istat
working
papers

N.12
2016

Strategie di imputazione delle variabili Redditi da lavoro dipendente e Consumi Intermedi per le Amministrazioni Comunali

Orietta Luzi, Roberta Varriale, Tiziana Pichiorri, Gerolamo Giungato

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.12
2016

Strategie di imputazione delle variabili Redditi da lavoro dipendente e Consumi Intermedi per le Amministrazioni Comunali

Orietta Luzi, Roberta Varriale, Tiziana Pichiorri, Gerolamo Giungato

Istat

Istat Working Papers

Strategie di imputazione delle variabili Redditi da lavoro dipendente e consumi Intermedi per le Amministrazioni Comunali

N. 12/2016

ISBN 978-88-458-1903-2

Strategie di imputazione delle variabili Redditi da lavoro dipendente e Consumi Intermedi per le Amministrazioni Comunali[1]

Orietta Luzi, Roberta Varriale, Tiziana Pichiorri, Gerolamo Giungato

Sommario

In questo lavoro viene presentato lo stato di avanzamento del progetto, avviato dall'Istat nel 2014, di costruzione di un sistema informativo della Pubblica Amministrazione, basato sull'uso integrato di microdati provenienti da fonti amministrative e statistiche. Nel documento si concentra l'attenzione sulle problematiche legate alla qualità e sulle soluzioni metodologiche adottate per trattare le informazioni mancanti delle variabili "Redditi da lavoro dipendente" e "Consumi intermedi" delle Amministrazioni Comunali. Fonte primaria è il Certificato del rendiconto al Bilancio, mentre informazioni ausiliarie nelle strategie di stima sono: il Censimento dell'Industria e dei Servizi 2011, il Censimento della popolazione 2011, l'indagine annuale Istat/Posas sulla popolazione residente, il Registro delle Istituzioni pubbliche. L'anno di riferimento è il 2012, mentre i dati relativi al 2011 sono utilizzati come informazione storica nei modelli longitudinali.

Parole chiave: Dati amministrativi, Pubblica Amministrazione, Bilanci, Imputazione.

Abstract

This report presents the state-of-the-art of an Istat research project started in 2014 to develop a statistical information system for the General Government Sector, based on the integrated use of microdata from different administrative and statistical sources. The report focuses on quality issues and methodological solutions adopted to deal with missing information of the variables "Compensation of employees" and "Intermediate consumption" of Municipalities. The Certificate of the financial statements represents the primary source, while auxiliary information used in the estimation strategy comes from statistical sources, such as 2011 General Census of Industry and Services, 2011 General Population and Housing Censuses, Resident municipal population by age, sex and marital status, Italian Statistical Register of Public Institutions. The reference year is 2012, while 2011 data are used as historical information in longitudinal models.

Keywords: Administrative data, General Government Sector, Balance sheet, Imputation.

[1] Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Indice

1. Introduzione

L'uso a fini statistici di dati di fonte amministrativa ha acquisito una rilevanza crescente nei settori produttivi dell'Istat, e subirà un ulteriore impulso grazie alla valorizzazione di fonti di dati secondarie determinato dal prossimo passaggio, previsto dalla modernizzazione dell'Istituto, ad un nuovo e più efficiente sistema di produzione statistica basato sull'uso di registri statistici integrati alimentati essenzialmente da fonti amministrative.

Le numerose esperienze di sfruttamento massivo di dati di fonte amministrativa condotte in Istat in ambito sia economico sia demo-sociale[2], se da un lato hanno mostrato le forti potenzialità di tale approccio alla stima in termini sia di riduzione dei costi di indagine sia di miglioramento della qualità complessiva dei risultati, dall'altro hanno evidenziato l'esistenza di nuove problematiche e criticità rispetto ai tradizionali processi di produzione statistica. Inoltre, l'uso integrato di fonti amministrative e fonti statistiche presuppone il disegno di nuove strategie di stima e nuove metodologie per la gestione delle componenti campionarie e non campionarie dell'errore (per una rassegna di tali problematiche vedere ad esempio Wallgren *et. al.* (2007), e Zhang (2012) e, fra i prodotti Istat, le *Linee guida per la qualità dei processi statistici di fonte amministrativa* (Istat, 2015a)).

In questo lavoro viene illustrata un'esperienza di uso integrato di fonti amministrative nell'area della Pubblica Amministrazione (PA nel seguito) ai fini della stima di variabili economiche che rivestono un ruolo fondamentale nell'ambito della Contabilità Nazionale e per le quali è divenuta sempre più pressante la necessità di miglioramento di qualità, trasparenza e tempestività.

Nel 2014 l'Istat ha avviato un progetto sperimentale finalizzato alla valutazione di benefici e problematiche connesse alla realizzazione di un sistema informativo statistico integrato di microdati nell'area della PA ("Frame PA") per la stima delle principali voci di conto economico, basato sull'uso integrato di fonti amministrative e statistiche. In questo ambito, oltre ai consueti problemi di integrabilità dei dati e di armonizzazione dei contenuti informativi delle fonti, vanno considerati aspetti specifici connessi alla particolare natura delle unità della popolazione obiettivo e dei fenomeni oggetto di stima: ad esempio, si rileva una marcata disomogeneità di comportamento di Enti pubblici appartenenti alla stessa tipologia amministrativa (ad es. Comuni, Provincie, Regioni) rispetto ai fenomeni economici oggetto di stima, con differenze nelle classificazioni delle poste in bilancio che possono determinare importanti effetti distorsivi sulle stime. Inoltre, in questo contesto le fonti sono caratterizzate da una forte variabilità sia nella loro tempistica (con forniture dei dati in tempi successivi con diversi livelli di copertura/qualità), sia nei contenuti (dovuta alla continua e frequente evoluzione normativa italiana degli ultimi anni), che rendono particolarmente difficile l'analisi longitudinale di alcuni fenomeni.

In questo lavoro sono descritte le metodologie utilizzate e le principali evidenze relative per l'analisi di una specifica tipologia di PA, le Amministrazioni Comunali, e di due fra le variabili principali del conto economico, i *Redditi da lavoro dipendente* e i *Consumi intermedi*. Come fonte primaria di informazione è stato utilizzato l'archivio di fonte Ministero dell'Interno "Certificato del rendiconto al Bilancio", mentre come fonti di informazione ausiliaria sono stati utilizzati il registro delle Amministrazioni Pubbliche di fonte Istat, e il Sistema Informativo sulle Operazioni degli Enti Pubblici. L'anno di riferimento è il 2012. Il documento è strutturato come segue. Nel paragrafo 2 sono sinteticamente descritte le principali caratteristiche delle fonti utilizzate nello studio delle Amministrazioni Comunali. Il paragrafo 3 contiene un'analisi del fenomeno della mancata risposta così come si è configurato nello studio in questione, mentre nel paragrafo 4 sono sinteticamente illustrati i metodi di imputazione sottoposti a valutazione sperimentale e la strategia adottata per la loro valutazione comparativa. Il paragrafo 5 riporta i risultati relativi alla variabile *Redditi da lavo-*

[2] Fra le più recenti, ricordiamo ad esempio i progetti basati sull'utilizzo integrato di fonti statistiche e amministrative per la stima dei conti economici delle piccole e medie imprese (Luzi et al., 2014), per la stima dell'occupazione (De Gregorio et al., 2015), per la realizzazione del Censimento permanente (Istat, 2015b).

ro dipendente, mentre i risultati relativi alla variabile *Consumi Intermedi* sono illustrati nel paragrafo 6. Nel paragrafo 7 viene esposta una valutazione del processo di imputazione basata sull'analisi di dati aggiornati della fonte Certificato del rendiconto al Bilancio. Il paragrafo 8 contiene infine alcune considerazioni conclusive e prospettive di lavoro future.

2. Le fonti informative e la strategia di stima

La prima fase del lavoro sperimentale è consistita nella ricognizione delle fonti amministrative e statistiche attualmente utilizzate o utilizzabili per la stima del conto economico delle Amministrazioni Comunali. Ai fini delle sperimentazioni descritte in questo lavoro, sono stati quindi utilizzati i seguenti archivi amministrativi e statistici:

- Registro delle Amministrazioni Pubbliche (fonte Istat, Lista S13 nel seguito), anni 2011 e 2012 (Istat, 2015c). Questo registro contiene informazioni relative a anagrafica, classificazione, stato di attività, etc. delle varie tipologie di Enti pubblici attivi nell'anno di riferimento. L'archivio contiene inoltre l'informazione relativa al numero dei dipendenti risultante dal 9° Censimento Industria e Servizi, anno 2011 (Istat, 2013).
- Certificato del rendiconto al Bilancio (fonte Ministero dell'Interno, CRB nel seguito), anni 2011 e 2012. Questo archivio contiene informazioni relative alle voci di bilancio consuntivo dei Comuni. Si tratta della fonte primaria utilizzata nella procedura di stima di variabili economiche per questa tipologia di AP. Nel corso dell'anno *t* il Ministero dell'Interno trasmette i dati dei Certificati riferiti all'esercizio *t-2* in due occasioni: dati provvisori a gennaio e dati definitivi a luglio/settembre. I primi vengono utilizzati sia ai fini della stima dei bilanci consuntivi delle Amministrazioni Comunali (Istat, 2014), sia dalla Contabilità Nazionale per l'elaborazione dei conti utili per la prima notifica dei conti pubblici a Eurostat, a marzo dell'anno *t*; quelli definitivi invece, poiché pervengono in ritardo rispetto alle scadenze europee, aggiornano i conti pubblici nell'anno *t+1*.
- Sistema Informativo sulle Operazioni degli Enti Pubblici (SIOPE nel seguito), anni 2011 e 2012. Questo archivio contiene informazioni relative alle riscossioni e ai pagamenti (flussi di cassa) dei Comuni, dettagliate per le principali voci economiche di entrata ed uscita. Tale sistema acquisisce in tempo reale le informazioni dai tesorieri degli enti locali, tuttavia le informazioni riferite ai periodi più recenti scontano problemi di provvisorietà in quanto un gran numero di operazioni vengono temporaneamente registrate e classificate nel sistema in una voce provvisoria in attesa di essere registrate correttamente nei mesi successivi. Gli enti, a partire dal 2014, possono modificare i prospetti inviati al sistema in tutte le loro articolazioni fino alla fine dell'anno successivo all'esercizio di riferimento. Il sistema è pubblico dal 2014, tuttavia per esigenze informative di maggior dettaglio, l'Istat acquisisce direttamente da Banca d'Italia l'intero database contenente i dati di ciascun singolo ente.
- Rendiconto delle Amministrazioni Comunali (SIRTEL). Negli ultimi anni, sulla base di un protocollo d'intesa fra Istat e la Corte dei Conti, l'Istat acquisisce il database costituito presso la Corte dei Conti. Il contenuto di questa fonte è analogo a quello del Ministero dell'Interno, anche se maggiormente analitico, ed è stato utilizzato dal 2014 sia per verificarne la coerenza con i dati di fonte Ministero dell'Interno, sia per effettuare approfondimenti sui singoli enti e su specifiche operazioni di bilancio. Sebbene il flusso di dati con la Corte dei Conti sia già a regime, questa fonte non è stata utilizzata nel presente lavoro poiché l'utilizzo dei relativi dati è ancora in fase sperimentale. Fermo restando l'utilizzo del CRB come fonte primaria per l'elaborazione dei conti dei Comuni, i dati dei Rendiconti della Corte dei Conti costituiscono potenzialmente una fonte rilevante per l'integrazione dei dati delle unità mancanti e per la correzione di eventuali errori ed anomalie nei dati di base.

Nell'analisi delle fonti a disposizione, è da considerare che il sistema attuale di fonti sui rendi-

conti delle amministrazioni locali a breve assumerà un assetto diverso in previsione dell'entrata a regime del Piano dei Conti integrato definito nell'ambito del processo di armonizzazione dei conti degli enti pubblici locali e della ridefinizione dei principi contabili (Ragioneria Generale dello Stato, 2013). Tali dati, a vari livelli di dettaglio, così come previsto dal Decreto legislativo 118 del 2011 e relativi aggiornamenti, alimenteranno la costituenda Banca Dati Unitaria della PA e le procedure di acquisizione, trattamento e elaborazione andranno adeguate alla nuova struttura dei conti[3].

Ai fini delle analisi riportate in questo lavoro sono state utilizzate anche alcune informazioni sulle caratteristiche della popolazione e sulle superfici comunali provenienti dalla rilevazione Istat sulla Popolazione residente comunale per sesso, anno di nascita e stato civile (POSAS), anni 2011 e 2012 (Istat, 2012). Sono state inoltre utilizzate informazioni provenienti dal Censimento delle Istituzioni Pubbliche (nell'ambito del 9° Censimento generale dell'Industria e dei Servizi e Censimento delle Istituzioni non profit), anno 2011.

La strategia di stima sperimentale adottata in questo lavoro ha riguardato alcune voci del conto economico dei Comuni, selezionate da un lato per la loro rilevanza economica, dall'altro perché maggiormente stabili (almeno a livello aggregato) da un punto di vista definitorio e di criteri contabili adottati dalle amministrazioni pubbliche. Per ogni Comune, i valori delle voci *Redditi da Lavoro Dipendente* e *Consumi Intermedi* sono stati ottenuti direttamente dalla fonte CRB. L'uso di tale fonte ha implicato da un lato la necessità di identificare correttamente e univocamente le unità statistiche (i Comuni) a partire dalla Lista S13, dall'altro di derivare le variabili oggetto di stima a partire dai contenuti informativi della fonte, attraverso un processo di armonizzazione delle definizioni adottate rispetto alle definizioni di Contabilità Nazionale (CN nel seguito). Come discusso nel paragrafo 3, problemi di sotto-copertura della fonte determinano l'assenza di informazioni di base su una quota di Comuni (assimilabili a unità non rispondenti): la ricostruzione statistica delle informazioni per questi Comuni può avvenire sia a livello aggregato sia a livello di microdati. In questo lavoro è stato adottato quest'ultimo tipo di approccio, per cui il completamento della base informativa è stato ottenuto utilizzando modelli di imputazione statistica di massa che sfruttano le informazioni ausiliarie sui Comuni contenute sia nella fonte CRB, sia in altre fonti amministrative e statistiche.

3. La mancata risposta

La "mancata risposta" corrisponde in questo studio all'assenza di informazioni nella fonte CRB per un sottoinsieme di Amministrazioni Comunali, originata da errori di risposta, di copertura (dovuti alla demografia dei Comuni, per cui un Comune da un anno al successivo può essere "inglobato" in un altro, oppure suddiviso in più Amministrazioni Comunali), alla tempistica della fonte (alcuni Comuni non inviano i dati di Bilancio in tempi compatibili con il rilascio della fonte all'Istat), a errori di identificazione delle unità (ad esempio uno stesso Comune può avere codici identificativi diversi da un anno all'altro, oppure possono esserci errori nei codici identificativi con conseguenti errori di abbinamento rispetto alla Lista S13, ecc.).

Nei due anni considerati, i Comuni presenti nella Lista S13 sono 8.092. Di questi, trovano corrispondenza nella fonte CRB 7.757 Comuni nel 2011 (96%) e 7.387 nel 2012 (91,3%), mentre la totalità si abbina al database SIOPE.

[3] Dal 2012 il Piano dei Conti è in sperimentazione presso un campione di amministrazioni locali, tra cui Regioni, Province e Comuni, e andrà in vigore su tutti gli enti nel 2016 con riferimento ai consuntivi dell'esercizio 2015.

Tavola 1 – Comuni mancanti per ripartizione geografica (anno 2012)

Ripartizione geografica	Totale	Valori mancanti	
		N	%
Nord-Ovest	3.059	136	4,4
Nord-Est	1.480	82	5,5
Centro	996	80	8,0
Sud	1.790	267	14,9
Isole	767	140	18,3
Totale	8.092	705	8,7

Tavola 2 – Comuni mancanti per classe di popolazione (anno 2012)

Classi di popolazione	Totale	Valori mancanti	
		N	%
< 1500	2.866	284	9,9
[1500,5000)	2.832	248	8,8
[5000,10000)	1.189	96	8,1
[10000,60000)	1.104	73	6,6
[60000,100000)	55	3	5,5
> 100000	46	1	2,2
Totale	8.092	705	8,7

Tavola 3 – Comuni mancanti per classe di popolazione e ripartizione geografica (anno 2012)

Classi di popolazione e ripartizione geografica		Totale	Valori mancanti	
			N	%
< 1500	Nord-Ovest	1.426	72	5,0
	Nord-Est	359	37	10,3
	Centro	268	30	11,2
	Sud	581	100	17,2
	Isole	232	45	19,4
[1500,5000)	Nord-Ovest	988	44	4,5
	Nord-Est	568	34	6,0
	Centro	355	27	7,6
	Sud	638	100	15,7
	Isole	283	43	15,2
[5000,10000)	Nord-Ovest	364	13	3,6
	Nord-Est	295	8	2,7
	Centro	160	15	9,4
	Sud	257	34	13,2
	Isole	113	26	23,0
[10000,60000)	Nord-Ovest	262	6	2,3
	Nord-Est	237	3	1,3
	Centro	191	8	4,2
	Sud	288	31	10,8
	Isole	126	25	19,8
[60000,100000)	Nord-Ovest	12	1	8,3
	Nord-Est	5	.	.
	Centro	14	.	.
	Sud	17	2	11,8
	Isole	7	.	.

Tavola 3 (segue) – Comuni mancanti per classe di popolazione e ripartizione geografica (anno 2012)

Classi di popolazione e ripartizione geografica		Totale	Valori mancanti	
			N	%
> 100000	Nord-Ovest	7	.	.
	Nord-Est	16	.	.
	Centro	8	.	.
	Sud	9	.	.
	Isole	6	1	16,7
Totale		8.092	705	8,7

Tavola 4 – Comuni mancanti per regione (anno 2012)

Regione	Totale	Valori mancanti	
		N	%
Piemonte	1.206	43	3,6
Valle D'Aosta	74	5	6,8
Lombardia	1.544	63	4,1
Trentino-Alto Adige	333	39	11,7
Veneto	581	24	4,1
Friuli-Venezia Giulia	218	12	5,5
Liguria	235	25	10,6
Emilia-Romagna	348	7	2,0
Toscana	287	5	1,7
Umbria	92	1	1,1
Marche	239	4	1,7
Lazio	378	70	18,5
Abruzzo	305	40	13,1
Molise	136	20	14,7
Campania	551	84	15,2
Puglia	258	34	13,2
Basilicata	131	13	9,9
Calabria	409	76	18,6
Sicilia	390	91	23,3
Sardegna	377	49	13,0
Totale	8.092	705	8,7

4. Strategia di valutazione

Per la ricostruzione delle voci di bilancio dei Comuni presenti nella Lista S13 nell'anno di riferimento, ma per i quali non è disponibile nessuna informazione nell'archivio CRB, si propone una procedura di imputazione supportata dall'uso di informazioni ausiliarie disponibili nelle altre fonti.

Per ciascuna delle variabili obiettivo sono stati valutati sperimentalmente diversi metodi: la scelta del metodo "ottimale" è stata basata su uno studio di simulazione che ha consentito di valutare l'accuratezza della ricostruzione dei dati a livello sia degli aggregati finali, sia di dati elementari. A questa valutazione si è affiancata la misurazione delle discrepanze fra gli aggregati ricostruiti sulla base delle "migliori" metodologie individuate e i corrispondenti aggregati pubblicati dall'Istat per l'anno di riferimento 2012 (Istat, 2014). Un'ulteriore valutazione ha sfruttato la disponibilità di parte delle informazioni mancanti in una fornitura successiva (aggiornata) della fonte CRB per il 2012.

Lo studio di simulazione è stato basato su un approccio di tipo Monte Carlo strutturato nei seguenti passi principali:

a) simulazione di una prefissata quota di valori mancanti, secondo un *modello Missing Completely At Random* (MCAR), sulla variabile obiettivo nel sottoinsieme dei Comuni rispon-

denti. La quota di valori mancanti generati artificialmente è stata scelta sulla base della percentuale di Comuni con dati mancanti osservata nel 2012;

b) ricostruzione dei dati mancanti mediante i metodi di imputazione oggetto di valutazione;
c) calcolo di misure di distanza (a livello elementare e aggregato) fra i valori imputati e i corrispondenti valori osservati;
d) Iterazione dei passi da a) a c) per k=1.000 volte;
e) Calcolo di indicatori di qualità di sintesi delle distribuzioni delle misure di cui al punto c).

4.1 Metodi di imputazione

I metodi oggetto di valutazione nel presente lavoro sono di tipo sia parametrico (prevedono cioè l'uso di modelli di tipo regressivo) sia non parametrico (basati su criteri da donatore di distanza minima), e sono stati sperimentati in combinazione sia con approcci di natura trasversale (con utilizzo esclusivo di informazioni relative all'anno di riferimento) che longitudinale (avvalendosi anche di informazioni storiche). I metodi sottoposti a valutazione sono sinteticamente illustrati di seguito.

– *Donatore di distanza minima trasversale per classi (Nearest Neighbour Donor, NND).*
Con questo metodo, noto in letteratura anche come *ratio hot-deck* (de Waal *et al.*, 2011), una volta individuato il donatore di distanza minima rispetto a opportune variabili di *matching*, il valore imputato Y_i^* della variabile obiettivo nel Comune mancante *i* si ottiene mediante la relazione:

$$Y_i^* = X_i \times \frac{Y_d}{X_d} = X_i \times Y_{d,pc} \qquad (1)$$

dove Y_d e X_d sono i valori rispettivamente della variabile obiettivo *Y* e della variabile ausiliaria *X*, nota per tutte le unità della popolazione, nel Comune donatore *d*, mentre $Y_{d,pc}$ è il valore pro-capite della variabile obiettivo calcolato rispetto alla variabile ausiliaria. Usualmente la ricerca del donatore avviene all'interno di classi omogenee di unità (celle di imputazione) definite sulla base di informazioni note su tutte le unità della popolazione obiettivo e considerate esplicative del meccanismo sottostante il fenomeno della mancata risposta. Tali informazioni sono tipicamente di tipo territoriale e/o dimensionale, di natura categorica o continua, ridotte in classi.

– *Predictive Mean Matching per classi (PMM).*
Il PMM è una tecnica di imputazione NND basata su funzioni di distanza nelle quali le variabili di *matching* sono ponderate in base al loro potere predittivo rispetto alle variabili da imputare. In un contesto multivariato, il PMM è generalmente applicato per abbinare ciascuna unità ricevente con il donatore avente media predittiva più vicina rispetto al valore previsto da un modello di regressione fra la variabile da imputare e un insieme di covariate. La selezione dei donatori si basa sulla distanza di Mahalanobis definita in termini di matrice di covarianza residua dal modello di regressione (Little, 1988). L'applicazione del metodo prevede due fasi. La prima fase consiste nello stimare un modello di regressione lineare multivariato sulle unità rispondenti della popolazione obiettivo suddivisa in classi omogenee (un modello per ogni classe):

$$Y_i^p = \alpha + \beta_1 X_i + \beta_2 Z_i + ... + e_i \qquad (2)$$

dove *X, Z,…* sono variabili ausiliarie correlate alla variabile obiettivo ed *e* è l'errore residuo su cui si assumono valide le usuali ipotesi distribuzionali. La stima dei parametri del modello di regressione è poi utilizzata per ottenere il valore Y_i^p previsto dal modello per tutte le unità (rispondenti e non rispondenti). Nella seconda fase, il valore imputato Y_i^* della variabile obiettivo è ottenuto dal donatore di distanza minima, scelto nella classe del Comune mancante e selezionato utilizzando come variabile di *matching* la variabile Y^p predetta al passo precedente.

– *Donatore di distanza minima longitudinale per classi*.
Si tratta di un metodo analogo al NND, in cui cioè si imputano preliminarmente i procapite della variabile obiettivo utilizzando il donatore di minima distanza per classi omogenee, ma in cui le variabili di *matching* includono informazioni sui Comuni riferite all'anno precedente (ad esempio, valori di variabili ausiliarie oppure valori pro-capite della variabile obiettivo riferite all'anno precedente):

$$Y_i^*(t) = X_i(t) \times \frac{Y_d(t-1)}{X_d(t-1)} = X_i(t) \times Y_{d,pc}(t-1) \qquad (3)$$

– *Metodi deterministici longitudinali*.
In generale, questi metodi sfruttano il valore della variabile obiettivo nell'anno precedente, aggiornandolo o con un trend individuale calcolato su una delle variabili ausiliarie disponibili (ad es. da fonte SIOPE, da Censimento, ecc.), oppure con un trend mediano per classi. Versioni "miste" di alcuni metodi combinano un approccio longitudinale deterministico con uno degli altri metodi descritti sopra (in particolare, il donatore longitudinale). I metodi utilizzati per ciascuna delle due variabili obiettivo sono descritti nei paragrafi 5.3 e 6.3.

4.2 Indicatori di sintesi

Gli indicatori utilizzati come sintesi delle distribuzioni delle misure di distanza sono (Luzi et al., 2007):

– *Relative Bias* (RB) - o *Relative estimation error due to imputation* - nel dominio *D*:

$$RB_Y^D = \frac{1}{I}\sum_{i=1}^{I} \frac{\left(\hat{T}_{Y,true}^D - \hat{T}_{Y,imp}^D(i)\right)}{\hat{T}_{Y,true}^D} \times 100 \qquad (4)$$

dove $\hat{T}_{Y,true}^D$, $\hat{T}_{Y,imp}^D$ sono le stime dei totali della variabile obiettivo calcolate rispettivamente sui valori osservati (veri) e sui valori imputati (per ogni iterazione *i, i=1,…, I=1.000*) nel dominio *D*.

– *Relative Root Mean Squared Error* (RMSE):

$$RMSE_Y^D = \frac{1}{I}\sum_{i=1}^{I} \sqrt{\frac{\left(\hat{T}_{Y,true}^D - \hat{T}_{Y,imp}^D(i)\right)^2}{\hat{T}_{Y,true}^D}} \times 100 \qquad (5)$$

– *Relative Imputation Error* (RIE);

$$RIE_Y = \frac{1}{I}\sum_{i=1}^{I}\frac{\sqrt{\sum_{k=1}^{n^*}(y_{true,k} - y_{imp,k})^2}}{\sum_{k=1}^{n^*} y_{true,k}} \qquad (6)$$

dove $y_{true,k}$, $y_{imp,k}$ sono rispettivamente il valore originale ed imputato della variabile obiettivo *Y* nell'unità *k*, ed *n** è il numero di unità rispondenti con valore simulato di mancata risposta.

Una ulteriore valutazione dei risultati è stata effettuata considerando le distanze relative fra i valori degli aggregati (a livello regionale) delle variabili obiettivo ottenuti utilizzando le metodologie di imputazione proposte, ed i corrispondenti valori ottenuti sulla base della procedura di stima corrente (Istat, 2014):

$$Diff^{D}_{\mathrm{Im}\,p,B} = \frac{\left(\hat{T}^{D}_{\mathrm{Im}\,p} - \hat{T}^{D}_{B}\right)}{\hat{T}^{D}_{B}} \times 100 \qquad (7)$$

dove $\hat{T}^{D}_{B}$, $\hat{T}^{D}_{\mathrm{Im}\,p}$ sono i valori dell'aggregato della variabile *Y* ottenuti rispettivamente con il metodo corrente di trattamento delle mancate risposte (Istat, 2014) e con un diverso metodo di imputazione.

Una valutazione aggiuntiva della qualità delle imputazioni ottenute sperimentalmente è stata basata sull'utilizzo dei dati contenuti in una fornitura aggiornata dell'archivio CRB (descritta nel paragrafo 7). Infatti, rispetto alla fornitura provvisoria della fonte utilizzata nello studio di simulazione, la fornitura aggiornata dell'archivio contiene, per un certo anno di riferimento, informazioni su un sottoinsieme di Comuni mancanti (non coperti) nella fornitura iniziale[4]. Questo ha consentito di misurare l'accostamento fra i valori inizialmente imputati e i corrispondenti valori successivamente "osservati" delle variabili obiettivo. La qualità del processo di imputazione è stata valutata in questo caso analizzando le discrepanze tra dato imputato e dato osservato nei Comuni mancanti nella prima fornitura ma presenti in quella aggiornata. L'indicatore utilizzato, opportunamente modificato, è il *RIE* (6) dove $y_{imp,k}$, $y_{true,k}$ sono rispettivamente i valori imputati (prima fornitura) ed osservati (fornitura definitiva) della variabile obiettivo *Y* nell'unità *k*, ed n^* è il numero di unità assenti nella prima fornitura ma presenti in quella aggiornata.

5. Stima della variabile *Redditi da lavoro dipendente*

In questo paragrafo sono descritti i metodi utilizzati e i risultati ottenuti per la variabile *Redditi da lavoro dipendente* (*Redditi* nel seguito). Tale variabile è definita come segue (Corradini, 2014):

I redditi da lavoro dipendente (SEC 1995, punto 4.02 e seguenti) sono definiti come il compenso complessivo, in denaro o in natura, riconosciuto da un datore di lavoro a un lavoratore dipendente

[4] Mentre non risultano modificati i valori già presenti nelle altre unità della popolazione.

quale corrispettivo per il lavoro svolto da quest'ultimo durante il periodo di riferimento. Esso si suddivide in retribuzioni lorde e contributi sociali a carico dei datori di lavoro.

La variabile *Redditi*, rilevata con criterio di competenza, è misurata in maniera diretta nel Quadro 4 del CRB tramite il *Totale delle Spese correnti per il Personale – Impegni*; tale variabile verrà nel seguito indicata con *Y*.

La variabile *Z*, data dal *Totale delle Spese correnti per il Personale – Pagamenti in Conto Competenza* più il *Totale delle Spese correnti per il Personale – Pagamenti in Conto Residui*, e la variabile *S*, data dalla somma di numerose voci dell'archivio SIOPE (vedere Appendice 1), misurano le spese per il personale con criterio di cassa. Tali variabili sono relative a voci di spesa concettualmente confrontabili e possono quindi essere utilizzate sia a fini di analisi, sia come informazioni ausiliarie nei modelli di imputazione della variabile dipendente *Y*.

5.1 Analisi esplorativa dei dati

In questo paragrafo sono illustrate alcune analisi di tipo esplorativo aventi il duplice obiettivo di descrivere la variabile oggetto di imputazione (*Redditi* e/o *Redditi pro-capite*) e le variabili ausiliarie utilizzabili nella procedura di imputazione, oltre che di evidenziare eventuali problematiche (problemi di abbinamento, discrepanze definitorie, ecc.) da trattare opportunamente.

Analizzando la presenza o assenza di informazione sulla variabile *Redditi* per fonte, si rileva che l'informazione da CRB sulle singole componenti è tutta presente o tutta assente, mentre l'informazione proveniente da SIOPE ha *pattern* diversi di risposta parziale. La Tavola 5 riporta i *pattern* di non risposta (complessiva) nelle due fonti per i due anni considerati. Il valore "1" indica la presenza dell'informazione, mentre il valore "0" ne indica l'assenza.

Tavola 5 – Variabili *Redditi*, *pattern* di dati mancanti nel 2011 e 2012

CRB 2012	CRB 2011	SIOPE 2012	SIOPE 2011	N.	%
1	1	1	1	7.121	88,00
1	1	1	0	3	0,04
1	1	0	1	2	0,02
1	1	0	0	85	1,05
1	0	1	1	165	2,04
1	0	0	0	11	0,14
0	1	1	1	534	6,60
0	1	0	0	3	0,04
0	0	1	1	167	2,06
0	0	0	0	1	0,01

Il numero totale di osservazioni con informazione completa nei due anni è pari a 7.121, mentre il numero di Comuni per cui è necessario procedere all'imputazione della variabile *Redditi* nel 2012 è pari a 705 (righe della tavola con CRB 2012 pari a 0), pari all'8,7% del totale delle Amministrazioni Comunali della Lista S13. Si noti che 168 Comuni (circa il 2%) non hanno informazioni nell'archivio CRB per entrambi gli anni di riferimento: su queste amministrazioni andrebbero effettuati controlli specifici per verificare la ragione della persistenza del problema nel tempo. Per uno di questi Comuni (ultima riga della Tavola 5) si registra addirittura l'assenza di informazioni anche nella base dati SIOPE in entrambi gli anni considerati.

Dal punto di vista statistico, il contenuto informativo delle variabili *Y*, *Z* e *S* risulta essere molto correlato. In particolare, qualora l'informazione è presente simultaneamente nelle diverse fonti, sono state calcolate le differenze relative percentuali tra *Y* e *Z* e tra *S* e *Z* utilizzando i semplici indici:

$$Diff_{i,YZ} = \frac{\left(Y_i - Z_i\right)}{Z_i} \times 100 \qquad (8)$$

$$Diff_{i,SZ} = \frac{\left(S_i - Z_i\right)}{Z_i} \times 100 \qquad (9).$$

La Tavola 6 mostra alcune statistiche descrittive delle distribuzioni delle variabili $Diff_{i,YZ}$ e $Diff_{i,SZ}$ per il 2012. Come anticipato, l'informazione presente nelle diverse variabili risulta essere quasi equivalente, soprattutto per quanto riguarda le variabili *S* e *Z* che misurano i *Redditi* con lo stesso criterio (cassa), diverso da quello (competenza) utilizzato dalla variabile *Y*.

Tavola 6 – Variabile *Redditi*, statistiche descrittive delle distribuzioni di $Diff_{i,YZ}$ e $Diff_{i,SZ}$ (anno 2012)

Variabile	N	Media	Min	Max	Mediana	Dev. Std.
$Diff_{YZ}$	7.379	1,78	-99,99	1161,04	1,13	15,68
$Diff_{SZ}$	7.289	-0,94	-77,79	47,88	0	3,78

Le Figure 1 e 2 mostrano i *box-plot* della variabile *Redditi pro-capite* (*Y/Ndip*, dove con *Ndip* si indicano il Numero dei dipendenti comunali nel 2011), costruiti, rispettivamente, considerando tutte le unità ed eliminando dall'insieme delle unità quelle con valori risultati "anomali" ad un'analisi grafica di tipo esplorativo. Ciò che si osserva è il numero limitato di situazioni fuori *range*. Inoltre, da un'analisi approfondita di tali valori, si rileva una probabile presenza di errori nel Numero dei dipendenti anziché nella variabile *Redditi*.

Relativamente alle informazioni disponibili da altre fonti, potenzialmente utilizzabili come variabili ausiliarie nella procedura di imputazione (per la definizione di opportune classi di imputazione oppure in qualità di covariate nei modelli di imputazione), sono state prese in considerazione le seguenti informazioni:

- Numero di dipendenti comunali (anno 2011);
- Ripartizione geografica;
- Popolazione residente;
- Capoluogo di provincia (si/no);
- Metropoli (variabile dicotomica; vale 1 per i Comuni con più di 500.000 abitanti);
- Altitudine;
- Comune montano (si/no);
- Comune litoraneo (si/no);
- Superficie;
- Zona altimetrica;

Per la selezione delle covariate statisticamente significative ai fini della stima della variabile *Redditi*, sono stati utilizzati modelli di regressione lineare stimati con procedura *stepwise*. Si tratta di una procedura iterativa che, partendo dal modello "nullo" senza alcuna covariata selezionata, ad ogni passo prevede l'inserimento o l'eliminazione di predittori nel modello di regressione sulla base del contributo di ciascuno alla spiegazione della variabilità della variabile dipendente *Y*, fino al raggiungimento di un prefissato criterio di arresto. Il processo di selezione può prevedere anche la rimozione di variabili precedentemente inserite nel modello. La procedura PROC REG del SAS utilizza come criterio di selezione il valore della statistica *F*: il processo finisce quando tutte le variabili nel modello producono una statistica *F* significativa, mentre nessuna delle variabili fuori dal modello produce una statistica *F* significativa.

Nei modelli di regressione che usano come variabile dipendente la variabile *Redditi pro-capite* (in scala logaritmica) risultano significative per il 2012 le variabili: *Superficie del Comune* e *Ripartizione geografica* di appartenenza, mentre nel modello in cui la variabile dipendente è la spesa complessiva per tutti i dipendenti (*Redditi*) risultano significative anche le variabili *Metropoli* e *Numero di dipendenti*. Questo suggerisce che, una volta considerato il reddito pro-capite (rispetto ai dipendenti), la dimensione del Comune in termini di popolazione non incide sulla variabilità del

reddito. Alla luce di questi risultati, per la voce *Redditi* si è scelto di utilizzare le seguenti variabili ausiliarie: *Superficie*, *Ripartizione geografica* e *Popolazione*.

Figura 1 – Box-plot della distribuzione della variabile *Redditi pro-capite* su tutte le unità (anno 2012)

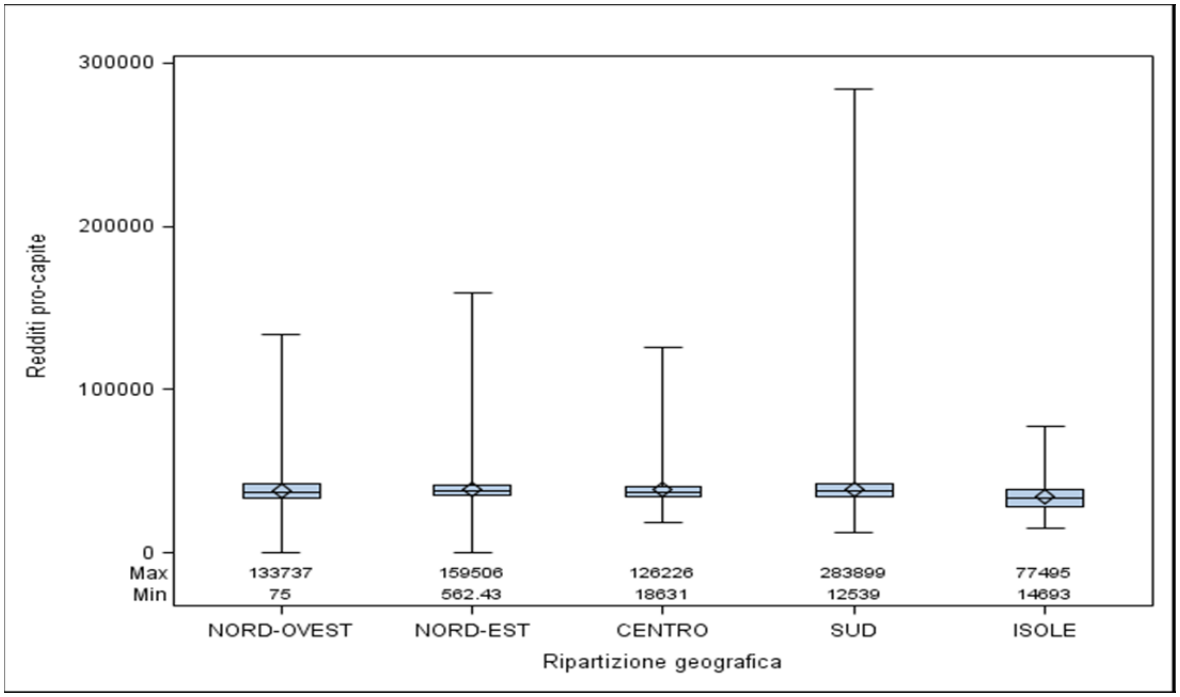
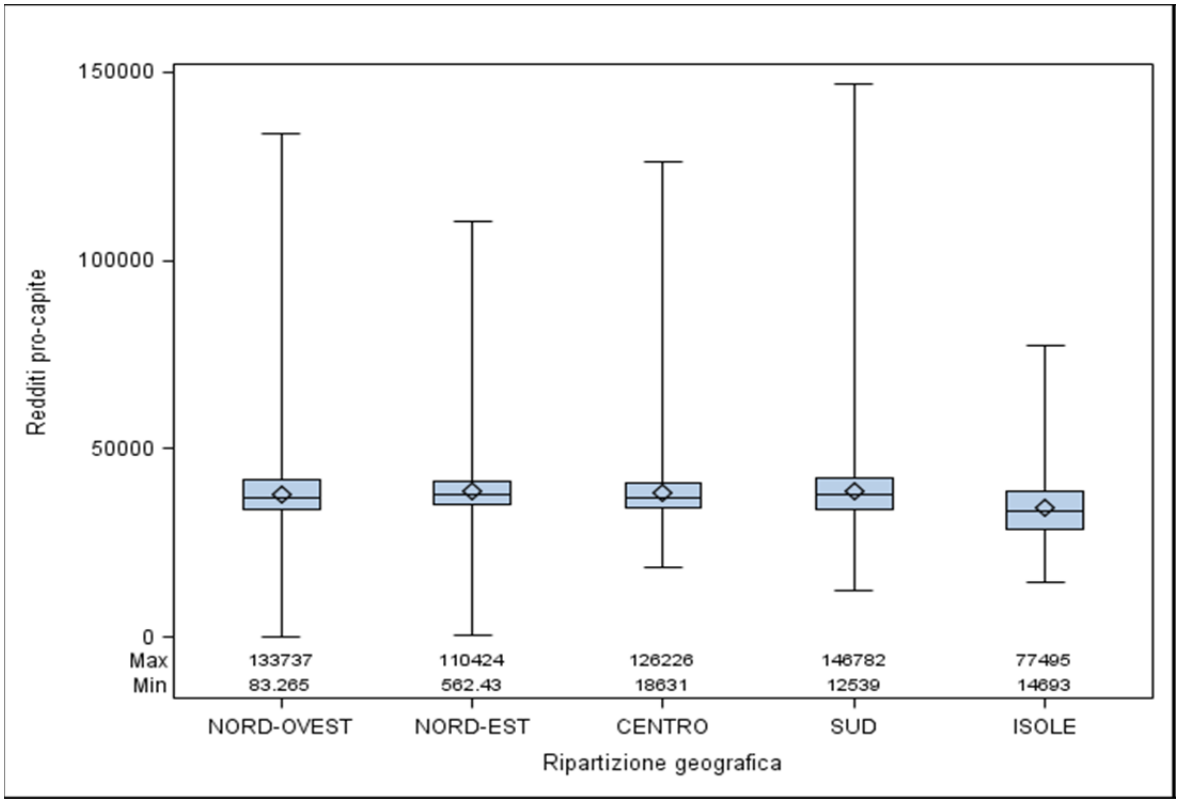


Figura 2 – Box-plot della distribuzione della variabile *Redditi pro-capite* escluse le unità con valori anomali (anno 2012)



Si noti che si è scelto di utilizzare comunque nel processo di imputazione la variabile *Popolazione* - nonostante questa non sia risultata significativa nell'analisi effettuata tramite modelli di regressione e sia molto correlata con la variabile *Numero di dipendenti* - per molteplici motivi:

- l'informazione relativa alla popolazione residente è aggiornata di anno in anno, a differenza della variabile sul numero dei dipendenti che attualmente non è disponibile con questa periodicità;
- l'informazione proveniente dalla variabile *Numero di dipendenti* entra già in alcuni dei modelli di imputazione attraverso l'utilizzo del *Reddito pro-capite*;
- è poco probabile che vi siano dati mancanti nei vari anni considerati.

In alcuni metodi la variabile *Popolazione* è stata utilizzata a livello di microdato; in altri modelli si è proceduto a una sua riduzione in classi (cercando di evitare celle di imputazione troppo "povere"): per la determinazione delle classi sono state valutate diverse alternative, utilizzate in diversi contesti Istat relativi all'analisi dei bilanci delle Amministrazioni Comunali, in modo da evitare classificazioni troppo arbitrarie.

5.2 La validazione dei dati

Preliminarmente alla fase di imputazione si è proceduto alla verifica della qualità della fonte CRB in termini di coerenza statistica dei dati. L'obiettivo era validare l'informazione proveniente dall'archivio in termini di presenza di possibili errori di misurazione. A tale scopo sono stati effettuati dei controlli di coerenza statistica fra la variabile obiettivo e la variabile *Numero di dipendenti*, per particolari domini. I casi risultati anomali sono stati trattati oppure temporaneamente esclusi dal processo di imputazione (ad es. per la stima dei parametri dei modelli o dal serbatoio dei potenziali donatori). Dall'analisi puntuale dei casi anomali è emerso che questi spesso sono dovuti o ad un'errata interpretazione da parte degli Enti della definizione di Dipendenti (ad esempio includendo erroneamente personale inquadrato in altre tipologie contrattuali), oppure alla mancata verifica di coerenza con le corrispondenti spese di personale in fase di compilazione delle scritture contabili.

Di seguito sono riportate le principali casistiche trattate:

a) Comuni con numero di dipendenti nullo o mancante. Al fine di calcolare la variabile *Redditi pro-capite*, è stata effettuata l'imputazione del numero di dipendenti con una tecnica da donatore di distanza minima per classi; quest'ultime sono state definite in termini di *Ripartizione geografica*, mentre le variabili di *matching* coinvolte nel processo sono state sono state *Popolazione* 2011 e 2012, *Superficie comunale*, *Redditi* 2011 e 2012.

b) Comuni con valore nullo della variabile *Redditi*. Essendo tali valori risultati non errati nei 15 casi sottoposti a verifiche puntuali, sono stati lasciati invariati, escludendoli però sia dal serbatoio dei possibili donatori, sia dal sottoinsieme di valori utilizzati per la stima dei modelli di imputazione (regressioni, mediane).

c) Comuni con valore di *Redditi pro-capite* anomalo. L'individuazione di questi Comuni è stata effettuata sulla base di modelli di regressione del tipo:

$$Y = b_0 + b_1 Ndip + e \qquad (10)$$

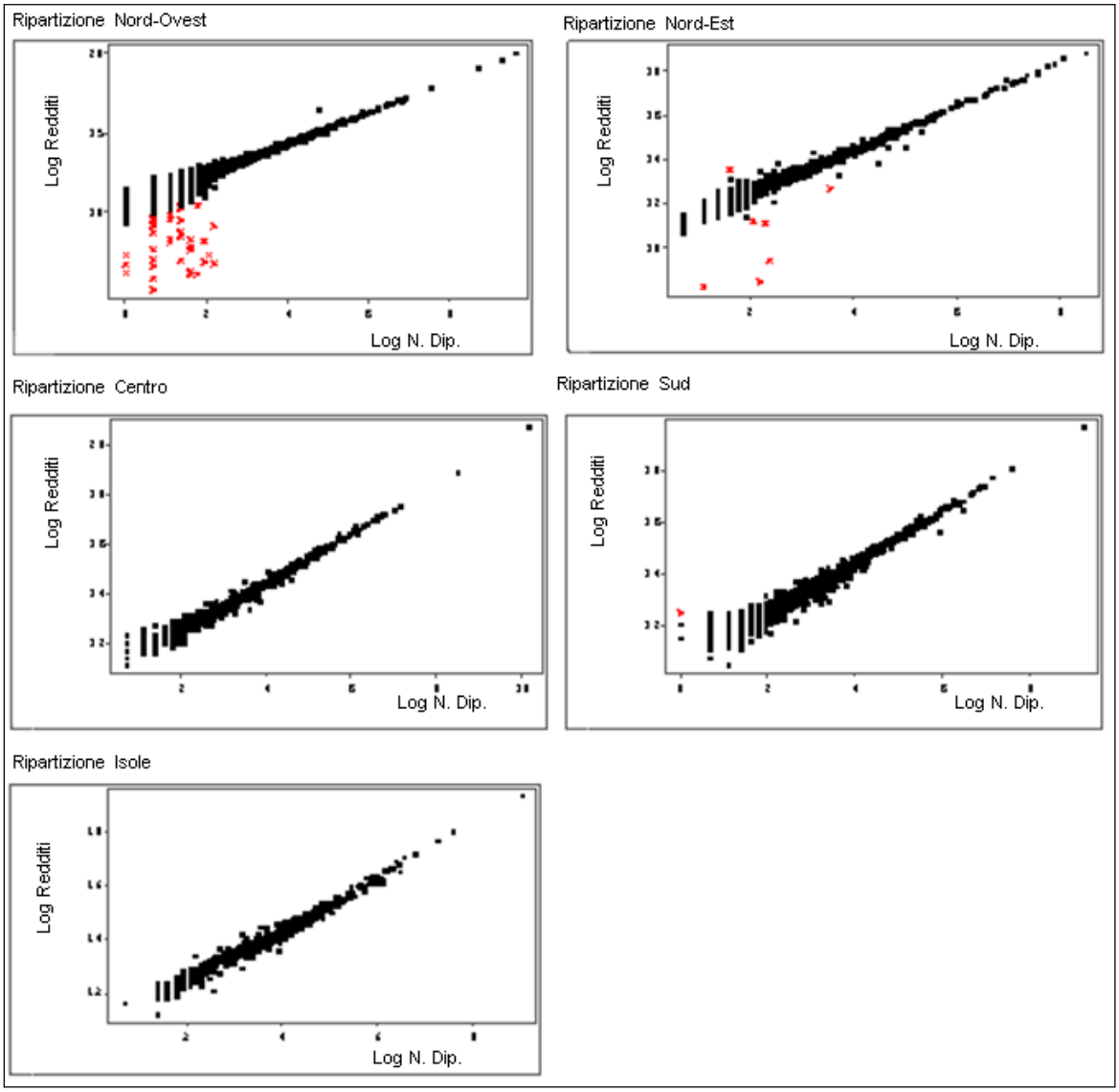
stimati in modo robusto, utilizzando un algoritmo iterativo (disponibile in ambiente SAS) che assicura una stima dei parametri di regressione non influenzata dalla presenza di *outlier* nei dati[5]. In estrema sintesi, la stima dei parametri del modello avviene escludendo le unità con residui superiori ad un certo valore di soglia *k* (*cutoff*) che agisce sulla distribuzione dei residui del modello secondo la regola:

[5] Metodo dei Least Trimmed Squares (LTS; *Rousseeuw et al., 2000).*

$$Y_i \equiv outlier \quad se \ |e_i| > ks^2 . \qquad (11)$$

Uno dei vantaggi dell'utilizzo dei modelli di regressione robusta risiede proprio nel fatto che essi consentono di identificare i casi anomali contestualmente alla fase di stima (robusta) dei parametri di regressione, al livello di tolleranza scelto (rappresentato da un intorno di ampiezza *k*). Tali modelli sono stati stimati per *Ripartizione geografica*, ed hanno portato all'individuazione di 45 Comuni (0,7%) con valori pro-capite osservati sensibilmente diversi in valore assoluto dai corrispondenti valori stimati dal modello di regressione. Negli *scatter-plot* della Figura 3 sono evidenziati i Comuni anomali per *Ripartizione geografica*. I Comuni selezionati, sottoposti a revisione manuale, sono stati comunque esclusi sia dal serbatoio dei possibili donatori, sia dal sottoinsieme di valori utilizzati per la stima dei modelli di imputazione.

Figura 3 – Scatter-plot delle variabili *Redditi* e *Numero di dipendenti* (scala logaritmica) (anno 2012)



5.3 Metodi di imputazione per la variabile *Redditi*

Sulla base delle analisi illustrate nei paragrafi precedenti, i metodi di imputazione descritti nel paragrafo 4 sono stati applicati alla variabile *Redditi* nel modo descritto di seguito.

a) *Donatore di distanza minima trasversale per classi* (*NND*).
Variabile oggetto di imputazione è *Redditi pro-capite* (calcolata rispetto al numero di dipendenti comunali nel 2011): $Y_{i,pc} = Y_i / Ndp_i$. Una volta individuato il donatore di distanza minima all'interno della cella di imputazione cui appartiene il Comune mancante, la variabile obiettivo viene ottenuta mediante la relazione:

$$Y_i^* = Ndip_i \times Y_{d,pc} = Ndip_i \times \frac{Y_d}{Ndip_d} \qquad (12)$$

dove Y_d e $Ndip_d$ sono i valori rispettivamente delle variabili *Redditi* e *Numero di dipendenti* del Comune donatore *d*, mentre Y_i^* è il valore imputato della variabile obiettivo nel Comune mancante *i*. Nell'applicare la procedura sono utilizzate le variabili *Redditi SIOPE 2012 pro-capite* (rispetto al Numero di dipendenti), *Superficie*, *Popolazione 2012* come variabili di *matching*; in qualità di variabile di classificazione la *Ripartizione geografica*.

b) *Predictive Mean Matching per classi* (*PMM*).
Il metodo è stato applicato stimando il modello regressione multivariato per ogni cella di imputazione:

$$Y_i^p = \alpha + \beta_1 S_{i,2012} / Ndip_i + \beta_2 Pop_{i,2012} + \beta_3 Superficie_i + e_i \qquad (13)$$

dove con $S_{i,2012}$ si intende il valore della variabile relativa ai *Redditi* presente nell'archivio SIOPE dell'anno 2012. L'imputazione della variabile obiettivo, laddove mancante, viene effettuata con donatore di distanza minima rispetto al valore predetto Y_i^p . Le celle di imputazione sono determinate in base alla variabile di classificazione *Ripartizione geografica*.

c) *Donatore di distanza minima longitudinale per classi* (*NND Long*).
In questo caso, i pro-capite della variabile *Redditi* (calcolati rispetto al *Numero di dipendenti*) sono imputati con metodo del donatore di distanza minima per classi utilizzando le seguenti informazioni: *Ripartizione geografica* per la formazione degli strati; informazioni di tipo longitudinale, quali $Y_{2011} / Ndip$, $S_{2011} / Ndip$, S_{2012} / S_{2011} , Pop_{2012} , *Superficie*, come variabili di *matching*.

d) *Metodi deterministici di tipo longitudinale.*
Si illustrano di seguito i 4 metodi oggetto di sperimentazione afferenti a questa classe di metodologie. L'applicazione dell'uno o dell'altro metodo è condizionata alla disponibilità delle informazioni ausiliarie e/o storiche. Tutti i metodi sono stati utilizzati stratificando in base alle variabili *Ripartizione geografica* e *Popolazione* 2012 (in classi[6]).

[6] Classe 0 = < 1500; classe 1 = [1500,5000); classe 2 = [5000,10000); classe 3 = [10000,60000); classe 4 = [60000,100000); classe 5 = > 100000.

– Metodo *Long CRB Sio*;
Se esiste il valore storico della variabile obiettivo $Y_{i,2011}$, allora:

$$Y_{i,2012} = Y_{i,2011} \times mediana(rappCRB)_s \qquad (14)$$

dove $mediana(rappCRB)_s$ è il valore mediano nello strato *s* della distribuzione del rapporto $Y_{i,2012} / Y_{i,2011}$. Nei casi in cui non è disponibile il valore della variabile $Y_{i,2011}$ ma esiste l'informazione in SIOPE nei due anni considerati, si ha:

$$Y_{i,2012} = Y_{i,2011}^{p} \times rappSIOPE_i \qquad (15)$$

dove $rappSIOPE_i$ è il rapporto $S_{i,2012} / S_{i,2011}$ nella fonte SIOPE, e $Y_{i,2011}^{p}$ è il valore della variabile $Y_{i,2011}$ previsto attraverso il modello di regressione robusta stimato per ogni ripartizione geografica: $Y_{i,2011} = \alpha + \beta\, S_{i,2011}$. Nei casi residuali:

$$Y_{i,2012} = mediana(Y_{i,2012})_s \qquad (16)$$

dove $mediana(Y_{i,2012})_s$ è il valore mediano nello strato *s* della distribuzione della variabile obiettivo $Y_{i,2012}$.

– Metodo *Long CRB Pop*;
Come nel metodo precedente, se esiste il valore della variabile $Y_{i,2011}$ nella fonte CRB, si applica la formula (14). Negli altri casi, si ha:

$$Y_{i,2012} = mediana(Y_{i,2011})_s \times rappPOP_i \qquad (17)$$

dove $mediana(Y_{i,2011})_s$ rappresenta il valore mediano della distribuzione della variabile $Y_{i,2011}$ nello strato *s* e $rappPOP_i$ è dato dal rapporto $Pop_{i,2012} / Pop_{i,2011}$ per il Comune *i*.

– Metodo *Long Pop*;
Se esiste il valore della variabile $Y_{i,2011}$ allora:

$$Y_{i,2012} = Y_{i,2011} \times rappPOP_i \qquad (18)$$

Negli altri casi si applica la formula (17).

– Metodo *Long Sio;*
Se esiste il valore della variabile obiettivo $Y_{i,2011}$ e vi è l'informazione in SIOPE nei due anni considerati, allora:

$$Y_{i,2012} = Y_{i,2011} \times rappSIOPE_i \qquad (19)$$

Qualora invece l'informazione in SIOPE nei due anni è mancante, si ha:

$$Y_{i,2012} = Y_{i,2011} \times mediana(rappSIOPE)_s \qquad (20)$$

Nei casi in cui non esiste il valore della variabile $Y_{i,2011}$ ma esiste l'informazione in SIOPE nei due anni considerati, si applica la formula (15), nei casi residuali la formula (16).

e) *Metodi "misti" di tipo longitudinale.*
Sono state sperimentate diverse tecniche di imputazione afferenti a questa classe di metodi. Ognuna di esse prevede l'esecuzione di un primo passo di tipo deterministico, nel quale il valore della variabile $Y_{i,2011}$ viene aggiornato utilizzando l'informazione ausiliaria disponibile in ottica longitudinale (vedi metodi descritti al punto d) precedente). La stratificazione adottata è basata sull'incrocio delle variabili *Ripartizione geografica* e *Popolazione 2012* (in classi). La procedura viene completata mediante un secondo passo di imputazione, non deterministica, basata sul metodo del donatore di distanza minima longitudinale per classi (metodo *NND Long* descritto al punto c) precedente), definite secondo la *Ripartizione geografica*. I tre metodi proposti si distinguono per il tipo di informazione ausiliaria utilizzata nel primo passo (Popolazione, SIOPE, CRB):

– Metodo *NND Long Misto Pop*

$$Y_{i,2012} = Y_{i,2011} \times rappPOP_i \qquad (21)$$

– Metodo *NND Long Misto CRB*

$$Y_{i,2012} = Y_{i,2011} \times mediana(rappCRB)_s \qquad (22)$$

– Metodo *NND Long Misto SIO*

$$Y_{i,2012} = Y_{i,2011} \times rappSIOPE_i \qquad (23)$$

5.4 Risultati della simulazione per la variabile *Redditi*

In questo paragrafo sono descritti i risultati della sperimentazione per la variabile *Redditi*. La Tavola 7 contiene i valori degli indicatori per ciascuno dei 10 metodi esaminati, ottenuti sulla base di *k=1.000* iterazioni del processo di simulazione illustrato nel paragrafo 4.

Tavola 7 – Variabile *Redditi*, indicatori di qualità per metodo di imputazione. Livello Italia (anno 2012)

Indice	Metodi di imputazione									
	NND	PMM	NND long	Long CRB Sio	Long CRB Pop	Long Pop	Long Sio	NND long Misto Pop	NND long Misto CRB	NND long Misto Sio
RB	0,045	0,366	0,044	-0,004	0,023	-0,320	-0,021	-0,354	-0,012	-0,028
RMSE	0,362	0,981	0,361	0,182	0,192	0,355	0,103	0,383	0,182	0,105
RIE	0,240	0,433	0,237	0,108	0,122	0,130	0,075	0,118	0,108	0,075

Come si può osservare, i metodi più efficaci in termini di RMSE sono quelli deterministici longitudinali. Fra questi, tenendo conto anche dei valori degli altri due indicatori, sono da preferire i metodi *Long Sio* e *NND long Misto Sio*, che sfruttano entrambi l'informazione ausiliaria contenuta nella fonte SIOPE.

La Tavola 8 contiene i valori dell'RMSE calcolato a livello regionale, al fine di evidenziare eventuali inefficienze dei metodi sperimentati a livello territoriale. Si conferma una prestazione migliore su tutte le regioni dei metodi longitudinali, in particolare di quelli che utilizzano le informazioni da fonte SIOPE.

Tavola 8 – Variabile Redditi, indicatore Relative MSE. Livello regionale (anno 2012)

Regione	Metodi di imputazione									
	NND	PMM	NND long	Long CRB Sio	Long CRB Pop	Long Pop	Long Sio	NND long Misto Pop	NND long Misto CRB	NND long Misto Sio
Piemonte	0,043	0,077	0,042	0,015	0,016	0,049	0,013	0,049	0,016	0,013
Valle D'Aosta	0,001	0,001	0,003	0,001	0,001	0,001	0,000	0,001	0,001	0,000
Lombardia	0,063	0,124	0,063	0,044	0,043	0,121	0,098	0,122	0,044	0,098
Trentino	0,026	0,010	0,026	0,007	0,044	0,041	0,009	0,003	0,008	0,009
Veneto	0,033	0,019	0,032	0,008	0,009	0,020	0,010	0,019	0,008	0,010
Friuli-Venezia Giulia	0,013	0,014	0,014	0,003	0,003	0,007	0,017	0,007	0,003	0,017
Liguria	0,025	0,034	0,025	0,004	0,004	0,022	0,004	0,022	0,004	0,004
Emilia-Romagna	0,073	0,024	0,071	0,029	0,028	0,050	0,017	0,050	0,029	0,017
Toscana	0,043	0,023	0,041	0,010	0,010	0,046	0,013	0,046	0,010	0,013
Umbria	0,012	0,008	0,012	0,006	0,006	0,016	0,008	0,016	0,006	0,008
Marche	0,005	0,009	0,004	0,005	0,005	0,008	0,005	0,008	0,005	0,005
Lazio	0,331	0,928	0,330	0,174	0,176	0,031	0,019	0,031	0,174	0,019
Abruzzo	0,010	0,010	0,009	0,003	0,003	0,009	0,003	0,008	0,003	0,003
Molise	0,003	0,006	0,003	0,003	0,004	0,005	0,001	0,005	0,003	0,001
Campania	0,078	0,126	0,080	0,046	0,049	0,084	0,018	0,087	0,048	0,015
Puglia	0,022	0,019	0,022	0,007	0,008	0,019	0,010	0,018	0,007	0,009
Basilicata	0,009	0,011	0,009	0,004	0,004	0,007	0,003	0,007	0,004	0,003
Calabria	0,016	0,014	0,012	0,004	0,004	0,012	0,011	0,012	0,004	0,011
Sicilia	0,100	0,112	0,100	0,019	0,020	0,039	0,020	0,040	0,019	0,020
Sardegna	0,012	0,008	0,011	0,008	0,009	0,014	0,006	0,012	0,008	0,006

5.4.1 Gli aggregati regionali

In questo paragrafo si presentano alcune analisi di confronto fra le stime della variabile *Redditi* ottenute sulla base dei metodi *Long Sio* e *NND long Misto Sio* (risultati migliori rispetto alle assunzioni fatte e agli indicatori utilizzati nella sperimentazione, vedere Tavola 8) e le corrispondenti stime Istat (Istat, 2014). Relativamente a queste ultime, *"la stima dei valori dell'universo dei Comuni viene ottenuta sulla base della popolazione residente al 31/12/2012, tramite coefficienti di espansione calcolati per ciascuna classe di popolazione residente di ciascuna regione"* (Istat, 2014).

Nella Tavola 9 sono riportate le stime regionali Istat (colonna "Bilanci"), le corrispondenti stime regionali ottenute aggregando i dati comunali imputati con i due metodi ottimali (colonne *Long Sio* e *NND long Misto Sio*) e le differenze assolute fra esse. Nella Tavola 10 sono riportate le differenze relative percentuali fra aggregati/stime regionali e le corrispondenti stime calcolate sui dati "originali" (ottenuti cioè sui dati affetti da dati mancanti). La tavola riporta, per una migliore analisi dei risultati, anche informazioni relative al numero di Comuni, alla popolazione e al numero di dipendenti dei Comuni osservati e imputati per ogni regione.

Si può osservare che con entrambi i metodi di imputazione le variazioni a livello di aggregati regionali rispetto ai corrispondenti aggregati ufficiali (Bilanci Istat) sono molto contenuti, con valori massimi per Piemonte, Abruzzo e Campania (Bolzano, con il valore massimo superiore al 2%, non è pubblicato separatamente in I.Stat). La differenza in valore assoluto, a livello Italia, fra stime imputate e stime da Bilanci Istat è pari a 38.317.284 (metodo *Long Sio*) e 31.212.445 (metodo *NND Long Misto Sio*).

Tavola 9 – Variabile *Redditi*, differenze assolute stime, per Regione (anno 2012)

Regione	Stime Totale Redditi			Differenze assolute	
	Long Sio	NND long Misto Sio	Bilanci	Bilanci-Imputati Long Sio	Bilanci-Imputati NNDlong Misto Sio
Piemonte	1.131.633.481	1.134.750.561	1.145.799.215	14.165.734	11.048.654
Valle D'Aosta	59.539.719	59.564.785	60.092.810	553.091	528.025
Lombardia	2.226.584.118	2.226.590.729	2.229.580.555	2.996.437	2.989.826
Veneto	1.014.192.672	1.014.187.518	1.010.967.062	-3.225.610	-3.220.456
Friuli-Venezia Giulia	398.380.155	398.380.155	396.918.512	-1.461.643	-1.461.643
Liguria	529.067.020	529.017.654	530.517.649	1.450.629	1.499.995
Emilia-Romagna	1.127.492.984	1.127.501.674	1.124.726.845	-2.766.139	-2.774.829
Toscana	1.021.791.847	1.021.791.847	1.022.274.967	483.120	483.120
Umbria	223.015.409	223.015.409	223.919.582	904.173	904.173
Marche	376.004.425	376.004.425	376.448.950	444.525	444.525
Lazio	1.675.847.016	1.676.263.654	1.672.838.079	-3.008.937	-3.425.575
Abruzzo	286.233.621	286.097.451	290.548.442	4.314.821	4.450.991
Molise	73.927.888	73.953.293	73.427.868	-500.020	-525.425
Campania	1.379.581.811	1.380.162.731	1.406.756.011	27.174.200	26.593.280
Puglia	701.188.054	701.497.826	697.660.489	-3.527.565	-3.837.337
Basilicata	139.496.986	139.699.772	140.527.520	1.030.534	827.748
Calabria	455.261.075	455.643.891	458.823.092	3.562.017	3.179.201
Sicilia	1.687.192.107	1.689.848.030	1.679.714.993	-7.477.114	-10.133.037
Sardegna	440.946.765	441.277.321	439.536.886	-1.409.879	-1.740.435
Bolzano	194.787.845	193.909.909	198.947.113	4.159.268	5.037.204
Trento	228.308.052	228.419.252	228.763.694	455.642	344.442
Totale	15.370.473.050	15.377.577.889	15.408.790.334	38.317.284	31.212.445

Tavola 10 – Variabile *Redditi*, differenze relative percentuali fra stime, percentuale Comuni, popolazione, dipendenti, per Regione (anno 2012)

Regione	Differenze percentuali					Percentuali imputati		
	Originali-Imputati Long Sio	Originali-Imputati NND long Misto Sio	Bilanci-Imputati Long Sio	Bilanci-Imputati NND long Misto Sio	Originali-Bilanci	Comuni	Popolazione	Dipendenti
Piemonte	3,16	3,43	1,25	0,97	4,36	3,57	3,35	3,37
Valle D'Aosta	4,58	4,62	0,93	0,89	5,46	6,76	5,45	4,23
Lombardia	1,48	1,48	0,13	0,13	1,62	4,08	2,17	1,48
Veneto	1,57	1,57	0,32	0,32	1,26	4,13	1,57	1,55
Friuli-Venezia Giulia	1,26	1,26	0,37	0,37	0,89	5,50	1,13	1,25
Liguria	2,90	2,89	0,27	0,28	3,17	10,64	3,43	2,88
Emilia-Romagna	0,41	0,41	0,25	0,25	0,16	2,01	0,65	0,40
Toscana	0,59	0,59	0,05	0,05	0,64	1,74	0,69	0,57
Umbria	1,45	1,45	0,41	0,41	1,85	1,09	2,12	1,41
Marche	0,54	0,54	0,12	0,12	0,66	1,67	0,74	0,52
Lazio	4,28	4,31	0,18	0,20	4,11	18,52	5,96	4,67
Abruzzo	7,17	7,12	1,51	1,56	8,54	13,11	8,42	6,93
Molise	15,47	15,50	0,68	0,71	14,89	14,71	14,88	17,19
Campania	7,95	7,98	1,97	1,93	9,72	15,25	10,59	8,01
Puglia	5,49	5,53	0,50	0,55	5,01	13,18	5,14	5,79
Basilicata	6,07	6,21	0,74	0,59	6,76	9,92	5,38	6,36
Calabria	15,58	15,65	0,78	0,70	16,24	18,58	16,04	15,24
Sicilia	21,55	21,67	0,44	0,60	21,20	23,33	21,23	21,92
Sardegna	6,92	6,99	0,32	0,39	6,62	13,00	6,39	6,92
Bolzano	13,79	13,40	2,14	2,60	15,59	16,38	15,55	13,81
Trento	4,97	5,01	0,20	0,15	5,16	9,22	5,47	4,90
Totale	5,77	5,82	0,25	0,20	6,01	8,71	5,91	6,22

Alla luce di questi risultati, si può concludere che per la voce *Redditi da lavoro dipendente* il passaggio dal metodo di stima corrente ad un nuovo metodo di imputazione longitudinale a livello comunale determinerebbe i seguenti vantaggi:

- non risultano variazioni significative sulle stime regionali; le differenze assolute a livello Italia sono intorno ai 30-40 milioni di euro;
- disporre di previsioni a livello di singola amministrazione delle variabili chiave del bilancio, in base alle quali sarà quindi possibile ricostruire le voci di dettaglio del bilancio stesso (ad es., attraverso metodi basati su proporzioni medie/donatori di proporzioni);
- garantire che la serie di valori di un Comune, a partire da un dato osservato, risulti coerente in termini longitudinali per costruzione, variando sulla base di un *trend* calcolato sui dati del Comune stesso da fonte SIOPE;
- estendibilità dell'approccio ad altre variabili chiave (capi-conto) del bilancio.

6. Stima della voce *Consumi intermedi*

In questa sezione sono descritti i modelli utilizzati e i risultati ottenuti per la variabile *Consumi intermedi* (*Consumi* nel seguito), misurata, in maniera diretta (variabile *Y*) con criterio di competenza nel Quadro 4 dei CRB, tramite la somma delle voci relative al Totale delle Spese correnti-impegni: *Acquisto di beni di consumo e/o di materie prime*, *Prestazioni di servizi* e *Utilizzo di beni di terzi*.

La variabile *Z*, data dalla somma delle voci del CRB relative al Totale delle Spese correnti-

pagamenti in conto competenza e in conto residui: *Acquisto di beni di consumo e/o di materie prime*, *Prestazioni di servizi* e *Utilizzo di beni di terzi* e la variabile *S*, data dalla somma di numerose voci dell'archivio SIOPE (vedere Appendice 2), misurano i consumi intermedi con criterio di cassa, e sono relative a spese concettualmente confrontabili: tali variabili derivate saranno pertanto utilizzate come informazioni ausiliarie ai fini dell'imputazione della variabile dipendente *Y*.

Analogamente a quanto fatto per la variabile *Redditi,* nel paragrafo 6.1 sono illustrate alcune analisi di tipo esplorativo aventi l'obiettivo di valutare il contenuto informativo di queste variabili rispetto alla variabile oggetto di imputazione (*Consumi* o *Consumi pro-capite*).

6.1 Analisi esplorativa dei dati

La Tavola 11 riporta i *pattern* di non risposta per le variabili *Y, Z* e *S*, nelle fonti CRB e SIOPE, per i due anni considerati. Il valore "1" indica la presenza dell'informazione, mentre il valore "0" ne indica l'assenza. Come per la variabile *Redditi*, l'informazione da CRB sulle singole voci è o tutta presente o tutta assente, mentre l'informazione proveniente da SIOPE ha *pattern* diversi di risposta parziale.

Tavola 11 – Variabili Consumi, *pattern* di dati mancanti nel 2011 e 2012

CRB 2012	CRB 2011	SIOPE 2012	SIOPE 2011	N.	%
1	1	1	1	7.132	88,14
1	1	0	0	79	0,98
1	0	1	1	165	2,04
1	0	0	0	11	0,14
0	1	1	1	534	6,60
0	1	0	0	3	0,04
0	0	1	1	167	2,06
0	0	0	0	1	0,01

Il numero totale di osservazioni con informazione completa nei due anni è pari a 7.132. Rispetto alla variabile *Redditi* rimane invariato il numero di Comuni per cui è necessario procedere all'imputazione della variabile *Consumi* nel 2012, pari a 705, ovvero l'8,7% del totale delle Amministrazioni Comunali della Lista S13.

Il contenuto informativo delle variabili *Y, Z* e *S* risulta essere molto correlato. Considerando di volta in volta solo il sottoinsieme di unità con presenza delle informazioni, sono state calcolate le differenze relative percentuali tra *Y* e *Z* e tra *S* e *Z* utilizzando i semplici indici già espressi nelle formule (8) e (9). La Tavola 12 mostra alcune statistiche descrittive delle distribuzioni delle variabili $Diff_{YZ}$ e $Diff_{SZ}$ per il 2012. A differenza di quanto osservato per i *Redditi*, la variabile *Consumi* presenta differenze più marcate fra la variabile *Y* e la variabile *Z*, dovuta alla forte eterogeneità della variabile in analisi, mentre si conferma un buon accostamento fra le variabili *Z* e *S* che forniscono una misurazione dei *Consumi* con lo stesso criterio (cassa).

Tavola 12 – Variabile Consumi, statistiche descrittive delle distribuzioni di $Diff_{i,YZ}$ e $Diff_{i,SZ}$ (anno 2012)

Variabile	N	Media	Min	Max	Mediana	Dev. Std.
$Diff_{YZ}$	7387	8,08	-78,89	545,88	4,82	20,49
$Diff_{SZ}$	7297	-0,02	-76,53	43,24	0	1,74

Le Figure 4 e 5 mostrano i *box-plot* della variabile *Consumi pro-capite* (*Y/Popolazione*) costruiti, rispettivamente, considerando tutte le unità e rimuovendo dall'insieme delle unità quelle con valori risultati "anomali" da un'analisi grafica di tipo esplorativo. La presenza di quest'ultimi è concentrata soprattutto nelle ripartizioni Nord-Ovest e Sud. La variabilità delle distribuzioni si mantiene elevata anche in assenza di *outliers*.

Figura 4 – Box-plot della distribuzione della variabile *Consumi pro-capite* su tutte le unità (anno 2012)

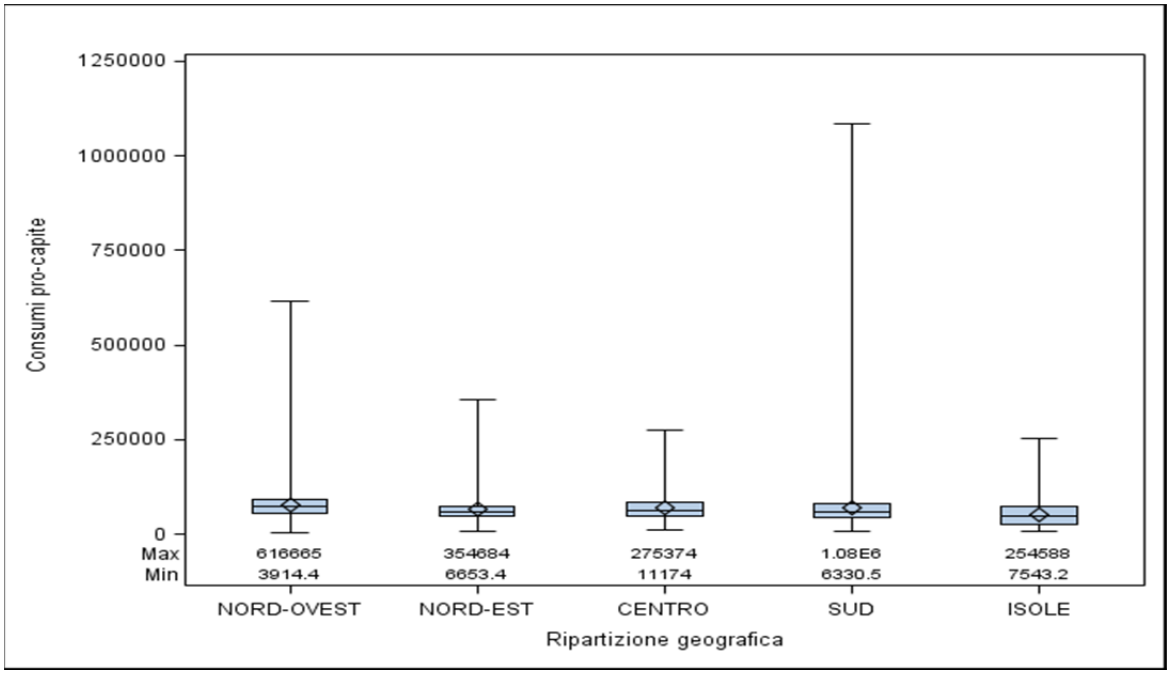
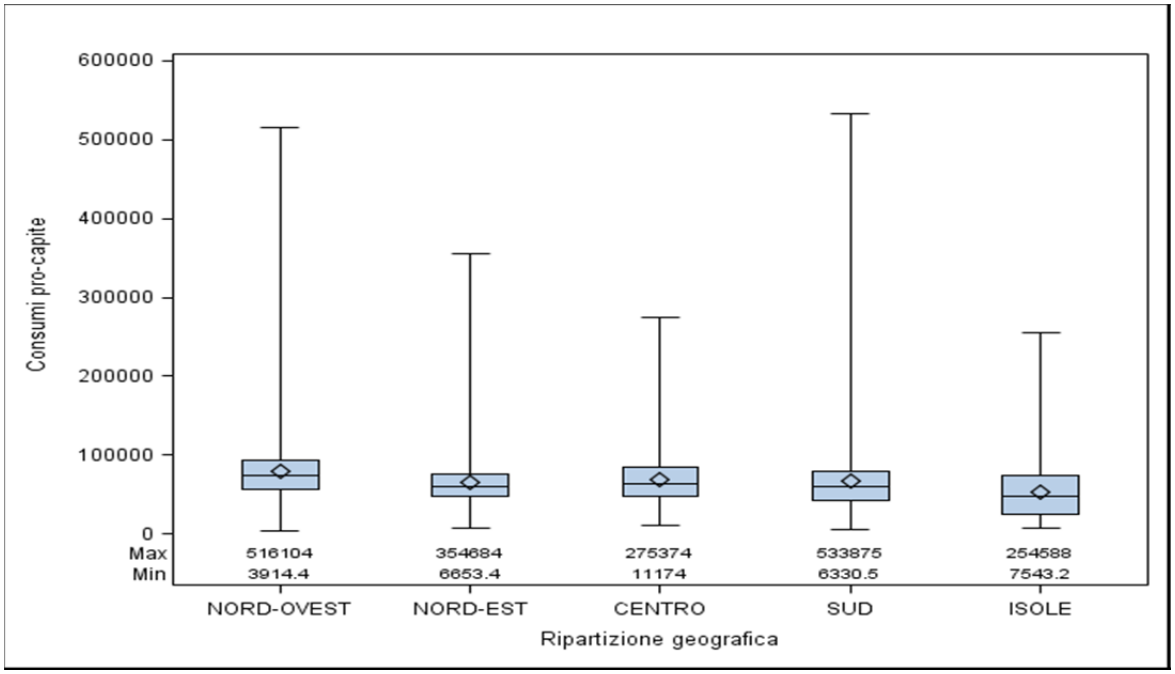


Figura 5 – Box-plot della distribuzione della variabile Consumi pro-capite escluse le unità con valori anomali (anno 2012)



Nel procedimento di scelta delle informazioni utilizzabili in qualità di variabili ausiliarie in fase di imputazione sono state considerate le stesse variabili esplorate nel caso della variabile *Redditi* (paragrafo 5.1). Anche in questo caso, per la selezione delle covariate statisticamente significative sono stati utilizzati modelli di regressione lineare con procedura *stepwise.*

Nei modelli di regressione che usano come variabile dipendente la variabile *Consumi pro-capite* (in scala logaritmica), dove il pro-capite è calcolato rispetto alla popolazione comunale, risultano significative le variabili *Ripartizione Geografica, Superficie, Zona altimetrica.*

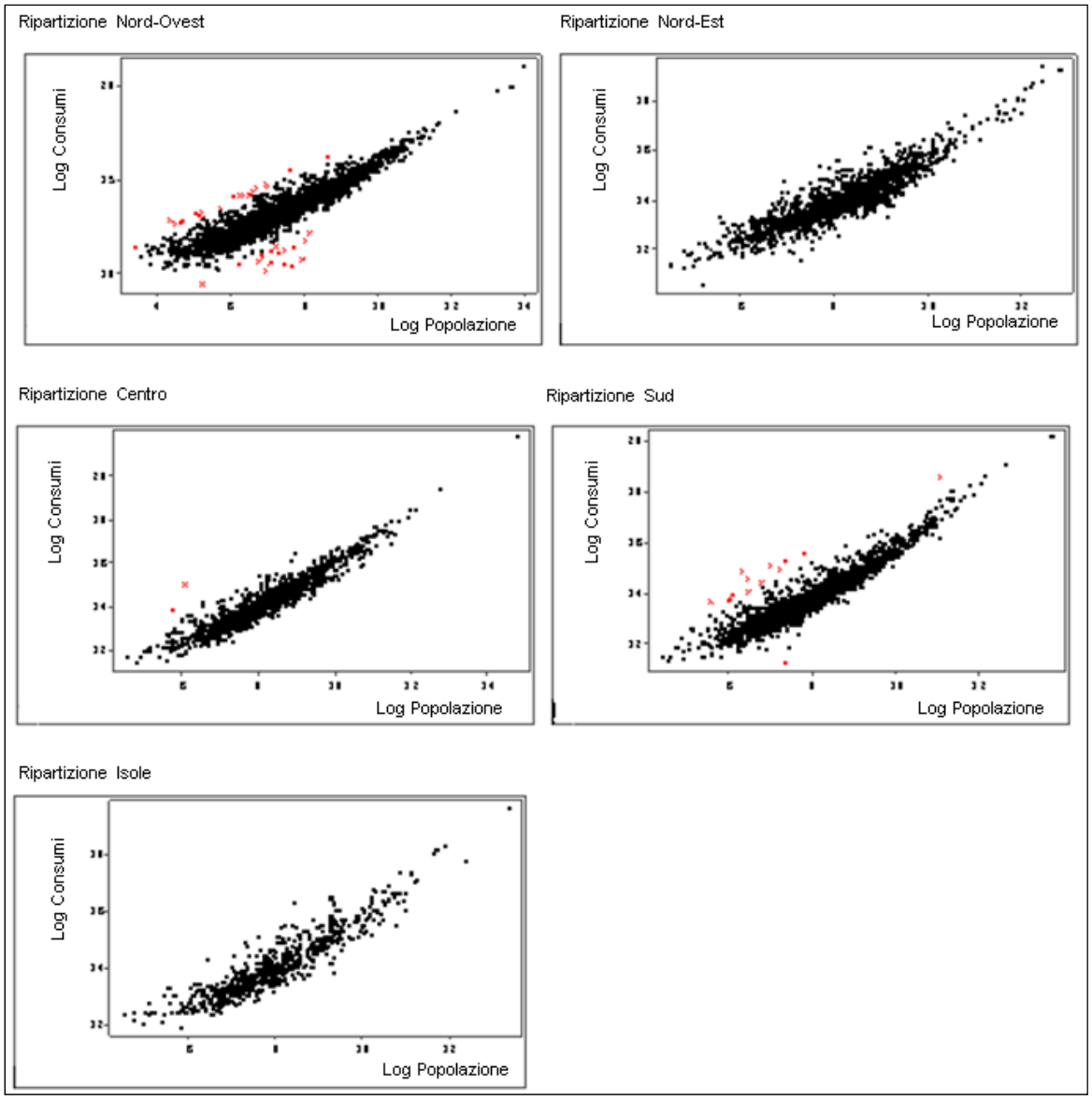
6.2 La validazione dei dati

Preliminarmente alla fase di imputazione, è stata verificata la presenza di possibili errori di misurazione sulla variabile obiettivo nei dati di fonte CRB. A tale scopo, sono stati individuati i Comuni con valore di *Consumi pro-capite* anomalo (rispetto alla popolazione) utilizzando, come per la variabile *Redditi*, modelli di regressione robusti del tipo (vedere paragrafo 5.2):

$$Y = \beta_0 + \beta_1 Pop_{2012} + \varepsilon \qquad (24)$$

separatamente per *Ripartizione geografica* e utilizzando prefissati valori di *cut-off*.

Figura 6 – Scatter-plot delle variabili *Consumi* e *Popolazione* (scala logaritmica) (anno 2012)



La procedura ha portato all'individuazione di 55 Comuni (0,7%) con valori pro-capite sensibilmente distanti in valore assoluto dai corrispondenti valori stimati dal modello di regressione. Negli

scatter-plot della Figura 6 sono evidenziati i Comuni anomali, per *Ripartizione geografica*. I Comuni selezionati, sottoposti a revisione manuale per la verifica delle possibile presenza di errori di misurazione, sono stati esclusi sia dal serbatoio dei possibili donatori, sia dal sottoinsieme di valori utilizzati per la stima dei modelli di imputazione.

Si osserva che i casi anomali si concentrano particolarmente nel Nord-est e nel Sud, a conferma delle evidenze già riscontrate in fase di analisi dei dati in termini di comportamenti anomali sulla variabile *Consumi pro-capite*. In questo caso, l'obiettivo della selezione è non solo verificare la possibile presenza di errori nella variabile obiettivo, ma soprattutto escludere tali valori dalla procedura di imputazione evitando il "propagarsi" di comportamenti anomali ad altre unità della popolazione obiettivo.

6.3 Metodi di imputazione per la variabile *Consumi intermedi*

Sulla base delle analisi descritte nel paragrafo precedente, i metodi descritti nel paragrafo 4.1 sono stati implementati utilizzando le informazioni ausiliarie come descritto di seguito.

a) *Donatore di distanza minima trasversale per classi (NND).*

La variabile da sottoporre a imputazione è *Consumi pro-capite* (rispetto alla popolazione comunale) $Y_{i,pc} = Y_i / Pop_{i,2012}$. Dopo aver individuato il donatore di distanza minima (*d*) rispetto a opportune variabili di *matching*, la variabile obiettivo *Consumi* viene ottenuta mediante la relazione:

$$Y_i^* = Pop_{i,2012} \times Y_{d,pc,2012} = Pop_{i,2012} \times (Y_{d,2012} / Pop_{d,2012}) \quad (25)$$

dove $Y_{d,2012}$ e $Pop_{d,2012}$ sono rispettivamente il valore osservato della variabile *Consumi* e la popolazione del Comune donatore *d* nell'anno di riferimento, mentre Y_i^* è il valore imputato della variabile *Consumi* nel Comune mancante *i*. Nell'applicare la procedura sono utilizzate le variabili *Redditi SIOPE 2012 pro-capite* (rispetto alla popolazione comunale), *Superficie* e *Pianura* come variabili di *matching*; in qualità di variabile di classificazione la *Ripartizione geografica*.

b) *Predictive Mean Matching per classi (PMM).*

Analogamente alla variabile *Redditi*, questo metodo prevede due passi:

- stima di un modello di regressione multivariato (per ogni classe di imputazione):

$$Y_i^p = \alpha + \beta_1 S_{i,2012} / Pop_{i,2012} + \beta_2 Superficie_i + \beta_3 Ndip_i + e_1 \quad (26)$$

  attraverso il quale calcolare il valore predetto Y_i^p per ogni Comune *i;*
- scelta del valore da imputare della variabile obiettivo Y_i^* mediante ricerca del donatore più vicino rispetto al valore predetto Y_i^p, all'interno di classi definite in base alla *Ripartizione geografica*.

c) *Donatore di distanza minima longitudinale per classi (NND long).*

I pro-capite della variabile *Consumi* $Y_i^* = Pop_{i,2012} \times Y_d / Pop_{d,2012}$ sono imputati con il metodo del donatore di distanza minima, in classi definite in base alla *Ripartizione geografica* e utilizzando come variabili di *matching* le informazioni di tipo longitudinale: Y_{2011} / Pop_{2011}, S_{2011} / Pop_{2011}, S_{2012} / S_{2011}, *Superficie, Pianura, Ndip*.

d) *Metodi deterministici di tipo longitudinale*.
Le tecniche di imputazione sperimentate sono analoghe a quelle descritte nel paragrafo 5.3. Le uniche differenze riguardano le variabili di classificazione utilizzate e il contenuto informativo delle variabili ausiliarie. In particolare:

- $rappCRB = Y_{i,2012} / Y_{i,2011}$ indica il rapporto tra il valore della variabile *Consumi* nell'anno di riferimento e nel 2011;
- $rappSIOPE = S_{i,2012} / S_{i,2011}$ indica il rapporto fra i valori della variabile *Consumi* da fonte SIOPE nell'anno di riferimento e nel 2011;
- i valori mediani sono relativi alle distribuzioni dei rapporti precedenti, all'interno di strati definiti in base alle variabili *Ripartizione geografica*, *Popolazione 2012* (in classi[7]) e *Pianura*.

e) *Metodi "misti" di tipo longitudinale.*
In questa classe di metodi sono state sperimentate le stesse tecniche descritte per la variabile *Redditi*. La prima fase del processo prevede un passo deterministico in cui il valore della variabile *Consumi* da fonte CRB relativamente al 2011, $Y_{i,2011}$, viene aggiornato utilizzando l'informazione ausiliaria disponibile in ottica longitudinale (vedi metodi descritti al punto d) precedente, con analogo significato per le variabili *rappCRB* e *rappSIOPE*), con stratificazione basata sull'incrocio delle variabili *Ripartizione geografica*, *Popolazione* (in classi) e *Pianura*. Nel secondo passo, la procedura viene completata da un'imputazione non deterministica effettuata mediante donatore di distanza minima longitudinale per classi (metodo *NND Long* descritto al punto c) precedente) definite dalla variabile *Ripartizione geografica*. Come nel caso dei *Redditi*, i tre metodi si distinguono per il tipo di informazione ausiliaria utilizzata nel primo passo (Popolazione, SIOPE, CRB).

6.4 Risultati della simulazione per la variabile *Consumi intermedi*

In questa sezione sono descritti i risultati della sperimentazione per la variabile *Consumi*. La Tavola 13 contiene i valori degli indicatori per ciascuno dei 10 metodi sperimentati, ottenuti per *k=1.000* iterazioni del processo sperimentale illustrato nel paragrafo 4.

Tavola 13 – Variabile *Consumi*, indicatori di qualità per metodo di imputazione. Livello Italia (anno 2012)

Indice	Metodi di imputazione									
	NND	PMM	NND long	Long CRB Sio	Long CRB Pop	Long Pop	Long Sio	NND long Misto Pop	NND long Misto CRB	NND long Misto Sio
RB	0,584	-0,418	0,601	0,284	0,298	0,359	0,002	0,321	0,260	-0,018
RMSE	1,975	1,680	1,968	1,159	1,163	0,491	0,338	0,464	1,153	0,340
RIE	1,238	0,816	1,227	0,171	0,176	0,219	0,275	0,216	0,172	0,277

Come si può osservare, anche nel caso della variabile *Consumi intermedi* i metodi ottimali in termini di RMSE sono quelli deterministici longitudinali e fra questi, sulla base dei valori degli altri due indicatori, sono da preferire quelli che sfruttano l'informazione ausiliaria contenuta nella fonte SIOPE (*Long CRB Sio* e *NND long Misto Sio*).

La Tavola 14 riporta i valori dell'RMSE per regione, al fine di evidenziare eventuali inefficienze dei metodi sperimentati a questo livello territoriale. Si conferma una prestazione migliore per tutte le regioni dei metodi longitudinali, in particolare di quelli che utilizzano le informazioni ausiliarie di fonte SIOPE.

[7] Classe 0 = < 1500; classe 1 = [1500,5000); classe 2 = [5000,10000); classe 3 = [10000,60000); classe 4 = [60000,100000); classe 5 = > 100000.

Tavola 14 – Variabile *Consumi*, valori dell'indicatore Relative MSE, per Regione (anno 2012)

Regione	Metodi di imputazione									
	NND	PMM	NND long	Long CRB Sio	Long CRB Pop	Long Pop	Long Sio	NND long Misto Pop	NND long Misto CRB	NND long Misto Sio
Piemonte	0,661	0,049	0,661	0,019	0,019	0,016	0,183	0,015	0,020	0,184
Valle D'Aosta	0,002	0,004	0,003	0,001	0,001	0,001	0,002	0,001	0,001	0,002
Lombardia	1,044	0,147	1,044	0,118	0,119	0,107	0,124	0,106	0,118	0,126
Trentino	0,035	0,014	0,029	0,007	0,023	0,023	0,009	0,006	0,008	0,009
Veneto	0,160	0,068	0,159	0,014	0,014	0,019	0,045	0,019	0,014	0,045
Friuli-Venezia Giulia	0,026	0,021	0,025	0,013	0,013	0,010	0,011	0,010	0,013	0,011
Liguria	0,500	0,077	0,500	0,015	0,016	0,015	0,017	0,015	0,015	0,017
Emilia-Romagna	0,135	0,054	0,131	0,039	0,039	0,046	0,037	0,045	0,039	0,037
Toscana	0,164	0,102	0,152	0,108	0,108	0,123	0,179	0,123	0,108	0,179
Umbria	0,038	0,030	0,032	0,004	0,004	0,010	0,043	0,010	0,004	0,043
Marche	0,018	0,026	0,016	0,009	0,009	0,017	0,011	0,017	0,009	0,011
Lazio	1,252	1,430	1,249	1,121	1,121	0,310	0,038	0,308	1,121	0,040
Abruzzo	0,019	0,024	0,023	0,007	0,007	0,009	0,011	0,009	0,007	0,011
Molise	0,008	0,008	0,006	0,003	0,003	0,003	0,007	0,003	0,003	0,007
Campania	0,068	0,717	0,069	0,043	0,047	0,043	0,127	0,036	0,046	0,122
Puglia	0,098	0,080	0,096	0,018	0,020	0,026	0,027	0,023	0,018	0,027
Basilicata	0,019	0,014	0,018	0,016	0,016	0,016	0,012	0,016	0,016	0,012
Calabria	0,061	0,039	0,051	0,013	0,014	0,012	0,033	0,013	0,014	0,033
Sicilia	0,242	0,298	0,242	0,077	0,076	0,047	0,110	0,046	0,076	0,111
Sardegna	0,065	0,048	0,065	0,017	0,017	0,013	0,011	0,013	0,017	0,010

6.4.1 Gli aggregati regionali

In questo paragrafo si presentano alcune analisi di confronto fra le stime della variabile *Consumi* ottenute sulla base dei metodi *Long Sio* e *NND long Misto Sio*, risultati migliori rispetto alle assunzioni fatte e agli indicatori utilizzati nella sperimentazione, e le corrispondenti stime Istat (Istat, 2014).

Nella Tavola 15 sono riportate le stime regionali Istat (colonna "Bilanci"), le corrispondenti stime regionali ottenute aggregando i dati comunali imputati con i due metodi ottimali (colonne *Long Sio* e *NND long Misto Sio*) e le differenze assolute fra esse. Nella Tavola 16 sono riportate le differenze relative percentuali (colonna *Diff percentuali*) fra aggregati/stime regionali e le corrispondenti stime calcolate sui dati "originali" (ottenuti cioè sui dati affetti da dati mancanti). La tavola riporta, per una migliore analisi dei risultati, anche informazioni relative alla percentuale di Comuni, Popolazione e Numero di dipendenti imputati per regione. Si può osservare che, con entrambi i metodi di imputazione, le variazioni a livello di aggregati regionali tra stime ottenute e le corrispondenti stime ufficiali (Bilanci Istat) sono molto contenute, con valori massimi in Piemonte, Valle D'Aosta, Molise e Bolzano. La differenza in valore assoluto fra stime imputate e stime da Bilanci Istat, a livello Italia, è pari a 22.389.746 per il metodo *Long Sio*, è pari a -8.636.671 per il metodo *NND Long Misto Sio*: tuttavia, la complessità della variabile in analisi, che risulta anche dagli andamenti a livello regionale molto differenziati, richiederà analisi più dettagliate a livello di singolo Comune.

Tavola 15 – Variabile *Consumi*, differenze assolute stime, per Regione (anno 2012)

Regione	Stime Totale Consumi			Differenze assolute	
	Long Sio	NND long Misto Sio	Bilanci	Bilanci-Imputati Long Sio	Bilanci-Imputati NND long Misto Sio
Piemonte	1.833.038.363	1.836.805.583	1.857.302.851	24.264.488	20.497.268
Valle D'Aosta	101.822.280	101.740.641	102.882.277	1.059.997	1.141.636
Lombardia	5.127.734.961	5.127.652.141	5.131.973.832	4.238.871	4.321.691
Veneto	1.772.252.839	1.772.174.877	1.765.642.754	-6.610.085	-6.532.123
Friuli-Venezia Giulia	662.051.882	662.051.882	659.803.345	-2.248.537	-2.248.537
Liguria	976.018.874	975.667.178	973.767.277	-2.251.597	-1.899.901
Emilia-Romagna	1.768.887.081	1.769.108.285	1.763.089.193	-5.797.888	-6.019.092
Toscana	1.735.302.765	1.735.302.765	1.735.615.824	313.059	313.059
Umbria	422.729.501	422.729.501	420.285.892	-2.443.609	-2.443.609
Marche	727.823.858	727.823.858	728.022.166	198.308	198.308
Lazio	4.438.805.753	4.441.137.676	4.429.346.276	-9.459.477	-11.791.400
Abruzzo	691.109.448	702.620.599	693.437.343	2.327.895	-9.183.256
Molise	139.639.830	139.812.358	142.450.555	2.810.725	2.638.197
Campania	2.352.370.525	2.356.436.201	2.374.065.767	21.695.242	17.629.566
Puglia	1.598.364.407	1.598.111.149	1.598.094.995	-269.412	-16.154
Basilicata	252.790.052	253.052.891	250.734.928	-2.055.124	-2.317.963
Calabria	756.681.974	759.278.659	753.900.879	-2.781.095	-5.377.780
Sicilia	1.879.984.658	1.887.527.424	1.878.251.615	-1.733.043	-9.275.809
Sardegna	964.747.545	964.580.309	964.456.864	-290.681	-123.445
Bolzano	228.138.669	227.893.014	230.620.641	2.481.972	2.727.627
Trento	298.554.630	298.369.321	297.494.366	-1.060.264	-874.955
Totale	28.728.849.894	28.759.876.311	28.751.239.640	22.389.746	-8.636.671

Alla luce dei risultati ottenuti si può concludere che il passaggio dal metodo di stima corrente ad un metodo di imputazione longitudinale a livello comunale, rendendo disponibili previsioni a livello di singola amministrazione delle variabili chiave del bilancio, determinerebbe i seguenti vantaggi:

- ricostruzione delle voci di dettaglio del bilancio stesso (ad es., attraverso metodi basati su proporzioni medie/donatori di proporzioni);
- disponibilità per ogni Comune di serie di valori coerenti longitudinalmente per costruzione (stime determinate sulla base del trend rilevato sui dati del Comune stesso da fonte SIOPE);
- estendibilità dell'approccio ad altre variabili chiave (capi-conto) del bilancio.

Tavola 16 – Variabile *Consumi*, differenze relative percentuali fra stime, percentuale Comuni, popolazione, dipendenti, per Regione (anno 2012)

Regione	Differenze percentuali					Percentuali imputati		
	Originali-Imputati Long Sio	Originali-Imputati NND long Misto Sio	Bilanci-Imputati Long Sio	Bilanci-Imputati NND long Misto Sio	Originali-Bilanci	Comuni	Popolazione	Dipendenti
Piemonte	2,67	2,88	1,31	1,10	4,03	3,57	3,35	3,37
Valle D'Aosta	4,21	4,12	1,03	1,11	5,29	6,76	5,45	4,23
Lombardia	1,49	1,49	0,08	0,08	1,58	4,08	2,17	1,48
Veneto	1,61	1,61	0,37	0,37	1,23	4,13	1,57	1,55
Friuli-Venezia Giulia	1,27	1,27	0,34	0,34	0,92	5,50	1,13	1,25
Liguria	3,64	3,60	0,23	0,20	3,40	10,64	3,43	2,88
Emilia-Romagna	0,43	0,45	0,33	0,34	0,11	2,01	0,65	0,40
Toscana	0,62	0,62	0,02	0,02	0,64	1,74	0,69	0,57
Umbria	2,79	2,79	0,58	0,58	2,19	1,09	2,12	1,41
Marche	0,61	0,61	0,03	0,03	0,64	1,67	0,74	0,52
Lazio	3,58	3,63	0,21	0,27	3,36	18,52	5,96	4,67
Abruzzo	7,85	9,65	0,34	1,32	8,22	13,11	8,42	6,93
Molise	16,42	16,56	1,97	1,85	18,76	14,71	14,88	17,19
Campania	9,14	9,33	0,91	0,74	10,15	15,25	10,59	8,01
Puglia	4,56	4,54	0,02	0,00	4,54	13,18	5,14	5,79
Basilicata	6,10	6,21	0,82	0,92	5,24	9,92	5,38	6,36
Calabria	19,60	20,01	0,37	0,71	19,16	18,58	16,04	15,24
Sicilia	26,33	26,84	0,09	0,49	26,22	23,33	21,23	21,92
Sardegna	6,79	6,77	0,03	0,01	6,76	13,00	6,39	6,92
Bolzano	16,42	16,29	1,08	1,18	17,69	16,38	15,55	13,81
Trento	6,23	6,17	0,36	0,29	5,85	9,22	5,47	4,90
Totale	4,75	4,85	0,08	0,03	4,82	8,71	5,91	6,22

7. Analisi basate sulla fornitura aggiornata della fonte CRB

In questo paragrafo vengono illustrati i risultati della valutazione della qualità delle imputazioni basata sull'utilizzo dei dati contenuti nella fornitura aggiornata dell'archivio CRB. Tale fornitura determina una situazione informativa differente in termini di mancata risposta: i Comuni presenti in entrambe le forniture sono 7.387, 297 Comuni presentano dati mancanti in entrambe le forniture, mentre per 408 Comuni il dato mancante della prima fornitura risulta osservato in quella aggiornata.

Ai fini delle analisi, i Comuni che presentano dati mancanti anche nella fornitura aggiornata sono stati imputati utilizzando il metodo *Long Sio*, risultato generalmente migliore rispetto agli altri. Tale metodo è di tipo deterministico per cui, a parità di informazioni provenienti da CRB nel 2012, i Comuni con dato imputato presentano lo stesso valore in entrambe le forniture. L'unica differenza che può verificarsi nel processo di imputazione che utilizza i dati della prima e seconda fornitura è di tipo indiretto, in quanto risiede nell'imputazione della variabile ausiliaria *Numero di dipendenti*, che utilizza tra le variabili di *matching* anche la variabile *Redditi*. Le differenze che sono risultate nei valori delle variabili obiettivo del 2012 per i 297 Comuni che sono stati imputati sia nella prima che nella seconda fornitura sono inferiori a 150 euro, facendo escludere quindi la possibilità di quantificare un eventuale effetto della revisione dei dati nel processo di imputazione.

L'efficienza del processo di imputazione è stata valutata analizzando le discrepanze tra dato imputato (prima fornitura) e dato osservato (seconda fornitura) su 408 Comuni sulla base dell'indicatore *RIE*. La Tavola 17 mostra i risultati ottenuti per le variabili *Redditi* e *Consumi*: i valori molto bassi dell'indicatore evidenziano la bontà del processo di imputazione. La Tavola 18 riporta le statistiche descrittive della distribuzione delle differenze assolute e percentuali tra il dato

imputato e il dato osservato per 408 Comuni per entrambe le variabili. A meno di valori anomali per cui sono necessarie analisi *ad hoc*, i risultati rilevano la bontà del processo di imputazione.

Come atteso, il processo di imputazione presenta delle problematicità maggiori per la variabile *Consumi*. Il numero delle osservazioni con valore della variabile differenze assolute minore del 1 e maggiore del 99 percentile sono 10: per 8 Comuni le differenze percentuali sono relativamente basse, mentre per i due Comuni con differenze percentuali più alte (superiori a 200) vi è verosimilmente un errore nelle variabili delle fonti ausiliarie utilizzate nel processo di imputazione.

Tavola 17 – RIE percentuale, variabili *Redditi* e *Consumi*.

Regione	RIE percentuale	
	Redditi	Consumi
Piemonte	0,01	0,08
Valle D'Aosta	0,15	0,28
Lombardia	0,01	0,02
Veneto	0,01	0,03
Friuli-Venezia Giulia	0,01	0,07
Liguria	0,03	0,12
Emilia-Romagna	0,01	0,20
Toscana	0,01	0,04
Umbria	0,04	0,27
Marche	0,01	0,02
Lazio	0,05	0,11
Abruzzo	0,08	0,19
Molise	0,12	0,43
Campania	0,05	0,32
Puglia	0,04	0,04
Basilicata	0,09	0,25
Calabria	0,15	0,83
Sicilia	0,45	2,07
Sardegna	0,07	0,11
Bolzano	0,26	0,76
Trento	0,01	0,01
Totale	0,05	0,14

Tavola 18 – Distribuzione differenze assolute e percentuali, variabili *Redditi* e *Consumi*.

Statistiche	Redditi		Consumi	
	Differenze assolute	Differenze percentuali	Differenze assolute	Differenze percentuali
minimo	-7.590.788	-47.32	-39.531.710	-72.57
1 percentile	-285.091	-21.56	-2.245.093	-43.35
1 quartile	-29.871	-4.62	-89.288	-11.26
media	-27.751	15.41	-83.118	3.57
mediana	-2.819	-0.52	-5.093	0.88
3 quartile	13.018	2.84	81.199	12.49
99 percentile	307.076	26.02	2.382.601	117.95
massimo	638.206	6577.30	4.892.893	241.28
deviazione std	387.766	325.77	2.059.548	28.31

8. Conclusioni, criticità e prospettive di lavoro

Le analisi e le sperimentazioni illustrate in questo lavoro, seppur condotte su due sole voci di bi-

lancio, hanno evidenziato le potenzialità connesse alla possibilità di stimare voci di conto economico mediante imputazione statistica. I risultati sono complessivamente incoraggianti in termini di livello di accuratezza delle previsioni dei valori mancanti per i Comuni non presenti nell'archivio di base, soprattutto nei casi in cui sia possibile sfruttare l'informazione longitudinale e quella proveniente dalla fonte SIOPE. Inoltre, rendendo disponibile l'intero set di microdati, esse rendono possibili elaborazioni intermedie, che rappresentano un input per i processi di produzione delle statistiche di Finanza pubblica di Contabilità Nazionale.

Alla luce di questi risultati, si può concludere che l'adozione di una strategia di stima basata sulla previsione statistica di microdati di variabili di fonte amministrativa, che sfrutti informazione ausiliaria anche di tipo longitudinale, può determinare una serie di vantaggi di tipo qualitativo rispetto all'attuale approccio adottato in Istat. Innanzi tutto, disponendo di previsioni a livello di singola amministrazione per le variabili chiave del bilancio, diventa possibile ricostruire coerentemente i microdati delle voci di dettaglio del bilancio stesso. Inoltre, la strategia proposta garantisce la coerenza longitudinale dei valori delle Amministrazioni Comunali. L'approccio appare poi teoricamente estendibile ad altre voci chiave (*capi-conto*) dei bilanci dei Comuni, tuttavia studi specifici saranno necessari per una valutazione accurata, tenendo conto di elementi come la qualità delle informazioni disponibili dalla fonte CRB, relazioni esistenti con le variabili ausiliarie, variabilità nel tempo dei fenomeni oggetto di stima, ecc. E' evidente infine che anche l'estendibilità ad altre tipologie di Amministrazioni Pubbliche andrà valutata caso per caso, tenendo conto ad esempio delle caratteristiche organizzative e funzionali degli Enti oggetto di analisi, della loro identificabilità e *demografia*, delle fonti utilizzabili e delle relative caratteristiche (contenuti informativi e relativa qualità, copertura).

Come era da attendersi, le analisi presentate nel lavoro hanno anche evidenziato alcune criticità di natura economica, concettuale e metodologica che richiederanno ulteriori valutazioni e sviluppi.

Da un punto di vista economico, l'utilizzo della fonte SIOPE, di cui ci si avvale nei metodi che sembrano mostrare i risultati migliori, presenta alcuni limiti. Oltre a quelli già citati riguardanti la provvisorietà dei dati, aspetto particolarmente problematico soprattutto per i dati recenti, vi è il problema della disomogeneità di comportamento dei Comuni nella compilazione dei documenti contabili. Tale aspetto potrebbe portare a distorsioni rilevanti nella stima delle variabili mancanti a livello del singolo Comune, soprattutto in considerazione della dimensione dell'ente; il fenomeno dovrebbe presentarsi in forma quantomeno attenuata nelle elaborazioni aggregate, per la consistente numerosità delle Amministrazioni Comunali. Altra cautela da tener presente relativamente all'uso dei dati di fonte SIOPE riguarda le metodologie utilizzate in CN. Per la produzione di stime viene impiegato un metodo analogo a quello utilizzato nelle procedure di imputazione deterministiche sperimentate in questo lavoro, basato sull'applicazione dei tassi di variazione tra due anni consecutivi di variabili da fonte SIOPE. Per le variabili di competenza, ovvero per impegni o accertamenti che sono considerati in contabilità nazionale come *proxy* della contabilità economica secondo il principio dell'*accrual*, l'applicazione dei tassi di variazione calcolati su aggregati di cassa potrebbe portare a stime imprecise, soprattutto per alcune voci economiche per le quali la dinamica di competenza sembra essere piuttosto diversa da quella di cassa. Questo si verifica in particolar modo negli ultimi anni, nei quali la dinamica dell'aggregato di cassa dei consumi intermedi è influenzata dal pagamento dei residui degli anni precedenti (debiti commerciali). Applicare la stessa dinamica alla competenza, senza depurare dall'effetto dei debiti commerciali, condurrebbe sicuramente a una sovrastima sia a livello aggregato e ancor più a livello dei singoli Comuni nei quali tale fenomeno si presenta piuttosto disomogeneo. A tal proposito, con l'entrata in vigore del Piano dei conti integrato (Ragioneria Generale dello Stato, 2013) e l'applicazione dei nuovi principi contabili, *in primis* il principio della competenza potenziata, ipotizzando comportamenti omogenei dei Comuni nella classificazione e registrazione delle poste in bilancio, la distorsione prima descritta dovrebbe ridursi in quanto le nuove regole contabili ridurranno la distanza fra la cassa e la competenza in virtù del fatto che per gli enti sarà possibile accertare o impegnare unicamente importi che siano esigibili entro lo stesso esercizio. Un altro aspetto da tenere in considerazione nella valutazione dell'efficacia dei metodi di imputazione presi in esame riguarda la non neutralità rispetto alle variabili economiche oggetto di analisi: sono state infatti oggetto di elaborazione le spese per il

personale e i consumi intermedi, che sono fra le voci più stabili e più regolari fra le spese correnti. Al contrario aggregati come i trasferimenti o le componenti della spesa in conto capitale mostrano nel tempo una variabilità molto alta e un aggiornamento poco tempestivo. Questo vale anche per le voci di entrata e in particolare per le entrate tributarie per le quali la continua e frequente evoluzione normativa degli ultimi anni, sopprimendo imposte o tasse e introducendone nuove, rende di fatto impossibile l'analisi longitudinale delle singole voci e il calcolo di tassi di variazione.

Da un punto di vista concettuale e metodologico, è evidente che il processo di stima di voci di conto economico di Amministrazioni Pubbliche deve poggiare su un *framework* di riferimento standardizzato (metadati) in cui le variabili oggetto di stima (sistema di voci del conto economico) siano derivate dalle diverse fonti sulla base di un impianto di regole di armonizzazione rispetto alle definizioni statistiche. Un *framework* di questo tipo consente da un lato di misurare in modo più accurato le discrepanze esistenti fra stesse variabili osservate in fonti diverse, e dunque di isolare eventuali effetti di misurazione/fonte, dall'altro di adottare approcci multivariati al problema dell'imputazione laddove possibile/opportuno. In effetti, nelle sperimentazioni presentate in questo documento si è proceduto in un'ottica univariata, dal momento che per la previsione di ciascuna voce non sono state sfruttate le eventuali relazioni statistiche con altre voci del conto economico. Un'analisi di tipo multivariato avrebbe richiesto un'analisi di tipo concettuale/definitorio anche sulle voci componenti gli aggregati considerati, oltre che la disponibilità di ulteriori variabili ausiliarie (ad esempio il numero di dipendenti). Ciò avrebbe implicato costi e risorse non compatibili con lo studio sperimentale condotto.

Una problematica rilevante è rappresentata, nel contesto applicativo analizzato in questo lavoro, dall'identificabilità delle unità statistiche appartenenti alla popolazione obiettivo, e dunque dall'integrabilità delle fonti. L'assenza di un codice identificativo standardizzato rispetto a quello della Lista S13 nelle fonti utilizzate nella sperimentazione ha reso infatti problematico il processo di identificazione e di abbinamento dei dati relativi ad uno stesso Comune. Ciò ha generato sia errori di (sotto-)copertura, con conseguente aumento della quota di Comuni da ricostruire mediante imputazione, sia errori di abbinamento da un anno al successivo (ad esempio a causa di cambiamenti nei codici identificativi dei Comuni dovuti a scorpori, fusioni, ecc.), con conseguente impossibilità di sfruttare per queste unità le relative informazioni longitudinali in fase di previsione. Questi aspetti, tuttavia, troveranno risposta nel nuovo contesto organizzativo e produttivo determinato dalla realizzazione in Istat di un sistema dei registri statistici basati sull'uso di fonti amministrative e dalla messa a regime del Sistema Integrato dei Microdati (Ambroselli, 2014). Una risposta statistica al problema è comunque rappresentata dal ricorso a metodologie di *record linkage* (Scanu, 2003) per l'abbinamento delle unità presenti nei vari archivi.

Un ulteriore aspetto di natura strettamente metodologica è rappresentato dalla necessità di valutare e monitorare la qualità dei dati, con riferimento sia alle singole fonti sia ai dati integrati. Gli errori di misurazione e/o di integrazione possono infatti dare origine a incoerenze nei dati e/o a valori anomali, la cui identificazione e trattamento pongono non solo problemi di natura metodologica, ma anche di costi e tempi. In particolare, vanno utilizzati metodi per l'ottimizzazione e la razionalizzazione dei controlli interattivi necessari sia per la verifica dei falsi positivi/falsi negativi negli abbinamenti fra unità da fonti differenti, sia dei valori anomali, potenzialmente dovuti a errori di misurazione nei dati di ogni fonte utilizzata. A questo scopo, dovranno essere valutate metodologie alternative a quella adottata in questo lavoro, come ad esempio tecniche di *editing* selettivo (Luzi, *et al.*, 2007), che consentono di minimizzare il numero di casi da sottoporre a controllo interattivo fissato il livello di accuratezza "atteso" per le stime obiettivo.

Appendice 1
Voce *Redditi da lavoro dipendente* nella fonte SIOPE

Al fine di disporre di una informazione ausiliaria dalla fonte SIOPE per la variabile *Redditi da lavoro dipendente* (disponibile nella fonte CRB)[8] occorre considerare il dato dei pagamenti di cassa (pagamenti totali come sommatoria dei pagamenti in conto competenza e in conto residui).

Tavola 19 – Spese per il personale. Fonte SIOPE, anno 2012.

Codici	Voci di spesa	Importi
1101	Competenze fisse per il personale a tempo indeterminato	9.801.096.485
1102	Straordinario per il personale tempo indeterminato	150.241.100
1103	Altre competenze ed indennità accessorie per il personale a tempo indeterminato	1.219.674.802
1104	Competenze fisse ed accessorie per il personale a tempo determinato	581.135.979
1105	Altre spese di personale (lavoro flessibile: personale con contratto di formazione e lavoro, lavoratori socialmente utili)	68.542.845
1106	Rimborsi spese per personale comandato	28.738.352
1107	Straordinario al personale per consultazioni elettorali	9.586.303
1109	Arretrati di anni precedenti	52.714.206
1110	Compensi per collaborazioni coordinate e continuative	11.321.484
1111	Contributi obbligatori per il personale	3.048.245.790
1112	Contributi previdenza complementare	15.641.391
1113	Contributi per indennità di fine servizio e accantonamenti TFR	141.874.470
1114	Contributi aggiuntivi	8.201.963
1115	Contributi relativi ad arretrati di anni precedenti	19.668.110
1121	Borse di studio e sussidi per il personale	1.050.852
1122	Centri attività sociali, sportive e culturali	421.085
1123	Contributi per prestazioni sanitarie	711.289
1124	Indennizzi	4.109.634
1131	Pensioni	8.038.457
1132	Pensioni integrative	15.723.995
1133	Altri oneri per il personale in quiescenza	29.687.298
1134	Arretrati di anni precedenti erogati al personale in quiescenza	2.137.472
	Totale (al netto di 1110 e 1105)	15.138.699.034

(a) I codici 1105 (Altre spese di personale con contratti di lavoro flessibile: personale con contratto di formazione e lavoro, lavoratori socialmente utili) e 1110 (Compensi per collaborazioni coordinate e continuative) sono stati esclusi dal momento che il SEC 2010 classifica queste forme contrattuali come acquisizione di servizi.

[8] Vedere G. Corradini (2014) "Manuale di lavoro per la realizzazione del frame PA". Documento interno gdl Frame PA.

Appendice 2
Voce “Consumi Intermedi” nella fonte SIOPE

La variabile *Consumi intermedi* è data dalla somma degli aggregati elencati nella Tavola 20.

Tavola 20 – Consumi intermedi. Fonte SIOPE, anni 2011 e 2012

Codici	Voci di spesa	Anno 2011	Anno 2012
	ACQUISTO BENI DI CONSUMO E MATERIE PRIME		
1201	Carta, cancelleria e stampati	x	x
1202	Carburanti, combustibili e lubrificanti	x	x
1203	Materiale informatico	x	x
1204	Materiale e strumenti tecnico-specialistici	x	x
1205	Pubblicazioni, giornali e riviste	x	x
1206	Medicinali, materiale sanitario e igienico	x	x
1207	Acquisto di beni per spese di rappresentanza	x	x
1208	Equipaggiamenti e vestiario	x	x
1209	Acquisto di beni di consumo per consultazioni elettorali	x	x
1210	Altri materiali di consumo	x	x
1211	Acquisto di derrate alimentari	x	x
1212	Materiali e strumenti per manutenzione	x	x
1213	Materiale divulgativo sui parchi, gadget e prodotti tipici locali		x
	PRESTAZIONI DI SERVIZI		
1302	Contratti di servizio per trasporto	x	x
1303	Contratti di servizio per smaltimento rifiuti	x	x
1304	Contratti di servizio per riscossione tributi	x	x
1305	Lavoro interinale	x	x
1306	Altri contratti di servizio	x	x
1307	Incarichi professionali	x	x
1308	Organizzazione manifestazioni e convegni	x	x
1309	Corsi di formazione per il proprio personale	x	x
1310	Altri corsi di formazione	x	x
1311	Manutenzione ordinaria e riparazioni di immobili	x	x
1312	Manutenzione ordinaria e riparazioni di automezzi	x	x
1313	Altre spese di manutenzione ordinaria e riparazioni	x	x
1314	Servizi ausiliari e spese di pulizia	x	x
1315	Utenze e canoni per telefonia e reti di trasmissione	x	x
1316	Utenze e canoni per energia elettrica	x	x
1317	Utenze e canoni per acqua	x	x
1318	Utenze e canoni per riscaldamento	x	x
1319	Utenze e canoni per altri servizi	x	x
1320	Acquisto di servizi per consultazioni elettorali	x	x
1321	Accertamenti sanitari resi necessari dall'attività lavorativa	x	x
1322	Spese postali	x	x
1323	Assicurazioni	x	x
1324	Acquisto di servizi per spese di rappresentanza	x	x
1325	Spese per gli organi istituzionali dell'ente - Indennità	x	x
1326	Spese per gli organi istituzionali dell'ente - Rimborsi	x	x
1327	Buoni pasto e mensa per il personale	x	x
1329	Assistenza informatica e manutenzione software	x	x
1330	Trattamento di missione e rimborsi spese viaggi	x	x
1331	Spese per liti (patrocinio legale)	x	x

Tavola 20 (segue) – Consumi intermedi. Fonte SIOPE, anni 2011 e 2012

Codici	Voci di spesa	Anno 2011	Anno 2012
1332	Altre spese per servizi	x	x
1333	Rette di ricovero in strutture per anziani/minori/handicap ed altri servizi	x	x
1334	Mense scolastiche	x	x
1335	Servizi scolastici	x	x
1336	Organismi e altre Commissioni istituiti presso l'ente	x	x
1337	Spese per pubblicità	x	x
1338	Global service	x	x
1339	Collaborazioni, coordinate e continuative (Co.co.co)	x	
1340	Rimborsi per il coordinamento nazionale dell'ambiente		x
	UTILIZZO BENI DI TERZI		
1401	Noleggi	x	x
1402	Locazioni	x	x
1403	Leasing operativo	x	x
1404	Licenze software	x	x
1499	Altri utilizzi di beni di terzi	x	x

Riferimenti bibliografici

Ambroselli, S. 2014. *I codici identificativi univoci all'interno del SIM (Sistema Integrato di Micro-dati)*. http://www.istat.it/it/files/2014/10/Paper_Sessione-III_Ambroselli.pdf

Corradini, G. 2014. Manuale di lavoro per la realizzazione del frame PA. Documento interno gruppo di lavoro Frame PA.

De Gregorio, C. and Giordano A., 2015. The heterogeneity of irregular employment in Italy: some evidence from the Labour force survey integrated with administrative data. *Istat Working Papers*. N.1/2015.

Istat, 2015a. *Linee guida per la qualità dei processi statistici di fonte amministrativa.* http://www.istat.it/it/files/2010/09/LineeGuida_v.1.0_Luglio_2015.pdf

Istat, 2015b. *Verso il censimento continuo delle istituzioni pubbliche alla luce delle principali evidenze della rilevazione del 2011*. http://www.istat.it/it/archivio/147427

Istat, 2015c. *Elenco delle unità istituzionali appartenenti al settore delle Amministrazioni Pubbliche*. G. U. n. 227 del 28 settembre 2012 e G. U. n. 228 del 30 settembre 2011. www.istat.it/it/archivio/6729

Istat, 2014. *I bilanci consuntivi delle Amministrazioni Comunali*. Nota metodologica. http://www.istat.it/it/archivio/121654.

Istat, 2013. 9° *Censimento industria e servizi, istituzioni e non profit: un Paese in profonda trasformazione*. http://www.istat.it/it/archivio/95481

Istat, 2012. *Rilevazioni sulla struttura per età della popolazione: informazioni sulla rilevazione*. http://www.istat.it/it/archivio/50362

Little, R.J.A. 1988. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6 (3): 287-296.

Luzi, O., Guarnera, U., Righi P. 2014. The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. *European Conference on Quality in Official Statistics (Q2014)*. Vienna, 3-5 June.

Luzi, O., M. Di Zio, U. Guarnera, A. Manzari, T. de Waal, J. Pannekoek, J. Hoogland, C. Tempelman, B. Hulliger and D. Kilchmann, 2007. *Recommended Practices for Editing and Imputation in Cross sectional Business Surveys*. EDIMBUS project report.

Ragioneria Generale dello Stato, 2013. *Piano dei conti integrato*. Decreto Ministeriale del 27 marzo 2013. http://www.rgs.mef.gov.it/VERSIONE-I/e-GOVERNME1/PianodeicontiIntegrato/.

Rousseeuw, P. J. and Van Driessen, K. 2000. An Algorithm for Positive-Breakdown Regression Based on Concentration Steps. *Data Analysis: Scientific Modeling and Practical Application, Gaul W., Opitz O. and Schader M. (Eds).* Berlin: Springer-Verlag: 335-346.

Scanu, M. 2003. *Metodi statistici per il record linkage*. Roma: Istat.

de Waal, T., J. Pannekoek and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation*. Wiley.