda familiare);
- degli impiegati, dirigenti, liberi professionisti (gruppo 5; 17% delle famiglie). Prevale nettamente in questo gruppo la tipologia della coppia (giovane - mediamente, il capofamiglia ha poco meno di 40 anni -) con figli (minori di 14 anni); a lavorare sono dunque entrambi i coniugi.

L'altro gruppo precedentemente individuato (quello degli operai) si scinde in due gruppi, tra i quali le principali differenze sono rappresentate dall'età dei coniugi e dal secondo membro della famiglia occupato:
- in un primo gruppo (1a; 21% delle famiglie) il secondo occupato della famiglia è il coniuge; si tratta di coppie molto giovani, con figli (se presenti) in tenera età, in cui entrambi i coniugi lavorano (sotto questo profilo, sono famiglie assimilabili a quelle del gruppo degli impiegati , dalle quali si distinguono per il più basso livello culturale e, di conseguenza, la diversa qualifica lavorativa);
- nell'altro (1b; 34% delle famiglie, il più numeroso) il secondo occupato della famiglia non è il coniuge, ma un altro membro, in genere un figlio/a. Prevale qui una tipologia lavorativa familiare mista, in cui è presente la figura della casalinga. Sono famiglie non più giovani (il capofamiglia, mediamente, ha quasi 50 anni), nelle quali non è infrequente trovare più di un membro della seconda generazione occupato.

In Lombardia, la caratterizzazione dei 5 gruppi che si ottengono (operai, ritirati dal lavoro, lavoratori in proprio, impiegati, dirigenti e liberi professionisti) è sostanzialmente la medesima illustrata per il complesso delle famiglie con almeno un occupato.

Se si effettua un raggruppamento a 6 gruppi, accade lo stesso fenomeno descritto per il Veneto: il gruppo degli operai si spacca in due tronconi, tra i quali le differenze più rilevanti sono anche qui quelle rispetto all'età del capofamiglia ed alla condizione del coniuge.

La Tab. 12 sintetizza le principali caratteristiche dei gruppi risultanti dalle analisi sulle famiglie con almeno due occupati.

Le analogie tra le due realtà indagate sono molto rilevanti: per i tre gruppi più numerosi (operai non più giovani, operai giovani, impiegati) i valori medi del numero di occupati per famiglia sono gli stessi nelle due regioni (Tab. 13), con una variabilità interna ai gruppi omologhi assai simile. Solo la dimensione media della famiglia nel suo complesso appare leggermente più elevata per le famiglie del Veneto.

6. *Alcune considerazioni conclusive*

Le similitudini nei risultati ottenuti per le due regioni esaminate confermano quanto illustrato nelle sezioni precedenti. In particolare, può essere utile proiettare i centri dei gruppi ottenuti nelle analisi sulle famiglie con almeno un occupato *sul piano* costituito dai primi due assi fattoriali di cui all'analisi delle corrispondenze multiple effettuata sui medesimi dati (Fig. 3).

Le considerazioni fin qui svolte trovano ulteriori conferme, soprattutto per quanto concerne:

Tab. 12: *Cluster analysis: famiglie con almeno due occupati*

Gruppo	Veneto			Lombardia		
	Famiglie N.	%	Inerzia intragruppo	Famiglie N.	%	Inerzia intragruppo
1a	97158	21,0	0,05746	211372	20,2	0,04730
1b	159239	34,4	0,07832	290845	27,8	0,05883
2	54862	11,9	0,11285	88581	8,5	0,09134
3	71516	15,4	0,07406	173710	16,6	0,07609
5 (5a)	79974	17,3	0,13450	208417	19,9	0,05052
(5b)				73557	7,0	0,03885
Tot.	462749	100,0		1046482	100,0	
Inerzia intragr.		0,74647			0,75801	
Inerzia totale		1,20366	= 0,62017		1,12094	= 0 ,67622

Tab. 13: *Caratteristiche demografiche ed occupazionali delle famiglie con almeno due occupati*

Gruppo	Eta' media del capo-famiglia		N. medio di occupati per famiglia		N. medio di non occupati (>14anni) per famiglia	
	Veneto	Lombardia	Veneto	Lombardia	Veneto	Lombardia
1a	33,0	35,4	2,02	2,02	0,04	0,03
1b	48,6	48,9	2,37	2,37	1,25	1,18
2	62,5	60,9	2,31	2,24	1,94	1,73
3	52,9	49,7	2,42	2,34	1,18	0,90
5 (5a)	39,7	38,0	2,04	2,04	0,42	0,32
(5b)		43,3		2,05		0,53
Complesso	46,1	44,7	2,24	2,19	0,92	0,73

- la netta separazione tra il gruppo delle famiglie con capofamiglia ritirato dal lavoro e tutti gli altri gruppi, in entrambe le regioni;
- la minore distanza tra le famiglie agricole (e, in definitiva, la loro meno chiara caratterizzazione rispetto ad altri settori di attività) e quelle operaie, in una realtà più avanzata;
- la differenziazione, invece, tra impiegati da un lato e dirigenti e liberi professionisti dall'altro in Lombardia, mentre in Veneto i due gruppi sono in pratica confusi in uno, pur se anche in questa regione emergono segnali di diversificazione;
- l'esistenza, in ciascuna regione, di un gruppo in posizione eccentrica, con caratteristiche legate fondamentalmente alla presenza in essi di famiglie senza nuclei (gli isolati, in primo luogo). Può essere l'indicazione

dell'affermarsi progressivo di modelli familiari diversi da quelli tradizionali, anche per aree cui spesso si associano stereotipi tradizionali.

Le notevoli somiglianze nei risultati ottenuti non fanno sembrare azzardato ipotizzare la possibilità di generalizzare le tipologie ad altre regioni, sia rispetto ad aspetti abbastanza scontati, quali il rilievo del bilancio formativo - individuale e familiare - nella segmentazione rispetto alla posizione nella professione del capofamiglia (operaio, impiegato, dirigente) e la diversità - peraltro in via di attenuazione - tra il settore agricolo e tutti gli altri, sia rispetto ad altri finora meno indagati e considerati, quali l'opportunità di una distinta considerazione delle famiglie il cui capofamiglia sia un ritirato dal lavoro e l'autonoma rilevanza, in analisi di questo tipo, di tutte quelle tipologie familiari caratterizzate da assenza di nuclei.

L'estensione dell'analisi ad altre regioni, con differenti caratterizzazioni occupazionali, potrà poi consentire di evidenziare alcuni fenomeni qui rimasti piuttosto in ombra: si pensi, ad esempio, ai fenomeni di familismo che, nelle regioni studiate, sono solo debolmente affiorati per il settore agricolo e per il commercio, ma che, altrove, sono presumibilmente ben più consistenti e generalizzati ad altri settori (industria, servizi, Pubblica Amministrazione).

15

DISOCCUPAZIONE E RICERCA DI LAVORO: ANALISI ESPLORATIVE DELL' 'ATTACHMENT' AL MERCATO DEL LAVORO E DELLA SUA DINAMICA

Enrico Rettore, Nicola Torelli e Ugo Trivellato

1. *Introduzione*

I criteri basilari per la misura degli aggregati che descrivono la partecipazione al lavoro si ritrovano nelle raccomandazioni dell'International Labour Office (ILO, 1983). In particolare, la definizione di disoccupazione (in senso lato, alla quale corrisponde nelle statistiche ufficiali italiane l'aggregato "persone in cerca di occupazione") si fonda su tre criteri. Sono considerate disoccupate le persone che nel corso del periodo di riferimento, generalmente una settimana, sono: (i) senza lavoro; (ii) disponibili a lavorare; (iii) alla ricerca di un lavoro, nel senso che hanno compiuto specifici passi nell'arco di un determinato periodo recente per cercare lavoro.

Tipicamente, le indagini sulle forze di lavoro dei vari Paesi utilizzano definizioni della disoccupazione di massima coerenti con le raccomandazioni dell'ILO. Eppure, la misura della disoccupazione resta una questione controversa. Anche lasciando da parte il basilare dibattito teorico sulla nozione di disoccupazione - frizionale e/o di disequilibrio, secondo la terminologia di Malinvaud (1984) - e restringendo l'attenzione ad approcci più operativi, le opinioni permangono significativamente diversificate. Per illustrare il punto è sufficiente richiamare due aspetti del dibattito: (i) i ricorrenti tentativi in favore di strategie di rilevazione che permettano flessibilità in sede di costruzione di indicatori della disoccupazione e/o che consentano di mettere a fuoco 'aree grigie' (cioè a dire, insiemi di persone di difficile collocazione in uno degli usuali tre stati della partecipazione al lavoro - occupazione, disoccupazione, inattività -); (ii) la controversia circa la rilevanza della distinzione fra disoccupazione e inattività, segnatamente per quanto attiene al comportamento dinamico.

Quanto al primo tema, se per un verso è indubbio che il tratto comune degli attuali dispositivi di misurazione dell'occupazione e della disoccupazione è l'accentuazione del ruolo svolto da convenzioni definitorie assai circostanziate, per un altro verso altrettanto palese è la crescente importanza attribuita a strategie di rilevazione che consentano di costruire gli aggregati in modo flessibile e di cogliere situazioni in prossimità delle frontiere fra

occupazione, disoccupazione e inattività. La proposta di un insieme di indicatori alternativi della disoccupazione, appropriati per diversi scopi, risale a Shiskin (1976). La stessa logica motiva la strategia dei *building blocks*, cioè a dire di costruzione di aggregati di specifico interesse aggiungendo o togliendo da un aggregato di riferimento - ad esempio, le persone in cerca di occupazione - particolari sottoinsiemi di persone in relazione a varianti definitorie, suggerita dall'ILO (1980). Palesemente, alla base di queste proposte è l'esigenza di nuovi concetti e di nuove classificazioni, capaci di cogliere con maggiore fedeltà e in modo più articolato i gradi della partecipazione al lavoro (per la considerazione di diverse situazioni ai confini fra gli usuali tre stati della partecipazione al lavoro, rilevanti sia a fini interpretativi che di politiche, vedi Malinvaud, 1986, pp. 32-59).

Il secondo tema, la rilevanza della distinzione fra disoccupati e inattivi, è stato affrontato soprattutto negli Stati Uniti (Clark e Summers, 1979; Flinn e Heckman, 1983; Poterba e Summers, 1984), ma ha chiaramente portata generale. In sostanza, i termini del dibattito possono essere così enunciati: per quanti mostrano, o hanno mostrato di recente, un *attachment* al mercato del lavoro (tipicamente, per coloro che sperimentano o nel recente passato hanno sperimentato un episodio di disoccupazione), la distinzione fra disoccupazione e inattività è debole, vuoi perché inficiata da errori di risposta vuoi perché le probabilità di transitare da questi due stati verso l'occupazione sono simili, oppure si tratta di due stati ben definiti e distinti, con differente propensione verso l'occupazione? La questione, si noti, non chiama direttamente in causa la plausibilità delle definizioni degli stati della partecipazione al lavoro e dei connessi criteri di classificazione. Essa ne mette tuttavia in discussione, almeno in parte, la valenza interpretativa.

Obiettivo di questo studio è di fornire qualche ulteriore lume alla comprensione delle caratteristiche della disoccupazione e dell'*attachment* al mercato del lavoro nel nostro Paese, tramite un insieme di analisi essenzialmente esplorative sui microdati della rilevazione trimestrale sulle forze di lavoro (nel seguito RTFL). La questione è affrontata da due punti di vista.

(a) In primo luogo, si muove dalle risposte fornite, in una singola occasione di indagine, al blocco di domande sulla ricerca di lavoro e si conduce un'analisi di classificazione esplorativa (Lebart, Morineau e Warwick, 1984), con lo scopo di individuare insiemi di persone omogenei quanto ad *attachment* al mercato del lavoro. Questo approccio, certo non usuale, non è alternativo alle procedure di classificazione *a priori* e non ha l'ambizione di sostituirle. Esso si prospetta tuttavia come una chiave di lettura integrativa, piuttosto potente e di particolare interesse proprio per mettere a fuoco le caratteristiche della partecipazione al lavoro di alcune delle 'aree grigie' or ora menzionate.

(b) In secondo luogo, si svolgono alcune prime analisi sui flussi di mobilità trimestrale tra i gruppi individuati con l'analisi di classificazione. Palesemente, lo scopo è di vagliare, sia pure in forma forzatamente semplificata (anche per i vincoli nelle informazioni disponibili), se i diversi gruppi sono caratterizzati da un differente comportamento dinamico: in altre parole, di accertare se la classificazione risultante dall'analisi esplorativa è di

rilievo per cogliere, e in definitiva predire, cambiamenti nel tempo della condizione rispetto al lavoro.

I precedenti di questo studio sono in Rettore, Torelli e Trivellato (1988). In tale lavoro abbiamo presentato, per la sola Lombardia, l'analisi statica di classificazione esplorativa, accompagnata da un vaglio dell'importanza di varianti definitorie, comunque riconducibili alle raccomandazioni dell'ILO, sulla misura dell'ammontare e del tasso di disoccupazione. Non ci soffermiamo qui su questo secondo aspetto, se non per ricordare come la scelta del confine definitorio tra disoccupati e inattivi, sia pure circoscritta a varianti nelle date e nelle caratteristiche delle azioni di ricerca compiute - ammissibili per essere considerati in cerca di lavoro -, abbia implicazioni tutt'altro che trascurabili sulla consistenza degli aggregati (vedi anche il cap. 1). È questa, infatti, un'evidenza che val la pena di tener presente, perchè motiva l'interesse per un più approfondito studio delle caratteristiche dell'*attachment* al mercato del lavoro.

In questa sede incentriamo l'attenzione sull'analisi di classificazione esplorativa e la estendiamo in una duplice direzione. Per un verso consideriamo, accanto alla Lombardia, una regione con caratteristiche socio-economiche marcatamente diverse - la Campania -, per avere riscontri sulla stabilità delle classificazioni finali rispetto a situazioni parecchio difformi del mercato del lavoro. Per un altro verso, come già detto, innestiamo sui risultati dell'analisi di classificazione un primo esame dei flussi di mobilità trimestrale, che consente di vagliare la rilevanza delle classificazioni per la dinamica di breve periodo nel mercato del lavoro. Per entrambe le regioni, l'analisi è condotta sui dati delle RTFL del 1986.I e del 1986.II relativi al sub-campione di persone che hanno dichiarato di essere alla ricerca di lavoro.

L'organizzazione del capitolo è la seguente. Nella sez. 2 presentiamo sinteticamente le tecniche di analisi esplorativa usate e quindi i risultati ottenuti, segnatamente per quanto attiene all'applicazione al campione lombardo per il secondo trimestre del 1986. Nella sez. 3 presentiamo e discutiamo primi risultati dell'analisi dei flussi di mobilità trimestrale tra i gruppi individuati nella sezione precedente, distintamente per Lombardia e Campania. Nella sez. 4, infine, prospettiamo alcune considerazioni conclusive.

2. *L'analisi di classificazione esplorativa delle risposte ai quesiti del questionario*

2.1. *Le tecniche esplorative adottate*

La via seguita in questo studio per la costruzione di gruppi omogenei di persone rispetto all'*attachment* al mercato del lavoro consiste nell'analizzare alcune variabili del questionario della RTFL, talune attinenti alla condizione dichiarata nella settimana di riferimento e altre riguardanti l'attività di ricerca di lavoro compiuta, tramite più tecniche di analisi dei dati opportunamente concatenate, con l'obiettivo di pervenire ad una classificazione finale suffi-

cientemente robusta.

Conviene ricordare che le variabili prese in esame sono qualitative a più modalità. Si rende pertanto opportuno il ricorso preliminare a tecniche di analisi fattoriale, segnatamente all'analisi delle corrispondenze multipla, al fine di rendere più agevole il trattamento delle informazioni nelle successive fasi di analisi classificatoria.

I gruppi, o classi, che verranno identificati risultano quindi dall'uso in sequenza di un analisi delle corrispondenze multipla, di una serie di *cluster analyses* non gerarchiche in via esplorativa e infine da una *cluster analysis* gerarchica[1] . Nel seguito, la procedura adottata viene tratteggiata sinteticamente, con riguardo sia alle tecniche utilizzate che alla strategia per un loro conveniente impiego combinato. Rinviamo a Lebart, Morineau e Tabard (1977) e Chandon e Pinson (1981) rispettivamente per l'analisi delle corrispondenze multipla e per le tecniche di *cluster analysis*, ed a Lebart, Morineau e Warwick (1984) e Griguolo e Palermo (1984) per una puntuale discussione sulle opportunità di impiego di tali tecniche in sequenza per la costituzione di gruppi stabili.

La procedura adottata si muove secondo le seguenti linee:

(a) analisi delle corrispondenze multipla delle variabili selezionate;
(b) generazione di un numero elevato di partizioni esplorative, tramite un algoritmo di *clustering* non gerarchico applicato ai fattori ottenuti al passo precedente;
(c) incrocio delle partizioni esplorative migliori (in termini di inerzia spiegata) al fine di ottenere un insieme di classi stabili, cioè di classi composte dagli individui che risultano sempre associati nelle diverse partizioni esplorative;
(d) classificazione gerarchica delle classi stabili, opportunamente ponderate, ottenute al passo precedente.

L'uso di tali tecniche, nella sequenza indicata, agisce nel senso di ridurre al minimo l'arbitrarietà connessa alle scelte necessarie in ciascuna delle fasi: ad esempio, del numero di fattori risultante dall'analisi delle corrispondenze, o del numero di partizioni esplorative, o ancora del numero di classi di queste ultime.

2.2. *Le variabili considerate e i risultati salienti dell'analisi esplorativa*

Le variabili impiegate nell'analisi esplorativa (o variabili attive) sono essenzialmente le stesse prese in considerazione dall'Istat, e più in generale dai criteri di classificazione *a priori* riconducibili alle raccomandazioni dell'ILO, per definire gli aggregati di occupati, disoccupati e inattivi. Si tratta della condizione dichiarata, dell'effettuazione o meno di ore di lavoro nella settimana di riferimento e di cinque variabili sulla disponibilità a lavorare e sulla ricerca di lavoro. Le modalità di tali variabili derivano da una opportuna

---

1 Le elaborazioni sono state eseguite con il *package* ADDAEST descritto in Griguolo e Vettoretto (1983).

ricodifica delle modalità originarie, e sono descritte nella Tab. 1. I criteri adottati per la ricodifica vanno ricondotti alla necessità di escludere le modalità scarsamente presenti nel campione, e ad un tempo di disporre di modalità il più possibile coerenti con quelle utilizzate nelle classificazioni *a priori*.

Tab. 1: *Variabili attive impiegate nell'analisi e loro codifica*

Variabili attive	Modalità	Sigla
Condizione unica o prevalente (autodichiarazione)	Occupato	cor1
	In cerca di nuova occupazione	cor2
	In cerca di prima occupazione	cor3
	In condizione non professionale	cor4
Ore di lavoro nella settimana di riferimento	Effettuate	ore1
	Non effettuate	ore2
Ricerca di occupazione	Cerca lavoro alle dipendenze o inizierà tra breve	cer1
	Occupato ma teme di perdere l'occupazione o questa è temporanea	cer2
	Occupato ma alla ricerca di una diversa occupazione o di un secondo lavoro	cer3
	Intende iniziare un lavoro in proprio	cer4
Disponibilità a lavorare	Sì	dis1
	No	dis2
Azioni di ricerca compiute	Iscrizione ad un ufficio di collocamento o partecipazione a concorso pubblico	taz1
	Ricerca presso privati	taz2
	Entrambe le precedenti	taz3
	Nessuna azione	taz4
Numero di azioni di ricerca	Nessuna	naz1
	Una	naz2
	Due	naz3
	Tre	naz4
	Più di tre	naz5
Epoca dell'ultima azione di ricerca	Ultimi 30 giorni	uaz1
	Da uno a sei mesi fa	uaz2
	Oltre sei mesi fa	uaz3
	Ricerca non iniziata	uaz4

Come già anticipato, l'analisi è condotta sulle persone che hanno dichiarato di essere alla ricerca di lavoro, indipendentemente dalla condizione dichiarata. Tale scelta merita due brevi commenti. È da notare innanzitutto che il criterio di identificazione del sub-campione oggetto dell'analisi, prescindendo dalla condizione dichiarata, consente di esplorare l'intensità dell'*attachment* al mercato del lavoro anche di occupati che si dichiarano in cerca, e quindi di avere verosimilmente qualche lume sull''area grigia' fra sottoccupazione e disoccupazione. D'altra parte, per il fatto che l'analisi è circoscritta a quanti si sono detti alla ricerca di lavoro, resta invece esclusa la possibilità di investigare l''area grigia' tra disoccupazione e inattività rappresentata dai cosiddetti 'lavoratori scoraggiati' (che è comunque problematico individuare, anche sulla scorta di criteri di classificazione *a priori*, sulla base della struttura del questionario della RTFL[2]).

Nel campione lombardo del 1986.II, le persone che hanno dichiarato di cercare lavoro sono 3.347. L'analisi delle corrispondenze ha portato alla selezione di un numero di fattori, pari a 13, tali da spiegare il 95% dell'inerzia totale.

A partire da questi fattori si è scelto di condurre, in via esplorativa, 20 *cluster analyses* non gerarchiche, fissando a 3 il numero di classi per ognuna di esse.

Le 6 partizioni per cui la quota di inerzia tra le classi è risultata più elevata sono state incrociate ottenendo così 26 classi, ognuna composta dagli individui che in tutte le partizioni esplorative risultavano nel medesimo gruppo.

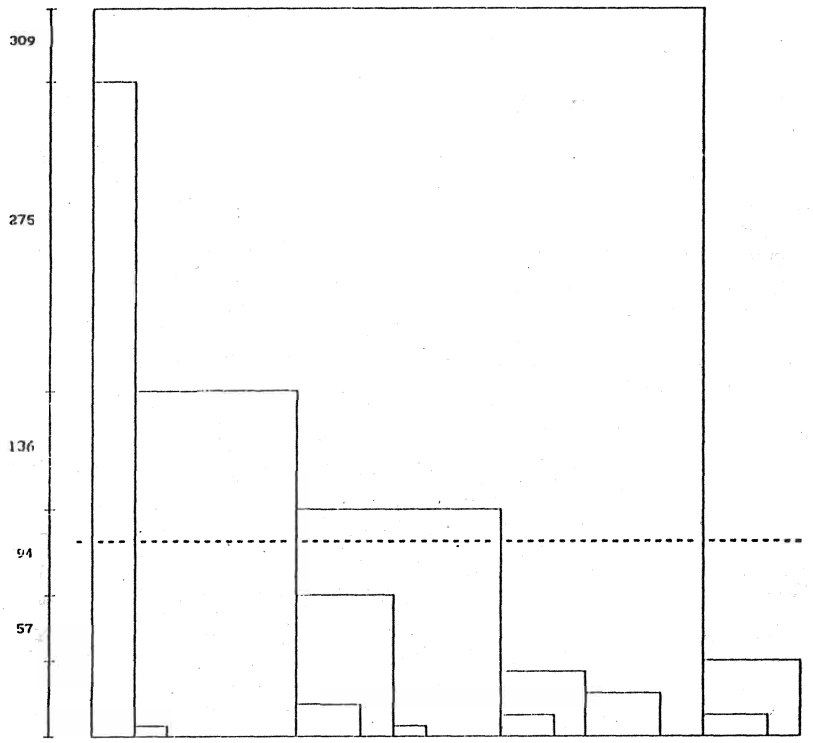
Le 26 classi così ottenute sono state oggetto di una *cluster analysis* gerarchica. Ogni classe è rappresentata dal vettore di medie che i 13 fattori assumono in essa, ed è pesata con la quota di persone che la compongono. La Fig. 1 mostra il dendogramma risultante. Appare evidente come la scelta più ragionevole sia quella di considerare 4 classi. L'incremento di inerzia spiegata che si ha aggiungendo un'ulteriore classe è, infatti, contenuto. Delle quattro classi che si ottengono 'tagliando' il dendogramma dove l'incremento di inerzia spiegata è ancora elevato, la prima è composta da occupati che svolgono azioni di ricerca di lavoro; la seconda è composta in gran parte da persone dichiaratesi disoccupate (in senso stretto) oppure in cerca di prima occupazione e caratterizzate da una vivace attività di ricerca; la terza si differenzia dalla precedente per l'inclusione di persone con attività di ricerca meno intensa; la quarta è formata sia da occupati che da non occupati, accomunati dal fatto che, pur avendo dichiarato di cercare lavoro, in realtà non hanno svolto attività di ricerca.

Al fine di avere evidenze sulla robustezza dei risultati ottenuti, la stessa analisi è stata condotta su insiemi di dati riferiti ad ambiti territoriali diversi: sempre per il 1986.II, l'analogo campione delle persone in cerca di occupazione del Veneto e della Campania. In tutti questi casi, percorrendo i passi già delineati, si è pervenuti sempre a quattro classi con caratteristiche

---

2 Per un'accurata disamina dei problemi di definizione, rilevazione e misura dei lavoratori scoraggiati, vedi OECD (1987, pp.142-170).

Fig. 1: *Dendogramma dell'aggregazione gerarchica di 26 classi stabili ottenute nella fase esplorativa: Lombardia, 1986.I-II* [(a)]



(a) Nella scala, sono riportate le variazioni nell'inerzia spiegata dovute alle variazioni nei livelli di aggregazione delle classi stabili. La linea tratteggiata mostra dove è stato 'tagliato' l'albero, così da produrre la partizione in quattro classi.

analoghe a quelle appena descritte (vedi la Tab. 2). E' questa una conferma, sì piuttosto circoscritta ma di ragguardevole interesse, della robustezza della classificazione cui si è pervenuti.

Pur con ovvie cautele, è ragionevole concludere che si è colta una struttura dei dati piuttosto 'forte': vuoi nel senso che, per la strategia di analisi esplorativa impiegata, i risultati sono scarsamente influenzati da opzioni in tema di utilizzazione di singole tecniche; vuoi perchè essa non sembra peculiare del campione lombardo in questione, e pare anzi presentare una qualche stabilità e generalità.

Tab. 2: *Risultati salienti dell'analisi esplorativa condotta su Lombardia, Veneto e Campania (1986.II)*

	Lombardia	Veneto	Campania
Numerosità campionaria	3.347	766	1.158
% di inerzia spiegata con 4 classi	37,6	36,6	36,1
Gruppi risultanti dall'analisi esplorativa[(a)]			
OCC: occupati in cerca di altra occupazione	25,3	27,2	19,4
INC1: non occupati in cerca di occupazione	35,7	33,8	30,4
INC2: non occupati con debole ricerca di occupazione	32,7	34,9	44,4
NONC: persone che non svolgono attività di ricerca	6,3	4,2	5,8

(a) L'interpretazione dei quattro gruppi è comune ai tre insiemi di dati.

La Tab. 3 presenta le distribuzioni delle variabili attive per ognuna delle quattro classi ottenute; l'ultima riga della tabella presenta la distribuzione marginale di tali variabili per il complesso dei dati. Passando in rassegna la distribuzione delle variabili attive è agevole cogliere la diversa caratterizzazione dei quattro gruppi.

La prima classe (nel seguito denotata OCCR) è formata quasi esclusivamente da persone che si sono dichiarate occupate. Mediamente mostrano una disponibilità ad accettare un eventuale (nuovo) lavoro elevata, ma inferiore a quella degli altri gruppi. Hanno inoltre compiuto almeno una azione di ricerca piuttosto di recente. In definitiva, si tratta di occupati alla ricerca di una nuova occupazione, ed abbastanza motivati in tal senso.

La seconda classe (nel seguito denotata INC1) è composta prevalentemente da persone dichiaratesi in cerca di prima o di nuova occupazione (90%). Le persone che appartengono a questa classe sono le più attive nella

Tab. 3: *Distribuzioni di frequenza delle sette variabili attive nelle quattro classi e distribuzione di frequenza marginale*

	OCCR	INC1	INC2	NONC	Totale
Condizione unica o prevalente					
cor1	0,98	0,04	0,02	0,50	0,30
cor2	0,01	0,27	0,22	0,11	0,18
cor3	-	0,63	0,43	0,18	0,38
cor4	0,01	0,06	0,33	0,21	0,15
Ore di lavoro nella settimana di riferimento					
ore1	0,92	0,06	0,02	0,45	0,29
ore2	0,08	0,94	0,98	0,56	0,71
Ricerca di occupazione					
cer1	0,22	0,88	0,88	0,52	0,69
cer2	0,28	-	-	0,10	0,08
cer3	0,50	-	-	0,31	0,15
cer4	-	0,12	0,12	0,08	0,09
Disponibilità a lavorare					
dis1	0,76	1,0	1,0	0,83	0,93
dis2	0,24	-	-	0,18	0,07
Azioni di ricerca compiute					
taz1	0,19	-	0,34	-	0,16
taz2	0,66	0,06	0,61	-	0,39
taz3	0,16	0,94	0,05	-	0,39
taz4	-	-	-	1,0	0,06
Numero di azioni di ricerca					
naz1	-	-	-	1,0	0,06
naz2	0,54	-	0,66	-	0,35
naz3	0,29	0,23	0,32	-	0,26
naz4	0,12	0,43	0,03	-	0,19
naz5	0,06	0,34	-	-	0,14
Epoca dell'ultima azione di ricerca					
uaz1	0,44	0,68	0,53	0,07	0,53
uaz2	0,41	0,27	0,30	0,05	0,30
uaz3	0,13	0,05	0,15	0,02	0,10
uaz4	0,02	-	0,03	0,86	0,07

ricerca di lavoro: hanno svolto 2 o più azioni di ricerca e, nel 70% dei casi, almeno una negli ultimi trenta giorni. E' questa la classe che può essere identificata con il 'nucleo forte' della disoccupazione.

La terza classe (nel seguito denotata INC2) si distingue dalla precedente per la presenza di una quota consistente di persone che si dichiarano in condizione non professionale. Si caratterizza inoltre per un'attività di ricerca

meno intensa, con riguardo sia al numero di azioni (prevalentemente una sola azione) che alla loro collocazione temporale (oltre il 45% svolte più di un mese prima dell'intervista). In sintesi, è un gruppo di non occupati con minor *attachment* al mercato del lavoro, in termini di percezione soggettiva della propria condizione così come di intensità della ricerca di lavoro. In tal senso, rispetto alle definizioni *a priori* suggerite dall'ILO, si colloca in una sorta di 'area grigia' fra disoccupazione e inattività, potendo confluire in uno o nell'altro dei due aggregati in relazione al modo con cui le definizioni vengono specificate e rese operative.

La quarta classe (nel seguito denotata NONC) comprende per il 70% persone in condizione non professionale o dichiaratesi occupate, e per il restante 30% persone che hanno invece dichiarato di essere in cerca di prima o nuova occupazione. A fronte di questa scarsa caratterizzazione rispetto alla condizione, sta il tratto distintivo costituito dal fatto che le persone del gruppo non svolgono alcuna apprezzabile attività di ricerca (solo il 15% dichiara di aver svolto azioni di ricerca oltre un mese prima dell'intervista). Si tratta in sostanza di persone che sono occupate e non cercano altro lavoro o che sono prive di lavoro e in tal caso possono ragionevolmente essere considerate inattive.

La Tab. 4 riporta i profili delle quattro classi rispetto ad alcune variabili supplementari, talune demografiche ed altre attinenti all'attività di ricerca di lavoro (durata della ricerca in corso, caratteristiche del lavoro cercato, condizione all'inizio della ricerca). Tali profili consentono di qualificare ulteriormente le diverse classi. In particolare, possiamo notare quanto segue:

(a) in OCCR vi è una prevalenza di maschi, mentre nelle rimanenti classi è più consistente la presenza femminile;
(b) la distribuzione per età mostra una massiccia presenza di giovani (in età compresa tra 14 e 29 anni) nella classe INC1. All'opposto in NONC è più consistente la presenza di persone in età adulta;
(c) le persone incluse in INC1 sono alla ricerca di lavoro da un tempo maggiore di quanto avvenga per le persone delle restanti classi. In NONC si ha invece una quota elevata di persone che solo da poco tempo hanno iniziato a cercare lavoro[3];
(d) le persone incluse in OCCR in gran parte esprimono forti preferenze per lavori alle dipendenze e a tempo pieno, mentre nei gruppi INC2 e NONC sono sovrarappresentati coloro che cercano lavori *part-time* o in proprio. La classe INC1 si caratterizza invece per comprendere persone che non esprimono forti opzioni quanto a caratteristiche del lavoro;
(e) per quanto attiene alla condizione all'inizio della ricerca, si ha un'apprezzabile presenza di occupati in OCCR e NONC, del resto affatto conforme

3 Quest'ultima evidenza non è necessariamente contraddittoria. Essa è infatti legata: in parte alle forzature imposte dal programma di correzione automatica impiegato dall'Istat, che imputa un periodo di ricerca non nullo a chi ha dichiarato di essere alla ricerca anche se questa non risulta ancora iniziata dalle risposte ai quesiti successivi; in parte alla reale dichiarazione dei rispondenti che coerentemente dicono di avere appena iniziato la ricerca (senza magari aver svolto alcuna attività) e quindi di essere alla ricerca solo da poche settimane.

Tab. 4: *Distribuzioni di frequenza di alcune variabili supplementari nelle quattro classi e distribuzione di frequenza marginale*

	OCCR	INC1	INC2	NONC	Totale
Sesso					
maschi	0,54	0,41	0,35	0,45	0,43
femmine	0,46	0,59	0,65	0,55	0,57
Età					
fino a 29 anni	0,63	0,81	0,63	0,54	0,69
da 30 a 39 anni	0,20	0,09	0,19	0,24	0,16
da 40 a 49 anni	0,13	0,06	0,12	0,15	0,10
50 anni e oltre	0,04	0,03	0,06	0,07	0,05
Relazione con il capofamiglia					
capofamiglia	0,29	0,10	0,14	0,23	0,17
altro	0,71	0,90	0,86	0,77	0,83
Durata della ricerca in corso					
da 0 a 6 mesi	0,30	0,19	0,22	0,49	0,25
da 7 a 12 mesi	0,29	0,27	0,30	0,26	0,28
da 13 a 24 mesi	0,24	0,25	0,24	0,16	0,24
più di 24 mesi	0,17	0,30	0,24	0,09	0,23
Caratteristiche del lavoro cercato					
indipendente	0,06	0,12	0,12	0,16	0,11
escl. dip. *full-t.*	0,49	0,30	0,32	0,31	0,35
escl. dip. *part-t.*	0,02	0,02	0,07	0,08	0,04
pref. dip. *full-t.*	0,23	0,32	0,20	0,20	0,25
pref. dip. *part-t.*	0,03	0,03	0,08	0,10	0,05
senza preferenze	0,17	0,22	0,20	0,14	0,20
Condizione all'inizio della ricerca					
occupato	0,95	0,14	0,11	0,51	0,36
studente	0,02	0,28	0,22	0,12	0,19
militare	-	0,05	0,03	0,01	0,03
casalinga	0,01	0,19	0,34	0,19	0,19
altro	0,02	0,33	0,30	0,17	0,23

alle attese, una prevalenza di studenti in INC1 e INC2 e la concentrazione di casalinghe in INC2.

La composizione delle classi con riguardo alle variabili supplementari è in definitiva coerente con quanto ci si poteva ragionevolmente attendere a seguito dell'interpretazione delle stesse risultante dall'analisi delle variabili attive, e vale anzi a caratterizzarle in maniera ancor più persuasiva.

Scontando qualche semplificazione forse drastica, la fisionomia dei quattro gruppi è così riassumibile in pochi, significativi tratti. OCCR comprende essenzialmente occupati maschi (come vedremo nel seguito, per una frazione non trascurabile sottoccupati) alla ricerca di un diverso lavoro, e selettivi

nelle caratteristiche dell'occupazione cercata - a tempo pieno e alle dipendenze -. INC1 è il 'nucleo forte' della disoccupazione: persone in cerca di nuova e soprattutto di prima occupazione, giovani, con intensa attività di ricerca, non selettive rispetto al lavoro desiderato. Meno marcate sono le caratteristiche di *attachment* al mercato del lavoro di INC2: una parte non esigua delle persone di questa classe si dichiara in condizione non professionale, e consta in particolare di casalinghe; l'attività di ricerca è presente, ma meno intensa e continuativa, e orientata selettivamente su una occupazione *part-time* o in proprio. Infine, NONC comprende persone in buona parte in età adulta, occupate o all'opposto inattive (quanto a condizione dichiarata), che sostanzialmente non cercano lavoro e che sono comunque selettive quanto alle caratteristiche dell'occupazione desiderata.

3. *I flussi tra gli stati: prime evidenze*

3.1. *La matrice dei flussi*

I risultati fin qui ottenuti, pur caratterizzando in modo convincente i quattro gruppi risultanti dalle analisi esplorative e facendo intuire l'esistenza di diversi atteggiamenti rispetto al lavoro, non mostrano ancora se e quanto alle diverse caratteristiche dei gruppi corrispondono diversi esiti della ricerca di lavoro. Per cercare di dare una prima risposta a tale quesito, si sono considerate le matrici di transizione relative ai flussi avvenuti tra il primo e il secondo trimestre del 1986, separatamente per Lombardia e Campania.

Preliminarmente, va spiegato come sono stati definiti gli stati tra i quali hanno luogo i flussi studiati. L'analisi di classificazione esplorativa descritta nella sez. 2 è stata applicata, separatamente in ognuno dei due trimestri e in ognuna delle due regioni, all'insieme composto da coloro che nel trimestre di riferimento hanno dichiarato di cercare lavoro e che nell'altro trimestre hanno fatto parte del campione degli intervistati[4]. Dato che il disegno di campionamento della RTFL è tale da garantire che il 50% degli intervistati in un'occasione sono intervistati anche all'occasione successiva (a meno di fenomeni di *attrition*), gli insiemi qui considerati sono di numerosità circa dimezzata rispetto agli insiemi di persone che hanno dichiarato di cercare lavoro. La Tab. 5 visualizza la composizione degli insiemi analizzati. Le analisi relative al primo trimestre sono state condotte sull'insieme degli appartenenti alle caselle a) e b); quelle relative al secondo, sull'insieme degli appartenenti alle caselle a) e c). Gli appartenenti alla casella d), vale a dire gli occupati e gli inattivi che hanno dichiarato di non cercare lavoro in entrambi i trimestri, pur non essendo sottoposti ad alcuna analisi esplorativa, sono inclusi nell'analisi dei flussi per l'ovvia esigenza di chiusura contabile della matrice.

---

4 Per l'abbinamento dei dati individuali nelle due occasioni di indagine, abbiamo utilizzato la procedura LINK descritta nel cap. 7.

Tab. 5: *Composizione degli insiemi sui quali sono state svolte le analisi esplorative relative agli abbinati 1986.I-II*

	1986.II	
	persone che hanno dichiarato di cercare lavoro	persone che hanno dichiarato di non cercare lavoro
persone che hanno dichiarato di cercare lavoro	a) inclusi nelle analisi esplorative relative a 1986.I e 1986.II	b) inclusi nella analisi esplorativa relativa a 1986.I
persone che hanno dichiarato di non cercare lavoro	c) inclusi nella analisi esplorativa relativa a 1986.II	d) non inclusi in alcuna analisi esplorativa

Le analisi esplorative condotte su questi gruppi hanno fornito risultati del tutto simili a quelli illustrati nella sez. 2. Per agevolare l'interpretazione dei risultati dell'analisi dei flussi, il quarto gruppo, vale a dire il gruppo composto da coloro che pur dichiarando di cercare un lavoro non hanno ancora svolto azioni di ricerca, è stato peraltro diviso in due, separando coloro che hanno un lavoro da coloro che non lavorano (nel seguito ONONC e INONC, rispettivamente). Gli stati tra i quali hanno luogo i flussi sono quindi sette: i cinque risultanti dalle analisi esplorative (OCCR, INC1, INC2, ONONC, INONC), più gli occupati e gli inattivi che in entrambe le occasioni hanno dichiarato di non cercare lavoro (nel seguito OCC e INA, rispettivamente).

Le Tabb. 6 e 7 riportano le matrici dei flussi relative alle due regioni. Per meglio studiarne le caratteristiche, nel seguito si fa uso di alcuni semplici indicatori tratti dalla letteratura sulle catene di Markov omogenee (vedi, ad es., Kemeny e Snell, 1960). Tali indicatori devono essere letti con molta prudenza, giacché nulla autorizza a pensare che il processo che governa le transizioni studiate sia markoviano omogeneo. Con queste avvertenze, l'analisi degli indicatori fornisce elementi di chiarimento sulla dinamica del fenomeno, in quanto ne coglie sinteticamente le implicazioni (sia pure sotto assunzioni di comodo, il cui realismo è dubbio).

Le Tabb. 8 e 9 presentano le probabilità che, partendo dagli stati cui sono intestate le righe, la prima transizione verso uno stato diverso da quello di origine avvenga verso gli stati cui sono intestate le colonne. Si nota innanzitutto che scendendo dalla riga intestata a ONONC alla riga intestata a INONC, la distribuzione di probabilità si sposta gradualmente verso destra. In altre parole, l'ordinamento degli stati proposto nelle tabelle appare essere un ordinamento secondo il grado di *attachment* al mercato del lavoro: passando da ONONC a INONC aumenta progressivamente la probabilità che, al momento di abbandonare lo stato di origine, si passi ad uno stato con basso *attachment* al mercato del lavoro.

Tuttavia, v'è da notare che tra le due regioni vi sono differenze non

Tab. 6: *Matrice dei flussi nel mercato del lavoro, con riferimento agli stati risultanti dall'analisi esplorativa sull''attachment': Lombardia, 1986.I-II* [(a)]

	OCC	ONONC	OCCR	INC1	INC2	INONC	INA
OCC	10.452	19	95	33	22	4	245
	0,96	..	*0,01*	..	..	..	*0,02*
ONONC	24	11	6	1	-	-	2
	0,55	*0,25*	*0,14*	*0,02*	-	-	*0,05*
OCCR	116	6	174	30	12	-	8
	0,34	*0,02*	*0,50*	*0,09*	*0,03*	-	*0,02*
INC1	68	-	37	345	46	1	30
	0,13	-	*0,07*	*0,65*	*0,09*	..	*0,06*
INC2	61	2	15	103	201	12	72
	0,13	..	*0,03*	*0,22*	*0,43*	*0,03*	*0,15*
INONC	1	-	-	6	11	9	11
	0,03	-	-	*0,16*	*0,29*	*0,24*	*0,29*
INA	231	1	4	51	63	13	10.555
	0,02	..	..	*0,01*	*0,01*	..	*0,97*

(a) Percentuali (per riga) in corsivo; ..: percentuale non significativa alla seconda cifra decimale.

Tab. 7: *Matrice dei flussi nel mercato del lavoro, con riferimento agli stati risultanti dall'analisi esplorativa sull''attachment': Campania, 1986.I-II* [(a)]

	OCC	ONONC	OCCR	INC1	INC2	INONC	INA
OCC	2.068	1	23	1	9	2	57
	0,96	..	*0,01*	..	..	..	*0,03*
ONONC	4	3	2	-	-	-	-
	0,44	*0,33*	*0,22*	-	-	-	-
OCCR	27	-	48	3	3	-	4
	0,32	-	*0,56*	*0,04*	*0,04*	-	*0,05*
INC1	14	-	8	87	28	-	23
	0,09	-	*0,05*	*0,54*	*0,17*	-	*0,14*
INC2	17	-	8	41	132	5	26
	0,07	-	*0,03*	*0,18*	*0,58*	*0,02*	*0,11*
INONC	1	-	-	1	4	10	3
	0,05	-	-	*0,05*	*0,21*	*0,53*	*0,16*
INA	71	-	4	13	37	7	2.398
	0,03	-	..	*0,01*	*0,01*	..	*0,95*

(a) Percentuali (per riga) in corsivo; ..: percentuale non significativa alla seconda cifra decimale.

trascurabili, prima fra tutte la scarsa caratterizzazione di INC1 rispetto a INC2 in Campania, a fronte delle marcate differenze che si riscontrano tra gli stessi due stati in Lombardia. In Lombardia, infatti, la probabilità che al momento di abbandonare INC1 la transizione sia verso OCC o OCCR è circa doppia dell'analoga probabilità quando si muova da INC2; in Campania, invece, le due probabilità sono praticamente eguali. Inoltre, è più facile passare direttamente da INA a OCC in Lombardia che non in Campania.

La Tab.10 presenta i tempi medi di permanenza in ognuno dei sette stati. Nel calcolo di tali valori risulta cruciale l'assunzione di markovianità (del primo ordine). Tale assunzione è palesemente irrealistica, dato che esclude, tra le altre cose, la possibilità che la probabilità di abbandonare lo stato dipenda dalla lunghezza del periodo trascorso nello stato. I valori ricavati forniscono quindi un'approssimazione alla realtà di larga massima, che serve essenzialmente per valutazioni ipotetiche ed a fini comparativi.

Emerge, riproposta da un altro punto di vista, la differenza esistente tra INC1 e INC2 in Lombardia, e come invece i due stati siano simili in Campania: in Lombardia le persone più motivate nella ricerca di lavoro (INC1) permangono nello stato mediamente più del doppio di quanto non accada per le persone meno motivate (INC2). Questa differenza di *attachment* allo stato, come si è notato commentando la Tab. 8, si riflette sulla probabilità di successo della ricerca: i più motivati hanno una probabilità molto maggiore che la ricerca vada a buon fine, si concluda cioè con un'occupazione; i meno motivati cambiano stato mediamente in meno tempo, ma con una eterogeneità di destinazioni molto più accentuata.

Le Tabb. 11 e 12 presentano i tempi medi necessari per raggiungere uno stato da ogni altro. Al pari dei tempi medi di permanenza nello stato, il loro calcolo dipende essenzialmente dall'assunzione di markovianità. L'analisi

Tab. 8: *Matrice delle probabilità di transizione condizionate all'abbandono dello stato di partenza: Lombardia (ricavata dalla matrice dei flussi 1986.I-II)*

	OCC	ONONC	OCCR	INC1	INC2	INONC	INA
OCC		0,05	0,23	0,08	0,05	-	0,59
ONONC	0,73		0,18	0,03	-	-	0,06
OCCR	0,68	0,04		0,18	0,06	-	0,04
INC1	0,37	-	0,20		0,26	-	0,17
INC2	0,23	0,01	0,06	0,39		0,05	0,27
INONC	0,04	-	-	0,21	0,38		0,38
INA	0,64	-	-	0,15	0,18	0,03	

Tab. 9: *Matrice delle probabilità di transizione condizionate all'abbandono dello stato di partenza: Campania (ricavata dalla matrice dei flussi 1986.I-II)*

	OCC	ONONC	OCCR	INC1	INC2	INONC	INA
OCC		0,01	0,25	0,01	0,10	0,02	0,61
ONONC	0,66		0,33	-	-	-	-
OCCR	0,73	-		0,08	0,08	-	0,11
INC1	0,20	-	0,11		0,40	-	0,30
INC2	0,18	-	0,08	0,42		0,05	0,27
INONC	0,11	-	-	0,11	0,44		0,33
INA	0,54	-	0,03	0,10	0,28	0,05	

Tab. 10: *Durata media, in trimestri, della permanenza negli stati in Lombardia e Campania (ricavata dalle matrici dei flussi 1986.I-II)*

	Lombardia	Campania
OCC	25,0	22,2
ONONC	0,3	0,5
OCCR	1,0	1,3
INC1	1,9	1,2
INC2	0,8	1,4
INONC	0,3	1,1
INA	29,1	18,2

delle due matrici evidenzia innanzitutto che i due stati INONC e ONONC possono essere considerati transitori. La loro esclusione dallo studio del processo non comporta quindi perdite di precisione apprezzabili. L'analisi delle due matrici fornisce poi un'ulteriore conferma che la sequenza dei sette stati proposta nelle tabelle, da OCC a INA, corrisponde ad un ordinamento naturale degli stessi, ordinamento che abbiamo visto essere interpretabile come grado di *attachment* al mercato del lavoro. Salvo tre eccezioni nella matrice relativa alla Campania, leggendo l'i-esima colonna dall'alto verso il basso si incontrano tempi medi decrescenti fino all'elemento i-esimo (vale a dire, fino al tempo medio per il primo ritorno allo stato i- esimo) e poi crescenti: in altre parole, più lo stato di origine è lontano dallo stato i-esimo

nell'ordinamento proposto, più alto è il tempo medio necessario per arrivare nello stato i-esimo. Come si è appena anticipato, nel caso della Campania si osservano tre violazioni a questa regola. Tali violazioni sono però facilmente spiegabili con la minore caratterizzazione dei gruppi in questa regione, già emersa da precedenti evidenze. Segnatamente, nel caso in esame NONC si distingue poco da OCC, come pure INC1 da INC2.

Tab. 11: *Matrice dei tempi medi necessari per raggiungere uno stato da ogni altro: Lombardia (ricavata dalla matrice dei flussi 1986.I-II)*

	OCC	ONONC	OCCR	INC1	INC2	INONC	INA
OCC	2	800	156	151	184	1.300	40
ONONC	6	588	129	142	179	1.297	38
OCCR	8	773	76	122	165	1.291	38
INC1	16	807	129	45	133	1.274	33
INC2	20	813	145	87	88	1.214	27
INONC	26	821	157	95	102	909	20
INA	37	834	180	145	170	1.271	2

Tab. 12: *Matrice dei tempi medi necessari per raggiungere uno stato da ogni altro: Campania (ricavata dalla matrice dei flussi 1986.I-II)*

	OCC	ONONC	OCCR	INC1	INC2	INONC	INA
OCC	2	4.213	119	143	92	454	33
ONONC	4	2.812	81	139	96	454	34
OCCR	9	4.222	53	126	88	450	30
INC1	20	4.233	109	51	57	438	19
INC2	20	4.234	111	76	32	422	20
INONC	23	4.236	120	94	46	206	17
INA	28	4.242	131	126	82	438	2

3.2. *Descrizione dei principali flussi*

Allo scopo di ottenere una migliore comprensione delle caratteristiche dei flussi tra i sette stati, si considerano ora le distribuzioni di alcune variabili condizionate ai flussi più significativi (Tabb. 13 e 14). Le variabili prese in esame sono il sesso, l'età, il titolo di studio, l'essere o meno sottoccupati all'inizio della ricerca. Si trascurano i flussi che, nella matrice relativa alla Lombardia, interessano meno di 11 persone. I commenti sono svolti con riferimento alla Lombardia e per confronto, ogni volta che la numerosità lo consente, alla Campania. (Essendo l'indagine lombarda sovracampionata, in molti casi a flussi sufficientemente consistenti in Lombardia corrispondono in Campania flussi troppo esigui per poter svolgere considerazioni appena affidabili.)

Le evidenze salienti possono essere così sintetizzate.

(a) Flussi da OCC. Sono considerati i flussi verso i tre stati che, nella sostanza, comprendono la totalità delle persone in qualche modo in cerca di lavoro, vale a dire OCCR, INC1 e INC2. È interessante osservare che quanto più alto nella scala di *attachment* al lavoro è lo stato di arrivo, tanto maggiore è la quota di persone con istruzione elevata, di donne e di sottoccupati. Coloro che passano da OCC a INC2 sono mediamente più anziani, mentre coloro che passano in INC1 sono più giovani.

(b) Flussi da ONONC. Sono considerati i flussi verso OCC e ONONC. Tra le persone che passano in OCC è maggiore la quota di giovani, di donne, di persone con un'istruzione elevata, di non sottoccupati.

(c) Flussi da OCCR. Sono considerati i flussi verso OCC, OCCR, INC1 e INC2. Anche in questo caso, l'ordinamento degli stati di arrivo secondo il grado di *attachment* determina un *pattern* facilmente individuabile nelle distribuzioni secondo il sesso e l'istruzione: al crescere dell'*attachment* cresce la quota di persone poco istruite e di maschi. (Il flusso verso INC1 è peraltro caratterizzato da una maggiore presenza di giovani.) Nel caso della Campania, tuttavia, le considerazioni appene svolte non valgono: il flusso verso OCC è composto da più donne e da persone più istruite del flusso verso OCCR.

(d) Flussi da INC1. In Lombardia i flussi verso OCC e verso INA sono contrassegnati da una presenza di maschi relativamente maggiore che non i flussi verso gli stati caratterizzati da qualche attività di ricerca. Ciò non succede invece, in Campania, dove l'incidenza delle femmine è nettamente prevalente nel flusso da INC1 a INA. Un'altra marcata differenza tra le due regioni è nella composizione per età del flusso INC1-INA: in Lombardia è il flusso dove sono relativamente dominanti i vecchi, mentre in Campania è quello dove prevalgono i giovani. Infine, sia in Lombardia che in Campania si nota una maggiore presenza di sottoccupati nel flusso INC1-OCCR.

(e) Flussi da INC2. In entrambe le regioni i flussi diretti verso stati con maggiore *attachment*, hanno tendenzialmente una maggiore presenza di maschi e di persone poco istruite. Tra i flussi della Campania, quello verso OCC è comparativamente dominato da persone in età avanzata,

Tab. 13: *Distribuzione di alcune variabili nei flussi più significativi della matrice di transizione della Lombardia 1986.I-II* [(a)]

		Stu1	Stu2	Stu3	Sex1	Sex2	Eta1	Eta2	Eta3	Eta4	Sot1	Sot2
OCC	- OCCR	0,02	0,68	0,29	0,51	0,49	0,57	0,19	0,19	0,05	0,21	0,79
	- INC1	-	0,88	0,12	0,61	0,39	0,67	0,12	0,12	0,09	0,09	0,91
	- INC2	-	0,91	0,09	0,77	0,23	0,45	0,14	0,27	0,14	0,09	0,91
ONONC	- OCC	-	0,79	0,21	0,54	0,46	0,50	0,21	0,21	0,08	0,12	0,87
	- ONONC	0,09	0,82	0,09	0,73	0,27	0,36	0,27	0,27	0,09	0,36	0,64
OCCR	- OCC	0,02	0,73	0,25	0,63	0,37	0,64	0,22	0,10	0,04	0,23	0,77
	- OCCR	0,02	0,70	0,28	0,55	0,45	0,57	0,26	0,13	0,04	0,30	0,70
	- INC1	-	0,77	0,23	0,47	0,53	0,73	0,07	0,10	0,10	0,03	0,70
	- INC2	-	0,67	0,33	0,33	0,67	0,67	0,25	0,08	-	0,33	0,67
INC1	- OCC	0,01	0,65	0,34	0,59	0,41	0,82	0,10	0,04	0,03	0,03	0,97
	- OCCR	-	0,59	0,41	0,38	0,62	0,76	0,14	0,08	0,03	0,19	0,81
	- INC1	0,01	0,67	0,32	0,34	0,66	0,84	0,10	0,05	0,02	0,03	0,97
	- INC2	-	0,80	0,20	0,35	0,65	0,72	0,15	0,11	0,02	0,04	0,96
	- INA	0,07	0,67	0,27	0,47	0,53	0,63	0,20	0,13	0,03	-	1,0
INC2	- OCC	-	0,85	0,15	0,57	0,43	0,62	0,15	0,13	0,10	0,07	0,93
	- OCCR	-	0,67	0,33	0,53	0,47	0,53	0,13	0,27	0,07	0,07	0,93
	- INC1	-	0,71	0,29	0,47	0,53	0,77	0,15	0,07	0,02	0,03	0,97
	- INC2	-	0,76	0,24	0,34	0,66	0,61	0,25	0,08	0,05	0,02	0,98
	- INONC	-	0,75	0,25	0,33	0,67	0,42	0,25	0,25	0,08	0,08	0,92
	- INA	0,04	0,64	0,32	0,25	0,75	0,51	0,22	0,18	0,08	-	1,0
INONC	- INC1	-	0,83	0,17	0,17	0,83	0,83	0,17	-	-	-	1,0
	- INC2	-	0,91	0,09	0,27	0,73	0,55	0,27	0,18	-	-	1,0
	- INA	-	0,64	0,36	0,18	0,82	0,45	0,36	-	0,18	-	1,0
INA	- INC1	0,06	0,61	0,33	0,59	0,41	0,78	0,14	0,04	0,04	-	1,0
	- INC2	-	0,81	0,19	0,25	0,75	0,57	0,24	0,13	0,06	-	1,0
	- INONC	-	0,62	0,38	0,38	0,62	0,54	0,15	0,15	0,15	-	1,0

(a) Stu1=nessun titolo; Stu2=licenza elementare o di scuola media inferiore; Stu3=diploma di scuola media superiore o di laurea; Sex1=maschio; Sex2=femmina; Eta1=fino a 29 anni; Eta2=da 30 a 39 anni; Eta3=da 40 a 49 anni; Eta4=oltre 49 anni; Sot1=sottoccupato all'inizio della ricerca;Sot2=non sottoccupato all'inizio della ricerca.

mentre in Lombardia i flussi dove più alta è l'incidenza dei vecchi sembrano essere quelli verso INA e INONC.

(f) Flussi da INONC. Si considerano solo i flussi verso INC2 e INA. In Lombardia gli inattivi con trascurabile attività di ricerca che al trimestre successivo mostrano qualche interesse a lavorare sono mediamente più giovani, meno istruiti, più facilmente di sesso maschile di quelli che transitano invece in INA.

(g) Flussi da INA. In entrambe le regioni le persone che passano da INA a INC1 sono nettamente più giovani delle altre. Il flusso verso INC2 è invece contraddistinto da una maggiore quota di donne e di persone con

Tab. 14: *Distribuzione di alcune variabili nei flussi più significativi della matrice di transizione della Campania 1986.I-II* [a]

		Stu1	Stu2	Stu3	Sex1	Sex2	Eta1	Eta2	Eta3	Eta4	Sot1	Sot2
OCC	- OCCR	0,09	0,70	0,22	0,57	0,43	0,52	0,22	0,17	0,09	0,22	0,78
OCCR	- OCC	-	0,67	0,33	0,70	0,30	0,41	0,41	0,19	-	0,15	0,85
	- OCCR	0,04	0,75	0,21	0,81	0,19	0,60	0,25	0,06	0,08	0,35	0,65
INC1	- OCC	0,07	0,64	0,29	0,86	0,14	0,79	0,14	-	0,07	0,07	0,93
	- INC1	0,02	0,52	0,46	0,53	0,47	0,85	0,11	0,02	0,01	-	1,0
	- INC2	0,04	0,39	0,57	0,57	0,43	0,89	0,07	0,04	-	-	1,0
	- INA	-	0,52	0,48	0,26	0,74	0,96	0,04	-	-	-	1,0
INC2	- OCC	0,06	0,76	0,18	0,82	0,18	0,47	0,18	0,24	0,12	0,02	0,98
	- INC1	0,02	0,76	0,22	0,41	0,59	0,88	0,10	-	0,02	-	1,0
	- INC2	0,02	0,70	0,27	0,41	0,59	0,81	0,11	0,05	0,02	0,02	0,98
	- INA	0,04	0,69	0,27	0,19	0,81	0,81	0,12	0,04	0,04	0,08	0,92
INA	- INC1	-	0,46	0,54	0,46	0,54	0,92	0,08	-	-	-	1,0
	- INC2	0,05	0,54	0,41	0,30	0,70	0,76	0,05	0,14	0,05	-	1,0

(a) Vedi nota alla Tab. 13.

basso livello di istruzione. Il flusso verso INONC presenta, al pari del flusso verso INC2, un'accentuata componente femminile, ma si distingue da questo per un maggiore peso di persone con livello di istruzione ed età elevate.

4. *Qualche riflessione conclusiva*

Le evidenze qui presentate e discusse, per quanto circoscritte nel tempo e nello spazio, e per gli aspetti dinamici limitate a primi approfondimenti, consentono di trarre alcune riflessioni di sintesi di qualche interesse.

Già dal confronto fra diverse classificazioni *a priori*, tutte riconducibili alle raccomandazioni dell'ILO, che avevamo condotto in Rettore, Torelli e Trivellato (1988), risaltava che la determinazione dell'aggregato dei disoccupati e del tasso di disoccupazione è sensibile a varianti definitorie apparentemente di non grande momento. È già questo un sintomo, sia pure indiretto, di fluidità nella collocazione di una frazione non trascurabile di persone tra disoccupazione e inattività.

L'esistenza di un'articolazione nell'*attachment* al mercato del lavoro, della quale le usuali classificazioni *a priori* danno solo parzialmente conto, emerge nitidamente dall'analisi esplorativa condotta sulle persone che si sono dichiarate in cerca di lavoro. I risultati ottenuti evidenziano l'esistenza: (i) da un lato, di un sottoinsieme di occupati, composto essenzialmente da occupati non dichiarati e/o sottoccupati, che è in buona sostanza assimilabile ai

disoccupati, vuoi nella percezione della propria condizione vuoi nell'intensità dell'attività di ricerca; (ii) dall'altro lato, di un insieme di persone che si sono dichiarate in cerca di lavoro e che in realtà non sono attribuibili in modo univoco ai disoccupati o agli inattivi. In particolare, l'esistenza di una cospicua 'area grigia' alla frontiera fra disoccupazione e inattività, contraddistinta da un *attachment* al mercato del lavoro piuttosto modesto e in prima approssimazione ricompresa in INC2, è un problema non meramente classificatorio, ma di rilievo per analisi e per politiche. I risultati ottenuti dall'analisi dei flussi mostrano infatti che, almeno in una delle due regioni considerate - la Lombardia -, questo insieme di disoccupati ha un comportamento dinamico apprezzabilmente diverso da quello dei disoccupati compresi in INC1: è nettamente più bassa la loro probabilità di ottenere un lavoro, ed è nettamente più alta la probabilità che cessino del tutto di cercare lavoro.

Quanto ciò dipenda da atteggiamenti soggettivi nei confronti del lavoro e quanto invece da differenze nelle opportunità di lavoro, i risultati fin qui ottenuti non consentono di dire. In verità, per poter distinguere queste due componenti, e più in generale fattori operanti rispettivamente dal lato dell'offerta e da quello della domanda di lavoro, sono verosimilmente necessari notevoli avanzamenti nell'analisi: tanto sul versante delle informazioni disponibili, quanto su quello dei modelli e metodi di analisi.

PARTE QUINTA:

MODELLI DI ANALISI DELLE FORZE DI LAVORO

16

DESTAGIONALIZZAZIONE DELLE SERIE STORICHE DELLE FORZE DI LAVORO

Silvano Bordignon

1. *Introduzione*

L'informazione fornita da una serie destagionalizzata gioca un ruolo rilevante nell'analisi delle condizioni economiche correnti, in particolare nel determinare lo stadio del ciclo dove l'economia si trova. Tale conoscenza è utile per prevedere i movimenti ciclici successivi e fornisce la base per intraprendere azioni di controllo dell'attività economica. Ciò diventa particolarmente importante attorno ai punti di svolta: ad esempio, la mancata individuazione di un picco nel movimento ciclico può portare all'adozione di politiche di freno dell'espansione quando, in realtà, ci si trova in una fase recessiva.

Data l'importanza dell'informazione fornita dalle serie destagionalizzate, il disporre di dati correttamente destagionalizzati è diventato uno dei principali obiettivi delle varie agenzie statistiche governative. Soprattutto negli ultimi anni, si sta assistendo ad una notevole produzione di studi tendenti a migliorare le procedure di destagionalizzazione attualmente in uso nei vari Paesi, in particolare per quanto riguarda la stima dei valori correnti delle componenti delle serie. Le tendenze più significative finora emerse sono essenzialmente due: (i) l'affermarsi della procedura X-11-ARIMA, introdotta e sperimentata in una varietà di situazioni applicative, specie nell'ambito delle serie del mercato del lavoro, da Dagum (1975, 1978, 1983); (ii) l'attenzione crescente verso procedure di destagionalizzazione secondo un approccio cosiddetto *model-based*, in cui la scelta di una particolare procedura di depurazione stagionale è legata alla identificazione del modello generatore della serie storica in esame (Burman, 1980; Hillmer e Tiao, 1982).

In questo quadro si inserisce il presente studio, che vuole valutare alcune recenti proposte in tema di destagionalizzazione di serie storiche, con riguardo a serie temporali del mercato del lavoro tratte dall'indagine trimestrale sulle forze di lavoro dell'Istat.

Le ragioni di questa scelta sono da ricondursi principalmente alla mancanza in Italia di un'esperienza vasta e continua su tale problema, analoga a quelle delle agenzie statistiche ufficiali dei Paesi anglosassoni (Statistics

Canada, U.S. Bureau of Labor Statistics, Central Statistical Office inglese)[1].

D'altra parte, è evidente che se si vuole sfruttare adeguatamente la periodicità trimestrale della rilevazione sulle forze di lavoro per valutare tempestivamente le tendenze del mercato del lavoro, è opportuno integrare il quadro conoscitivo fornito dall'esame delle serie grezze con un'analisi delle caratteristiche delle componenti delle serie stesse. Tale analisi può permettere di cogliere con maggiore precisione l'andamento ciclico dell' occupazione e della disoccupazione; può inoltre evidenziare la differente reattività ciclica e stagionale di specifiche componenti dell'occupazione e della disoccupazione. Da tale operazione possono infine derivare utili informazioni per la specificazione di modelli interpretativi sul mercato del lavoro.

Ora, ciò è possibile solo se si dispone di un'adeguata procedura di decomposizione. Al riguardo è opportuno chiedersi se una procedura empirica e per lo più automatica, come l'X-11-ARIMA, che si è rivelata sufficientemente adeguata per le serie mensili delle forze di lavoro di Stati Uniti e Canada, è adatta alle caratteristiche delle serie trimestrali italiane, o se procedure più sofisticate ma meno automatizzabili, proposte recentemente in seno all'approccio *model-based*, possono rivelarsi più soddisfacenti per questo specifico settore. Lo studio intende fornire degli elementi per una risposta a tali quesiti, valutando empiricamente due procedure di destagionalizzazione, X-11-ARIMA e MSX (Burman, 1980), su un campione di serie storiche relativo alle forze di lavoro italiane.

Dopo una breve descrizione delle caratteristiche salienti di alcune procedure di destagionalizzazione (sez. 2), nella sez. 3 sono esaminati problemi quali l'identificabilità della componente stagionale, l'adozione di un modello additivo o moltiplicativo e la specificazione di un adeguato modello ARIMA, che tipicamente il produttore di dati destagionalizzati si trova ad affrontare prima di effettuare la destagionalizzazione. Nella sez. 4 sono presentati e discussi alcuni risultati relativi alla bontà dell'aggiustamento stagionale, relativamente alle procedure di destagionalizzazione adottate. Particolare attenzione viene dedicata al problema della stabilità della stima della componente stagionale, e conseguentemente della serie destagionalizzata, data la rilevanza che assume la destagionalizzazione dei dati correnti ai fini dell'analisi economica. Un'ulteriore questione che tipicamente si pone riguarda la destagionalizzazione di una serie aggregata, cioè a dire di una serie risultante dalla combinazione (tipicamente per somma e/o per rapporto) di altre serie. E' questo il caso, ad esempio, della serie storica della forza lavoro che è la somma di due componenti distinte, rispettivamente gli occupati e i disoccupati, le quali a loro volta risultano dall'aggregazione di ulteriori componenti la cui numerosità varia a seconda del livello di dettaglio considerato (sesso, settore, territorio, ecc.). Il tasso di disoccupazione rappresenta un altro esempio di serie aggregata, ottenuta questa volta come rapporto di due

---

1 Un'eccezione è rappresentata dal progetto DESEC (Piccolo, 1985), nel cui ambito è stata condotta una vasta sperimentazione empirica sulla destagionalizzazione che ha interessato serie storiche italiane relative a vari settori dell'economia, tra cui alcune riguardanti il mercato del lavoro.

serie. Tale questione viene affrontata nella sez. 5, mentre alla sez. 6 sono affidate alcune considerazioni conclusive.

2. *Procedure di destagionalizzazione*

La maggior parte delle procedure di destagionalizzazione possono essere classificate in due categorie: le procedure empiriche e quelle *model-based.* Appartengono alla prima categoria le procedure di carattere essenzialmente descrittivo, nel senso che nessun modello parametrico è esplicitamente formalizzato per le componenti della serie. Tali procedure, basate per lo più su metodi di filtraggio, quali, ad esempio, le medie mobili, si sono sviluppate nel corso degli anni facendo leva più sulla sperimentazione empirica che su considerazioni teoriche, con la conseguenza che le loro proprietà statistiche risultano non ben definite. La maggioranza delle procedure di destagionalizzazione adottate dalle agenzie statistiche ufficiali nei vari Paesi rientra in questa categoria. Fra di esse, la più collaudata nel tempo e nella versatilità delle opzioni è senz'altro l'X-11 del U.S. Bureau of the Census (Shiskin, Young e Musgrave, 1967), che, tuttavia, non sempre ha fornito buoni risultati in termini di stima delle componenti relativamente alle osservazioni più recenti. Per ovviare a tale limitazione, Dagum (1975) ha proposto una versione modificata dell'X-11, l'X-11-ARIMA, consistente nell'estendere la serie originaria con un certo numero di previsioni, ottenute da un modello ARIMA adattato alla serie, e nel decomporre quindi la serie estesa con l'X-11. Tale procedura aumenta il grado di affidabilità relativo alla stima corrente delle componenti della serie, riducendo significativamente la grandezza delle revisioni, una volta che si rendono disponibili nuovi dati.

La critica principale che viene mossa alle procedure appena menzionate riguarda la mancanza di un modello esplicito concernente la decomposizione della serie, a cui possa essere collegato un qualche criterio statistico di ottimalità. Questa insoddisfazione per la natura empirica di tali procedure si è fatta più evidente negli anni recenti, in conseguenza dello sviluppo e dell'uso di modelli per serie storiche, quali soprattutto i modelli ARIMA. Ciò ha portato molti ricercatori a sviluppare delle procedure di decomposizione che assumono esplicitamente modelli statistici per la serie e/o per ciascuna componente. Esse si basano generalmente su una decomposizione additiva del tipo $Z_t = S_t + N_t$, dove Z_t è la serie da decomporre, o una qualche sua trasformata (ad es., logaritmica), mentre S_t e N_t rappresentano rispettivamente le componenti stagionale e non stagionale (N_t può essere ulteriormente scomposta in due parti per rappresentare distintamente il trend-ciclo e la componente irregolare), e utilizzano modelli statistici per Z_t , S_t e N_t . Il modello per Z_t può essere stimato dalle osservazioni, ma dato che S_t e N_t sono non osservabili, i rispettivi modelli poggiano su assunzioni arbitrarie, per quanto plausibili. I vari metodi differiscono a seconda del tipo di modello adottato per Z_t e a seconda delle assunzioni utilizzate nello specificare i modelli per S_t e N_t .

Si possono distinguere sostanzialmente due approcci allo sviluppo di

procedure *model-based* per la decomposizione di una serie. Nel primo si assume che Z_t segua un modello ARIMA, dal quale, sotto alcune assunzioni, vengono dedotti univocamente modelli della stessa classe per S_t e N_t . Tali componenti sono quindi stimate utilizzando la teoria di estrazione del segnale. Box, Hillmer e Tiao (1978), partendo da un modello ARIMA $(0,1,1)(0,1,1)_{12}$ per Z_t , hanno derivato modelli per le componenti coerenti col modello di partenza facendo uso di certe assunzioni, tra cui quella della massimizzazione della varianza della componente erratica. Questo approccio, esteso successivamente a modelli ARIMA più generali da Burman (1980) e da Hillmer e Tiao (1982), completato con opzioni per il trattamento degli *outliers*, ha raggiunto ormai uno stato di sviluppo tale da risultare operativo. Esistono infatti dei *packages*, quale, ad es., MSX di Burman, che utilizzano appunto questa procedura per la decomposizione.

Il secondo approccio, noto anche come approccio strutturale, data la similarità col tema dell'identificazione di una classe di strutture coerenti con i dati nei modelli econometrici, implica la specificazione di modelli parametrici per le componenti della serie (modelli strutturali), che a sua volta porta ad un modello per Z_t soggetto a dei vincoli (forma ridotta). I problemi consistono nello stimare il modello per Z_t soggetto a tali vincoli e nella identificabilità dei parametri dei modelli strutturali per le componenti. Una volta risolti questi problemi, la stima delle componenti viene effettuata tramite l'estrazione del segnale. La maggiore difficoltà di questo approccio sta nel fatto che la stima del modello complessivo per Z_t , soggetto ai vincoli imposti dalla struttura dei modelli per le componenti, può risultare un compito abbastanza arduo. In effetti Engle (1978), partito da modelli ARIMA sufficientemente generali per le componenti, non essendo in grado di stimare il modello complessivo soggetto a tutti i vincoli, è costretto alla fine ad allentarne alcuni. L'operatività di questo approccio è quindi condizionata, almeno per il momento, alla semplificazione delle strutture imposte per i modelli delle componenti della serie: in questa direzione sono da collocare le recenti proposte di Akaike (1980), Harvey (1984) e Kitagawa e Gersch (1984). Da un punto di vista applicativo, resta da verificare se tale semplificazione compromette la possibilità per i modelli strutturali di rappresentare adeguatamente le serie che si incontrano nella pratica. Alcune prime esperienze applicative (Akaike e Ishiguro, 1983; Harvey e Todd, 1983) sembrano, al riguardo, piuttosto incoraggianti.

Questo rapido sguardo alle recenti tendenze sul problema della decomposizione di una serie storica serve ad inquadrare la scelta delle procedure messe a confronto in questa sede. Esse sono l'X-11-ARIMA, attualmente la procedura più rappresentativa nella categoria delle procedure empiriche, e MSX di Burman che, al momento, nell'ambito dell'approccio *model-based* rappresenta l'alternativa più valida da un punto di vista operativo[2].

---

2 Per una descrizione dettagliata di queste procedure si rinvia ai lavori originali, mentre le principali caratteristiche operative si possono trovare in Bordignon e Masarotto (1987), dove, tra l'altro, sono considerate e sperimentate altre proposte nell'ambito dell'approccio *model-based*, come la procedura BAYSEA di Akaike.

3. *Analisi preliminari*

3.1. *Serie analizzate*

Le serie utilizzate sono desunte dalle indagini sulle forze di lavoro condotte dall'Istat. Si tratta di un insieme di serie trimestrali che concorrono a formare degli aggregati, quali gli occupati, i disoccupati e il tasso di disoccupazione, i quali in varia misura costituiscono degli importanti indicatori delle tendenze del mercato del lavoro. Per gli occupati abbiamo utilizzato la classificazione per settore di attività economica e per i disoccupati la disaggregazione secondo la condizione rispetto alla ricerca di lavoro. Tutte le serie considerate sono distinte per sesso. L'elenco completo delle serie è riportato nella Tab. 1. Il periodo temporale a cui esse fanno riferimento va dal I trimestre del 1970 al I trimestre del 1987[3] .

Tab. 1: *Serie analizzate (periodo 1970.I-1987.I)*

OAM	Occupati nell'agricoltura - maschi
OIM	Occupati nell'industria - maschi
OAAM	Occupati in altre attività - maschi
OM	Totale occupati - maschi
DM	Disoccupati in senso stretto - maschi
PM	Persone in cerca di prima occupazione - maschi
AM	Altre persone in cerca di occupazione - maschi
TOTM	Totale persone in cerca di occupazione - maschi
FLM	Forze di lavoro - maschi
TDM	Tasso di disoccupazione - maschi
OAF	Occupati nell'agricoltura - femmine
OIF	Occupati nell'industria - femmine
OAAF	Occupati in altre attività - femmine
OF	Totale occupati - femmine
DF	Disoccupati in senso stretto - femmine
PF	Persone in cerca di prima occupazione - femmine
AF	Altre persone in cerca di occupazione - femmine
TOTF	Totale persone in cerca di occupazione - femmine
FLF	Forze di lavoro - femmine
TDF	Tasso di disoccupazione - femmine
TD	Tasso di disoccupazione - totale

3 Per il periodo 1970.I-1976.IV, abbiamo fatto uso delle serie ricostruite da Sanetti e Settani (1979), a seguito delle modificazioni introdotte a partire dal 1977 nell'indagine sulle forze di lavoro.

3.2. *Verifica della presenza di stagionalità*

Prima di procedere alla destagionalizzazione di una serie è buona norma verificare se e in che misura la stagionalità è presente nella serie stessa. Infatti per molte serie economiche, mensili o trimestrali, la presenza di evidente stagionalità non è una regola, vuoi perchè le fluttuazioni stagionali mancano del tutto, vuoi perchè l'influenza della stagionalità è talmente poco significativa da risultare irrilevante per la serie, specie quando il suo peso è piccolo se paragonato a quello della componente irregolare.

In questi casi l'applicazione alla serie di una procedura di destagionalizzazione potrebbe portare a risultati fuorvianti, come hanno dimostrato alcuni esperimenti di simulazione condotti da Dagum (1976).

Per poter stimare con sufficiente accuratezza la componente stagionale, qualsiasi sia il metodo di destagionalizzazione adottato, è perciò importante accertare preliminarmente se c'è sufficiente stagionalità nella serie.

Più precisamente, si richiede alla componente stagionale il cosiddetto requisito dell'identificabilità: ciò significa che le variazioni stagionali, oltre che essere presenti nella serie, debbono essere 'grandi', e quindi distinguibili, rispetto a quelle attribuibili alla componente irregolare.

La presenza di stagionalità identificabile nelle serie delle forze di lavoro è stata verificata in vari modi[4]. Per compattezza di esposizione riteniamo sufficiente presentare in questa sede solo i risultati relativi ad una statistica di tipo R^2 , che misura la frazione di varianza della serie detrendizzata spiegata da variabili *dummy* stagionali. Più precisamente, se indichiamo con Z_t la serie originale e con Y_t la corrispondente stima del trend, la statistica utilizzata è

$$\frac{\text{varianza di } X_t \text{ spiegata dalle } \textit{dummies} \text{ stagionali}}{\text{varianza totale di } X_t}, \qquad (1)$$

dove $X_t = Z_t - Y_t$ se l'aggiustamento è additivo, mentre $X_t = Z_t / Y_t$ se l'aggiustamento è moltiplicativo. Questo indice, che è essenzialmente analogo al test utilizzato dal programma X-11 per verificare la presenza di stagionalità, misura le differenze esistenti tra le medie di X_t nei vari trimestri. Ha perciò come punto di riferimento un modello di stagionalità costante e quindi fornisce un limite inferiore al peso delle stagionalità.

La Tab. 2 riporta per tutte le serie considerate e separatamente per le due procedure di destagionalizzazione i valori dell'indice, che, come si può notare, sono sempre molto elevati e quindi segnalano la presenza di una forte componente stagionale in tutte le serie.

---

4 Per alcune indicazioni circa i modi di verificare la presenza di stagionalità identificabile in una serie, vedi Dagum (1979).

Tab. 2: *Identificabilità della stagionalità* [(a)]

Serie	X-11-ARIMA	MSX	Serie	X-11-ARIMA	MSX
OAM	0,970	0,949	DM	0,988	0,979
OIM	0,959	0,941	PM	0,955	0,947
OAAM	0,985	0,983	AM	0,957	0,950
OM	0,992	0,996	TOTM	0,959	0,960
FLM	0,990	0,990	TDM	0,960	0,963
OAF	0,991	0,981	DF	0,958	0,931
OIF	0,920	0,863	PF	0,985	0,976
OAAF	0,959	0,979	AF	0,953	0,963
OF	0,98	0,984	TOTF	0,969	0,966
FLF	0,98	0,978	TDF	0,950	0,945
TD	0,52	0,950			

(a) Per la statistica utilizzata, vedi l'equazione (1).

3.3. *Scelta tra modello additivo o moltiplicativo*

Un altro problema che richiede una certa attenzione quando si vuole destagionalizzare una serie, riguarda la scelta tra specificazione additiva o moltiplicativa del modello di composizione del trend-ciclo, della stagionalità e dell'errore.

In generale, non c'è alcuna motivazione teorica che faccia preferire l'una o l'altra specificazione: semplicemente esse riflettono assunzioni differenti circa il meccanismo generatore della serie. In un modello additivo, le componenti della serie sono indipendenti tra di loro e pertanto l'effetto stagionale non è influenzato dal livello dell'attività economica, descritto dal trend-ciclo. Al contrario, in un modello moltiplicativo l'effetto stagionale è proporzionale al trend-ciclo. Ciò significa che, se anche i fattori stagionali sono costanti, più alto è il livello della serie destagionalizzata, più grande è l'effetto stagionale.

Una scelta errata della specificazione non è senza conseguenze per quanto riguarda la stima della componente stagionale. Ad esempio, questa può risultare di tipo evolutivo quando in realtà si tratta di un effetto stagionale piuttosto stabile nel tempo, e ciò porta inevitabilmente ad una cattiva destagionalizzazione. La scelta corretta del modello è importante soprattutto per quanto riguarda la destagionalizzazione dei dati più recenti di serie storiche che manifestano una rapida crescita nel trend-ciclo.

Le serie delle forze di lavoro sono state sottoposte a verifica per determinare quale delle due specificazioni alternative, l'additiva o la moltiplicativa, possiede la componente stagionale più stabile, tramite un test preliminare suggerito da Dagum (1979). Il test è costruito nel modo seguente. Si stima il trend-ciclo della serie, unitamente ad una componente stagionale stabile. Viene quindi formata la parte sistematica della serie, sia additivamente che moltiplicativamente. Le due parti sistematiche così ottenute sono sottratte dalla serie originale, ed è considerato migliore il modello per il quale la

somma dei quadrati dei residui è significativamente più piccola.

L'applicazione di tale test non sempre ha portato univocamente alla scelta di una specificazione, data la non rilevante differenza tra le due quantità (Dagum suggerisce che il test è da considerarsi inconclusivo se la differenza tra i due valori è inferiore al 20%). Nelle situazioni dubbie abbiamo condotto la destagionalizzazione con entrambe le specificazioni, decidendo infine per quella che ha prodotto l'insieme migliore delle statistiche di controllo sulla qualità della destagionalizzazione. Le specificazioni adottate per ciascuna serie a seguito di queste operazioni sono riportate nella prima colonna della Tab. 3.

Tab. 3: *Identificazione dei modelli*

Serie	Tipo di modello[a]	Modello ARIMA[b]		
		AUT	OWN	MSX
OAM	A	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$
OIM	M	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$
OAAM	M	$(0,2,2)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$
OM	M	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$	$(2,1,0)(0,1,0)_4$
DM	M	$(0,1,1)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$
PM	A	$(0,1,1)(0,1,1)_4$	$(0,1,0)(1,0,1)_4$	$(0,1,0)(0,1,1)_4$
AM	M	nessuno	$(1,0,0)(1,0,1)_4$	$(1,0,0)(0,1,1)_4$
TOTM	A	$(0,1,1)(0,1,1)_4$	$(0,1,0)(1,0,1)_4$	$(0,1,0)(1,0,1)_4$
FLM	A	$(0,1,1)(0,1,1)_4$	$(3,0,0)(0,1,1)_4$	$(3,0,0)(0,1,1)_4$
TDM	M	$(0,1,1)(0,1,1)_4$	$(0,1,0)(1,0,1)_4$	$(0,1,0)(1,0,1)_4$
OAF	M	$(0,1,1)(0,1,1)_4$	$(0,1,2)(0,1,1)_4$	$(0,1,2)(0,1,1)_4$
OIF	M	$(0,2,2)(0,1,1)_4$	$(0,1,0)(1,0,0)_4$	$(0,1,0)(1,0,0)_4$
OAAF	A	$(0,2,2)(0,1,1)_4$	$(2,1,0)(0,1,2)_4$	$(2,1,0)(0,1,2)_4$
OF	M	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,0)_4$
DF	A	$(2,1,2)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$
PF	M	$(0,1,1)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$	$(2,0,0)(0,1,1)_4$
AF	M	$(0,1,1)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$
TOTF	M	$(0,1,1)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$	$(0,1,0)(0,1,1)_4$
FLF	A	$(0,1,1)(0,1,1)_4$	$(0,1,2)(0,1,1)_4$	$(0,1,2)(0,1,1)_4$
TDF	A	$(0,1,1)(0,1,1)_4$	$(1,1,0)(0,1,1)_4$	$(1,1,0)(0,1,1)_4$
TD	A	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$	$(0,1,1)(0,1,1)_4$

(a) M indica il modello moltiplicativo, A quello additivo.

(b) In un modello ARIMA stagionale $(p,d,q)\ (P,D,Q)_s$, p e P rappresentano il numero dei parametri autoregressivi non stagionali e stagionali, d e D il numero delle differenze non stagionali e stagionali, q e Q il numero dei parametri a media mobile non stagionali e stagionali, s il periodo stagionale.

3.4. *Identificazione dei modelli ARIMA*

L'applicazione della procedura MSX richiede l'identificazione preliminare di un modello ARIMA per la serie da decomporre. Tale identificazione può servire anche da input per la procedura X-11-ARIMA, nel caso l'utilizzatore voglia fornire un proprio modello ARIMA, anzichè sfruttare l'opzione per l'identificazione automatica. Nel seguito indicheremo con la sigla AUT la versione automatica dell'X-11-ARIMA e con OWN l'altra.

Pertanto per tutte le serie prese in considerazione sono stati identificati e stimati dei modelli ARIMA secondo la procedura standard di Box e Jenkins (1976). I risultati dell'identificazione sono presentati nella Tab. 3. Si può rilevare come la stagionalità delle serie abbia richiesto una differenziazione stagionale per la maggior parte di esse, mentre per ottenere delle serie stazionarie il più delle volte è risultata sufficiente una singola differenziazione non stagionale. Inoltre quasi tutti i modelli includono un parametro a media mobile stagionale, che ha il compito di catturare la parte di stagionalità stocastica, rimasta dopo l'operazione di differenziazione. La frequenza della struttura $(0,1,1)_4$ per la parte stagionale, riscontrata sulle serie del mercato del lavoro italiano, concorda con i risultati ottenuti in analogo ambiente applicativo nei paesi anglosassoni. Come si nota dalla Tab. 3, non sempre i modelli identificati per MSX e per X-11-ARIMA sono gli stessi. Per tratteggiare una spiegazione, è utile ricordare la diversa funzione che hanno i modelli ARIMA nelle due procedure. Nell'X-11-ARIMA il modello ARIMA ha una funzione esclusivamente previsiva, dato che serve solo ad estendere la serie originale, mentre il nucleo della procedura di decomposizione poggia sui filtri a media mobile dell'X-11. Al contrario, in MSX la procedura di decomposizione si basa proprio sulle caratteristiche del modello ARIMA adattato ai dati. Ora la procedura di identificazione di Box e Jenkins non sempre porta ad un modello unico per una serie storica: spesso fornisce diverse alternative, equivalenti da un punto di vista di adattamento ai dati, tra le quali non è facile discriminare. Inoltre, il programma MSX non accetta, in qualche circostanza, l'identificazione originaria del modello ARIMA da noi proposta, per la presenza di *outliers* nei residui e/o per i vincoli imposti dal programma stesso alla struttura del modello. Sulla base di queste considerazioni abbiamo riportato nella terza e quarta colonna della Tab. 3 le strutture dei modelli che, a nostro avviso, meglio possono assolvere la funzione loro assegnata all'interno di ciascuna procedura di decomposizione, compatibilmente con i vincoli imposti per le due procedure dai rispettivi programmi.

La seconda colonna della Tab. 3 presenta invece le caratteristiche dei modelli ARIMA scelti automaticamente dal programma X-11-ARIMA. L'identificazione automatica differisce in qualche caso da quella da noi proposta. In aggiunta, se il programma non è in grado di trovare un modello adeguato secondo i propri criteri, la procedura di decomposizione coincide con quella dell'X-11. Tale evento, indicato da 'nessun modello' nella Tab. 3, accade per la serie AM.

4. *Analisi 'ex-post' sulle serie destagionalizzate*

Una volta destagionalizzata una serie, si rende opportuna un'analisi *ex-post* per valutare l'accuratezza della stima della stagionalità e l'adeguatezza della procedura di destagionalizzazione.

Vari criteri sono stati proposti ed utilizzati nella letteratura per valutare la bontà di un aggiustamento stagionale e/o per confrontare procedure di destagionalizzazione alternative. Generalmente tali criteri forniscono ciascuno informazioni su aspetti particolari. Gli aspetti che sono comunemente privilegiati sono i seguenti: (i) lisciamento (*smoothness*) della serie destagionalizzata ed eventualmente del trend-ciclo; (ii) assenza di residui stagionali nella componente irregolare; (iii) stabilità nella stima della componente stagionale. Nel seguito consideriamo ordinatamente questi tre aspetti.

4.1. *Grado di lisciamento ('smoothness')*

Uno degli scopi principali della destagionalizzazione è quello di fornire una serie di interpretazione più agevole rispetto a quella originale, che evidenzi in particolare la componente trend-ciclo. Poiché si assume che tale componente sia una funzione liscia del tempo, una misura della bontà della procedura di destagionalizzazione è appunto costituita dal grado di lisciamento della serie destagionalizzata.

Tra le varie misure del grado di lisciamento esistenti in letteratura, utilizziamo una versione relativa dell'indice basato sulle variazioni assolute, data da

$$C_1 = \frac{100}{n-1} \sum_t \left| \frac{Z_t - Z_{t-1}}{Z_{t-1}} \right| , \qquad (2)$$

dove n è il numero delle osservazioni, mentre Z_t rappresenta o la stima del trend-ciclo o la serie destagionalizzata o la serie originale.

Nella Tab. 4 sono riportati i valori dell'indice di lisciamento relativi alla serie originale, al trend-ciclo e alla serie destagionalizzata per alcune serie delle forze di lavoro e per ciascuna procedura di destagionalizzazione.

Dall'esame della tabella si possono trarre alcune considerazioni. Innanzi tutto e secondo le attese, la serie destagionalizzata e, a maggior ragione, il trend-ciclo risultano generalmente più lisci della corrispondente serie originale, seppure in modo diversificato a seconda della serie considerata. In linea di massima le serie dei disoccupati, disaggregate secondo la condizione rispetto alla ricerca di lavoro e per sesso, manifestano un minor grado di lisciamento sia rispetto al loro totale che rispetto alle restanti serie. Ciò risulta comprensibile, considerando le maggiori irregolarità a cui sono soggette le serie disaggregate rispetto ai loro totali, dove tali irregolarità possono

Tab. 4: *Media delle variazioni assolute percentuali per il trend (a) e per la serie destagionalizzata (b).*

Serie	Serie originale	AUT		OWN		MSX	
		(a)	(b)	(a)	(b)	(a)	(b)
Maschi							
DM	15,622	4,510	7,447	4,513	7,409	3,867	6,932
PM	7,963	3,122	4,043	3,121	4,040	2,916	4,267
AM	12,711	5,200	9,059	5,165	8,959	5,186	10,167
TOTM	7,286	2,382	3,405	2,392	3,421	2,215	3,261
OM	0,902	0,237	0,423	0,237	0,423	0,313	0,483
FLM	0,871	0,212	0,381	0,211	0,381	0,213	0,352
TDM	7,162	2,398	3,456	2,409	3,483	2,217	3,306
Femmine							
DF	11,902	5,453	8,817	5,462	8,821	4,004	9,493
PF	9,556	2,807	4,257	2,811	4,255	2,353	4,224
AF	11,078	3,681	6,878	3,756	6,837	3,394	6,862
TOTF	8,151	2,282	4,133	2,286	4,122	2,003	4,076
OF	2,062	0,705	1,094	0,705	1,094	0,872	1,289
FLF	2,255	0,752	1,162	0,751	1,163	0,588	1,195
TDF	6,648	1,094	3,624	1,902	3,614	1,627	3,373

trovare compensazione. In secondo luogo le due varianti della procedura X-11-ARIMA forniscono un grado di lisciamento molto simile (ovviamente, è esattamente lo stesso quando i modelli ARIMA coincidono). Anche questo è un risultato atteso, dato che in sede di analisi storica della procedura di decomposizione assumono maggior rilievo i filtri centrali simmetrici propri dell'X-11, che in entrambi i casi sono gli stessi. Infine, per quanto riguarda la procedura MSX è da notare come questa produca, nella maggior parte dei casi, trend e serie destagionalizzate con indici di lisciamento più piccoli rispetto a quelli ottenuti con l'X-11-ARIMA. Tale caratteristica, tipica delle procedure *model-based*, è legata in gran parte alle ipotesi e ai vincoli che vengono posti sulle componenti di una serie, e trova riscontro in analoghi risultati ottenuti in altri studi empirici (Burman, 1980; den Butter, Coenen e van de Gevel 1985).

4.2. *Esame della componente irregolare*

La presenza di stagionalità residua in una serie destagionalizzata è certamente una caratteristica non desiderabile: è questo, infatti, un segnale di cattiva qualità dell'aggiustamento stagionale. Un altro criterio utile per verificare la bontà della destagionalizzazione è perciò costituito dall'esame della componente irregolare. In generale è auspicabile che la stima della

componente irregolare prodotta dalla procedura di decomposizione non contenga residui sistematici, periodici e non, appartenenti alle altre componenti.

Per tale scopo abbiamo utilizzato: (i) la funzione di autocorrelazione della stima della componente irregolare I_t , definita da

$$r_k = \frac{\Sigma (I_t - \bar{I})(I_{t-k} - \bar{I})}{\Sigma (I_t - \bar{I})^2}, \qquad k = 0,1,..n-k, \tag{3}$$

con particolare attenzione per i valori di r_1 e r_4 ; (ii) il test di Ljung-Box

$$Q = n(n+2) \sum_{j=1}^{16} r_j^2/(n-j), \tag{4}$$

e la corrispondente versione stagionale

$$Q_4 = n(n+2) \sum_{j=1}^{4} r_{4j}^2/(n-4j), \tag{5}$$

Nella Tab. 5 sono riportate le statistiche Q e Q_4 sulla componente irregolare per alcune delle serie delle forze di lavoro, sempre con riguardo alle due procedure di destagionalizzazione adottate. Per avere un'idea dell'ordine di grandezza di Q e Q_4 , considerato che sono state utilizzate le prime 16 autocorrelazioni per calcolare Q e le prime quattro autocorrelazioni stagionali per il calcolo di Q_4 , abbiamo confrontato tali quantità rispettivamente con dei χ^2_{15} e χ^2_3, ai livelli di significatività $\alpha = 0,05$ e 0,01. Dall'esame della tabella, per quanto riguarda Q risulta che la componente irregolare ottenuta con la procedura X-11-ARIMA, in entrambe le sue varianti (AUT e OWN), manifesta evidenti segnali di autocorrelazione per quasi tutte le serie esaminate, mentre la componente irregolare prodotta da MSX sembra complessivamente, con alcune eccezioni, più in linea con le caratteristiche di un *white noise*. Guardando poi alla statistica Q_4 , particolarmente utile per valutare una procedura di destagionalizzazione in quanto considera solo le autocorrelazioni ai ritardi stagionali, si può notare il buon comportamento complessivo delle procedure X-11-ARIMA e MSX, che, specie per le serie delle femmine, producono una componente irregolare senza segni evidenti di stagionalità residua.

L'esame della funzione di autocorrelazione della componente irregolare conferma queste indicazioni. Di tale funzione nella Tab. 6 sono riportati i valori ai ritardi 1 e 4. E' da notare che r_1 e r_4 risultano entrambi negativi nella quasi totalità dei casi: in presenza di autocorrelazioni significative, e questo succede soprattutto per r_1 in X-11-ARIMA e r_4 in qualche caso di MSX, ciò indica che la procedura di decomposizione ha probabilmente sovrastimato il trend e la componente stagionale, producendo così autocorrelazioni negative nella componente residua.

Tab. 5: *Tests Q e Q_4 sulla componente irregolare*

Serie	AUT		OWN		MSX	
	Q	Q_4	Q	Q_4	Q	Q_4
Maschi						
DM	40,329*	3,876	40,113*	4,144	20,312	14,951*
PM	46,373*	2,519	47,044*	2,609	11,718	0,518
AM	17,583	1,207	17,460	1,281	6,157	1,260
TOTM	42,062*	11,573*	37,815*	9,103+	65,742*	17,705*
OM	41,750*	11,373*	41,750*	11,333	10,332	1,550
FLM	36,115*	4,031	35,634*	4,002	3,882	0,216
TDM	38,435*	6,771	37,284*	6,097	59,253*	16,545*
Femmine						
DF	35,205*	4,622	35,180*	4,594	17,131	4,067
PF	71,731*	11,698*	75,900*	12,472*	11,605	1,606
AF	23,814	5,236	23,246+	5,375	32,325*	5,016
TOTF	45,518*	5,594	47,885*	5,731	14,469	1,970
OF	36,981*	3,256	36,942*	3,247	6,713	1,051
FLF	27,469+	2,298	27,238+	2,246	16,795	3,664
TDF	42,424*	4,031	42,888*	4,250	54,578	4,811

* significativo rispetto a un χ^2_{15} (χ^2_3) $\alpha = 0,01$

\+ significativo rispetto a un χ^2_{15} (χ^2_3) $\alpha = 0,05$

Tab. 6: *Autocorrelazioni della componente irregolare: (a) r_1 ; (b) r_4*

Serie	AUT		OWN		MSX	
	(a)	(b)	(a)	(b)	(a)	(b)
Maschi						
DM	-0,546	-0,132	-0,540	-0,143	-0,078	-0,425
PM	-0,582	-0,149	-0,583	-0,150	-0,111	-0,078
AM	-0,368	-0,020	-0,377	-0,020	-0,134	0,082
TOTM	-0,512	-0,092	-0,486	-0,086	-0,644	-0,276
OM	-0,357	-0,176	-0,357	-0,176	-0,211	0,035
FLM	-0,437	-0,038	-0,439	-0,039	-0,061	-0,030
TDM	-0,527	-0,059	-0,520	-0,055	-0,624	-0,274
Femmine						
DF	-0,505	-0,176	-0,506	-0,175	-0,171	-0,068
PF	-0,606	-0,275	-0,614	-0,278	-0,187	0,017
AF	-0,383	-0,103	-0,411	-0,101	-0,452	-0,191
TOTF	-0,516	-0,143	-0,533	-0,151	-0,097	-0,028
OF	-0,454	-0,013	-0,454	-0,013	-0,126	0,014
FLF	-0,432	-0,141	-0,418	-0,137	-0,096	-0,194
TDF	-0,589	-0,104	-0,593	-0,106	-0,319	-0,199

È chiaro quindi che se lo scopo principale della procedura di decomposizione è ottenere una serie da cui risultino eliminate le variazioni stagionali, allora l'X-11-ARIMA pare più adeguato. Se al contrario lo scopo della procedura è di ottenere delle componenti interpretabili come trend-ciclo, stagionalità e componente irregolare, allora i risultati sia sul lisciamento del trend-ciclo che sulla autocorrelazione della componente irregolare sembrano favorire MSX, che manifesta complessivamente un comportamento più equilibrato.

4.3. *Analisi della stabilità delle stime della componente stagionale*

Generalmente le stime della componente stagionale, e quindi le serie destagionalizzate, sono soggette a revisioni. Infatti, la necessità di disporre di valutazioni il più possibile tempestive degli andamenti dei vari aggregati costringe a pubblicare una prima valutazione delle serie aggiustate appena diventa disponibile il dato grezzo. Con l'accumulo di ulteriori osservazioni risulta però possibile rivedere i valori già pubblicati. Al proposito c'è da osservare che l'utilità di una procedura di aggiustamento stagionale è notevolmente ridotta se le revisioni apportate sono grandi: revisioni rilevanti indicano la non affidabilità delle prime stime pubblicate, che viceversa sono probabilmente quelle più interessanti per l'utente.

Il criterio della stabilità serve per vagliare in quale misura le stime della stagionalità cambiano quando si aggiungono nuove osservazioni. Indicando con $S_t(l)$ la stima della componente stagionale (o della serie destagionalizzata) riferita al tempo t ed effettuata con t+*l* osservazioni, allora $R_t(l,r)=S_t(r)-S_t(l)$, r>l, viene chiamata revisione e rappresenta l'errore compiuto nello stimare S_t con t+*l* osservazioni, rispetto alla stima effettuata aggiungendone altre r-*l*. Come criterio di stabilità abbiamo utilizzato un indice basato sulle revisioni che si ottengono al variare di r e/o *l*, dato da

$$C_2=\frac{100}{m-1}\sum_r\left|\frac{S_t(r)-S_t(l)}{S_t(r)}\right| . \qquad (6)$$

L'analisi di stabilità è stata poi articolata nei seguenti tre aspetti:

(a) un confronto, essenzialmente interno ai singoli metodi, tra le diverse strategie utilizzabili per produrre dei dati aggiustati in corso d'anno;
(b) un confronto tra i metodi concernente le revisioni apportate nel breve e medio periodo: per fissare le idee, possiamo pensare alle revisioni dei primi due anni;
(c) una analisi della velocità di convergenza delle stime ad un valore definitivo.

4.3.1. *Aggiustamento in corso d'anno*

Con riferimento all'aggiustamento stagionale in corso d'anno è possibile distinguere due strategie differenti (vedi, ad es., McKenzie, 1984):

(a) Aggiustamento basato sui fattori stagionali previsti. I dati vengono aggiustati alla fine di ogni anno, ottenendo oltre alle serie aggiustate relative agli anni passati anche delle previsioni dei fattori stagionali per l'anno successivo. Questi ultimi sono poi usati per destagionalizzare le nuove osservazioni man mano che queste diventano disponibili;

(b) Aggiustamento concorrente. I fattori stagionali sono ristimati tutte le volte che una nuova osservazione diventa disponibile. Per l'aggiustamento dei dati in corso d'anno si utilizzano quindi dei fattori stagionali calcolati su tutte le osservazioni, comprese quelle dell'anno in corso.

Il primo approccio richiede meno calcoli e garantisce inoltre una maggiore trasparenza, visto che i fattori stagionali sono determinati prima della loro applicazione. Il secondo approccio, viceversa, utilizza tutte le informazioni disponibili e quindi può risultare più efficiente, in particolare per serie che presentano una stagionalità che cambia nel tempo. La seconda strategia permette poi di ottenere delle stime in corso d'anno non solo della serie destagionalizzata, ma anche del trend-ciclo e della componente irregolare.

Per quanto concerne il confronto, si osservi che le serie aggiustate utilizzando i due approcci diventano uguali ogni qualvolta le osservazioni relative ad un determinato anno sono complete. Le differenze riguardano perciò solamente i valori pubblicati durante l'anno. Si tratta quindi di un confronto tra le revisioni apportate alle stime in corso d'anno man mano che le osservazioni si accumulano. Per effettuarlo abbiamo calcolato, per gli anni 1984-85, le serie aggiustate utilizzando ambedue le strategie. I valori così ottenuti sono stati poi comparati con le serie destagionalizzate calcolate alla fine dell'anno: alcune misure delle discrepanze sono riportate nella Tab. 7. E' immediato osservare come, indipendentemente dalla scelta del metodo di destagionalizzazione, i risultati suggeriscano l'utilizzo della strategia basata sull'aggiustamento concorrente.

4.3.2. *Un'analisi più accurata della stabilità delle stime concorrenti*

Visti questi risultati, abbiamo deciso di analizzare più in dettaglio le proprietà di stabilità delle stime concorrenti. Il periodo preso in considerazione per il confronto va dal 1983.II al 1985.I. Per ogni serie e per ogni trimestre del periodo di riferimento abbiamo messo a confronto le seguenti stime delle serie destagionalizzate: (1) stima concorrente; (2) stima ottenuta alla fine del trimestre successivo; (3) stima ottenuta dopo due trimestri; (4) stima ottenuta dopo un anno; (5) stima ottenuta dopo due anni. Per fare un esempio, le stime confrontate che si riferiscono al 1985.I sono quelle ottenute utilizzando tutte le osservazioni fino al: (1) 1985.I; (2) 1985.II; (3) 1985.III; (4) 1986.I; (5) 1987.I.

Tab. 7: *Confronto tra aggiustamento basato su fattori stagionali previsti (a) e aggiustamento concorrente (b) (anni 1984 e 1985)* (*)

Serie	AUT		OWN		MSX	
	(a)	(b)	(a)	(b)	(a)	(b)
Maschi						
DM	1,28	0,50	1,21	0,66	3,68	1,90
PM	0,69	0,28	0,41	0,23	0,62	0,30
AM	1,26	0,59	1,63	0,81	1,34	0,79
TOTM	0,77	0,52	0,65	0,36	0,68	0,38
OM	0,08	0,04	0,08	0,04	0,02	0,01
FLM	0,05	0,02	0,07	0,06	0,03	0,06
TDM	0,82	0,46	0,61	0,27	0,08	0,40
Femmine						
DF	0,99	0,73	1,47	0,74	0,43	0,27
PF	0,53	0,43	0,80	0,49	1,75	1,47
AF	1,33	0,55	0,86	0,64	1,16	1,28
TOTF	0,78	0,26	0,76	0,26	1,07	0,35
OF	0,13	0,09	0,10	0,07	0,03	0,02
FLF	0,19	0,10	0,20	0,09	0,24	0,09
TDF	0,76	0,23	0,78	0,24	0,88	0,40

(*) Rapporti rispetto alle serie destagionalizzate calcolate alla fine dell'anno.

Tab. 8: *Medie assolute percentuali delle revisioni (periodo 1983.I-1985.I)*

Serie	AUT					OWN				
	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
Maschi										
DM	0,85	0,44	0,64	0,94	2,01	1,24	0,33	0,79	0,57	2,08
PM	0,43	0,49	0,55	0,49	1,45	0,42	0,39	0,47	0,27	0,88
AM	1,14	0,52	1,12	0,92	1,31	1,73	0,93	1,62	1,58	2,51
TOTM	0,61	0,14	0,51	0,44	1,18	0,48	0,16	0,43	0,43	0,86
OM	0,06	0,04	0,06	0,04	0,07	0,05	0,04	0,06	0,03	0,07
FLM	0,04	0,02	0,06	0,05	0,09	0,10	0,05	0,06	0,07	0,13
TDM	0,54	0,28	0,37	0,37	1,16	0,46	0,18	0,32	0,28	0,84
Femmine										
DF	0,65	0,48	1,00	1,00	1,94	0,64	0,38	0,57	0,92	1,65
PF	0,40	0,25	0,52	0,41	0,66	0,42	0,21	0,55	0,45	0,83
AF	1,18	0,60	1,02	1,39	2,00	1,23	0,38	0,73	1,06	1,59
TOTF	0,57	0,39	0,38	0,40	0,72	0,50	0,46	0,26	0,33	0,76
OF	0,06	0,04	0,10	0,15	0,21	0,05	0,03	0,10	0,15	0,20
FLF	0,10	0,06	0,13	0,21	0,22	0,10	0,07	0,14	0,22	0,21
TDF	0,42	0,32	0,30	0,33	0,59	0,45	0,28	0,31	0,28	0,66

(a) Confronto della stima concorrente con la stima un trimestre dopo.

(b) Confronto della stima un trimestre dopo con la stima due trimestri dopo.

(c) Confronto della stima due trimestri dopo con la stima un anno dopo.

(d) Confronto della stima un anno dopo con la stima due anni dopo.

(e) Confronto della stima concorrente con la stima due anni dopo.

Segue Tab. 8: *Medie assolute percentuali delle revisioni (periodo 1983.I-1985.I)*

Serie	MSX				
	(a)	(b)	(c)	(d)	(e)
Maschi					
DM	4,17	2,98	3,66	2,58	3,84
PM	1,99	1,82	0,93	0,74	1,44
AM	1,55	0,53	0,91	1,18	2,22
TOTM	0,57	0,25	0,44	0,23	0,77
OM	0,02	0,03	0,02	0,01	0,02
FLM	0,15	0,17	0,10	0,05	0,15
TDM	0,59	0,27	0,45	0,24	0,92
Femmine					
DF	0,59	0,23	0,43	0,49	0,77
PF	0,79	0,72	0,77	0,68	1,38
AF	1,78	1,60	1,60	1,18	2,21
TOTF	0,73	0,45	0,37	0,39	0,89
OF	0,03	0,01	0,02	0,04	0,07
FLF	0,11	0,10	0,12	0,20	0,19
TDF	0,78	0,52	0,60	0,69	0,95

La Tab. 8, dove sono riportate le medie assolute percentuali delle revisioni, suggerisce le seguenti considerazioni.

(a) Le due versioni dell'X-11-ARIMA sono quelle che rivedono meno nel corso del primo anno. Durante il secondo anno le revisioni apportate dalla versione automatica dell'X-11-ARIMA sono però le più grandi. Migliori sono invece le prestazioni della versione basata su un modello specificato dall'utente.

(b) MSX rivede nel primo anno più delle due versioni dell'X-11-ARIMA, mentre nel secondo anno le revisioni sono confrontabili con quelle della versione OWN dell'X-11-ARIMA.

(c) Per quanto riguarda il confronto tra la stima concorrente e la stima due anni dopo, i metodi che apportano le revisioni minori sono le due versioni dell'X-11-ARIMA.

4.3.3. *Velocità di convergenza delle stime al valore definitivo*

Un'ulteriore importante questione riguarda la lunghezza dell'intervallo temporale in cui le stime sono da considerarsi ancora provvisorie. Per analizzare questa questione abbiamo confrontato le stime della serie aggiustata relative agli anni 1980 e '81 ottenute rispettivamente alla fine degli anni 1984, '85 e '86. La Tab. 9 riporta le medie assolute percentuali delle revisioni.

Le evidenze salienti sono sensibilmente differenti per le due procedure di destagionalizzazione adottate. Le stime prodotte utilizzando le due versioni

Tab. 9: *Medie assolute percentuali delle revisioni: (a) stime tre anni dopo contro stime quattro anni dopo: (b) stime tre anni dopo contro stime cinque anni dopo: (c) stime quattro anni dopo contro stime cinque anni dopo (periodo 1980.I-1981.IV)*

Serie	AUT			OWN			MSX		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
Maschi									
DM	0,077	0,102	0,141	0,057	0,197	0,161	7,397	7,645	0,390
PM	0,197	0,197	0,007	0,123	0,124	0,004	0,776	0,735	0,132
AM	0,409	0,386	0,046	0,157	0,088	0,140	0,457	0,799	1,102
TOTM	0,188	0,141	0,054	0,185	0,068	0,117	0,055	0,061	0,010
OM	0,008	0,007	0,004	0,008	0,007	0,004	0,015	0,014	0,004
FLM	0,006	0,011	0,005	0,009	0,020	0,013	0,014	0,019	0,007
TDM	0,026	0,026	0,000	0,021	0,021	0,000	0,072	0,090	0,026
Femmine									
DF	0,057	0,083	0,044	0,298	0,195	0,107	1,018	1,270	0,385
PF	0,033	0,032	0,051	0,040	0,029	0,047	0,118	0,081	0,043
AF	0,115	0,271	0,162	0,107	0,224	0,173	0,683	0,268	0,612
TOTF	0,142	0,087	0,059	0,109	0,075	0,047	0,142	0,109	0,061
OF	0,060	0,057	0,011	0,050	0,048	0,011	0,055	0,084	0,064
FLF	0,036	0,030	0,027	0,040	0,031	0,028	0,097	0,082	0,070
TDF	0,255	0,202	0,053	0,247	0,202	0,045	0,190	0,085	0,176

dell'X-11-ARIMA, infatti, sono sì soggette a delle revisioni anche dopo tre anni, ma la dimensione delle correzioni apportate è quasi sempre irrilevante, specialmente per la versione OWN.

Per quanto riguarda MSX, invece, i risultati indicano la possibilità che anche dopo tre anni alcune serie siano soggette a revisioni importanti. Nel nostro caso, per esempio, questo accade per i disoccupati in senso stretto ed è particolarmente rilevante per la serie dei maschi. Alla base delle revisioni apportate c'è una certa instabilità delle stime dei parametri del modello utilizzato per interpretare la serie, quindi una conseguente instabilità dei filtri usati per estrarre le componenti.

5. *Destagionalizzazione di serie aggregate*

Quando si affronta il tema della destagionalizzazione di una serie aggregata, tradizionalmente vengono seguiti due approcci. Uno, il metodo diretto, consiste nell'applicare la procedura di aggiustamento stagionale ai dati grezzi della serie aggregata. L'altro, il metodo indiretto, consiste nel destagionalizzare ciascuna componente e nel formare quindi l'aggregato destagiona-

lizzato aggregando semplicemente le serie componenti destagionalizzate[5] .

Generalmente i due metodi di aggiustamento portano a risultati differenti: ciò è dovuto essenzialmente alla non linearità della procedura di aggiustamento stagionale e/o alla non linearità dell'operazione di aggregazione delle componenti. Infatti, il metodo diretto fornisce gli stessi risultati di quello indiretto solo quando l'aggregato è una somma algebrica di componenti, per ciascuna componente è adottato un modello additivo e, infine, la procedura di aggiustamento è lineare[6]. Questo caso è tuttavia molto restrittivo, dato che molti dei più importanti indicatori economici sono sotto la forma di tasso e che le procedure di aggiustamento stagionale di maggior impiego, soprattutto presso le agenzie statistiche ufficiali, sono di natura non lineare.

Poichè un aggregato può essere classificato in vari modi, col metodo indiretto sussiste inoltre un'ovvia incertezza circa la classificazione da scegliere per l'aggiustamento stagionale dell'aggregato. Per esempio, il tasso di disoccupazione rimane per costruzione lo stesso anche a seguito di diverse classificazioni delle varie componenti dell'occupazione e della disoccupazione, ma il tasso di disoccupazione destagionalizzato col metodo indiretto può cambiare a seconda della classificazione adottata per le componenti. Tali differenze possono risultare più o meno significative e in alcuni casi possono portare anche a erronee conclusioni sull'evoluzione del tasso.

Il problema di trovare la maniera più conveniente per destagionalizzare alcuni degli aggregati finora considerati è dunque non banale. Abbiamo perciò condotto varie analisi empiriche su tali aggregati, non solo per scegliere tra aggiustamento diretto e indiretto, ma anche per identificare possibili combinazioni di componenti che portino alla 'migliore' destagionalizzazione indiretta. Un riassunto del quadro applicativo in cui si collocano tali analisi è nella Tab. 10, dove sono elencati gli aggregati presi in considerazione per l'aggiustamento con i due metodi.

Accanto agli aggregati destagionalizzati direttamente (prima colonna), compaiono le varie combinazioni dell'aggregato destagionalizzate col metodo indiretto (seconda colonna). Come si nota, per ogni aggregato, diciamo di ordine superiore, abbiamo considerato per l'aggiustamento indiretto combinazioni di componenti che interessano sia la massima disaggregazione possibile, compatibilmente con l'insieme delle serie di partenza, sia aggregazioni intermedie, diciamo di ordine inferiore. Ad esempio, alla serie della forza di lavoro maschile (FLM) aggiustata col metodo diretto, abbiamo fatto corrispondere due combinazioni dell'aggregato candidabili per l'aggiustamento indiretto, corrispondenti rispettivamente alla somma di tutte le com-

---

5 Nell'ambito delle procedure di destagionalizzazione cosiddette ottimali, una soluzione efficiente al problema della destagionalizzazione di un aggregato può essere trovata, sotto certe condizioni, purchè si sfrutti adeguatamente l'informazione relativa alla struttura di correlazione di tutte le serie componenti l'aggregato stesso (Geweke,1982). Per una discussione circa le caratteristiche di questa procedura e la sua applicabilità pratica, vedi Bordignon e Masarotto (1988).

6 Alcune verifiche empiriche si possono trovare in Lothian e Morry (1977) per la procedura X-11 e in Bordignon e Masarotto (1988) per l'X-11-ARIMA.

Tab. 10: *Schema riassuntivo degli aggregati destagionalizzati col metodo diretto e indiretto* [(a)]

Aggregato (Metodo diretto)	Composizioni dell'aggregato (Metodo indiretto)
OM	OAM+OIM+OAAM
TOTM	DM+PM+AM
FLM	1. OAM+ OIM+OAAM+DM+PM+AM 2. OM+TOTM
TDM	1. $\frac{DM+PM+AM}{OAM+OIM+OAAM+DM+PM+AM}$ 2. $\frac{TOTM}{OM+TOTM}$ 3. $\frac{TOTM}{FLM}$ 4. $\frac{FLM-OM}{FLM}$
TD	1. $\frac{DM+PM+AM+DF+PF+AF}{OAM+OIM+OAAM+DM+PM+AM+OAF+OIF+OAAF+DF+PF+AF}$ 2. $\frac{TOTM+TOTF}{DM+TOTM+OF+TOTF}$ 3. $\frac{TOTM+TOTF}{FLM+FLF}$ 4. $\frac{(FLM+FLF)-(OM+OF)}{FLM+FLF}$

(a) Lo stesso schema, riportato per gli aggregati maschili, vale anche per quelli femminili.

ponenti distinte dell'occupazione e della disoccupazione o semplicemente alla somma per l'aggiustamento indiretto di questi due totali. Per quanto riguarda i tassi di disoccupazione, ci è sembrato interessante prendere in considerazione per l'aggiustamento indiretto un'ulteriore alternativa, proposta originariamente da Brittain (1959) in contrapposizione alla procedura ufficiale adottata dal U.S. Bureau of Labor Statistics. Tale alternativa, nota anche *come metodo* residuale, consiste nel destagionalizzare la forza di lavoro e l'occupazione separatamente, ottenere quindi la disoccupazione destagionalizzata per differenza e rapportare quest'ultima alla forza di lavoro destagionalizzata.

Alcuni dei risultati relativi al confronto tra aggiustamento diretto ed indiretto sono riassunti nelle Tabb. 11 e 12. Le procedure di destagionalizzazione utilizzate sono, al solito, le due versioni dell'X-11-ARIMA (AUT e OWN) e MSX. Come criteri per il confronto sono state utilizzate le seguenti quantità:

(a) una versione relativa della misura del grado di lisciamento di una serie proposta da Dagum (1979), basata su stime della componente trend-ciclo (Y) tramite medie mobili di Henderson:

$$H_0 = \frac{1}{n} \sum_t \left(\frac{Y_t - HY_t}{Y_t} \right)^2 , \qquad (7)$$

dove H indica il filtro di Henderson con un numero di termini che dipende dalla cadenza temporale della serie;

(b) un indice della stabilità della stima della serie destagionalizzata, basato sulle revisioni $N_t\ (r)$-$N_t\ (l)$, dato da

$$R = \frac{100}{8n} \sum_t \sum_{l=t+1}^{t+8} \left| \frac{N_t\ (l) - N_t\ (l-1)}{N_t\ (l-1)} \right| , \qquad (8)$$

dove il numero 8 indica che R misura la media delle revisioni successive apportate nei primi due anni di una serie trimestrale;

(c) un indice della dissomiglianza tra serie destagionalizzate secondo i due approcci - diretto e indiretto - e secondo vari livelli di disaggregazione delle componenti, data da

$$D = \frac{100}{n} \sum \left| \frac{N\ (a)_t - N\ (b)_t}{\frac{1}{2} (N\ (a)_t - N\ (b)_t)} \right| . \qquad (9)$$

I risultati relativi al confronto tra destagionalizzazione diretta e indiretta non indicano la chiara prevalenza di un metodo rispetto all'altro. Piuttosto, essi sembrano dipendere dalle caratteristiche dell'aggregato e dalle sue componenti, dalla procedura di destagionalizzazione adottata e dal criterio di confronto utilizzato. Ad esempio, se si considera l'aggregato TOTM, secondo il criterio basato sulle revisioni l'aggiustamento indiretto risulta leggermente preferibile quando si adotta l'X-11-ARIMA, mentre è nettamente peggiore di quello diretto nel caso della procedura MSX. Secondo il criterio basato sul lisciamento, invece, si ha una lieve prevalenza dell'aggiustamento diretto per entrambe le procedure di destagionalizzazione.

Come si è appena rilevato, non sempre le indicazioni fornite dagli indici di lisciamento sono in accordo con quelle suggerite dagli indici basati sulle revisioni. Ciò è spiegabile se si ricordano alcuni risultati teorici di Maravall (1984), secondo i quali una procedura di decomposizione che cerca di ottimizzare il grado di lisciamento del trend-ciclo non necessariamente porta a minimizzare l'errore quadratico medio delle revisioni. Pertanto, quando i due indici sono discordi occorre una certa cautela nell'interpretazione dei risultati, contrariamente alle indicazioni fornite da Lothian e Morry (1977), i quali impropriamente suggeriscono di utilizzare gli indici di lisciamento come sostituti, e non complementi, degli indici di stabilità.

Tab. 11: *Confronto fra metodo diretto e metodo indiretto di aggiustamento stagionale, per il totale delle persone in cerca, gli occupati e le forze di lavoro*

Serie e metodo	H_0			R			D		
	AUT	OWN	MSX	AUT	OWN	MSX	AUT	OWN	MSX
OM:									
Diretto	0,190	0,190	0,210	0,036	0,033	0,015			
Indir.	0,194	0,194	0,193	0,040	0,032	0,031	0,069	0,069	0,103
TOTM:									
Diretto	1,643	1,641	1,386	0,300	0,279	0,248			
Indir.	1,746	1,727	1,679	0,276	0,216	1,017	0,649	0,669	0,891
FLM:									
Diretto	0,179	0,179	0,168	0,032	0,055	0,075			
Indir.1	0,199	0,197	0,192	0,046	0,031	0,076	0,062	0,061	0,047
Indir.2	0,172	0,172	0,195	0,039	0,038	0,020	0,039	0,039	0,109
OF:									
Diretto	0,520	0,520	0,594	0,068	0,060	0,017			
Indir.	0,516	0,516	0,550	0,077	0,086	0,070	0,095	0,097	0,279
TOTF:									
Diretto	2,234	2,232	2,230	0,313	0,301	0,321			
Indir.	2,288	2,283	2,176	0,309	0,241	0,501	0,568	0,582	0,563
FLF:									
Diretto	0,554	0,554	0,568	0,077	0,080	0,092			
Indir.1	0,539	0,537	0,546	0,099	0,101	0,127	0,093	0,103	0,120
Indir.2	0,542	0,543	0,588	0,087	0,084	0,065	0,123	0,129	0,294

In generale, per quanto riguarda la procedura X-11-ARIMA, in entrambe le sue varianti (AUT e OWN), le differenze tra aggiustamento diretto ed indiretto sono relativamente contenute. Sia l'indice di dissomiglianza che gli altri indici suggeriscono che è irrilevante, da un punto di vista pratico, destagionalizzare direttamente o indirettamente: di conseguenza, la scelta tra i due tipi di aggiustamento può essere fatta sulla base di altre motivazioni dipendenti dal contesto applicativo e dalle finalità dell'utilizzatore. (E' chiaro che ciò vale per gli aggregati e le corrispondenti combinazioni di componenti considerate in questa analisi empirica, mentre rimangono da esplorare altre disaggregazioni, ad esempio per età e per sesso.) Un'eccezione a questo comportamento generale è rappresentata dall'aggiustamento indiretto dei tassi di disoccupazione mediante il metodo residuale di Brittain (1959), che, specie per i maschi e il totale, fornisce risultati decisamente peggiori.

Lievemente diverso, infine, è il discorso per la procedura MSX. In tal caso, almeno per alcuni aggregati, ad esempio TOTM, le differenze tra destagionalizzazione diretta e indiretta paiono più rilevanti, specie secondo il criterio della stabilità.

Tab. 12: *Confronto fra metodo diretto e metodo indiretto di aggiustamento stagionale dei tassi di disoccupazione*

Serie e metodo	H_0			R			D		
	AUT	OWN	MSX	AUT	OWN	MSX	AUT	OWN	MSX
TDM:									
Diretto	1,564	1,566	1,390	0,296	0,248	0,255			
Indir.1	1,694	1,676	1,644	0,251	0,211	0,928	0,561	0,589	0,820
Indir.2	1,646	1,647	1,378	0,290	0,253	0,224	0,339	0,331	0,145
Indir.3	1,626	1,624	1,387	0,299	0,258	0,243	0,343	0,331	0,085
Indir.4	2,130	2,132	2,333	0,400	0,637	1,116	0,921	0,913	2,407
TDF:									
Diretto	1,982	1,981	1,808	0,236	0,235	0,502			
Indir.1	2,046	2,044	1,908	0,236	0,181	0,407	0,321	0,335	0,529
Indir.2	1,987	1,985	2,022	0,255	0,258	0,265	0,495	0,493	0,659
Indir.3	1,989	1,991	2,019	0,269	0,275	0,266	0,527	0,518	0,595
Indir.4	2,069	2,077	2,446	0,332	0,278	0,386	0,827	0,856	2,327
TD:									
Diretto	1,584	1,584	1,470	0,260	0,237	0,258			
Indir.1	1,656	1,644	1,585	0,186	0,163	0,548	0,314	0,268	0,556
Indir.2	1,592	1,586	1,540	0,233	0,240	0,231	0,348	0,368	0,395
Indir.3	1,566	1,568	1,540	0,239	0,231	0,219	0,373	0,405	0,423
Indir.4	1,799	1,805	1,930	0,332	0,305	0,594	0,647	0,647	0,817

6. *Conclusioni*

In questo capitolo sono state valutate alcune procedure di destagionalizzazione su un insieme di serie storiche del mercato del lavoro. Le procedure messe a confronto sono le due varianti dell'X-11-ARIMA (rispettivamente AUT e OWN) e MSX di Burman (1980). I risultati del confronto portano alle seguenti considerazioni conclusive.

(a) La procedura X-11-ARIMA conferma le sue buone proprietà. In particolare, è rimarchevole la capacità di questa procedura, essenzialmente empirica e non basata su particolari proprietà di ottimalità, di produrre risultati mai troppo peggiori della migliore delle alternative. Merita inoltre di essere sottolineato come le *performances* della variante AUT siano inferiori a quelle della versione OWN solamente per quanto riguarda la stabilità delle stime e come, anche con riferimento a questo aspetto, le differenze non siano grandi. Ciò significa che i modelli ARIMA incorporati nella versione automatica dell'X-11-ARIMA risultano anche per la maggior parte delle serie del mercato del lavoro sufficientemente adeguati al compito loro assegnato, che è quello di estendere la serie originale con un insieme di previsioni.

(b) Sostanzialmente buone si sono rivelate anche le proprietà della procedura MSX. Tuttavia, la centralità della fase di identificazione di un ap-

propriato modello ARIMA fa sì che MSX sia la procedura che richiede di più all'utilizzatore sia in termini di tempo che di preparazione. Inoltre il reperimento di un modello adatto a rappresentare una particolare serie è a volte complicato dal fatto che la ricerca deve avvenire all'interno della classe dei modelli decomponibili, che non coincide con l'insieme di tutti i modelli ARIMA. Queste difficoltà suggeriscono di limitare l'utilizzo della versione attuale di MSX solamente alle serie che sembrano decomposte in maniera inadeguata da altri metodi, e nello stesso tempo di esplorare la possibilità di utilizzare per l'identificazione di un modello delle procedure automatiche. Si osservi inoltre come il punto debole di MSX, messo bene in luce dai risultati presentati, sia la stabilità delle stime. Il risultato era parzialmente atteso. Infatti, come mostrato da Maravall (1984), la scelta di fondo di MSX di massimizzare la varianza della componente irregolare da un lato permette di ottenere stime del trend e della serie destagionalizzata più lisce, ma dall'altro massimizza la varianza delle revisioni. Le analisi presentate suggeriscono quindi di esplorare la possibilità di ottenere delle procedure *model-based* che realizzino un migliore compromesso tra questi due aspetti (vedi Piccolo, 1985, e i riferimenti ivi contenuti per indicazioni su questo punto).

(c) Per quanto riguarda il problema dell'aggiustamento stagionale dei dati in corso d'anno, le analisi empiriche effettuate indicano chiaramente che, indipendentemente dalla procedura di destagionalizzazione adottata, è preferibile l'approccio concorrente, basato cioè su fattori stagionali stimati ogni volta che arriva una nuova osservazione.

(d) Guardando infine al problema dell'aggiustamento stagionale di serie aggregate, che risultano cioè dalla combinazione (tipicamente per somma e/o rapporto) di più serie componenti, le analisi empiriche da noi condotte portano ad una sostanziale equivalenza tra i due metodi, quello diretto e quello indiretto, usualmente impiegati.

17

ANALISI MULTIVARIATE DINAMICHE DI SERIE STORICHE RELATIVE AL MERCATO DEL LAVORO

Giuliana Passamani e Marina Schenkel *

1. *Premessa*

L'analisi multivariata di alcune serie aggregate delle forze di lavoro si giustifica non soltanto per l'interesse intrinseco, ma anche per la possibilità di proseguire nella linea indicata dalle applicazioni dell'approccio VAR (*Vector Auto-Regression*) al mercato del lavoro. A partire dall'articolo originario di Ashenfelter e Card (1982), vari autori (per l'Italia, vedi Del Boca, 1987) si sono proposti di confrontare i modelli empirici derivati dall'analisi dei dati (i ' fatti '), con quelli dedotti dalle varie ipotesi interpretative (le ' teorie '). Tratto comune di questi studi è la considerazione di una sola, o al massimo di due serie relative al mercato del lavoro: usualmente si tratta dell'occupazione totale, oppure del tasso di disoccupazione.

La semplice ispezione grafica di tali serie distinte per sesso, e ulteriormente per settore per quanto riguarda gli occupati, e per condizione per quanto riguarda i disoccupati, solleva però qualche dubbio sulla legittimità di considerare in maniera indifferenziata il mercato del lavoro.

D'altronde,benchè l'esistenza di più mercati sia data per scontata in letteratura, altrettanto scontata è la difficoltà di trattare i problemi teorici e empirici che da questa ipotesi derivano.

Le analisi da noi svolte non hanno avuto ambizioni in questo senso; piuttosto, mirano ad indicare l'opportunità di considerare distintamente i principali subaggregati che concorrono a determinare l'occupazione e la disoccupazione totale. A tal fine si sono esaminate alcune serie aggregate tratte dall'indagine trimestrale delle forze di lavoro, sia nel loro comportamento univariato, sia ponendole in relazione con il salario orario nell'industria, l'indice dei prezzi al consumo e il tasso di interesse[1].

---

* Il capitolo è frutto della collaborazione degli autori. In particolare, G. Passamani ha curato la stesura delle sezz. 2 e 3, mentre M. Schenkel quella delle sezz. 1 e 4.

1 Le serie della rilevazione trimestrale delle forze di lavoro sono state tratte da varie pubblicazioni dell'Istat. Fino al 1977 sono state ricostruite da Sanetti e Settanni (1979) tenendo conto delle revisioni delle definizioni di occupato, disoccupato, ecc. introdotte nel 1979 con il cambiamento del questionario. Ricordiamo che nel 1984 è stato cambiato il modello di rilevazione ed è stato rivisto il riporto all'universo per adeguarlo ai risultati del Censimento della Popolazione del 1981, e che nel 1986. Il è cambiata la definizione di persona

I risultati ottenuti permettono di concludere che occupazione e disoccupazione non sono da considerarsi equivalenti, o meglio interscambiabili, come risulterebbe invece dai due lavori appena citati. I loro comportamenti divergono in maniera tale da doverli attribuire all'esplicarsi di fenomeni diversi. Ciò vale in particolare per l'occupazione alle dipendenze nell'industria e per le diverse condizioni di disoccupazione.

In questo studio le serie della disoccupazione vengono analizzate in dettaglio, mentre verranno considerate solo brevemente le serie dell'occupazione, già ampiamente analizzate in un precedente lavoro (Passamani e Schenkel, 1988), al quale rinviamo . Per quanto riguarda l'occupazione, ricordiamo soltanto che mentre fra le variabili di prezzo - salari, prezzi, tasso di interesse - sono apparse relazioni ben riconoscibili e stabili, i legami fra prezzi e quantità differiscono a seconda delle serie: l'occupazione femminile risente dell'azione delle variabili di prezzo, ed in particolare del saggio di interesse; viceversa, è l'occupazione maschile a esercitare un'azione sul tasso di interesse. In generale, peraltro, risulta che tutte le serie relative all'occupazione dipendono significativamente dal loro passato.

Lo scopo di questo capitolo è, in sostanza, la presentazione di alcuni risultati relativi alle serie riguardanti le persone in cerca di occupazione: senza proporci la confutazione o la dimostrazione di alcuna teoria, pensiamo di poter fornire qualche contributo al dibattito sulle cause e i rimedi della disoccupazione.

Riteniamo infatti che alcuni fenomeni emblematici di questi ultimi anni — la disoccupazione giovanile, la crisi dell'occupazione industriale, la crescita dell'offerta di lavoro femminile — indichino la necessità di una visione complessa, e non parziale, del fenomeno.

2. *Analisi descrittiva delle serie*

L'analisi empirica delle serie univariate, i cui risultati sono riportati in Passamani e Schenkel (1988), introduceva per le stesse serie rese stazionarie una rappresentazione autoregressiva troncata ad un certo ordine, la quale consentiva di esplicitare le caratteristiche dinamiche delle singole variabili mediante l'analisi della loro dipendenza temporale.

Ora estendiamo e completiamo l'analisi empirico-descrittiva delle singole serie studiandone il comportamento temporale anche mediante un modello

---

[segue nota]

in cerca di occupazione. Nel corso dell'analisi si è tenuto conto del fatto che alcuni valori anomali presenti nelle serie derivano da questi cambiamenti, e si è quindi proceduto a un adeguato aggiustamento. Le serie del salario orario nell'industria in senso stretto (stabilimenti con un numero di addetti superiore ai 50; fonte: Ministero del Lavoro), dell'indice dei prezzi al consumo base 1970 (fonte: ISTAT) e del tasso di interesse sui BOT a tre mesi (media mensile ponderata delle aste quindicinali nei mesi di marzo, giugno, settembre, dicembre; fonte: Banca d'Italia) ci sono state fornite da Prometeia, a meno della serie del salario dal 1986.I, che abbiamo ricostruito. Tutte le serie utilizzate sono trimestrali e non destagionalizzate. Il periodo considerato va dal primo trimestre 1970 al quarto trimestre 1988 per le analisi descrittive, e dalla stessa data al quarto trimestre 1986 per le analisi causali.

strutturale, in cui si distinguono e si stimano le componenti non direttamente osservabili di una serie, quali il trend, la stagionalità e l'errore.

In questa sezione ci limitiamo a presentare i risultati delle stime della componente di trend di alcune delle serie di interesse. Questi risultati torneranno poi utili per capire il diverso comportamento delle serie storiche che formano l'aggregato delle persone in cerca di occupazione, quando poste in relazione con le serie di prezzo scelte per le analisi multivariate. Si omette volutamente la presentazione delle stime relative alle serie dell'occupazione in quanto l'importanza di queste ultime nelle analisi a più variabili è apparsa, *a posteriori*, limitata.

In accordo con la letteratura[2], definiamo modello strutturale il modello utilizzato per descrivere le componenti delle serie storiche,in particolare la componente di trend, che ci permette di cogliere l'evoluzione delle serie al di là dell'effetto della componente stagionale. Il modello strutturale di base consiste di una componente di trend e di una componente stagionale che sono supposte variare lentamente nel tempo, e di una componente d'errore che varia casualmente. In termini formali (Harvey e Todd, 1983, p. 300):

$$Y_t = \mu_t + \gamma_t + \varepsilon_t \ , \qquad t = 1,2,...,n, \tag{1}$$

dove y_t è la serie osservata e μ_t , γ_t ed ε_t , sono le componenti di trend, stagionale ed erratica. Rispetto al modello (1) si assume un trend di tipo locale, cioè un trend i cui parametri di livello e di pendenza sono stimati attribuendo relativamente più peso alle osservazioni più recenti. Il processo che genera il trend sarà perciò della forma:

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_\tau \ , \tag{2}$$

$$\beta_t = \beta_{t-1} + \xi_t \ , \tag{3}$$

dove η_τ e ξ_t sono sequenze casuali, incorrelate nel tempo ed indipendenti fra di loro, normalmente distribuite, con media nulla e varianze τ_η^2 e τ_ξ^2 rispettivamente. Sempre rispetto al modello (1) si assume che la componente stagionale vari nel tempo, sebbene sia soggetta al vincolo che gli effetti stagionali si annullino nell'arco del periodo della stessa componente. Il processo che genera γ_t sarà perciò della forma:

$$\gamma_t = -\sum_{j=1}^{(s-1)} \gamma_{t-j} + \omega_t \ , \tag{4}$$

dove $\omega_t \sim$ NID $(0,\sigma_\omega^2)$ ed s è il numero delle stagioni nell'arco del periodo.

---

2 Ci limitiamo qui a citare i lavori di Harvey e Todd (1983) e di Harvey (1984 1985).

Per la stima dei parametri del modello (1), (2) e (3) usiamo il metodo della massima verosimiglianza esatta[3].

Nelle Figg. da 1 a 9[4] è presentata la stima *smoothed* della componente di trend delle serie storiche di maggior interesse rispetto allo scopo del lavoro.

La serie del salario orario nominale presenta una componente di trend con una pendenza più accentuata fra la fine degli anni '70 e l'inizio degli anni '80, mentre la serie del salario orario reale presenta una forte crescita fino all'inizio degli anni' 80 e quindi una certa stabilità almeno fino al 1986.

Le serie del tasso di disoccupazione totale, maschile e femminile mostrano un'evoluzione abbastanza simile, ma la terza presenta valori molto più elevati. Sostanziali differenze si colgono dall'osservazione degli aggregati che formano la disoccupazione: in particolare, la disoccupazione in senso stretto e le altre persone in cerca di occupazione. Per quanto riguarda la disoccupazione in senso stretto, si nota che l'evoluzione relativa alla popolazione maschile è stata diversa dall'evoluzione relativa alla popolazione femminile nel corso degli anni '70, mentre con gli anni '80 si riscontra una maggiore similarità evolutiva. Per quanto riguarda le altre persone in cerca di occupazione, l'evoluzione relativa alla popolazione femminile si presenta sempre sensibilmente diversa e con valori di gran lunga più elevati rispetto a quella maschile.

Si evidenzia inoltre che, sebbene si sia in presenza di una certa stabilità del salario reale nel corso degli anni '80[5] gli aggregati che determinano la disoccupazione, in particolare la disoccupazione in senso stretto e la ricerca di prima occupazione, subiscono evidenti incrementi.

3. *Analisi causale delle serie: alcuni risultati*

L'analisi empirica delle serie multivariate è stata condotta, lo ripetiamo, mediante un approccio vettoriale autoregressivo[6], in cui si sono sempre considerate congiuntamente quattro serie, opportunamente trasformate per renderle stazionarie, la prima delle quali, relativa alle forze di lavoro, viene di volta in volta sostituita, mentre le altre tre, cioè il salario nominale, l'indice dei prezzi al consumo ed il tasso di interesse sui BOT vengono mantenute costantemente presenti.

Nel seguito richiamiamo brevemente i principali risultati delle analisi VAR.

Presentiamo e commentiamo, invece, con qualche dettaglio i risultati

---

3 I risultati della stima sono stati ottenuti utilizzando il programma STAMP (*Structural Time Analyser Modeller Predictor*) distribuito da ESRC, Centre for Economic Computing, London School of Economics and Political Science, Londra.

4 Le figure e le tabelle sono riportate nell'Appendice al presente capitolo.

5 Fra i lavori più noti riguardanti la relazione fra costo del lavoro, salario e disoccupazione in Italia, ricordiamo che Modigliani, Padoa Schioppa e Rossi (1986) non analizzano gli anni più recenti.

6 Per una presentazione semplificata dell'approccio rinviamo a Passamani e Schenkel (1988) ed, eventualmente, ai lavori ivi citati.

delle simulazioni dinamiche degli stessi modelli VAR relativamente alle sole variabili legate alla disoccupazione.

Dall'insieme delle analisi autoregressive si rileva che le serie delle forze di lavoro risultano spiegate soprattutto dal loro passato. Fanno eccezione due serie relative alla forza lavoro femminile: una è l'occupazione dipendente nell'industria, nei confronti della quale si manifesta una causalità diretta del tasso di interesse; l'altra è la disoccupazione in senso stretto femminile, nei confronti della quale vi è una causalità diretta sia del salario nominale che dei prezzi. E interessante riscontrare il verificarsi di questa relazione solo per la popolazione femminile, quasi che per la popolazione maschile non si possa parlare di relazione fra variazione del salario e variazione della disoccupazione.

Ritornando ai risultati, la serie dei prezzi risulta sempre direttamente causale nei confronti del salario nominale. Sulla serie dei prezzi, oltre ai prezzi stessi, risulta spesso causale il tasso di interesse, e questo anche istantaneamente[7]. La serie del tasso di interesse risulta determinata da se stessa, dall'occupazione maschile e, di conseguenza, dall'occupazione totale.

Sia la disoccupazione in senso stretto maschile che la ricerca di prima occupazione maschile esercitano un'effetto sui prezzi. Anche il tasso di disoccupazione maschile mostra il suo effetto sui prezzi, ma la sua significatività è solo del 10%.

Per quei modelli vettoriali autoregressivi in cui si sono riscontrate correlazioni contemporanee significativamente diverse da zero, si è approfondita l'analisi del comportamento dinamico delle serie considerate, al fine di capire come rispondono le diverse variabili analizzate congiuntamente ad una eventuale perturbazione in una di esse. Nella sostanza, si è trattato di simulare delle perturbazioni e di vedere come risponde il modello. A questo scopo, abbiamo derivato la rappresentazione a media mobile, troncata ad un certo ordine, per tutti quei modelli autoregressivi la cui matrice delle correlazioni dei residui mostrava almeno una correlazione contemporanea significativamente diversa da zero. Questi sono i modelli contenenti le variabili disoccupazione in senso stretto sia maschile che femminile, ricerca di prima occupazione maschile, tasso di disoccupazione maschile, femminile e totale. La visualizzazione grafica delle risposte delle diverse variabili a perturbazioni in una di esse[8] ha messo in evidenza comportamenti di risposta molto simili se si tiene fermo il sesso e comportamenti di risposta molto diversi al variare del sesso. In altre parole, non si rilevano differenze importanti nelle risposte dei maschi o delle femmine al variare dei modelli, ma si rilevano differenze interessanti fra gli stessi modelli riferiti ai maschi oppure alle femmine.

---

7 L'esistenza di causalità istantanea si ha quando si riscontrano correlazioni significativamente diverse fra le serie dei residui, o innovazioni, delle diverse equazioni che compongono il modello multivariato.

8 In ogni grafico sono rappresentate le risposte delle diverse variabili ad una perturbazione unitaria, al tempo zero, in una determinata innovazione ortogonalizzata del processo AR(p) multivariato. Le risposte relative a ciascuna variabile sono state divise per la radice della loro varianza di innovazione, in modo da esprimerle tutte in termini di frazioni di deviazioni standard. Le risposte delle diverse variabili sono quindi confrontabili fra di loro.

Tenendo presente che le variabili sono espresse in termini di variazioni per ragioni di stazionarietà - più precisamente le serie analizzate sono differenze prime di differenze di periodo quattro - dall'osservazione delle figure emerge un insieme di interessanti riscontri. Ci limitiamo ad illustrare le risposte del sistema a perturbazioni nelle variabili tasso di disoccupazione maschile e femminile e nella variabile salario orario nominale. Innanzitutto, come si può immediatamente cogliere dall'esame della fig. 10, una perturbazione nella variabile tasso di disoccupazione maschile induce delle oscillazioni nell'indice dei prezzi che si smorzano solo lentamente. Una perturbazione nella variabile tasso di disoccupazione femminile induce, invece, dei movimenti abbastanza ampi nel salario orario nominale (vedi la fig.11).D'altra parte, una perturbazione nella variabile salario orario nominale non induce movimenti di un qualche rilievo nelle altre variabili nel caso essa sia analizzata congiuntamente alla disoccupazione in senso stretto maschile, mentre nel caso essa sia analizzata congiuntamente alla disoccupazione in senso stretto femminile porta con sè relativamente ampi movimenti iniziali nella stessa disoccupazione (vedi le figg. 12 e 13).

L'interpretazione della dinamica causale di un modello multivariato non è sempre facile. Sebbene lo studio delle reazioni del modello ad eventuali perturbazioni possa fornire spesso indicazioni utili, esso non sempre consente di cogliere in maniera distinta effetticausali diversi. Per questo motivo diventa interessante proseguire l'analisidi simulazione dinamica andando a vedere gli effetti della scomposizione della varianza dell'errore di previsione relativo ad ogni variabile, sulla base delle innovazioni ortogonalizzate del modello[9]. Al riguardo, conviene ricordare che una variabile è esogena se tutta la sua varianza è dovuta alle sue stesse innovazioni.

Innanzitutto, per la popolazione maschile (veda la Tab. 1), l' 84,92 % della varianza di previsione della disoccupazione in senso stretto risulta spiegata dalle innovazioni in sè stessa, mentre la stessa variabile risulta spiegare nel tempo il 46,08 % della varianza di previsione dell'indice dei prezzi al consumo. Questo risultato conferma, peraltro, quanto avevamo già riscontrato con la stima dei modelli autoregressivi vettoriali.

Per la popolazione femminile (vedi la Tab. 2), invece, la varianza della disoccupazione in senso stretto risulta spiegata solamente per il 48,25 % dalle innovazioni in se stessa, mentre per il 24,28 % risulta spiegata dalle innovazioni nell'indice dei prezzi al consumo. Il 24,15 % della varianza del tasso di interesse risulta spiegato dalle innovazioni nella disoccupazione in senso stretto. È da notare che fra queste due variabili si era riscontrata la correlazione contemporanea più elevata in valore assoluto. A differenza della popolazione maschile, la varianza dell'indice dei prezzi al consumo risulta spiegata per il 59,18 % dall'indice stesso. Anche il salario orario nominale

---

9 Questa scomposizione consente ancora di evidenziare le risposte del sistema a perturbazioni in determinate variabili, ma, essendo gli errori del sistema trasformato ortogonali fra di loro, si può attribuire agli specifici errori la variabilità che il sistema mostra in risposta alle perturbazioni. In altre parole si può scomporre la varianza degli errori di previsione di ogni variabile ai diversi orizzonti temporali in parti attribuibili alle innovazioni in sè stessa e nelle altre variabili (vedi Sims, 1981, p. 285).

mostra una percentuale di varianza del 59,98 % spiegata da se stesso, mentre una percentuale del 25,46 % risulta spiegata dalle innovazioni nell'indice dei prezzi. La maggior parte della varianza di quest'ultimo non spiegata da se stesso risulta spiegata dal tasso di interesse.

In sostanza, mentre per la popolazione maschile la variabile disoccupazione in senso stretto può essere considerata esogena rispetto al modello, per la popolazione femminile il comportamento della stessa variabile risulta legato ai comportamenti delle altre.

Nel modello in cui compare la variabile tasso di disoccupazione totale (vedi la Tab. 3), le innovazioni nell'indice dei prezzi al consumo spiegano solo in misura del 21,67 % la variabilità dello stesso tasso ed il 47,57 % della variabilità del salario orario. E interessante notare che le innovazioni nel salario non compaiono mai a spiegare in maniera significativa le altre variabili. L'ultima considerazione che vogliamo fare è che i risultati ottenuti appaiono robusti rispetto a cambiamenti nell'ordinamento delle variabili, e questo non fa che confortare le analisi e le interpretazioni anche in presenza di correlazioni contemporanee significative fra le innovazioni.

4. *Ipotesi interpretative*

Data l'ottica esplorativa che caratterizza questo studio, si cercherà ora di individuare, sulla base delle analisi svolte, quali siano i legami la cui natura può costituire oggetto di ulteriori analisi econometriche, e quali invece risultino non confermati dai dati. Nella sezione precedente , analizzando distintamente le serie che formano l'aggregato disoccupazione, è emerso che mentre tutte le serie relative alla disoccupazione maschile agiscono soprattutto sui prezzi, e possono essere considerate esogene rispetto al modello, le serie relative alla disoccupazione femminile - salvo le persone in cerca di prima occupazione - hanno relazioni reciproche con le altre variabili.

Per i maschi non risulta quindi confermato un meccanismo à la Phillips, dato che il legame fra disoccupazione e prezzi non passa attraverso l'influsso sul salario. Per quanto riguarda le donne invece la disoccupazione in senso stretto e il tasso di disoccupazione complessivo influenzano e sono influenzati da variabili di prezzo. Dato che non si e' riscontrato un effetto del salario sulll'occupazione femminile (Passamani e Schenkel,1988), ciò implica che il salario agisce sulla ricerca di lavoro e sulla partecipazione femminile. Questo risultato è coerente con l'influsso del salario sul tasso di partecipazione riscontrato ultimamente per l'Italia da Quintieri e Rosati (1988).

In ogni caso l'interpretazione di questi risultati non sembra del tutto immediata. Forse se si tiene conto del fatto che il tasso di interesse, variabile che risente particolarmente dei provvedimenti delle autorità monetarie e allo stesso tempo è fortemente indicativa del ciclo e degli orientamenti generali della politica economica, presenta una forte correlazione contemporanea con salario e disoccupazione femminile in senso stretto, si può pensare che vi sia un legame fra disoccupazione e provvedimenti di politica economica,

e fra questi e il salario. Ovviamente , però, un effetto della disoccupazione in senso stretto femminile sui salari industriali si potrebbe anche interpretare nel senso di Phillips.

Rimane inoltre incerto il significato del legame fra disoccupazione femminile e tasso di interesse. Anche se fosse presente sostituzione intertemporale (peraltro negata da alcune evidenze empiriche per l'Italia: vedi Lucifora, 1987), è difficile pensare che la risposta dell'offerta di lavoro non si manifesti con qualche ritardo (Andrews e Nickell, 1986). Del resto, dall'insieme delle analisi effettuate risultano inapplicabili alla realtà italiana versioni semplici di tale teoria[10],secondo le quali la disoccupazione dovrebbe risultare, come l'offerta di lavoro, da una combinazione di ritardi sul salario e sul tasso di interesse, e la disoccupazione passata dovrebbe avere effetto sulla futura solo attraverso il suo influsso sul salario. Tutte le serie relative alla disoccupazione e all'occupazione invece dipendono significativamente dal loro passato.

I nostri strumenti di analisi empirica non ci permettono però di affermare che esiste persistenza, o addirittura che essa è dovuta a una causa piuttosto che ad un'altra. Ad esempio, secondo i nostri risultati, non si può escludere isteresi nel caso italiano, ipotesi che è invece rifiutata da altri autori (Gagliardi e Coe, 1985; Bodo e Visco, 1987; Baici, 1987). Risulta infatti non smentita dai dati l'interpretazione secondo la quale i livelli passati contribuiscono a bloccare il tasso di disoccupazione corrente su livelli più alti. Resta però da determinare se l'innalzamento del tasso di disoccupazione sia stabile. Una volta accertata la natura dell'aumento, si potranno indagarne le cause, e in particolare se risultino adeguate le spiegazioni in termini di insufficienza di capitale fisico e di deterioramento di quello umano che sembrano meglio applicabili al caso in esame di quella, più nota, fornita dai modelli *Insiders-Outsiders*. Secondo i nostri risultati, questi sono scarsamente adattabili alla realtà italiana, data la ridotta capacità esplicativa del salario rispetto alle altre variabili.

Sempre a questo proposito, vi è da notare che il salario risulta determinato quasi unicamente dai prezzi. Ciò è perfettamente congruente con la costanza del salario reale a partire dal 1980. In questa situazione non è facile discernere un ruolo del salario reale sulla domanda di lavoro (Kennan, 1988). E invece possibile che le variazioni del salario nominale abbiano un effetto solo sull'occupazione, e non sul salario reale (Blanchard e Summers, 1986). Dato però, ancora una volta, che l'influenza del salario reale sulle altre variabili è dubbia, è più pertinente il richiamo al meccanismo secondo il quale, se vi è rigidità dei prezzi relativi, un aumento della spesa pubblica in stagflazione può aumentare i prezzi senza alcun effetto sulla disoccupazione (Fitoussi e Le Cacheux, 1988)[11]. Se, infatti, all'attività lavorativa dei maschi viene attribuito un maggior significato sociale, un aumento dei disoccupati (e non delle disoccupate) può portare alla messa in moto di politiche espan-

---

10 Non però quelle che si rifanno ad una versione più sofisticata, che tenga conto della possibilita' di correlazione fra termini d'errore di equazioni individuali di offerta di lavoro (vedi Ashenfelter e Card, 1982)

sive, da cui l'effetto sui prezzi.

In conclusione, riteniamo che il risultato più interessante del lavoro compiuto sta nel fatto che la prima intuizione, di un comportamento differenziato fra i vari segmenti che compongono gli aggregati occupazione e disoccupazione, è sostanzialmente confermata. Alcune relazioni valide per le serie disaggregate si riflettono anche sugli aggregati maggiori, ma in nessuno dei casi qui studiati una relazione si verifica sia per l'aggregato che per tutte le singole serie dalle quali esso risulta composto.

I risultati ottenuti finora indicano due linee per la ricerca futura: da una parte, approfondire lo studio della persistenza delle serie, specialmente dal punto di vista dei metodi; e dall'altra, raffinare la disaggregazione delle serie, per arrivare a definire meglio quali segmenti si dimostrino sensibili all'influenza dei vari fattori di prezzo considerati.

E d'altronde auspicabile estendere le analisi multivariate oltre l'ambito puramente esplorativo, anche per poter ampliare e approfondire gli obiettivi interpretativi e le implicazioni teoriche delle analisi effettuate.

Appendice: figure e tabelle

Legenda delle variabili:

DISSM = Disoccupati in senso stretto, maschi
DISSF = Disoccupate in senso stretto, femmine
RIC1OCCM = Persone in cerca di prima occupazione, maschi
RIC1OCCF = Persone in cerca di prima occupazione, femmine
ALPERCEM = Altre persone in cerca di occupazione, maschi
ALPERCEF = Altre persone in cerca di occupazione, femmine
TASSODIS = Tasso di disoccupazione totale
TASSODISM = Tasso di disoccupazione, maschi
TASSODISF = Tasso di disoccupazione, femmine
WA =Salario orario nominale nell'industria
RWA= Salario orario reale nell'industria
PC70Z= Indice dei prezzi al consumo per le famiglie di operai ed impiegati (base 1970 = 100)
RBOT = Tasso di interesse sui BOT a tre mesi

---

11 Notiamo che questa rigidità può aversi sia per effetto dell'indicizzazionedei salari che per effetto dellacostanza del *mark-up* (Fitoussi e Le Cacheux, 1988).

Fig. 1 - *Componente di trend della variabile WA*

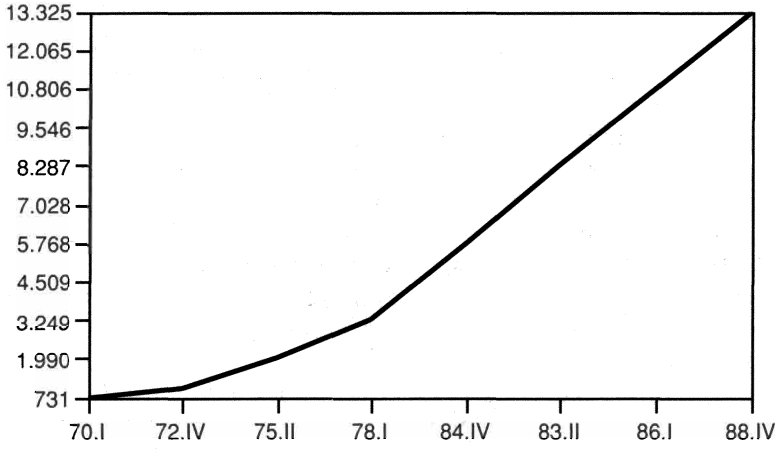


Fig. 2 - *Componente di trend della variabile RWA*

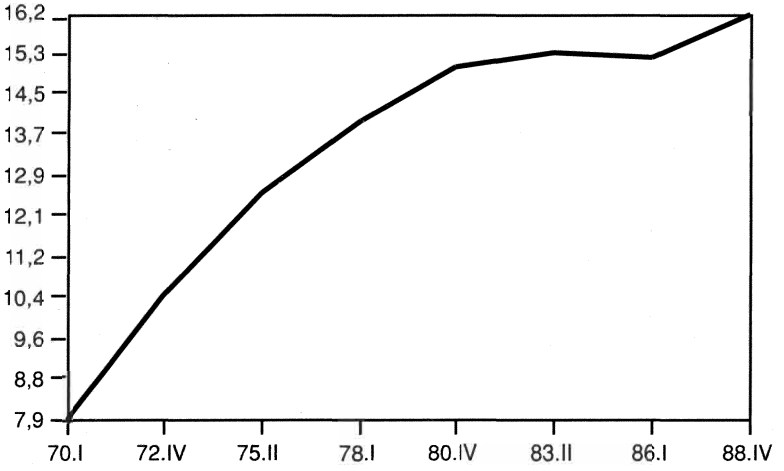


Fig. 3 - *Componente di trend della variabile TASSODIS*

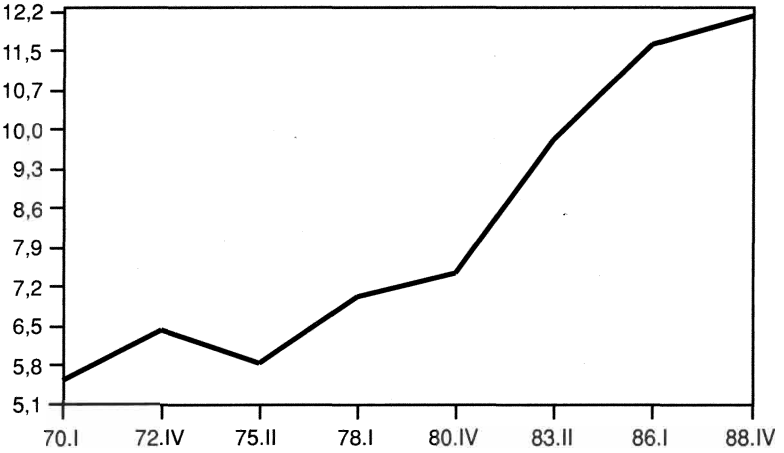


Fig. 4 - *Componente di trend della variabile* ***TASSODISM***

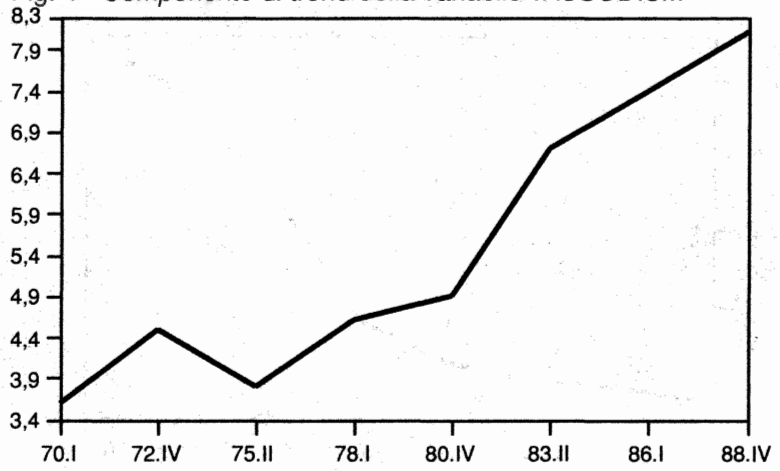


Fig. 5 - *Componente di trend della variabile TASSODISF*

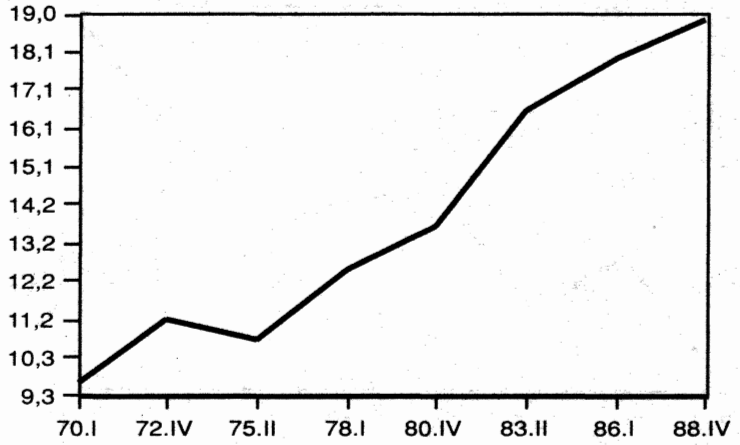


Fig. 6 - *Componente di trend della variabile DISSM (valore x 1000)*

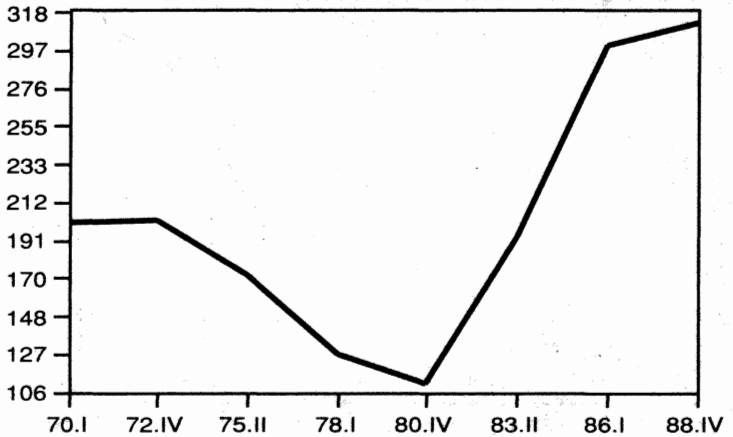


Fig. 7 - *Componente di trend della variabile DISSF (valore x 1000)*

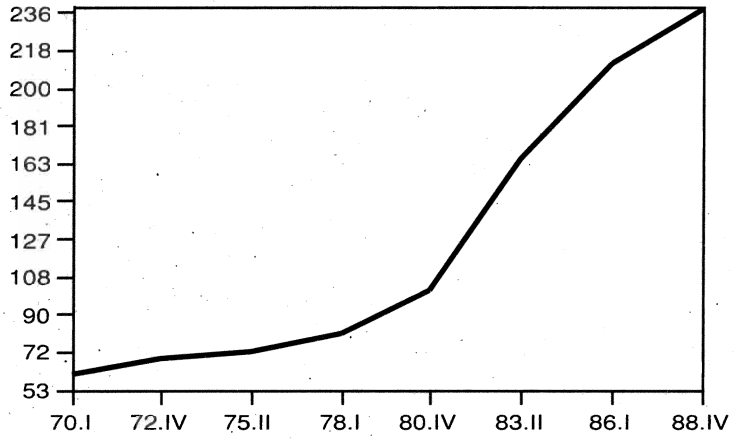


Fig. 8 - *Componente di trend della variabile ALPERCEM (valore x 1000)*

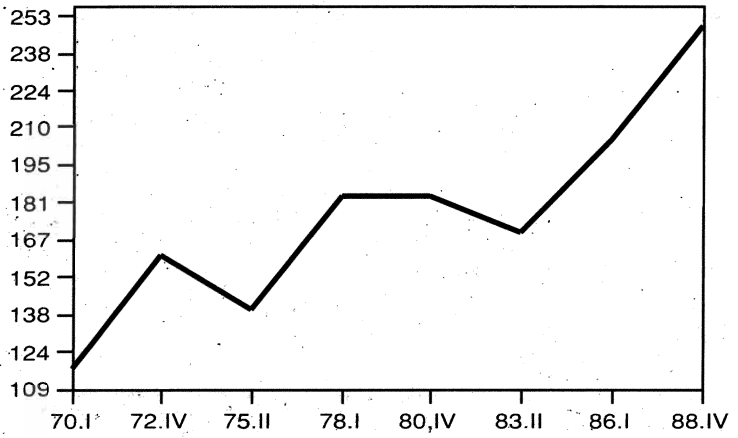


Fig. 9 - *Componente di trend della variabile ALPERCEM (valore x 1000)*

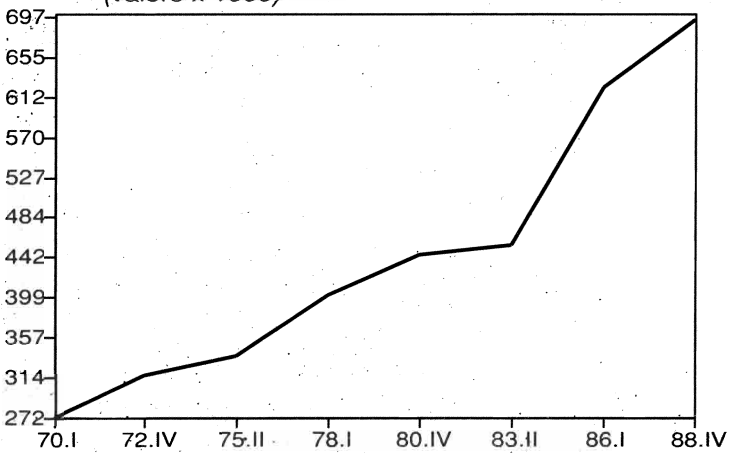


Fig. 10 - *Perturbazione nella variabile TASSODISM*

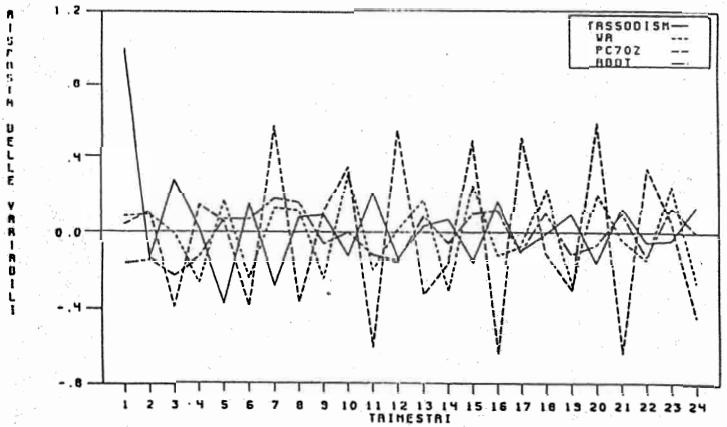


Fig. 11 - *Perturbazione nella variabile TASSODISF*

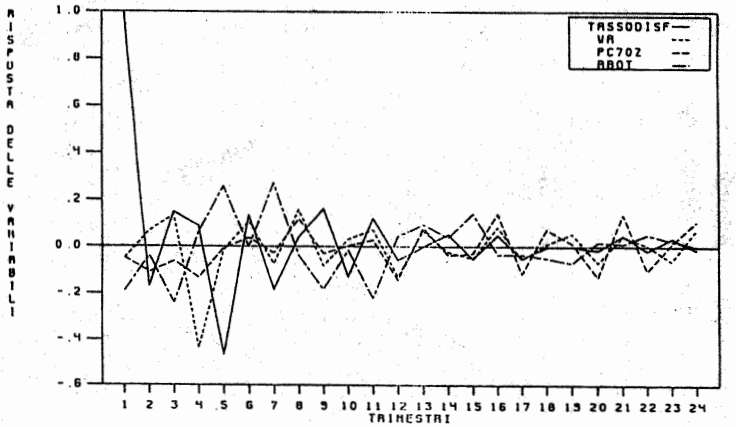


Fig. 12 - *Perturbazione nella variabile WA*

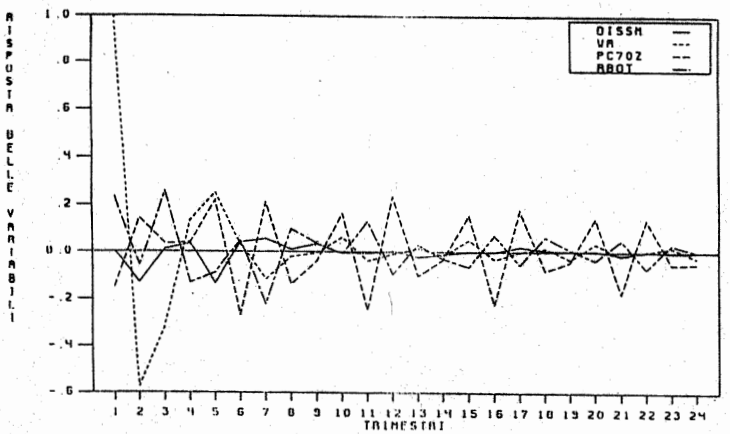
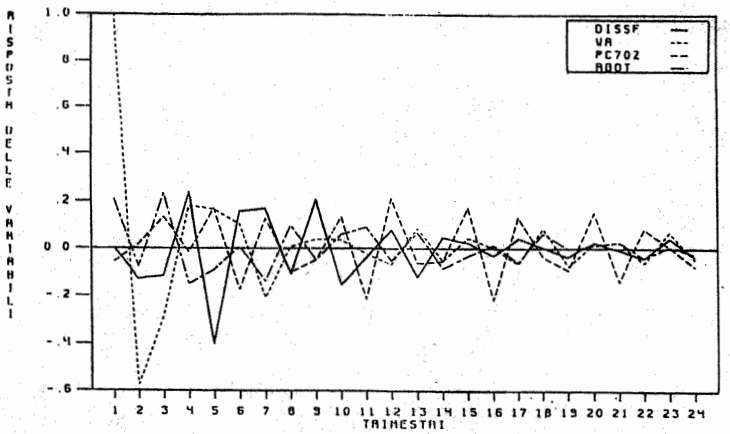


Fig. 13 - *Perturbazione nella variabile WA*



Tab. 1: *Percentuali di varianza dell'errore di previsione, alla distanza di τ periodi, imputabili alle innovazioni ortogonalizzate relative alle diverse variabili del modello* [a]

Errore di previsione in:	τ	Innovazioni ortogonalizzate in:			
		DISSM	WA	PC70Z	RBOT
DISSM [b]	4	90,53	1,61	6,49	1,36
	8	86,97	3,07	7,90	2,06
	12	85,97	3,10	8,14	2,80
	24	84,92	3,15	8,29	3,65
WA	4	2,11	83,91	9,57	4,41
	8	11,25	70,74	12,13	5,88
	12	15,44	65,54	12,57	6,45
	24	17,74	62,92	12,56	6,78
PC70Z	4	25,82	2,28	59,22	12,68
	8	36,69	5,92	45,16	12,23
	12	40,36	7,18	40,79	11,67
	24	46,08	7,93	34,51	11,48
RBOT	4	7,98	11,45	14,79	65,78
	8	8,45	12,07	17,60	61,88
	12	9,76	12,03	19,46	58,75
	24	16,08	11,30	20,61	52,01

(a) Tutte le serie sono state filtrate mediante (1-B)(1-B**4) e sono state quindi ricalcolate in termini di scarti dalla loro media.

(b) per la *legenda* delle variabili, vedi a pag. 347.

Tab. 2 : *Percentuali di varianza dell'errore di previsione, alla distanza di τ periodi, imputabili alle innovazioni ortogonalizzate relative alle diverse variabili del modello* [(a)]

Errore di previsione in:	τ	Innovazioni ortogonalizzate in:			
		DISSF	WA	PC70Z	RBOT
DISSF	4	76,86	6,51	14,77	1,86
	8	64,00	14,98	16,89	4,14
	12	57,58	15,77	19,23	7,42
	24	48,25	13,86	24,28	13,61
WA	4	0,98	82,58	12,89	3,55
	8	1,33	70,49	21,73	6,44
	12	1,65	64,04	24,90	9,41
	24	2,06	59,98	25,46	12,50
PC70Z	4	3,69	1,26	74,43	20,62
	8	4,65	3,57	68,38	23,40
	12	5,15	5,72	64,29	24,84
	24	5,87	8,80	59,18	26,15
RBOT	4	14,33	10,59	9,55	65,54
	8	19,97	9,88	11,50	58,65
	12	22,77	9,83	12,88	54,52
	24	24,15	10,51	13,73	51,61

(a) Vedi le note della Tab. 1.

Tab. 3 : *Percentuali di varianza dell'errore di previsione, alla distanza di τ periodi, imputabili alle innovazioni ortogonalizzate relative alle diverse variabili del modello* [(a)]

Errore di previsione in:	τ	Innovazioni ortogonalizzate in:			
		TASSODIS	WA	PC70Z	RBOT
TASSODIS	4	95,33	0,73	2,66	1,28
	8	84,68	7,50	6,07	1,76
	12	79,46	8,86	9,18	2,50
	24	64,76	9,43	21,67	4,14
WA	4	9,29	75,57	13,99	1,15
	8	9,56	61,87	26,47	2,10
	12	11,56	51,59	33,50	3,35
	24	13,66	31,76	47,57	7,01
PC70Z	4	2,16	1,23	87,30	9,32
	8	3,63	1,48	83,26	11,63
	12	5,38	1,88	80,01	12,73
	24	7,92	3,00	76,24	12,84
RBOT	4	8,12	11,84	9,85	70,19
	8	12,20	10,38	10,46	66,96
	12	14,27	9,70	10,77	65,26
	24	15,40	9,55	11,05	64,00

(a) Vedi le note della Tab. 1.

18

LA STIMA DEI FLUSSI E DI MATRICI DI TRANSIZIONE

Lorenzo Bernardi e Susanna Zaccarin *

1. *Introduzione*

Il processo di parziale rotazione delle unità statistiche del primo stadio di campionamento (comuni) e delle unità di secondo stadio (famiglie) adottato nella rilevazione trimestrale italiana sulle forze di lavoro (RTFL), consente di analizzare comportamenti dinamici della popolazione nel mercato del lavoro a livello individuale, tramite la misurazione delle transizioni tra le diverse condizioni occupazionali, assicurando per questa via elementi informativi preziosi tanto per una maggior comprensione delle dinamiche del settore quanto per indirizzi di programmazione politico-economica.

I vantaggi conoscitivi che derivano da questa opportunità sono molteplici e risiedono, in particolare, nella possibilità di individuare i movimenti, realizzatisi nell'arco del periodo di osservazione, tra stati. Più precisamente, si considerano, per ciascun individuo, le condizioni alla prima e alla successiva rilevazione, non essendo possibile con le informazioni attualmente rilevate, riconoscere numero dei movimenti e durata delle permanenze all'interno del periodo di osservazione. Pur con questi limiti, le informazioni che si possono ottenere risultano di gran lunga più apprezzabili dei flussi netti desumibili dal confronto tra i dati di *stock* raccolti in due occasioni successive (Veevers e Macredie, 1983; Moriani, 1981).

Tali opportunità interpretative aggiuntive, tuttavia, sono condizionate dall'insieme di scelte che si compiono per la costruzione della matrice dei flussi. Per la loro determinazione, infatti, insorgono, da un lato, nuovi e specifici problemi metodologici e, dall' altro, risultano aggravati fenomeni di disturbo già presenti nella definizione delle informazioni di consistenza alle due date.

Obiettivo di questo capitolo è quello di fornire una prima ricognizione dei problemi che devono essere affrontati per la costruzione di matrici di flusso tra condizioni occupazionali nell'ambito dell' indagine italiana sulle forze di lavoro, nel tentativo, in particolare, di sfruttare tutte le informazioni che si rendono disponibili dal confronto di due rilevazioni. Oltre a quelle relative

---

* Il lavoro è frutto dell'impegno comune degli autori. Per quanto riguarda la stesura, L. Bernardi ha scritto le sezz. 1, 3.2 e 4 e S. Zaccarin le sezz. 2, 3.1, 5, 6 e 7.

agli individui presenti in entrambe le occasioni, si hanno anche informazioni parziali sui soggetti per i quali manca l'osservazione, per vari motivi, ad una delle due occasioni. Poichè anche tali individui, per la maggior parte, appartengono alla popolazione oggetto di studio e non è possibile escludere *a priori* l'esistenza di relazioni tra il mancato contatto e comportamenti differenziali rispetto al mercato del lavoro, pare ragionevole incorporare nella procedura di stima anche i dati ad essi riferiti.

La stima dei flussi è inoltre prerequisito importante per avviare prospettive alternative di classificazione degli stati occupazionali, progressivamente proiettate a cogliere con maggiore sensibilità i processi interni al mercato del lavoro e ad evidenziare comportamenti specifici di particolari gruppi sociali[1].

Il capitolo è organizzato come segue. Nella sez. 2 sono descritti i principali fattori di disturbo che in varia misura possono distorcere le stime dei flussi, unitamente ai metodi proposti in letteratura per il loro trattamento. Nella sez. 3 sono esaminate le caratteristiche dell'indagine italiana (sez. 3.1) e la procedura di stima dei flussi adottata dall'Istat (sez. 3.2). L'individuazione, e l'analisi, degli aggregati che compongono l'insieme dei dati disponibili dal confronto tra due indagini (compresenti, non trovati alla 2^ indagine, non trovati alla 1^ indagine) si rivela operazione complessa, ma necessaria (sez. 4) al fine di poter applicare ai dati italiani l'impianto teorico per la riallocazione delle informazioni parziali e le conseguenti procedure di stima dei flussi in presenza di dati mancanti (sez. 5). Nella sez. 6 è presentato un esempio di applicazione di una classe di modelli che, nell'ambito di tale impianto, esplicita le relazioni tra non risposta e variabili ogggetto di studio. Nella sez. 7, infine, sono riportate alcune considerazioni conclusive, che sottolineano la necessità di informazioni aggiuntive nell'indagine italiana al fine di disporre di dati più adeguati.

2. *Principali aspetti di disturbo dei flussi da indagini longitudinali*

La determinazione dei flussi di popolazione nel mercato del lavoro, mediante il confronto tra le condizioni dichiarate in due occasioni successive d'indagine, appare influenzata da alcuni fattori di disturbo: da un lato, legati ai criteri adottati per il confronto e alle soluzioni operative predisposte per realizzarlo; dall'altro, tipici di rilevazioni su larga scala come quelle sulle forze di lavoro, la cui incidenza risulta aggravata nel caso dei flussi.

L' esperienza maturata, a partire dai primi anni '60, negli U.S.A. e in Canada con appositi studi sulla qualità dei flussi delle forze di lavoro, ha permesso di individuare alcuni di tali fattori che in dettaglio, riguardano (Hogue, 1985):

(a) *il piano di rotazione.* Il calcolo dei flussi è basato su di un sottoinsieme del campione originario di individui presenti ad ogni rilevazione, cioè a dire il campione costituito dalla frazione di unità sovrapposte in due

---

1 Per alcune analisi in questo senso vedi Bernardi e Zaccarin (1987).

indagini successive, determinata dal piano di rotazione. Il sub-campione presenta un errore campionario diverso e più elevato (essendo di dimensioni inferiori) del campione originario.

(b) *le mancate risposte.* Durante la realizzazione operativa dell' indagine la frazione teorica di sovrapposizione viene ridotta a causa delle mancate interviste ad individui nel campione (per trasferimenti, assenze, rifiuti, ecc.; alla stessa problematica attiene, in questo ambito, il caso della mancata risposta a quesiti di interesse). Poichè non è possibile escludere, *a priori*, che le cause della mancata intervista non siano, in qualche modo, collegate ai comportamenti differenziali rispetto al mercato del lavoro, la non inclusione di questi individui nel calcolo dei flussi può provocare distorsione nei risultati;

(c) *gli errori di risposta.* L'errata classificazione della condizione lavorativa dovuta a: (c1) errata comprensione del quesito, (c2) errata percezione della condizione effettiva, (c3) errata conoscenza della condizione dell'interessato da parte del rispondente (nel caso di *proxy respondent*) può distorcere il volume dei flussi, segnalando movimenti tra condizioni di fatto mai avvenuti. Tali errori, mentre possono tendere alla compensazione nei dati di *stock,* vengono amplificati, poichè si cumulano, nei dati di flusso (National Commission on Employment and Unemployment Statistics, 1979; Macredie, 1983).

(d) *la distorsione dovuta al gruppo di rotazione ('rotation group bias').* I *patterns* di risposta relativi all'attività lavorativa esibiti da individui intervistati la prima volta possono risultare molto diversi da quelli di individui intervistati più volte (tendenzialmente nel senso di riportare incidenze maggiori dei fenomeni di disoccupazione; Bailar, 1975). Possibili spiegazioni di tale comportamento differenziato sono state individuate in fenomeni di *telescoping* e di condizionamento da reintervista;

(e) *la procedura di abbinamento.* I criteri che possono essere adottati per operare il confronto tra i *records* individuali inevitabilmente danno luogo a due tipi di errori: l'inclusione di falsi positivi e l'esclusione di falsi negativi (Giusti, Marliani e Torelli, 1987 e 1988; vedi anche il cap. 7), i cui effetti riverberano nella stima dei flussi.

L'effetto più evidente dell' azione congiunta di tali fattori si riscontra nelle differenze in valore, e spesso anche nella direzione, tra i saldi ottenuti dai dati di flusso e quelli di *stock*.

Metodi di stima alternativi, esistenti in letteratura per il trattamento di alcuni dei problemi segnalati, sono stati proposti con soluzioni specifiche all' analisi dei flussi.

La tecnica di aggiustamento più semplice è quella dell' *iterative proportional fitting* (Bishop, Fienberg e Holland, 1975), che consiste nel riproporzionare le celle della matrice dei flussi, imponendo l'uguaglianza tra i totali marginali della matrice e i totali ottenuti dalle indagini *cross-section.* L'applicazione di tale metodo, che pur risolve l'incoerenza tra i totali di flusso e quelli di *stock*, non implica necessariamente un miglioramento nell'accuratezza dei flussi. Tale accorgimento è infatti usato, generalmente, insieme ad altre tecniche di aggiustamento (Moriani, 1981; Fienberg e Stasny, 1983).

Specificamente per il trattamento degli errori di risposta, la cui incidenza appare particolarmente severa, sono stati proposti due metodi (vedi Abowd e Zellner, 1985, Fuller e Chua, 1985, e Poterba e Summers, 1985, per il primo; Statistics Canada, 1979, e Wong, 1983, per il secondo; per una analisi critica dei due metodi vedi poi Meyer, 1988). Tali metodi usano stime dell'errore di classificazione, calcolate sulla base di informazioni provenienti da reinterviste più accurate alle medesime unità (per es., la *Reinterview Survey,* indagine sulla qualità dei dati condotta correntemente in associazione con la CPS - *Current Population Survey* - statunitense; vedi Flaim e Hogue, 1985).

Infine, le tecniche più comuni per tener conto delle mancate interviste (*missing data*) sono quelle basate su procedure di ponderazione e/o di sostituzione. Nella prima, nel calcolo delle quantità riferite alla popolazione si tiene conto, oltre alla probabilità di estrazione, anche della probabilità di risposta, mentre nella seconda l'unità non rispondente viene sostituita con un'altra estratta da un campione di riserva. Tale pratica può essere vista come un caso particolare delle più generali tecniche di imputazione per il trattamento dei *missing data* (Little e Rubin, 1987). L'assunzione sottostante all'uso di tali tecniche è quella di assoluta casualità per le non risposte, ovvero di indipendenza tra le variabili di interesse e le cause di non risposta. Un approccio alternativo è quello basato su modelli che esplicitano in modo diretto le relazioni tra il fenomeno sotto studio (nel nostro caso il comportamento rispetto al mercato del lavoro) e la non risposta: esso non assume tale indipendenza, e l'uso di questi modelli appare molto proficuo soprattutto nell' ambito di dati di *panel,* come quelli per l'analisi dei flussi (Little, 1985; Stasny, 1988). Nell' ambito dei metodi per il trattamento delle non risposte deve essere ricordata anche la procedura proposta da Abowd e Zellner (1985) per i dati della CPS, basata su un modello moltiplicativo di allocazione delle informazioni parziali relative ai soggetti non rispondenti ad una delle due occasioni. La stima dei parametri del modello è ottenuta rendendo minima la discrepanza tra le marginali della matrice dei flussi e i totali di stock.

In chiusura di questa sezione dedicata alla rassegna dei problemi che devono essere affrontati nella stima dei flussi, pare il caso di sottolineare che gran parte delle soluzioni proposte hanno ancora carattere esplorativo piuttosto che di metodologie consolidate, correntemente incorporate nel procedimento di stima. A questo proposito, le scelte adottate dai principali enti produttori di statistiche sulle forze di lavoro sono varie e molto differenziate[2].

---

2 L' U.S. Bureau of the Census, per esempio, ha ripreso solo dal 1982 la pubblicazione su base annuale dei dati di flusso provenienti dalla CPS, sospesa dal 1953 a causa delle incoerenze rilevate. Nessun intervento correttivo viene applicato ai dati, mentre continua la promozione dell' attivita di ricerca. La procedura sviluppata da Statistics Canada, invece, prevede una serie di passaggi successivi per pervenire alla stima dei flussi (Fienberg e Stasny, 1983). Il primo passo effettua la correzione per le non risposte, pesando opportunamente i soli rispondenti alle due occasioni considerate. Nel secondo passo sono riproporzionati i flussi di entrata e uscita dalla popolazione rispetto ai valori di stime censuarie esterne. Nel terzo è corretta la distorsione introdotta dagli errori di risposta (Wong, 1983). Il quarto passo, infine, prevede l'applicazione dell'iterative proportional fitting per la quadratura dei totali della matrice di flusso con i totali di stock.

Anche l'Istat pubblica dal 1979 tavole di flusso ottenute sulla scorta di una complessa metodologia che verrà discussa nel seguito.

3. *La RTFL e i flussi*

3.1. *Le principali caratteristiche*

Prima di procedere alla descrizione della metodologia adottata dall'Istat per la stima dei flussi, è opportuno richiamare brevemente alcune caratteristiche dell'indagine trimestrale italiana, particolarmente cruciali per gli effetti che possono indurre sulla qualità dei flussi (per maggiori ragguagli, vedi il cap. 1).

Secondo lo schema di rotazione previsto (del tipo 2-2-2), una frazione teorica pari al 50% del campione di famiglie è comune tra due rilevazioni consecutive e tra due rilevazioni corrispondenti di due anni consecutivi. In realtà, a causa della rotazione dei comuni nell'indagine di luglio, la percentuale di sovrapposizione si riduce al 41% circa per due indagini distanziate di un anno o per i trimestri aprile-luglio. Inoltre, le cadute di unità di 1^ e 2^ stadio che si accompagnano alla realizzazione dell'indagine possono ridurre ulteriormente la frazione di sovrapposizione (Parenti, 1979).

La procedura di rilevazione prevede che, qualora una famiglia non sia reperita, essa sia sostituita con un'altra reperibile. Tale pratica, che pur assicura il raggiungimento della numerosità campionaria prevista, tende a sovrastimare le famiglie con particolari caratteristiche di residenza (e con specifiche caratteristiche verso il lavoro) e a sottostimare le famiglie (e i modelli di comportamento) associati a situazioni di elevata mobilità.

La conduzione dell'indagine non prevede programmi di controllo sulla qualità dei dati mediante reinterviste, per cui le possibilità di verifiche dell'incidenza dell'errore di risposta appaiono molto limitate. A questo proposito, devono essere ricordati anche gli effetti del piano di correzione automatica, la cui adozione può generare movimenti fittizi tra condizioni, in quanto assicura la compatibilità tra le risposte fornite occasione per occasione ma non rispetto al confronto tra occasioni. In questo senso, l'uso dei dati abbinati potrebbe fornire valide indicazioni sulla qualità delle risposte anche a livello di singola indagine (vedi il cap. 8).

Infine, come evidenziano le analisi del cap. 9, l'adozione dello schema ruotato non sembra produrre particolari effetti di *rotation group bias.*

3.2. *Le matrici di flusso dell'Istat*

L'Istat pubblica dal 1979 matrici di transizione tra condizioni occupazionali mediante un complesso programma di calcolo, il cui impianto generale è illustrato in Moriani (1981). I passi principali della procedura possono essere così riassunti:

(a) abbinamento dei dati individuali relativi alle due rilevazioni a confronto;
(b) riporto all'universo;
(c) elaborazione delle stime su dati esterni e 'chiusura' della matrice con il metodo RAS.

All'interno della procedura vi sono alcuni passaggi di grande rilievo per i potenziali riflessi sulle stime dei flussi che è conveniente ricordare:

(a) le regole per il collegamento dei soggetti appaiono particolarmente restrittive, essendo costruite sull'uguaglianza dell' età - con la tolleranza necessaria per tener conto del tempo tra le due indagini - del sesso, della relazione con il capofamiglia e dell' istruzione (vedi il cap. 7);
(b) le operazioni per il riporto all'universo dei flussi paiono non del tutto convincenti[3] in quanto:
  b1) trascurano completamente i non abbinati che, tramite la procedura di costruzione dei coefficienti di riporto, di fatto vengono assimilati all'insieme degli abbinati del proprio strato;
  b2) i comportamenti degli strati mancanti vengono, a loro volta, assimilati a quelli della classe di comuni cui appartengono entro la regione.
(c) l'applicazione del metodo RAS si effettua a partire dalla matrice dei flussi grezza costruita con l'attribuzione dei pesi ai soggetti abbinati, tenendo come vincoli:
  (c1) i vettori della popolazione per strato al tempo 1 e 2;
  (c2) i vettori dei morti e degli emigrati al tempo 1 e dei nati e degli immigrati al tempo 2, riallocati (tranne i nati) nelle categorie socio-economiche previste dalla matrice, in parte sulla scorta di informazioni esterne (coefficienti di mortalità per sesso e condizione professionale), in parte nell'ipotesi di proporzionalità alle persone temporaneamente all'estero rilevate con le indagini campionarie al tempo 1 e 2 (per la redistribuzione rispettivamente degli immigrati e degli emigrati).

---

3 Nel calcolo dei pesi per il riporto all'universo, la popolazione di riferimento è quella relativa alla seconda occasione osservata, depurata idealmente dei nati e degli immigrati dall'estero nel periodo intercorso tra la 1^ e la 2^ indagine. In realtà, mancando una fonte corrente capace di assicurare queste informazioni al livello territoriale desiderato e con la tempestività ed accuratezza necessaria, si adottano valori nazionali fissi, in ogni trimestre, di nati e di immigrati, desunti da stime o osservazioni degli anni precedenti, per definire un coefficiente di correzione che viene applicato direttamente alla matrice dei flussi, prima di assoggettarla al metodo RAS. I coefficienti di riporto utili alla costruzione della matrice dei flussi sono ottenuti con tre successive fasi di elaborazione:

(a) determinazione dei primi coefficienti di riporto, forniti dal rapporto tra la popolazione ottenuta secondo i passi precedenti e individui complessivamente abbinati: questa operazione viene compiuta per sesso all'interno di ciascun strato;

(b) a causa della caduta di alcune unità di 1^ stadio (tipicamente comuni di tipo B) e/o della irrilevanza del numero di accoppiamenti effettuati in altre unità di 1^ stadio (la soglia è pari a 5), può talvolta avvenire che l'uso dei coefficienti di riporto costruiti in (a) non generi l'intera popolazione prevista. Si adotta pertanto un fattore regionale di correzione, ottenuto dal rapporto tra popolazione regionale attesa e popolazione regionale calcolata con i coefficienti di riporto: questa operazione è effettuata per i tre tipi di comuni considerati (capoluoghi > 20.000 ab., non capoluoghi > 20.000 ab., < 20.000 ab.);

(c) calcolo dei coefficienti per i dati di flusso, forniti dal rapporto (per strato) della popolazione al tempo 2 e numero dei soggetti abbinati, corretto con il fattore regionale per classe di comuni come illustrato in (b).

In conclusione sembra di poter dire che l'insieme dei criteri adottati rischia di svolgere una funzione di enfasi degli errori indicati dalla letteratura sull'argomento, in relazione soprattutto a due scelte. (i) L'estensione all'intera popolazione delle sole informazioni sugli abbinati può essere un fattore di distorsione piuttosto rilevante, perchè fondato sull'assunto, poco plausibile, dell'indipendenza dei comportamenti sul mercato del lavoro dai motivi di non abbinamento. (ii) Nella stessa direzione sembrano muoversi i numerosi correttivi imposti per giungere alla 'chiusura' della matrice.

4. *Un tentativo di classificazione degli aggregati da considerare per la stima dei flussi nell'indagine italiana*

La ricognizione compiuta sulle procedure adottate dall'Istat per la costruzione delle matrici di flusso, i numerosi esperimenti e le soluzioni approntate per l'abbinamento dei *records* individuali (vedi il cap. 7), le prove fatte per riconoscere i comportamenti rispettivamente dei compresenti alle due indagini e dei cosiddetti *missing* in una di esse (Bernardi e Zaccarin, 1987), hanno posto in evidenza la necessità di pervenire alla corretta individuazione, e analisi, dei singoli aggregati che si possono generare nell'intero processo di conduzione dell'indagine e di formazione dei dati di base, per poter poi affrontare l'uso di modelli per la stima dei flussi mediante lo sfruttamento delle informazioni relative a soggetti osservati in due occasioni e di quelle relative a soggetti osservati in una sola. Riesaminando i problemi evocati nella sez. 2 sotto una diversa luce, è opportuno distinguere gli aggregati formati dall'uso combinato di due indagini nel modo seguente:

(a) *aggregato dei compresenti*, per il quale nel cap. 7 si mostrano vantaggi e svantaggi derivanti dall'uso di criteri alternativi di abbinamento. Non si esclude peraltro che a formare l'aggregato possano concorrere casi di falsi positivi o che nella determinazione dei passaggi tra condizioni siano presenti errori di classificazione. Per appurare tali eventualità non pare esistano modalità alternative a sondaggi con reinterviste.

(b1) *aggregato dei non trovati alla 2^ indagine* (soggetti per cui si hanno informazioni solo alla 1^). Esso può essere costituito perlomeno da tre sottoinsiemi:

- (i) non risposte, dovute a cadute di unità di 1^ e di 2^ stadio, per qualsiasi motivo, non oggetto di sostituzione (è questo lo specifico insieme che sarà considerato nel seguito, insieme con gli abbinati, per la stima dei flussi);
- (ii) eventuali falsi negativi, soggetti cioè esclusi dall'abbinamento per errore (essi dovrebbero trovare il corrispettivo nel successivo aggregato (b2) o, con minor probabilità, in eventuali soggetti che hanno concorso a formare un falso positivo nell' aggregato dei compresenti);
- (iii) sostituiti, cioè soggetti appartenenti al campione di famiglie, originale o di riserva, intervistate solo alla 1^ indagine, per le quali si è provveduto, per vari motivi, alla sostituzione nella 2^.

(b2) *aggregato dei non trovati alla 1^ indagine* (soggetti per cui si hanno informazioni solo alla 2^ indagine), costituito dagli stessi tre sottoinsiemi di (b1):
(i) non risposte;
(ii) falsi negativi;
(iii) sostituiti, a loro volta provenienti dal campione originale o di riserva.

Ciò che emerge da questa classificazione è che un nodo cruciale è dato dal procedimento di sostituzione, che si manifesta in due indagini contigue in quattro forme:

	indagine 1	indagine 2
1)	Xo	0
2)	Xr	0
3)	0	Xr
4)	0	Xo

dove: X = osservazione,
0 = mancata osservazione,
o = campione originale,
r = campione di riserva.

L'effetto immediato che esso produce è la duplicazione dei soggetti che per definizione, non possono essere collegati. Più in generale sembra di poter sostenere che essi non possono essere sommariamente considerati irrilevanti al fine di stimare i movimenti tra condizioni.

Il loro riconoscimento e il confronto delle loro caratteristiche socio-economiche (in rapporto anche agli abbinati) sono condizioni irrinunciabili per una scelta sul loro trattamento, che potrebbe portare alla loro totale esclusione o all'uso dei soli individui appartenenti al gruppo Xo.

Per gli altri aggregati spuri (falsi negativi), alle condizioni attuali, salvo cioè possibili verifiche da attuare con indagini suppletive, sembra ragionevole assumere che possano essere assimilati al gruppo dei compresenti, essendo plausibile l'indipendenza tra la causa di non abbinamento e le variabili di studio.

5. *Modelli per la stima dei flussi con dati mancanti*

Nel considerare soltanto il sottocampione degli abbinati, ovvero degli individui rispondenti ad entrambe le occasioni, per la costruzione della matrice dei flussi, si assume implicitamente che i non rispondenti siano un campione casuale dei rispondenti. Ciò equivale ad assumere l'indipendenza tra le non risposte e le variabili di interesse (*missing completely at random,* MCAR, secondo la definizione di Rubin, 1976).

Allo stesso modo, l'introduzione di sistemi di pesi per tener conto delle mancate risposte assume, più o meno implicitamente, ipotesi analoghe ri-

spetto al meccanismo di non risposta.

La condizione MCAR è molto forte e molto spesso poco realistica, per cui appare interessante indagare le possibilità d'analisi offerte da approcci che utilizzano, oltre alle informazioni sugli abbinati, anche quelle parziali fornite dai rispondenti ad una sola occasione.

Lo sfruttamento di tutti i dati disponibili porta, in generale, a stime più efficienti e inoltre può permettere il controllo delle relazioni tra i meccanismi che generano la non risposta e le variabili oggetto di studio.

Limitando l'attenzione, per esempio, all'usuale sistema a tre stati per la classificazione dello stato occupazionale degli intervistati - occupato (OCC), in cerca di occupazione (INC) e non forza lavoro (NFL) - i dati disponibili per la stima dei flussi possono essere organizzati come in Tab. 1, dove:

X_{ij} , i,j = 1,2,3: individui nello stato i in t-1 e nello stato j in t;
R_i , i = 1,2,3: individui nello stato i in t-1 non rispondenti in t;
C_j , j = 1,2,3: individui nello stato j in t non rispondenti in t-1.

Tab. 1: *Dati osservati per la stima dei flussi*

	t			
t-1	OCC	INC	NFL	Supplementi di riga
OCC				
INC		X_{ij}		R_i
NFL				
Suppl. di colonna		C_j		

Utilizzare anche le informazioni contenute in R_i e C_j significa, sostanzialmente, formulare un modello per la riallocazione dei valori dei supplementi di riga e colonna entro la matrice 3x3, tenendo conto anche della natura del fenomeno che le ha prodotte.

Il problema può essere adeguatamente impostato nell' ambito della teoria per la stima a massima verosimiglianza dei valori delle celle di una tabella di contingenza con dati parzialmente classificati (Chen e Fienberg, 1974). In generale, le procedure di stima si basano su algoritmi iterativi del tipo EM (Dempster, Laird e Rubin, 1977).

Anche la classe dei modelli loglineari gerarchici può essere impiegata per l' analisi della distribuzione congiunta delle variabili categoriali (stato occupazionale nei due momenti) e della variabile indicatrice della non risposta (Fay, 1986; Little, 1985; Little e Rubin, 1987). Applicazioni di tali metodi per la stima dei flussi dai dati di panel della CPS e della *Canadian Labour Force Survey* si trovano in Stasny e Fienberg (1985) e Stasny (1986, 1987, e 1988).

La classe di modelli proposta da Stasny (1988), che estende l'impianto metodologico di Chen e Fienberg (1974), esplicita in modo diretto la relazione tra le variabili di studio e i meccanismi di non risposta. Tale approccio considera i dati osservati di Tab. 1 come la realizzazione finale di un processo a due stadi.

Nel primo stadio, non osservato, gli individui sono allocati nelle 9 celle della matrice secondo uno schema multinomiale con probabilità ω_{ij}.

Nel secondo stadio ogni osservazione della tavola può perdere la propria classificazione di riga con probabilità φ_{ij} (non risposta in t-1) o la propria classificazione di colonna con probabilità ψ_{ij} (non risposta in t).

Assumendo pari a zero la probabilità di non risposta in entrambe le occasioni[4] e che tutte le non risposte debbano essere classificate entro le 9 celle, dopo il secondo stadio sono osservati appunto i dati in Tab. 1. La probabilità di osservare un individuo nella cella (i,j) è perciò data da $(1 - \varphi_{ij} - \psi_{ij})$, come illustrato in Tab. 2.

Tab. 2: *Probabilità per i dati osservati*

	t			
t-1	OCC	INC	NFL	Supplementi di riga
OCC				
INC		$(1-\varphi_{ij}-\psi_{ij})\,\omega_{ij}$		$\Sigma_i\,\omega_{ij}\,\psi_{ij}$
NFL				
Suppl. di colonna		$\Sigma_j\,\omega_{ij}\,\varphi_{ij}$		

Indicando con m_{ij} le numerosità attese secondo l'allocazione multinomiale, la funzione di verosimiglianza per i dati osservati risulta proporzionale a:

$$\Pi_i \Pi_j \left(\left(1 - \varphi_{ij}\,\psi_{ij} \right) m_{ij} \right)^{X_{ij}} \left(\Pi_i \left(\Sigma_j\,\psi_{ij}\,m_{ij} \right)^{R_i} \right) \left(\Pi_j \left(\Sigma_i\,\varphi_{ij}\,m_{ij} \right)^{C_j} \right) \qquad (1)$$

con i,j = 1,2,3.

Le 15 osservazioni disponibili sono insufficienti per la stima dei 27 parametri ignoti contenuti nella (1). E' necessario, perciò, ridurre il numero dei parametri formulando dei modelli parsimoniosi per le probabilità di non risposta φ_{ij} e ψ_{ij}. Intuitivamente, tali probabilità possono dipendere dalla condizione occupazionale nelle due occasioni e/o dal momento stesso in cui è effettuata l' indagine. Stasny (1986 e 1988) propone 3 specificazioni:

A: $\varphi_{ij} = \lambda_{t-1}$ $\quad \psi_{ij} = \lambda_t$ $\quad$ g.d.l. = 4,
B: $\varphi_{ij} = \lambda_j$ $\quad \psi_{ij} = \lambda_i$ $\quad$ g.d.l. = 3,
C: $\varphi_{ij} = \lambda_i$ $\quad \psi_{ij} = \lambda_j$ $\quad$ g.d.l. = 3.

4 L'assunzione può essere rilasciata qualora si rendano disponibili informazioni anche su questi soggetti.

In A, la probabilità di non risposta ad una occasione dipende solo dal mese di osservazione, in B dipende dallo stato occupazionale nell'occasione in cui l'individuo viene intervistato, mentre in C si ipotizza la dipendenza rispetto allo stato occupazionale (ignoto) nell'occasione in cui risulta *missing*.

I modelli A e B specificano la relazione rispetto alla non risposta solo in termini delle variabili osservate. Il meccanismo di non risposta può perciò essere "ignorato" (*missing at random,* MAR, nel senso di Rubin, 1976): condizionatamente allo stato occupazionale in t-1 si assume che i rispondenti e i non rispondenti abbiano lo stesso comportamento rispetto al mercato del lavoro al tempo t e viceversa. Tale assunzione implica che le stime dei flussi dai due modelli siano identiche. Ciò deriva essenzialmente dalla possibilità di separare la funzione di verosimiglianza in (1), per entrambe le formulazioni, nel prodotto di due fattori di cui il primo dipendente dai soli parametri m_{ij} e il secondo dai soli λ_t o λ_i. Le due quantità possono perciò essere massimizzate autonomamente (vedi Stasny, 1985).

Se l'ipotesi MAR può essere ritenuta plausibile per i dati osservati, pur fornendo i modelli A e B la stessa stima dei flussi di interesse, le *performances* dei due modelli in termini di adattamento ai dati osservati (compresi R_i e C_j) permettono di indagare più a fondo l'insieme delle cause che possono provocare i *missing*. Tali conoscenze possono essere adeguatamente impiegate per migliorare o controllare la rilevazione sul campo.

Il meccanismo di non risposta espresso nella formulazione C è, invece, "non ignorabile", in quanto viene ipotizzata una relazione di dipendenza tra la non risposta ad una occasione e la condizione occupazionale di quella occasione. In questo caso la funzione (1) non è più separabile e le stime dei parametri devono essere trovate simultaneamente ricorrendo a tecniche di massimizzazione.

Nel seguito è illustrato un esempio di applicazione dei 3 modelli a 5 matrici trimestrali per il Veneto (rilevazioni del 1985 e le prime due del 1986) e 2 per la Lombardia (le prime due rilevazioni del 1985 e del 1986), abbinate secondo la procedura LINK (vedi il cap. 7).

6. *Un esempio di utilizzazione delle non risposte per la stima dei flussi*

L'applicazione dei modelli descritti nella sez. 5 ai dati RTFL è stata condotta principalmente a scopo illustrativo al fine di confrontare le tre formulazioni, piuttosto che con l' obiettivo di fornire le stime dei flussi tra condizioni per le due regioni nei periodi indicati.

È il caso infatti di ricordare che le informazioni R_i e C_j (supplementi di riga e colonna) fornite dalla procedura LINK non rappresentano i reali casi di non abbinamento per non risposta (NR) ad una delle due occasioni (che in teoria dovrebbero essere noti prima dell'abbinamento e scartati dai *files* degli abbinabili), bensì in totale i casi non abbinati per i motivi più vari (errori

di trascrizione nei codici, codifica, ecc.)[5]. Sulla base delle informazioni attuali, riconoscere e depurare i casi di sostituzione in fase di abbinamento risulta alquanto complesso. La frazione di non abbinamenti imputabile a tale circostanza, sul totale dei non abbinati, è stimata intorno al 40% circa (Giusti, Marliani, Torelli, 1987). Si sono perciò ridotti i valori relativi alle non risposte di tale quota. Il criterio, che ovviamente non altera la distribuzione rispetto alla stato occupazionale, dovrebbe, almeno, riportare il fenomeno alle dimensioni originali. Le stime dei parametri λ e i valori della statistica χ^2 sono riportati in Tab. 3, mentre i dati utilizzati e le stime dei flussi (solo per 1985.I-II) sono riportati in Tab. 4.

Nel Veneto il modello A, in cui la probabilità di non risposta dipende dal momento in cui è effettuata l'indagine, presenta i valori più bassi di χ^2, tranne nel primo trimestre 1986. Al contrario, per la Lombardia l'adattamento degli altri due modelli, in particolare del modello B, sembra senz'altro migliore.

Le stime fornite da B e C risultano sostanzialmente simili, sia rispetto ai valori dei parametri λ, sia rispetto alle stime dei flussi. Nel modello C, il valore di λ_{inc} associato alle persone in cerca di occupazione appare generalmente più elevato, oltre che rispetto ai parametri relativi agli altri stati, anche al corrispondente in B, e assegna qualche individuo in più tra le persone che rimangono in cerca di occupazione nei due periodi.

La scelta del modello più appropriato, solo sulla scorta delle stime ottenute e del limitato periodo di osservazione, in particolare per la Lombardia, non è immediata. Di difficile interpretazione è, per esempio, l'apparente divario tra le due regioni rispetto alle cause di generazione delle non risposte.

Tab. 3: *Stime dei parametri di modelli sulle probabilità di non risposta*

Regione e occasione (iniziale)	Mod. A			Mod. B				Mod. C			
	$\hat{\lambda}_{t-1}$	$\hat{\lambda}_{t}$	χ^2_4	$\hat{\lambda}_{occ}$	$\hat{\lambda}_{inc}$	$\hat{\lambda}_{nfl}$	χ^2_3	$\hat{\lambda}_{occ}$	$\hat{\lambda}_{inc}$	$\hat{\lambda}_{nfl}$	χ^2_3
Veneto											
I	0,084	0,089	1,59	0,085	0,093	0,087	1,89	0,085	0,095	0,087	1,93
85 II	0,101	0,113	4,61	0,104	0,128	0,108	4,98	0,103	0,138	0,107	5,54
III	0,068	0,071	1,07	0,068	0,064	0,070	1,14	0,068	0,062	0,070	1,23
IV	0,108	0,115	2,10	0,114	0,117	0,109	2,09	0,115	0,120	0,108	2,14
86 I	0,084	0,089	10,71	0,092	0,111	0,080	0,99	0,092	0,120	0,079	1,59
Lombardia											
85 I	0,123	0,125	14,57	0,126	0,138	0,120	3,47	0,126	0,142	0,119	4,02
86 I	0,094	0,098	11,32	0,094	0,114	0,097	3,83	0,094	0,119	0,097	4,69

5 Nell'ipotesi di usare tali informazioni come *proxy* per le non risposte, il fenomeno di sostituzione agisce in pratica raddoppiando (o quasi) il numero di casi mancanti.

Tab. 4: *Dati osservati e stima dei flussi con modelli per le non risposte*

	Veneto 85.I-II				Lombardia 85.I-II			
	OCC	INC	NFL	NR	OCC	INC	NFL	NR
OCC	2.298	19	100	254	9.929	71	323	1.747
	95,1	*0,8*	*4,1*		*96,2*	*0,7*	*3,1*	
INC	41	169	34	25	186	573	100	152
	16,8	*69,3*	*13,9*		*21,7*	*66,7*	*11,7*	
NFL	76	42	2.514	288	375	119	10.199	1.727
	2,9	*1,6*	*95,5*		*3,5*	*1,1*	*95,4*	
NR	241	29	266		1.775	155	1.631	
	Mod.A/B				Mod.A/B			
OCC	2.769	23	121		13.290	98	427	
	95,1	*0,8*	*4,2*		*96.2*	*0,7*	*3,1*	
INC	49	208	41		100	250	792	
	16,4	*69,8*	*13,8*		*21,3*	*67,4*	*11,3*	
NFL	92	52	3.042		499	163	13.410	
	2,9	*1,6*	*95,5*		*3,5*	*1,2*	*95,3*	
	Mod. C				Mod. C			
OCC	2.768	23	121		13.291	98	427	
	95,1	*0,8*	*4,2*		*96,2*	*0,7*	*3,1*	
INC	49	208	41		133	252	801	
	16,4	*69,9*	*13,7*		*21,2*	*67,5*	*11,3*	
NFL	92	52	3.042		499	164	13.397	
	2,9	*1,6*	*95,5*		*3,5*	*1,2*	*95,3*	

D'altra parte, la redistribuzione dei casi con informazioni parziali lascia sostanzialmente immutati i tassi di transizione tra stati calcolati sui soli dati abbinati (Tab. 4). Tale risultato, tipicamente italiano e non riscontrato nelle applicazioni in altri contesti, invita ad una ulteriore riflessione sulla natura dei dati utilizzati. Le distribuzioni per condizione occupazionale degli abbinati e dei non abbinati non paiono, infatti, così diverse da poter escludere del tutto l' ipotesi di completa casualità (MCAR) per i dati utilizzati come non risposte (Tab. 5), in particolare nel caso del Veneto.

Questo risultato non è necessariamente sorprendente se si considera, ancora una volta, il concetto spurio di non risposta contenuto nelle informazioni usate e gli effetti della pratica di sostituzione nel senso di stabilizzare le distribuzioni.

Tab. 5: *Valori del χ^2_4 per l'ipotesi di indipendenza tra le distribuzioni marginali degli abbinati e dei non abbinati*

	1985				1986
Regione	I	II	III	IV	I
Veneto	1,54	4,70	1,07	2,17	10,55
Lombardia	15,33	–	–	–	12,24

Tab. 6: *Tavola sinottica (i) dei fattori di disturbo nella costruzione dei flussi, (ii) dei metodi proposti in letteratura per il loro trattamento e (iii) delle caratteristiche della RTFL e delle soluzioni previste o prevedibili*

Fattori di disturbo	Metodi proposti	RTFL	Alternative per RTFL
rotazione		schema 2-2-2: fraz. abbinati 50%	
mancate risposte	- ponderazione - imputazione/ sostituzione - approccio basato su: modelli loglin. gerarchici (Little e Rubin, 1987); modelli causali (Fay, 1986); modelli di allocazione (Stasny, 1988; Abowd e Zellner, 1985)	- sostituzione - sistema di pesi per il riporto all'universo	introduzione di un codice di avvenuta sostituzione per l'individuazione piu' precisa delle mancate risposte
errori di risposta	stime dell'errore di risposta basate su dati da reinterviste		- utilizzo anche delle informazioni longitudinali - introduzione di un programma di reinterviste
rotation group bias	analisi di Bailar (1975) e Solon (1985)	vedi cap. 9	
procedura di abbinamento	metodi ad hoc	criteri seguiti: uguaglianza di eta' (+1), sesso, relazione capofam., istruzione	- vedi cap. 7 - semplificazione del sistema di codifica del codice familiare
differenze tra saldi flussi e saldi stock	*iterative proportional fitting*	RAS	in generale: attento controllo sullo svolgimento dell' indagine

7. *Alcune considerazioni di sintesi*

Nell' affrontare il tema dello studio dei flussi delle forze di lavoro, dopo aver considerato le indicazioni presenti in letteratura sulle principali minacce alla validità dell'analisi, se ne sono esaminati gli effetti nel particolare contesto dell'indagine italiana. L'esame dei modelli per il trattamento delle non risposte ha condotto ad una applicazione i cui risultati evidenziano alcune caratteristiche peculiari della metodologia dell'indagine italiana e delle fasi di sua conduzione, che devono essere attentamente vagliate prima di porre in atto procedure per la stima dei flussi (vedi Tab. 6).

In particolare, pare prematuro affrontare il tema della stima dei flussi senza una attività di separazione degli aggregati disponibili (sez. 5), ottenuta o con indagini supplementari e/o con modifiche delle procedure d'indagine. Almeno in parte tale distinzione è stata recentemente resa possibile con l'inserimento di un codice di avvenuta sostituzione della famiglia, introdotto con la prima indagine del 1988, ma effettivamente utilizzabile a partire dalla prima indagine del 1989. Anche l'adozione di sistemi più semplificati per la codifica dei codici familiari può apportare notevoli contributi verso la corretta identificazione degli aggregati di riferimento. Inoltre, nell'ipotesi di sfruttamento delle informazioni parziali, la costruzione della base di dati appropriata è comunque condizionata ad una maggiore attività dell'Istat di supporto ai comuni durante lo svolgimento dell' indagine per assicurare sia una maggiore qualità delle informazioni sia minori cadute di unità di 1^ stadio.

Infine, nell'eventualità che anche per l'indagine italiana il fenomeno della non risposta ad una occasione assuma rilevanza (per dimensione e caratteristiche) ai fini delle analisi del mercato del lavoro, pare opportuno individuare l'ambito in cui muoversi per selezionare i metodi più appropriati per la riallocazione di questi individui, risolvendo contestualmente i delicati aspetti del riporto all'universo: con modelli che esplicitano le relazioni tra variabili di studio e motivi di non risposta (eventualmente specificati per gruppi omogenei di popolazione) o con procedure *ad hoc* fondate, per esempio, sul principio della compatibilità delle stime ottenute con quelle di consistenza.

19

MODELLI DI DURATA PER DATI DA INDAGINI SULLE FORZE DI LAVORO: DISOCCUPAZIONE GIOVANILE E DIPENDENZA DALLA DURATA

Nicola Torelli e Ugo Trivellato

1. *Introduzione*

Nell'arco degli ultimi due decenni, la letteratura sui metodi di analisi di segmenti di storie lavorative (e, più in generale, di storie individuali nelle scienze sociali) ha conosciuto un imponente sviluppo, ed ha connotato in senso marcatamente innovativo lo studio del comportamento dinamico nel mercato del lavoro. La storia lavorativa dell'individuo è trattata come la realizzazione di un processo stocastico a stati discreti e a tempo continuo, il cui spazio degli stati consta, almeno nella formulazione più semplice (accolta anche in questo studio), di due soli stati - disoccupazione ed occupazione -. L'apparato metodologico per lo studio della durata della disoccupazione e/o dell'occupazione è stato in larga parte mutuato dall'analisi di dati di sopravvivenza, consolidatasi nell'ambito della biostatistica e della teoria dell'affidabilità (vedi, ad es., Lawless, 1982, e Cox e Oakes, 1984). Per un altro verso, nell'analisi di segmenti di storie lavorative (e, più in generale, di storie individuali nelle scienze sociali) si è chiamati a confrontarsi con problemi in qualche misura peculiari, in parte per trattare specificazioni e restrizioni del modello stocastico suggerite dalla teoria economica, e soprattutto per tener conto dei dati longitudinali tutt'altro che ideali di cui tipicamente si dispone. Recenti rassegne di questi problemi, e degli interessanti sviluppi che hanno stimolato, sono in Heckman e Singer (1986) e Kiefer (1988).

In questo capitolo, esaminiamo le questioni salienti che si pongono per lo studio della dinamica della disoccupazione sulla base di dati individuali tratti dalla Rilevazione Trimestrale delle Forze di Lavoro (nel seguito RTFL), e presentiamo i risultati di alcune analisi empiriche sulla durata della disoccupazione giovanile.

Le opportunità di documentare la dinamica della disoccupazione risiedono in due caratteristiche della RTFL: (i) l'informazione sulla durata dell'episodio di ricerca di lavoro in corso alla data dell'indagine, informazione raccolta in ogni rilevazione tramite un quesito retrospettivo posto a tutti gli

individui disoccupati [1] (si noti che un analogo quesito, sulla durata dell'episodio di occupazione in corso, non è invece posto agli occupati); (ii) il piano di campionamento con rotazione, che consente di abbinare i dati di individui presenti in successive occasioni d'indagine (al riguardo, vedi il cap. 7).

Già a partire dai dati di una singola rilevazione, cioè utilizzando la sola informazione (i), è possibile condurre analisi di durata della disoccupazione. Naturalmente, si ha un'informazione longitudinale più ricca quando si sfrutta anche la caratteristica (ii), abbinando i dati di individui presenti in due o più indagini. In ogni caso, tuttavia, i dati longitudinali tratti dalla RTFL soffrono di due limitazioni cruciali. In primo luogo, essi sono ristretti ad un breve segmento della storia lavorativa degli individui, essenzialmente quello coperto dalla sequenza delle indagini abbinate (di massima, quindi, 15 mesi). In secondo luogo, essi documentano questo segmento in modo frammentario, sicché in generale non è possibile ricostruire la sequenza degli stati occupati nel periodo di permanenza nel campione e la durata del soggiorno negli stessi. L'incompletezza nella descrizione della storia lavorativa emerge per molteplici aspetti:

(a) La durata della disoccupazione è rilevata soltanto per gli episodi che sono in corso alla data dell'indagine, il che implica che verosimilmente i dati sono affetti da *length bias*. Il fenomeno si manifesta perché, detto semplicemente, nella maggior parte delle situazioni episodi di disoccupazione più lunghi hanno maggiore probabilità di essere interrotti da un piano di osservazione puntuale.

(b) Vi sono episodi di disoccupazione ancora in corso alla data dell'ultima occasione: parte dei dati è quindi censurata a destra (ovviamente, nel caso si utilizzino i dati di una singola rilevazione tutti gli episodi sono tali).

(c) Sfruttando la struttura longitudinale del campione, si possono osservare episodi di disoccupazione completi. Tuttavia, non si osserva l'istante preciso in cui si conclude l'episodio, e l'informazione sulla sua conclusione si ricava dal confronto delle risposte a due successive indagini. La durata degli episodi completi è perciò nota con approssimazione, essendo compresa tra t_p e $t_p + s$, dove t_p è la durata della disoccupazione rilevata alla prima indagine e s è il periodo che separa le indagini (quindi 3 o 9 mesi).

(d) L'informazione sul tempo di permanenza nello stato di disoccupazione è verosimilmente soggetta ad errori di memoria, essendo rilevata con un quesito retrospettivo.

(e) Con riguardo al segmento della storia lavorativa coperto dalle indagini abbinate, oltre alla durata dell'episodio (o degli episodi) di disoccupazione

---

1 Due precisazioni si impongono. (i) Il quesito sulla durata della ricerca di lavoro è in realtà posto a tutti coloro che hanno dichiarato di essere alla ricerca (quindi, ad es., anche agli occupati in cerca di un diverso o ulteriore lavoro). Nel seguito, restringeremo tuttavia l'attenzione sui soli disoccupati, sicché nel contesto di questo studio è equivalente parlare di 'disoccupazione' o di 'ricerca di lavoro'. (ii) Con la RTFL viene raccolta anche un'altra informazione retrospettiva sulla durata: si tratta, per i non occupati già occupati, del periodo trascorso dalla conclusione dell'ultima occupazione.

è osservata soltanto la sequenza discreta degli stati - occupazione o disoccupazione o inattività - in cui si trova l'individuo alla data di riferimento delle varie occasioni d'indagine.

(f) Quanto a variabili concomitanti che possono influire sulla dinamica della partecipazione al lavoro, è tipicamente rilevato, sempre alla data delle varie occasioni d'indagine, uno scarno insieme di caratteristiche demografiche e sociali dell'individuo e di altri componenti la famiglia. È quindi plausibile vi siano variabili rilevanti omesse, ovvero - con terminologia tipica dell'analisi di dati di durata - vi sia 'eterogeneità non osservata'.

Dopo alcuni essenziali richiami ai modelli e metodi per l'analisi di dati di sopravvivenza (sez. 2), nella sez. 3 fissiamo l'attenzione essenzialmente su tre dei problemi or ora elencati, cruciali per la costruzione di una modello di durata della disoccupazione, cioè nell'ordine: (i) la presenza di eterogeneità non osservata; (ii) la presenza di *length bias*; (iii) la scarsa accuratezza nella registrazione delle informazioni sulla durata [2] . Sulla scorta di queste riflessioni metodologiche, nella sez. 4 procediamo poi a specificare e stimare un modello di durata della disoccupazione per un campione di giovani lombardi in età 14-29 anni, utilizzando i dati abbinati della RTFL relativi alle indagini di gennaio e aprile 1986. I risultati, pur tutt'altro che conclusivi, offrono interessanti elementi per la comprensione, in un ottica dinamica, del comportamento dell'offerta di lavoro giovanile.

2. *Modelli di rischio ad uno stato: alcuni richiami essenziali*

Qui e nel seguito, consideriamo il caso più semplice, noto come *single-spell one-state model*, in cui il sistema di stati è dicotomico e i flussi si verificano in un'unica direzione (nella fattispecie, dalla disoccupazione all'occupazione). Conveniamo inoltre di denotare gli stati rispettivamente con 0 e 1, sicché le transizioni riguardano il passaggio allo stato 1. Per ogni unità campionaria si osserva il tempo t trascorso nello stato 0 fino al momento della transizione.

Il tempo t, di sopravvivenza nello stato 0, viene assimilato alla determinazione di una variabile casuale continua non negativa T, con funzione di ripartizione G(t) e funzione di densità g(t). A fianco della funzione di ripartizione o di densità, ciascuna sufficiente a definire completamente il comportamento stocastico della variabile casuale T, di particolare interesse per il loro valore interpretativo sono le seguenti funzioni, definite in relazione alle precedenti:

(a) la funzione di sopravvivenza S(t), cioè la probabilità che la transizione allo stato 1 si verifichi oltre il tempo t:

---

2 Per una più generale trattazione delle diverse questioni derivanti da uno schema di osservazione caratterizzato dalla combinazione di quesiti retrospettivi e di un campione con rotazione, rinviamo a Trivellato e Torelli (1989).

$$S(t) = Pr\,(T \geq t) = 1 - G(t) = \int_t^{\infty} g(t)\,dt; \qquad (1)$$

(b) la funzione di rischio, definita come

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t+\Delta t \mid T \geq t)}{\Delta t} = \frac{g(t)}{S(t)}. \qquad (2)$$

Palesemente, h(t) rappresenta il tasso istantaneo di passaggio allo stato 1 al tempo t, condizionatamente alla sopravvivenza nello stato 0 fino a t. In particolare, $h(t)\Delta t$ è la probabilità approssimata di transizione in $(t, t+\Delta t)$, data la sopravvivenza fino a t.

La funzione di rischio è forse la quantità che meglio illustra la struttura della dipendenza dal tempo trascorso nello stato 0. Più precisamente:

– Se $\frac{dh(t)}{dt} > 0$ per $t = t_0$,

si parla di dipendenza positiva dalla durata in t_0 .

– Se $\frac{dh(t)}{dt} < 0$ per $t = t_0$,

in t_0 è presente dipendenza negativa dalla durata.

Per alcune distribuzioni si dà che h(t) è monotona crescente (decrescente), per cui è presente dipendenza positiva (negativa) dalla durata per ogni t.

– Non si ha, invece, dipendenza dalla durata quando risulta

$$\frac{dh(t)}{dt} = 0 \text{ per ogni } t. \qquad (3)$$

Il tasso istantaneo di transizione è in tal caso indipendente da t, cioè a dire la probabilità di cambiare stato è la stessa quale che sia il tempo già trascorso nello stato 0.

Spesso è opportuno specificare la distribuzione di durata in funzione di un vettore x(t) di variabili osservate e di una componente residua non osservabile $\vartheta(t)$, che rappresenta appunto l'eterogeneità non osservata. Per il momento, trascuriamo $\vartheta(t)$ (ovvero assumiamo che tutte le variabili rilevanti siano osservate) e poniamo per semplicità x(t)=x, cioè a dire il vettore x invariante nel tempo. Si tratta dunque di fornire una specificazione di G(t|x) o, vista l'equivalenza tra le diverse quantità introdotte, della funzione di rischio condizionata

$$h(t|x) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t+\Delta t \mid T \geq t, x)}{\Delta t}. \qquad (4)$$

Una specificazione di h(t|x) frequentemente utilizzata è il modello di rischio proporzionale, nel quale l'effetto delle variabili x sulla funzione di rischio è moltiplicativo:

$$h(t|x) = h_0(t)\,\mu(\beta'x), \qquad (5)$$

dove $h_0(t)$ è la funzione di rischio di base e le covariate agiscono appunto moltiplicativamente su di essa tramite $\mu(\beta'x)$, essendo β un vettore di parametri ignoti. La forma privilegiata per $\mu(\beta'x)$ è $\exp(\beta'x)$, che tra l'altro garantisce valori non negativi per $h(t|x)$.

La stima di modelli di rischio ad uno stato può essere condotta secondo differenti approcci. In buona parte delle applicazioni nelle scienze sociali, l'approccio adottato è parametrico: si presuppone la conoscenza della funzione $h_0(t)$, o dei suoi equivalenti, a meno di un insieme di parametri (o comunque si formula un'assunzione del genere). La diffusa preferenza per l'approccio parametrico è dovuta al fatto che esso consente di affrontare, in maniera abbastanza soddisfacente, le difficoltà connesse alla presenza di dati incompleti e/o poco accurati e di eterogeneità non osservata, mentre tipicamente ciò non si dà per approcci non parametrici (del tipo stima limite-prodotto di Kaplan-Meyer) o parzialmente parametrici (quale il modello di regressione introdotto da Cox, 1972).

L'approccio parametrico è appunto quello che adottiamo in questo studio. Al riguardo, si impongono tuttavia almeno un paio di osservazioni. Innnanzitutto, è bene tener presente che la forma funzionale di $h_0(t)$ non è usualmente giustificata da stringenti suggerimenti della teoria economica, sicché la scelta della specificazione è dettata in larga parte da considerazioni empiriche - di semplicità e/o di adattamento ai dati - e il modello va comunque interpretato come una forma ridotta (vedi, ad es., Heckman e Singer, 1986). In secondo luogo, in letteratura vi sono recenti sviluppi nella direzione di specificazioni parametriche estremamente flessibili, prossime a quella parzialmente parametrica di Cox, che consentono di affrontare in maniera promettente i tipici problemi di incompletezza dei dati longitudinali disponibili nelle scienze sociali (vedi, ad es., Han e Hausman, 1989).

3. *Modelli di rischio ad uno stato per dati dalla RTFL*

3.1. *Il problema dell'eterogeneità non osservata*

Si è fin qui trascurata la presenza di variabili non osservabili, ovvero di forme di eterogeneità (non osservata) nella popolazione che abbiano qualche relazione con la distribuzione di sopravvivenza nello stato. In assenza di eterogeneità, se la forma funzionale ipotizzata per $G(\cdot)$ è plausibile, le formulazioni proposte permettono di cogliere in maniera soddisfacente sia la struttura di dipendenza dalla durata sia l'effetto delle variabili x. Quando i dati non sono ottenuti sotto il controllo sperimentale, e in particolare quando

risultano da un piano di osservazione quale quello della RTFL, è tuttavia difficile ipotizzare che tutte le variabili rilevanti siano osservate e quindi incluse tra le esplicative.

Ora, nel contesto di modelli di rischio trascurare l'eterogeneità ha notoriamente implicazioni serie. Infatti, si ha una distorsione nella stima della dipendenza dalla durata così come dei parametri β, e ciò anche se le variabili non osservate sono incorrelate con le variabili x incluse nel modello (Lancaster e Nickell, 1980)[3]. Specificamente, si mostra che la mancata considerazione dell'eterogeneità provoca una distorsione della funzione di rischio di base verso la dipendenza negativa dalla durata (vedi, ad es., Heckman e Singer, 1986, pp. 1705-1710). Questo risultato è dovuto ad un processo di selezione dinamica: in sostanza, le persone con maggior propensione alla mobilità (che presentano cioè valori elevati in una variabile, omessa, in relazione positiva con la probabilità di transitare allo stato 1) sono le prime ad uscire, sicché con l'andar del tempo tra i sopravviventi nello stato 0 restano proporzionalmente più numerose le persone con minor propensione alla mobilità, il che crea l'illusione di un tasso di uscita decrescente con t anche quando esso non è tale (o comunque di un tasso più decrescente, o meno crescente, di quanto in realtà non sia).

E' allora necessario considerare la funzione di ripartizione condizionata $G(t|x,\vartheta)$[4]. Vale la seguente equivalenza, che pone in relazione il modello con la controparte osservata $G_*(t|x)$:

$$G_*(t|x) = \int_\theta g(t|x,\vartheta)\, dH(\vartheta) , \qquad (6)$$

dove $H(\vartheta)$ è la ripartizione dell'eterogeneità non osservata e $\vartheta \in \theta \subset \Re$. È evidente che l'interesse dell'analista è rivolto alla stima di $G(t|x,\vartheta)$, o meglio, quando si suppone l'appartenenza di $G(\bullet)$ ad una famiglia di distribuzioni nota, alla stima degli ignoti parametri che la specificano completamente. Le osservazioni si riferiscono, invece, a $G_*(t|x)$. Ciò rende chiaro come il controllo dell'eterogeneità sia tutt'altro che agevole. Il problema può essere posto in termini di identificazione, chiedendosi cioè se e quando sia possibile, dalla conoscenza di $G_*(t|x)$, separare la componente $G(t|x,\vartheta)$ e l'eterogeneità (vedi, tra gli altri, Elbers e Ridder, 1982, e Heckman e Singer, 1986).

Nelle prime applicazioni di modelli di durata della disoccupazione (Lancaster, 1979; Nickell, 1979; Heckman e Borjas, 1980), la pratica corrente per il controllo dell'eterogeneità è consistita nell'assumere per $G(t|x,\vartheta)$ e $H(\vartheta)$ specifiche forme funzionali, note a meno di un numero finito di parametri. In tal caso, la scelta della famiglia parametrica per $G(t|x,\vartheta)$ e $H(\vartheta)$ viene

---

3 Le conseguenze sono dunque più severe che nel modello di regressione lineare, nel quale l'omissione di variabili induce distorsione nei coefficenti stimati di quelle incluse solo quando le variabili omesse (non osservate o non riconosciute come rilevanti per la spiegazione della variabile endogena) sono correlate con le variabili incluse.

4 Si noti che, qui e nel seguito, $\vartheta(t)$ è assunto invariante nel tempo. L'assunzione è obbligata per il piano di osservazione della RTFL, e in generale per modelli per singoli episodi (vedi Flinn e Heckman, 1982).

spesso condotta in base alla convenienza computazionale, procurando cioè di rendere valutabile la (6) in modo da evitare l'integrazione numerica (in molti casi, ciò ha portato ad adottare per $H(\vartheta)$ la specificazione della ripartizione di una variabile casuale Gamma).

Heckman e Singer (1984) hanno vigorosamente criticato questa pratica, argomentando che sovraparametrizza il modello di durata, e producendo inoltre risultati empirici che mostrano la scarsa stabilità delle stime dei parametri del modello in corrispondenza di diverse specificazioni della distribuzione dell'eterogeneità. In alternativa, essi hanno proposto una rappresentazione non parametrica di $H(\vartheta)$, per mezzo di una funzione di ripartizione a gradini con un numero I finito di punti di incremento. È allora possibile riscrivere la (6) come segue:

$$G_*(t|x) = \sum_{i=1}^{I} G(t|x,\vartheta_i)\, p_i \,, \qquad (7)$$

con $p_i = dH(\vartheta_i)$. Chiaramente, oltre che i parametri di $G(t|x,\vartheta_i)$, è necessario stimare anche i valori ϑ_i e p_i . Occorre considerare, poi, che in genere il valore I è ignoto, sicché si deve ricorrere a opportuni accorgimenti nella procedura di stima. Una procedura di stima di massima verosimiglianza iterativa, che sfrutta l'algoritmo EM (*Expectation Maximization*), è in Heckman e Singer (1984, pag. 304 e segg.). I risultati delle numerose simulazioni svolte da questi autori hanno mostrato che la procedura conduce a buone stime dei parametri di $G(t|x,\vartheta)$ (e poco importa che, in queste simulazioni, i valori $\hat{p}_i$ e $\hat{\vartheta}_i$ forniscano una stima poco accurata di $H(\vartheta)$, perché si tratta di parametri di disturbo, senza alcun interesse interpretativo). Le prove empiriche effettuate hanno mostrato inoltre, nella totalità dei casi, che sono sufficienti valori di I non elevati (<5) per produrre risultati soddisfacenti [5].

Le evidenze di Heckman e Singer sono, tuttavia, tutt'altro che conclusive. Altri lavori (ad es., Trussel e Richards, 1985, e Manton, Stallard e Vaupel, 1986) suggeriscono infatti che le stime dei parametri del modello sono tipicamente più sensibili alla specificazione della funzione di sopravvivenza che a quella dell'eterogeneità. Pertanto, salvo si abbiano ragioni *a priori* per confidare nella fondatezza della specificazione parametrica di $G(t|x,\vartheta)$, i tentativi di controllare l'eterogeneità non osservata vanno valutati con cautela ed eventualmente convalidati con opportune analisi di specificazione del modello.

3.2. *Il problema del 'length bias'*

Per indagini longitudinali in ambito economico e sociale, raramente si dà

---

5 Non va trascurato, d'altra parte, che la procedura è di laboriosa implementazione. Inoltre, è sensibile ai valori iniziali e non ne è garantita la convergenza ad un massimo globale. Ciò suggerisce ovvie cautele nella sua applicazione.

la situazione ideale in cui le osservazioni cominciano in coincidenza con la data d'inizio della permanenza in uno stato. Comunemente, invece, i dati sono riferiti ad un campione di soggetti che alla prima intervista sono presenti da un certo periodo nello stato. Come già abbiamo osservato, tale è appunto il campione delle persone in cerca di lavoro della RTFL. Un siffatto campione è affetto dal cosiddetto *length bias problem*: nei casi di più rilevante interesse applicativo, il piano di osservazione tende a favorire l'ingresso nel campione di coloro che hanno episodi di disoccupazione più lunghi.

Indichiamo con 0 il momento dell'inizio dell'osservazione, con t_p l'istante di tempo (misurato rispetto a 0) in cui ha avuto inizio l'episodio di permanenza nello stato 0, con t_d l'istante di transizione all'altro stato (sempre misurato rispetto a 0)[6]. Le due situazioni rilevanti per il piano di osservazione della RTFL sono le seguenti:

(a) Si osserva t_p e non t_d . È questo il caso di dati rilevati in una singola indagine *cross-section*. Specificamente, corrisponde ai dati di durata della disoccupazione tratti da una singola RTFL.

(b) Si osservano t_p e t_d . È questo il caso di un'indagine longitudinale che, a partire dal campione dei presenti nello stato all'inizio dell' osservazione da t_p unità di tempo, permette poi di rilevare il tempo t_d in cui transitano all'altro stato. Corrisponde pertanto ai dati longitudinali sulla durata della disoccupazione tratti dalla RTFL sfruttando la struttura rotante del campione (a meno delle complicazioni, che per ora trascuriamo, di presenza di censura a destra o alternativamente di misura approssimata della durata completa).

A partire da g(•), funzione di densità degli episodi nella popolazione, e non solo di quelli osservati in quanto sopravviventi all'istante 0 (o equivalentemente da G(•) o h(•) o S(•)), si tratta dunque di determinare la funzione di densità delle quantità osservate nei due casi in questione.

Nel primo caso, la densità di un episodio osservato in t=0 di lunghezza t_p, diciamola $f'(t_p)$, può essere ottenuta come rapporto tra la proporzione di individui sopravviventi entrati nello stato t_p periodi prima e il totale degli episodi non conclusi all'istante di osservazione. Più precisamente, il totale dei sopravviventi al tempo t=0 è pari a

$$P_0 = \int_0^\infty k\,(1-G(t_p))\,dt_p, \tag{8}$$

dove k è la quota, che per semplicità è assunta costante nel tempo, di episodi iniziati t_p periodi prima. Quindi, è

$$f'(t_p) = \frac{k\,(1-G(t_p))}{P_0}. \tag{9}$$

---

6 I deponenti p e d stanno evidentemente per 'prima' e 'dopo' l'istante di osservazione iniziale 0.

Dall'integrazione per parti di P_0 e definendo $M = \int_0^\infty t_p\, g(t_p)\, dt_p$, si ha:

$$f'(t_p) = \frac{1-G(t_p)}{M}. \qquad (10)$$

La densità degli episodi osservati in t=0, dunque, è in generale diversa dalla densità g(•). È facile verificare che solo quando g(•) è esponenziale si verifica la coincidenza con f'(•). Infatti, se per ipotesi è

$$f'(t_p) = \frac{1-G(t_p)}{M} = g(t_p),$$

si ha che la funzione di rischio

$$h(t_p) = \frac{g(t_p)}{1-G(t_p)} = M$$

è costante, sicché g(•) è appunto la densità dell'esponenziale.

Nel secondo caso, si tratta di ottenere la densità. di $t^* = t_p + t_d$, diciamola f''(t*). A tale scopo, conviene dapprima definire la densità condizionata

$$g(t^*|\, t_p) = \frac{g(t^*)}{1-G(t_p)}. \qquad (11)$$

Sfruttando la (11), si ottiene

$$f''(t^*) = \int_0^{t^*} g(t^*|t_p)\, f'(t_p)\, dt_p = \int_0^{t^*} \frac{g(t^*)}{M}\, dt_p = \frac{t^* g(t^*)}{M}. \qquad (12)$$

Si può ancora notare che

$$E(t^*) = \int_0^\infty t^*\, f''(t^*)\, dt^* = \frac{1}{M}\int_0^\infty t^{*2}\, g(t^*)\, dt^* =$$

$$= \frac{\sigma^2 + M^2}{M} = M\left(1+\frac{\sigma^2}{M^2}\right),$$

dove σ^2 è la varianza di t* . Pertanto, la media delle durate complete degli episodi osservati in t=0 è superiore alla media delle durate complete nella popolazione (è questo il caso in cui propriamente si parla di *length bias*).

Tali risultati sono difficilmente generalizzabili quando si rinuncia all'ipotesi di tasso costante di ingresso nello stato o si introduce l'eterogeneità osservata e/o non osservata nel modello. Alcune parziali estensioni fornite da Heckman e Singer (1986) mostrano ulteriormente la rilevanza del proble-

ma, mentre una sua discussione per dati longitudinali tratti da indagini con campionamento ruotato è in Trivellato e Torelli (1989).

In particolare, v'è ancora da osservare che la presenza di *length bias* rende difficile soprattutto l'analisi di durata a partire dai dati di una singola rilevazione. Muovendo da precise assunzioni sulla famiglia parametrica per $g(\bullet)$, si può sì ricorrere a opportune correzioni, sulla scorta della (10). È questo, in sostanza, l'approccio seguito da Nickell (1979), e con riguardo ad un campione tratto proprio della rilevazione italiana sulle forze di lavoro da Flinn (1986). Con riferimento alla (10), tuttavia, non è agevole l'introduzione diretta di variabili esplicative nel modello[7], e ancor più è problematico considerare forme generali di eterogeneità. In quanto possibile, e per la RTFL ciò è consentito appunto dalla struttura rotante del campione, conviene dunque cercare di disporre di dati su t_p e t_d e ricorrere alla densità condizionata (11).

3.3. *L'inaccuratezza della durata riportata della disoccupazione*

Un ulteriore problema che si incontra nell'analisi della durata della disoccupazione riguarda la scarsa accuratezza con cui nella RTFL, rispondendo ad un quesito retrospettivo, gli intervistati riportano la durata della permanenza nello stato.

Il fenomeno è stato diffusamente analizzato nel cap. 10, per un campione dell'intera popolazione di disoccupati. Le evidenze salienti di quell'analisi possono essere così sintetizzate: (i) vi è una forte polarizzazione delle risposte intorno al numero di mesi corrispondente all'anno, all'anno e mezzo, ai due anni e così via (effetto *heaping*, o ammucchiamento che dir si voglia); (ii) per coloro i quali, in due indagini successive, sperimentano un episodio non interrotto di ricerca di lavoro, vi è una consistente quota di incoerenze fra le durate rilevate nelle due occasioni. Queste incoerenze sono dovute essenzialmente ad un effetto telescopio in avanti, che porta a collocare l'evento 'inizio della ricerca di lavoro' ad una data più prossima a quella dell'indagine, e ancora all'effetto *heaping*.

Analoghe evidenze si hanno, in verità, per durate rilevate con quesiti retrospettivi in svariati tipi di indagini: con specifico riguardo alla durata della disoccupazione e/o dell'occupazione, tanto in indagini con campione ruotato (Bowers e Horvath, 1984), quanto in indagini genuinamente longitudinali (seppur, in tal caso, con un *pattern* delle inaccuratezze in parte diverso: vedi, ad es., Hill, 1988). Si è dunque in presenza di un fenomeno di indole generale, che chiama in causa l'influenza del processo di memoria e produce diffuse e sistematiche inaccuratezze nelle durate riportate.

In via di principio, si potrebbe sviluppare un modello di misura, attinente al processo di memoria, e combinarlo col modello di interesse, sulla durata

---

7 Se il campione è sufficientemente numeroso, si preferisce utilizzare le variabili esplicative per stratificarlo, e condurre quindi l'analisi distintamente sui diversi sub-campioni. Vedi, ad es., Flinn (1986).

della disoccupazione. È questa una linea di ricerca promettente, ma ardua. In questa sede, ci limitiamo a considerare più modeste correzioni *ad hoc*, che sono state usate o proposte in letteratura soprattutto per controllare l'effetto *heaping* (che peraltro dà conto anche, almeno in parte, di altre incoerenze), o comunque per verificare se esso induce rilevanti distorsioni nelle stime dei parametri del modello di interesse. Alcune di queste correzioni sono le seguenti:

(a) uso di variabili *dummy*, per catturare l'effetto ammucchiamento su certi valori della distribuzione di frequenza delle durate (Hujer e Schneider, 1989);
(b) sostituzione delle durate rilevate con intervalli più ampi, scelti in modo da scontare l'entità degli errori di misura presenti nei dati (Little, 1988);
(c) lisciamento della distribuzione di frequenza delle durate attorno ai picchi indotti dall'*heaping*, mediante una procedura che redistribuisce ragionevolmente le risposte inaccurate per tale motivo (Torelli e Trivellato, 1989).

4. *Un modello di durata della disoccupazione giovanile*

4.1. *Il campione e le variabili*

Ci proponiamo ora di utilizzare l'apparato metodologico appena tratteggiato per analizzare la durata della disoccupazione giovanile. Specificamente, siamo interessati a stabilire se e in che modo la probabilità di transitare allo stato di occupazione dipende dalla lunghezza del periodo di ricerca, una volta che si sia tenuto conto dell'eterogeneità osservata (cioè a dire, di un insieme di variabili esplicative) e eventualmente di quella non osservata.

La questione è di rilievo per la comprensione delle dinamiche del mercato del lavoro, e cruciale a fini di politiche. Val quindi la pena di chiarirne, seppur sinteticamente, i termini. L'interrogativo essenziale può essere così enunciato: *ceteris paribus*, l'esperienza della disoccupazione agisce nel senso di ridurre le opportunità di trovare un lavoro, e quindi la probabilità di transitare dalla disoccupazione all'occupazione diminuisce a causa del prolungarsi dell'episodio di disoccupazione? Se ciò accade, in altri termini se vi è dipendenza negativa dalla durata, è giustificato il ricorso a misure di politica economica che, favorendo il contatto anche occasionale con il mercato del lavoro, sottraggano i giovani a lunghi periodi di disoccupazione. All'opposto, se vi è una dipendenza positiva dalla durata risultano corroborati gli schemi interpretativi della *job-search theory*, secondo i quali il tempo di ricerca serve per accumulare utili informazioni sul mercato del lavoro, sicché non vi è motivo per un intervento dell'operatore pubblico in materia.

Cerchiamo di rispondere a questo interrogativo adottando un approccio, quello della specificazione e stima di una funzione di rischio, che può essere definito in forma ridotta. In altre parole, non ricorriamo ad una specificazione strutturale del processo di ricerca di lavoro desumibile dalla teoria economica (come, ad, es., in Narendranathan e Nickell, 1985): specificazione discutibile

per le assunzioni estremamente restrittive che imporrebbe, e per di più impraticabile date le caratteristiche e le limitazioni dei dati traibili dalla RTFL (in assenza della cruciale informazione sui salari accettati, il modello strutturale non è infatti identificato). Piuttosto, la teoria economica viene utilizzata *ex-post* per cercare di interpretare i coefficienti stimati del modello. Specificamente, l'effetto della durata della disoccupazione sulla funzione di rischio può essere considerato come la risultante di due effetti contrastanti: l'uno, positivo, sulla probabilità di accettare un'offerta di lavoro (attraverso una riduzione del salario di accettazione); l'altro, negativo, sulla probabilità di ricevere un'offerta di lavoro (a seguito di un deterioramento del capitale umano o perché i datori di lavoro interpretano la durata della disoccupazione come indicatore, inverso, di potenziale produttività).

Il campione per le nostre analisi è tratto dal *file* di dati abbinati della RTFL per la Lombardia, relativo alle indagini di gennaio e aprile 1986 (per la procedura di abbinamento, vedi il cap. 7). Il campione è costituito dagli individui in età compresa fra i 14 ed i 29 anni, che alla prima rilevazione risultavano disoccupati.

Le ragioni che hanno indotto a condurre l'analisi su questo sottoinsieme dei disoccupati sono sostanzialmente due: (i) da un punto di vista sostanziale, lo studio della disoccupazione giovanile riveste particolare interesse, perché in Italia, e in misura del tutto analoga in Lombardia, vi è una abnorme polarizzazione dei rischi di disoccupazione sui giovani (per scarne, ma illuminanti evidenze in tal senso, vedi Trivellato, 1990); (ii) d'altra parte, l'adozione di criteri piuttosto stringenti per l'individuazione del campione - quanto ad area geografica ed a classe di età - garantisce che esso è ragionevolmente omogeneo (ad es., in termini di precedente storia lavorativa si tratta di persone in larga maggioranza in cerca di prima occupazione), condizione questa essenziale per un appropriato impiego dei metodi di analisi della durata con dati *length biased*.

Il campione consta di 678 individui: 267 maschi e 411 femmine. Si dispone della durata della ricerca, in mesi, riportata alla prima intervista. Considerata la seconda rilevazione, si può valutare se l'episodio di ricerca si è o meno concluso nell'intervallo di tre mesi che la separa dalla prima. Il campione comprende quindi individui con:

- episodi di disoccupazione completi. Ciò si dà per quanti transitano all'occupazione nell'intervallo fra le due indagini, cioè a dire: coloro che hanno dichiarato di essere occupati alla seconda occasione; coloro che, pur essendo disoccupati alla seconda occasione, dichiarano di essere in cerca di lavoro da meno di quattro mesi;
- episodi di disoccupazione in corso, censurati a destra. Ciò si dà per quanti, alla seconda occasione, sono disoccupati e riportano una durata della ricerca di lavoro pari a quattro mesi o più.

Le variabili esplicative considerate sono descritte nella Tab. 1. Come si nota, si utilizzano prevalentemente variabili *dummy*, per evitare di imporre la linearità, o comunque una specifica forma funzionale, al *pattern* con cui influenzano la durata. In larga parte, si tratta di variabili che attengono a caratteristiche personali e familiari. Sono relative a caratteristiche personali le variabili ETÀ1, ETÀ2, ISTR, FORM e SPOS. L'ambiente familiare è de-

scritto, piuttosto grossolanamente invero, dalle variabili ISCF, DISF, OCCF e DIMF. Una variabile, GOCC, coglie la precedente esperienza lavorativa dell'individuo. La sola variabile quantitativa è il tasso di disoccupazione della provincia di residenza (DISL), che si propone di catturare, sia pure in maniera rudimentale, le condizioni della domanda di lavoro locale. Tutte le variabili sono riferite alla data della prima indagine.

Palesemente, vi sono diverse variabili ipoteticamente rilevanti che mancano, o sono rappresentate da *proxies* piuttosto grezze. L'ovvia spiegazione è nelle limitazioni delle informazioni raccolte con la RTFL, che non rileva alcunché sul reddito (né, direttamente, sull'ambiente familiare). D'altronde, l'assenza di talune variabili può forse rivelarsi meno cruciale, una volta che si tenga conto degli stringenti criteri adottati per la selezione del campione. Infine, una variabile chiave, il sesso, non compare fra le esplicative semplicemente perché l'analisi è svolta distintamente per maschi e femmine.

Tab. 1: *Caratteristiche demografiche e sociali del campione accoppiato di disoccupati in età 14-29 anni per sesso. Lombardia 1986.I*

		Maschi (N=267)	Femmine (N= 411)
Variabili 'dummy'		%	%
ETÀ1	(1 se in età 20-24 anni)	33,1	38,1
ETÀ2	(1 se in età 25-29 anni)	15,7	19,0
ISTR	(1 se con titolo sc. sup.)	30,6	34,5
FORM	(1 se in attività formative)	8,2	9,2
GOCC	(1 se già occupato)	28,0	17,8
SPOS	(1 se sposato)	3,0	20,0
ISCF	(1 se istr. capofam. > 8 anni)	35,4	33,2
DISF	(1 se disoccupati in fam. > 1)	24,2	18,3
OCCF	(1 se nessuno occupato in fam.)	18,3	11,3
DIMF	(1 se dimensione fam. > 3)	60,1	58,6
Variabili continue		*media (dev. st.)*	*media (dev. st.)*
DISL	(tasso di disoccup. locale)	7,8 (1,1)	7,8 (1,1)
DURA	(durata della disoccupazione)		
	- episodi completi	14,2 (15,8)	15,3 (12,3)
	- episodi in corso	16,7 (15,1)	19,8 (15,7)
	- totale	16,1 (15,3)	19,0 (15,3)

4.2. *Specificazioni alternative e risultati*

Una sensata applicazione del modello di rischio ad uno stato al campione di dati in questione richiede che siano affrontate le questioni tratteggiate nella sez. 3.

Quanto alla scrittura di una funzione di verosimiglianza appropriata ai dati di durata di cui si dispone, conviene muovere dalla densità condizionata (11) (vedi anche Lancaster, 1979). La verosimiglianza del campione risulta quindi

$$L = \prod_{i=1}^{NO} \frac{G(t_i + 3 \mid x_i) - G(t_i \mid x_i)}{1 - G(t_i \mid x_i)} \prod_{j=1}^{ND} \frac{1 - G(t_j + 3 \mid x_j)}{1 - G(t_j \mid x_j)}, \tag{13}$$

dove:

- t_i : durata riportata della disoccupazione (in mesi), alla prima intervista, per l'individuo i-esimo;
- NO: numero di individui con episodio di disoccupazione completo, cioè con durata compresa tra t_i e $t_i + 3$;
- ND: numero di individui con episodio di disoccupazione censurato a destra, che sono cioè ancora alla ricerca di occupazione alla seconda occasione.

Si noti che la (13) non solo fornisce la corretta verosimiglianza per un campione di durate *length biased*, ma risponde anche a due ulteriori complicazioni nei dati di durata tratti dalla RTFL, che sinora avevamo lasciato nell'ombra: la censura a destra o alternativamente la misura approssimata degli episodi completi.

Per G(t|x), adottiamo il modello di rischio proporzionale (5). Assumiamo inoltre

$$\mu(\beta'x) = \exp(\beta'x) \tag{14}$$

e una funzione di rischio di base Weibull con parametro di forma $\alpha>0$:

$$h_0(t) = \alpha t^{\alpha-1}. \tag{15}$$

Questa specificazione parametrica, seppur non sostenuta da stringenti argomentazioni, è la più frequente in applicazioni alla ricerca di lavoro, per la sua immediata interpretabilità e per la semplicità computazionale. Val la pena di ricordare che vi è una dipendenza positiva dalla durata per $\alpha>1$, dipendenza negativa dalla durata per $\alpha<1$, mentre non vi è dipendenza dalla durata per $\alpha=1$ (in tal caso, la funzione di rischio di base si riduce a quella di un'esponenziale). È da segnalare inoltre che per la distribuzione Weibull il modello di durata è, sotto condizioni generali, identificato (Heckman e Singer, 1986).

È infine immediato ottenere la seguente espressione per la funzione di sopravvivenza del modello rischio proporzionale (14)-(15):

$$1 - G(t|x) = \exp\left[-\int_0^t h(u|x)\,du\right] = \exp\left[-\mu(\beta'x)\int_0^t h_0(u)\,du\right]$$
$$= \exp\left[-\exp(\beta'x)\,t^{\alpha}\right]. \qquad (16)$$

Tab. 2: *Stime di massima verosimiglianza di modelli Weibull di rischio proporzionale della durata della disoccupazione, per maschi e femmine in età 14-29 anni. Lombardia 1986.I-II* [(a)]

	Maschi (N=267)			Femmine (N=411)		
	(A)	(B)	(C)	(A)	(B)	(C)
DURA (α)	0,73** (0,16)	0,63** (0,16)	0,64** (0,16)	0,50** (0,16)	0,53** (0,15)	0,57** (0,15)
Costante	0,63 (1,02)	0,81 (1,37)	0,80 (1,35)	-0,31 (1,52)	-0,30 (1,28)	-1,08* (0,64)
ETÀ1	0,10 (0,31)	0,07 (0,31)	-	0,45 (0,29)	0,50* (0,28)	0,38 (0,27)
ETÀ2	0,43 (0,46)	0,39 (0,44)	0,27 (0,34)	0,83** (0,43)	0,98** (0,40)	0,82** (0,37)
STR	-0,26 (0,33)	-0,27 (0,34)	-0,31 (0,29)	-0,24 (0,28)	-0,19 (0,28)	-
FORM	-0,08 (0,55)	-0,03 (0,49)	-	-2,16** (1,02)	-2,19** (1,01)	-2,16** (1,01)
GOCC	0,13 (0,33)	0,09 (0,33)	-	-0,39 (0,33)	-0,33 (0,33)	-
SPOS	-0,42 (0,81)	-0,34 (0,81)	-	-1,71** (0,54)	-1,63** (0,51)	-1,53** (0,48)
ISCF	-0,15 (0,29)	-0,13 (0,30)	-	0,01 (0,29)	-0,03 (0,29)	-
DISF	-0,10 (0,39)	-	-	0,07 (0,32)	-	-
OCCF	-0,46 (0,38)	-	-	0,37 (0,37)	-	-
DIMF	-0,15 (0,26)	-	-	-0,38 (0,29)	-	-
DISL	-0,22* (0,11)	-0,23 (0,14)	-0,23 (0,14)	-0,01 (0,09)	-0,06 (0,14)	-
log veros.	-137,15	-141,54	-141,85	-170,37	-172,44	-173,28

(a) Errori standard asintotici tra parentesi. **: significativo al livello 5%; *: significativo al livello 10% (test asintotico unilaterale, $H_0 : \alpha=1$ oppure $H_0 : \beta_3=0$).

La Tab. 2 presenta le stime di massima verosimiglianza di alcune varianti del modello Weibull di rischio proporzionale, senza eterogeneità (non osservata): da una specificazione comprensiva di tutte le variabili esplicative a formulazioni via via più parsimoniose. Vi è una chiara evidenza di dipendenza negativa dalla durata tanto per i maschi che per le femmine ($\hat{\alpha}$ è significativamente inferiore a 1 in tutte le specificazioni), e il fenomeno è leggermente più marcato per le femmine.

Un più attento esame dei risultati segnala ulteriori apprezzabili differenze fra maschi e femmine. Per il campione di giovani disoccupati, tutte le variabili esplicative della specificazione iniziale sono non significative, salvo il tasso di disoccupazione locale, il cui parametro è ai margini della regione di accettazione (ed ha l'atteso segno negativo). E questi riscontri non cambiano quando si adottano specificazioni più parsimoniose. Per le giovani donne disoccupate, invece, almeno tre variabili concorrono in modo significativo a spostare la funzione di rischio: ETÀ2, con un effetto positivo, e FORM e SPOS, con effetti negativi. Queste evidenze possono ovviamente essere razionalizzate e interpretate, ma in una certa misura sono inevitabilmente ambigue.

In definitiva, guardando alla Tab. 2 il risultato più chiaro e significativo resta la stabilità della dipendenza negativa della durata, al variare delle specificazioni e tanto per i maschi che per le femmine. Una siffatta conclusione richiede peraltro qualche cautela, per almeno due motivi: sinora non abbiamo considerato le inaccuratezze nella durata riportata della disoccupazione, né la possibile presenza di eterogeneità non osservata.

In tema di inaccuratezze nella durata riportata della ricerca di lavoro, v'è da osservare innanzitutto che le evidenze riscontrate nel cap. 10 per un campione dell'intera popolazione di disoccupati, succintamente richiamate nella sez. 3.3, si ripropongono in termini sostanzialmente inalterati per il campione di giovani disoccupati lombardi (al riguardo, vedi Trivellato, Marliani e Torelli, 1989). Una consistente quota di queste inaccuratezze, già lo abbiamo notato, è attribuibile all'*heaping*.

Per avere lumi sulla robustezza delle stime dei parametri della (16) rispetto all'effetto ammucchiamento, ricorriamo ad una procedura che redistribuisce ragionevolmente le risposte *heaped* (Torelli e Trivellato, 1989). Le durate riportate t_i corrispondenti a 6k (k=1,2,...) vengono ripartite nell'intervallo

(1) $(t_i \mp 2)$ se $t_i \leq 36$,

(2) $(t_i \mp 3)$ se $t_i > 36$,

secondo la regola seguente. Si generano determinazioni indipendenti di una variabile casuale Z equidistribuita (0,N), con N=4 per il caso (1) e N=6 per il caso (2). Ai valori originari t_i si sostituiscono quindi i valori $\tilde{t}_i$, come segue:

$\tilde{t}_i = t_i - 2 + z_i$ se $t_i = 6k$, $k = 1,2,...,6$,
$\tilde{t}_i = t_i - 3 + z_i$ se $t_i = 6k$, $k>6$,
$\tilde{t}_i = t_i$ altrimenti.

Le stime e il valore della funzione di verosimiglianza del modello in questione, ottenuti dopo aver sostituito $\tilde{t}_i$ a t_i, sono risultati praticamente identici a quelli della Tab. 2. Si sono riscontrate, infatti, differenze solo oltre la terza cifra decimale. Ciò induce a concludere che, perlomeno entro questo spettro di ragionevoli correzioni dell'effetto *heaping* (e, com'è ovvio, limitatamente ai dati analizzati ed al modello adottato), l'inaccuratezza nella durata riportata non impedisce di cogliere le caratteristiche salienti della distribuzione delle durate di ricerca.

L'altro motivo di cautela nel concludere, dalla Tab. 2, in favore di una dipendenza negativa dalla durata attiene alla possibile presenza di eterogeneità non osservata nel campione, eterogeneità che, se trascurata, porta appunto ad una distorsione in tale senso. Per tener conto di un'eventuale componente di eterogeneità, abbiamo quindi rispecificato il modello Weibull di rischio proporzionale come segue:

$$h(t|x,\vartheta) = \vartheta \alpha t^{\alpha-1} \mu(\beta'x) . \qquad (17)$$

Palesemente, tale formulazione implica che l'eterogeneità modifichi la durata degli episodi individuali di ricerca di lavoro, mentre resta costante la struttura di dipendenza dalla durata. Seguendo Heckman e Siger (1984), abbiamo quindi adottato una rappresentazione non parametrica della distribuzione dell'eterogeneità, secondo la (7), ed abbiamo approntato la pertinente procedura di stima. Non essendo provato che la procedura converge ad un massimo globale, per proteggerci contro l'eventualità di massimi locali per la funzione di verosimiglianza abbiamo applicato la procedura più volte, partendo da valori iniziali diversi.

Il risultato di queste prove è stato che, tanto per il campione di maschi che per quello di femmine, si è sempre trovato un numero di punti di incremento pari ad 1, il che sta a significare che non vi è evidenza di eterogeneità non osservata [8]. Date le limitazioni nell'insieme di variabili esplicative impiegato, il risultato può destare forse qualche sorpresa. In parte, esso può tuttavia essere dovuto agli stringenti criteri adottati per l'individuazione del campione, criteri che valgono appunto ad identificare un insieme di soggetti ragionevolmente omogeneo. D'altronde, concordi riscontri empirici si hanno anche da altri studi sulla dinamica del mercato del lavoro giovanile (vedi, ad es., Flinn e Heckman, 1982, pp. 64-72; van Opstal e Theuwees, 1985; Lynch, 1986), i quali suggeriscono che una componente di eterogeneità non osservata è tipicamente rilevante per le transizioni dall'occupazione alla disoccu-

8 È interessante notare che in precedenti analisi dello stesso campione, condotte utilizzando un insieme sensibilmente più ristretto di variabili esplicative (ed una diversa metrica per le stesse), si è avuta evidenza di eterogeneità non osservata per le femmine (vedi Torelli, 1990).

pazione, ma non per quelle in direzione opposta.

Nel merito, per il campione di giovani disoccupati lombardi la presenza di dipendenza negativa dalla durata pare in definitiva ragionevolmente argomentata. E in una diversa prospettiva, attenta alle opportunità di analisi (ed agli interrogativi metodologici che sollevano), lo studio di caso svolto suggerisce che le potenzialità della RTFL per valutare, per di più in modo ricorrente, anche questi aspetti della dinamica del mercato del lavoro meritano forse di essere più pienamente utilizzate.

20

UN MODELLO DELL'OFFERTA DI LAVORO FEMMINILE IN PRESENZA DI VINCOLI ISTITUZIONALI SULL'ORARIO DI LAVORO

Enrico Rettore

1. *Introduzione*

Il problema della stima della funzione di offerta di lavoro, affrontato già negli anni '30, ha ricevuto grande attenzione negli ultimi quindici anni. A partire da Heckman (1974), il primo dei lavori empirici che Heckman, Killingsworth e MaCurdy (1981) definiscono "della seconda generazione", in tutti i Paesi industrializzati si sono moltiplicati gli studi econometrici sul comportamento rispetto al lavoro. Il salto di qualità compiuto in questi studi, rispetto ai precedenti della prima generazione, è stato reso possibile da un lato dall' esplicito riferimento alle indicazioni derivanti dalla teoria neoclassica del comportamento del consumatore, nella specificazione del modello e nella scelta delle tecniche di stima appropriate; dall' altro, dalla crescente disponibilità di basi di dati ricche di informazioni a livello individuale.

In tempi più recenti è cresciuta l'attenzione all'esigenza di rendere più realistici i modelli dell'offerta di lavoro, per quel che riguarda il riconoscimento del ruolo giocato dai 'vincoli' (cioè dalle caratteristiche istituzionali del mercato del lavoro e dallo stato dell' economia) nella determinazione delle condizioni rispetto al lavoro. L' assunzione accessoria (nel senso di non implicata necessariamente dalla teoria neoclassica), ma cruciale per le verifiche empiriche, che i comportamenti osservati in termini di partecipazione e ore di lavoro sono esclusivamente il risultato delle scelte operate dalle singole persone, formulata in gran parte dei lavori della seconda generazione, è stata rilassata in vari modi (Arellano e Meghir, 1990; Blundell, Ham e Meghir, 1987 e 1988; Colombino, 1984; Ham, 1982; Moffitt, 1982).

In Italia gli studi empirici sull' offerta di lavoro a livello micro non hanno avuto grande sviluppo, a causa della ridotta disponibilità di informazioni individuali. Gli studi esistenti (Colombino, 1984; Del Boca e Flinn, 1984; Del Boca, 1984) hanno fatto uso dei dati rilevati nel corso di un' indagine speciale sulla città di Torino (vedi Martinotti, 1980).

Questo capitolo si inserisce nel panorama delineato con l'intento di dare un contributo all'analisi del comportamento rispetto al lavoro in Italia. I problemi affrontati sono di due ordini: di specificazione e stima di un modello appropriato alle peculiarità del mercato del lavoro italiano; di sviluppo di

tecniche che permettano di utilizzare al meglio le informazioni atipiche correntemente rilevate in Italia.

Il primo problema è dato dal fatto che in Italia, più che altrove, la condizione rispetto al lavoro appare determinata dall'intrecciarsi di vari fattori. In particolare, oltre che le preferenze individuali e le condizioni della domanda, giocano un ruolo importante nella determinazione degli orari di lavoro gli accordi collettivi tra le organizzazioni dei datori di lavoro e dei lavoratori.

Quanto al secondo problema, la particolarità del caso italiano consiste nel fatto che nessuna delle indagini svolte correntemente rileva tutte le informazioni necessarie. Tali informazioni sono invece sparse in basi di dati diverse. Rispetto a questo problema, l' obiettivo perseguito è quello di utilizzare simultaneamente queste informazioni sparse. Senza nulla togliere al valore di studi condotti facendo uso di informazioni rilevate da indagini occasionali, sembra importante valutare l' utilizzabilità dei dati resi disponibili da indagini correnti, perchè questi consentirebbero di replicare gli studi in tempi e per aree diverse.

L' organizzazione del capitolo è la seguente. Nella sez. 2 è presentato un modello della decisione di partecipazione al lavoro in presenza di forti vincoli istituzionali sull'orario di lavoro. Nella sez. 3 sono discussi i problemi di stima posti dai dati disponibili in Italia. Nella sez. 4 sono presentati i risultati di un' applicazione all' offerta di lavoro delle donne sposate residenti in Lombardia o Veneto. Nella sez. 5 sono infine presentate alcune conclusioni.

2. *Il modello*

Nella loro versione più semplificata, i modelli dell'offerta di lavoro di derivazione neoclassica esistenti in letteratura constano della sola equazione delle ore di lavoro desiderate:

$$H^{*}=\alpha_1(x)+\alpha_2(x)w+\alpha_3(x)y+e, \qquad (1)$$

corredata dagli assunti $E\{e\}=0$ e $var\{e\}=\sigma^2$. In tale modello, w e y denotano rispettivamente il saggio salariale e il reddito non da lavoro; x sono variabili socio-demografiche individuali che modificano la forma della relazione esistente tra le ore di lavoro desiderate, H^{*}, e w e y; e è una v.c. che coglie gli effetti dell'eterogeneità campionaria non catturati dalle variabili incluse in x. Le variabili w, y e x sono assunte esogene.

L'equazione (1) è rilevante per spiegare sia la decisione di partecipare o meno, che la quantità di lavoro offerta una volta che si sia deciso di partecipare. Il modello assume infatti che se $H^{*}<0$ la decisione presa è quella di non partecipare; se invece $H^{*}>0$ la decisione presa è quella di lavorare H^{*} ore. Le ore di lavoro osservabili H sono assunte essere coincidenti con le ore di lavoro desiderate: se $H^{*}>0$, vale l'eguaglianza $H=H^{*}$.

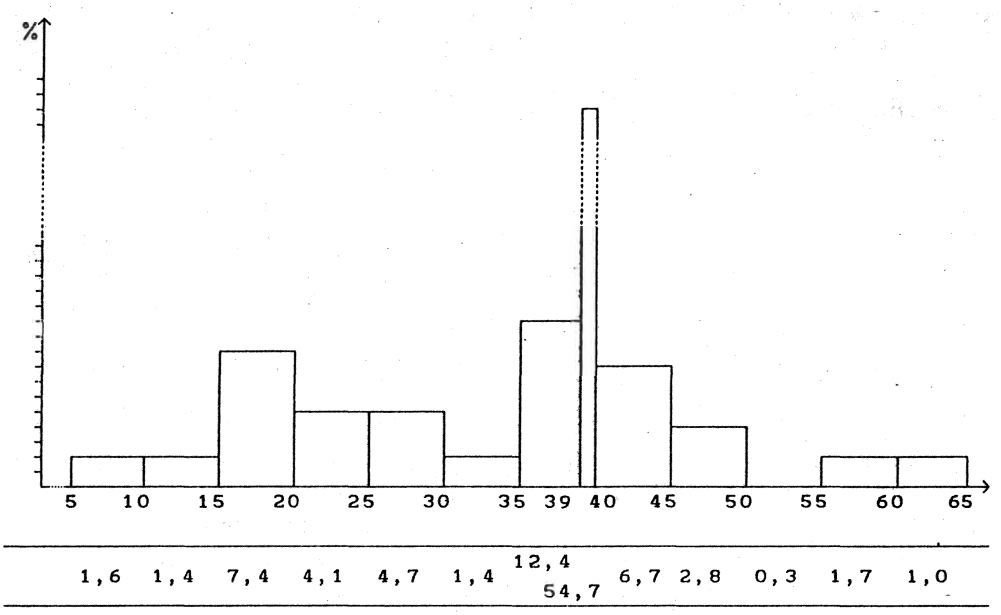
Dal punto di vista statistico, la stima dell'equazione (1) è riconducibile al problema di stima dei parametri di una regressione (eventualmente non

lineare) con dati censurati.

Ciò che desta maggiori perplessità in questo modello è l' eguaglianza $H=H^*$ assunta valere se $H^*> 0$. Da un lato non tutte le decisioni di partecipare sono destinate a realizzarsi (Blundell, Ham e Meghir, 1987 e 1988, hanno esteso il modello ammettendo l'esistenza sia di lavoratori disoccupati che di lavoratori scoraggiati); dall'altro, anche quando chi si offre trova un lavoro, non è detto che riesca a lavorare quanto vorrebbe.

A questo riguardo, la Fig. 1 presenta la distribuzione di frequenza delle ore di lavoro settimanali abituali delle donne sposate in Lombardia e Veneto (Aprile 1984).Risulta evidente che, anche se gli orari di lavoro osservati coprono un ampio arco di possibilità, la distribuzione è fortemente concentrata attorno alle 40 ore. Ciò è plausibilmente dovuto ai meccanismi collettivi di determinazione dell'orario di lavoro.Quale che sia l'importanza del ruolo giocato da tali meccanismi, la loro esistenza impedisce di stabilire l'eguaglianza $H=H^*$; in altre parole, impedisce di interpretare le stime ottenute regredendo H su w,y e *x* come stime dei parametri della funzione di offerta di lavoro.

Fig. 1 *Distribuzione percentuale delle ore di lavoro abituali settimanali delle donne sposate in Lombardia e Veneto nell'aprile 1984 (dati Istat, indagine sulle forze di lavoro)*



Una possibile via d'uscita consiste nel raccogliere ulteriori informazioni sulle persone che compongono il campione, informazioni che permettano di stabilire se vale l'eguaglianza $H=H^*$ e, nel caso non valga, di accertare il verso della diseguaglianza. E' questa la strada seguita da Colombino (1984). Tali informazioni pongono però problemi di accuratezza superiori alla norma, per il maggiore ricorso fatto a valutazioni di natura soggettiva dell'intervistato (vedi Del Boca e Flinn, 1984).

La soluzione proposta in questa sede consiste invece in un cambiamento del modello di riferimento. Sia:

$$H_p=\alpha_1(x,w,y)+\alpha_2(t)+u, \qquad (2)$$

con $E\{u\}=0$ e $var\{u\}=\sigma_u^2$, l'equazione dell'orario di lavoro associato alle proposte di lavoro che una persona riceve. Tale orario dipende dalle variabili *t*, che colgono gli effetti dei vincoli istituzionali vigenti sull'orario di lavoro, ma almeno in parte, dipende anche dal sistema di preferenze individuale, dal salario orario e dal reddito non da lavoro. In generale infatti, anche se è realistico ritenere che i vincoli istituzionali giochino un ruolo decisivo nella determinazione degli orari di lavoro, è poco realistico assumere che tutte le persone siano totalmente incapaci di influenzarlo anche solo parzialmente. Di qui l'inclusione di *x*, w e y nell'equazione. Sia poi:

$$H_{max}=\beta_1(x)+\beta_2(x)w+\beta_3(x)y+e, \qquad (3)$$

con $E\{e\}=0$ e $var\{e\}=\sigma_e^2$, l'equazione che lega il numero massimo di ore che una persona è disponibile a lavorare al saggio salariale e al reddito non da lavoro. La v.c. e coglie l'eterogeneità campionaria delle preferenze individuali. Per quanto detto riguardo la specificazione dell'equazione (2), in generale e ed u sono correlate.

Le implicazioni del sistema di equazioni (2)-(3) sul comportamento osservato rispetto al lavoro sono immediate. Se l'orario di lavoro proposto è inferiore ad H_{max}, la persona decide di lavorare e le ore di lavoro osservate sono determinate dalla (2); viceversa, se vale $H_{max}< H_p$ la persona decide di non lavorare. Il modello (2)-(3) ricade nella classe che Maddala (1983, pp. 174- 178 e 228-230) definisce "dei modelli a soglia di censura casuale non osservabile". In questo senso, è formalmente eguale ai modelli proposti da Cogan (1981) e da Gronau (1974). E' immediato anche mostrare che si tratta di un modello "Tobit di 2° tipo" (Amemiya, 1985, pp. 385-389).

Il modello proposto rinuncia quindi a modellare le preferenze individuali mediante le ore di lavoro, dato che nel contesto esaminato le ore di lavoro non sono oggetto di scelta (o lo sono molto poco). Porta invece l'attenzione su una grandezza, le ore massime, cruciale per l'unica vera scelta esercitabile liberamente, vale a dire la decisione di partecipare o meno alle condizioni salariali e di orario vigenti.

L'equazione di partecipazione implicata dal modello (2)-(3) è data semplicemente dalla differenza tra H_{max} e H_p:

$$Z^{*}=H_{max}-H_{p}. \qquad (4)$$

La variabile Z^{*} non è osservabile direttamente, ma se ne osserva il segno: positivo se la persona lavora, negativo in caso contrario. Quanto all'impatto di *x*, w e y sulle variabili osservabili orario di lavoro (H_p) e condizione rispetto al lavoro (sign Z^{*}), ci si attende quanto segue. Se veramente i vincoli sull'orario di lavoro sono difficilmente eludibili, tali variabili dovrebbero essere scarsamente significative nell'equazione relativa ad H_p. Al contrario, dovrebbero risultare significative nell' equazione di partecipazione (4), a causa del loro effetto su H_{max}.

Si noti che il modello dà luogo a due tipi logicamente distinti di inattività: da un lato l' inattività classica caratterizzata dalla condizione $H_{max} < 0$, determinata dalla non disponibilità a lavorare alle condizioni salariali vigenti (è facile mostrare che vale la doppia implicazione $H_{max} < 0 <=> H^{*} < 0$); dall'altro l'inattività caratterizzata dalla condizione $H_p > H_{max} > 0$ determinata dai vincoli istituzionali vigenti sull'orario di lavoro. In altre parole, il modello prevede l'esistenza di persone disponibili a lavorare al saggio salariale vigente, che scelgono di non lavorare perchè gli orari di lavoro disponibili sono troppo lunghi.

3. *Problemi posti dai dati disponibili*

Negli studi empirici sull'offerta di lavoro è usuale disporre, tra le altre, di informazioni sulle ore di lavoro svolte in un determinato periodo e sui redditi da lavoro percepiti nello stesso periodo. Una misura del saggio salariale w si ottiene banalmente dal rapporto redditi/ore.

In questi studi il principale problema è dato dal fatto che w non è osservabile per i non occupati. Tale problema è stato risolto da Heckman (1974) specificando accanto all'equazione (1)[1] un'equazione dei salari del seguente tipo:

$$W=\mu' z+v, \qquad (5)$$

con $E\{v\}=0$ e $var\{v\}=\sigma_v^2$. Nel modello composto dalle equazioni (1) e (5) si assume che il salario sia noto ai diretti interessati, i quali prendono quindi le loro decisioni in condizioni di certezza completa. La stima congiunta delle due equazioni permette di risolvere in modo esatto il problema della parziale non osservabilità di w. La funzione di verosimiglianza si ottiene osservando che il contributo di una persona che lavora è dato dalla funzione di densità della v.c. bivariata (H^{*},W), $f_{H^{*},W}(h,w)$, mentre il contributo di una persona che non lavora è dato dalla funzione di ripartizione della v.c. marginale H^{*} calcolata nel punto 0, $F_{H^{*}}(0)$.

---

1 In realtà nel modello di Heckman l'eq. (1) è parametrizzata in modo diverso, ma ciò non ha rilevanza per quanto qui interessa.

Nel caso del modello introdotto nella sezione precedente,il problema si pone in modo del tutto simile: H_p e W sono osservabili solo per coloro che lavorano. Sembra ragionevole affiancare all'assunzione che ogni persona conosce il suo saggio salariale, l'assunzione che ogni persona conosce l'orario di lavoro che può ottenere se decide di lavorare. Ciò essenzialmente perchè i vincoli istituzionali, riducendo drasticamente la variabilità degli orari disponibili, eliminano anche ogni incertezza al riguardo da parte di coloro che devono prendere la decisione di offrirsi o meno. Anche in questo caso, la stima congiunta delle tre equazioni (2), (3) e (5) permette di risolvere il problema. I contributi alla funzione di verosimiglianza di una persona che lavora e di una che non lavora sono rispettivamente:

$$P(H_{max} > H_p | H_p=h, W=w)\ P\ (H_p=h, W=w)=(1-F_{Hmax|Hp,W}(h|H_p=h, W=w))$$

$$f_{Hp,W}(h,w),$$

$$P(H_{max} < H_p)=F_{Hmax-Hp}(0), \qquad (6)$$

dove F ed f sono le funzioni di ripartizione e di densità delle v.c. indicate dai deponenti.

E' possibile ottenere stime consistenti dei parametri sia del modello basato sulle ore di lavoro desiderate che del modello basato sulle ore massime, facendo uso della procedura a più passi proposta da Heckman (1979) o di sue semplici estensioni.

Il problema aggiuntivo posto dai dati disponibili correntemente in Italia è dato dalla mancanza di indagini che rilevino contemporaneamente informazioni sui redditi (da lavoro e non) e sull' orario di lavoro: la rilevazione trimestrale delle forze di lavoro svolta dall' Istat rileva l'orario di lavoro; l'indagine annuale sui bilanci delle famiglie svolta dalla Banca d'Italia (Bdl nel seguito) rileva i redditi da lavoro e non. La principale conseguenza è che non è possibile calcolare il saggio salariale come rapporto tra redditi da lavoro e ore di lavoro.

La soluzione adottata riprende la proposta di Meghir e Arellano (1988) e consiste nell'utilizzare congiuntamente le due fonti, che prese singolarmente non permetterebbero di stimare il modello. Altri autori (Colombino, 1986; Cannari e Lemmi, 1988) si sono occupati dell'integrazione di più fonti di dati in riferimento allo stesso problema qui considerato, proponendo soluzioni in qualche modo basate su procedure di imputazione. La tesi implicita nel lavoro di Meghir e Arellano (1988) e qui accolta, è che la fase di integrazione delle varie fonti non può essere logicamente disgiunta dalla fase di costruzione e stima del modello di comportamento (per una discussione di questo punto vedi Rettore, 1989, cap.4).

Il modello viene specificato nel modo seguente:

$$\ln W=\mu' z+v \qquad (7)$$

$$\ln H_p=\alpha_1' x+\alpha_2' t+\alpha_3' z+\alpha_4 \ln Y+u \qquad (8)$$

$$\ln H_{max}=\beta_1' x+\beta_2 \ln W+\beta_3 \ln Y+e. \qquad (9)$$

Si noti in particolare che la variabile esplicativa ln W è stata eliminata dalla (8), sostituendovi l'equazione (7)[2].

Il punto cruciale della procedura corrisponde alla stima delle equazioni (7) e (8). Le informazioni disponibili non permettono di stimare l'equazione (7), dato che W non è mai osservabile. E' però possibile stimare congiuntamente le equazioni (7) e (8), utilizzando contemporaneamente le informazioni rilevate dalle due indagini. Siano:

$$K_i = \begin{cases} \ln H_i \\ \ln H_i^* W_i \end{cases} \quad e \quad D_i = \begin{cases} 0 & \text{se i appartiene al campione Istat,} \\ 1 & \text{se i appartiene al campione Bdl.} \end{cases}$$

Combinando le due equazioni (7) e (8) si ottiene:

$$K=\alpha_1' x + \alpha_2' t + \alpha_3' z + \alpha_4 \ln Y + \mu' zD + u + vD \qquad (10)$$

I residui di tale equazione sono palesemente eteroschedastici. Inoltre l'utilizzo dei dati relativi ai soli occupati pone i consueti problemi derivanti dalla selezione non casuale del campione. Stime consistenti di α e μ si ottengono applicando lo stimatore GLS all'equazione (10) opportunamente modificata secondo la procedura di Heckman (1979) per ovviare ai problemi di selezione non casuale del campione utilizzato per la stima[3].

Utilizzando le due equazioni stimate nel modo descritto, è possibile attribuire ad ognuna delle persone intervistate i valori previsti di $\ln H_p$ e ln W, grazie ai quali è possibile stimare l'equazione delle ore massime (9). Tali stime si ottengono massimizzando la funzione di verosimiglianza, alla quale gli occupati contribuiscono con:

$$P(\ln H_{max} > \ln H_p)=1-F_{Hmax}(\ln H_p) \qquad (11)$$

e gli inattivi con:

$$P(\ln H_{max} < \ln H_p)=F_{Hmax}(\ln H_p) \qquad (12)$$

(ulteriori dettagli sulla procedura di stima sono in Rettore, 1989).

---

2 Il modello include anche un'equazione dei redditi non da lavoro, per ovviare al fatto che tali redditi sono rilevati solo nell'indagine Bdl. Dato che il trattamento di tale equazione non pone alcun problema di rilievo, nel seguito per brevità si ragiona come se i redditi non da lavoro fossero rilevati da entrambe le indagini.

3 Il procedimento corrisponde al passaggio dalla v.c. bivariata (Hp,W), la cui distribuzione si presta ad essere parametrizzata alla luce delle considerazioni economiche svolte, alla v.c. bivariata (Hp, Hp W). Le osservazioni disponibili sono determinazioni di una o dell'altra delle due v.c. marginali Hp, Hp W.

4. *Un'applicazione all'offerta di lavoro delle donne sposate*

L'applicazione si riferisce ad un campione di donne sposate, di età inferiore a 60 anni, il cui marito lavora, residenti in Lombardia o Veneto. Il periodo di riferimento è il mese di Aprile del 1984 o l' intero 1984, secondo che le donne intervistate sono incluse nel campione dell' indagine trimestrale Istat sulle forze di lavoro oppure nel campione dell' indagine annuale Banca d' Italia sui bilanci delle famiglie. Per alleggerire il carico delle elaborazioni, dall'insieme delle donne del campione Istat soddisfacenti il criterio di selezione (circa 15.000) è stato estratto casualmente un sottocampione di dimensione pari a 1/10 del campione originario. Le dimensioni del campione utilizzato per la stima sono comunque analoghe a quelle di gran parte dei lavori empirici esistenti sull' argomento.

Una misura del reddito da lavoro settimanale è stata ottenuta dividendo il reddito annuo per 52.

Accanto alla già ricordata mancanza delle informazioni sui redditi da lavoro e non da lavoro nell' indagine Istat e delle informazioni su ore di lavoro e ricerca di lavoro nell' indagine Bdl, vi è da notare che l'indagine Bdl non rileva l' istruzione di coloro che non lavorano. Tale variabile, come si vedrà nel seguito, compare in quasi tutte le equazioni che compongono il modello adottato. Il problema è stato risolto includendo nel campione solo le occupate dell' indagine Bdl[4].

Il modello stimato è composto dalle tre equazioni relative ai logaritmi delle ore massime settimanali di lavoro, delle ore di lavoro settimanali e dei salari orari.

Le variabili esplicative incluse nel vettore x sono le seguenti: Fi0-2, Fi3-5, Fi6-10, Fi > 10 sono quattro *dummies* che assumono il valore 1 se il figlio più giovane appartiene alla classe di età che compare nel nome della variabile; Età, Età^2 denotano rispettivamente l'età in anni compiuti della donna e il suo quadrato; Ele, Matu, Lau sono tre *dummies* che assumono il valore 1 se la donna ha conseguito rispettivamente, la licenza elementare (al più), il diploma di scuola media superiore e la laurea.

Le variabili incluse nel vettore **t** che descrivono il regime degli orari di lavoro a cui sono soggette le donne, sono le ore medie settimanali nella provincia di residenza (Hmed) e una *dummy* che assume il valore 1 per le residenti in Lombardia (Regio). Si assume che la variabilità da provincia a provincia di Hmed sia ascrivibile esclusivamente alla diversità degli orari di lavoro disponibili in aree diverse[5].

---

4 E' noto che tale modo di procedere in generale non è neutrale rispetto al modello. Dà infatti luogo ad un campione stratificato endogenamente al modello. Ciò ha delle ripercussioni sulla forma della funzione di verosimiglianza dei parametri da stimare. (vedi Maddala, 1983, pp.170-174);Heckman e MaCurdy, 1986, pp. 1926- 1929). Tuttavia, in considerazione del fatto che il campione Bdl è notevolmente più piccolo del sottocampione Istat utilizzato per la stima, si può ritenere che i risultati ottenuti ignorando la stratificazione endogena del campione siano prossimi ai risultati esatti.

5 L'adozione dell'indicatore Hmed non è esente da critiche, dato che i confini provinciali non sono necessariamente rilevanti per la decisione di partecipazione.

Nell'equazione dei salari orari sono incluse le variabili (z) Fi0-2, Fi3-5, Fi6-10, Fi>10, Età, Età^2 , Ele, Matu, Lau , Regio già definite in precedenza e le variabili Età m., Età^2 m. (età in anni compiuti del marito e il suo quadrato), Ele m., Matu m., Lau m. (hanno il significato di Ele, Matu, Lau ma sono relative al marito).

Un ulteriore, non banale problema è posto dall'esistenza di persone che dichiarano di essere disponibili a lavorare e di non essere (ancora) riuscite a trovare un lavoro, cioè di coloro che le statistiche ufficiali definiscono e rilevano come persone in cerca di occupazione. Le posizioni possibili nei loro confronti sono due. Da un lato nei modelli più semplici della seconda generazione, riconducibili all'equazione (1), si assimilano i disoccupati agli inattivi. In alternativa si può assumere che la disponibilità a lavorare dichiarata corrisponda alla condizione $H_{max>}H_p$; questa assunzione porta ad includere i disoccupati tra gli attivi, come è usuale nelle statistiche sulle forze di lavoro. Nel primo caso il contributo di un disoccupato alla funzione di verosimiglianza è dato dalla (12), nel secondo dalla (11). Nel seguito sono presentati i risultati ottenuti aggiungendo il gruppo in questione agli attivi. In Rettore (1989) sono presentati anche i risultati relativi all'altra specificazione. Le differenze sono complessivamente trascurabili, anche per le ridotte dimensioni dell'insieme delle persone in cerca di occupazione.

La Tab. 1 presenta le stime dei parametri delle equazioni di partecipazione e delle ore di lavoro. Il fatto più appariscente è lo scarso impatto delle caratteristiche individuali sulle ore di lavoro, a fronte del forte impatto che queste hanno sulla probabilità di partecipazione. I due insiemi di stime non sono confrontabili quanto a valori assoluti, dato che i parametri dell'equazione di partecipazione sono stimabili a meno di un fattore di scala, ma i valori della statistica t sono eloquenti.

Nell'equazione di partecipazione le variabili si comportano nel modo atteso: i figli abbassano la probabilità di partecipazione (ma l'influenza di Fi0-2 non è ben definita); l' età ha un effetto positivo fino a 36 anni, poi negativo; l' istruzione ha un effetto positivo. Il reddito non da lavoro, conformemente alle attese, ha un effetto negativo, ma il coefficiente non è ben determinato. Tale coefficiente va forse valutato congiuntamente ai coefficienti relativi all'istruzione del marito, che si può ritenere una buona *proxy* del reddito non da lavoro della moglie. Se ne ricava l'impressione che il reddito non da lavoro faccia decrescere in modo più che lineare la variabile latente dipendente dell' equazione, e che tale effetto sia colto in parte dalla variabile ln Y e per la parte rimanente dalle tre *dummies* Ele m., Matu m. e Lau m..

Nell'equazione delle ore di lavoro nessuna delle variabili appare significativa. Il parametro meglio determinato è quello relativo all'orario medio nella provincia di residenza, che ha pure un effetto negativo ben determinato sulla probabilità di partecipazione. Questo risultato è coerente con l'ipotesi che quanto più lunghi sono gli orari di lavoro disponibili tanto più alta è la probabilità di non partecipazione $P(H_{max}< H_p)$. Un discorso analogo potrebbe valere anche per la *dummy* regionale, in considerazione del fatto che il suo parametro ha segno opposto nelle equazioni di partecipazione e delle ore

Tab. 1: *Stime di massima verosimiglianza relative all'equazione di partecipazione e stime OLS corrette per la selezione non casuale del campione relative all'equazione delle ore di lavoro* [a]

	equazione di partecipazione		equazione delle ore di lavoro	
intercetta	5,9580	(2,8925)	1,1878	(1,1479)
ln Y	-0,2899	(-1,6689)	0,1018	(1,4349)
Fi0-2	-0,3016	(-1,8427)	-0,0740	(-1,0936)
Fi3-5	-0,4376	(-3,1566)	-0,0822	(-0,9204)
Fi6-10	-0,5412	(-4,1201)	-0,1043	(-0,9135)
Fi>10	-0,4690	(-3,8416)	-0,0878	(-0,8569)
Età	0,1336	(2,5199)	-0,0135	(-0,4332)
Età^{2}[b]	-0,1868	(-2,9785)	0,0230	(0,5259)
Ele	-0,0642	(-0,6691)	-0,0205	(-0,4780)
Matu	0,4391	(3,3052)	-0,0943	(-1,0664)
Lau	1,4838	(4,9598)	-0,4052	(-1,4077)
Età m.	-0,0995	(-1,8210)	0,0256	(1,0263)
Età^{2} m.[b]	0,1077	(1,8175)	-0,0387	(-1,3545)
Ele m.	-0,1442	(-1,5366)	-0,0025	(-0,0557)
Matu m.	0,1032	(0,7972)	-0,0595	(-1,3396)
Lau m.	-0,5555	(-2,3066)	0,0543	(0,4106)
Regic	0,2019	(2,1264)	-0,0200	(-0,3772)
Hmec	-0,0721	(-2,3195)	0,0337	(1,7772)
Sele[c]			-0,0620	(-0,1787)

(a) tra parentesi i valori della statistica t.

(b) moltiplicato per 100.

(c) termine di correzione per la selezione non casuale.

di lavoro (anche se in quest'ultimo caso la stima è scarsamente affidabile): la variabile potrebbe cogliere l' effetto dell' orario di lavoro prevalente nell' area geografica sulla partecipazione, non colto dalla variabile Hmed a causa della dubbia rilevanza dei confini provinciali per il problema in questione.

Le stime relative all' equazione dei salari orari sono presentate nella prima colonna della Tab. 2. Si nota innanzi tutto l'effetto forte e ben definito dell' istruzione: le persone in possesso al più della licenza elementare non differiscono in modo significativo da quelle in possesso della licenza media; ma dal gruppo delle persone in possesso di titoli di studio inferiori si staccano in modo netto le diplomate e ancora più le laureate. Le *dummies* relative ai figli e l'età dovrebbero cogliere l'effetto dell'esperienza lavorativa sul salario orario: le prime perchè legate a periodi di astensione dal lavoro, la seconda perchè legata in modo *diretto* al numero di anni di lavoro (condizionatamente al fatto che la persona abbia avuto una storia lavorativa regolare). I risultati non sono in linea con queste attese. La presenza di figli appare del tutto

irrilevante. Il profilo del salario orario secondo l'età risulta invece crescente fino a 38 anni e poi decrescente. Presumibilmente tale profilo risulta dalla combinazione dell'effetto 'esperienza lavorativa' con effetti generazionali. Questi ultimi non possono essere colti in modo appropriato disponendo di dati sezionali, per cui anche il ruolo dell'esperienza lavorativa rimane nascosto.

Per quel che riguarda l' equazione delle ore massime (seconda colonna della Tab. 2), le sole variabili che sembrano rilevanti sono le quattro *dummies* relative all' età del figlio più giovane (peraltro con un'imprevedibile prevalenza della terza classe di età sulla prima e sulla seconda), il salario orario e, in minor misura, il reddito non da lavoro. A queste sono forse da aggiungere Età e Età^2 , i cui valori t sono poco indicativi a causa della forte collinearità esistente tra loro: l' effetto combinato delle due variabili dà le ore massime

Tab. 2: *Stime OLS corrette per la selezione non casuale relative all'equazione dei salari e stime di massima verosimiglianza relative all'equazione delle ore massime* [(a)]

	equazione dei salari orari		equazione delle ore massime [(b)]	
intercetta	1,9933	(1,5705)	8,1142	(3,3177)
ln W			0,4864	(2,1369)
ln Y			-0,2274	(-1,4255)
Fi0-2	0,0225	(0,1445)	-0,4135	(-2,4628)
Fi3-5	-0,1611	(-0,8292)	-0,5130	(-3,3067)
Fi6-10	-0,0182	(-0,0836)	-0,7159	(-4,8659)
Fi>10	-0,1743	(-0,9940)	-0,5526	(-3,7336)
Età	0,1328	(1,3837)	0,0303	(0,8194)
Età$^{2(c)}$	-0,1728	(-1,5541)	-0,0654	(-1,4147)
Ele	-0,1740	(-1,0193)	-0,0325	(-0,2719)
Matu	0,2892	(2,0591)	0,0446	(0,3006)
Lau	0,6540	(2,1171)	0,0823	(0,2287)
Età m.	-0,0833	(-0,9903)		
Età^2 m.$^{(c)}$	0,1011	(1,1697)		
Ele m.	-0,3295	(-1,6956)		
Matu m.	0,1706	(1,4675)		
Lau m.	-0,0284	(-0,1513)		
Regio	0,0781	(0,6825)		
Sele$^{(d)}$	3,7184	(0,3358)		
$1/\sigma_e$			1,8788	(2,9619)

(a) tra parentesi i valori della statistica t.

(b) le stime si riferiscono alla parametrizzazione β/σ_e.

(c) moltiplicato per 100.

(d) termine di correzione per la selezione non casuale.

crescenti fino a circa 23 anni e poi decrescenti. Il livello di istruzione appare del tutto irrilevante. E' possibile quindi affermare che il forte effetto dell' istruzione e dell' età sulla partecipazione, documentato in Tab. 1, è dovuto all'influenza di queste variabili sul salario orario e del salario orario sulle ore massime. Al netto di quest' effetto rimane poca cosa. Il contrario avviene per i figli, che sono poco rilevanti nella determinazione del salario orario e molto rilevanti nella determinazione delle ore massime.

La Tab. 3 presenta i valori medi e le deviazioni standard delle ore massime di lavoro, delle ore di lavoro e dei salari orari previsti dal modello, calcolati nei dodici gruppi che si ottengono combinando le modalità delle tre variabili figli (sì,no), età (< 29, (20,50], 50) e istruzione (licenza elementare o media, maturità o laurea). Si nota innanzi tutto la ridotta variabilità tra gruppi delle ore di lavoro (min.=33, max.=41) a fronte della variabilità molto più marcata delle ore massime (min.=24, max.=64). Inoltre, le ore di lavoro appaiono risentire solo delle differenze di istruzione, mentre le ore massime mostrano una sensibilità marcata a variazioni in tutte le variabili considerate. Questo è evidentemente attribuibile alla scarsa flessibilità degli orari di lavoro. Si nota poi che, a parità di altre condizioni, una maggiore istruzione comporta ore massime più elevate (e quindi una maggiore disponibilità a lavorare) e orari di lavoro più corti.

Il confronto tra ore massime e ore di lavoro permette di individuare il comportamento rispetto al lavoro dei gruppi considerati. Se ne ricava che tipicamente le inattive sono donne sopra i 50 anni, indipendentemente da ogni altra caratteristica, oppure donne tra i 29 e i 50 anni che hanno figli e un livello di istruzione non superiore all'obbligo. Quest'ultimo è il gruppo di gran lunga più numeroso tra quelli considerati.

Guardando ai valori assunti dalle ore massime, si potrebbe essere tentati di concludere che un forte aumento nella flessibilità degli orari, oppure anche solo un forte aumento dei posti di lavoro *part- time*, produrrebbe un sostanziale incremento del tasso di partecipazione femminile. A prima vista, infatti, gran parte dell'inattività sembra dovuta ai vincoli vigenti sugli orari. Almeno in parte però, le ore massime delle donne inattive sono sovrastimate dal modello, a causa della sovrastima del loro salario orario. Infatti, come si è visto in precedenza, l'esperienza lavorativa è misurata nell'equazione dei salari dall'età, che è plausibilmente un buon predittore del salario orario di una donna con una storia lavorativa regolare, ma è destinata inevitabilmente a sovrastimare il salario orario delle donne giunte a una certa età senza alcuna esperienza lavorativa. Dato l'effetto positivo del salario orario sulle ore massime, ne discende la sovrastima di queste ultime. Vi è da notare peraltro che questo è un difetto comune a molti dei lavori esistenti sull'argomento.

Infine, quanto ai salari orari, si nota il forte effetto dell'istruzione (minimo 10.000 lire in più,a parità di altre condizioni, per le donne con istruzione elevata). Il salario orario delle donne con istruzione bassa decresce con l' età, mentre quello delle donne con istruzione elevata è prima crescente e poi decrescente. La presenza di figli diminuisce il salario orario, ma le dif-

Tab. 3: *Ore massime di lavoro settimanali, ore di lavoro settimanali e salari orari* [(a)]. *Media e deviazione standard dei valori previsti dal modello in alcuni gruppi socio-demografici* [(b)]

gruppi socio-demografici	ore massime settimanali	ore settimanali	salari orari
con figli, età giovane, istruzione bassa (124)[(c)]	41,0 (3,5)	38,2 (2,0)	13,00 (2,96)
senza figli, età giovane, istruzione bassa (56)	55,9 (3,4)	41,2 (2,0)	14,74 (3,20)
con figli, età media, istruzione bassa (759)	33,5 (4,1)	37,9 (2,1)	12,04 (3,64)
senza figli, età media, istruzione bassa (74)	46,7 (6,1)	41,2 (2,2)	13,26 (3,64)
con figli, età matura, istruzione bassa (141)	24,4 (2,2)	37,1 (2,7)	8,64 (2,28)
senza figli, età matura, istruzione bassa (39)	33,9 (2,8)	39,9 (3,2)	10,12 (2,87)
con figli, età giovane, istruzione alta (33)	48,1 (3,0)	35,2 (1,7)	22,61 (4,02)
senza figli, età giovane, istruzione alta (32)	64,4 (3,3)	36,3 (3,3)	24,24 (5,74)
con figli, età media, istruzione alta (125)	42,8 (4,9)	33,6 (4,0)	24,75 (6,11)
senza figli, età media, istruzione alta (13)	64,1 (4,7)	34,5 (4,1)	29,98 (6,95)
con figli, età matura, istruzione alta (9)	27,9 (2,4)	34,6 (3,0)	17,91 (3,37)
senza figli, età matura, istruzione alta (1)	35,7 (0,0)	38,9 (0,0)	19,58 (0,00)

(a) migliaia di lire.

(b) le tre classi di età sono <=29, (29,50], >50; istruzione bassa=al più licenza media, istruzione alta=maturità o laurea.

(c) numerosità dei gruppi.

ferenze il più delle volte sono contenute.

5. *Conclusioni*

L'ipotesi attorno alla quale si sviluppa questo capitolo è che le ore di lavoro osservate, diversamente da quanto si è tradizionalmente assunto nei modelli empirici dell'offerta di lavoro, non riflettano se non in minima parte le preferenze individuali con riguardo all'allocazione del tempo, tra tempo riservato al lavoro e tempo libero. Tale ipotesi è vagliata specificando un modello in cui la decisione di partecipare o meno è presa confrontando le ore di lavoro che si è costretti a svolgere se si decide di lavorare (assunte note alle dirette interessate) e le ore massime che si è disponibili a lavorare alle condizioni retributive vigenti.

I risultati ottenuti sono coerenti con l'ipotesi formulata: (i) l'orario di lavoro medio nell'area geografica di residenza, assunto descrivere i vincoli istituzionali vigenti, influenza positivamente gli orari di lavoro osservati e negativamente la probabilità di partecipazione; (ii) il reddito non da lavoro e le altre variabili socio-demografiche che, direttamente o tramite il saggio salariale, la teoria economica suggerisce essere rilevanti nella determinazione dell'offerta di lavoro, risultano tali nell'equazione di partecipazione, ma del tutto irrilevanti nell'equazione delle ore di lavoro. Il modello proposto, esprimendo l'equazione di partecipazione come differenza tra l'equazione delle ore massime di lavoro e l'equazione delle ore di lavoro, fornisce una razionalizzazione di tali evidenze.

Quanto alle determinanti delle ore massime di lavoro,le evidenze sono abbastanza nette. Il saggio salariale, secondo quanto suggerito dalla teoria, influisce positivamente. Il coefficiente del reddito non da lavoro ha segno negativo, ma risulta significativamente diverso da zero solo ad un livello pari circa a 0.15. Le variabili socio-demografiche considerate si comportano in modo polarizzato: l'età e il livello di istruzione hanno un effetto marcato sulle ore massime, ma prevalentemente tramite il loro effetto sul saggio salariale, mentre gli effetti diretti appaiono trascurabili; la presenza di figli nel nucleo famigliare, al contrario, non sembra avere grandi effetti sul saggio salariale (effetti che comunque, quando presenti, sono negativi, conformemente alle attese), ma abbassa in modo drastico le ore massime.

PARTE SESTA:

INDAGINI SUPPLETIVE

21

L'ERRORE DELL'INTERVISTATORE NELL'INDAGINE SULLE FORZE DI LAVORO VALUTATO MEDIANTE COMPENETRAZIONE DELLE ASSEGNAZIONI DEGLI INTERVISTATORI

Lorenzo Bernardi, Luigi Fabbris, Ippolito Sanetti e M. Antonia De Marchis

1. *Motivi ed obiettivi*

Nel corso del progetto FOLA è parso opportuno avviare la discussione sul ruolo degli strumenti e degli attori del processo di rilevazione dei dati sulle forze di lavoro in Italia, mediante l'accertamento empirico degli effetti che questi e quelli hanno sulle stime delle grandezze prodotte nell'ambito dell'indagine.

In una indagine complessa dal punto di vista della rilevazione, qual è quella trimestrale sulle forze di lavoro (RTFL), assume un particolare rilievo l'effetto negativo che gli intervistatori hanno sulle stime di buona parte delle grandezze stimate. Ciò si verifica durante l'intervista quando il rilevatore è disattento o ha scarsa conoscenza degli strumenti di rilevazione, quando frappone - anche inconsciamente - tra il dato reale e quello rilevato propri pregiudizi o erronee attese sui risultati dell'indagine, quando la capacità di comunicare le domande o di recepire le risposte è scarsa, quando le istruzioni predisposte in calce al modello di rilevazione sono inadeguate o la qualità dello stesso questionario è scarsa. Il condizionamento delle risposte può, infine, derivare dalla sola presenza fisica del rilevatore di fronte al rispondente, in eventuale interazione con il modo di presentarsi.

In genere, questa forma di incanalamento delle risposte determina, per le risposte ottenute da un rilevatore, una variabilità mediamente inferiore a quella osservabile in assenza del rilevatore. La perdita di variabilità tra individui conseguente all'effetto degli intervistatori determina una proporzionale caduta di efficienza delle stime. La perdita di efficienza varia a seconda della variabile, della categoria di rispondenti, delle modalità di intervista.

Per misurare l'effetto dei rilevatori sulle stime è particolarmente indicata la tecnica della compenetrazione delle assegnazioni degli intervistatori (Mahalanobis, 1946). Questa tecnica consiste nel suddividere a caso l'insieme di unità da rilevare in sottoinsiemi di uguale ampiezza e nell'associare casualmente i sottoinsiemi (assegnazioni) agli intervistatori. Dato che ogni assegnazione è un campione casuale della popolazione, le differenze tra i valori medi delle risposte registrate dai singoli rilevatori e il valor medio

globale sono addebitabili alla sola azione dei rilevatori.

La compenetrazione delle assegnazioni si giustifica come strumento per la rilevazione dell'errore degli intervistatori perché, se intelligentemente applicata, è un sistema di controllo della qualità della rilevazione *silenzioso* (nel senso che gli addetti alla rilevazione non se ne accorgono), *poco costoso* (il numero di rilevazioni è lo stesso che si avrebbe se non si applicasse) e *affidabile* (le stime che si ottengono sono stabili e si basano su pochi assunti) (Fabbris, 1989, cap. 9).

Il piano di raccolta dei dati per l'indagine di cui si tratta è descritto dettagliatamente nella sez. 2.

Mediante la compenetrazione dei rilevatori si possono stimare quasi correttamente due indicatori dell'effetto dei rilevatori: la varianza dell'intervistatore e il coefficiente di correlazione intra-intervistatore (sez. 3). Il primo misura in termini assoluti l'incremento di varianza globale addebitabile all'effetto degli intervistatori: componente che è da sommare a quella campionaria, agli errori che i rispondenti commettono indipendentemente dalla presenza dei rilevatori, ai vari effetti di interazione tra le fonti di errore.

Il secondo indicatore è una misura dello stesso effetto resa indipendente dalla scala di misura delle variabili. Anche se, come si precisa nel seguito, la stima ottenibile con il disegno di rilevazione adottato è solo una buona approssimazione del valore da stimare, il coefficiente di correlazione intra-intervistatore permette di confrontare l'effetto su variabili diverse o su sottoclassi di dimensione variabile.

Il modello di stima è stato applicato su insiemi di dati rilevati in alcuni comuni rappresentativi di due regioni. Quantunque gli insiemi campionari osservati non ambiscano a rappresentare l'intera popolazione italiana, l'analisi dei risultati permette di:

(a) misurare l'entità delle diverse componenti della variabilità delle stime in alcuni strati rappresentativi, separando dalle restanti la componente di varianza attribuibile agli intervistatori;
(b) identificare le variabili particolarmente esposte al rischio di errore dell'intervistatore e individuare sottogruppi di rispondenti che, a vario titolo, tendono ad essere classificati in modo difforme in zone diverse del Paese da rilevatori con certe caratteristiche;
(c) suggerire modalità di rilevazione (definizioni, analiticità delle modalità di risposta, istruzioni per i rilevatori) e/o di elaborazione dei dati che contengono l'errore degli intervistatori.

2. *Il disegno di rilevazione*

L'indagine ha interessato le regioni Lombardia e Campania ed è stata svolta contestualmente alla RTFL di ottobre 1988.

La scelta dell'ambito territoriale sul quale effettuare l'esperimento è stata dettata dall'essere

- il campione per la rilevazione delle forze di lavoro ampliato, in entrambe le regioni,

– le due regioni differenziate dal punto di vista organizzativo e per i comportamenti attesi dei rilevatori e della popolazione oggetto di studio.

Per un più efficace controllo dello svolgimento delle operazioni sul campo, l'indagine è stata svolta su pochi comuni con i quali l'Istat aveva stretto rapporti di collaborazione particolarmente felici e nei quali era possibile seguire da vicino le operazioni di rilevazione. Requisito indispensabile era che nel Comune operassero almeno due rilevatori, e in questo ambito sono stati scelti i Comuni che si sapeva essere meglio organizzati e più disponibili.

2.1. *L'organizzazione dell'indagine*

Le norme di rilevazione delle RTFL prevedono che, per rendere più agevole agli intervistatori il contatto con le famiglie campione, i comuni suddividano prima dell'estrazione delle famiglie dall'anagrafe il territorio in aree il più possibile omogenee dal punto di vista socio-economico ed assegnino un'area a ciascun rilevatore, evitando così che questi debba spostarsi in tutta la città.

La tecnica della compenetrazione dei campioni prevede, invece, di suddividere in maniera casuale le famiglie campione delle aree sulle quali, di norma, operano più intervistatori, assegnando a ciascun intervistatore una partizione casuale del campione. L'ottimo, suggerito da Biemer e Stokes (1985) per una indagine su più stadi, è che si formino strati entro i quali esercitano la propria attività due rilevatori. Nella nostra indagine, per limitare gli spostamenti, sono stati scelti solo comuni con almeno due rilevatori.

Dal punto di vista pratico, sarebbe stata ottima l'individuazione di strati territoriali intracomunali sui quali ogni rilevatore avrebbe operato intervistando metà della propria e metà dell'altrui assegnazione abituale di famiglie. Per vari motivi (numero dispari di aree, mancata corrispondenza tra rilevatori ed aree, suddivisione delle aree secondo una logica non geografica *etc.*), al disegno stratificato dentro i comuni è stata preferita la compenetrazione delle famiglie dell'intero comune. In ogni comune, dunque, ogni assegnazione è un campione casuale delle famiglie campione, associata casualmente ad un rilevatore operante per la RTFL. A differenza di quanto accade per l'indagine corrente, per questa indagine suppletiva sono stati registrati anche i modelli P51 relativi alle sostituzioni delle famiglie che non rispondono, utilizzabili per un controllo più approfondito del lavoro svolto sul campo sia dagli uffici comunali che dai rilevatori.

L'Istat ha provveduto attraverso i suoi organi centrali e periferici ai lavori di preparazione dell'indagine, garantendone anche la supervisione ed il controllo in fase esecutiva[1] . La dimensione e l'articolazione dell'indagine

---

1 La preparazione dell'indagine suppletiva è iniziata con la rilevazione corrente del luglio 1988. Sono state attuate le seguenti operazioni: (i) visita ed informazione ai Comuni prescelti nel corso delle ispezioni per la rilevazione corrente; (ii) aumento del compenso ai Comuni, in ragione di £ 3.000 a modello, in considerazione del maggiore impegno richiesto; (iii) opera di convincimento degli uffici comunali e degli intervistatori per l'espletamento delle operazioni preparatorie e della rilevazione sotto la costante guida di personale qualificato dell'Istat, nel periodo settembre-ottobre 1988; (iv) assegnazione di un codice ad

sono illustrati nella Tab. 1. Si tratta di 3 comuni capoluogo di provincia, altrettanti non capoluogo sopra i 20 mila abitanti e 4 con popolazione inferiore ai 20 mila abitanti in Lombardia: di 4 capoluoghi di provincia, 2 non capoluogo con popolazione superiore a 20 mila abitanti e 3 con popolazione inferiore a 20 mila abitanti in Campania. Si è deciso di escludere i comuni di Milano e Napoli, per la vastità del territorio e per il numero di rilevatori che vi operano (40 e 17, rispettivamente). La dimensione del campione è di 3.200 famiglie, di cui circa 2.000 in Lombardia e circa 1.200 in Campania.

La numerosità campionaria risultante, in termini di individui, e il carico medio di lavoro per intervistatore sono sinteticamente presentati nella Tab.2.

2.2. *Resoconto sull'esperienza di raccolta dei dati*

Dal punto di vista dell'organizzazione, lo sforzo impiegato nella gestione dell'indagine è stato ampiamente compensato dall'esperienza accumulata e dalle informazioni ottenute. Sono così state confermate le impressioni precedentemente accumulate sulla conduzione delle operazioni sul campo. E' risultato che essa non è univoca nei vari comuni a causa soprattutto della scorretta interpretazione o scarsa conoscenza delle norme di rilevazione quanto a:

- divisione del territorio in aree ed assegnazione delle aree e delle famiglie ai rilevatori,
- selezione delle famiglie dall'anagrafe,
- reclutamento ed addestramento dei rilevatori.

Va sottolineato che l'indagine in questione è la prima dell'Istat, collegata alla RTFL, nel corso della quale sono stati particolarmente curati i rapporti con i Comuni, le istruzioni ai rilevatori sono state impartite direttamente dal centro, sono stati sperimentati vari sistemi di controllo. L'evidenza è che l'indagine abbia ottenuto un discreto grado di rispondenza[2] e sia pertanto affidabile.

[segue nota]

ogni rilevatore; (v) raccolta di informazioni sul numero di aree, sul numero di famiglie per area, sui criteri di ripartizione topografica delle aree, sul numero di rilevatori e sull'abbinamento area-rilevatore.
Nel mese di settembre 1988 il Servizio Famiglie dell'Istat ha tenuto riunioni di istruzione agli ispettori Istat, nel corso delle quali sono stati esposti i motivi e le modalità di espletamento dell'indagine suppletiva. Per agevolare il lavoro sul campo, è stato anche predisposto un manuale di istruzioni. Gli ispettori, a loro volta, hanno provveduto ad istruire i responsabili comunali e i rilevatori a predisporre i modelli P/48, comprendenti gli elenchi delle famiglie da consegnare ai rilevatori per l'indagine dell'ottobre 1988.
Nel corso della rilevazione, gli ispettori sono rimasti presso i Comuni per 10-15 giorni per svolgere efficacemente le operazioni di assistenza, supervisione e controllo. In ogni comune sono state effettuate verifiche sul 10% delle famiglie campione per quanto concerne: (i) la consegna giornaliera del materiale compilato; (ii) il controllo dell'avvenuta intervista, quando possibile per telefono oppure mediante ritorno presso alcune famiglie già intervistate; (iii) la revisione di alcuni modelli alla presenza dell'intervistatore (vedi la Tab. 1).

2 Come già per l'indagine corrente, è stata adottata per l'indagine suppletiva la tecnica di sostituire le famiglie che non collaborano. La percentuale di famiglie che non hanno collaborato è quasi uguale nelle due Regioni: 4,5% in Lombardia e 4,7% in Campania. I motivi di sostituzione sono parzialmente differenti: rifiuto nel 30% dei casi in Lombardia e nel 16% in Campania, irreperibilità nel 35% delle mancate collaborazioni in ambedue le regioni, morte dell'intera famiglia o domicilio in altro comune nel 30% dei casi in Lombardia e nel 46% in Campania.

Tab. 1: *Famiglie rilevate e addetti alla rilevazione nei comuni nei quali è stata svolta l'indagine suppletiva con compenetrazione delle assegnazioni agli intervistatori*

	Famiglie rilevate	Rilevatori	Controlli			
			Totale famiglie contr.	Avvenuta interv.		Revisione modelli
				telefono	ritorno	
Cinisello B.	140	2	25	10	5	10
Gonzaga	92	2	14	6	8	-
Lecco	156	2	25	10	5	10
Ponte S. Pietro	124	2	13	4	3	6
Como	288	4	48	30	8	10
Vigevano	308	7	36	18	6	12
Peschiera B.	112	2	58	12	6	40
Pavia	404	2	30	14	16	-
Bressana B.	84	2	84	10	6	68
Varese	272	3	91	31	30	30
Lombardia	1.980	28	424	145	93	186
Sant'Arpino	108	3	9	3	3	3
Aversa	84	2	9	3	4	2
Capua	104	4	12	4	4	4
Caserta	108	3	18	6	6	6
Casaluce	108	2	21	4	13	4
Benevento	280	4	24	5	12	7
Salerno	220	5	20	10	5	5
Nocera Inf.	60	2	8	3	3	2
Avellino	152	2	22	8	6	8
Campania	1.224	27	143	46	56	41

Tab. 2: *Numerosità campionaria, numero di intervistatori e carico medio di lavoro per intervistatore, per strato*

Strato	Numerosità campionaria	Numero intervistatori	Carico medio di lavoro
Lombardia:			
capoluoghi	2.442	9	271,3
altri	2.865	19	150,8
Campania:			
capoluoghi	2.359	14	168,5
altri	2.865	13	220,4
Totale	10.531	55	191,5

3. *Stimatori*

L'effetto degli intervistatori sull'attendibilità delle stime è funzione della *varianza dell'intervistatore*, la quale è scomponibile nel prodotto della varianza elementare di risposta, del coefficiente di correlazione intraclasse (intra-intervistatore) e di un fattore correttivo. Denotando con σ_r^2 la varianza elementare di risposta, ρ il coefficiente di correlazione intraclasse, n il numero di interviste assegnate a k intervistatori, la varianza dell'intervistatore $\sigma^2(\bar{\alpha})$ è data da:

$$\sigma^2(\bar{\alpha}) = \rho\, \sigma_r^2 \left(\frac{1}{k} - \frac{1}{n} \right) = \rho\, \sigma_r^2 \, (\tilde{n} - 1)/n \,, \tag{1}$$

dove $\tilde{n} = n/k$.

Uno stimatore di $\rho\, \sigma_r^2$, corretto nell'ipotesi di campionamento casuale semplice entro l'area dove è avvenuta la compenetrazione delle assegnazioni e di correlazione nulla tra errori campionari e di risposta e tra errori di risposta di intervistatori diversi, è dato da (Kish, 1962):

$$\hat{\rho}\, s_r^2 = k\, s^2 (\bar{y}) - \frac{1}{n(\tilde{n} - 1)} \sum_i^k \sum_j^{\tilde{n}} (y_{ij} - \bar{y})^2, \tag{2}$$

dove

$$s^2 (\bar{y}) = \frac{1}{k(k-1)} \sum_i (\bar{y}_i - \bar{y})^2, \tag{3}$$

y_{ij} è il valore della variabile y rilevato dall'intervistatore i presso l'unità campionaria j ; $\bar{y}_i$ è il valor medio dei valori rilevati dall'intervistatore i ; $\bar{y}$ è il valor medio globale.

Se la variabile y assume solo i valori 0 e 1, è

$$\hat{\rho}\, s_r^2 = k\, s^2 (p) - \frac{1}{k(\tilde{n} - 1)} \sum p_i (1 - p_i) \,. \tag{4}$$

Uno stimatore approssimato del coefficiente di correlazione intra-intervi-

---

[segue nota]

Nei comuni con almeno 20 mila abitanti le sostituzioni sono dovute soprattutto ad espliciti rifiuti a collaborare (30-35%), a domicilio in altro comune (un altro 35-40%), mentre nei comuni più piccoli il 60% delle sostituzioni consegue all'irreperibilità dei rispondenti, ragionevolmente a causa della minore frequenza del telefono e alla maggiore distanza tra famiglie campione nelle aree rurali. Per quanto riguarda il numero di visite, nei comuni capoluogo è stata effettuata in media una sola visita, negli altri comuni con oltre 20 mila abitanti la media è stata 1,11, nei comuni con meno di 20 mila abitanti la media è stata 1,88.

statore è proposto da Kish (1962) come rapporto tra la (4) e la varianza elementare

$$\hat{\rho}\, s_r^2/(s_r^2 + s_c^2), \tag{5}$$

dove s_c^2 è lo stimatore della varianza campionaria e

$$s_r^2 + s_c^2 = \frac{1}{n-1} \sum_i \sum_j (y_{ij} - \bar{y})^2 \tag{6}$$

è lo stimatore della varianza elementare (campionaria e di risposta), corretto sotto le identiche assunzioni per la correttezza della (2).

Va tuttavia precisato, come rileva Fellegi (1964), che la (5) stima in realtà ρI, dove

$$I = \sigma_r^2/(\sigma_r^2 + \sigma_c^2) \tag{7}$$

è un indice detto "di incoerenza" (Pritzker e Hanson, 1962), considerato che misura la frazione di varianza di risposta sul totale delle varianze elementari. Pertanto, la (5) sottostima il vero valore del coefficiente di correlazione intraclasse.

3. *Analisi dei risultati*

La valutazione dell'effetto dell'intervistatore sulle stime si basa prevalentemente sui dati riportati nell'ultima colonna delle Tabb. 3-6 e nella Tab. 7. Le principali considerazioni traibili dall'analisi di questi dati sono riassunte nei punti che seguono.

(a) I valori negativi o nulli dei coefficienti di correlazione intraclasse appartengono a variabili oggettive che è quasi impossibile rilevare in modo scorretto, come il sesso del rispondente o l'essere la persona di riferimento in servizio di leva. Nel complesso, si tratta di un numero esiguo di variabili.

(b) Errori del rilevatore particolarmente bassi, e in alcuni strati anche nulli, si trovano per le modalità dello *stato civile* "celibe o nubile" e "separato, divorziato, già coniugato"; per le modalità della *condizione occupazionale* "disoccupato", "in cerca di prima occupazione", "inabile"; per le modalità della *posizione nella professione* "lavoratore in proprio", "operaio", "apprendista" e per la variabile "in cerca di lavoro pur essendo attualmente occupato".

(c) I valori calcolati di ρ non sono grandi. Nei comuni capoluogo della Lombardia, infatti, nessuno supera 0,02; 9 su 40 variabili esaminate lo superano negli altri comuni della stessa regione; 5 sempre su 40 e 14 su 40 lo superano, rispettivamente, nei comuni capolugo e negli altri comuni della Campania.

Tab. 3: *Stima, varianza di stima, varianza elementare e varianza dell'intervistatore delle principali grandezze rilevate con l'indagine sulle forze di lavoro nei comuni capoluogo della Lombardia* [a]

Variabili	stima	varianza stima		varianza elementare	varianza interv.	
	$(\bar{x})$	$s^2(\bar{x})$	$100\frac{s(\bar{x})}{\bar{x}}$	$s_c^2+s_r^2$	$s^2(\bar{\alpha})$	$100\frac{s^2(\bar{\alpha})}{s^2(\bar{x})}$
Sesso: femmine	0,530	1,30 (-4)	2,2	0,25	VN	VN
Grado istruzione (> 6 anni) (b): Analfabeti	0,256 (-2)	2,17 (-6)	57,5	0,26 (-2)	1,06 (-6)	49,0
Alfabeti privi di titolo	0,799 (-1)	0,48 (-4)	8,7	0,74 (-1)	1,70 (-5)	35,5
Licenza elementare	0,323	0,88 (-4)	2,9	0,22 (-1)	VN	VN
Licenza media	0,306	9,38 (-5)	3,2	0,22 (-1)	2,95 (-6)	3,1
Diploma scuola media superiore	0,218	1,51 (-4)	5,6	0,17	7,84 (-5)	51,9
Laurea	0,701 (-1)	3,36 (-5)	8,3	0,65 (-1)	5,85 (-6)	17,4
Stato civile: Celibe, nubile	0,250	1,34 (-4)	5,4	0,19	1,21 (-5)	9,0
Coniugati	0,600	3,2 (-4)	3,0	0,24	9,73 (-5)	30,4
Vedovi	0,130	1,8 (-4)	10,3	0,11	6,08 (-5)	33,8
Separati/divorziati	0,018	6,4 (-6)	14,1	0,18 (-1)	VN	VN
Condizione (> 14 anni) (c): Occupati	0,470	1,25 (-4)	2,4	0,25	8,41 (-6)	6,7
Disoccupati	0,982 (-2)	5,46 (-6)	23,8	0,97 (-2)	9,59 (-7)	17,6
In cerca I occup.	0,206 (-1)	3,37 (-6)	8,9	0,20 (-1)	VN	VN
Servizio leva (solo maschi)(d)	0,910 (-2)	1,02 (-5)	35,1	0,90 (-2)	9,77 (-7)	9,6
Casalinghe (solo femmine)(e)	0,285	3,53 (-4)	6,6	0,20	1,75 (-4)	49,6
Studenti	0,105	4,50 (-5)	6,4	0,94 (-1)	8,74 (-7)	1,9
Inabili	0,122 (-1)	0,66 (-5)	21,1	0,12 (-1)	1,02 (-6)	15,4
Ritirati dal lavoro	0,221	2,16 (-4)	6,6	0,17	1,37 (-4)	62,8
Altro	0,374 (-2)	0,21 (-5)	38,7	0,37 (-2)	3,98 (-7)	19,0

segue Tab.3: *Stima, varianza di stima, varianza elementare e varianza dell'intervistatore delle principali grandezze rilevate con l'indagine sulle forze di lavoro nei comuni capoluogo della Lombardia* [(a)]

Variabili	stima $(\bar{x})$	varianza stima $s^2(\bar{x})$		$100\frac{s(\bar{x})}{\bar{x}}$	varianza elementare $s_c^2+s_r^2$		varianza interv. $s^2(\bar{\alpha})$		$100\frac{s^2(\bar{\alpha})}{s^2(\bar{x})}$
Effettuato ore lav. (> 14 anni) (c):	0,465	1,82	(-4)	2,9	0,25		6,56	(-4)	36,1
Ore lavorate (chi ha lavorato almeno 1 ora)(g)	39,4	5,28		0,6	0,75	(+2)	1,74	(-2)	32,2
Occupati:									
a tempo pieno (f)	0,967	8,00	(-5)	0,9	0,32	(-1)	4,82	(-5)	60,3
permanenti	0,966	2,63	(-5)	0,5	0,33		1,27	(-5)	48,3
solo un'attività	0,997	0,14	(-5)	0,1	0,28	(-2)	VN		VN
Ore di lavoro abituali	39,0	12,6	(-1)	0,9	0,65	(+2)	6,17	(-2)	4,9
Occupati per posizione nella professione (f):									
Imprenditore	0,589	(-2) 1,25	(-5)	60,0	0,52	(-2)	1,01	(-5)	81,0
Libero professionista	0,343	(-1) 0,25	(-4)	14,6	0,33	(-1)	1,13	(-5)	45,4
Lavoratore in proprio	0,108	0,90	(-5)	2,8	0,96	(-1)	VN		VN
Coadiuvante	0,406	(-1) 0,28	(-4)	13,0	0,39	(-1)	2,75	(-6)	9,8
Dirigente, funzionario	0,279	(-1) 0,42	(-4)	23,3	0,27	(-1)	1,45	(-5)	34,6
Impiegato, intermedio	0,436	1,60	(-4)	2,9	0,25		7,71	(-6)	4,8
Operaio	0,352	9,04	(-5)	2,7	0,23		VN		VN
Apprendista	0,839	(-2) 0,38	(-5)	23,2	0,83	(-2)	VN		VN
Lavorante a domicilio conto imprese	0,701	(-2) 0,61	(-5)	35,2	0,70	(-2)	3,27	(-6)	53,6
Cerca lavoro:									
- dipendente	0,332	(-1) 1,22	(-5)	10,5	0,32	(-1)	VN		VN
- ma ha già una occupazione	0,514	(-2) 0,25	(-5)	30,8	0,51	(-2)	4,09	(-7)	16,4
- in proprio	0,701	(-2) 0,61	(-5)	35,2	0,70	(-2)	3,21	(-6)	52,6
Non cerca lavoro:									
- ma potrebbe lavorare a part. condizioni	0,280	(-2) 0,53	(-5)	82,2	0,28	(-2)	4,09	(-6)	77,2
- perché ne ha uno	0,952	2,55	(-5)	0,2	0,46	(-1)	6,64	(-6)	26,0

(a) VN: valore nullo o negativo; (b) n=2.340 ; (c) n=2.138 ; (d) n=989 ; (e) n=1.149 ; (f) n=1.005 ; (g) n=994.

Tab. 4: *Stima, varianza di stima, varianza elementare e varianza dell'intervistatore delle principali grandezze rilevate con l'indagine sulle forze di lavoro negli altri comuni della Lombardia* [(a)]

Variabili	stima	varianza stima		varianza elementare		varianza interv.	
	$(\bar{x})$	$s^2(\bar{x})$	$100\frac{s(\bar{x})}{\bar{x}}$	$s_c^2+s_r^2$		$s^2(\bar{\alpha})$	$100\frac{s^2(\bar{\alpha})}{s^2(\bar{x})}$
Sesso: femmine	0,509	4,39 (-5)	1,3	0,25		VN	VN
Grado istruzione (> 6 anni) (b): Analfabeti	0,262 (-2)	1,06 (-6)	39,4	0,26	(-2)	7,99 (-8)	7,5
Alfabeti privi di titolo	0,115	1,06 (-4)	9,0	0,10		6,74 (-5)	63,6
Licenza elementare	0,348	2,84 (-4)	4,8	0,23		1,99 (-4)	70,1
Licenza media	0,325	2,94 (-4)	5,3	0,22		2,01 (-4)	72,2
Diploma scuola media superiore	0,175	1,43 (-4)	6,8	0,14		8,95 (-5)	62,6
Laurea	0,344 (-1)	2,14 (-5)	13,5	0,33	(-1)	8,99 (-6)	42,0
Stato civile: Celibe, nubile	0,268	8,32 (-5)	3,4	0,20		1,46 (-5)	17,6
Coniugati	0,654	2,21 (-4)	2,3	0,23		1,44 (-4)	64,7
Vedovi	0,657 (-1)	9,11 (-5)	14,5	0,61	(-1)	7,00 (-5)	76,8
Separati/divorziati	0,124	0,72 (-5)	21,7	0,12	(-1)	2,97 (-6)	41,3
Condizione (> 14 anni) (c): Occupati	0,530	5,12 (-4)	4,3	0,25		4,10 (-4)	80,7
Disoccupati	0,136 (-1)	1,44 (-5)	27,8	0,13	(-1)	8,88 (-6)	61,7
In cerca I occup.	0,161 (-1)	1,41 (-5)	23,3	0,16	(-1)	7,57 (-6)	53,8
Servizio leva (solo maschi)(d)	0,144 (-1)	1,20 (-5)	23,9	0,14	(-1)	VN	VN
Casalinghe (solo femmine)(e)	0,273	2,68 (-4)	6,0	0,20		1,09 (-4)	40,0
Studenti	0,777 (-1)	7,38 (-5)	11,1	0,71	(-1)	4,42 (-5)	59,9
Inabili	0,123 (-1)	0,79 (-5)	22,7	0,12	(-1)	2,84 (-6)	36,0
Ritirati dal lavoro	0,198	4,98 (-4)	0,6	0,16		4,37 (-4)	87,7
Altro	0,537	1,55 (-6)	23,2	0,53	(-2)	VN	VN

Segue Tab.4: *Stima, varianza di stima, varianza elementare e varianza dell'intervistatore delle principali grandezze rilevate con l'indagine sulle forze di lavoro negli altri comuni della Lombardia* [a]

Variabili	stima	varianza stima		varianza elementare		varianza interv.	
	$(\bar{x})$	$s^2(\bar{x})$	$100\frac{s(\bar{x})}{\bar{x}}$	$s_c^2+s_r^2$		$s^2(\bar{\alpha})$	$100\frac{s^2(\bar{\alpha})}{s^2(\bar{x})}$
Effettuato ore lav. (> 14 anni) (c)	0,512	3,35 (-4)	3,6	0,25		2,32 (-4)	69,3
Ore lavorate (chi ha lavorato almeno 1 ora)(g)	0,406 (+2)	7,58 (-2)	0,7	0,95	(+2)	4,31 (-2)	56,8
Occupati:							
a tempo pieno (f)	0,945	7,79 (-5)	0,9	0,52	(-1)	3,70 (-5)	47,5
permanenti	0,966	3,16 (-5)	0,6	0,32	(-1)	6,06 (-6)	19,2
solo un'attività	0,988	4,48 (-5)	0,7	0,12(	-1)	3,60 (-5)	80,5
Ore di lavoro abituali	0,399	3,66 (-1)	1,5	0,11	(+3)	2,83 (-1)	77,3
Occupati per posizione nella professione (f):							
Imprenditore	0,125 (-1)	1,11 (-5)	26,6	0,12	(-1)	1,39 (-6)	12,5
Libero professionista	0,312 (-1)	4,34 (-5)	21,1	0,30	(-1)	1,97 (-5)	45,9
Lavoratore in proprio	0,172	2,63 (-4)	9,4	0,14		1,51 (-4)	57,5
Coadiuvante	0,452 (-1)	4,37 (-5)	14,6	0,43	(-1)	9,77 (-6)	22,4
Dirigente, funzionario	0,117 (-1)	1,90 (-5)	37,3	0,11	(-1)	9,93 (-6)	52,3
Impiegato, intermedio	0,331	2,93 (-4)	5,2	0,22		1,20 (-4)	41,0
Operaio	0,368	3,17 (-4)	4,8	0,23		1,35 (-4)	42,6
Apprendista	0,156 (-1)	1,39 (-5)	23,9	0,15	(-1)	1,82 (-6)	13,1
Lavorante a domicilio conto imprese	0,125 (-1)	2,51 (-5)	40,1	0,12	(-1)	1,54 (-5)	61,6
Cerca lavoro:							
- dipendente	0,347 (-1)	3,94 (-5)	18,1	0,33	(-1)	2,55 (-5)	64,9
- ma ha già una occupazione	0,165 (-1)	8,57 (-6)	17,7	0,16	(-1)	1,82 (-6)	21,4
- in proprio	0,785 (-2)	6,98 (-6)	33,6	0,78	(-2)	3,76 (-6)	53,9
Non cerca lavoro:							
- ma potrebbe lavorare a part. condizioni	0,144 (-1)	7,45 (-5)	59,7	0,15	(-1)	6,88 (-5)	92,3
- perché ne ha uno	0,926	1,09 (-4)	1,1	0,68	(-1)	8,12 (-5)	74,5

(a) VN: valore nullo o negativo; (b) n=2.676 ; (c) n=2.421 ; (d) n=1.147; (e) n=1.247 ; (f) n=1.282 ; (g) n=1.240.

Tab. 5: *Stima, varianza di stima, varianza elementare e varianza dell'intervistatore delle principali grandezze rilevate con l'indagine sulle forze di lavoro nei comuni capoluogo della Campania* [(a)]

Variabili	stima	varianza stima		varianza elementare		varianza interv.	
	$(\bar{x})$	$s^2(\bar{x})$	$100\,\frac{s(\bar{x})}{\bar{x}}$	$s_c^2 + s_r^2$		$s^2(\bar{\alpha})$	$100\,\frac{s^2(\bar{\alpha})}{s^2(\bar{x})}$
Sesso: femmine	0,517	0,86 (-4)	1,8	0,25		VN	VN
Grado istruzione (> 6 anni) (b): Analfabeti	0,405 (-2)	0,39 (-5)	48,8	0,40	(-2)	2,10 (-6)	53,9
Alfabeti privi di titolo	0,136	1,32 (-4)	8,5	0,12		8,16 (-5)	61,8
Licenza elementare	0,252	1,89 (-4)	5,5	0,19		1,04 (-4)	55,2
Licenza media	0,273	2,12 (-4)	5,3	0,20		1,23 (-4)	58,0
Diploma scuola media superiore	0,266	0,66 (-4)	3,1	0,20		VN	VN
Laurea	0,693 (-1)	3,46 (-5)	8,5	0,65	(-1)	5,47 (-6)	15,8
Stato civile: Celibe, nubile	0,344	1,96 (-4)	4,0	0,14	(-2)	1,03 (-4)	52,3
Coniugati	0,593	2,91 (-4)	2,9	0,24		1,89 (-4)	65,1
Vedovi	0,593 (-1)	7,42 (-5)	46,1	0,56	(-1)	5,07 (-5)	68,3
Separati/divorziati	0,413	1,31 (-6)	27,8	0,41	(-2)	VN	VN
Condizione (> 14 anni) (c): Occupati	0,382	2,83 (-4)	4,4	0,24		1,62 (-4)	57,1
Disoccupati	0,413 (-2)	6,26 (-6)	60,7	0,4	(-2)	4,13 (-6)	66,4
In cerca I occup.	0,768 (-1)	6,82 (-5)	10,8	0,71	(-1)	3,16 (-5)	46,4
Servizio leva (solo maschi) (d)	0,324 (-2)	2,38 (-6)	47,6	0,32	(-2)	VN	VN
Casalinghe (solo femmine) (e)	0,454	5,77 (-4)	5,3	0,25		3,34 (-4)	57,8
Studenti	0,156	2,28 (-4)	9,7	0,13		1,60 (-4)	70,1
Inabili	0,774 (-2)	0,39 (-5)	25,5	0,77	(-2)	VN	VN
Ritirati dal lavoro	0,125	1,48 (-4)	9,7	0,11		9,13 (-5)	61,7
Altro	0,980	9,47 (-6)	31,4	0,97	(-2)	4,45 (-6)	46,9
Effettuato ore lav. (> 14 anni) (c)	0,379	2,69 (-4)	4,3	0,24		1,48 (-4)	55,0

Segue Tab.5: *Stima, varianza di stima, varianza elementare e varianza dell'intervistatore delle principali grandezze rilevate con l'indagine sulle forze di lavoro nei comuni capoluogo della Campania* [(a)]

Variabili	stima $(\bar{x})$		varianza stima $s^2(\bar{x})$		$100\frac{s(\bar{x})}{\bar{x}}$	varianza elementare $s_c^2+s_r^2$		varianza interv. $s^2(\bar{\alpha})$		$100\frac{s^2(\bar{\alpha})}{s^2(\bar{x})}$
Ore lavorate (chi ha lavorato almeno 1 ora) (g)	0,381	(+2)	5,96	(-2)	0,6	0,95	(+2)	1,08	(-2)	18,1
Occupati:										
a tempo pieno (f)	0,974		1,43	(-4)	1,2	0,25	(-1)	1,10	(-4)	76,8
permanenti	0,979		5,57	(-5)	0,8	0,21	(-1)	4,64	(-5)	83,4
solo un'attività	1,000		-		-	-		-		-
Ore di lavoro abituali	0,380		1,57	(-1)	1,0	0,97	(+2)	23,7	(-2)	15,1
Occupati per posizione nella professione (f):										
Imprenditore	0,534	(-2)	0,58	(-5)	45,1	0,53	(-2)	3,57	(-6)	61,6
Libero professionista	0,534	(-1)	4,73	(-5)	12,9	0,51	(-1)	2,56	(-5)	54,1
Lavoratore in proprio	0,138		5,12	(-5)	5,2	0,12		3,22	(-7)	0,6
Coadiuvante	0,267	(-1)	0,19	(-4)	16,3	0,26	(-1)	8,28	(-6)	43,6
Dirigente, funzionario	0,334	(-1)	1,91	(-5)	13,1	0,32	(-1)	5,26	(-6)	27,5
Impiegato, intermedio	0,481		1,29	(-4)	2,4	0,25		2,22	(-5)	17,2
Operaio	0,253		1,15	(-4)	4,2	0,19		3,33	(-5)	29,0
Apprendista	0,106	(-1)	1,74	(-5)	25,7	0,11	(-1)	2,94	(-6)	39,8
Lavorante a domicilio conto imprese	0,464	(-2)	1,12	(-5)	72,1	0,47	(-2)	9,13	(-6)	81,5
Cerca lavoro:										
-dipendente	0,104		1,06	(-4)	9,9	0,93	(-1)	6,61	(-4)	62,4
-ma ha già una occupazione	0,250	(-2)	0,25	(-5)	63,2	0,51	(-2)	2,62	(-7)	10,5
-in proprio	0,464	(-2)	1,12	(-5)	72,1	0,47	(-2)	9,24	(-6)	82,5
Non cerca lavoro:										
-ma potrebbe lavorare a part. condizioni	0,920	(-2)	2,91	(-4)	58,6	0,92	(-2)	2,52	(-5)	86,8
-perché ne ha uno	0,880		1,50	(-4)	1,4	0,11		1,06	(-4)	70,4

(a) VN: valore nullo o negativo. (b) n=2.223; (c) n=1.939; (d) n=925; (e) n=1.147; (f) n=740; (g) n=735.

Tab. 6: *Stima, varianza di stima, varianza elementare e varianza dell'intervistatore delle principali grandezze rilevate con l'indagine sulle forze di lavoro negli altri comuni della Campania* [(a)]

Variabili	stima	varianza stima		varianza elementare		varianza interv.	
	$(\bar{x})$	$s^2(\bar{x})$	$100\frac{s(\bar{x})}{\bar{x}}$	$s_c^2+s_r^2$		$s^2(\bar{\alpha})$	$100\frac{s^2(\bar{\alpha})}{s^2(\bar{x})}$
Sesso: femmine	0,518	4,89 (-5)	1,3	0,25		VN	VN
Grado istruzione (> 6 anni) (b): Analfabeti	0,737 (-2)	8,30 (-6)	39,1	0,38	(-4)	2,19 (-6)	26,5
Alfabeti privi di titolo	0,180	3,32 (-4)	10,1	0,15		1,73 (-4)	52,2
Licenza elementare	0,315	2,75 (-4)	5,3	0,22		8,88 (-5)	32,3
Licenza media	0,337	3,46 (-4)	5,5	0,22		1,59 (-4)	45,9
Diploma scuola media superiore	0,136	3,06 (-4)	12,8	0,12		1,70 (-4)	55,4
Laurea	0,243 (-1)	2,79 (-5)	21,7	0,24	(-1)	7,98 (-6)	28,6
Stato civile: Celibe, nubile	0,313	1,43 (-4)	3,8	0,22		VN	VN
Coniugati	0,611	4,56 (-4)	3,5	0,24		2,98 (-4)	65,4
Vedovi	0,746 (-1)	1,59 (-4)	16,9	0,69	(-1)	1,14 (-4)	71,5
Separati/divorziati	0,176 (-2)	1,53 (-6)	70,4	0,17		3,45 (-7)	22,5
Condizione (> 14 anni) (c): Occupati	0,337	4,33 (-4)	6,2	0,22		2,39 (-4)	55,3
Disoccupati	0,158 (-1)	1,66 (-5)	10,5	0,16	(-1)	2,93 (-6)	17,3
In cerca I occup.	0,119	1,89 (-4)	11,6	0,13	(-2)	9,70 (-5)	51,3
Servizio leva (solo maschi) (d)	0,127 (-1)	2,67 (-5)	40,7	0,13	(-1)	3,70 (-6)	13,7
Casalinghe (solo femmine) (e)	0,534	1,21 (-3)	6,5	0,25		8,08 (-4)	66,8
Studenti	0,112	1,11 (-4)	9,4	0,99	(-1)	2,31 (-5)	21,0
Inabili	0,878 (-2)	2,42 (-5)	56,0	0,88	(-2)	1,66 (-5)	69,4
Ritirati dal lavoro	0,790 (-1)	2,28 (-4)	19,1	0,73	(-1)	1,63 (-4)	71,7
Altro	0,448 (-1)	2,27 (-4)	33,6	0,43	(-1)	1,89 (-4)	83,3

Segue Tab.6: *Stima, varianza di stima, varianza elementare e varianza dell'intervistatore delle principali grandezze rilevate con l'indagine sulle forze di lavoro negli altri comuni della Campania* [a]

Variabili	stima	varianza stima		varianza elementare		varianza interv.	
	$(\bar{x})$	$s^2(\bar{x})$	$100\frac{s(\bar{x})}{\bar{x}}$	$s_c^2+s_r^2$		$s^2(\bar{\alpha})$	$100\frac{s^2(\bar{\alpha})}{s^2(\bar{x})}$
Effettuato ore lav. (>14 anni) (c)	0,320	5,67 (-4)	7,4	0,22		3,77 (-4)	66,5
Ore lavorate (chi ha lavorato almeno 1 ora)(g)	0,397 (+2)	3,92 (-2)	0,5	81,85		VN	VN
Occupati:							
a tempo pieno (f)	0,940	5,31 (-4)	2,5	0,57	(-1)	3,95 (-4)	74,4
permanenti	0,911	7,49 (-4)	3,0	0,81	(-1)	5,55 (-4)	74,1
solo un'attività	0,969	2,48 (-4)	1,6	0,3	(-1)	1,74 (-4)	70,2
Ore di lavoroia abituali	0,371 (+2)	6,39 (-1)	2,2	158,73		2,29 (-1)	35,9
Occupati per posizione nella professione (f):							
Imprenditore	0,521 (-2)	9,31 (-6)	58,6	0,52	(-2)	VN	VN
Libero professionista	0,313 (-1)	6,60 (-5)	26,0	0,30	(-1)	VN	VN
Lavoratore in proprio	0,190	2,95 (-4)	9,0	0,15		VN	VN
Coadiuvante	0,182 (-1)	7,98 (-5)	49,1	0,18	(-1)	3,34 (-5)	41,8
Dirigente, funzionario	0,130 (-1)	4,02 (-5)	48,8	0,13	(-1)	6,60 (-6)	16,5
Impiegato, intermedio	0,289	1,02 (-3)	11,1	0,21		4,97 (-4)	48,7
Operaio	0,435	7,16 (-4)	6,2	0,25		6,44 (-4)	9,0
Apprendista	0,182 (-1)	3,87 (-5)	34,2	0,18	(-1)	VN	VN
Lavorante a domicilio conto imprese	-	-	-	-		-	-
Cerca lavoro:							
- dipendente	0,213	2,49 (-4)	7,4	0,17		1,22 (-4)	49,0
-ma ha già una occupazione	0,167 (-1)	1,49 (-5)	23,1	0,16	(-1)	4,56 (-7)	3,1
- in proprio	0,149 (-1)	2,13 (-5)	31,0	0,15	(-1)	8,37 (-6)	39,9
Non cerca lavoro:							
- ma potrebbe lavorare a part. condizioni	0,219 (-1)	4,53 (-5)	30,7	0,21	(-1)	2,66 (-5)	59,2
- perché ne ha uno	0,733	6,02 (-4)	3,3	0,19		4,33 (-4)	72,0

(a) VN: valore nullo o negativo. (b) n=2.676 ; (c) n=2.421 ; (d) n=1.174; (e) n=1.247 ; (f) n=1.282 ; (g) n=776.

Tab. 7: *Coefficente di correlazione intra-intervistatore (x 1.000) delle principali grandezze rilevate con l'indagine sulle forze di lavoro, per strato* [(a)]

Variabili	Strato			
	Lombardia		Campania	
	Capoluoghi	Altri	Capoluoghi	Altri
Sesso: femmine	-0,4	-3,4	-1,2	-6,2
Grado istruzione (> 6 anni):				
Analfabeti	3,8	0,6	7,3	5,2
Alfabeti privi di titolo	2,1	12,8	9,5	19,6
Licenza elementare	-0,2	16,7	7,8	6,9
Licenza media	0,1	18,4	8,7	10,7
Diploma scuola media superiore	4,1	11,9	-1,6	24,3
Laurea	0,8	5,2	1,2	5,7
Stato civile:				
Celibe, nubile	0,6	1,4	6,2	-0,1
Coniugati	3,6	12,1	11,0	16,3
Vedovi	5,0	21,8	12,7	21,4
Separati/divorziati	-0,3	4,7	-1,5	2,6
Condizione (> 14 anni): Occupati	0,3	31,4	9,7	13,9
Disoccupati	0,8	12,5	14,1	2,5
In cerca I occup.	-2,7	9,1	6,3	12,1
Servizio leva (solo maschi)	1,0	-0,5	4,9	3,9
Casalinghe (solo femmine)	7,7	10,5	19,1	41,8
Studenti	0,1	11,8	17,0	3,1
Inabili	0,8	4,6	-0,1	24,8
Ritirati dal lavoro	7,1	52,2	11,5	29,4
Altro	1,0	-2,4	6,4	58,5

Segue Tab.7: *Coefficente di correlazione intra-intervistatore (x 1.000) delle principali grandezze rilevate con l'indagine sulle forze di lavoro, per strato* [(a)]

Variabili	Strato			
	Lombardia		Campania	
	Capoluoghi	Altri	Capoluoghi	Altri
Effettuato ore lav.(>14 anni):	2,4	17,8	8,9	226,8
Ore lavorate (chi ha lavorato almeno 1 ora)	2,1	8,6	1,6	-2,5
Occupati :				
a tempo pieno	13,7	13,6	61,9	90,2
permanenti	3,5	3,6	32,2	89,6
solo un'attività	-0,4	58,7	-	76,9
Ore di lavoro abituali	8,6	50,7	3,5	18,8
Occupati per posizione nella professione:				
Imprenditore	15,8	2,2	9,6	-11,7
Libero professionista	3,1	12,6	7,2	-6,5
Lavoratore in proprio	-2,4	20,4	0,0	-10,2
Coadiuvante	0,6	4,4	4,5	24,4
Dirigente, funzionario	4,9	16,5	2,3	6,8
Impiegato, intermedio	0,3	10,5	1,3	30,9
Operaio	-1,0	11,2	2,5	3,4
Apprendista	-0,6	2,3	4,0	-6,7
Lavorante a domicilio conto imprese	4,3	24,3	27,9	-
Cerca lavoro:				
-dipendente	-0,3	14,6	10,1	10,7
-ma ha già una occupazione	0,7	2,2	0,7	0,4
-in proprio	4,3	9,2	27,9	7,5
Non cerca lavoro:				
-ma potrebbe lavorare a part. condizioni	13,4	89,8	38,6	16,4
-perché ne ha uno	1,3	22,7	14,2	29,0

(a) Stima ottenuta applicando lo stimatore (5).

Il carico medio di lavoro degli intervistatori è peraltro considerevole (Tab. 2), anche se mediamente non si discosta da quello attuale delle RTFL. E' proprio il carico di lavoro un'importante causa del considerevole effetto del rilevatore sulla inefficienza delle stime: pur con apprezzabili diversità tra gli strati considerati, l'incidenza mediana dell'errore dell'intervistatore sulle stime si aggira attorno al 44% (corrispondente ad un valore mediano di ρ=0,0023). Il raddoppio del numero di intervistatori ridurrebbe questo valore, *ceteris paribus*, al 28% della varianza di stima e la varianza di stima stessa al 78% dell'attuale valore.

(d) Le variabili, o modalità, più colpite da errori di classificazione attribuibili all'intervistatore sono:

(d1) quelle con frequenze molto basse (non superiori all'1%) o molto alte (superiori al 92%) nella popolazione. Tra i valori più elevati si trovano infatti
– nei comuni capoluogo della Lombardia, la modalità "imprenditori" e "non cerca lavoro ma potrebbe lavorare a particolari condizioni", che hanno frequenze inferiori all'1%, e la modalità "occupati che esercitano l'attività a tempo pieno", con frequenza 97%;
– negli altri comuni lombardi, le modalità "non cerca attivamente un lavoro ma potrebbe lavorare a particolari condizioni" (frequenza 1%), "svolge soltanto un'attività" (frequenza 99%) e "ha già un lavoro e non ne cerca un altro (frequenza 93%);
– nei comuni capoluogo campani, le modalità "lavorante a domicilio per conto di imprese", "cerca lavoro in proprio" (frequenze inferiori all'1%) e "ha un'occupazione permanente" e "esercita l'attività a tempo pieno" (frequenze attorno al 98%), "non cerca lavoro ma potrebbe lavorare a particolari condizioni" e "ha già un lavoro e non ne cerca un altro" (frequenze 88-92%);
– negli altri comuni campani, le modalità "ha un'occupazione permanente", "esercita un'attività a tempo pieno" e "svolge soltanto un'attività" (frequenze tra il 91 e il 97%). La ragione per cui alla frequenza estrema dell'accadimento di un fenomeno è associato un alto rischio di errore dell'intervistatore è eminentemente tecnica, essendo minima la varianza naturale del fenomeno con cui si rapporta quella introdotta dai rilevatori. Naturalmente, esistono le debite eccezioni, sia entro la classe di variabili evidenziate, sia tra le altre variabili caratterizzate da frequenze intermedie e da forte condizionamento del rilevatore (vedi oltre);

(d2) le modalità residuali di variabili nominali: tra quelle esaminate la modalità "altra condizione", concernente la *condizione occupazionale*, e in un certo senso la modalità "perché ha già un lavoro" tra le motivazioni per cui un individuo *non ha cercato lavoro* nel periodo di riferimento (è una modalità essenzialmente residuale quando sia percepita, dal rispondente e/o dal rilevatore, come una scappatoia offerta per fornire comunque una motivazione);

(d3) la variabile *ore di lavoro abituali*, verosimilmente interpretabile in più modi perché riguarda un intervallo temporale aperto a sinistra. Va ricordato che nelle norme per la compilazione del modello di rilevazione si prevedono possibilità di comportamenti discrezionali, giacché si specifica: "Per le persone che non possono fornire la risposta perché le ore prestate variano

considerevolmente da una settimana all'altra o da un mese all'altro indicare 00". Si prevede inoltre che "Per i cassa-integrati [si debba] indicare l'orario abituale dell'unità locale cui sono addetti", il che lascia ulteriore, notevole spazio all'indecisione dell'intervistato ed espone la sua risposta al rischio del condizionamento in ragione dell'interpretazione datane dall'intervistatore;

(d4) la modalità "ritirato dal lavoro", probabilmente per motivi definitori. Nelle norme per la compilazione del modello di rilevazione si precisa che è *ritirato dal lavoro* "Chi ha cessato un'attività lavorativa per raggiunti limiti di età, invalidità o altra causa. La figura del ritirato dal lavoro non coincide necessariamente con quella del pensionato in quanto non sempre il ritirato dal lavoro gode di una pensione". È immediato notare una qualche sovrapposizione con la definizione di *inabile al lavoro*, che sarebbe "Chi è fisicamente impossibilitato a svolgere un'attività lavorativa", senza una dovuta specificazione sulla precedente attività. Per comprendere se durante la rilevazione le due figure in questione siano rimaste in molti casi indistinte e, pertanto, siano state collocate in una delle due categorie di risposta secondo l'arbitrio del rilevatore, è stata ricalcolata la varianza dell'intervistatore accorpando le due categorie[3]. Il confronto con i valori delle Tabb. 3-7, che sono addirittura inferiori, indica che la difficoltà di distinguere le due posizioni non è la causa principale di un così grande impatto dell'errore dell'intervistatore sulle stime. Ci sembra, comunque, che il comprendere tra i ritirati dal lavoro anche gli invalidi necessiti per lo meno di una precisazione sull'aver precedentemente lavorato, oppure sulle cause dell'invalidità (civile, da lavoro), in modo da distinguere le due categorie in questione;

(d5) la modalità "vedovo" inerente allo *stato civile*. E' arduo proporre una ipotesi verosimile sull'origine dell'errore dell'intervistatore per questa modalità (sempre ammesso che la domanda sia posta agli intervistati); può darsi che il coniuge sopravissuto di una coppia già convivente (non coniugata) non si dichiari vedovo;

(d6) le modalità più frequenti della variabile *titolo di studio*: "licenza elementare", "licenza di scuola media" e "diploma di scuola media inferiore". Le modalità di risposta si presentano nel questionario in una successione monotòna e non dovrebbero dare adito ad errori, se si seguisse l'avvertenza posta in calce al modello di indicare il più alto titolo di studio conseguito (con la precisazione esemplificativa che se una persona ha studiato fino alla terza o alla quarta elementare va indicato "Nessun titolo", se ha studiato fino alla seconda liceo va indicato "Licenza media inferiore").

La presenza di un così massiccio errore del rilevatore (superiore al 50% della varianza globale per il diploma di scuola media nei comuni capoluogo della Lombardia, superiore al 60% per tutte le tre modalità negli altri comuni lombardi, quasi il 60% per la licenza elementare e la licenza di scuola media

---

3 I valori dell'incidenza percentuale della varianza dell'intervistatore sulla varianza globale di stima sono: 64,8 (corrispondente a ρ=0,0078) nei comuni capoluogo della Lombardia, 87,5 (ρ=0,0547) negli altri comuni lombardi, 63,4 (ρ=0,0126) nei comuni capoluogo della Campania, e 75,9 (ρ=0,0360) negli altri comuni campani.

nei comuni capoluogo della Campania, attorno al 50% per i titoli di scuola media negli altri comuni campani) fa sospettare che la indubbia ordinalità dello schema classificatorio sia compromessa dalla omissione, o dalla mancata specificazione nelle avvertenze, dei gradi di istruzione intermedi tra quelli oggi canonici (tra gli altri, la 3ª e la 4ª elementare quando il superamento con successo di questi gradi costituiva proscioglimento dall'obbligo scolastico, i diplomi professionali di 2 o 3 anni ottenibili dopo la licenza media, le qualifiche professionali del sistema extrascolastico, i diplomi universitari e para-universitari) o di quelli paragonabili a quelli elencati nel questionario (per esempio, l'avviamento professionale al posto della scuola media inferiore, i titoli ottenuti all'estero).

Si può supporre che le persone che hanno percorso sistemi scolastici diversi, dubitando della correttezza dell'auto-classificazione, possano aver interagito con l'intervistatore ed abbiano lasciato a questi l'onere della scelta. Una indicazione indiretta dei motivi della inaccuratezza nella classificazione in funzione del sistema scolastico si ricava suddividendo l'insieme delle persone con almeno 6 anni di età in due categorie: "età fino a 49 anni" (rappresenta, non senza qualche sovrapposizione con la classe complementare, gli italiani assoggettati all'obbligo di frequentare la scuola almeno fino all'ottenimento della licenza elementare) e "età 50 anni e più" per lo studio della modalità *possesso della licenza elementare*; "età fino a 39 anni" (questa classe rappresenta, in prima approssimazione, le persone obbligate all'ottenimento della licenza di scuola media) ed "età 40 e più" per lo studio della modalità *possesso della licenza di scuola media*. L'incidenza della variabilità causata dagli intervistatori sulla variabilità globale di stima e il valore dei coefficienti di correlazione intra-intervistatore nelle 4 classi descritte sono generalmente inferiori a quelli registrati sull'intero insieme di unità.

Il valore mediano delle stime di ρ nelle sottoclassi è più contenuto di quello calcolato per l'intero insieme di unità: per la modalità *possesso della licenza elementare* scende da 0,0074 a 0,006 per la sottoclasse fino a 49 anni e a 0,002 per la sottoclasse complementare; per la modalità *possesso della licenza di scuola media* scende da 0,0097 a 0,0052 per la sottoclasse fino a 49 anni e a 0,0090 per la sottoclasse complementare. Anche se i motivi delle diversità non traspaiono dai valori calcolati, è fuori discussione l'esistenza di una relazione tra l'effetto rilevatore e l'età delle persone intervistate.

(e) Esiste una notevole disparità tra strati per quanto riguarda l'incidenza dell'errore del rilevatore. Il valore mediano della percentuale di errore è 17,5 nei comuni capoluogo della Lombardia, che sono risultati quelli con il tasso d'errore dell'intervistatore più contenuto, 53,8 negli altri comuni della Lombardia, 54 nei comuni capoluogo della Campania e 41,8 negli altri comuni della Campania. La disparità è ancor più evidente se, invece della varianza dell'intervistatore - la quale, come è evidente dalla (1), misura l'effetto degli intervistatori sulle stime relativizzato quasi unicamente con il numero di rilevatori -, si considera la varianza correlata di

risposta denotata nella sez. 3 con ρ_{or}^2 o l'analoga misura data dal coefficiente di correlazione intra-intervistatore stimato con la (5).
I coefficienti di correlazione intraclasse evidenziano ancor più le differenze di qualità della rilevazione nei comuni capoluogo della Lombardia rispetto agli altri e un generale miglior esito della rilevazione nei comuni capoluogo (sia della Lombardia, sia della Campania). Infatti, il coefficiente di correlazione mediano è 0,001 nei comuni capoluogo e 0,012 negli altri comuni della Lombardia, e 0,007 nei comuni capoluogo e 0,011 negli altri comuni della Campania.

5. *Considerazioni riepilogative e proposte*

Le analisi svolte indicano in modo inequivocabile che l'effetto dell'intervistatore sulle stime inerenti alle forze di lavoro è così rilevante, ma anche così diverso da stima a stima, da dover essere necessariamente considerato nella valutazione dell'attendibilità e, prima ancora, nella predisposizione di strumenti di rilevazione adeguati a farvi fronte.

Tra gli strumenti di rilevazione ci sembrano perfezionabili sia il questionario che le istruzioni per i rilevatori. Vari suggerimenti minuti sono deducibili dai commenti con cui si è accompagnata l'analisi dei risultati dell'indagine. Si tratta, tuttavia, di indicazioni inerenti per lo più ai picchi dell'effetto dell'intervistatore. Pertanto, come si è evidenziato nell'introduzione, l'esperimento di cui si è dato conto ha soprattutto il valore di una testimonianza della necessità di investimenti intellettuali da parte dell'Istat e degli studiosi intenzionati a riflettere sul tema del controllo dell'errore di rilevazione. Ci sembra dunque indispensabile estendere la compenetrazione su ambiti territoriali più vasti, onde rappresentare l'effetto dell'intera rete di rilevazione attiva per la RTFL, tentando di misurare anche la relazione tra l'errore di rilevazione, varie modalità di svolgimento dell'indagine (rispondente designato o chi per lui e correlato effetto dell'ottenimento delle risposte da un solo adulto per famiglia, orari e luogo di svolgimento della rilevazione, numero di occasioni di rilevazione presso la stessa famiglia, numero di tentativi a vuoto prima di ottenere l'intervista, *etc.*) e caratteristiche della struttura di rilevazione (tipo di organizzazione locale, dimensione dell'area interessata, caratteristiche sociali ed economiche dei comuni, *etc.*). Con riferimento alla riduzione dell'errore del rilevatore, va parallelamente avviato il processo di identificazione delle caratteristiche personali e di formazione dei rilevatori connesse ai valori più bassi dell'errore.

L'esame dell'errore dei rilevatori non può essere disgiunto dalle altre fonti d'errore, in particolare da quello dei rispondenti, per il quale è necessario studiare modalità di accertamento. Va comunque detto che i disegni di accertamento dell'errore di rilevazione mirano a quantificare il livello d'errore, ma non indicano, se non indirettamente, come debbono essere formulati i quesiti.

Per ridefinire i quesiti 'a rischio', si rendono indispensabili indagini spe-

rimentali sul campo, nelle quali si pongono a confronto modi alternativi di porre i quesiti (e ancora di ordinarli, di evidenziarli e di collegarli nel questionario), elenchi di modalità di risposta con vari gradi di analiticità *etc.*. Tutto ciò che, per ragioni pratiche, non può essere precisato nel questionario deve essere descritto analiticamente nelle istruzioni per i rilevatori.

22

UN'INDAGINE SUPPLETIVA ALLA RILEVAZIONE SULLE FORZE DI LAVORO INCENTRATA SULLA STORIA LAVORATIVA

Ugo Trivellato, Ignazio De Nicola, Ersilia Di Pietro, Giulio Ghellini, Enrico Rettore e Nicola Torelli *

1. *Introduzione*

Nel corso degli anni '70 e '80, nei Paesi sviluppati si sono avute significative innovazioni nell'informazione statistica sull'occupazione e l'offerta di lavoro: nei suoi contenuti, così come negli strumenti di rilevazione.

Hanno fatto da sfondo e da stimolo a tali innovazioni le profonde trasformazioni avvenute, e tuttora in atto, nel mercato del lavoro (e più in generale nel sistema sociale), e il parallelo riesame dei paradigmi interpretativi dell'economia e della sociologia del lavoro. Sia pure in estrema sintesi, non possiamo non ricordare almeno tre tratti salienti di queste modificazioni, di specifico rilievo per l'argomento affrontato in questa sede: (i) il dinamismo e la differenziazioni nella quantità e nella qualità delle prestazioni lavorative richieste, e più in generale "un contesto [del mercato del lavoro] divenuto più articolato, denso di interrelazioni spesso insospettate, maggiormente dominato dall'incertezza" (Bruno, 1987, p. 111); (ii) la crescente importanza dei fattori operanti dal lato dell'offerta, cioè a dire l'accresciuto ruolo decisionale di individui e famiglie, nel concorrere a determinare la situazione del mercato del lavoro; (iii) il maggior peso assunto dalle politiche sociali, di *welfare* o specificamente di promozione dell'occupazione, a fronte degli squilibri nel mercato del lavoro.

Le implicazioni salienti di queste trasformazioni, e della connessa riconsiderazione degli schemi interpretativi, sull'informazione statistica in tema di mercato del lavoro sono state considerate, in termini generali, nel cap. 1. Conviene ora incentrare l'attenzione sulle esigenze più marcatamente inno-

---

* Una versione preliminare del progetto di indagine suppletiva è stata presentata al Seminario "Progetto di indagine suppletiva per l'acquisizione di informazioni aggiuntive alla rilevazione sulle forze di lavoro", organizzato dall'Istat e tenutosi a Milano il 30 maggio 1988. I commenti e le critiche emerse nel corso del Seminario sono stati quanto mai utili. In particolare, siamo grati a D. Ciravegna, O. Chillemi, U. Colombino, C. D' Apice, P. De Sandre, C.J. Flinn, L. Frey, R. Leoni, A. Martini e M. Schenkel per osservazioni e suggerimenti su precedenti versioni del questionario. La definitiva messa a punto dell'indagine è, infine, avvenuta nell'ambito di un'apposito gruppo di lavoro istituito presso l'Istat, del quale facevano parte, insieme con noi, L. Bernardi, S. Bordignon, E. Caperdoni, A. Ciriello, M.A. De Marchis, L. Fabbris, P. Manfroni, M. Masselli, I. Sanetti e A. Zuliani (coordinatore).

vative, in materia di contenuti e di modi di rilevazione. Scontando drastiche semplificazioni, non è forse azzardato ricondurre i fabbisogni informativi prevalenti a due aspetti (o meglio, ad un aspetto, del quale vanno evidenziate due dimensioni complementari): l'attenzione alla *dinamica* ed alle *determinanti* dei comportamenti dei soggetti - individui e/o famiglie - sul mercato del lavoro. Non ci dilunghiamo a motivare questa proposizione, perché se per un verso l'intero capitolo poggia su di essa, per un altro verso nel seguito avremo modo di presentare varie argomentazioni che, direttamente o indirettamente, la corroborano. D'altra parte, è appena ovvio osservare che l'importanza di informazioni, e di analisi, sulla dinamica e le determinanti delle scelte dei soggetti rispetto al lavoro discende implicitamente dalle modificazioni del mercato del lavoro cui abbiamo inizialmente accennato.

È poi palese che informazioni di questo tipo non sono raccolte, o lo sono in misura affatto inadeguata, dalle usuali rilevazioni sull'occupazione e l'offerta di lavoro. Emerge pertanto la necessità di procedere a nuove indagini (o comunque all'approntamento di nuove basi di dati, combinando convenientemente quelli esistenti), che si caratterizzano di massima per fornire *dati longitudinali* sul mercato del lavoro *integrati* con altre informazioni: cioè a dire, (i) dati riferiti agli stessi individui in più unità di tempo, (ii) e relativi, oltre che alla partecipazione al lavoro e ai tradizionali caratteri ascrittivi (sesso, età, ecc.), ad altre variabili di potenziale rilievo per l'analisi del comportamento dei soggetti rispetto al lavoro (tipicamente, l'istruzione conseguita e in corso, il salario e il reddito, variabili attinenti alla famiglia, variabili attinenti al mercato del lavoro locale)[1].

Esempi di produzione di basi di dati di questo tipo si hanno per numerosi Paesi, fra i quali spiccano gli Stati Uniti e non manca l'Italia (per alcune rassegne vedi Kalachek, 1979, Ashenfelter e Solon, 1982, e con specifico riguardo alle indagini sulla transizione scuola-lavoro Trivellato, 1980).

Oggetto di questo capitolo è la presentazione di un'indagine suppletiva alla rilevazione trimestrale sulle forze di lavoro (nel seguito RTFL), condotta in via sperimentale in Lombardia nel maggio-giugno 1989, indagine che si qualifica appunto per l'acquisizione di informazioni longitudinali integrate sul mercato del lavoro.

L'indagine suppletiva è presentata sia nei suoi aspetti e finalità generali, sia in alcune caratteristiche salienti, segnatamente per quanto riguarda la struttura del questionario. Per cogliere le specificità dell'indagine in questione, ed apprezzarne correttamente potenzialità e limiti, nella sez. 2 sono brevemente discussi i principali tipi di fonti di dati longitudinali integrati sul mercato del lavoro. La sez. 3 è dedicata all'illustrazione delle finalità e dei tratti salienti dell'indagine suppletiva. Nella sez. 4 è presentato e discusso con qualche dettaglio il questionario. La sez. 5 dà brevemente conto dell'iter seguito per la messa a punto dell'indagine e della modalità di suo svolgi-

---

1 È di qualche interesse ricordare che, ai fini dello studio dell'evoluzione della struttura economica e della connessa definizione di politiche, una parallela attenzione si manifesta per dati longitudinali riferiti alle imprese (vedi U.S. Bureau of the Census, 1982, e McGuckin, 1990).

mento. La conclusiva sez. 6, infine, tratteggia sommariamente le opportunità di analisi che l'indagine consente e prime linee di lavoro definite in proposito.

2. *Fonti di dati longitudinali integrati sul mercato del lavoro e loro usi*

Secondo Ashenfelter e Solon (1982, pp. 109-110), la ragione principale dello sviluppo di informazioni longitudinali sta nel fatto che "una convincente ricerca su numerose questioni di politica pubblica richiede dati longitudinali. Invero, senza dati longitudinali alcuni importanti temi di ricerca non possono essere affrontati affatto. Per esempio, un'appropriata politica nei riguardi della povertà e della disoccupazione [....] poggia in parte sul fatto che l'esperienza di questi stati da parte delle famiglie o degli individui sia tipicamente transitoria o cronica. Istantanee sezionali dei poveri o dei disoccupati, che mettono a fuoco individui diversi in tempi diversi, non possono rivelare quanti di questi poveri o disoccupati a un certo tempo restano poveri o disoccupati in tempi successivi. Siffatte questioni di persistenza in uno stato necessitano di seguire longitudinalmente gli stessi individui"[2].

Se Ashenfelter e Solon mettono l'accento sull'informazione longitudinale (perché questo è l'argomento del loro scritto, e forse anche perché nell'esperienza statunitense l'integrazione dei dati sulla partecipazione al lavoro con dati sul reddito è patrimonio acquisito), Siesto (1982, p. 117) richiama, condividendoli, diffusi orientamenti sull' "astrattezza della definizione di forza di lavoro avulsa dal contesto familiare e sociale e dalle esperienze conseguite nel ciclo di vita precedente, e quindi [sul]la necessità di indagare congiuntamente sulla partecipazione al lavoro e sui redditi personali e familiari".

L'esigenza di informazioni integrate, lì prospettata in termini generali, diventa poi ancor più pregnante per disegnare politiche sociali e valutarne gli effetti. Infatti, "è proprio il fatto di dover intervenire più frequentemente e più articolatamente che nel passato a suggerire più che l'opportunità il dovere, per i soggetti di *policy*, di raccogliere sistematicamente informazioni nell'ambito nel quale gli interventi vengono attuati". E ancora, di raccogliere informazioni sufficientemente complete per consentire la misura e l'analisi a livello micro, giacché gli interventi di *policy* si collocano "in sistemi in cui i soggetti sono divenuti più 'smaliziati', più capaci di un agire strategico in grado di anticipare molte delle possibili mosse dei soggetti di regolazione" (Bruno, 1987, pp.117-123)[3].

L'utilità di dati longitudinali integrati sul mercato del lavoro risalta in definitiva per tre tipi di ricerca: (i) la misura e l'analisi dei cambiamenti di stato degli individui nel tempo; (ii) lo studio dei processi e delle determinanti che caratterizzano il comportamento dei soggetti, individui o famiglie, rispetto

---

2 L'altro fattore di sviluppo di basi di dati longitudinali è, sempre secondo Ashenfelter e Solon (1982, p. 110), nelle possibilità tecniche e nei costi non più proibitivi per il loro approntamento, tramite sistemi computerizzati di gestione dei dati.

3 Per più generali riflessioni sulle connessioni fra problemi di politica sociale, bisogni conoscitivi e basi di informazione in società complesse, vedi, ad es., Martinotti (1987).

al lavoro; (iii) le analisi che richiedono il controllo degli effetti di variabili non osservabili (tipicamente, caratteristiche non osservate specifiche degli individui). Questo elenco può sembrare scarno e piuttosto astratto. Esempi di ciascun tipo di utilizzazione, che sovente mostrano il considerevole rilievo pratico di siffatte analisi, sono tuttavia assai numerosi in letteratura (vedi, ad es., Ashenfelter e Solon, 1982, Solon, 1989, e Duncan, Juster e Morgan, 1987, per brevi rassegne; Heckman e Singer, 1985, e Blundell e Walker, 1986, per raccolte di saggi).

Se si guarda alle maggiori fonti di dati longitudinali integrati sul mercato del lavoro, e al modo con cui sono state generate, emergono peraltro differenze non trascurabili, che si riflettono poi in maniera significativa sulle opportunità di analisi che esse offrono. Con qualche semplificazione, è possibile distinguere quattro tipologie principali:

(a) dati longitudinali ottenuti tramite abbinamento di *records* amministrativi riguardanti soggetti coinvolti in programmi di sicurezza sociale (tipicamente, lavoratori dipendenti) o di *welfare* (tipicamente, disoccupati cui è corrisposta un'indennità di disoccupazione);
(b) dati longitudinali raccolti tramite rilevazioni retrospettive che acquisiscono informazioni sull'esperienza passata dei soggetti, rilevazioni tipicamente collegate a indagini correnti sulle forze di lavoro;
(c) dati longitudinali raccolti tramite indagini *panel*, cioè tramite rilevazioni genuinamente longitudinali, che comportano interviste periodiche degli stessi individui. V'è da osservare che è questa una tipologia dallo spettro particolarmente largo. Fra le varie indagini *panel* si registrano infatti marcate differenzazioni, segnatamente con riguardo alla popolazione obiettivo, alla lunghezza del periodo di tempo sul quale si estende la rilevazione longitudinale, all'ampiezza ed ai contenuti prevalenti delle informazioni acquisite;
(d) dati longitudinali acquisiti in connessione con quasi-esperimenti di politica economica e sociale (Fienberg, Singer e Tanur, 1985). L'esempio emblematico al riguardo è rappresentato dai *negative income tax programs* condotti in alcune aree degli Stati Uniti a partire dalla fine degli anni '60.

Una persuasiva presentazione delle caratteristiche di differenti fonti di dati longitudinali integrati, e dei relativi vantaggi e limitazioni, è in Ashenfelter e Solon (1982) e, con specifica attenzione ai problemi di disegno delle indagini e di analisi statistica dei risultati, in Duncan e Kalton (1987). Ci limitiamo qui a poche notazioni sul tipo di fonte di maggior interesse ai nostri fini, cioè sulle rilevazioni retrospettive collegate a indagini correnti sulle forze di lavoro.

Le esperienze più mature, che è istruttivo avere presenti, sono quelle collegate alla *Current Population Survey* (CPS) statunitense ed alla *Labour Force Survey* canadese. V'è da osservare, innanzitutto, che entrambe queste indagini hanno un piano di campionamento con rotazione, cioè tale per cui una frazione del campione di *households* di una data indagine è intervistata più volte in indagini successive. Già le rilevazioni correnti presentano, dunque, un aspetto di *panel*, e possono fornire dati longitudinali previo collegamento delle informazioni fornite da un individuo nella sequenza di indagini

cui partecipa. Alle rilevazioni correnti si aggiungono poi svariate indagini suppletive, fra le quali spiccano appunto quelle orientate a completare in chiave retrospettiva le informazioni sulla storia lavorativa e ad integrare i dati sulla partecipazione al lavoro con dati sui salari ed i redditi e sulla partecipazione a programmi pubblici di politica sociale.

Negli Stati Uniti, assolve a questo scopo il *CPS March Supplement*, rilevazione suppletiva condotta annualmente a marzo con periodo di riferimento l'anno precedente, e attenta in particolare alla rilevazione del reddito (U.S. Bureau of the Census, 1987). A partire dal 1983, a questa rilevazione si è affiancata una nuova indagine dal disegno e dagli obiettivi più ambiziosi: la *Survey of Income and Program Participation* (SIPP) (per una diffusa presentazione, vedi David, 1985). La SIPP si qualifica propriamente come una indagine *panel* affatto autonoma e assai impegnativa, che segue il campione di individui per un periodo di 32 mesi attraverso 9 occasioni di intervista distanziate l'una dall'altra di 4 mesi. La sua genesi è peraltro strettamente intrecciata con la riflessione critica sui limiti informativi del *CPS March Supplement*, soprattutto in relazione ai fabbisogni conoscitivi per la definizione e la valutazione degli effetti di programmi di politica sociale. Essa mantiene inoltre la caratteristica di indagine su larga scala condotta su un campione probabilistico dell'intera popolazione degli Stati Uniti. Non è dunque improprio richiamarla in questo contesto, non fosse altro che per trarre lumi da un'esperienza di indagini suppletive che si è venuta significatamente evolvendo nell'arco del secondo dopoguerra.

Nel Canada, sulla scorta di un'esperienza decennale dal 1987 l'indagine suppletiva sulla storia lavorativa è stata completamente ridisegnata. La nuova indagine, denominata *Labour Market Activity Survey* (LMAS), acquisisce informazioni su un periodo di osservazione di 24 mesi tramite 2 successive interviste, ciascuna con periodo di riferimento l'anno solare precedente (anche questa è, dunque, una indagine *panel*, sia pure limitata a 2 occasioni). Quanto ai contenuti, essa si caratterizza per la maggior enfasi data alla storia lavorativa e per un approccio di rilevazione *employer-specific*, incentrato cioè sugli episodi di occupazione identificati col datore di lavoro (Statistics Canada, 1987).

Alcuni dei vantaggi e delle limitazioni delle rilevazioni retrospettive collegate alle indagini correnti sulle forze di lavoro, rispetto ad altre fonti di dati longitudinali integrati, emergono con sufficiente chiarezza considerando gli aspetti di disegno e di contenuti che le caratterizzano. È palese, innanzitutto, che si tratta di indagini relativamente poco costose, in quanto sfruttano il *frame*, il piano di campionamento e parte dell'impianto delle operazioni sul campo delle rilevazioni correnti[4].

---

4 Naturalmente il giudizio di minore onerosità vale rispetto a dati longitudinali raccolti con apposite indagini. Va da sé che la tipologia di generazione di dati longitudinali meno costosa è costituita dalla collazione longitudinale di dati già raccolti nel processo di amministrazione di programmi di sicurezza sociale o di *welfare*. I dati longitudinali di origine amministrativa, peraltro, soffrono sovente di severe limitazioni, conseguenti: (i) per un verso, al fatto che possono riferirsi ad una popolazione diversa da quella obiettivo (tipicamente, riguardano solo i beneficiari del programma in questione, sicché manca il gruppo di controllo per corrette valutazioni dell'impatto distributivo del programma - o di sue modificazioni - e delle interrelazioni

In secondo luogo, esse presentano il vantaggio di acquisire informazioni longitudinali su un campione solitamente piuttosto grande, pertinente all'intera popolazione. Esse consentono quindi una valutazione completa dello stato, della dinamica e delle determinanti comportamentali dell'offerta di lavoro, così come analisi differenziate sull'intero spettro dei sottogruppi di interesse. A fronte di questi vantaggi, il rovescio della medaglia sta nel fatto che l'informazione longitudinale si estende su un periodo relativamente breve - l'anno o giù di lì -, ed inoltre che la rilevazione è circoscritta ad un insieme di variabili ragionevolmente contenuto (e sovente misurato in maniera non molto 'fine').

Infine, la raccolta di informazioni longitudinali tramite una rilevazione retrospettiva invece che a mezzo di un'indagine panel presenta sì sicuri meriti, ma altrettanto indubbi svantaggi. Tra i meriti di una singola intervista retrospettiva sono da annoverare la rapidità di acquisizione dell'informazione longitudinale, il contenimento dei costi, l'eliminazione del problema di progressiva riduzione del campione (e di conseguente distorsione connessa alla selezione), la minore vulnerabilità ad alcuni tipi di errore di risposta (ad es., la diversa descrizione di uno stesso lavoro in due successive interviste, con conseguente erronea registrazione di un cambiamento dell'occupazione). D'altra parte, i problemi di distorsione da selezione del campione possono presentarsi sotto altra forma in indagini suppletive retrospettive collegate alle rilevazioni correnti sulle forze di lavoro, quando - ed è questa la regola - l'indagine retrospettiva sia condotta su un gruppo di rotazione che è già stato presente varie volte nel campione: in tal modo, infatti, risultano esclusi quanti hanno cambiato residenza dopo la loro entrata nel campione, cioè a dire soggetti verosimilmente caratterizzati da un comportamento più dinamico rispetto al lavoro. E, soprattutto, l'acquisizione di informazioni per via retrospettiva è soggetta ad errori di memoria, tanto più marcati e diversificati quanto più lungo è il periodo di riferimento.

Queste brevi notazioni in chiave comparativa, su potenzialità e limiti delle rilevazioni retrospettive collegate alle indagini correnti sulle forze di lavoro, tornano utili per mettere a fuoco gli obiettivi conoscitivi che si possono ragionevolmente perseguire con questo strumento di rilevazione, e insieme le condizioni per conseguirli in maniera soddisfacente.

3. *L'indagine suppletiva alla rilevazione sulle forze di lavoro: finalità e caratteristiche salienti*

3.1. *Orientamenti generali*

Nel motivare l'indagine suppletiva e nel delinearne le caratteristiche

---

[segue nota]

del programma stesso con la dinamica comportamentale dei soggetti); (ii) per un altro verso, al fatto che l'informazione raccolta per scopi amministrativi è tipicamente ridotta, e talora 'deformata' dalla specifica finalità amministrativa, rispetto a quanto richiesto a fini di ricerca, sia essa genericamente applicata o di supporto alla definizione e alla valutazione di politiche.

salienti, è necessario prendere le mosse da alcuni riferimenti alla RTFL (vedi Istat, 1978, e per ragguagli essenziali il cap. 1). I rischi di richiamare l'ovvio sono naturalmente dietro l'angolo, e ci scusiamo sin d'ora se la brevità degli accenni non varrà ad evitarli del tutto. Com'è noto, la RTFL ha cadenza trimestrale e adotta un disegno campionario con rotazione del tipo 2-2-2: cioè a dire, ogni famiglia resta nel campione per due indagini consecutive, esce dal campione per le due successive, quindi rientra per altre due indagini. Salvi mutamenti di residenza o altri fenomeni di *attrition*, ciascuna famiglia è quindi intervistata quattro volte nell'arco di sedici mesi: ai trimestri 1, 2, 5 e 6.

Tenendo presente la struttura del questionario, l'informazione raccolta tramite la RTFL si qualifica dunque per i seguenti tratti: (i) la condizione rispetto al lavoro è rilevata con riferimento alla prima settimana di quattro mesi dell'anno - gennaio, aprile, luglio e ottobre -, e per parte dei soggetti è completata da succinte informazioni retrospettive (sui precedenti lavorativi per i non occupati, sulla durata della ricerca di occupazione per quanti hanno dichiarato di cercare attivamente un lavoro); (ii) la rilevazione di variabili contestuali riguarda i tradizionali caratteri demografici (sesso, età, relazione col capofamiglia, presenza o assenza dal comune di residenza), il titolo di studio e le eventuali attività formative seguite nella quattro settimane precedenti quella dell'intervista; (iii) le possibilità di ricostruire la dinamica della condizione rispetto al lavoro nell'arco dei 16 mesi di presenza degli individui nel campione sono interessanti, ma limitate all'osservazione della permanenza/cambiamento di stato ad intervalli trimestrali e annuale[5].

La contestuale considerazione di questi tratti dell'informazione corrente sulle forze di lavoro e delle riflessioni svolte in precedenza in tema di dati longitudinali integrati sul mercato del lavoro, offre convincenti motivazioni e plausibili orientamenti per un'indagine suppletiva retrospettiva[6]. In termini ancora piuttosto generali, gli indirizzi informatori dell'indagine suppletiva sono riassumibili in poche proposizioni, attinenti per un verso ai contenuti e per l'altro verso al disegno.

Circa i contenuti, appare ragionevole incentrare la rilevazione su due aspetti: (i) la ricostruzione della storia lavorativa - cioè della durata della permanenza nei singoli stati e della sequenza delle transizioni da stato a stato - su un arco di tempo contenuto, l'anno o al più il periodo di 16 mesi

---

5 Più propriamente, si tratta di stima - e non di osservazione - della permanenza/cambiamento di stato, perché tale informazione dipende dalla procedura di accoppiamento dei dati individuali in indagini successive (vedi Moriani, 1981, e il cap. 7) ed è particolarmente sensibile ad errori di misura (vedi Poterba e Summers, 1986, e il cap. 18). Si noti ancora che, comportando il piano di rotazione l'uscita della famiglia dal campione ai trimestri 3 e 4 della sequenza, non si può generare un *panel* di dati equispaziati esteso a 6 occasioni d'indagine.

6 Sintetiche, ma importanti indicazioni sull'argomento si trovano anche nel rapporto conclusivo della Commissione di studio dell'Istat per un sistema informativo del lavoro (Istat, 1984). In particolare, è significativo il seguente passo: "Va fortemente potenziata la pratica di coordinare alla rilevazione campionaria trimestrale sulle forze di lavoro indagini saltuarie volte a investigare temi collaterali. Tali indagini potrebbero servire per approfondimenti monografici ricorrenti ovvero occasionali (ad es., *ricostruzione retrospettiva delle esperienze di lavoro di un anno*, [...] proseguimento dell'osservazione longitudinale su un sub-campione di famiglie, ecc.) e inoltre anche per obiettivi metodologici" (Istat, 1984, p. 57; l'enfasi del corsivo è nostra).

di presenza (in parte saltuaria) di una famiglia nel campione; (ii) l'acquisizione di informazioni sul reddito da lavoro e sull'insieme degli altri redditi personali e familiari, di massima per lo stesso periodo di riferimento. All'attenzione in via prioritaria su questi due aspetti, è poi plausibile affiancare la rilevazione di un ridotto insieme di altre variabili, riguardanti l'individuo e la famiglia, che consenta una sintetica descrizione del contesto entro il quale si colloca la dinamica della partecipazione dei singoli al mercato del lavoro[7].

Circa il disegno dell'indagine, occorre osservare che l'acquisizione di informazioni sulla storia lavorativa e sui redditi si presenta ad un tempo complessa e delicata: per l'intrinseca difficoltà di rilevazione di buona parte delle variabili in questione; per gli specifici problemi che pone il ricorso all'interrogazione retrospettiva su un arco di tempo dell'ordine di 12-16 mesi; per verosimili atteggiamenti di reticenza degli intervistati rispetto ai quesiti sul reddito. La soddisfacente riuscita di una siffatta indagine dipende quindi in maniera cruciale da elevati standards dell'impianto metodologico, in tutte le sue fasi e segnatamente nella definizione del questionario, nel reclutamento e nella formazione degli intervistatori, nelle operazioni sul campo. Da qui discende l'orientamento a realizzare l'indagine suppletiva su un segmento territorialmente circoscritto del campione della rilevazione principale, caratterizzandola anche in senso sperimentale: per saggiarne le potenzialità conoscitive e per valutare impegni e problemi nella sua effettuazione, in vista della possibilità che tale indagine (com'è ovvio, convenientemente riveduta proprio sulla base di questa esperienza) venga estesa all'intero territorio nazionale e diventi ricorrente.

Vediamo ora meno sommariamente gli aspetti salienti dell'indagine suppletiva, con riguardo nell'ordine alle informazioni che ci si è proposti di acquisire e all'impianto tecnico della rilevazione.

3.2. *Contenuti informativi*

Già si è detto che le informazioni da acquisire in via prioritaria vertono su due aspetti: (i) la storia lavorativa; (ii) il reddito da lavoro e l'insieme dei redditi personali e familiari.

Quanto alla ricostruzione della storia lavorativa, è parso opportuno porsi nella condizione di poter utilizzare al meglio i dati tratti dalla RTFL sfruttandone la parziale struttura longitudinale: vuoi a fini di confronto e di parziale controllo delle informazioni sulla condizioni rispetto al lavoro, vuoi eventualmente per integrare le informazioni fornite dall'indagine suppletiva con i dati

---

7 Obiettivi più ambiziosi, quanto a lunghezza della storia lavorativa e/o quanto a completezza e dettaglio delle informazioni (sulle prestazioni lavorative, sui flussi di reddito, sul contesto individuale e familiare) e/o quanto a documentazione delle interrelazioni fra i programmi di politica sociale e dinamica dei comportamenti di individui e famiglie, a nostro avviso rimandano a indagini con disegno parecchio più complesso, di tipo longitudinale, delle quali la SIPP è forse l'esempio più significativo.

delle precedenti rilevazioni trimestrali[8]. Ciò ha indotto ad estendere la rilevazione retrospettiva della storia lavorativa oltre l'anno, al periodo di 16 mesi di (teorica) permanenza dell'individuo nel campione della RTFL. Tale periodo è stato poi dilatato a 17 mesi, perché, alla luce della complessità di alcune operazioni previe di raccordo con la RTFL (vedi la sez. 5), si è deciso di svolgere l'indagine suppletiva non già immediatamente a ridosso della rilevazione di aprile 1989, bensì a cavallo fra maggio e giugno 1989, sicché è parso conveniente estendere la ricostruzione della storia lavorativa appunto anche a maggio.

In sede di rilevazione retrospettiva della condizione rispetto al lavoro, per di più su un periodo non breve, è apparso d'altra parte problematico mantenere rigorosamente gli stessi criteri adottati nella RTFL per l'identificazione degli occupati e delle persone in cerca di occupazione. In particolare, ovvi dubbi sono emersi circa la possibilità di rilevare la condizione rispetto al lavoro secondo gli usuali tre stati - occupato, persona in cerca di occupazione (o *tout court* disoccupato), non forza di lavoro (o inattivo) - avendo come periodo di riferimento la settimana[9]. Ed è risultato francamente impraticabile impiegare la complessa griglia di quesiti, ben sei, utilizzati nella RTFL per giungere a classificare le persone nei tre stati (vedi, ad es., Rettore, Torelli e Trivellato, 1988, pp. 74-75, e inoltre De Nicola, 1989). Pur restando a definizioni concettualmente coerenti con quelle dell'indagine trimestrale, è stato in definitiva inevitabile introdurre due lievi varianti.

(a) La prima variante consiste nell'assumere come periodo elementare di riferimento per la ricostruzione della storia lavorativa il mese. Di conseguenza, si modificano le definizioni di occupato e di disoccupato. Si considera occupato nel mese chi "ha svolto un lavoro retribuito anche solo per alcuni giorni" (a fronte della convenzione corrente, che classifica occupato chi "ha effettuato ore di lavoro nella settimana di riferimento"). Si considera in cerca di occupazione nel mese chi, essendo "senza lavoro retribuito, ha cercato attivamente un lavoro anche solo per un breve periodo" (a fronte della convenzione corrente, che richiede che la persona sia senza lavoro, disponibile a lavorare, ed abbia compiuto un'azione concreta di ricerca - peraltro anche non nell'immediato passato, cioè anche più di sei mesi prima dell'intervista -).

(b) La seconda variante consiste nell'assumere come disgiuntive sin dall'inizio, cioè dalla formulazione delle domande sulla condizione in cui il soggetto si trova, le condizioni di occupato e di persona in cerca di

---

8 Ciò vale, naturalmente, per il sub-campione dei 'sempre presenti' nella sequenza 1, 2, 5 e 6 delle rilevazioni correnti. Primi esperimenti di accoppiamento di dati individuali esteso alle 4 occasioni indicano peraltro che la frazione degli accoppiati (rispetto al campione teoricamente accoppiabile in base al piano di rotazione, in assenza di mutamenti di residenza e di altri fenomeni di *attrition*) è piuttosto elevata, dell'ordine dell'80% (vedi il cap. 7). Si noti ancora che di proposito si parla di parziali possibilità di controllo delle risposte sulla condizione rispetto al lavoro: per lievi varianti nella definizione di occupato e di persona in cerca di occupazione introdotte nell'indagine suppletiva, di cui si dirà tra poco nel testo principale; per i problemi che si pongono in generale nel confrontare una sequenza di dati rilevati longitudinalmente con informazioni parallele desunte da una rilevazione retrospettiva.

9 Un'interessante esperienza in tal senso è stata peraltro condotta in un'occasionale indagine suppletiva alla RTFL, condotta in Emilia-Romagna (vedi Brusco, Gennari, Marchesini e Salinas, 1986).

occupazione (invece di pervenire alla classificazione disgiuntiva sulla base delle risposte a molteplici quesiti - e non solo a quello sulla condizione dichiarata -, come accade nella RTFL). Ciò non esclude, peraltro, che nello stesso mese vi possa essere la compresenza di occupazione e di ricerca di lavoro. Con la variante definitoria in questione, infatti, resta sì preclusa la possibilità di rilevare periodi di ricerca *on-the-job*. Ma è certamente possibile che il periodo elementare di riferimento - il mese - risulti superiore a singoli periodi di occupazione o di ricerca di lavoro, e comunque non vi è motivo di attendersi che le transizioni da uno stato all'altro avvengano in corrispondenza dei mesi di calendario. Come avremo modo di chiarire nel seguito (vedi la sez. 4), si può dunque registrare compresenza dei due stati nello stesso mese, quando la transizione da uno stato all'altro avvenga nel corso del mese, oppure quando nell'arco del mese si avvicendino almeno tre diversi episodi di occupazione e di ricerca di cui uno completo.

Quanto al secondo aspetto di rilievo prioritario per l'indagine suppletiva, cioè la rilevazione del reddito da lavoro e dell'insieme degli altri redditi personali e familiari, è notoriamente di particolare complessità, per la spiccata reticenza dei rispondenti a quesiti su questi temi e per i problemi talora acuti di attendibilità che affliggono i dati sui redditi originati da *surveys*[10]. La diffusa pratica di evitare quesiti sul reddito nelle indagini correnti sulle forze di lavoro è dunque giustificata da ovvie ragioni, ma si scontra con la consapevolezza che l'interpretazione della condizione rispetto al lavoro di una persona non può prescindere dalla situazione economica della stessa e della sua famiglia. È parso quindi ragionevole sperimentare quesiti aggiuntivi sull'argomento, sulla scorta anche dell'esperienza del *CPS March Supplement* e dell'ampia riflessione che si è recentemente avuta nel nostro Paese sulle indagini nostrane sui bilanci delle famiglie (vedi in particolare Fabbris, Leti e Zuliani, 1986). Le scelte salienti operate al riguardo attengono:

(a) al periodo di riferimento. Per i dati sui redditi, è l'anno solare, che appare il più adatto per cogliere in maniera sufficientemente sintetica anche le componenti stagionali ed erratiche dei redditi, da lavoro e non da lavoro;
(b) alla rilevazione dei soli redditi. Di massima, sono quindi esclusi quesiti sul patrimonio e le sue variazioni;
(c) al grado di dettaglio nella classificazione dei redditi. Ci si allontana di proposito dall'unica domanda sul reddito totale, per l'ovvia necessità di disporre del reddito da lavoro, al fine di poterlo collegare con la quantità di lavoro prestato nell'anno, e per la diffusa consapevolezza che un'unica domanda produce risultati particolarmente inattendibili. Si mantiene tuttavia un grado di dettaglio delle domande piuttosto contenuto.

---

10 Anche nell'esperienza statunitense, che è forse la più matura sia per il contesto culturale collaborativo nei confronti di *surveys* sui temi più disparati e delicati sia per la lunga e attenta pratica di indagini sui redditi, i problemi di attendibilità dei dati sui redditi restano piuttosto severi: vedi, ad es., Herriot e Spiers (1975) e Duncan e Hill (1985).

3.3. *Caratteristiche tecniche*

Le scelte operate in tema di impianto tecnico dell'indagine suppletiva sono state dettate, oltre che dai suoi contenuti, dall'opportunità che essa non operasse come elemento perturbatore della sequenza delle rilevazioni correnti sulle forze di lavoro, stante anche la delicatezza di alcuni degli aspetti toccati. Gli indirizzi che è parso ragionevole adottare per il suo svolgimento sono risultati, in definitiva, i seguenti:

(a) Conduzione dell'indagine sul cosiddetto 'quarto uscente', cioè sulle famiglie comprese nella sezione che è presente per l'ultima volta nel campione. La motivazione basilare di questa scelta sta nel fatto che essa riduce l'eventualità di interazioni le quali possano concorrere a distorcere i dati rilevati correntemente con la RTFL (eventualità che viene poi del tutto esclusa dall'ulteriore indirizzo (c)). Essa sola, inoltre, consente di poter collegare ai dati dell'indagine retrospettiva quelli di fino a quattro precedenti rilevazioni correnti. È doveroso segnalare che tale scelta non è priva di controindicazioni, quali la distorsione da selezione che presenta la sezione uscente dal campione e i possibili effetti del *panel conditioning* (Kalton, Kasprzyk e McMillen, 1989). Il bilancio dei pro e dei contro non ci ha comunque indotto a metterla in discussione.

(b) Effettuazione dell'indagine in maniera coordinata alla RTFL di aprile 1989 (e nella prospettiva di svolgerla ricorrentemente, sempre in connessione con la rilevazione di aprile). In tal modo, il periodo retrospettivo di 17 mesi viene ad iniziare con gennaio 1988 (e in generale con gennaio dell'anno prima) e risulta agevole raccordare la raccolta di informazioni sulla storia lavorativa estese a 17 mesi con quella di dati sui redditi riguardanti l'anno solare precedente.

(c) Svolgimento dell'indagine alcune settimane dopo la rilevazione corrente, di massima nell'arco di due settimane a cavallo fra maggio e giugno 1989. Questa opzione è parsa preferibile all'alternativa di condurla come supplemento che segue senza soluzione di continuità l'usuale intervista della RTFL, per svariate ragioni. Essa fornisce ulteriori garanzie che non si verifichino indesiderate interazioni con la rilevazione corrente. Evita poi i rischi di un eccessivo carico per i rispondenti, e ancor più di scarsa cooperazione degli stessi nei confronti di un'intervista supplementare parecchio più impegnativa della principale. Consente infine di utilizzare la rilevazione principale per la rigorosa identificazione del campione di famiglie su cui condurre l'indagine suppletiva (per l'appunto il 'quarto uscente' effettivamente intervistato, con esclusione della possibilità di operare sostituzioni), ed inoltre per contatti che servano a motivare un elevato tasso di risposta e un atteggiamento collaborativo dei rispondenti.

(d) Limitazione dell'indagine, in via sperimentale, alla regione Lombardia. Tenendo conto che si tratta di una regione dove la RTFL è largamente 'sovracampionata', e pur scontando alcune riduzioni rispetto al 'quarto uscente' dettate da vincoli di bilancio (vedi la sez. 5), il campione dell'indagine suppletiva consta di circa 4.500 famiglie, quindi di 12-13 mila individui. A prima vista, può sembrare un campione esorbitante. Non va

dimenticato, tuttavia, che vi sono scarti notevoli fra campione di riferimento per i basilari dati demografici (tutti i componenti le famiglie campione), campione oggetto del contatto (le persone in età superiore a 14 anni alla data di svolgimento dell'indagine), campione di interesse sul quale si dilunga l'intervista (essenzialmente coloro che hanno sperimentato l'occupazione e/o la disoccupazione nei 17 mesi precedenti). È plausibile attendersi che le persone con episodi di occupazione e/o disoccupazione siano 6-7 mila, e che un gruppo di particolare interesse quali i disoccupati si aggiri sulle 700 unità. A fronte del numero di variabili rilevate, e dell'interesse a studiarne le interrelazioni, si tratta dunque di una numerosità campionaria ragionevole. D'altra parte, in tal modo ci si è posti nella condizione di sperimentare compiutamente il disegno dell'indagine anche nelle sue implicazioni organizzative, ed è questo un aspetto certo non secondario.

(e) Impegno ad assicurare elevati standards qualitativi nelle modalità di svolgimento delle operazioni sul campo, in relazione anche alla complessità del questionario. A tale scopo, ha assolto un importante ruolo l'attività di progettazione e di costante supervisione dell'insieme delle operazioni da parte di un gruppo di lavoro costituito presso l'Istat, attività della quale riferiamo brevemente nella sez. 5.

4. *Il questionario*

4.1. *La struttura generale del questionario*

Veniamo ora alla presentazione del questionario[11]. Esso consta di un frontespizio e di cinque Sezioni di quesiti, siglate da A ad E. Il frontespizio ricalca quello dell'attuale questionario della RTFL. Con la Sezione A sono raccolte notizie generali sulla composizione familiare. I quesiti delle Sezioni B e C sono dedicati all'acquisizione di informazioni sulla storia lavorativa: essi vertono ordinatamente sulla sintetica ricostruzione di tale storia e sulla descrizione dei principali lavori (e/o delle eventuali situazioni di disoccupazione e/o di inattività) nell'intero periodo di riferimento. Le ultime due Sezioni, D e E, riguardano infine la rilevazione dei redditi, rispettivamente a livello individuale e familiare.

È da tener presente che delle Sezioni A ed E si è prevista la somministrazione al solo capofamiglia, mentre le Sezioni B, C e D vanno somministrate a ciascun componente la famiglia in età superiore a 14 anni (o, secondo opportune regole, ad un *proxy*). Un semplice quadro sinottico delle cinque Sezioni secondo i contenuti ed i rispondenti è nella Tab. 1. La sequenza delle Sezioni risponde ad un ovvio ordine logico, con le domande sui redditi

11 Chi fosse interessato, può ottenere copia del questionario richiedendola all'Istat, Servizio Indagini sulle Famiglie.

collocate a conclusione dell'intervista. La sezione E, che è opportunamente l'ultima, vista la delicatezza delle informazioni richieste, va peraltro somministrata al capofamiglia, che è pure il primo intervistato, in quanto fornisce le notizie della Sezione A necessarie per identificare le altre persone da intervistare. Ciò ha richiesto particolare cura nel definire la strategia di conduzione delle interviste nell'ambito della famiglia, e nell'addestrare conseguentemente gli intervistatori.

Prima di procedere alla disamina delle diverse Sezioni, torna utile esplicitare alcuni orientamenti di fondo ai quali ci si è attenuti nella stesura del questionario. In buona sostanza, bastano brevi proposizioni per enunciarli e motivarli:

(a) di massima, ogni informazione viene rilevata con la formulazione diretta di una domanda, che l'intervistatore deve leggere. È invece evitata la predisposizione di prospetti che l'intervistatore deve compilare, interagendo in modi non controllati col rispondente. Palesemente, l'esito che ci si attende è una riduzione dell''effetto intervistatore';

(b) il questionario presenta una struttura di salti piuttosto complessa. Questo elaborato *skip pattern* consente il trattamento differenziato, all'interno del questionario, di sub-campioni con particolari percorsi lavorativi. Il risultato è un'accresciuta capacità di ricostruire storie lavorative complicate (e per loro natura più soggette a errori di rilevazione). D'altra parte, l'allungamento del questionario è solo apparente, riguarda cioè il testo, non i tempi di somministrazione;

(c) è diffuso il ricorso a domande di controllo per gli intervistatori (evidenziate in grassetto nel questionario). Lo scopo di queste domande, così come delle istruzioni di rinvio che affiancano le modalità di risposta di alcuni quesiti, è palesemente di guidare l'intervistatore a seguire lo *skip pattern* del questionario ed a svolgere correttamente l'intervista.

Tab. 1: *Quadro sinottico delle Sezioni del questionario, secondo i contenuti ed i rispondenti*

Contenuti prevalenti	Rispondenti	
	Capofamiglia	Persona>14 anni
Notizie generali sui componenti la famiglia	A	-
Storia lavorativa	-	B, C
Redditi	E	D

4.2. *Le notizie generali sui componenti la famiglia*

Prospetti e domande della Sezione A rispondono a due scopi principali: (i) identificare la composizione della famiglia alla data dell'indagine e raccogliere le usuali informazioni socio-demografiche della RTFL su tutti i componenti; (ii) ricostruire sinteticamente le variazioni nella composizione della famiglia nei 17 mesi precedenti.

Al primo scopo serve il Blocco A1. Di esso si è prevista la precompilazione sulla base dei dati della RTFL di aprile 1989. In sede di indagine suppletiva, l'intervistatore è quindi chiamato a registrare soltanto le sporadiche variazioni intervenute da metà aprile a fine maggio 1989.

La rilevazione della dinamica della composizione familiare avviene tramite i quesiti del Blocco A2. L'obiettivo conoscitivo è circoscritto ai movimenti sull'intervallo di 17 mesi, con esclusione dei movimenti che si compensano all'interno dell'intervallo. In altre parole si rilevano soltanto: (i) in uscita, le persone che facevano parte della famiglia il 1 gennaio 1988 e che non ne fanno parte a fine maggio 1989; (ii) in entrata, le persone attualmente componenti della famiglia che non ne facevano parte il 1 gennaio 1988. Pur con palesi limitazioni, queste informazioni colgono tratti salienti della dinamica della composizione familiare, che possono concorrere in maniera significativa a spiegare la storia lavorativa individuale.

Il problema non banale di una definizione ad un tempo persuasiva e praticabile di continuità della famiglia, che consenta di distinguere il caso della famiglia che permane (con eventuali movimenti di individui in entrata e/o uscita) dal caso di cambiamento della famiglia *tout court*, è risolto adottando un criterio di continuità parecchio restrittivo, ma coerente con la procedura di identificazione della famiglia nella RTFL. Tale criterio è dato dalla permanenza della stessa persona quale capofamiglia.

4.3. *La storia lavorativa*

Alla ricostruzione della storia lavorativa, e ad una meno sommaria descrizione di alcuni episodi o situazioni, è dedicata la parte centrale del questionario. La strategia di rilevazione si ispira liberamente, semplificandolo parecchio, al questionario della SIPP, con il quale ha in comune soprattutto: (i) lo *skip pattern* per l'identificazione di sub-campioni con diversi percorsi lavorativi; (ii) la rilevazione della sequenza degli stati per cui è transitato il rispondente e dell'associata collocazione temporale, mediante l'uso di un calendario contenente l'indicazione esplicita dei mesi che coprono il periodo di riferimento.

Questa strategia di rilevazione è chiaramente più persuasiva di quella che procede per sintetiche domande sulle durate trascorse nei diversi strati, adottata dal *CPS March Supplement*: sì più snella, ma con mediocri risultati in termini di completezza e di attendibilità dei dati che produce sulla storia lavorativa. D'altra parte, essa si mantiene abbastanza semplice, il che induce a preferirla all'approccio *employer-specific* della LMAS canadese: forse in

grado di evocare meglio le capacità mnemoniche del rispondente, ma di certo parecchio più impegnativo per l'intervistatore.

La ricostruzione della storia lavorativa avviene con la Sezione B. Essa comprende sette quesiti collegati da una struttura a salti, che portano alla compilazione di un calendario, in cui per ciascuno dei mesi del periodo di riferimento viene segnalato il verificarsi o meno di una data condizione (o stato). Il sistema di stati è limitato agli usuali tre - occupazione, ricerca di lavoro, inattività -, considerati disgiuntivi. In particolare, la struttura dei quesiti porta a compilare il calendario per gli eventi episodi di occupazione e/o ricerca di lavoro. Lo stato di non appartenenza alle forze di lavoro risulta dunque in maniera residuale, dall'assenza delle altre due condizioni nel mese. Una visualizzazione del sentiero dei quesiti, e degli esiti classificatori che producono, è nella Fig. 1.

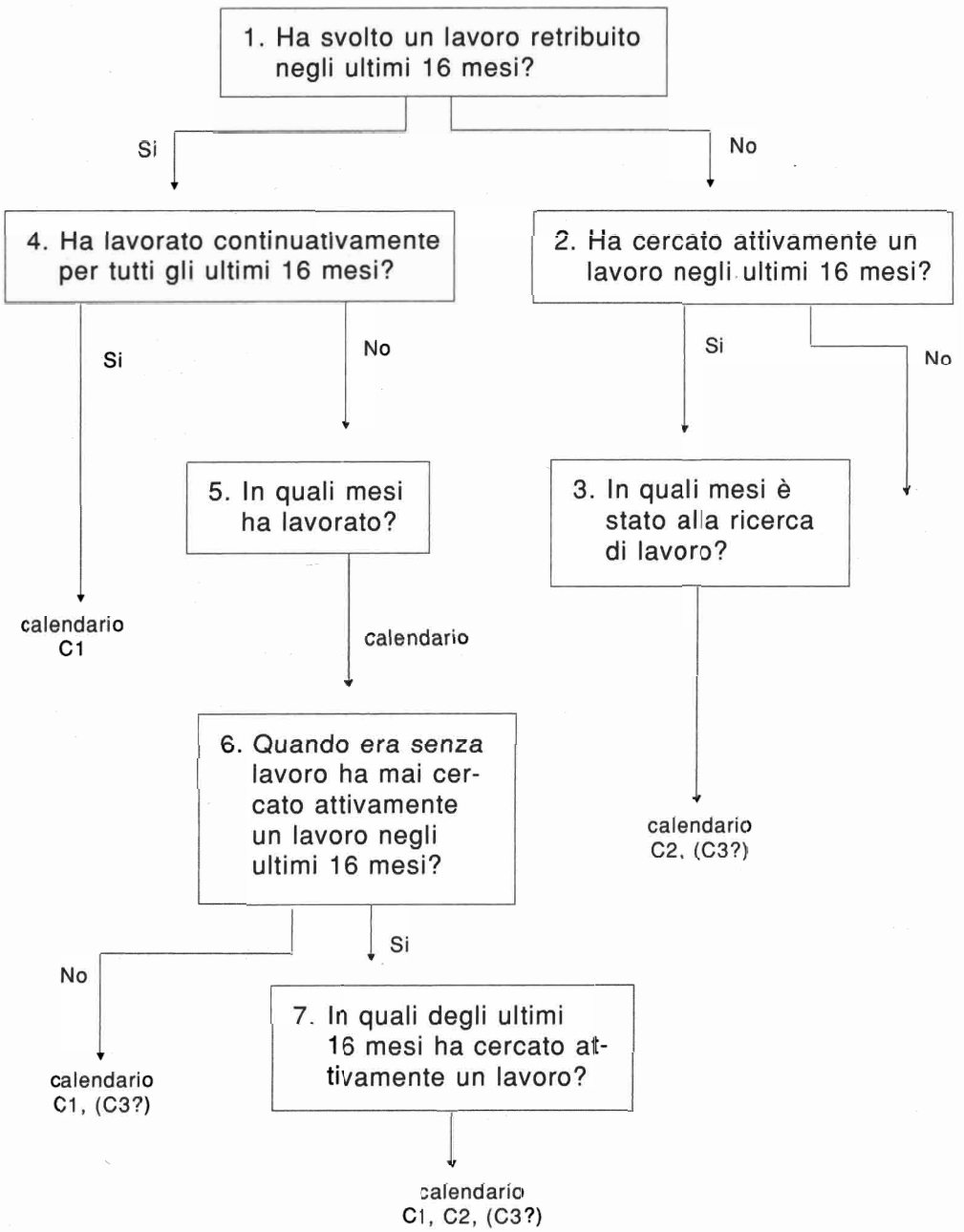
Le convenzioni definitorie dell'occupazione e della disoccupazione (vedi la sez. 3.2) e questa struttura dei quesiti hanno un'importante conseguenza sul dettaglio nella ricostruzione della storia lavorativa: è ammessa la possibilità di compresenza in uno stesso mese di episodi di occupazione e di ricerca di lavoro che si succedono l'un l'altro; è invece esclusa la possibilità che in uno stesso mese siano compresenti la condizione di inattività ed una delle restanti due (e tantomeno entrambe).

Al termine della compilazione del calendario, e sulla base di questo, è previsto che l'intervistatore compili una 'griglia di controllo', la quale riassume gli stati per cui il rispondente è transitato nel periodo di riferimento: se sia cioè stato occupato (C1), e/o alla ricerca di lavoro (C2), e/o inattivo (C3), e/o se in alcuni mesi abbia sperimentato la compresenza - in successione - delle condizioni di occupazione e disoccupazione.

La successiva Sezione C, ai cui Blocchi rinviano appunto le sigle di identificazione da C1 a C3, si propone di descrivere con qualche maggiore dettaglio tali situazioni. La scelta basilare è di procedere in modo asimmetrico per le diverse situazioni: in presenza di periodi di occupazione, si raccolgono informazioni sui singoli lavori - fino ad un massimo di tre -; per periodi di ricerca di lavoro e/o di inattività, le domande vertono sull'intero periodo trascorso nello stato, indipendentemente dal fatto che esso coincida con un episodio o consti di più episodi.

Agli approfondimenti sui lavori è dedicato il Blocco C1. La raccolta di maggiori informazioni è limitata ai due principali lavori alle dipendenze ed al principale lavoro in proprio. È importante notare che i singoli lavori alle dipendenze sono individuati col criterio del datore di lavoro, e ancora che si intende per principale il lavoro che "nel periodo di riferimento ha impegnato l'intervistato per il maggior numero di mesi (o di ore complessive, se si tratta di episodi lavorativi di durata inferiore al mese)". I criteri di identificazione dei lavori risultano perciò tre: la dicotomia lavoro dipendente/in proprio, per il lavoro dipendente il datore di lavoro, la durata in mesi. Guardare ai lavori nel periodo di riferimento (e non più agli episodi di occupazione che si snodano in successione temporale) ha un'importante conseguenza: i lavori possono sovrapporsi temporalmente. Per questa via possono dunque essere rilevati fenomeni ed episodi di 'doppio lavoro', ovviamente quando il 'secondo

Fig. 1: *Struttura a salti delle domande per la rilevazione della storia lavorativa* [(a)]



(a) C1 = episodi(o) di occupazione; C2 = episodi(o) di ricerca di lavoro; C3 = episodi(o) di inattività..

lavoro' non sia trascurabile nel quadro delle esperienze lavorative.

I quesiti sui singoli lavori riguardano innanzitutto le date di inizio e di (eventuale) conclusione e le usuali caratteristiche - posizione nella professione, settore di attività, attività a tempo pieno o parziale, carattere permanente o temporaneo -, di massima rilevate con le stesse classificazioni della RTFL. Essi si estendono poi all'orario medio settimanale e alla retribuzione media mensile (o, nel caso di lavoro in proprio, al reddito medio mensile).

Il Blocco C1 si conclude poi con alcune domande dettate da una finalità diversa: avere informazioni sulla quantità di lavoro prestata, da poter poi porre in relazione con i dati sul reddito da lavoro. Ciò da conto del cambiamento del periodo di riferimento, che diventa l'anno solare precedente. L'ultima domanda del Blocco è la sola dell'intero questionario che verte su dimensioni soggettive: si chiede al rispondente se, "a parità di guadagno orario avrebbe preferito lavorare più ore, guadagnando quindi di più, o lavorare meno ore guadagnando di meno", con l'intento di misurare squilibri indotti da vincoli istituzionali e/o da razionamento dal lato della domanda di lavoro.

Come già abbiamo anticipato, passando ad approfondimenti sulla situazione di ricerca di lavoro le domande perdono il riferimento a (eventuali) singoli episodi e riguardano l'intero periodo di ricerca. L'intento è di acquisire informazioni complementari a quelle della RTFL. L'enfasi del Blocco C2 è sull'intensità dell'attività di ricerca, che si cerca di cogliere in modo piuttosto semplice, incrociando tipo e frequenza delle azioni di ricerca.

Alla stessa logica di descrizione della situazione sull'intero periodo di inattività si ispira il successivo Blocco C3. Ed anche in tal caso l'attenzione è focalizzata sulla rilevazione di un fenomeno non colto dall'indagine trimestrale, quello dei 'lavoratori scoraggiati'. I tre quesiti per identificare gli aggregati riguardano (i) il desiderio di lavorare, (ii) la disponibilità a lavorare e (iii) il motivo prevalente per l'assenza di ricerca di lavoro, e sono formulati sulla falsariga di consolidate esperienze straniere (vedi OECD, 1987, pp. 142-170).

4.4. *I redditi individuali e familiari*

Le domande sui redditi sono contenute nelle ultime due Sezioni, D e E. Come abbiamo già avuto modo di precisare, per tali domande muta il riferimento temporale, che diventa l'anno solare 1988.

La Sezione D comprende le domande relative ai redditi da lavoro - distintamente dipendente e in proprio - e da trasferimenti, che vanno poste rispettivamente a tutti coloro che hanno lavorato nell'anno e a tutti i componenti la famiglia in età superiore ai 14 anni[12].

I quesiti della Sezione E riguardano i redditi non da lavoro riferiti all'intera famiglia, non ai singoli componenti, e sono posti preferibilmente al capofa-

12 Sono quindi trascurati gli eventuali redditi da trasferimenti in favore di componenti con meno di 14 anni.

miglia. Nell'organizzazione della sequenza delle domande, così come nella formulazione delle stesse, è immediato riconoscere non marginali somiglianze con parti del questionario dell'indagine della Banca d'Italia sui bilanci delle famiglie italiane. Le riduzioni, le semplificazioni e le aggregazioni operate rispetto al questionario della Banca d'Italia sono tuttavia drastiche. Esse rispondono all'ovvio scopo di contenere le dimensioni e la complessità del questionario.

Un'ultima osservazione si impone per spiegare il modo forse non del tutto convenzionale con cui si propone di tener conto degli effetti, in termini di reddito familiare, della proprietà o meno dell'abitazione. Piuttosto che rilevare il fitto figurativo per chi è proprietario della propria abitazione - per poi aggiungerlo agli altri redditi -, si preferisce rilevare l'affitto pagato da chi non è proprietario dell'abitazione - in vista di detrarlo dal totale degli altri redditi -. Le ragioni di questa scelta sono tre: (i) per lo studio dell'influenza del reddito sul comportamento dell'offerta di lavoro, sono importanti le differenze di reddito fra individui e famiglie più che i livelli assoluti del reddito: da questo punto di vista, dunque, le alternative di rilevazione sono equivalenti; (ii) è preferibile rilevare una grandezza nota al rispondente, quale il canone di affitto, piuttosto che una di problematica valutazione, collegata ad un ipotetico 'prezzo di mercato' dell'abitazione, quale il fitto figurativo; (iii) le vistose imperfezioni del mercato delle abitazioni nel nostro Paese non spostano i termini della questione. (Infatti, la rilevazione o l'imputazione di un fitto figurativo non risolverebbe il problema di valutazione del flusso di reddito generato dal bene 'casa abitata' secondo criteri omogenei per proprietari e non proprietari, a meno di non stimare differenziali di fitto figurativo per i non proprietari che pagano 'prezzi amministrati' indotti dalla legislazione sull'equo canone!). La rilevazione del flusso di spesa effettivo, appunto il canone di affitto, appare in definitiva l'alternativa più semplice e persuasiva[13].

5. *Progettazione operativa e modalità di svolgimento dell'indagine*

Il completamento della progettazione e lo svolgimento dell'indagine suppletiva sono stati curati da un apposito gruppo di lavoro costituito presso l'Istat, e per gli aspetti operativi dal Servizio Indagini sulle Famiglie e dall'Ufficio Regionale della Lombardia dell'Istituto. Ci limitiamo qui a pochi, essenziali cenni su tappe ed aspetti di particolare rilievo, cioè a dire: (i) l'indagine-pilota; (ii) gli strumenti di informazione/formazione e di rilevazione che hanno affiancato il questionario; (iii) il campione; (iv) l'iter delle operazioni sul campo.

Notevole importanza ha rivestito, per la progettazione conclusiva dell'indagine suppletiva, un'indagine-pilota su piccola scala, condotta a novembre 1988. Tale indagine ha coinvolto un campione di circa 300 famiglie, tratte

---

13 Va naturalmente scontata una qualche distorsione nella stima dei redditi non da lavoro in favore dei non proprietari, perché l'abitazione di proprietà ha un valore mediamente superiore all'abitazione in affitto.

dal 'quarto uscente' della RTFL di ottobre in tre comuni della Lombardia (Milano, Vigevano e Casteldario), scelti in maniera ragionata sì da riflettere situazioni del mercato del lavoro sufficientemente diversificate. Oltre allo scopo primario di saggiare la bozza del questionario, essa è stata orientata anche a raccogliere indicazioni sull'intero spettro degli aspetti operativi del progetto. Essa ha fornito riscontri complessivamente positivi sulla fattibilità dell'indagine suppletiva, e utili lumi per l'affinamento del questionario e per l'approntamento di altri strumenti di supporto alla rilevazione. In tema di questionario, le valutazioni positive hanno riguardato soprattutto lo *skip pattern* e il diffuso ricorso a domande di controllo per gli intervistatori (che ne rendono tutto sommato snella la somministrazione), mentre concordi evidenze sulla ridondanza di alcuni quesiti hanno condotto qua e là a utili alleggerimenti[14]. L'indagine-pilota ha inoltre evidenziato come il raccordo con la RTFL immediatamente precedente, per la formazione del campione e per la precompilazione del Blocco A1 del questionario, comportasse operazioni laboriose e delicate, che hanno richiesto una più puntuale definizione di procedure e di strumenti aggiuntivi di rilevazione. Tra l'altro, la laboriosità di queste operazioni ha indotto a dilatare l'intervallo tra periodo di effettuazione della RTFL e periodo di effettuazione dell'indagine suppletiva dalle due-tre settimane inizialmente ipotizzate a sei settimane, con conseguente estensione a 17 mesi del periodo di ricostruzione retrospettiva della storia lavorativa.

Anche a seguito delle risultanze dell'indagine-pilota, oltre al testo definitivo del questionario è stato messo a punto un insieme di altri strumenti di informazione/formazione e di rilevazione, costituiti essenzialmente:

(a) per l'informazione/formazione da: (a1) una nota di istruzioni ai Comuni sulle modalità di esecuzione dell'indagine; (a2) un manuale per gli intervistatori; (a3) una lettera di presentazione dell'indagine alle famiglie, da parte del Presidente dell'Istat, da consegnare alla conclusione della RTFL di aprile 1989;
(b) per la rilevazione da: (b1) un modello per la registrazione, a livello di comune e distintamente per ciascun rilevatore, dell'elenco della famiglie da intervistare; (b2) una scheda-cartoncino per gli intervistatori, con la codifica delle modalità di risposta a taluni quesiti del questionario; (b3) una scheda-calendario da consegnare ai rispondenti, con il dettaglio dei 17 mesi precedenti l'indagine, per agevolare il ricordo.

È appena ovvio notare che la predisposizione di questo insieme di strumenti non è stata mera operazione tecnica, ma ha richiesto ulteriori chiarimenti su delicati aspetti attinenti alle modalità di svolgimento della rilevazione suppletiva. Al riguardo può forse bastare un esempio. Le istruzioni ai rilevatori per la conduzione dell'intervista hanno implicato fosse chiarito che, e a quali condizioni (nella fattispecie, dopo due ritorni presso la famiglia per cercare di completare il questionario con i diretti interessati), erano accet-

---

14 Rispetto alla versione somministrata nell'indagine-pilota, nel questionario definitivo non figurano tre insiemi di quesiti: sulla partecipazione a precedenti rilevazioni sulle forze di lavoro; su un eventuale lavoro in proprio secondario; sulla compresenza di episodi di occupazione e ricerca di lavoro nello stesso mese.

tabili risposte tramite *proxy*.

La progettazione operativa dell'indagine è stata infine completata con la definizione del campione. In proposito, è stato sì mantenuto il riferimento di massima al 'quarto uscente' della RTFL di aprile 1989 in Lombardia. Vincoli di bilancio hanno tuttavia imposto una riduzione della dimensione campionaria. Tale riduzione è stata realizzata con un ragionevole compromesso tra due esigenze contrastanti: intaccare il meno possibile il disegno campionario delle unità di primo stadio - i comuni -, per assicurarne la rappresentatività e la rispondenza con la RTFL; ridurre la dispersone territoriale delle interviste, per contenere i costi. Specificamente, rispetto al 'quarto uscente' di aprile 1989 si sono considerati soltanto i comuni del campione originario, escludendo quelli inclusi a seguito del 'sovracampionamento'; (ii) sono stati poi esclusi i comuni per i quali il numero di famiglie nel 'quarto uscente' del campione era pari o inferiore a 3. Ne è risultato un campione di 190 comuni, e di 4.497 famiglie (vedi la Tab. 2).

Per quanto riguarda le operazioni sul campo, vanno ricordati perlomeno la cura dedicata alla selezione e all'addestramento degli intervistatori, e l'impegno nell'insieme delle attività di supervisione e controllo. Verosimilmente si deve a questi fattori, oltre che all'accurata progettazione dell'indagine, l'elevato tasso di risposta (vedi la Tab. 2, ultima colonna), tanto più ragguardevole se si riflette sulla delicatezza di alcuni temi del questionario e sul fatto che non erano previste sostituzioni.

6. *Cenni sulle possibilità di analisi*

Gli obiettivi conoscitivi che l'indagine suppletiva consente soddisfare sono iscritti nei suoi contenuti e nel suo disegno. Può tuttavia non essere inutile cercare, in via esemplificativa, di segnalarne alcuni.

La ricostruzione della storia lavorativa consente innanzitutto di cogliere partecipazioni al lavoro stagionali o marginali, che possono sfuggire alle 'istantanee' delle rilevazioni correnti scattate su quattro settimane dell'anno.

Tab. 2: *Campione della RTFL di aprile 1989 per la Lombardia e dell'indagine suppletiva*

Unità di campionamento e unità di rilevazione	RTFL aprile 1989		Indagine suppletiva	
	Totale	'Quarto uscente'	Campione teorico	Campione effettivo
Comuni	337	337	190	185
Famiglie	24.768	6.192	4.497	4.127
Individui	66.787	16.796	11.752	10.806

Analoga argomentazione vale, ovviamente, anche per gli episodi di ricerca di occupazione. In definitiva, risultano possibili stime del numero di persone che nell'arco di un anno (o di 17 mesi) sperimentano l'occupazione, così come di quelle che sperimentano la disoccupazione.

In secondo luogo, le informazioni raccolte con l'indagine suppletiva permettono di misurare ed analizzare per l'intero campione di individui e in modo sufficientemente completo, anche se per un periodo piuttosto breve, la dinamica rispetto al lavoro. Ciò apre la prospettiva per analisi dei flussi e della mobilità, così come per studi sui fenomeni di persistenza in uno stato e di influenza che essa dispiega sulle probabilità di permanervi cronicamente. Il rilievo di queste opportunità di analisi per temi quali la disoccupazione giovanile, la stabilità/mobilità occupazionale, l'entrata/uscita delle donne dal mercato del lavoro, neppure necessita di essere evocato tanto è palese.

Più in generale l'integrazione di dati alla scala micro sulla condizione dell'individuo rispetto al lavoro - e la sua dinamica - con informazioni sui principali episodi di occupazione, con il volume del lavoro prestato, con i redditi da lavoro, con gli altri redditi individuali e familiari, con indicatori del *background* formativo e familiare, configura possibilità di analisi dei comportamenti a spettro assai largo: a livello di individui e di famiglie, sia in chiave statica che dinamica, a fini descrittivo-interpretativi ma anche di supporto conoscitivo in vista della definizione di politiche. Più che azzardare possibili percorsi di ricerca, che inevitabilmente rifletterebbero le nostre propensioni e curiosità (e le nostre ignoranze), vorremmo semplicemente richiamare un esempio: quello dell'indagine su qualità della vita e strategie di comportamento familiare del 'progetto Torino' (Martinotti, 1982). È un esempio che ci pare emblematico. Il campione di 1.000 famiglie torinesi, per propri meriti e forse anche perché ha a lungo sperimentato la condizione di 'figlio unico' della ricerca orientata alla produzione di basi di dati integrati sul mercato del lavoro, ha ormai un suo posto non marginale nel sentiero di crescita di consapevolezze empiriche, schemi interpretativi e metodi di analisi sull'offerta di lavoro in Italia.

Queste indicazioni, che speriamo prospettate senza un pedaggio eccessivo ai toni enfatici, sono naturalmente condizionate a un grande 'se' sulla qualità dei dati: sulla loro pertinenza rispetto a diversificati obiettivi conoscitivi, e ancor più sulla loro attendibilità. Anche un diverso ordine di considerazioni è, tuttavia, pertinente. A nostro avviso, un obiettivo tutt'altro che trascurabile dell'indagine suppletiva è proprio quello di consentire un suo vaglio sul terreno della qualità dei dati. L'attendibilità di informazioni sulla storia lavorativa rilevate retrospettivamente, le possibilità ed i limiti di confronto/integrazione di tali informazioni con i dati raccolti correntemente con la RTFL, l'attendibilità dei dati di reddito, le procedure di controllo e di correzione che possono essere approntate per i diversi sottoinsiemi di dati, sono altrettanti temi di riflessione e di ricerca di rilevante interesse pratico. Approfondimenti in queste direzioni possono tornare utili sia all'Istat, per migliorare il disegno di questa indagine e per acquisire esperienze proficuamente trasferibili ad indagini similari, sia agli studiosi interessati ad un uso migliore - quindi più intensivo e ad un tempo più consapevolmente critico -

dei dati.

Naturalmente, un adeguato sfruttamento delle potenzialità conoscitive dell'indagine suppletiva non può certo essere assicurato da un pur ricco piano di spogli, ma richiede l'accesso al *file* di dati elementari da parte degli studiosi interessati. L'argomento ha risvolti delicati, per la doverosa tutela del segreto statistico recentemente disciplinata dal DL 322/1989. L'orientamento adottato dall'Istat in materia è, peraltro, estremamente promettente. L'Istituto ha infatto recentemente definito due linee di lavoro:

(a) la predisposizione di un rapporto di presentazione dei risultati salienti dell'indagine, nella forma usuale di un organico insieme di tabelle;
(b) l'approntamento e la messa a disposizione di un *public use file* (com'è ovvio, previa eliminazione o oscuramento dei caratteri di identificazione, per tutelare il segreto statistico).

Verosimilmente, i tempi non saranno brevi. Lo spoglio dei dati è infatti condizionato ad una attività di *editing* e di imputazione che si prospetta particolarmente laboriosa, per il carattere innovativo-sperimentale dell'indagine e per la ricchezza e complessità delle informazioni rilevate. D'altra parte, la predisposizione di un *public use file* non solo (e non tanto) richiede ulteriori operazioni tecniche, ma soprattutto si presenta come compito affatto nuovo per la statistica ufficiale italiana: essa comporta quindi una riflessione, su contenuti e modi di distribuzione della base di dati, che di necessità coinvolgerà aspetti generali della politica di diffusione dell'informazione statistica. A fronte dell'importanza di una tale innovazione, non ci pare tuttavia irragionevole mettere temporaneamente la sordina a comprensibili impazienze.

Riferimenti bibliografici

Abowd J.M. e A. Zellner (1985), 'Estimating gross labor-force flows', *Journal of Business & Economic Statistics*, 3, pp. 254- 283.

Aitkin M. e R. Healey (1985), 'Statistical modelling of unemployment rates from the EEC Labour Force Survey', *Journal of the Royal Statistical Society*, A, 148, pp. 45-56.

Akaike H. (1980), 'Seasonal adjustment by a bayesian modelling', *Journal of Time Series Analysis*, 1, pp. 1-13.

Akaike H. e M. Ishiguro (1983), 'Comparative study of the X-11 and BAYSEA procedures of seasonal adjustment', in A. Zellner (ed.), *Applied time series analysis of economic data*, Washington, D.C., U.S. Department of Commerce, Bureau of the Census, pp. 17-30.

Altham P.M.E. (1976), 'Discrete variable analysis for individual grouped into families', *Biometrika*, 63, pp. 263-269.

Amemiya T. (1985), *Advanced econometrics*, Oxford, Basil Blackwell.

Andrews M. e S. Nickell (1986), 'A disaggregated disequilibrium model of the labour market', *Oxford Economic Papers*, 38, pp. 386-402.

Arellano M. e C. Meghir (1990), 'Labour supply and hours constraints', in J.P. Florens, M. Ivaldi, J.J. Laffont e F. Laisney (eds.), *Microeconometrics: surveys and applications*, Oxford, Basil Blackwell, pp. 213-230.

Ashenfelter O. e D. Card (1982), 'Time series representations of economic variables and alternative models of the labour market', *Review of Economic Studies*, 49 (Special Issue), pp. 761-782.

Ashenfelter O. e G. Solon (1982), 'Longitudinal labor market data. Sources, uses and limitations', in *Conference on the assessment of labor force measurements for policy formulation*, Washington, D.C., U.S. Government Printing Office, pp. 109-126.

Baddeley A. (1979), 'The limitations of human memory: implications for the design of retrospective surveys', in L. Moss e H. Goldstein (eds.), *The recall method in social surveys*, London, University of London Institute of Education, pp. 13-27.

Baici E. (1987), 'Un modello *insider-outsider* di determinazione del salario in Italia', *Rivista Internazionale di Scienze Sociali*, 95 (2), pp. 107-122.

Bailar B.A. (1975), 'The effects of rotation group bias on estimates from panel surveys', *Journal of the American Statistical Association*, 70, pp. 23-30.

Bedrick E.J. (1983), 'Adjusted chi-squared tests for cross -classified tables of survey data', *Biometrika*, 70, pp. 591-596.

Bernard H.R., P. Killworth, D. Kronenfeld e L. Sailer (1984), 'The problem of informant accuracy: the validity of retrospective data', *Annual Review of Anthropology*, 13, pp. 495-517.

Bernardi L., P. Manfroni, I. Sanetti e U. Trivellato (1987), 'Proposta conclusiva in tema di indagini suppletive alla rilevazione sulle forze di lavoro', Nota interna FOLA n. 37, Dipartimento di Scienze Statistiche, Università di Padova, (cicl.).

Bernardi L. e S. Zaccarin (1987), 'Classificazione di stati per l'analisi della mobilità

occupazionale', Nota interna FOLA n. 32, Dipartimento di Scienze Statistiche, Università di Padova, (cicl.).

Bethelhem J.K. e W.J. Keller (1987), 'Linear weighting of sample survey data', *Journal of Official Statistics*, 3, pp 141-153.

Bianchi S.M (1987), 'Living at home: young adult's living arrangements in the 1980s', paper presented at the Annual Meeting of the American Sociological Association, Chicago, August 1987, (mimeo).

Biemer P.P. e Stokes S.L. (1985), 'Optimal design of interviewer variance experiments in complex surveys', *Journal of the American Statistical Association*, 80, pp. 158-166.

Bishop Y.M.M., S.E. Fienberg e P.W. Holland (1975), *Discrete multivariate analysis: theory and practice*, Cambridge, Mass., MIT Press.

Blanchard O.J. e L.H. Summers (1986), 'Hysteresis and the European unemployment problem', in S. Fisher (ed.), *NBER Macroeconomics Annual 1986*, Cambridge, Mass., MIT Press, pp. 15-78.

Blangiardo G., S. Lauro e D. Semisa (1986), *Stima della popolazione residente nelle province della Lombardia per sesso ed età negli anni 1981-1991*, Collana di documentazione statistica, 20/2, Milano, Regione Lombardia.

Blight B.J.N. e A.J. Scott (1973), 'A stochastic model for repeated surveys', *Journal of the Royal Statistical Society*, B, 35, pp. 61- 66.

Blundell R., J. Ham e C. Meghir (1987), 'Unemployment and female labour supply', *The Economic Journal*, 97, pp. 44-64.

Blundell R., J. Ham e C. Meghir (1988), *Unemployment, discouraged workers and female labour supply*, UCL discussion paper, London, University College.

Blundell R. e I. Walker (eds.) (1986), *Unemployment, search and labour supply*, Cambridge, Cambridge University Press.

Bodo G. e I. Visco (1987), *La disoccupazione in Italia: un'analisi con il modello econometrico della Banca d'Italia*, Temi di discussione n. 91, Roma, Banca d'Italia.

Bolasco S. (1983), 'Analisi sulle componenti strutturali dei mercati del lavoro su base territoriale. Una mappa delle regioni italiane al 1980', *Economia & Lavoro*, 17 (2), pp. 41-65.

Bolasco S. e R. Coppi (1987), 'Primi risultati sull'analisi strutturale e dinamica dell'occupazione in cinque regioni italiane dal 1978 al 1986', Nota interna FOLA n. 22, Dipartimento di Scienze Statistiche, Università di Padova, (cicl.).

Bolasco S. e R. Coppi (1988), 'Analisi strutturale dell'occupazione mediante matrici di dati a più indici', Nota interna FOLA n. 51, Dipartimento di Scienze Statistiche, Università di Padova, (cicl.).

Bordignon S. e G. Masarotto (1987), *Confronto tra alcuni metodi di destagionalizzazione sulle serie storiche del mercato del lavoro*, Working paper n. 1, Progetto finalizzato CNR "Struttura ed evoluzione dell'economia italiana", tema di ricerca I/01/C, Padova, Cleup.

Bordignon S. e G. Masarotto (1988), *Metodi di destagionalizzazione di un aggregato con riferimento al tasso di disoccupazione e alle sue componenti*, Working paper n. 6, Progetto finalizzato CNR "Struttura ad evoluzione dell'economia italiana", tema di ricerca I/01/C, Padova, Cleup.

Bordignon S.e U. Trivellato (1989), 'The optimal use of provisional data in foreca-

sting with dynamic models', *Journal of Business & Economic Statistics*, 7, pp. 275 - 286

Bowers N. e F.W. Horvath (1984), 'Keeping time: an analysis of errors in the measurement of unemployment duration', *Journal of Business & Economic Statistics*, 2, pp. 140-149.

Box G.E.P., S.C. Hillmer e G.C. Tiao (1978), 'Analysis and modelling of seasonal time series', in A. Zellner (ed.), *Seasonal analysis of economic time series*, Washington, D.C., U.S. Department of Commerce, Bureau of the Census, pp. 309-334.

Box G.E.P. e G.M. Jenkins (1976), *Time series analysis: forecasting and control*, S. Francisco, Holden-Day.

Brier S.E. (1980), 'Analysis of contingency tables under cluster sampling', *Biometrika*, 67, pp. 591-596.

Brittain J.A. (1959), 'A bias in the seasonally adjusted unemployment series and a suggested alternative', *Review of Economics and Statistics*, 41, pp. 405-411.

Bruno S. (1987), 'L'informazione statistica come strumento indispensabile per decidere: il mercato del lavoro', *Economia & Lavoro*, 21 (2), pp. 111-118.

Brusco S., P. Gennari, L. Marchesini e G. Solinas (1986), 'Giovani in cerca di occupazione e disoccupati in Emilia Romagna', Bologna, Regione Emilia Romagna, (cicl.).

Burman J.P. (1980), 'Seasonal adjustment by signal extraction', *Journal of the Royal Statistical Society*, A, 143, pp. 321-337.

den Butter F.A.G., R.L. Coenen e F.J.J.S. van de Gevel (1985), 'The use of ARIMA models in seasonal adjustment', *Empirical Economics*, 10, pp. 209-230.

Butz W.P. e T.J. Plewes (1989), 'A Current Population Survey for the 21th century', in Bureau of the Census, *Fifth Annual Research Conference. March 19-22, 1989. Proceedings*, Washington, D.C., U.S. Department of Commerce, Bureau of the Census, pp. 3-13.

Cannari L. e A. Lemmi (1988), *L'integrazione dell'indagine sulle forze di lavoro con informazioni sui salari e sui redditi: procedure di stima e di imputazione*, Rapporto di ricerca FOLA n. 8, Padova, Cleup.

Carroll J.D. e P. Arabie (1980), 'Multidimensional scaling', *Annual Review of Psycology*, 31, pp. 607-649.

Castellini M. (1982), 'Il campionamento ruotato: vantaggi derivanti dall'impiego di stimatori composti e possibili distorsioni', Università di Padova, Facoltà di Scienze Statistiche Demografiche ed Attuariali, (tesi di laurea non pubblicata).

Chambless L.E. e K.E. Boyle (1985), 'ML methods for complex sample data: logistic regression and discrete proportional hazard models', *Communications in Statistics. Theory and Methods*, 14, pp. 1377-1392.

Chandon J.L. e S. Pinson (1981), *Analyse typologique: théories et application*, Paris, Masson.

Chapman D.W. (1966), 'An approximate test of independence based on replications of a complex sample survey design', Cornell University, (unpublished M.Sc. thesis).

Chen T. e S.E. Fienberg (1974), 'Two-dimensional contingency tables with both completely and partially cross-classified data', *Biometrics*, 30, pp. 629-642.

Choudhry G.H. e M.A. Hidiriglou (1987), 'Small area estimation: some investiga-

tions at Statistics Canada', *Bulletin of the International Statistical Institute*, vol. LII, book 4, pp. 451-468.

Cielo R. (1989), 'L'accuratezza dei dati di durata della ricerca di lavoro nell'indagine italiana sulle forze di lavoro', Università di Padova, Facoltà di Scienze Statistiche Demografiche ed Attuariali, (tesi di laurea non pubblicata).

Clark K.B. e L.H. Summers (1979), 'Labor market dynamics and unemployment: a reconsideration', *Brooking Papers on Economic Activity*, 10 (1), pp. 13-60.

Cocchi D. (1990), *Stimatori per campioni ruotati: esperienze e proposte per l'indagine*, Rapporto di ricerca FOLA n. 22, Padova, Cleup.

Cocchi D. e M. Castellini (1988), *Stimatori composti in campioni ruotati per l'indagine italiana sulle forze di lavoro*, Rapporto di ricerca FOLA n. 7, Padova, Cleup.

Cochran W.G. (1977), *Sampling techniques*, London, Wiley.

Cogan J. (1981), 'Fixed costs and labor supply', *Econometrica*, 49, pp. 945-963.

Cohen J.E. (1976), 'The distribution of the chi-squared statistics under cluster sampling from contingency tables', *Journal of the American Statistical Association*, 71, pp. 665-670.

Coletti F. (1923), *Studi sulla popolazione italiana in pace e in guerra*, Bari, Laterza.

Colombino U. (1979), 'Mercato del lavoro', in M. Carmagnani e A. Vercelli (a cura di), *Il mondo contemporaneo. Economia e storia*, Firenze, La Nuova Italia, vol. 1, pp. 428-450.

Colombino U. (1984), 'Costi di squilibrio sul mercato del lavoro e programmi di riequilibrio. Un esercizio di microsimulazione', in M. Schenkel (a cura di), *L'offerta di lavoro in Italia*, Venezia, Marsilio Editori, pp. 367-396.

Colombino U. (1986), 'Stima di modelli strutturali di offerta e domanda di lavoro', Nota interna FOLA n. 9, Dipartimento di Scienze Statistiche, Università di Padova, (ciclo.).

Copeland K.R., F.K. Peitzmeier e C.E. Hoy (1986), 'An alternative method of controlling Current Population Survey estimates to population counts', Washington, D.C., U.S. Bureau of Labor Statistics, (mimeo).

Coppi R. (1986), 'Analysis of three-way data matrices based on pairwise relation measures', in *COMPSTAT 1986 - Proceedings in computational statistics*, Heidelberg, Physica-Verlag, pp. 129- 139.

Coppi R. e S. Bolasco (eds.) (1989), *Multiway data analysis*, Amsterdam, North Holland.

Cox D.R. (1972), 'Regression models and life tables', *Journal of the Royal Statistical Society*, B, 34, pp. 187-220 (with discussion).

Cox D.R e D. Oakes (1984), *Analysis of survival data*, London, Chapman and Hall.

Cox L.H. e R.F. Boruch (1988), 'Record linkage, privacy and statistical policy', *Journal of Official Statistics*, 4, pp. 3-16.

Dagum E.B. (1975), 'Seasonal factor forecasts from ARIMA models', *Bulletin of the International Statistical Institute*, vol. XLVI, book 3, pp. 203-216.

Dagum E.B. (1976), 'Recent developments in seasonal adjustment methods and applications', in *Selected Papers from North American Conference on Labor Statistics*, Washington, D.C., U.S. Department of Labor, pp. 146-161.

Dagum E.B. (1978), *A comparison and assessment of seasonal adjustment methods for employment and unemployment statistics*, National Commission on Employment and Unemployment Statistics, Background paper n. 5, Washin-

gton, D.C., U.S. Government Printing Office.
Dagum E.B. (1979), *The seasonal adjustment of economic time series aggregates: a case study on the unemployment rate*, National Commission on Employment and Unemployment Statistics, Background paper n. 31, Washington, D.C., U.S. Government Printing Office.
Dagum E.B. (1983), *The X-11-ARIMA seasonal adjustment method*, Ottawa, Statistics Canada.
David M. (ed.) (1985), *Survey of Income and Program Participation*, Special Issue of *Journal of Economic and Social Measurement*, 13 (3 e 4).
Del Boca D. (1984), 'Differenziali salariali e alternative di orario', *Economia & Lavoro*, 18 (4), pp. 3-16.
Del Boca D. (1987), 'Facts and theories of the Italian labour market (1970-1985): a time series analysis of wages, unemployment and prices', *Labour*, 1 (3), pp. 163-179.
Del Boca D. e C.J. Flinn (1984), 'Self-reported reservation wages and the labor market participation decision', *Ricerche Economiche*, 38, pp. 263-283.
Del Boca D. e M. Turvani (1979a), *Famiglia e mercato del lavoro*, Bologna, Il Mulino.
Del Boca D. e M. Turvani (1979b), 'Il processo di nuclearizzazione della famiglia e il mercato del lavoro: un'analisi sui dati censuari 51, 61, 71', *Note Economiche*, 5, pp. 115-124.
Dempster A.P., N.M. Laird e D.B. Rubin (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society*, B, 39, pp. 1-38.
De Nicola I. (1989), 'Alcune considerazioni sul calcolo dei tassi di disoccupazione', in Ministero del Lavoro e della Previdenza Sociale, *Rapporto 1989. Lavoro e politiche della occupazione in Italia*, Roma, Fondazione Brodolini e Centro Europa Ricerche, pp. 229-232.
De Sandre P. (1988), *La famiglia 'lunga' del giovane adulto*, Collana "Studi Interdisciplinari sulla famiglia", Milano, Vita e Pensiero.
Dippo C.S. (1989), 'The use of cognitive laboratory techiniques for investigating memory retrival errors in retrospective surveys', *Bulletin of the International Institute*, vol. LIII, book 2, pp. 363-382.
von Dosselaar P.G.W.M., G.J.H.H. van den Hurk e A.Z. Isral (1989), 'Analysis of memory effects at the Netherlands Central Bureau of Statistics', *Bulletin of the International Statistical Institute*, vol. LIII, book 2, pp. 383-402.
Duncan G.J. e D.H. Hill (1985), 'An investigation of the extent and consequences of measurement error in labor-economic survey data', *Journal of Labor Economics*, 3, pp. 508-532.
Duncan G.J. e D.H. Hill (1989), 'Assessing the quality of household panel data: the case of the panel study of income dynamics', *Journal of Business & Economic Statistics*, 7, pp. 441-452.
Duncan G.J. e M.S. Hill (1985), 'Conception of longitudinal households: fertile or futile?', *Journal of Economic and Social Measurement*, 13, pp. 361-376.
Duncan G.J., F.T. Juster e J.N. Morgan (1987), 'The role of panel studies in research on economic behavior', *Transportation Research*, 21A, pp. 249-263.
Duncan G.J. e G. Kalton (1987), 'Issues of design and analysis of surveys across time', *International Statistical Review*, 55, pp. 97-117.

Eckler A.R. (1955), 'Rotation sampling', *The Annals of Mathematical Statistics*, 26, pp. 664-685.

Elbers C. e G. Ridder (1982), 'True and spurious duration dependence. The identifiability of the proportional hazard model', *Review of Economic Studies*, 49, pp. 403-410.

Engle R.F. (1978), 'Estimating structural models of seasonality', in A. Zellner (ed.), *Seasonal analysis of economic time series*, Washington, D.C., U.S. Department of Commerce, Bureau of the Census, pp. 281-308.

Escofier B. e J. Pages (1988), *Analyses factorielles simples et multiples*, Paris, Dunod.

Eurostat (1985), *Indagine per campione sulle forze di lavoro. Metodi e definizioni*, Serie gialla, Lussemburgo.

Eurostat (1990), 'Stato di avanzamento della revisione della indagine comunitaria sulle forze di lavoro', Doc. E1/651/90-IT, Lussemburgo, (cicl.).

Fabbris L. (1989), *L'indagine campionaria. Metodi, disegni e tecniche di campionamento*, Roma, La Nuova Italia Scientifica.

Fabbris L. e L. Bernardi (1986), 'Disegno e caratteristiche dell'indagine sulle forze di lavoro', Nota interna FOLA n. 2, Dipartimento di Scienze Statistiche, Università di Padova, (cicl.).

Fabbris L., G. Leti, E. Zaghini e A. Zuliani (1986), 'Aspetti metodologici delle indagini campionarie sui bilanci delle famiglie italiane', in *Le indagini campionarie sui bilanci delle famiglie italiane*, numero speciale dei *Contributi all'analisi economica*, Roma, Banca d'Italia, pp. 11-62.

Fabbris L., A. Russo e I. Sanetti (1988), *Storia e prime proposte in tema di sovra-campionamento a livello regionale, provinciale e sub-provinciale per indagini sulle forze di lavoro*, Rapporto di ricerca FOLA n. 4, Padova, Cleup.

Falorsi P.D. (1990), 'Analisi dell'efficienza di uno stimatore composto nella rilevazione trimestrale Istat sulle forze di lavoro', in Società Italiana di Statistica, *Atti della XXXV Riunione Scientifica*, Padova, Cedam, vol. 2, pp. 249-256.

Falorsi P.D. e A. Russo (1987), 'Un metodo di stima sintetica per piccoli domini territoriali nelle indagini Istat sulle famiglie', in Società Italiana di Statistica, *Convegno 1987 - Informazione ed analisi statistica per aree regionali e subregionali*, Perugia, Galeno Editrice, pp. 175-183.

Fay R.E. (1979), 'On adjusting the Pearson chi-square statistic for clustered sampling', in American Statistical Association, *Proceedings of the Social Statistics Section*, pp. 402-406.

Fay R.E. (1982), 'Contingency tables analysis for complex survey designs: CPLX', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 44-53.

Fay R.E. (1984), 'Application of linear and log-linear models to data from complex samples', *Survey Methodology*, 10, pp. 82-96.

Fay R.E. (1985), 'A jackknifed chi-squared test for complex samples', *Journal of the American Statistical Association*, 80, pp. 148-157.

Fay R.E. (1986), 'Causal models for patterns of nonresponse', *Journal of the American Statistical Association*, 81, pp. 354- 365.

Fellegi I.P. (1963), 'Sampling with varying probabilities without replacement: rotating and non-rotating samples', *Journal of the American Statistical Association*,

58, pp 183-201.
Fellegi I.P. (1964), 'Response variance and its estimation', *Journal of the American Statistical Association*, 59, pp. 1016-1041.
Fellegi I.P. (1980), 'Approximate tests of independence and goodness-of-fit based on stratified multistage samples', *Journal of the American Statistical Association*, 75, pp. 261- 268.
Fellegi I.P. e A. Sunter (1969), 'A theory for record linkage', *Journal of the American Statistical Association*, 64, pp.1183-1210.
Fienberg S.E. (1980), *The analysis of cross-classified data*, Cambridge, Mass., MIT Press.
Fienberg S.E., B. Singer e J.M. Tanur (1985), 'Large-scale social experimentation in the United States', in A.C. Atkinson e S.E. Fienberg (eds.), *A celebration of Statistics*, New York, Springer-Verlag, pp. 287-326.
Fienberg S.E. e E.A. Stasny (1985), 'Estimating monthly gross flows in labour force participation', *Survey Methodology*, 9, pp. 77-102.
Fitoussi J.P e J. Le Cacheux (1988), 'On theories of unemployment persistence: a quick look at recent developments', *Labour*, 2 (2), pp. 3-20.
Flaim P.O. e C.R. Hogue (1985), 'Measuring labor force flows: a conference examines the problems', *Monthly Labor Review*, 108 (July), pp. 7-17.
Flinn C.J. (1986), 'Econometric analysis of CPS-type unemployment data', *Journal of Human Resources*, 21, pp. 456-484.
Flinn C.J. e J.J. Heckman (1982), 'Models for the analysis of labor force dynamics', in R. Basmann e G. Rhodes (eds.), *Advances in Econometrics.1*, Greenwich, CT, JAI Press, pp. 35-95.
Flinn C.J. e J.J. Heckman (1983), 'Are unemployment and out of the labor force behaviorally distinct categories?', *Journal of Labor Economics*, 1, pp. 28-42.
Frey L. (1988), *La disoccupazione in Italia: i punti di vista*, Quaderni di Economia del Lavoro/36, Milano, Franco Angeli.
Fuller W.A. e T. Chua (1985), 'Gross change estimation in the presence of response error', *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, Washington, D.C., U.S. Department of Commerce and U.S. Department of Labor, pp. 65- 79.
Fürst H. (1988), 'Problems in the statistical measurement of unemployment in the Community', *Economia & Lavoro*, 22 (2), pp. 59-87.
Gagliardi F. e D.T. Coe (1985), 'Nominal wage determination in ten OECD economies', Working paper, Economics and Statistics Department, Paris, OECD, (mimeo).
Geweke J. (1978), 'The temporal and sectoral aggregation of seasonally adjusted time series', in A. Zellner (ed.), *Seasonal analysis of economic time series*, Washington, D.C., U.S. Department of Commerce, Bureau of the Census, pp. 411-427.
Ghangurde P.D. (1982), 'Rotation group bias in the LFS estimates', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 421-426.
Giommi A. (1988), *Un'analisi della qualità dei dati basata sul confronto dei 'records' individuali in più occasioni di indagine*, Rapporto di ricerca FOLA n. 9, Padova, Cleup.

Giommi A., A. Giusti e N. Torelli (1987), 'Valutazioni preliminari sulla qualità dei dati basate sul confronto di *records* individuali in successive occasioni', Nota interna FOLA n. 16, Dipartimento di Scienze Statistiche, Università di Padova, (cicl.).

Giusti A., G. Marliani e N. Torelli (1987), *L'abbinamento dei dati individuali di successive rilevazioni nell'indagine trimestrale delle forze di lavoro*, Rapporto di ricerca FOLA n. 2, Padova, Cleup.

Giusti A., G. Marliani e N. Torelli (1988), *Una procedura probalistica per l'abbinamento dei dati individuali delle forze di lavoro*, Rapporto di ricerca FOLA n. 10, Padova, Cleup.

Golini A. (1987), 'Famille et ménage dans l'Italie récente', *Population*, 42, pp. 699-713.

Gonzalez M.E. e C. Hoza (1978), 'Small area estimation with application to unemployment and housing estimates', *Journal of the American Statistical Association*, 73, pp. 7-15.

Gonzalez M.E. e J. Waksberg (1973), 'Estimation of the error of synthetic estimates', paper presented at the First meeting of the International Association of Survey Statisticians, Vienna, August 18-25, 1973, (mimeo).

Griguolo S. e P.C. Palermo (1984), *Nuovi problemi e metodi di analisi territoriale*, Milano, Franco Angeli.

Griguolo S. e L. Vettoretto (1983), *ADDAEST: un 'package' per l'analisi di dati territoriali*, Venezia, DAEST.

Grizzle J.E., C.F. Starmer e G.G. Koch (1969), 'Analysis of categorical data by linear models', *Biometrics*, 25, pp. 489-504.

Gronau R. (1974), 'Wage comparisons. A selectivity bias', *Journal of Political Economy*, 82, pp. 1119-1143.

Gurney M. e J.F. Daly (1965), 'A multivariate approach to estimation in periodic sample surveys', in American Statistical Association, *Proceedings of the Social Statistics Section*, pp. 242-257.

Ham J. (1982), 'Estimation of a labor supply model with censoring due to unemployment and underemployment', *Review of Economic Studies*, 49, pp. 335-354.

Han A. e J.A. Hausman (1990), 'Flexible parametric estimation of duration and competing risk models', *Journal of Applied Econometrics*, 5, pp. 1-28.

Hansen M.H., W.N. Hurwitz, H. Nisselson e J. Steinberg (1955), 'The redesign of the Census Current Population Survey', *Journal of the American Statistical Association*, 50, pp. 701-719.

Harvey A.C. (1984), 'A unified view of statistical forecasting procedures', *Journal of Forecasting*, 3, pp. 245-275.

Harvey A.C. (1985), 'Trends and cycles in macroeconomic time series', *Journal of Business & Economic Statistics*, 3, pp. 216- 227.

Harvey A.C. e P.J.H. Todd (1983), 'Forecasting economic time series with structural and Box-Jenkins models: a case study', *Journal of Business & Economic Statistics*, 1, pp. 299-315.

Heckman J.J. (1974), 'Shadow prices, market wages and labor supply', *Econometrica*, 42, pp. 679-694.

Heckman J.J. (1979), 'Sample selection bias as a specification error', *Econometrica*, 47, pp. 153-161.

Heckman J.J. e G.J. Borjas (1980), 'Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence', *Economica*, 47, pp. 247-283.

Heckman J.J. e V.J. Hotz (1989), 'Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of Manpower Training', *Journal of the American Statistical Association*, 84, pp. 862-874.

Heckman J.J., M. Killingsworth e T. MaCurdy (1981), 'Empirical evidence on static labour supply models: a survey of recent developments', in Z. Hornstein, J. Grice e A. Webb (eds.), *The economics of the labour market*, London, HMSO, pp. 75-122.

Heckman J.J. e T. MaCurdy (1986), 'Labor econometrics', in Z. Griliches e M.D. Intriligator (eds.), *Handbook of econometrics*, Amsterdam, North-Holland, vol. 3, pp. 1917-1977.

Heckman J.J. e B. Singer (1984), 'A method for minimizing the impact of distributional assumptions in economic models for duration data', *Econometrica*, 52, pp. 271-320.

Heckman J.J. e B. Singer (eds.) (1985), *Longitudinal analysis of labor market data*, Cambridge, Cambridge University Press.

Heckman J.J. e B. Singer (1986), 'Econometric analysis of longitudinal data', in Z. Griliches e M.D. Intriligator (eds.), *Handbook of econometrics*, Amsterdam, North-Holland, vol. 3, pp. 1689-1763.

Herriot R.A. e E.M. Spiers (1975), 'Measuring the impact on income statistics of reporting differences between the Current Population Survey and amministrative sources', in American Statistical Association, *Proceedings of the Social Statistics Section*, pp. 147-158.

Hill D.H. (1988), 'Response errors around the seam: analysis of change in a panel with overlapping reference periods', paper presented at the SSRC Conference on "Individuals and Families in Transition: Understanding Change trough Longitudinal Data", Annapolis, MD, March 16-18, 1988, (mimeo).

Hillmer S.C. e G.C. Tiao (1982), 'An ARIMA-model-based approach to seasonal adjustment', *Journal of the American Statistical Association*, 77, pp. 63-70.

Hogue C.R. (1985), 'History of the problems encountered in estimating gross flows', in *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, Washington, D.C., U.S. Department of Commerce and U.S. Department of Labor, pp. 1-7.

Holt D., A.J. Scott e P.D. Ewings (1980), 'Chi-squared tests with survey data', *Journal of the Royal Statistical Society*, A, 143, pp. 303-320.

Holt D. e T.M.F. Smith (1979), 'Post-stratification', *Journal of the Royal Statistical Society*, A, 142, pp. 33-46.

Howe G.R. e J. Lindsay (1981), 'A generalized iterative record linkage computer system for use in medical follow-up studies', ristampa in *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies*, Washington, D.C., U.S. Department of the Treasury, Internal Revenue Service, Statistics of Income Division, Publ. 1299 (2-86).

Huang E.T. e L.R. Ernst (1981), 'Comparison of an alternative estimator to the current composite estimator in CPS', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 303-308.

Hujer R. e H. Schneider (1989), 'The analysis of labor market mobility using panel data', *European Economic Review*, 33, pp. 530-536.

Hussmanns R., F. Merhan e S.M. Verma (1990), *Surveys of economically active population, employment and underemployment: an ILO manual on concepts and definitions*, Geneva, ILO.

ILO (1980), 'A building block approach for international comparisons of employment and unemployment statistics', Geneva, (mimeo).

ILO (1983), 'Resolution concerning statistics of the economically active population, employment, unemployment and underemployment', *Bulletin of Labour Statistics*, 3, pp. IX-XV.

ILO (1986), *Statistical sources and methods. Volume 3. Economically active populations, employment and hours of work (household surveys)*, Geneva.

Imrey P.B., G.C. Koch e M.E. Stokes (1982), 'Categorical data analysis: some reflections on the log-linear model and logistic regression. Part II: data analysis', *International Statistical Review*, 50, pp. 35-63.

Imrey P.B., E. Sobel e M. Francis (1980), 'Modelling contingency tables from complex surveys', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 212-217.

Isaki T.C. (1990), 'Small-area estimation of economic statistics', *Journal of Business & Economic Statistics*, 8, pp. 435-441.

Istat (1978), *Rilevazioni campionarie sulle forze di lavoro*, Metodi e Norme, serie A, n. 15, Roma.

Istat (1984), 'Commissione di studio per un sistema informativo del lavoro. Rapporto conclusivo', Roma, (mimeo).

Istat (1985a), 'Programma di lavoro del Reparto Studi per la messa a punto di una metodologia generalizzata per indagini campionarie sulla popolazione', Roma, (cicl.).

Istat (1985b), *Indagine sulle strutture e i comportamenti familiari*, Roma.

Istat (1989), 'Giornata di studio su "Disegno progettuale della nuova rilevazione delle forze di lavoro". Materiali', Roma, 15 dicembre 1989, (cicl.).

Jabine T.B. e F.G. Scheuren (1986), 'Record linkage for statistical purposes: methodological issues', *Journal of Official Statistics*, 2, pp. 255-277.

Jagers P. (1986), 'Post-stratification against bias in sampling', *International Statistical Review*, 54, pp. 159-167.

Jagers P., A. Odén e L. Trulsson (1985), 'Post-stratification and ratio estimation: usages of auxiliary information in survey sampling and opinion polls', *International Statistical Review*, 53, pp. 221-238.

Jones R.G. (1980), 'Best linear unbiased estimators for repeated surveys', *Journal of the Royal Statistical Society*, B, 42, pp. 221-226.

Kalachek E. (1979), 'Longitudinal survey and labor market analysis', in National Commission on Employment and Unemployment Statistics, *Data collection, processing and presentation: national and local*, Washington, D.C., U.S. Government Printing Office, pp. 160-187.

Kalton G. (1983), *Compensating for missing survey data*, Working paper, Survey Research Center, Ann Arbor, University of Michigan.

Kalton G. (1986), 'Handling wave nonresponse in panel surveys', *Journal of Official Statistics*, 2, pp. 303-314.

Kalton G., D. Kasprzyk e D. McMillen (1989), 'Nonsampling errors in panel surveys', in D. Kasprzyk, G. Duncan, G. Kalton e M.P. Singh (eds.), *Panel surveys*, New York, Wiley, pp. 249-270.

Kalton G. e J.M. Lepkowski (1985), 'Following rules in the Survey of Income and Program Participation', *Journal of Economic and Social Measurement*, 13, pp. 319-329.

Kalton G. e E.M. Miller (1986), 'Effects of adjustment for wave nonresponse on panel survey estimates', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 194-199.

Kaplan B., I. Francis e J. Sedransk (1979), 'A comparison of methods and programs for computing variances of estimators from complex sample surveys', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 97- 100.

Kelley P.R. (1985), 'Advances in record linkage methodology: a method for determining the best blocking strategy', in *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies*, Washington, D.C., U.S. Department of the Treasury, Internal Revenue Service, Statistics of Income Division, Publ. 1299(2-86).

Kemeny J.G. e J.L. Snell (1960), *Finite Markov chains*, Princeton, D. Van Nostrand.

Kennan J. (1988), 'An econometric analysis of fluctuations in aggregate labor supply and demand', *Econometrica*, 56, pp. 317- 333.

Kiefer N.M. (1988), 'Econometric duration data and hazard functions', *Journal of Economic Literature*, 26, pp. 646-679.

Kiernan K.E. (1986), 'Leaving home: living arrangements of young people in six West-European countries', *European Journal of Population*, 2, pp. 177-184.

Kirkendall N.J. (1985), 'Weights in computer matching: applications and an information theoretic point of view', in *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies*, Washington, D.C., U.S. Department of the Treasury, Internal Revenue Service, Statistics of Income Division, Publ. 1299(2-86).

Kish L. (1962), 'Studies of interviewer variance for attitudinal variables', *Journal of the American Statistical Association*, 57, pp. 92-115.

Kish L. (1965), *Survey sampling*, New York, Wiley.

Kitagawa G. e W. Gersch (1984), 'A smoothing prior state space modelling of time series with trend and seasonality', *Journal of the American Statistical Association*, 79, pp. 378-389.

Kobrin Goldscheider F. e L.J. Waite (1987), 'Nest-leaving patterns and the transition to marriage for young men and women', *Journal of Marriage and the Family*, 49, pp. 507-516.

Koch G.C., D.H. Jr. Freeman e J. Freeman (1975), 'Strategies in the multivariate analysis of data from complex surveys', *International Statistical Review*, 43, pp. 59-76.

Kroonenberg P.M. (1983), *Three-mode principal component analysis*, Leiden, DSWO Press.

Kumar S. e H. Lee (1985), *Evaluation of composite estimation for the Canadian Labour Force Survey*, CHSM-067E, Ottawa, Statistics Canada.

Kumar S. e J.N.K. Rao (1984), 'Logistic regression analysis of Labour Force Survey

data', *Survey Methodology*, 10, pp. 62-81.
Lancaster T. (1979), 'Econometric methods for the duration of unemployment', *Econometrica*, 47, pp. 939-956.
Lancaster T. (1990), *The econometric analysis of transition data*, Cambrige, Cambrige University Press.
Lancaster T. e S. Nickell (1980), 'The analysis of re-employment probabilities for the unemployed', *Journal of the Royal Statistical Society*, A, 143, pp. 141-165 (with discussion).
Lavit C. (1988), *Analyse conjointe de tableaux quantitatifs*, Paris, Masson.
Law H.G., C.W. Snyder, J.A. Hattie e R.P. Mc Donald (eds.) (1984), *Research methods in multimode data analysis*, New York, Praeger.
Lawless J.F. (1982), *Statistical models and methods for lifetime data*, New York, Wiley.
Lebart L., A. Morineau *et al.* (1985), *SPAD-Système Portable pour l'Analyse des Données*, Paris, C.E.S.I.A..
Lebart L., A. Morineau e N. Tabard (1977), *Techniques de la description statistique, méthodes et logiciels pour l'analyse des grands tableaux*, Paris, Dunod.
Lebart L., A. Morineau e K.M. Warwick (1984), *Multivariate descriptive statistical analysis, correspondence analysis and related techniques for large matrices*, New York, Wiley.
Lee H. (1987), 'Estimation of panel correlations in the Canadian Labour Force Survey', Social Survey Methods Division, Ottawa Statistics Canada, (mimeo).
Lepkowski J.M.(1982), 'The use of OSIRIS IV to analyse complex sample survey data', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 38- 43.
Little R.J.A. (1985), 'Nonresponse adjustments in longitudinal surveys: models for categorical data', *Bulletin of the International Statistical Institute*, vol. LI, book 3, pp. 15.I.1-18.
Little R.J.A. (1986), 'Survey nonresponse adjustment for estimates of means', *International Statistical Review*, 54, pp. 139-157.
Little R.J.A. (1988), 'Incomplete data in event history analysis', paper presented at the IUSSP "Seminar on Event History Analysis", Paris, March 14-17, 1988, (mimeo).
Little R.J.A. e M.H. David (1983), 'Weighting adiustment for nonresponse in panel surveys', in W.G. Madow e I. Olkin (eds.), *Incomplete data in sample surveys*, New York, Academic Press, vol 2, pp. 143-184.
Little R.J.A. e D.B. Rubin (1987), *Statistical analysis with missing data*, New York, Wiley.
Lothian J. e M. Morry (1977), *The problem of aggregation: direct or indirect seasonal adjustment*, Research paper, Seasonal Adjustment and Time Series Staff, Ottawa, Statistics Canada.
Lovison G. e P.D. Falorsi (1990), *La selezione di modelli log-lineari parsimoniosi per l'analisi delle tabelle sulle forze di lavoro*, Rapporto di ricerca FOLA n. 16, Padova, Cleup.
Lucifora C. (1987), 'An aggregate disequilibrium model of the labour market: an empirical investigation', *Labour*, 1 (1), pp. 83- 107.
Lunardi A. (1989), 'L'impiego di OSIRIS IV nella modellazione di dati qualitativi da

disegni campionari complessi', Università di Padova, Facoltà di Scienze Statistiche Demografiche ed Attuariali, (tesi di laurea non pubblicata).
Lynch L.M. (1986), *The youth labor market in the '80s: determinants of re-employment probabilities for young men and women*, Working paper No. 2021, Cambrige, Mass., NBER.
Macredie I. (1983), *The impact of response errors in the estimation of labour market flows*, Labour Force Activity Section, Economic Characteristics Division, Ottawa Statistics Canada.
Macredie I. (1987), 'The Canadian Labour Force Survey: a future that builds in the past', paper presented at the Seminar "The Community Labour Force Survey in the 1990s", Luxembourg, Eurostat, 12-14 October 1987, (mimeo).
Maddala G. (1983), *Limited dependent and qualitative variables in econometrics*, Cambridge, Cambridge University Press.
Mahalanobis P.C. (1946), 'Recent experiments in statistical sampling in the Indian Statistical Institute', *Journal of the Royal Statistical Society*, A, 109, pp. 325-370.
Malinvaud E. (1984), *Mass unemployment*, Oxford, Basil Blackwell.
Malinvaud E. (1986), *Sur les statistiques de l'emploi et du chômage*, Paris, La Documentation Française.
Mamberti Pedullà G., C. Pascarella e C. Abate (1987), 'Le nuove stime dell'occupazione presente in contabilità nazionale: concetti, metodi e risultati', in U. Trivellato (a cura di), *Attendibilità e tempestività delle stime di contabilità nazionale*, Padova, Cleup, pp. 199-232,
Manoussakis (1977), 'Repeated sampling with partial replacement of units', *The Annals of Statistics*, 5, pp. 795-802.
Manton K.G., E. Stallard e J.W. Vaupel (1986), 'Alternative models for the heterogeneity of mortality risk among the aged', *Journal of the American Statistical Association*, 81, pp. 635-644.
Maravall A. (1984), 'Comment on "Issues involved with the seasonal adjustment of economic time series" of W.R. Bell e S.C. Hillmer', *Journal of Business & Economic Statistics*, 2, pp. 337-339.
Martini A. (1987), 'The discouraged worker effect: a reappraisal using spell duration data', University of Wisconsin-Madison, Department of Economics, (mimeo).
Martinotti G. (a cura di) (1982), *La città difficile. Equilibri e diseguaglianze nel mercato urbano*, Milano, Franco Angeli.
Martinotti G. (1987), 'Bisogni conoscitivi per la società italiana degli anni 90', *Economia & Lavoro*, 21 (2), pp. 3-29.
Masselli M. (1985), 'Nota sul progetto "Qualità dei dati" ', Roma, Istat, (cicl.).
Masselli M. (1988), 'L'errore di identificazione delle unità e il sistema di controllo di un'indagine statistica. Un'applicazione all'indagine sulle forze di lavoro', in Società Italiana di Statistica, *Atti della XXXIV Riunione Scientifica*, Siena, La Nuova Immagine, vol. 2, tomo 1, pp. 169-176.
Masselli M. (1989), 'L'*error profile* dell'indagine sulle forze di lavoro', Roma , Istat, (cicl.).
Mathiowetz N.A. e G.J. Duncan (1988), 'Out of work, out of mind: response errors in retrospective reports of unemployment', *Journal of Business & Economic Statistics*, 6, pp. 221-229.
McCarthy P.J. (1966), *Replication: an approach to the analysis of data from complex*

surveys, N.C.H.S. Vital and Health Statistics, ser. 2, n. 14, Washington, D.C., U.S. Government Printing Office.

McCarthy P.J. (1969), 'Pseudo-replication: half samples', *International Statistical Review*, 37, pp. 239-264.

McGuckin R.H. (1990), *Longitudinal economic data at the Census Bureau: a new database yields fresh insights on some old issues*, CES Discussion paper 90-1, Washington, D.C., U.S. Department of Commerce, Bureau of the Census.

McKenzie S. (1984), 'Concurrent seasonal adjustment with Census X-11', *Journal of Business & Economic Statistics*, 2, pp. 235- 249.

Meghir C. e M. Arellano (1988), *Using complementary data sources: an application to labour supply and job search*, UCL discussion paper, London, University College.

Meyer B.D. (1988), 'Classification error models and labor-market dynamics', *Journal of Business & Economic Statistics*, 6, pp. 385-390.

Micali A. (1990). 'La disoccupazione in Italia: livello e composizione interna', *Economia & Lavoro*, 24 (1), pp. 61-78.

Modigliani F., F. Padoa Schioppa e N. Rossi (1986), 'Aggregate unemployment in Italy, 1960-1983', *Economica*, 53 (210S, Supplement: Unemployment), pp. S245-S273.

Moffitt R. (1982), 'The Tobit model, hours of work and istitutional constraints', *The Review of Economics and Statistics*, 84, pp. 510-515.

Moriani C. (1981), 'Forze di lavoro e flussi di popolazione', Supplemento al *Bollettino Mensile di Statistica*, n. 15, Roma, Istat, pp. 5-15.

Mosimann J.E. (1962), 'On the compound multinomial distribution, the multivariate beta-distribution and correlation among proportions', *Biometrika*, 49, pp. 65-82.

Narendranathan W.S. e S. Nickell (1985), 'Modelling the process of job search', *Journal of Econometrics*, 28, pp. 71-84.

Nathan G. (1969), 'Tests of independence in contingency tables from stratified samples', in N.L. Johnson e H. Smith (eds.), *New developments in survey sampling*, New York, Wiley, pp. 578-600.

Nathan G. (1973), *Approximate tests of independence in contingency tables from complex stratified samples*, N.C.H.S. Vital and Health Statistics, ser. 2, n. 53, Washington D.C., U.S. Government Printing Office.

National Commission on Employment and Unemployment Statistics (1979), *Counting the labor force*, Washington, D.C., U.S. Governement Printing Office.

Newcombe H.B. (1988), *Handbook of record linkage*, Oxford, Oxford University Press.

Newcombe H.B. *et al.* (1959), 'Automatic linkage of vital records', ristampa in *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies*, Washington, D.C., U.S. Department of the Treasury, Internal Revenue Service, Statistics of Income Division, Publ. 1299(2-86).

Newcombe H.B. *et al.* (1983), 'Reliability of computerized versus manual death searches in a study of the health of Eldorado uranio workers', ristampa in *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies*, Washington, D.C., U.S. Department of the Treasury, Internal Revenue Service, Statistics of Income Division, Publ. 1299(2- 86).

Nickell S. (1979), 'Estimating the probability of leaving unemployment', *Econome-*

trica, 47, pp. 1247-1266.
OECD (1987), *Employment outlook 1987*, Paris.
OECD (1990), 'International comparisons of employment and unemployment', MAS/WS7/(89)10, Paris, (mimeo).
Ongaro F. (1990), *Riflessioni sulle tipologie familiari desumibili dall'indagine sulle forze di lavoro*, Rapporto di ricerca FOLA n. 17, Padova, Cleup.
van Opstal R. e J. Theeuwes (1985), 'Duration of unemployment in the Dutch youth labour market', The Hague, Central Planning Bureau, (mimeo).
Paass G. (1984), 'Statistical record linkage methodology', *Bulletin of the International Statistical Institute*, vol. LI, book 2, pp. 9.3.1-16.
Parenti, G. (1979), 'Evaluation of labor force's interindustry flows estimates based on a rotating sample', *Bulletin of the International Statistical Institute*, vol. XLVIII, book 2, pp. 409- 415.
Passamani G. e M. Schenkel (1988), *Forze di lavoro, salari, prezzi: analisi multivariate dinamiche disaggregate (1970-1986)*, Working paper n. 8, Progetto finalizzato CNR: "Struttura ed evoluzione dell'economia italiana", tema di ricerca I/01/C, Padova, Cleup.
Patterson H.D. (1950), 'Sampling on successive occasions with partial replacement of units', *Journal of the Royal Statistical Society*, B, 12, pp. 241-255.
Piccolo D. (1985), 'Progetto DESEC: un'esperienza di ricerca sulle serie storiche stagionali', *Quaderni di Statistica e Econometria*, 8, pp. 5-39.
Pollastri A. (1976a), 'La stima di matrici di transizione nell'ambito della mobilità delle forze di lavoro in Italia', *Rivista di Statistica Applicata*, 9, pp. 17-45.
Pollastri A. (1976b), 'The mover-stayer model for the occupational mobility in Italy', in M. Zenga (a cura di), *Saggi di statistica metodologica ed applicata*, Contributi del Centro Interdipartimentale di Studi Statistici, vol. 1, Cosenza, Università della Calabria.
Poterba J.M. e L.H. Summers (1984), 'Response variation in the CPS: caveats for the unemployment analyst', *Monthly Labor Review*, 107 (March), pp. 37-43.
Poterba J.M. e L.H. Summers (1985), 'Adjusting the gross change data: implications for labor market dynamics', in *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, Washington, D.C., U.S. Department of Commerce and U.S. Department of Labor, pp. 81-98.
Poterba J.M. e L.H. Summers (1986), 'Reporting errors and labor market dynamics', *Econometrica*, 54, pp. 1319-1388.
Pritzker L. e R.H. Hanson (1962), 'Measurement errors in the 1960 Census of Population', in American Statistical Association, *Proceedings of the Social Statistics Section*, pp. 80-90.
Purcell N.J. e S. Linacre (1976), 'Techniques for the estimation of small area characteristics', paper presented at the Australian Statistical Conference, Melbourne, August 18-20, 1976, (mimeo).
Quintieri B. e F.C. Rosati (1988), *Politica fiscale ed offerta di lavoro in Italia: un'analisi su serie temporali*, Working paper, Progetto finalizzato CNR: "Struttura ed evoluzione dell'economia italiana", tema di ricerca III/03/A, Roma, LUISS.
Rao J.N.K. (1985), *Analysis of categorical data from sample surveys*, Technical report, Laboratory for Research in Statistics and Probability, Ottawa, Carleton

University.
Rao J.N.K. e J.E. Graham (1964), 'Rotation designs for sampling on repeated occasions', *Journal of the American Statistical Association,* 59, pp. 492-509.
Rao J.N.K. e A.J. Scott (1981), 'The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables', *Journal of the American Statistical Association*, 76, pp. 221-230.
Rao J.N.K. e A.J. Scott (1984), 'On chi-squared tests for multiway tables with cell proportions estimated from survey data', *Annals of Statistics*, 12, pp. 46-60.
Rao J.N.K. e A.J. Scott (1985), *On simple adjustments to chi-square tests with sample survey data*, Technical report, Laboratory for Research in Statistics and Probability, Ottawa, Carleton University.
Rettore E. (1989), *Dati e modelli per l'analisi statica dell'offerta di lavoro a livello individuale. Con un'applicazione all'offerta di lavoro femminile*, Università di Padova, Dipartimento di Scienze Statistiche, (tesi di dottorato).
Rettore E., N. Torelli e U. Trivellato (1988), 'Disoccupazione e ricerca di lavoro: convenzioni definitorie e analisi esplorative sull'*attachment* al mercato del lavoro', *Economia & Lavoro*, 22 (3), pp. 71-94.
Royall R.M. e W.G. Cumberland (1981), 'An empirical study of the ratio estimator and estimators of its variance', *Journal of the American Statistical Association*, 76, pp. 66-88.
Rubin D.B. (1976), 'Inference and missing data', *Biometrika*, 63, pp. 581-592.
Russo A. (1988), 'A comparative analysis of some estimators utilized in the sample surveys of the household', in International Association for Official Statistics, *1st Conference, Rome, 4-7 October, 1988. Pre-proceedings*, pp. 237-241.
Russo A. (1990), 'Un metodo di stima composta AK per l'indagine Istat sulle forze di lavoro', in Società Italiana di Statistica, *Atti della XXXV Riunione Scientifica*, Padova, Cedam, vol. 2, pp. 257-264.
Russo A. e P.D. Falorsi (1990), *Un'analisi comparativa di alcune tecniche di stima per piccole aree per l'indagine sulle forze di lavoro*, Rapporto di ricerca FOLA n.18, Padova, Cleup.
Russo A., P.D. Falorsi, G. Coccia e G. D'Angiolini (1988), *Una metodologia per la valutazione degli effetti stratificazione, stadificazione, ponderazione e dell'effetto complessivo del disegno di campionamento per l'indagine Istat sulle forze di lavoro*, Rapporto di ricerca FOLA n. 3, Padova, Cleup.
Salant S.W. (1977), 'Search theory and duration data: a theory of sorts', *Quarterly Journal of Economics*, 91, pp. 39-57.
Sanetti I. e I. Settanni (1979), *Una metodologia di raccordo per le serie statistiche sulle forze di lavoro*, Note e Relazioni, n. 56, Roma, Istat.
Sanna F., I. Santini e S. Lauro (1988), 'Tipologie familiari ed approccio al mercato del lavoro', Nota interna FOLA n. 53, Dipartimento di Scienze Statistiche, Università di Padova, (cicl.).
Santini I., F. Sanna e S. Lauro (1989), *Relazioni tra caratteristiche occupazionali e forme familiari*, Rapporto di ricerca FOLA n. 15, Padova, Cleup.
Satterthwaite F.E. (1946), 'An approximate distribution of estimate of variance components', *Biometrics*, 2, pp. 110-114.
Schuster J.J. e D.J. Downing (1976), 'Two-way contingency tables for complex

sampling schemes', *Biometrika*, 63, pp. 271-276.
Scott A.J. e T.M.F. Smith (1974), 'Analysis of repeated surveys using time series methods', *Journal of the American Statistical Association*, 69, pp. 674-678.
Scott A.J., T.M.F. Smith e R.G. Jones (1977), 'The application of time series methods to the analysis of repeated surveys', *International Statistical Review*, 45, pp. 13-28.
Sestito P. (1989), *Misurazione dell'offerta di lavoro e tasso di disoccupazione*, Temi di discussione n. 132, Roma, Banca d'Italia.
Shack-Marquez J. (1985), *Interview group bias: effects of repeated interviewing on estimation of labour force status*, BLS Working paper, n. 154, Washington, D.C., U.S. Department of Labor.
Shah B.V. e L.M. LaVange (1982), 'Software for inference on linear models from survey data', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 34- 37.
Shiskin J. (1976), 'Employment and unemployment: the doughnut or the hole?', Monthly Labor Review, 99 (Febr.), pp. 3-10.
Shiskin J., A.H. Young e J.C. Musgrave (1967), *The X-11 variant of the Census method II seasonal adjustment program*, Technical paper n. 15, Washington, D.C., U.S. Department of Commerce, Bureau of the Census.
Siesto V. (1980), 'Le capacità informative delle nuove rilevazioni delle forze di lavoro', in L. Frey *et al.*, *Le informazioni quantitative sull'occupazione e la disoccupazione in Italia*, Milano, Franco Angeli, pp. 54-94.
Siesto V. (1982), 'Idee per un potenziamento dell'indagine campionaria dell'Istat sulle forze di lavoro', in U. Trivellato e A. Zuliani (a cura di), *Informazione statistica su scuola e mercato del lavoro e sulle politiche per l'occupazione giovanile*, Roma, Istituto della Enciclopedia Italiana, pp. 117-128.
Sikkel D. (1985), 'Models for memory effects', *Journal of the American Statistical Association*, 80, pp. 835-841.
Sims C.A. (1981), 'An autoregressive index model for the U.S., 1948-1975', in J. Kmenta e J.B. Ramsey (eds.), *Large scale macro-econometric models*, Amsterdam, North Holland, pp. 283- 327.
Singh M.P., J.D. Drew (1981), 'Redesigning continuous survey in a changing environment', *Survey Methodology*, 7, pp. 44-73.
Singh M.P., J.D. Drew e G.H. Choudhry (1984), 'Post '81 censal redesign of the Canadian Labour Force Survey', *Survey Methodology*, 10, pp. 127-140.
Singh M.P. e R. Tessier (1976), 'Some estimators for domain totals', *Journal of the American Statistical Association*, 71, pp. 322- 325.
Smith T.M.F. (1978), 'Principles and problems in the analysis of repeated surveys', in N.K. Namboodiri (ed.), *Current topics in survey sampling*, New York, Academic Press, pp. 201-216.
Solon G. (1985), 'Effects of rotation group bias on estimation of unemployment', in *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, Washington, D.C., U.S. Department of Commerce and U.S. Department of Labor, pp. 15-23.
Solon G. (1989), 'The value of panel data in economic research', in D. Kasprzyk, G. Duncan, G. Kalton e M.P. Singh (eds.), *Panel surveys*, New York, Wiley, pp. 486-496.

Stasny E.A. (1986), 'Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey', *Journal of the American Statistical Association*, 81, pp. 42-47.

Stasny E.A. (1987), 'Some Markov-chain models for nonresponse in estimating gross labour force flows", *Journal of Official Statistics*, 3, pp. 359-373.

Stasny E.A. (1988), 'Modeling non-ignorable nonresponse in categorical panel data: an example in estimating gross labor-force flows', *Journal of Business & Economic Statistics*, 6, pp. 207-219.

Stasny E.A. e S.E. Fienberg (1985), 'Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse', in *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, Washington, D.C., U.S. Department of Commerce and U.S. Department of Labor, pp. 25-39.

Statistics Canada (1977), *Methodology of the Canadian Labour Force Survey 1976*, Catalogue 71-526, Occasional, Ottawa.

Statistics Canada (1979), *Estimation of gross flows from Canadian Labour Force Survey*, Working paper, Economics Characteristics Staff of Labour Force Activities Section, Toronto.

Statistics Canada (1987), *Labour Market Activity Survey information manual*, ISS 1121, Ottawa.

Sudman S. e N.M. Brandburn (1973), 'Effects of time and memory factors on response in surveys', *Journal of the American Statistical Association*, 68, pp. 805-815.

Tam S.M. (1987), 'Analysis of repeated surveys using a dynamic linear model', *International Statistical Review*, 55, pp. 63-74.

Tepping B.J. (1968), 'A model for optimum linkage of records', *Journal of the American Statistical Association*, 63, pp. 1321- 1332.

Thomas D.R. e J.N.K. Rao (1984), 'A MonteCarlo study of exact levels of goodness-of-fit statistics under cluster sampling', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 207-211.

Thomsen I. e D. Tesfu (1988), 'On the use of models in sampling from finite populations' in P.R. Krishnaiah e C.R. Rao (eds.), *Sampling*, Amsterdam, North Holland, pp. 369-398.

Tomberlin T.J. (1979), 'The analysis of contingency tables of data from complex samples', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 152- 157.

Tomberlin T.J. (1980), 'A model-based approach to the analysis of contingency tables of data from complex samples', in American Statistical Association, *Proceedings of the Section on Survey Research Methods*, pp. 230-236.

Torelli N. (1990), 'L'analisi dei dati di durata nelle scienze sociali: riflessioni metodologiche e una applicazione alla durata della disoccupazione', *Rivista di Statistica Applicata*, 2, pp. 359-378.

Torelli N. e U. Trivellato (1989), 'Youth unemployment duration from the Italian labour force survey: accuracy issues and modelling attempts', *European Economic Review*, 33, pp. 407- 415.

Tramblay V. (1986), 'Pratical criteria for definition of weighting classes', *Survey Methodology*, 12, pp. 85-97.

Trivellato U. (1980), 'Transition from school to working life. Review and methodo-

logical summary of the main surveys carried out in some Member Countries', Luxembourg, Eurostat, (mimeo).

Trivellato U. (1981), 'Le fonti sul mercato del lavoro ai vari livelli territoriali: un'analisi critica', *Economia & Lavoro*, 15 (2), pp. 59-77.

Trivellato U. (1986), 'Sulla valutazione dell'accuratezza di stime provvisorie di aggregati economici', in Autori Vari, *Studi in onore di Silvio Vianelli*, Palermo, Università degli studi di Palermo, vol. II, pp. 1589-1622.

Trivellato U. (1990), 'Le rilevazioni dell'occupazione e delle forze di lavoro', in Autori Vari, *La statistica italiana per l'Europa del 1993*, Roma, (in corso di pubblicazione).

Trivellato U., Bernardi L. e Fabbris L. (1987), *FOLA: Revisiting the survey and modelling data on labour forces. A research project on the Italian labour force sample survey*, Rapporto di ricerca FOLA, n. 1, Padova, Cleup.

Trivellato U., G. Marliani e N. Torelli (1989), 'Longitudinal analysis of unemployment duration from a household survey with rotating sample: a case study with Italian labour force data', in Bureau of the Census, *Fifth Annual Research Conference March 19-22, 1989. Proceedings*, Washington, D.C., U.S. Department of Commerce, Bureau of the Census, pp. 408-427.

Trivellato U. e N. Torelli (1989), 'Analysis of labour force dynamics from rotating panel survey data', *Bulletin of the International Statistical Institute*, vol. LIII, book 2, pp. 425-444.

Trussel J. e T. Richards (1985), 'Correcting for unobserved heterogeneity in hazard models: an application of the Heckman- Singer procedure to demographic data', in N.B. Tuma (ed.), *Sociological Methodology 1985*, San Francisco, Jossey Bass, pp. 242-276.

U.S. Bureau of the Census (1978), *The Current Population Survey. Design and methodology*, Technical paper n. 40, Washington, D.C., Department of Commerce.

U.S. Bureau of the Census (1982), *Development and use of longitudinal establishment data*, Workshop on the Development and Use of Longitudinal Establishment Data, Economic Research Report n. 4, Washington D.C., U.S. Government Printing Office.

U.S. Bureau of the Census (1987), 'March 1987 CPS interviewer's instructions', Washington D.C., U.S. Department of Commerce, (mimeo).

U.S. Department of Labor (1980), *Using the Current Population Survey as a longitudinal data base*, Bureau of Labor Statistics, Report 608, Washington, D.C..

Veevers R. e I. Macredie (1983), *Estimating gross flows from the Canadian Labour Force Survey*, Labour Force Activity Section, Economic Characteristics Division, Ottawa, Statistics Canada.

Verma V., C. Scott e C. O'Muircheartaigh (1980), 'Sample design and sampling errors for the World Fertility Survey', *Journal of the Royal Statistical Society*, A, 143, pp. 86-118.

Williams W.H. e C.L. Mallow (1970), 'Systematic biases in panel surveys due to differential nonresponse', *Journal of the American Statistical Association*, 65, pp 1338-1349.

Wolter K.M. (1979), 'Composite estimation in finite populations', *Journal of the*

American Statistical Association, 74, pp. 604- 613.

Wong F. (1983), 'A technique to correct the response bias in the 4x4 labour force gross flow matrix', Technical report, Ottawa, Statistics Canada, (mimeo).

Woodruff R.S. (1971), 'A simple method for approximating the variance of a complicated estimate', *Journal of the American Statistical Association*, 66, pp. 411-414.

Yates F. (1981), *Sampling methods for censuses and surveys*, London, Griffin.

Young C. (1987), *Young people leaving home in Australia*, Australian Family Formation Project, Monograph n. 9, Melbourn, Australian Institut of Family Studies.

Zaccarin S. e L. Bernardi (1984), 'Indicatori di mobilità: applicazione di un modello markoviano ai dati della rilevazione trimestrale delle forze di lavoro', *Economia & Lavoro*, 17 (4), pp. 91-103.

PUBBLICAZIONI ISTAT

BOLLETTINO MENSILE DI STATISTICA
La più completa ed autorevole raccolta di dati congiunturali concernenti l'evoluzione dei fenomeni demografici, sociali, economici e finanziari
Abbonamento annuo L. 115.000 (Estero L. 139.000) Ogni fascicolo L. 15.000
INDICATORI MENSILI
Forniscono dati riassuntivi e tempestivi sull'andamento mensile dei principali fenomeni interessanti la vita nazionale
Abbonamento annuo L. 29.000 (Estero L. 35.000) Ogni fascicolo L. 3.700
NOTIZIARI ISTAT
Forniscono i primi risultati delle rilevazioni ed elaborazioni statistiche riguardanti l'attività produttiva, i prezzi, il commercio interno, gli scambi internazionali come pure lo stato ed il movimento della popolazione e le sue caratteristiche sociali e sanitarie.
I dati, esposti in grafici e tabelle, sono accompagnati da commenti, illustrazioni e note interpretative.
Serie 1 - Statistiche demografiche e sociali
Abbonamento annuo L. 22.000 (Estero L. 29.000) Una copia L. 1.600
Serie 2 - Statistiche dell'attività produttiva
Abbonamento annuo L. 64.000 (Estero L. 85.000) Una copia L. 1.600
Serie 3 - Statistiche del lavoro, delle retribuzioni e dei prezzi
Abbonamento annuo L. 22.000 (Estero L. 29.000) Una copia L. 1.600
Serie 4 - Argomenti vari
Abbonamento annuo L. 13.000 (Estero L. 17.000) Una copia L. 1.600
Abbonamento annuo a tutte le serie L. 106.000 (Estero L. 144.000).
INDICATORI TRIMESTRALI
Conti economici trimestrali
Abbonamento annuo L. 11.000 (Estero L. 13.000) Ogni fascicolo L. 3.700
STATISTICA DEL COMMERCIO CON L'ESTERO
Documentazione statistica ufficiale, a periodicità trimestrale, sul commercio dell'Italia con l'estero; fornisce, per tutte le merci comprese nella classificazione merceologica della tariffa dei dazi doganali, l'andamento delle importazioni e delle esporazioni da e per i principali Paesi
Abbonamento annuo L. 99.000 (Estero L. 112.000) Ogni fascicolo L. 31.000

Abbonamento annuo cumulativo a tutti i periodici, compresa la "Statistica del commercio con l'estero": L. 300.000 (Estero L. 390.000); esclusa la "Statistica del commercio con l'estero" L. 209.000 (Estero L. 286.000)
Gli abbonamenti decorrono dal 1° gennaio anche se sottoscritti nel corso dell'anno. In tal caso l'abbonato riceverà i numeri dell'annata già pubblicati. L'abbonato ai periodici ISTAT ha diritto a ricevere gratuitamente i fascicoli non pervenutigli soltanto se ne segnalerà il mancato arrivo entro 10 giorni dal ricevimento del fascicolo successivo. Decorso tale termine, si spediscono solo contro rimessa dell'importo. Le variazioni di indirizzo devono essere segnalate dall'abbonato per iscritto. Nel sottoscrivere l'abbonamento cumulativo, gli interessati possono chiedere che l'ISTAT provveda, senza ulteriori richieste, all'invio di tutte le pubblicazioni non periodiche non appena liberate dalle stampe, contro assegno o con emissione di fattura, con lo sconto del 30%. Le singole pubblicazioni possono essere richieste direttamente all'Istituto nazionale di statistica (Via Cesare Balbo, 16 - 00100 Roma) versando il relativo importo, maggiorato del 10% per spese di spedizione, sul c/c postale n. 619007.
Tutti i prezzi sono riferiti all'anno 1991.

ANNUARIO STATISTICO ITALIANO - Edizione 1990 - L. 46.000
Sintetizza in semplici tabelle numeriche di facile lettura ed attraverso appropriate note illustrative e rappresentazioni grafiche, i dati fondamentali della vita economica, demografica e sociale e fornisce un quadro panoramico della corrispondente situazione degli altri principali Paesi del mondo.
COMPENDIO STATISTICO ITALIANO - Edizione 1991 - L. 24.000
Sintetizza i risultati delle rilevazioni ed elaborazioni statistiche di maggior interesse nazionale.
ITALIAN STATISTICAL ABSTRACT - Edition 1991 - L. 24.000
Fornisce i principali risultati delle rilevazioni ed elaborazioni statistiche concernenti la situazione sociale ed economica italiana - Edizione in lingua inglese.
I CONTI DEGLI ITALIANI - Vol. 25, edizione 1991 - L. 17.000
Illustra in forma divulgativa i principali aspetti quantitativi dell'economia italiana.
LE REGIONI IN CIFRE - Edizione 1991 - Distribuzione gratuita
Fornisce i dati delle singole regioni e delle due grandi ripartizioni geografiche: Nord-Centro e Mezzogiorno.

ANNUARI

STATISTICHE DEMOGRAFICHE
n. 34 - Anno 1985
Tomo 1, parte prima - Movimento e calcolo della popolazione secondo gli atti anagrafici - L. 11.000
Tomo 1, parte seconda - Iscrizioni e cancellazioni anagrafiche per trasferimento di residenza, 1984 - Espatriati e rimpatriati, 1985 - L. 9.500
n. 33/34 - Anni 1984 e 1985
Tomo 2, parte prima - Nascite e decessi - L. 38.000
Tomo 2, parte seconda - Matrimoni, separazioni e divorzi - L. 15.000
n. 35 - Anno 1986
Tomo 1, parte prima - Popolazione residente e movimento anagrafico dei Comuni - L. 11.500
Tomo 1, parte seconda - Iscrizioni e cancellazioni anagrafiche, 1985 e 1986 - Espatriati e rimpatriati, 1986 - L. 15.800
n. 36 - Anno 1987
Tomo 1, parte prima - Popolazione residente e movimento anagrafico dei Comuni - L. 18.900
Tomo 1, parte seconda - Iscrizioni e cancellazioni anagrafiche - Espatri e rimpatri, 1987 - L. 15.000
n. 35/36 - Anni 1986 e 1987
Tomo 2, parte prima - Nascite e decessi *(in preparazione)*
Tomo 2, parte seconda - Matrimoni, separazioni e divorzi - L. 16.000
Raccoglie i dati sulla dinamica demografica italiana, sia naturale che migratoria, nonchè dei dati sintetici sul movimento annuale della popolazione residente anagrafica comunale e sul suo ammontare.

POPOLAZIONE E MOVIMENTO ANAGRAFICO DEI COMUNI - n. 2 - Anno 1989 - L. 20.000
Riporta i dati dell'ammontare della popolazione residente, desunti dall'analisi del movimento naturale e di quello migratorio, nonchè la stima della popolazione residente per sesso ed età a livello regionale.

STATISTICHE DELLA SANITA' - n. 4 - Anno 1988 - L. 25.000
Riunisce le statistiche sulle strutture e sull'attività degli Istituti di cura, sulle malattie infettive e diffusive soggette a denuncia obbligatoria, sulle interruzioni volontarie della gravidanza e sugli aborti spontanei.

CAUSE DI MORTE - n. 4 - Anno 1988 - L. 29.000
Raccoglie i dati relativi alle statistiche sulle cause di morte e di nati-mortalità.

STATISTICHE DELLA PREVIDENZA, DELLA SANITA' E DELL'ASSISTENZA SOCIALE
n. 29 - Anni 1988, 1989 - L. 22.000
Vengono illustrate alcune forme di attività svolte dai vari Istituti nel settore della previdenza sociale, i conti economici delle Unità Sanitarie Locali e degli Istituti ospedalieri pubblici, nonchè i principali aspetti dell'assistenza sociale.

STATISTICHE DELL'ISTRUZIONE - n. 40 - Anno scolastico 1986-87
Tomo 1 - Dati analitici: nazionali, regionali e provinciali - L. 23.000
Tomo 2 - Dati riassuntivi comunali - L. 18.000
Quadro statistico completo ed aggiornato della situazione scolastica del Paese, attraverso dati sui vari rami d'insegnamento esaminati sotto i più interessanti aspetti dell'ordinamento degli studi e dei risultati conseguiti dagli iscritti.

STATISTICHE CULTURALI - n. 30 - Anno 1988 - L. 16.000 (in corso di stampa)
Documentazione ufficiale completa sulle principali attività culturali concernenti, tra l'altro, la produzione libraria, la pubblicazione di riviste scientifiche, la stampa periodica e le biblioteche.

STATISTICHE GIUDIZIARIE - n. 37 - Anno 1989 - L. 44.000 (in corso di stampa)
Ampia documentazione statistica dell'attività giudiziaria nonchè dei principali fenomeni in materia civile e penale nel campo della criminalità e degli Istituti di prevenzione e pena.

STATISTICHE DELL'AGRICOLTURA, ZOOTECNIA E MEZZI DI PRODUZIONE - n. 36 - Anno 1988 - L. 41.000
Contiene i dati relativi ai vari aspetti dell'agricoltura nazionale, nonchè i dati sulla consistenza e produttività degli allevamenti.

STATISTICHE FORESTALI - n. 41 - Anno 1988 - L. 16.000
Fornisce un quadro completo sulla struttura delle foreste italiane e delle relative utilizzazioni legnose, unitamente ad alcuni aspetti economici.

STATISTICHE METEOROLOGICHE - n. 24 - Anno 1983 - L. 15.800
Raccoglie i dati relativi alle temperature, piovosità e altri fattori climatici rilevati da una rete di stazioni ed osservatori distribuiti nel territorio nazionale.

STATISTICHE DELLA CACCIA E DELLA PESCA - n. 4 - Anno 1988 - L. 12.000
Raccoglie i dati sull'attività della pesca e sulla consistenza del relativo naviglio, nonchè su alcuni aspetti del settore venatorio.

STATISTICHE INDUSTRIALI - n. 28 - Anni 1986 e 1987 - L. 41.000
Nel suo genere, unica e veramente preziosa pubblicazione in cui sono organicamente raccolte tutte le informazioni statistiche fondamentali concernenti il complesso ed importante settore dell'industria.

STATISTICHE DELL'ATTIVITA' EDILIZIA - n. 3 - Anno 1988 - L. 22.000 *(in corso di stampa)*
Fornisce i risultati del settore dell'attività edilizia relativamente ai fabbricati residenziali e non residenziali.

STATISTICHE DELLE OPERE PUBBLICHE - n. 3 - Anno 1988 - L. 12.000
Statistica ufficiale delle opere pubbliche effettuate dallo Stato e da Enti pubblici, nonchè da privati con finanziamento parziale dello Stato.

STATISTICHE DEL COMMERCIO INTERNO - n. 31 - Anni 1988, 1989 - L. 12.000
Fornisce i risultati delle rilevazioni correnti relativi al fenomeno della distribuzione. Vi figurano gli indici mensili delle vendite al minuto, nonchè la più recente distribuzione per Comune delle licenze di esercizio.

STATISTICHE DEL TURISMO - n. 4 - Anno 1989 - L. 12.000
Descrive il sistema delle informazioni statistiche sul turismo ed espone, in un quadro organico, statistiche, dati ed indicatori aventi per oggetto i principali aspetti di questo fenomeno.

STATISTICHE DELLA NAVIGAZIONE MARITTIMA - n. 43 - Anno 1988 - L. 20.000
Contiene i dati statistici sul movimento dei natanti e del relativo carico avvenuto nei porti marittimi e negli altri approdi autorizzati del territorio nazionale.

STATISTICA DEGLI INCIDENTI STRADALI - n. 38 - Anno 1990 - L. 22.000
La più completa ed aggiornata raccolta di dati su una materia di viva attualità.

STATISTICA ANNUALE DEL COMMERCIO CON L'ESTERO - n. 44 - Anno 1987
Tomo 1 - Dati generali e riassuntivi - L. 41.000
Tomo 2 - Merci per Capitoli merceologici e Paesi
- Parte prima: da Cap. 1 a Cap. 24 - L. 14.000
- Parte seconda: da Cap. 25 a Cap. 40 - L. 18.000
- Parte terza: da Cap. 41 a Cap. 67 - L. 21.000
- Parte quarta: da Cap. 68 a Cap. 83 - L. 18.000
- Parte quinta: da Cap. 84 a Cap. 85 - L. 25.000
- Parte sesta: da Cap. 86 a Cap. 99 - L. 18.000
- Appendice: L. 10.000

Riporta i dati definitivi sull'andamento delle importazioni e delle esportazioni con l'analisi completa del movimento per merci e per Paesi. Nel tomo primo è riportata, tra l'altro, un'ampia documentazione sul movimento delle merci nei depositi doganali e sul commercio di transito.

STATISTICHE DEI BILANCI DELLE AMMINISTRAZIONI REGIONALI, PROVINCIALI E COMUNALI - n. XXVII - Anno 1982 - L. 14.000
Espone i dati relativi ai bilanci delle Amministrazioni, tenendo conto dell'aspetto contabile, funzionale ed amministrativo dei documenti contabili. Per le Amministrazioni provinciali e comunali è stata dedicata particolare attenzione ai dati riguardanti i servizi sociali, i settori d'intervento nel campo economico ed il personale.

STATISTICHE DEL LAVORO - n. 26 - Anno 1984 - L. 12.000
Organica ed aggiornata documentazione statistica su tutti i principali aspetti del mondo del lavoro.

CONTABILITA' NAZIONALE - n. 15 - Anni 1960-85 - L. 17.000
Contiene i dati sulla struttura e sulla evoluzione delle principali grandezze del sistema economico italiano.

COLLANA D'INFORMAZIONE

Anno 1991

n. 11 - COMMERCIO, ALBERGHI E SERVIZI VARI PER COMUNE AL 31 DICEMBRE 1988 - L. 22.000
n. 12 - STATISTICHE DELL'ISTRUZIONE - Dati sommari dell'anno scolastico 1988-89 - L. 22.000
n. 13 - STATISTICHE DELL'ISTRUZIONE - Dati sommari dell'anno scolastico 1989-90 - L. 21.000
n. 14 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (CAMPANIA) - L. 12.000
n. 15 - LAVORO E RETRIBUZIONI - Anno 1989 - L. 12.000
n. 16 - STATISTICHE DELLA ZOOTECNIA E DEI MEZZI DI PRODUZIONE IN AGRICOLTURA - Anno 1989 - L. 12.000
n. 17 - CONTI ECONOMICI DELLE IMPRESE CON ADDETTI DA 10 A 19 - Anno 1988 - L. 12.000
n. 18 - ACQUEDOTTI E RETI DI DISTRIBUZIONE DELL'ACQUA POTABILE IN ITALIA - Anno 1987 - L. 22.000
n. 19 - STATISTICHE SUI TRATTAMENTI PENSIONISTICI AL 31 DICEMBRE 1989 - L. 12.000
n. 20 - APPROVVIGIONAMENTO IDRICO, FOGNATURE E IMPIANTI DI DEPURAZIONE IN ITALIA - Anno 1987 - L. 25.000
n. 21 - LA DISTRIBUZIONE QUANTITATIVA DEL REDDITO IN ITALIA NELLE INDAGINI SUI BILANCI DI FAMIGLIA - Anno 1989 - L. 12.000
n. 22 - STATISTICA ANNUALE DELLA PRODUZIONE INDUSTRIALE - Anno 1988 - L. 12.000
n. 23 - BILANCI CONSUNTIVI DELLE REGIONI E DELLE PROVINCE AUTONOME - Anno 1988 - L. 25.000
n. 24 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (VENETO) - L. 12.000
n. 25 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (EMILIA-ROMAGNA) - L. 12.000 *(in corso di stampa)*
n. 26 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (LAZIO) - L. 12.000
n. 27 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (SARDEGNA) - L. 12.000 *(in corso di stampa)*
n. 28 - BILANCI CONSUNTIVI DELLE AMMINISTRAZIONI PROVINCIALI E COMUNALI - Anno 1987 - L. 25.000
n. 29 - STATISTICHE DELLA RICERCA SCIENTIFICA - CONSUNTIVO 1988 - PREVISIONE 1989 e 1990 - L. 12.000
n. 30 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (FRIULI-VENEZIA GIULIA) - L. 12.000
n. 31 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (PIEMONTE) - L. 12.000
n. 32 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (PUGLIA) - L. 12.000
n. 33 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (SICILIA) - L. 12.000
n. 34 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (TOSCANA) - L. 12.000
n. 35 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (BASILICATA) - L. 12.000 *(in corso di stampa)*
n. 36 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (LOMBARDIA) - L. 12.000 *(in corso di stampa)*
n. 37 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 (MARCHE) - L. 12.000 *(in corso di stampa)*
n. 38 - CONTI ECONOMICI NAZIONALI - Anni 1970-90 - L. 12.000
n. 39 - VALORE AGGIUNTO DELL'AGRICOLTURA PER REGIONE - Anni 1980-90 - L. 16.000 *(in corso di stampa)*

NOTE E RELAZIONI

Anno 1989

n. 1 - MANUALE DI TECNICHE DI INDAGINE (n. 7 fascicoli)
1. Pianificazione della produzione dei dati - L. 10.000 - 2. Il questionario: progettazione, redazione e verifica - L. 11.000 - 3. Tecniche di somministrazione del questionario - L. 11.000 - 4. Tecniche di campionamento: teoria e pratica - L. 20.000 - 5. Tecniche di stima della varianza campionaria - L. 16.000 - 6. Il sistema di controllo della qualità dei dati *(in corso di stampa)* - 7. Le rappresentazioni grafiche di dati statistici - L. 15.000
n. 2 - DISTRIBUZIONE PER ETA' DELLA POPOLAZIONE SCOLASTICA - Anno scolastico 1984-85 - L. 10.000
n. 3 - LA CRIMINALITA' ATTRAVERSO LE STATISTICHE - Anni 1971-87 - L. 14.000
n. 4 - PREVISIONI DELLA POPOLAZIONE RESIDENTE PER SESSO, ETA' E REGIONE - Base 1-1-1988
Tomo 1 - L. 18.000
Tomo 2 - L. 38.000
n. 5 - STATISTICHE SUI MINORENNI - Anni 1984-86 - L. 18.000
n. 6 - ANALISI DELLE FONTI STATISTICHE PER LA MISURA DELL'IMMIGRAZIONE STRANIERA IN ITALIA: ESAME E PROPOSTE - L. 10.000
n. 7 - NUMERI INDICI DEI PREZZI ALLA PRODUZIONE DEI PRODOTTI INDUSTRIALI - Base 1980 = 100 - L. 10.000

Anno 1990

n. 1 - METODOLOGIA E ANALISI DEI RISULTATI DELL'INDAGINE SULLE COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 - L. 11.000

n. 2 - LA MORTALITA' DIFFERENZIALE SECONDO ALCUNI FATTORI SOCIO-ECONOMICI - Anni 1981-82 - L. 11.000

Anno 1991

n. 1 - GLI IMMIGRATI PRESENTI IN ITALIA - UNA STIMA PER L'ANNO 1989 - L. 12.000
n. 2 - INDAGINE STATISTICA SULLE CONDIZIONI DI SALUTE DELLA POPOLAZIONE E SUL RICORSO AI SERVIZI SANITARI (Novembre 1986 - Aprile 1987) - L. 22.000

METODI E NORME

Serie A

n. 18 - NUMERI INDICI DEL COSTO DI COSTRUZIONE DI UN FABBRICATO RESIDENZIALE: Base 1976 = 100 - L. 1.500
n. 20 - NUMERI INDICI DEI PREZZI: Base 1980 = 100 - L. 4.500
n. 21 - NUMERI INDICI DEI PREZZI DEI PRODOTTI VENDUTI E DEI BENI ACQUISTATI DAGLI AGRICOLTORI: Base 1980 = 100 - L. 5.000
n. 23 - NUMERI INDICI DEI PREZZI AL CONSUMO: Base 1985 = 100 - L. 6.300
n. 25 - NUMERI INDICI DELLA PRODUZIONE INDUSTRIALE: Base 1985 = 100 - L. 11.000
n. 26 - NUMERI INDICI DEI PREZZI ALLA PRODUZIONE DEI PRODOTTI INDUSTRIALI: Base 1980 = 100 - L. 11.000
n. 27 - NUMERI INDICI DEL FATTURATO, DEGLI ORDINATIVI E DELLA CONSISTENZA DEGLI ORDINATIVI: Base 1985 = 100 - L. 11.000
n. 28 - NUMERI INDICI DEI PREZZI PRATICATI DAI GROSSISTI: Base 1989 = 100 - L. 12.000

Serie B

n. 21 - ISTRUZIONI PER LA RILEVAZIONE STATISTICA DEL MOVIMENTO DELLA POPOLAZIONE - L. 4.000
n. 22 - ISTRUZIONI PER LA RILEVAZIONE DEI DATI DELLE STATISTICHE FORESTALI - L. 6.000
n. 23 - ISTRUZIONI PER LA RILEVAZIONE DELL'ATTIVITA' EDILIZIA - L. 8.400
n. 24 - ISTRUZIONI PER LE RILEVAZIONI DELLE STATISTICHE GIUDIZIARIE
Tomo 1 - Procedura di rilevazione - L. 15.800
Tomo 2 - Modelli di rilevazione - L. 15.800
n. 25 - MANUALE PER LA PROGETTAZIONE DEI DATI STATISTICI - L. 10.000
n. 26 - ISTRUZIONI PER LE COMMISSIONI COMUNALI DI CONTROLLO DELLE RILEVAZIONI DEI PREZZI AL CONSUMO - L. 10.000
n. 27 - ISTRUZIONI PER LA RILEVAZIONE DELLE OPERE PUBBLICHE - L. 11.000
n. 28 - ISTRUZIONI PER LA RILEVAZIONE STATISTICA DEGLI INCIDENTI STRADALI - L. 11.000

Serie C

n. 10 - CLASSIFICAZIONI DELLE MALATTIE, TRAUMATISMI E CAUSE DI MORTE - Ristampa 1986
Vol. 1: Introduzione e parte sistematica - L. 16.000
Vol. 2: Indici alfabetici - L. 25.000
n. 11 - CLASSIFICAZIONE DELLE ATTIVITA' ECONOMICHE - Edizione 1991 - L. 25.000
n. 12 - CLASSIFICAZIONE DELLE PROFESSIONI - Edizione 1991 - L. 22.000

ANNALI DI STATISTICA

Serie IX

Vol. 1 - ATTI DEL 2° CONVEGNO SULL'INFORMAZIONE STATISTICA IN ITALIA (Roma, 17-19 giugno 1981) - L. 10.000
Vol. 3 - STUDI STATISTICI SUI CONSUMI - Dati dal 1959 al 1974 - L. 9.500
Vol. 5 - ATTI DEL SEMINARIO SULLA VALUTAZIONE DEI RISULTATI E DELLA METODOLOGIA DEI CENSIMENTI (Roma, 7-11 maggio 1984) - L. 25.000
Vol. 6 - ATTI DEL CONVEGNO "LA FAMIGLIA IN ITALIA" (Roma, 29-30 ottobre 1985) - L. 14.000
Vol. 7 - ATTI DEL CONVEGNO SULL'INFORMAZIONE STATISTICA E I PROCESSI DECISIONALI (Roma, 11-12 dicembre 1986) - L. 15.000
Vol. 8 - ATTI DEL SEMINARIO SULLE STATISTICHE ECOLOGICHE (Roma, 28 marzo-1 aprile 1988) - L. 23.000
Vol. 9 - NUOVA CONTABILITA' NAZIONALE - L. 23.000
Vol.10 - ATTI DELLA GIORNATA DI STUDIO SUL CAMPIONAMENTO STATISTICO (Roma, 27 Aprile 1989) - L. 25.000

CENSIMENTI

12° CENSIMENTO GENERALE DELLA POPOLAZIONE - 25 ottobre 1981
DATI SULLE CARATTERISTICHE STRUTTURALI DELLA POPOLAZIONE E DELLE ABITAZIONI - Campione al 2% dei fogli di famiglia - Dati provvisori - L. 5.000
Vol. I - Primi risultati provinciali e comunali sulla popolazione e sulle abitazioni *(dati provvisori)* - L. 6.500

Vol. II - Dati sulle caratteristiche strutturali della popolazione e delle abitazioni:
Tomo 1 - Fascicoli provinciali - Prezzi vari
Tomo 2 - Fascicoli regionali - Prezzi vari
Tomo 3 - Fascicolo nazionale - Italia - L. 25.000
Vol. III - Popolazione delle frazioni geografiche e delle località abitate dei comuni - Fascicoli regionali e nazionale - Prezzi vari
Vol. IV - Atti del censimento - L. 26.500
Vol. V - Relazione generale sul censimento - L. 25.000

POPOLAZIONE LEGALE DEI COMUNI - L. 8.000

6° CENSIMENTO GENERALE DELL'INDUSTRIA, DEL COMMERCIO, DEI SERVIZI E DELL'ARTIGIANATO - 26 ottobre 1981
Vol. I - Primi risultati sulle imprese e sulle unità locali - Dati provvisori
Tomo 1 - Dati nazionali, regionali e provinciali *(esaurito)*
Tomo 2 - Dati comunali *(esaurito)*
Vol. II - Dati sulle caratteristiche strutturali delle imprese e delle unità locali
Tomo 1 - Fascicoli provinciali - Prezzi vari
Tomo 2 - Fascicoli regionali - Prezzi vari
Tomo 3 - Fascicolo nazionale - Italia - L. 14.000
Vol. III - Atti del censimento - L. 11.000
Vol. IV - Relazione generale sul censimento - L. 26.500

3° CENSIMENTO GENERALE DELL'AGRICOLTURA - 24 ottobre 1982
CARATTERISTICHE STRUTTURALI DELLE AZIENDE AGRICOLE - L. 14.000
Vol. I - Primi risultati provinciali e comunali - Dati provvisori - L. 8.000
Vol. II - Caratteristiche strutturali delle aziende agricole:
Tomo 1: Fascicoli provinciali - Prezzi vari
Tomo 2: Fascicoli regionali - Prezzi vari
Tomo 3: Fascicolo nazionale - Italia - L. 11.000
Vol. III - Atti del censimento - L. 33.500
Vol. IV - Relazione generale sul censimento - L. 22.000 *(in corso di stampa)*
TIPOLOGIA DELLE AZIENDE AGRICOLE - Campione al 10% dei questionari d'azienda - L. 6.000
INDAGINE SULLE SUPERFICI A VITE
Vol. I - Caratteristiche delle aziende con vite
Tomo 1: Dati provinciali, regionali e nazionali - L. 33.500
Tomo 2: Dati comunali - L. 15.000
Vol. II - Caratteristiche dei vitigni - L. 33.500

4° CENSIMENTO GENERALE DELL'AGRICOLTURA - 21 ottobre 1990
CARATTERISTICHE DELLE AZIENDE AGRICOLE - Fascicolo nazionale - Risultati provvisori - L. 30.000

L'ITALIA DEI CENSIMENTI - L. 10.000

ALTRE
INFORMAZIONE STATISTICA - Parliamone con l'ISTAT - Edizione 1988 - L. 12.000
CONOSCERE L'ITALIA - INTRODUCING ITALY - Edizione 1991 - Distribuzione gratuita
SOMMARIO DI STATISTICHE STORICHE - 1926-1985 - L. 35.000
ATLANTE STATISTICO ITALIANO 1988 - L. 50.000
COMUNI, COMUNITA' MONTANE, REGIONI AGRARIE AL 31 DICEMBRE 1988 - Edizione 1990 - L. 20.000
ELENCO DEI COMUNI AL 31 MAGGIO 1991 - Edizione 1991 - L. 15.000
STATISTICHE AMBIENTALI - Vol. 2, - Edizione 1991 - L. 22.000
POPOLAZIONE RESIDENTE E PRESENTE DEI COMUNI - Censimenti dal 1861 al 1981 - L. 14.000
SOMMARIO STORICO DI STATISTICHE SULLA POPOLAZIONE - Anni 1951-87 - L. 41.000
IMMAGINI DELLA SOCIETA' ITALIANA - Edizione 1988 - L. 30.000
SINTESI DELLA VITA SOCIALE ITALIANA - Edizione 1990 - L. 15.000
CENSIMENTO DEGLI IMPIANTI SPORTIVI 1989 - Edizione 1991
Volume 1 - Italia L. 22.000
Volume 2 - Fascicoli regionali L. 12.000
MORTALITA' PER CAUSA E UNITA' SANITARIA LOCALE - Anni 1980-82 - L. 35.000
ELEZIONI DELLA CAMERA DEI DEPUTATI E DEL SENATO DELLA REPUBBLICA, 14 giugno 1987 - L. 10.000
45 ANNI DI ELEZIONI IN ITALIA 1946-90 - Edizione 1990 - L. 20.000
IL VALORE DELLA LIRA DAL 1861 al 1982 - L. 5.000
STATISTICHE SULLA AMMINISTRAZIONE PUBBLICA - Anni 1985-87 - L. 21.000
CONTI ECONOMICI REGIONALI - Anno 1988 - Edizione 1991 - L. 3.700

Stampato a Soveria Mannelli dalle Arti Grafiche Rubbettino
con i tipi fotocomposti della System Data Computer Srl - Roma
Novembre 1991