istat
working
papers

N .4
2014

Sperimentazione di un sistema di controllo e correzione per la codifica dell'attività economica

Francesca della Ratta-Rinaldi, Mauro Tibaldi

Istat
Istituto Nazionale
di Statistica

istat working papers

N .4
2014

Sperimentazione di un sistema di controllo e correzione per la codifica dell'attività economica

Francesca della Ratta-Rinaldi, Mauro Tibaldi

Istat

Istat Working Papers

Sperimentazione di un sistema di controllo
e correzione per la codifica
dell'attività economica

N. 4/2014

ISBN 978-88-458-1802-8

Sperimentazione di un sistema di controllo e correzione per la codifica dell'attività economica*

Francesca della Ratta-Rinaldi, Mauro Tibaldi

Sommario

Per migliorare la qualità dei dati il Servizio Istruzione, formazione e lavoro dell'Istat ha messo in campo una procedura di Text Mining per il controllo e la correzione della codifica dell'attività economica effettuata dai rilevatori nell'indagine sulle Forze di lavoro. La strategia è resa possibile grazie alla presenza nel file dati di un campo testuale in cui i rilevatori riportano le caratteristiche dell'attività economica svolta, così come descritta dagli intervistati, accompagnato dalla codifica relativa all'attività economica. La procedura, attiva dal quarto trimestre 2011, si basa sul confronto tra il vocabolario adottato dai rispondenti e il vocabolario specifico di ciascuna divisione, desunto dalla classificazione ufficiale (classificazione Ateco 2007, Nace rev. 2). I record nelle cui descrizioni compaiono parole non presenti nel vocabolario ufficiale della divisione specifica sono automaticamente estratti e controllati. In caso di errori ricorrenti, le istruzioni di ricerca sono eseguite automaticamente nelle future sessioni di correzione. Il processo, che ha permesso di migliorare la qualità dei prodotti, si completa attraverso una verifica puntuale della coerenza degli stessi che, in ogni sessione, prevede sempre un intervento manuale di esperti per la validazione delle correzioni effettuate e la definitiva attribuzione del codice corretto.

Parole chiave: analisi statistica dei dati testuali; qualità dei dati; controllo dei dati; classificazioni statistiche; Indagine sulle forze di lavoro.

Abstract

In order to improve data quality, Istat Division "Education, Training and Labour" experienced a text mining procedure for editing and imputation of the economic activity coding carried out by interviewers in the Labour Force Survey. This strategy is made possible by the presence of a text field in the data file in which the interviewers report economic activity features and the occupation as described by respondents. The procedure has been operating since the fourth quarter of 2011 and is based on a comparison between respondent vocabulary and the specific dictionary of each official classification division (Ateco 2007 classification, originated from Nace rev. 2). Records containing words used by the interviewers and not reported in the official dictionary are automatically extracted and verified. In the case of recurring errors, queries can be saved to be used in future editing and imputation sessions. The process is completed by a thorough examination of data consistency in each session, to validate the editing and imputation work done and for the final assignment of proper codes.

Keywords: text mining; data quality; data checking; statistical classifications; Labour Force Survey.

* Il paper è frutto di un lavoro comune. In particolare Francesca della Ratta ha redatto i paragrafi 1, 6 e 7, Mauro Tibaldi la Premessa e i paragrafi 2, 3 e 4.

Indice

Premessa

Il processo delle indagini della statistica ufficiale è continuamente investito da interventi mirati a migliorare l'accuratezza del dato, specie nel campo del contenimento degli errori non campionari che, come noto, sono legati alle procedure di misurazione e possono insorgere ad ogni passo del processo di produzione delle informazioni (Istat, 2011). In particolare, gli errori di misurazione assumono un grande rilievo in quanto l'informazione è disponibile ma non è corretta. Pertanto, la fase dei "controlli a caldo" eseguita nel corso del trattamento dei dati statistici può essere intesa come un insieme di azioni predisposte per individuare gli errori che insorgono durante il processo di produzione[1].

In questo quadro, a partire dal 2011 il Servizio Istruzione, formazione e lavoro dell'Istat ha messo in campo una procedura di Text Mining per il controllo e correzione della codifica dell'attività economica effettuata dai rilevatori nell'indagine sulle Forze di lavoro.

Questo sistema è stato progettato in seguito all'introduzione della nuova classificazione delle attività economiche Ateco 2007 (Nace Rev. 2) nella rilevazione sulle forze di lavoro (RFL). Considerata la tempestività con cui devono essere prodotti e diffusi ogni trimestre i dati definitivi dell'indagine (60 giorni dalla conclusione del trimestre) e l'elevata numerosità campionaria (circa 150 mila record individuali a trimestre), per garantire la sostenibilità del sistema di controllo si è ritenuto opportuno implementare una procedura che, sulla base di analisi preventive, procedesse a controlli mirati e selettivi compatibili con i tempi di diffusione dei dati trimestrali. Nella fase preliminare alla sperimentazione, l'analisi dei dati grezzi sulle coerenze tra le dinamiche occupazionali dei diversi settori di attività rispetto alla congiuntura economica, a cui si sono aggiunti controlli a campione sulle codifiche, ha permesso di individuare nell'Agricoltura e nella branca dell'Amministrazione pubblica, difesa e assicurazione sociale (d'ora in avanti denominata servizi generali della PA), i settori su cui procedere con il nuovo sistema di controllo. Pertanto, sono stati approntati controlli finalizzati a verificare la congruenza tra la codifica Ateco inserita dal rilevatore e la descrizione libera dell'attività economica fornita dall'intervistato nel corso dell'intervista. In tal modo è stato possibile individuare alcune incongruenze sulle quali si è deciso di intervenire in sede di correzione del dato. Esemplificativa la casistica che si riscontra nei servizi generali della PA, all'interno della quale spesso i rilevatori fanno confluire attività riconducibili invece alle divisioni della sanità (Asl, ospedali, ecc.), dell'istruzione (scuole di diverso ordine e grado), dello smaltimento dei rifiuti o dei trasporti. L'errata assegnazione di questi record determina un incremento occupazionale del comparto che non risponde né alle informazioni effettivamente rilevate, né alle dinamiche occupazionali che emergono da altre fonti. Dopo alcuni test di verifica, il controllo è stato esteso ad altre sezioni di attività economica (costruzioni, istruzione, sanità e assistenza sociale), nelle quali sono state individuate incongruenze ricorrenti.

Appurata l'esistenza di un certo quantitativo di record mal codificati dai rilevatori, per superare la gravosità connessa alla fase di revisione manuale si è deciso di mettere in piedi una sperimentazione finalizzata all'individuazione e alla correzione degli errori più rilevanti, valorizzando al massimo l'informazione contenuta nella stringa testuale utilizzata per la descrizione dell'attività economica dell'ente/azienda presso cui l'intervistato svolge l'attività lavorativa e, contestualmente,

[1] Nel caso della codifica della professione, peraltro, da tempo è stato realizzato un lavoro di controllo a campione della qualità delle codifiche dei rilevatori, finalizzato a migliorare le loro *performances* e a fornire elementi utili per la formazione e i *debriefing* (in merito si veda Gallo e Scalisi, 2012).

della professione esercitata. In parallelo, sono state definite alcune incongruenze tra il codice della professione e l'attività economica, utili per individuare alcuni errori specifici, specie nel settore dell'agricoltura e nell'istruzione. Nel caso delle incongruenze più stringenti è stato possibile definire alcune regole soft inserite nel questionario elettronico per ridurre i margini di errore.

A partire dal IV trimestre 2011, è stata quindi messa a punto una procedura di controllo, in parte automatizzata, basata sulla congruenza tra il linguaggio del dizionario ufficiale Ateco e il linguaggio dei rispondenti, finalizzata all'individuazione e alla correzione dei record mal classificati. Le caratteristiche della procedura e l'entità delle correzioni effettuate sono descritte nei paragrafi successivi.

Inoltre, per migliorare l'accuratezza delle stringhe descrittive inserite dai rilevatori (a prescindere dall'esattezza o meno della codifica) dal settembre del 2012 è stata avviata un'operazione di monitoraggio continuo, che consiste nell'invio quotidiano alla ditta che gestisce la rete CAPI di un file contenente le interviste che contengono descrizioni imprecise da trasmettere tempestivamente ai rilevatori interessati.

Al tempo stesso si è intervenuti in più occasioni in sede formativa con *debriefing* presso i rilevatori (nel luglio 2011, a marzo e dicembre del 2012, a marzo e dicembre del 2013) e presso i responsabili di *field*.

Parallelamente, nel corso del 2012 sono state compiute alcune operazioni di controllo applicando il software di codifica automatica ACTR sui record relativi all'industria in senso stretto e al commercio, divisioni che presentano tassi di errore leggermente inferiori a quelli inizialmente riscontrati nelle sezioni di attività economica per le quali si procede a correzione. Nel 2013 è stato inoltre eseguito un controllo manuale su un campione di 3.000 record rappresentativo delle 21 sezioni della classificazione Ateco, che ha fatto emergere altri settori critici.

In questo documento sono presentate le metodologie operative adottate e i tipi di errore più ricorrenti, insieme alla valutazione degli esiti della sperimentazione e dei possibili sviluppi della procedura di correzione. È utile ricordare che la sperimentazione condotta è stata realizzata grazie alla collaborazione di personale con competenze differenti e trasversali a diverse unità, costituendo probabilmente anche solo per questo un esempio di buona prassi organizzativa[2] finalizzata al processo di continuo miglioramento della qualità del dato statistico.

1. La procedura di controllo "Dizionari"

Le analisi a campione condotte durante il 2011, l'esperienza maturata nel corso del monitoraggio sulla qualità dei dati e la prima fase di correzione manuale hanno mostrato che i problemi di errata classificazione riguardano in particolare cinque divisioni: agricoltura, costruzioni, istruzione, sanità e assistenza sociale e soprattutto i servizi generali della PA. Dal IV trimestre 2011, pertanto, si è provveduto a implementare una procedura semi automatica per l'individuazione e la correzione degli errori (tuttora in corso), a partire dalla stringa testuale inserita dai rilevatori per la descrizione

[2] Si ringraziano in particolar modo Angela Ferrillo, per la consulenza sull'elaborazione dei dati con ACTR, Filomena de Filippo, per la costanza con cui ha analizzato i risultati di ACTR, Antonio Discenza, Carlo Lucarelli, Miriam de Santis, Antonella Iorio, Alessandra Lugli e Teresa Lizi per l'indispensabile apporto in sede di implementazione delle correzioni, Francesca Gallo e Pietro Scalisi per il confronto serrato sulle diverse opzioni di correzione dei dati, Rita Ranaldi, Emanuela Vergura e Rita Lima per il supporto nelle attività formative, Gianlorenzo Bagatta, Mario Albisinni, Silvia Loriga, Alessandro Martini, Federica Pintaldi, Maria Elena Pontecorvo e Vincenzo Triolo per il confronto critico nel corso della sperimentazione.

sia dell'attività economica dell'ente/azienda presso cui viene prestata l'attività lavorativa, sia della professione svolta. Nonostante l'oggetto della sperimentazione fosse limitato soltanto alla validazione della codifica dell'attività economica, si è ritenuto indispensabile considerare congiuntamente i *corpora* costituiti dalle risposte alle due domande aperte presenti nel questionario, relative a professione e settore di attività economica[3]. Difatti, soprattutto in alcuni settori, come quelli che erogano servizi per la collettività, la descrizione della professione può contribuire in alcuni casi a individuare l'unità locale presso cui viene svolta l'attività economica e quindi a valorizzare in modo più appropriato la codifica Ateco[4]. Ad esempio le stringhe contenenti le informazioni "raccolta rifiuti (C11) comune (C15)", oppure "cuoco di cooperativa ristorazione (C11) scuola dell'infanzia (C15)" - unitamente alla presenza di informazioni su un'azienda o ente organizzata in più sedi - ci consentono di individuare errori di codifica difficilmente evidenziabili analizzando la sola informazione contenuta nella risposta sull'attività economica ("comune" o "scuola dell'infanzia"), di per sé non necessariamente errata, ma con un ampio margine di indeterminatezza perché eccessivamente generica (al riguardo si veda il paragrafo 2). In questi casi la descrizione della professione consente, pertanto, di raccogliere informazioni aggiuntive utili alla descrizione dell'attività economica. In sostanza, la lettura integrata di queste variabili permette di specificare il contesto di svolgimento dell'attività economica[5], rendendo di fatto più agevole l'individuazione della codifica più pertinente.

Le interviste codificate dai rilevatori in ciascuna delle cinque divisioni sono analizzate in distinte sessioni di analisi, in modo da studiare esclusivamente il linguaggio utile a descrivere una specifica attività economica.

La procedura si basa sul confronto tra il vocabolario adottato dai rispondenti (X) e un vocabolario specifico, riferito nel nostro caso alla classificazione ufficiale dell'Ateco (Y). Tale vocabolario può essere facilmente ottenuto a partire dall'insieme delle denominazioni ufficiali della classificazione Ateco 2007, da cui sono state estrapolate le cinque divisioni in esame in modo da definire un primo lessico di riferimento per ciascuna sessione di analisi. Si tratta, peraltro, dello stesso vocabolario contenuto nel navigatore della classificazione che i rilevatori consultano nel corso dell'intervista per codificare le risposte degli intervistati.

L'intersezione tra i due insiemi determina tre blocchi logici di parole (Figura 1). Il sottoinsieme di vocaboli contenuti esclusivamente nel vocabolario Ateco (A) non entra nella procedura, in quanto contiene le parole riferite a settori specifici non presenti nel campione di interviste in esame o una terminologia specialistica non utilizzata dai rispondenti. In ogni caso, i termini presenti esclusivamente nel vocabolario dell'Ateco costituiscono un indicatore della distanza tra il linguaggio ufficiale della classificazione e quello degli intervistati (si veda il paragrafo 5). Il sottoinsieme di termini comuni ai due vocabolari (B) è quello meno problematico, in quanto è costituito da parole specifiche che caratterizzano con elevate probabilità il settore di riferimento, oltre che da quelle strumentali della lingua (articoli, preposizioni ecc.) che non sono di interesse per la presente analisi. Di particolare interesse è invece il sottoinsieme (C) composto dalle parole riportate dai rilevatori e non presenti nel vocabolario Ateco della divisione specifica.

[3] Si tratta delle domande C11 *"Può dirmi il nome della sua professione e in che cosa consiste il suo lavoro?"* e C15 *"Cosa fa l'Ente o l'azienda presso la quale lavora? (Indichi i principali beni e/o servizi prodotti).*

[4] In realtà si tratta di due classificazioni e due operazioni di codifica ben distinte per i rilevatori; nel caso di descrittivi troppo generici, tuttavia, è necessaria la loro integrazione nella fase di correzione per ridurre l'indeterminatezza del contesto produttivo, al fine di codificare con maggior accuratezza l'attività economica. In alcuni casi, ad esempio, la sede presso la quale l'intervistato lavora viene esplicitata nella descrizione della professione invece che nel descrittivo dell'attività economica. Al contrario, in presenza di campi descrittivi valorizzati in maniera appropriata il ricorso alla lettura integrata non è necessario.

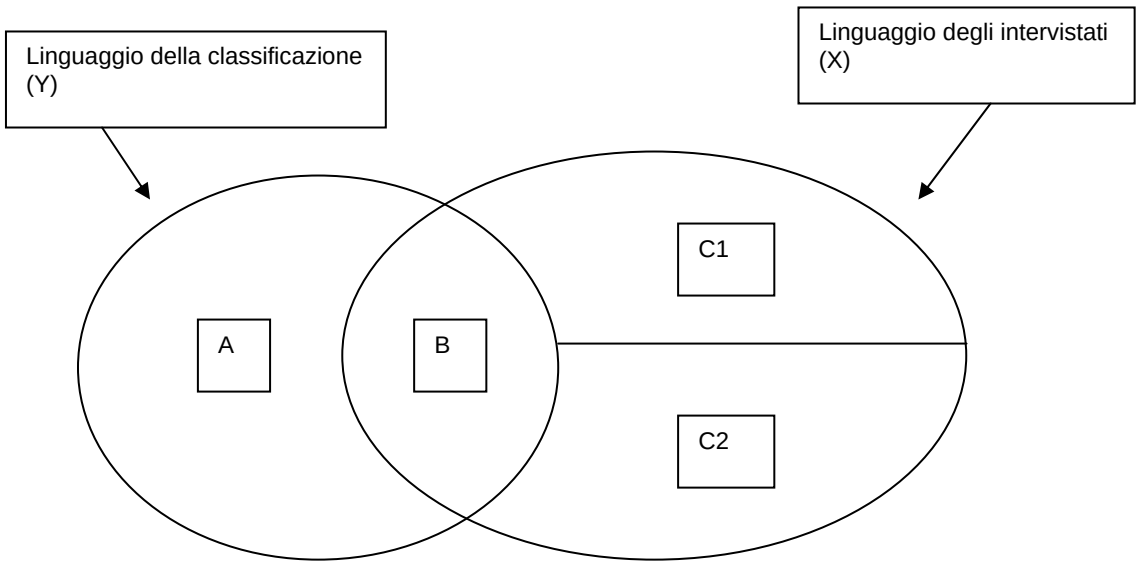
[5] Come già rilevato, a tale risultato concorre anche l'informazione contenuta nella domanda C14, ovvero se l'ente/azienda presso cui l'intervistato lavora ha una o più sedi.

Questo sottoinsieme può essere a sua volta scomposto in due blocchi: da un lato la terminologia "pertinente" (C1) che gli intervistati utilizzano per descrivere la propria attività pur con vocaboli non presenti nella classificazione ufficiale (ad esempio laddove la classificazione parla di 'coltivazione di agrumi' gli intervistati potrebbero parlare direttamente di 'limoni' o di 'arance'); dall'altro invece la terminologia "non pertinente" (C2) che probabilmente nasconde la presenza di errori. Ad esempio termini quali 'imbottigliamento', 'confezionamento', 'commercio' nelle interviste classificate dai rilevatori nel settore dell'agricoltura, rivelano attività che si possono più probabilmente ricondurre ad attività industriali o di commercializzazione.

I due sottoinsiemi di parole non presenti nel vocabolario Ateco (d'ora in poi denominati 'Unici') costituiscono il punto di partenza della procedura di individuazione degli errori: da un lato i termini da noi attribuiti al sottoinsieme C1 vengono aggiunti di volta in volta al vocabolario di partenza dell'Ateco, in modo da ridurre i tempi dell'analisi per le sessioni di controllo dei trimestri successivi (ampliando quindi progressivamente il sottoinsieme B); dall'altro l'insieme di termini del sottoinsieme C2 sono utilizzati per le procedure di controllo e correzione.

Il software di analisi testuale utilizzato per costruire i vocabolari e calcolare le intersezioni, Taltac2[6], consente di rintracciare in modo piuttosto semplice tutti i record che contengono una determinata parola o una combinazione di esse, visualizzando sia il contenuto del campo testuale sia l'insieme di variabili a esso associate in modo da consentire la valutazione dell'eventuale errore di codifica.

Figura 1 - Intersezione tra il linguaggio della classificazione e il linguaggio degli intervistati



Una volta appurato che la presenza di un determinato termine o combinazione di termini permette di individuare un errore di classificazione, è possibile indicare il codice Ateco corretto aggiungendolo al file di partenza in una nuova variabile personalizzata. In presenza di un errore ricor-

[6] Taltac2 è un software specifico per l'analisi lessicale e testuale messo a punto da Sergio Bolasco, Francesco Baiocchi e Adolfo Morrone (www.taltac.it). Si veda anche Bolasco (2013).

rente, che si presume possa ripresentarsi anche nella correzione del successivo file trimestrale, l'istruzione di ricerca, comprensiva dell'indicazione della codifica dell'attività economica corretta, può essere eseguita automaticamente nelle successive sessioni di correzione[7], in modo da individuare contemporaneamente sia l'errore sia la codifica pertinente.
In questo modo le sessioni di correzione successive alla prima "apprendono" dalle operazioni effettuate in precedenza, sia perché viene ampliato il vocabolario specifico per ciascuna divisione Ateco riducendo di volta in volta il numero di vocaboli del sottoinsieme C da controllare, sia perché gli errori di codifica più frequenti possono essere individuati e corretti in automatico. Dopo oltre un anno e mezzo dalla messa a punto della strategia di controllo, è stato dunque possibile semplificare ulteriormente - con una conseguente riduzione dei tempi - la fase di controllo testuale, grazie alla progressiva automatizzazione della stessa e alla possibilità di "apprendere" dalle sessioni di correzione precedenti. Difatti, da un lato le *query* che girano in automatico consentono di individuare circa il 60% dei record da correggere, dall'altro la progressiva riduzione del sottoinsieme di parole da controllare (insieme C2) rende piuttosto agevole etichettare direttamente tutti i record che contengono una parola anomala, controllando poi direttamente nel file finale l'effettiva esattezza della codifica proposta dal rilevatore. In pratica, per ciascuna sessione, con una semplice procedura di *tagging* semantico viene assegnata l'etichetta "VERIFICA" a tutte le parole anomale del sottoinsieme C2. Piuttosto che costruire *query* specifiche per ciascuna di queste parole - spesso presenti con frequenza molto ridotta nel *corpus* - la funzione di Ricerca Entità consente di assegnare automaticamente un'etichetta (ad esempio "VERIFICARE") a tutti i record che contengono una delle parole anomale presenti nella lista. Nel corso della successiva fase di verifica del risultato finale è possibile abbinare il codice Ateco corretto ai record che risultano effettivamente classificati in maniera erronea.

In ogni caso il processo di correzione si completa attraverso una verifica puntuale sulla coerenza dei dati; in ogni sessione, difatti, è sempre necessario un intervento manuale da parte di esperti per la validazione delle correzioni effettuate e per la definitiva attribuzione del codice Ateco a 6 digit.

2. Il problema a monte: la qualità delle stringhe descrittive

L'analisi delle stringhe testuali ha consentito di far emergere il problema della qualità del testo inserito dai rilevatori, considerato che descrizioni troppo generiche non consentono di codificare l'attività economica con il dettaglio richiesto dall'indagine sulle Forze di Lavoro (sesto digit per l'Ateco e quinto per le professioni). Per comprendere la portata di questo problema è necessario considerare le caratteristiche delle classificazioni in uso, che utilizzano variabili di tipo categoriale e sottostanno a criteri di ordinamento astratti che rendono piuttosto complessa l'attività di codifica (Vicari, Ferrillo, Valery, 2009). Mentre la classificazione delle professioni, tra le due, appare relativamente meno problematica poiché richiede di descrivere un insieme di attività lavorative concretamente svolte da ciascun individuo, la classificazione Ateco è di più complesso utilizzo. Quest'ultima, infatti, è una classificazione piatta, non ordinabile e molto analitica (prevede in tutto 1.224 sottocategorie) che rimanda a un livello più astratto, ovvero l'attività svolta dall'ente/impresa presso cui il soggetto presta la propria opera, attività che non sempre gli intervistati conoscono in

[7] Si tratta di una funzione particolarmente innovativa presente in Taltac2, la Ricerca entità (RE), che consente di considerare come unità di analisi non solo la singola parola presente in un testo, come avviene nell'analisi lessicale classica, ma anche l'intero record in cui questa compare. Attraverso *query* specifiche, che si avvalgono di operatori logici e booleani e operatori di distanza tra parole, è inoltre possibile individuare anche combinazioni di parole (ad. es. 'asilo nido' o 'cooperativa di ristorazione') o insiemi di parole accomunate da una specifica annotazione in uno dei campi del vocabolario, ad esempio la categoria semantica. Sui record che rispondono positivamente alle *query* è possibile effettuare operazioni di etichettatura o conteggio, assegnando una nuova variabile al data set di partenza (nel nostro caso Ateco corretta). Per ulteriori dettagli si veda della Ratta (2010a) e (2009).

maniera sufficiente. Un contributo fondamentale al processo di codifica è pertanto fornito dal rilevatore, che attraverso la propria esperienza e capacità relazionali dovrebbe tendere a ottenere il maggior numero di informazioni riguardo le variabili indagate, in modo da rendere il più possibile pertinente ed esaustiva la codifica.

Sono invece stati individuati numerosi casi in cui nelle stringhe testuali inserite mancava la descrizione del tipo di attività, oppure non era evidenziato il bene o il servizio oggetto di tale attività, o ancora era assente il materiale del bene prodotto. Per tali ragioni le descrizioni fornite risultano troppo generiche e non consentono né di costruire una codifica pertinente né di controllare l'esattezza del codice Ateco inserito dal rilevatore. Un altro problema su cui si è deciso di intervenire riguarda la copiatura delle stringhe descrittive presenti nel navigatore, utilizzato in fase di consultazione da parte dei rilevatori, nel campo destinato alla descrizione dell'attività economica.

Per limitare questi fenomeni, dal settembre del 2012 è stata avviata un'operazione di monitoraggio continuo, che consiste nell'invio quotidiano di un file contenente proprio i casi che contemplano descrizioni troppo generiche o ricopiate dal motore di ricerca. Mentre nel caso delle professioni si è attinto a una lista costituita da 120 voci generiche costruita nel corso dei monitoraggi finora effettuati, nel caso dell'Ateco le descrizioni generiche sono state individuate a partire da tutte le stringhe inserite dai rilevatori nei primi due trimestri del 2012 composte da meno di 13 caratteri. A partire da questo insieme è stato definita una lista di circa 800 descrittivi eccessivamente generici per una corretta codifica, cui è stato aggiunta l'indicazione dei motivi per cui tali descrizioni sono da ritenersi inadeguate o il suggerimento sul modo più appropriato di inserire le informazioni. Per incrementare l'efficienza di questo tipo di controllo e arricchirlo altresì di un contenuto formativo, da marzo 2013 l'invio è stato corredato da brevi spiegazioni ed esempi in merito agli errori che si verificano con maggior frequenza. Anche se una descrizione generica non comporta necessariamente un errore di codifica, si ritiene che richiamare l'attenzione dei rilevatori sull'importanza di una descrizione esaustiva possa migliorare l'accuratezza della codifica stessa.

Peraltro, un elemento critico che permane nell'insieme delle operazioni di descrizione e codifica, e che è stato già affrontato in modo più analitico nell'incontro di *debriefing* del dicembre 2013, è la mancata individuazione/esplicitazione dell'unità locale presso la quale l'intervistato lavora, fattore che genera spesso una descrizione dell'attività economica non pertinente e quindi un errore di codifica[8]. Le regole di attribuzione dell'attività economica, infatti, prevedono che se un ente/azienda dispone di più sedi/unità locali, deve essere rilevata l'attività principale della sede presso cui l'intervistato lavora, che può essere diversa da quella dell'impresa o ente madre. Il caso più frequente, come sottolineato, riguarda gli enti locali e in particolare i comuni, che erogano una pluralità di servizi ai cittadini cui sottostanno attività economiche differenti[9]. Nei comuni di una certa dimensione tali attività vengono esercitate in unità locali diverse, per cui il rilevatore dovrebbe raccogliere questo tipo di informazione in modo da descrivere l'attività economica dell'unità locale e codificarla correttamente.

Nei primi 9 mesi del 2013 sono state inviate circa 3.500 segnalazioni riferite alla descrizione dell'attività economica e 2.800 alla professione. Il 14% circa delle segnalazioni riguarda il proble-

[8] Mentre l'individuazione delle unità locali eventualmente presenti è agevole per le imprese, non lo è altrettanto quando sono gli individui a doverla effettuare.

[9] Rimane il fatto che nei comuni di piccole dimensioni, per motivi organizzativi e di opportunità, spesso i servizi e quindi le attività sono erogate dalla stessa sede del comune, in tal caso l'attività economica principale è unica e corrisponde a quella delle attività generali dell'amministrazione pubblica (codice 841110).

ma del descrittivo copiato e i restanti casi quello della descrizione generica[10]. Analizzando l'insieme delle codifiche generiche riferite all'Ateco è interessante osservare la loro concentrazione in alcuni settori, quali la sanità, i servizi generali della PA, le costruzioni, le attività dei servizi per la persona, l'agricoltura, la ristorazione, l'istruzione e il commercio. Nel corso dei mesi, tuttavia, la numerosità delle segnalazioni si è più che dimezzata, passando da 1.965 del primo trimestre a 949 del terzo trimestre 2013. La riduzione del volume delle segnalazioni costituisce un risultato molto positivo che mostra l'efficacia della strategia fin qui messa in atto e che verrà quindi rafforzata nei successivi incontri formativi.

3. Le correzioni effettuate e gli errori più ricorrenti

Nei primi nove trimestri in cui è stata utilizzata, la procedura di correzione si è andata di volta in volta affinando. In particolare, per evitare di intervenire più volte sullo stesso record, sono stati esclusi dal controllo i record già corretti in precedenza e riproposti nel file di intervista nelle *wave* successive alla prima. Si ricorda, infatti, che nelle re-interviste svolte con tecnica Cati, le informazioni relative ad Ateco e professione sono nuovamente rilevate soltanto nei casi in cui nel periodo trascorso tra la prima e le successive tre interviste si sia modificato lo status occupazionale o il tipo di lavoro svolto. Se non sono intervenuti cambiamenti, all'intervistato viene chiesto semplicemente di confermare le informazioni (testuali) dichiarate nel corso della prima intervista.

Le correzioni individuate nei dati relativi a uno specifico trimestre sono implementate automaticamente nel file dati definitivo. L'insieme delle correzioni indicate nella Tabella 1 è pertanto da riferirsi esclusivamente alle nuove correzioni da noi introdotte nei nove trimestri analizzati, con riferimento alle prime interviste, o alle interviste successive alla prima nel caso di cambiamenti intervenuti nella condizione occupazionale.

Tabella 1 - **Codifiche errate, record controllati e tasso di errore per divisione Ateco IV trimestre 2011-IV trimestre 2013**

Divisione	Codifiche errate	Record controllati	Errori in %
Agricoltura	335	22.828	1,5
Costruzioni	442	39.483	1,1
Servizi generali della PA	1.430	34.907	4,1
Istruzione	459	37.631	1,2
Sanità	444	41.801	1,1
Totale	**3.110**	**176.650**	**1,8**

Come evidenziato dalla Tabella 1, nel complesso, tra il IV trimestre 2011 e il IV trimestre 2013, sono stati corretti 3.110 record, pari all'1,8% dei circa 177 mila controllati. Le divisioni più critiche risultano i servizi generali della PA (con il 4,1% di errori), l'agricoltura (1,5%) e l'istruzione (1,2%)[11].

[10] Le elaborazioni sulle segnalazioni automatiche sono state realizzate da Rita Lima.

[11] La possibilità di confermare nelle successive interviste realizzate con tecnica CATI le informazioni su attività economica e professione rilevate nel corso della prima intervista CAPI rende problematico distinguere il tasso di errore per tecnica.

L'analisi puntuale delle correzioni effettuate consente di sintetizzare gli errori più ricorrenti per ciascuna divisione, individuando in tal modo una casistica che costituisce la base per la progettazione di attività formative per i successivi *debriefing* presso i rilevatori.

Agricoltura. Gli errori di codifica in questo settore possono generalmente ricondursi alla valutazione del prodotto e non del processo nel quale il prodotto è inserito. Si tratta di attività da ricondurre alla trasformazione industriale di prodotti agricoli (industria alimentare, macellazione carni, trasformazione, imballaggio e confezionamento) o alla loro commercializzazione (vendita di frutta, piante, mercato ortofrutticolo, ecc.). Inoltre, sono erroneamente classificate in agricoltura (branca della silvicoltura) anche le attività della Pubblica amministrazione riguardanti il controllo del patrimonio forestale (Corpo Forestale dello Stato) e le attività di servizi per il paesaggio come la manutenzione del verde e le attività di giardinaggio.

Costruzioni. Circa la metà dei record mal classificati sono in realtà da attribuire ad attività manifatturiere, poiché riguardano soprattutto la produzione industriale di materiali utilizzati successivamente in edilizia (ad esempio pavimenti, prefabbricati, tettoie). Un altro tipo di errore riguarda l'errata codifica in questa sezione di attività di vendita di prodotti per l'edilizia, da attribuire invece al commercio. Vi sono poi attività da codificare nella gestione di reti idriche (gestione di acquedotto), nelle attività immobiliari e nelle attività professionali tecniche (architetti).

Servizi generali della Pubblica amministrazione. Come già osservato, si tratta della divisione nella quale si riscontra il maggior tasso di errore. L'errore più frequente riguarda le codifiche da attribuire al settore della sanità. Si tratta di personale sanitario o amministrativo (medici, infermieri, impiegati, operatori) che lavora in strutture sanitarie, in prevalenza Asl, erroneamente classificato nei servizi generali della Pa (divisione 84) invece che nelle corrispondenti Ateco della divisione sanità (86). Altro errore ricorrente è quello degli asili nido (comunali o privati), erroneamente codificati dai rilevatori nella Pubblica amministrazione o nell'istruzione anziché nell'assistenza sociale. Un tipo di errore abbastanza frequente, che coinvolge soprattutto coloro che si dichiarano dipendenti dai comuni, riguarda la codifica delle attività connesse alla gestione dei rifiuti e delle reti idriche, al trasporto pubblico locale o alla gestione delle biblioteche, da codificare rispettivamente come attività dell'industria (divisioni 38 e 36), dei trasporti (divisione 49) e delle biblioteche (divisione 91) in presenza di più sedi dell'ente. Sempre in ambito comunale vi sono poi le attività relative alla manutenzione del verde e dei cimiteri (divisione 81), mentre gli enti pubblici di ricerca andrebbero classificati nell'attività a loro più pertinente (divisione 72).

Istruzione. In questa divisione l'errore più frequente riguarda la codifica degli asili nido, erroneamente attribuiti all'istruzione, dove erano collocati nella vecchia classificazione Ateco 2002. Vi sono poi le attività riferibili alle mense o ai servizi di pulizia che, quando non sono esercitate direttamente dall'ente (scuola o comune), dovrebbero essere codificate come attività di ristorazione (divisione 56) e dei servizi di pulizia (divisione 81). Frequenti anche gli errori riguardanti le attività connesse alla ricerca (divisione 72), specie se riferita agli enti pubblici.

Sanità. Gli errori più frequenti riguardano le attività di ristorazione erogate per gli ospedali o strutture sanitarie che, qualora non svolte direttamente dagli stessi enti, vanno classificate come attività di ristorazione (divisione 56), insieme alle attività di pulizia svolte da cooperative o altre società, che vanno classificate come attività di pulizia e disinfestazione (divisione 81)[12]. Vi sono poi

[12] L'esternalizzazione di alcuni servizi, in particolare la somministrazione di pasti e le pulizie, che si è progressivamente realizzata nella scuola e nella sanità rende più difficoltosa la codifica dell'attività economica, giacché la mancata esplicitazione del soggetto che fornisce tali servizi può creare un ampio margine di incertezza e quindi di errore. Se tali servizi vengono prodotti all'interno delle stesse istituzioni tramite proprio

attività da attribuire all'istruzione perché riferite alla scuola dell'infanzia (soltanto gli asili nido devono essere infatti codificati nelle attività assistenziali, mentre se il nido è inserito in un istituto comprensivo che comprende la scuola dell'infanzia la codifica resta nell'istruzione), e altre riferite ai servizi alle persone (massaggi e centri estetici da attribuire alla divisione 96).

4. Il ruolo della formazione

L'analisi degli errori ricorrenti e l'introduzione delle azioni di monitoraggio sulla qualità delle stringhe descrittive costituiscono un ulteriore tassello di un processo rivolto a migliorare la qualità dei dati, la cui architettura poggia su periodici ritorni formativi presso i rilevatori sia CATI sia CAPI. La formazione, infatti, svolge un ruolo fondamentale perché consente di migliorare ex-ante la qualità dei dati prodotti attraverso un'opera di prevenzione, finalizzata a ridurre in maniera significativa l'insorgere di errori. Per dispiegare appieno i suoi effetti, tuttavia, deve essere progettata e programmata con particolare attenzione orientandola verso un flusso che si dipana nel tempo, piuttosto che a un singolo evento isolato. Per tale ragione sono stati programmati dei ritorni formativi presso i rilevatori, con l'obiettivo di consolidare e incrementare le conoscenze utili a supportare una corretta codifica dell'attività economica.

In tali occasioni, la discussione degli errori più frequenti rappresenta il mezzo per rimettere a fuoco alcuni punti chiave inerenti la valorizzazione esauriente del descrittivo relativo all'Ateco. In particolare, l'identificazione dell'attività e dei beni/servizi prodotti, l'individuazione di eventuali unità locali dell'ente/impresa, i casi nei quali un'impresa svolga l'attività economica nella sede di un'altra impresa o ente, le regole di prevalenza nel caso di più attività svolte all'interno di uno stesso ente o impresa. Si tratta, in sostanza, di tutti gli elementi che possono generare incongruenze tra descrizione dell'attività economica e codifica della stessa da parte del rilevatore.

E' opportuno rilevare che nella rete di rilevazione CAPI la qualità del processo di codifica non è del tutto soddisfacente. Poiché gran parte delle prime interviste viene effettuata da questa rete, si può innescare un effetto "moltiplicatore" per cui la bassa qualità delle codifiche si riflette sulle re-interviste, dove il processo di correzione è più difficoltoso da attuare.

Tra i fattori che contribuiscono a tale discrasia è da segnalare la tipologia di ritorno formativo, che nel caso della rete CAPI avviene con minor frequenza e generalmente a distanza tramite teleconferenza, uno strumento che non facilita l'interazione personale. Inoltre è utile ricordare che nel processo d'intervista della RFL, per ridurre i tempi di intervista è possibile lasciare la codifica in sospeso e procedere all'individuazione dei codici Ateco e professione a intervista conclusa. Mentre la codifica effettuata nel corso dell'intervista consente di approfondire con l'intervistato il tipo di attività economica o la professione, la codifica ex-post è interamente affidata al rilevatore e si basa esclusivamente sulla qualità e l'esaustività descrittiva della stringa testuale. Se da un lato gli intervistatori CATI possono confrontarsi con i colleghi e i supervisori di sala nel caso di codifiche complesse, dall'altro questa possibilità non è realisticamente praticabile per i rilevatori CAPI. Per tale ragione, le nuove procedure di segnalazione automatica sopra descritte sono state ideate per la rete CAPI, con l'obiettivo di intensificare l'invio di informazioni personalizzate a rilevatori dislocati su

personale, infatti, si tratterebbe di attività ausiliare di supporto a quella principale, risultando quindi corretta una classificazione che li attribuisse a una branca dell'istruzione o della sanità. Viceversa se fossero offerti da una società esterna, andrebbero classificati nelle rispettive sezioni di appartenenza. Per tale motivo si è più volte raccomandato ai rilevatori una valorizzazione esaustiva delle stringhe descrittive.

tutto il territorio nazionale. Alla luce di quanto esposto, nel dicembre 2013 è stato realizzato un intervento formativo frontale con tutti i rilevatori della rete CAPI, a tre anni di distanza dall'ultimo intervento di formazione diretto dedicato all'introduzione della nuova classificazione. Questo evento ha consentito, attraverso un'interazione diretta, di comprendere le dinamiche reali del processo di codifica, i problemi sottostanti e le difficoltà incontrate in particolari contesti territoriali o economici. Tra gli altri, è emerso da parte della maggioranza dei rilevatori un atteggiamento di quasi totale delega nei confronti del navigatore, considerato come unico strumento deputato alla codifica mediante la mera digitazione dei descrittivi rilevati, che esclude invece il ricorso alle regole e alle conoscenze che presiedono al processo stesso di codifica.

L'intervento si è sviluppato da una parte sull'approfondimento dei contenuti e della logica della classificazione dell'attività economica e sulle regole che presiedono all'individuazione, descrizione e codifica dell'Ateco, dall'altra su simulazioni di interviste ed esercitazioni pratiche di codifica accompagnate da sessioni di discussione e verifica dei risultati.

Vista l'indubbia efficacia emersa da questo tipo d'intervento, si sta progettando di replicare la formazione frontale con tutti i rilevatori almeno a cadenza biennale. Nel frattempo, è stata avviata la somministrazione di esercitazioni di codifica per via telematica a livello trimestrale, proprio per consolidare le conoscenze acquisite nell'ultimo incontro formativo.

Un altro aspetto critico sul quale si è ulteriormente intervenuti è costituito dall'interazione fra rilevatori e supervisori, che probabilmente rappresentava l'anello debole della catena formativa. Ciò è avvenuto da un lato coinvolgendo attivamente i supervisori della rete di rilevazione come destinatari del processo formativo, grazie a un incontro svoltosi a marzo 2013, dall'altro incentivando il flusso comunicazionale tra i rilevatori e questi ultimi che, con la preparazione acquisita, possono fungere da "filtro" per ridurre le impurità contenute nelle codifiche.

Quanto esposto finora non deve far prefigurare una sorta di "abbandono" della rete CATI. Anche su questo versante, infatti, si sta progettando il coinvolgimento dei supervisori di sala in un ritorno formativo specifico, mentre a gennaio 2014 è stata effettuata un'esercitazione di codifica per via telematica per i rilevatori di questa rete.

5. Linguaggio degli intervistati e linguaggio della classificazione[13]

Il confronto del linguaggio degli intervistati con quello della classificazione ufficiale ha dunque evidenziato un limite dello strumento di codifica, che utilizza un vocabolario piuttosto astratto e lontano dal modo di esprimersi degli intervistati (della Ratta, Gallo, Loré, 2011). Nel corso della loro attività di codifica i rilevatori hanno a disposizione un motore di ricerca della classificazione (navigatore) che consente, attraverso l'inserimento di stringhe di testo, di individuare la codifica più appropriata nella classificazione Ateco. Se la parola menzionata dall'intervistato non è presente nel dizionario ufficiale, il rilevatore dovrà ricorrere esclusivamente alle sue nozioni sulla classificazione per individuare il codice giusto, esponendosi a maggiori rischi di errore. Al contrario una classificazione arricchita dalla terminologia impiegata dai rispondenti potrebbe facilitare il lavoro di codifica riducendo gli errori.

[13] Il paragrafo è stato redatto da Maria Elena Pontecorvo.

Per valutare la distanza tra i due linguaggi si è proceduto all'intersezione tra i due vocabolari (vedi Figura 1), quello della classificazione ufficiale e quello riguardante le risposte aperte inserite dai rilevatori in uno specifico trimestre (il II del 2012) relative alla sola codifica Ateco, circoscrivendo l'analisi alle cinque sezioni soggette a procedura di controllo[14].

La misura più utile a valutare il grado di indipendenza lessicale tra i due linguaggi è l'indice di connessione lessicale (Cortelazzo, Tuzzi 2008). Nel nostro caso l'indice di indipendenza del linguaggio della classificazione (Y) rispetto al linguaggio degli intervistati (X) è dato dal rapporto tra le forme di Y non presenti in X (V_A) rispetto al totale delle forme di Y (V_A+V_B). La maggiore o minore dipendenza tra i due vocabolari fornisce un'indicazione sul grado di utilità della classificazione ufficiale per codificare l'attività economica. Se si guarda al valore dell'indice (Tab. 2) si nota che le sezioni denotate da maggiore indipendenza[15] sono proprio l'agricoltura e l'istruzione, due settori nei quali il tasso di errore è piuttosto elevato. In questi settori l'indice di connessione lessicale è pari rispettivamente a 0,54 e 0,47. La relazione tra errore e indipendenza è invece meno evidente nel settore affetto dal maggior livello di errore, i servizi generali della PA, in cui l'indice assume uno dei valori più bassi (0,34). Probabilmente in questo settore gli errori di classificazione sono da ricondurre non tanto all'inadeguatezza dello strumento di codifica, quanto piuttosto alla notevole complessità del comparto, che ha fatto emergere un evidente deficit concettuale dei rilevatori, da colmare in sede formativa. Peraltro, è proprio in questo settore che la descrizione della professione (non considerata nell'esperimento di confronto in quanto non presente nel dizionario Ateco) consente di individuare molti errori di codifica. Ad esempio, se l'intervistato dichiara di svolgere la professione di "autista di mezzi pubblici" presso il comune, è proprio la descrizione della professione che consente di comprendere che si tratta di attività da classificare nei servizi di trasporto di passeggeri e non nei servizi generali della PA, mentre la semplice stringa "Comune" è soltanto generica ma in sé non errata.

Tabella 2 - Linguaggio degli intervistati e della classificazione. Connessione lessicale e distanza intertestuale - II trimestre 2012

Settore di attività economica	Vocabolario intervistati (X)		Dizionario classificazione (Y)		Parole presenti solo in Y (V_A)	Forme comuni $V_{(B)}$	Connessione lessicale (V_A/V_A+V_B)	Distanza Labbé
	Occorrenze (N_X)	Forme grafiche (V_X)	Occorrenze (N_Y)	Forme grafiche (V_Y)				
Agricoltura	7.942	740	2.259	460	247	213	0,54	0,38
Costruzioni	13.558	1.227	1.723	363	138	225	0,38	0,36
Istruzione	9.508	728	779	199	94	105	0,47	0,41
Servizi generali PA	7.710	856	717	182	61	121	0,34	0,47
Sanità	11.734	1.016	707	235	78	157	0,33	0,35

Oltre all'indice di connessione lessicale è stata utilizzata un'ulteriore misura di similarità, la distanza intertestuale di Labbè (Labbè & Labbè 2003), che consente di valutare la similarità tra i corpus nella frequenza di impiego delle parole comuni, tenendo conto della differenza di ampiezza dei

[14] Per un esempio di confronto tra vocabolari provenienti dalle descrizioni fornite dagli intervistati e vocabolari generati da liste precodificate, si veda anche Tuzzi e Zaccarin (2004).

[15] L'indice raggiunge il valore zero quando i due testi sono uguali (cioè perfettamente «dipendenti» uno dall'altro).

due linguaggi a confronto (Cortelazzo, Tuzzi, Nadalutti 2012)[16]. In questo caso le maggiori distanze si osservano nelle divisioni della PA e dell'istruzione, che sono gli stessi settori caratterizzati da una più elevata percentuale di errori di codifica, con valori dell'indice pari a 0,47 e 0,41. Il settore dell'istruzione sembra dunque affetto sia da una maggiore indipendenza tra linguaggio degli intervistati e classificazione, sia da una maggiore differenza nella frequenza con cui anche le parole comuni sono utilizzate, mentre quello della PA, che pure aveva un grado di dipendenza maggiore tra vocabolario degli intervistati e della classificazione, mostra una distanza più elevata in termini di utilizzo delle forme grafiche comuni.

Per approfondire ulteriormente la natura della distanza tra i due linguaggi, è utile analizzare anche per ciascun settore i termini presenti solo nel linguaggio degli intervistati. La Tabella 3 riporta in ordine di occorrenza le prime dieci forme grafiche pertinenti[17] per ciascun settore, il cui lemma non è presente nel dizionario della classificazione.

Ad esempio, in agricoltura troviamo forme grafiche che fanno riferimento a specifici oggetti dell'attività agricola (vigneto/i, campo, viti, oliveto ecc.) che non sono compresi nel vocabolario Ateco. Nelle costruzioni spiccano parole inerenti il tipo di sede (impresa, ditta, azienda), e vocaboli comuni che connotano l'attività edilizia (ristrutturazione, abitazioni, case, interni). Nel settore dell'istruzione gli intervistati utilizzano molto gli aggettivi che specificano il tipo di scuola utilizzando il linguaggio comune al posto di quello istituzionale (materna, tecnico, alberghiero, industriale, agrario ecc), mentre nella PA le parole più diffuse specificano soprattutto il tipo di amministrazione (comunale, provinciale ecc.). Infine, in sanità le parole non presenti nel dizionario Ateco riguardano in particolar modo le Asl, come si evince sia dalla specifica forma grafica, sia dai termini "azienda" e "locale", oppure l'attività svolte dalle cooperative e le cure dentistiche.

Tabella 3 - Linguaggio degli intervistati: lemmi non presenti nel dizionario Ateco per settore e numero di occorrenze. II trimestre 2012

Agricoltura		Costruzioni		Istruzione		PA		Sanità	
F.G.	Occ	F.G.	Occ.	F.G.	Occ.	F.G.	Occ.	F.G.	Occ.
vigneto/i	62	impresa	450	materna	312	comunale	160	azienda	191
campo	29	ristrutturazione/i	386	tecnico	172	esercito	89	generale	152
viti	23	ditta	126	inferiore	83	agenzia	88	cooperativa	138
coltiva	23	abitazioni	100	pubblica	59	entrate	61	centro	123
oliveto	22	residenziali	73	statale	45	ente	54	asl	114
olio	19	case	64	comprensivo	44	provinciale	49	dentistico	99
polli	15	termoidraulici	50	alberghiero	44	marina	42	locale	99
agricoltore	15	azienda	49	privata	40	inps	40	privata	58
fieno	12	condizionamento	36	industriale	39	ufficio	38	civile	52
cooperativa	12	interni	35	agrario	20	aeronautica	32	soccorso	21

[16] Dati i due corpus Y e X con rispettive numerosità $N_Y<N_X$ la distanza intertestuale si calcola sulle parole comuni ai due corpus secondo la formula:

$d(Y,X)=\frac{\sum_{i\in Vy\cup x}|f_{iY}-f^{*}_{iX}|}{2N_y}$ dove f_{iY} sono le frequenze delle parole del corpus più piccolo Y e f^{*}_{iX} sono pari a f_{iX}*Ny/Nx, ovvero le frequenze del corpus più grande X "ridotte" in ragione della dimensione del testo più piccolo Y. L'indice varia tra 0 e 1, dove 0 indica il massimo della similarità tra i due corpus.

[17] Come visto nel paragrafo 1, tra i termini presenti esclusivamente nel linguaggio degli intervistati incontriamo sia la terminologia non pertinente indizio di errore (il sottoinsieme C2) sia termini pertinenti (C1) che specificano ulteriormente l'attività economica. L'analisi condotta di seguito è naturalmente riferita solo a questi termini.

Questo tipo di risultato suggerisce l'opportunità di arricchire il vocabolario attualmente in uso dai rilevatori: considerati i legami tra il tasso di errore e la distanza tra i linguaggi, una maggiore aderenza del linguaggio utilizzato dallo strumento di codifica non potrà che facilitare tutto il processo. Un'applicazione di questo esercizio estesa a tutti i settori di attività e costruita su una base di dati più consistente potrebbe costituire il punto di avvio per un processo più ampio, in cui il materiale testuale raccolto sul campo nelle indagini sulle famiglie possa contribuire ad arricchire il navigatore utilizzato per la codifica con esempi tratti dal modo concreto di esprimersi degli intervistati.

6. E gli altri settori? Un controllo con ACTR[18]

Al fine di ottimizzare l'automazione della procedura di individuazione e correzione degli errori è stata effettuata una sperimentazione utilizzando ACTR (*Automatic Coding by Text Recognition*), un sistema in uso nell'Istat dalla fine degli anni '90 che consente, nella sua versione aggiornata, di individuare le possibili codifiche per l'Ateco 2007 a fronte di una descrizione testuale. Il software utilizza un apposito algoritmo per misurare la similarità tra i testi presenti nel proprio dizionario e le descrizioni libere inserite dai rilevatori[19].

L'obiettivo del test era di valutare le potenzialità dello strumento, correntemente utilizzato dall'Istat per la codifica dell'Attività economica anche in indagini sulle famiglie, anche per le operazioni di controllo e correzione, in modo da velocizzarne i tempi di esecuzione ed estendere l'attività di correzione ad altre divisioni dell'Ateco.

Il test effettuato ha riguardato due settori finora esclusi dal processo di controllo, l'industria in senso stretto e il commercio, e uno invece già compreso, le costruzioni (inclusa in quanto condivide con il commercio il primo digit del codice Ateco). In base al livello di similarità riscontrato, i possibili abbinamenti del sistema ACTR si possono suddividere in unici, possibili, multipli e falliti. In particolare i record codificati come "unici" sono quelli di migliore qualità, quelli cioè per i quali ACTR individua un'unica possibile codifica dell'Ateco. L'incidenza dei record unici è piuttosto elevata, pari al 70,8% nel commercio e costruzioni e al 57,5% nell'industria in senso stretto. Di contro, ACTR non riesce in nessun modo a codificare il 9,6% dei record dell'industria e il 4,9% di quelli del commercio e delle costruzioni.

Passando alla corrispondenza tra la codifica indicata dal rilevatore e quella suggerita da ACTR, è stato invece riscontrato un elevato numero di discordanze al primo digit, vale a dire record che ACTR e rilevatore attribuiscono a divisioni differenti. Si tratta del 18,4% del totale dei record dell'industria in senso stretto e del 14,7% di quelli attribuiti al commercio e alle costruzioni, un risultato che all'apparenza sembra costituire un serio indizio di scarsa qualità delle codifiche. In realtà, a una verifica puntuale dei casi discordanti si è appurato che in circa 9 casi su 10 la codifica esatta è quella inserita dal rilevatore. Il software, infatti, non sembra possedere la flessibilità necessaria a interpretare le stringhe spesso eccessivamente sintetiche e/o generiche inserite dai rilevatori sulla base del linguaggio degli intervistati e, non supportando il dizionario delle professioni, non può avvalersi quando necessario della lettura integrata delle due informazioni. In particolare, al

[18] Alla stesura del paragrafo ha partecipato Filomena De Filippo, che ha curato l'attività di controllo con ACTR.

[19] Su ACTR si vedano (De Angelis, Macchia, Mazza, 2000; Macchia, Murgia, Talucci, 2008; Vicari, 2009). Senza entrare in dettagli tecnici, l'attività di codifica è preceduta da una fase di standardizzazione dei testi chiamata *parsing*, finalizzata a rimuovere tutte le varianti grammaticali e sintattiche, in modo da rendere uguali due descrizioni con lo stesso contenuto semantico originariamente diverse. Come sulla risposta da codificare, il *parsing* viene effettuato anche sulle descrizioni del dizionario (*reference file*). I testi così trattati vengono confrontati tra di loro: se si realizza un match perfetto (*direct match*) viene assegnato un unico codice, altrimenti il software tramite un algoritmo individua nel dizionario i testi più simili a quello da codificare (*indirect match*).

termine dei controlli sono risultati errati appena 107 record attribuiti all'industria in senso stretto (pari all'1,1% del totale dei record esaminati) e 177 record al commercio e alle costruzioni (pari all'1,5%). In quest'ultimo raggruppamento gli errori sono stati ripartiti tra le due sezioni, individuando un tasso di errore più elevato nelle costruzioni (1,8%) rispetto al commercio (1,4%), a conferma della bontà della decisione di inserire la prima tra quelle soggette a controllo sistematico ogni trimestre. In seguito a tale risultato si è tuttavia stabilito di non intervenire sui record dell'industria in senso stretto e del commercio, in considerazione del non elevato[20] tasso di errore rilevato in questi settori.

Il test, se da un lato ha evidenziato la presenza di un certo tasso di errore (seppur inferiore ai livelli inizialmente riscontrati nelle divisioni sottoposte a correzione sistematica), ha mostrato al contempo le difficoltà che comporterebbe l'impiego sistematico di ACTR come strumento di controllo e correzione in quanto poco compatibile con i tempi molto stretti richiesti dal processo di correzione, anche a causa del rischio di incorrere in falsi negativi. Infatti, nonostante il dizionario in dotazione sia molto ampio (circa 33.700 voci), questo non sembra ancora in grado di catalogare l'ampiezza, la varietà e l'ambiguità linguistica riscontrate tra i rispondenti all'indagine sulle Forze di lavoro. Il dizionario in uso, infatti, anche se arricchito nel corso del tempo con esempi tratti dalle indagini, risulta comunque orientato a recepire descrizioni testuali precise e tecniche, considerando che la base informativa impiegata è in gran parte costituita da vocaboli tratti dalle rilevazioni sulle imprese. D'altra parte, invece, lo strumento risulta poco flessibile nel recepire la variabilità e l'indeterminatezza del linguaggio espresso dai soggetti intervistati, peraltro non sempre pienamente consapevoli dell'attività economica svolta dall'ente/impresa presso cui lavorano, proprio perché il vocabolario utilizzato dal software si adatta meglio ad interpretare il linguaggio delle imprese più che quello degli individui. Va poi considerato che nell'indagine sulle forze di lavoro (pur con i limiti discussi nei paragrafi precedenti, che riguardano in ogni caso una quota residuale dei record elaborati), si è in presenza di una codifica effettuata da rilevatori già addestrati all'uso della classificazione e sensibilizzati alla combinazione delle informazioni ottenute nel corso dell'intervista per effettuare una codifica a un livello molto dettagliato (6° digit). Quello che è necessario per il processo in atto, quindi, non è uno strumento di codifica, sicuramente prezioso per la prima fase di classificazione delle informazioni, ma uno strumento finalizzato al controllo della codifica. Tale strumento dovrebbe consentire l'individuazione degli errori coniugando le informazioni contenute nel campo dedicato alla descrizione dell'attività economica con altre informazioni fornite nel corso dell'intervista (soprattutto la professione e il numero di unità locali dell'ente/azienda presso cui lavora l'intervistato), considerando che proprio per le divisioni più critiche la combinazione tra Ateco e professione costituisce la chiave per risolvere le descrizioni più ambigue. D'altro canto, invece, potrebbe rivelarsi sicuramente proficua un'integrazione del dizionario implementato in Actr nel navigatore utilizzato dai rilevatori per la fase di codifica, come illustrato nel paragrafo precedente.

[20] Nei primi trimestri della sperimentazione tutte le divisioni sottoposte a correzione presentavano tassi di errore più alti di quelli riscontrati nelle divisioni sottoposte a controllo con ACTR.

7. Vantaggi, svantaggi e sviluppi operativi

La procedura fin qui descritta presenta il vantaggio di concorrere al miglioramento progressivo della qualità dei dati dell'indagine sulle forze di lavoro, eliminando un numero rilevante di record mal classificati.

Si tratta di un'attività che ormai viene svolta sistematicamente[21] ogni trimestre, in modo da assicurare sia la qualità sia la coerenza della dinamica occupazionale nei diversi settori di attività economica. Rimane da valutare con attenzione l'eventualità di estendere il campo di correzione ad altri settori. Va considerato, difatti, che avviare un'operazione di controllo sistematico su tutti i dati richiederebbe un impiego di risorse consistente. Da un lato occorrerebbe un investimento iniziale per costruire i vocabolari specifici per ogni sezione/divisione di attività economica, attraverso cui confrontare la coerenza delle stringhe descrittive, dall'altro le procedure richiedono comunque una verifica finale di congruità da parte di esperti in materia. Non da ultimo è poi necessario tener conto dei tempi stringenti di diffusione dei dati dell'indagine RFL (60 giorni dalla fine del trimestre)[22].

É opportuno ricordare, inoltre, che la procedura di correzione dà origine a sezioni Ateco che 'donano' e 'ricevono' record errati (quelli sotto esame) e altri che ne 'ricevono' soltanto, pur contenendo presumibilmente al proprio interno record che potrebbero essere attribuiti ad altri settori. Mentre per alcuni comparti specifici (ad esempio servizi generali della PA, sanità e istruzione) si osserva un certo tasso di compensazione degli errori, che tendono a rimanere all'interno del macroaggregato riferibile ai servizi verso la collettività, nelle sezioni che ricevono record corretti senza la contropartita delle modifiche di quelli errati che fuoriescono, può invece manifestarsi il rischio che le variazioni nella consistenza settoriale degli occupati possano dipendere in qualche misura anche dalla parzialità della procedura di correzione.

Le soluzioni ipotizzabili sono molteplici, ma la loro eventuale adozione deve necessariamente tener conto di una valutazione complessiva di vantaggi e svantaggi che ognuna di esse comporta.

Una di queste, che potrebbe in parte attenuare la gravosità di una procedura allargata di correzione dei dati, consiste nel circoscrivere tale attività alle sezioni che risultano coinvolte dall'attuale controllo. In sostanza si tratterebbe di intervenire, dopo opportune verifiche, sulle divisioni che "ricevono" solamente record (al momento industria in senso stretto e commercio), in maniera tale da compensare i record in entrata con quelli in uscita con l'obiettivo di minimizzare la presenza di codifiche spurie in quelle divisioni.

Una strada alternativa potrebbe essere invece la definizione di una soglia minima accettabile dell'errore e abbandonare l'attività di correzione per le divisioni che scendono sotto tale soglia. Ad esempio, se tale soglia fosse fissata all'1,5%, si potrebbero già interrompere le attività di correzione nelle costruzioni, nella sanità e nell'istruzione. Tale decisione potrebbe essere supportata dai risultati delle attività di controllo effettuate a campione, finalizzate all'individuazione delle sezioni Ateco più problematiche. Da una parte la sperimentazione condotta con ACTR ha mostrato un tasso di errore più consistente nel commercio rispetto all'industria in senso stretto, suggerendo la priorità dell'intervento in questo comparto. Dall'altra il controllo eseguito su un campione di circa 3 mila

[21] Dal 2013 l'attività è stata inserita nel Piano annuale delle attività con la denominazione "Controllo e implementazione della codifica dell'attività economica nelle Forze lavoro" ob2650.

[22] In tale quadro, andrebbe attentamente considerato il rischio di introdurre variazioni nelle stime generate dall'estensione dell'attività stessa di correzione. Per tale ragione, si potrebbe optare in futuro per un approccio di correzione graduale, ad esempio dei soli record riferiti alla prima *wave*.

record del secondo trimestre 2013 ha segnalato altre sezioni critiche, quali organizzazioni e organismi extra territoriali, le attività immobiliari e le altre attività di servizi.

Le soluzioni prospettate comportano comunque un aggravio nel processo di correzione, che in tal modo interviene solamente *ex-post* sugli errori esistenti.

Per questa ragione la strada da percorrere dovrebbe essere invece quella della prevenzione dell'errore ex-ante, e in tale prospettiva la formazione assume un ruolo centrale, in quanto strumento più efficace dal punto di vista dei risultati. All'interno del processo produttivo del dato, infatti, svolge una funzione fondamentale perché concorre fortemente a prevenire gli errori di codifica. Per tale ragione è necessario continuare a investire sulla "manutenzione" delle conoscenze dei rilevatori, ampliando e intensificando le occasioni formative rivolte a tutti gli attori coinvolti nel processo di produzione dei dati.

Un ulteriore percorso da intraprendere, che completerebbe quello appena descritto, riguarda la possibilità di ampliare il dizionario dell'Ateco, un'operazione che consentirebbe di fornire uno strumento più utile ai rilevatori nella fase della codifica. Tale obiettivo potrebbe essere raggiunto arricchendo il vocabolario al momento disponibile con le informazioni che provengono dall'attività economica così come viene "raccontata" nelle indagini presso le famiglie. Una maggiore aderenza dello strumento attualmente utilizzato nel corso dell'intervista (navigatore) al linguaggio degli intervistati, difatti, potrebbe sicuramente contribuire a ridurre l'insorgenza degli errori. E' opportuno rilevare che il navigatore delle professioni è da tempo continuamente aggiornato con le nuove professioni rilevate in sede di indagine, quindi in grado di approssimarsi sempre di più al linguaggio impiegato nel mercato del lavoro, mentre quello delle attività economiche è basato esclusivamente sull'elenco delle attività ufficiali contenute nel vocabolario della classificazione perché costruito e utilizzato per le indagini rivolte alle imprese. Questo navigatore risente pertanto di una maggiore distanza dal linguaggio reale e richiederebbe un'analoga procedura di attualizzazione, proprio per facilitare la complessa fase della codifica da parte dei rilevatori.

In definitiva, è stato implementato un processo produttivo innovativo che ha consentito di elevare il livello di qualità statistica dei dati dell'indagine, un'attività a regime e ormai irrinunciabile per i risultati conseguiti. Per la sua prosecuzione e consolidamento, tuttavia, è indispensabile individuare risorse aggiuntive con l'obiettivo di continuare ad assicurare e, se possibile, migliorare la qualità dei dati prodotti per la collettività.

Bibliografia

Bolasco S. 2013. *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma: Carocci.

Bolasco S. 2005. *Statistica testuale e text mining: alcuni paradigmi applicativi*. Quaderni di Statistica. Napoli: Liguori.

Bolasco S. 1999. *Analisi multidimensionale dei dati*. Roma: Carocci.

Cortelazzo M., Tuzzi A. 2008. *Metodi statistici applicati all'italiano.* Bologna: Zanichelli.

Cortelazzo M., Tuzzi A., Nadalutti P. 2012."Una versione iterativa della distanza intertestuale applicata a un corpus di opere della letteratura italiana contemporanea". In: A. Diester, D. Longrée, G. Purnelle (eds), *Actes des 11es Journées internationales d'Analyse statistique des Données Textuelles.* Université de Liège, Facultés Universitaires Saint-Louis, Bruxelles, pp. 295-307.

De Angelis R., Macchia S., Mazza L. 2000. "Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale". Roma: Istat, Quaderni di ricerca - Rivista di statistica Ufficiale, n.1/2000, pp. 29-54.

della Ratta Rinaldi F. 2009. "Il trattamento dei dati", in F. Gallo, P. Scalisi, C. Scarnera. *L'indagine sulle professioni. Anno 2007, Contenuti, metodologia e organizzazione*. Metodi e Norme, n. 42, cap. 7, pp. 73-89. Roma: Istat.

della Ratta-Rinaldi F., Loré B. 2010. "Il lavoro e i suoi contenuti. Un'applicazione di Text Mining per categorizzare le attività dettagliate di lavoro nell'indagine campionaria sulle professioni Istat"; in S. Bolasco, I. Chiari, L. Giuliano (a cura di). *Statistical Analysis of Textual Data. Proceedings of 10th International Conference 9-10 June 2010;* pp. 195-202, 917-928 e 929-937. Roma:Led.

della Ratta-Rinaldi F., Gallo F., Loré B. 2011. "How do you name your occupation? A text mining application on the language used by workers and by the standard occupational classification", in *Cladag 2011, 8th Scientific Meeting of the Data Analysis Group of the Italian Statistical Society*, Pavia, settembre 2011. Pavia: University press.

della Ratta-Rinaldi F., Tibaldi M., Pontecorvo M. E., "Strategie di Text Mining per il controllo e la correzione della codifica dell'attività economica nell'indagine Istat sulle forze di lavoro", in E. Née, J.M. Daube, M. Valette, S. Fleury, (eds), *Actes des 12es Journées internationales d'Analyse statistique des Données Textuelles.* Paris: Sorbonne Nouvelle – Inalco, 2014..

Ferrillo A., Macchia S., Mazza L., Valery A., Vicari P. 2012. *La funzione su web per l'individuazione del codice ATECO sulla base di una descrizione sintetica e monitoraggio delle performance.* Roma: Istat working papers, n. 4.

Gallo F., Scalisi P. 2012. *Technical report on Isco08 and Nace rev.2 data quality.* Eurostat.

Gallo F., Loré B. 2012. *Training on the new occupational classification: the Italian experience.* Roma: Istat Working paper n. 12.

Gazzelloni S. (a cura di). 2006. *La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione.* Metodi e norme n. 32. Roma: Istat.

Giuliano L. 2004. *L'analisi automatica dei dati testuali. Software e istruzioni per l'uso.* Milano: LED.

Istat (Aa. Vv.). 2011. Linee guida per la qualità dei processi statistici, Roma. (http://www.istat.it/it/files/2010/09/Linee-Guida-Qualit%C3%A0-_v.1.1_IT.pdf).

Labbé C., Labbé D. 2003. *La distance intertextuelle*. *Corpus*, 2:95-118.

Lebart L. 1982. *Exploratory Analysis of Large Sparse Matrices, with Application to Textual Data.* Vienna: Compstat, Physica Verlag, pp. 67-76.

Macchia S., Murgia M., Talucci V., "Coding the spoken language through the integration of different approaches of textual analysis", in *JADT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles.* Lyon: 12-14 mars 2008.

Macchia S., Mastroluca S. 2013. *Il trattamento delle variabili testuali nel 15° Censimento generale della popolazione.* Roma*:* Istat working papers, n. 3.

Poibeau T. 2003. *Extraction automatique d'information (du texte brut au web sémantique).* Paris: Hermes-Lavoisier.

Tuzzi. A., Zaccarin S. 2004. "Il lavoro raccontato dai laureati: analisi lessico-testuale delle professioni". In: E. Aureli Cutillo. *Strategie metodologiche per lo studio della transizione Università lavoro*. pp. 357-373. Padova: Cleup.

Vicari P., Ferrillo A., Valery A. (a cura di). 2009. *Classificazione delle attività economiche - Ateco 2007.* Metodi e norme n. 40. Roma: Istat.

Vicari P. (a cura di). 2009. *L'ambiente di codifica automatica dell'Ateco 2007.* Metodi e norme n. 41. Roma: Istat.

Informazioni per gli autori

La collana è aperta ad autori dell'Istat e del Sistema statistico nazionale, e ad altri studiosi che abbiano partecipato ad attività promosse dal Sistan (convegni, seminari, gruppi di lavoro, ecc.). Da gennaio 2011 essa sostituirà Documenti Istat e Contributi Istat.

Coloro che desiderano pubblicare sulla nuova collana dovranno sottoporre il proprio contributo alla redazione degli Istat Working Papers inviandolo per posta elettronica all'indirizzo iwp@istat.it. Il saggio deve essere redatto seguendo gli standard editoriali previsti, corredato di un sommario in italiano e in inglese; deve, altresì, essere accompagnato da una dichiarazione di paternità dell'opera. Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Per gli autori Istat, la sottomissione dei lavori deve essere accompagnata da una mail del proprio dirigente di Servizio/Struttura, che ne assicura la presa visione. Per gli autori degli altri enti del Sistan la trasmissione avviene attraverso il responsabile dell'ufficio di statistica, che ne prende visione. Per tutti gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione. Tutti i lavori saranno sottoposti al Comitato di redazione, che valuterà la significatività del lavoro per il progresso dell'attività statistica istituzionale. La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line.

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Salvo diversa indicazione la riproduzione è libera, a condizione che venga citata la fonte.