rivista di statistica ufficiale

n. 3
2014

Temi trattati

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 3
2014

Temi trattati

Istat

Direttore responsabile
Patrizia Cacioli

Comitato scientifico

Giorgio Alleva	Emanuele Baldacci	Francesco Billari
Tommaso Di Fonzo	Andrea Mancini	Roberto Monducci
Fabrizio Onida	Linda Laura Sabbadini	Antonio Schizzerotto

Comitato di redazione

Alessandro Brunetti	Romina Fraboni	Marco Fortini
Stefania Rossetti	Daniela Rossi	Maria Pia Sorvillo

Segreteria tecnica
Daniela De Luca, Laura Peci, Marinella Pepe, Gilda Sonetti

Per contattare la redazione o per inviare lavori scrivere a:
Segreteria del Comitato di redazione della Rivista di Statistica Ufficiale
All'attenzione di Gilda Sonetti
Istat – Via Cesare Balbo, 16 – 00184 Roma
e-mail: rivista@istat.it

rivista di statistica ufficiale

n. 3/2014

Periodico quadrimestrale
ISSN 1828-1982

Registrato presso il Tribunale di Roma
n. 339 del 19 luglio 2007

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Stampato nel mese di febbraio 2015
da Centro stampa e riproduzione S.r.l.
Via di Pietralata, 157 – Roma
Copie 260

FSC
www.fsc.org
MISTO
Carta da fonti gestite in maniera responsabile
FSC® C109152

Il ruolo della proiezione estera nella performance delle imprese manifatturiere italiane durante la crisi[1]

Maria Serena Causo[2], Stefano Costa[3] , Francesca Luchetti[4], Roberto Monducci[5]

Sommario

L'Italia, insieme alla Spagna, è il paese che nel corso della fase recessiva del 2011-2013 ha sperimentato la maggiore divaricazione tra andamento della domanda interna – fortemente negativo – e aumento delle vendite all'estero – in linea con l'evoluzione del commercio mondiale. Queste dinamiche hanno rappresentato da un lato un potente elemento di selezione delle imprese; dall'altro una causa di riorientamento complessivo dell'apparato industriale. Al di là delle esportazioni, negli anni recenti le diverse modalità di internazionalizzazione delle imprese sembrano aver giocato un ruolo importante nel determinare più o meno intense spinte alla crescita. In questo quadro, nuove basi di dati d'impresa integrate sviluppate dall'Istat consentono: a) di definire una "mappa" delle imprese manifatturiere in base ai loro profili strategici, evidenziando il peso della proiezione estera; b) di classificare le imprese con relazioni con l'estero in base al grado di complessità della loro esposizione; c) di valutare la performance esportativa delle singole unità alla luce della tipologia, delle strategie e del profilo economico dell'impresa.

Parole chiave: internazionalizzazione, esportazioni, competitività, integrazione di basi di dati microeconomici.

Abstract

In line with the global market evolution, during 2011-2013 recession, Italy and Spain showed the largest gaps between (strongly decreasing) domestic and (increasing) foreign demand. In Italy such trends have been both a key factor in selecting enterprises and a cause for a readjustment in the overall manufacturing sector. Beside export, other forms of enterprise internalisation have also played a crucial role in stimulating economic growth in recent years. In this vein, some innovative and integrated dataset at enterprise level provided by Istat make it possible to: a) define a "map" of the manufacturing sector on the basis of enterprise strategic profiles, emphasising their propensity to compete

[1] L'articolo pubblicato impegna esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Gli autori ringraziano Stefano De Santis, Maria Moscufo, Carmine Pappalardo, Stefania Rossetti e Claudio Vicarelli per gli utili suggerimenti, rimanendo comunque i soli responsabili del contenuto e di eventuali errori.

[2] Ricercatore (Istat), e-mail: causo@istat.it.

[3] Ricercatore (Istat), e-mail: scosta@istat.it.

[4] Ricercatore (Istat), e-mail: luchetti@istat.it.

[5] Direttore di dipartimento (Istat), e-mail: monducci@istat.it.

internationally; b) classify enterprises with foreign relationships according to the complexity of this propensity; c) assess enterprise exporting performance in the light of their characteristics, strategies and economic profiles.

Keywords: internationalisation, exports, competitiveness, microdata, database integration

1. Introduzione

In un contesto di forte flessione delle componenti interne di domanda e di conseguente forte caduta delle importazioni, il contributo alla crescita del Pil da parte delle esportazioni nette è stato elevato sia nel 2011 sia nel 2012. Nel 2013 la crescita delle esportazioni ha invece registrato un forte rallentamento rispetto all'anno precedente, coerente con l'andamento del commercio mondiale. L'impatto di queste dinamiche sul sistema manifatturiero è stato rilevante, innescando profondi fenomeni di ricomposizione settoriale dell'output e forti divergenze nelle dinamiche individuali delle imprese.

Il tema del potenziale di crescita delle imprese italiane associato a un aumento del grado di internazionalizzazione del sistema produttivo si propone, nella fase attuale, come centrale sia a livello macro – per le prospettive di tenuta e di ripresa della nostra economia – sia a livello micro, poiché la letteratura economica ha evidenziato l'esistenza di una relazione positiva tra la competitività delle imprese e il loro grado di internazionalizzazione.[6]

Di seguito vengono dapprima individuati i profili strategici prevalenti tra le imprese industriali italiane, con specifica attenzione alla proiezione internazionale. Successivamente, l'analisi si concentra sulle modalità con cui le nostre imprese operano sui mercati internazionali, dalle forme più "semplici" quali l'attività esclusiva di esportazione su un numero relativamente contenuto di mercati, a quelle più complesse, come l'esportazione su scala globale o l'internazionalizzazione produttiva. Infine, vengono analizzate le caratteristiche delle imprese che, dal 2010 al 2013, presentano una persistente spinta alla crescita dell'export (le imprese "vincenti") e di quelle che, nello stesso periodo, hanno registrato un declino delle vendite all'estero, in modo da ricavare indicazioni sulle strategie aziendali o sui fattori strutturali più promettenti ai fini della ricerca di spazi di crescita all'estero.

Le analisi presentate sono in gran parte basate su ampie basi di dati microeconomici recentemente costruite dall'Istat in risposta alle crescenti esigenze informative e di analisi sull'internazionalizzazione del sistema produttivo italiano. L'approccio seguito è stato da un lato quello dello sfruttamento dei dati già raccolti per la produzione delle diverse statistiche sulle imprese, integrandoli tra di loro per aumentare le possibilità di analisi della competitività; dall'altro la progettazione e realizzazione di nuove indagini.

[6] Per una visione d'insieme si rimanda, tra gli altri, a Melitz (2003), Barba Navaretti e Venables (2004), Melitz e Ottaviano (2008), Chaney (2008), Wagner (2011).

2. Il peso della proiezione estera nei profili strategici delle imprese manifatturiere italiane

La fase recessiva avviatasi nel 2008 e che, dopo una breve fase di ripresa, ha interessato gran parte del 2013, ha inciso notevolmente sulle performance economiche delle imprese manifatturiere, determinando differenze significative nelle dinamiche settoriali. La struttura organizzativa e le strategie delle imprese sono cambiate e continuano a modificarsi, con ricadute sulla flessibilità produttiva, sull'orientamento degli investimenti, sul posizionamento delle singole unità all'interno delle catene del valore e sulla capacità di rivolgersi ai mercati più dinamici. Peraltro, la crisi sta determinando cambiamenti sostanziali anche nelle imprese esposte sui mercati esteri, le quali hanno risentito del forte rallentamento della domanda internazionale manifestatosi nel 2013.

Caduta della domanda interna e forti instabilità di quella internazionale hanno quindi determinato, successivamente al 2007, una elevata pressione sulle condizioni produttive e di mercato delle imprese, modificando i fattori rilevanti per la sopravvivenza e la crescita delle singole unità produttive.

La competitività e il potenziale di crescita delle imprese italiane sono influenzati dalla combinazione delle scelte aziendali relative ad un insieme estremamente ampio di strategie produttive, organizzative, di mercato che a volte non trovano riscontro nemmeno nell'informazione statistica ufficiale. Da questo punto di vista, pressoché tutta la letteratura empirica si scontra con una generale limitatezza di informazioni. La consapevolezza di questi problemi ha indotto l'Istat a dedicare una parte della rilevazione del 9° censimento dell'industria e dei servizi all'approfondimento della misurazione degli elementi di modernizzazione e competitività del sistema delle imprese, con un focus particolare sulle imprese di minori dimensioni.

Alcune analisi hanno consentito di individuare, attraverso metodologie di analisi delle corrispondenze multiple,[7] tre "profili strategici", ciascuno a sua volta rappresentativo di un aspetto fondamentale della competitività delle imprese (cfr. prospetto 1).

Prospetto 1 - Principali profili strategici delle imprese manifatturiere italiane – 2011

DINAMISMO	INTERNAZIONALIZZAZIONE	COMPLESSITA' ORGANIZZATIVA
- Innovazione di prodotto o di servizio	- Il mercato di riferimento è internazionale	- Gestione di tipo manageriale
- Mercati di riferimento (locale, nazionale, internazionale)	- Esportazioni	- Appartenenza a un gruppo
- Innovazione di processo	- Innovazioni organizzative	- Controllo familiare
- Innovazione di marketing	- Relazioni con le imprese estere	- Assunzione di personale ad elevata qualifica
- Innovazioni organizzative	- Strategie volte ad aumentare le attività all'estero	- Qualità dei prodotti
- Strategie volte ad accedere a nuovi mercati		- Accordi di tipo formale (esclusi subfornitura e commessa)

[7] Per maggiori dettagli si veda Istat (2013.2).

I profili così ottenuti hanno le caratteristiche seguenti:

a. *dinamismo aziendale*: include l'attività innovativa in senso ampio − non solo innovazioni di processo e prodotto, ma anche di marketing e organizzazione aziendale – e la propensione a espandersi verso nuovi mercati, nazionali e internazionali;

b. *proiezione estera*: esprime le diverse strategie di presenza sui mercati internazionali − esportazioni, attivazione di relazioni o accordi con imprese estere, delocalizzazione produttiva – e la propensione a introdurre innovazioni nell'organizzazione dell'impresa;

c. *complessità organizzativa*: individua i modelli di governance delle imprese, a seconda che queste presentino una gestione di tipo manageriale, appartengano o meno ad un gruppo (eventualmente in posizione di controllanti), siano caratterizzate da meccanismi di controllo familiare, abbiano il socio principale di nazionalità estera, abbiano attivato *joint ventures*, consorzi o accordi formali con altre imprese. A questi aspetti se ne aggiungono altri di diversa natura, quale la circostanza che l'impresa abbia un punto di forza competitivo nella qualità dei prodotti e servizi offerti, o abbia assunto prevalentemente personale a elevata qualifica professionale.

A seconda di come questi profili si combinano nel caratterizzare l'attività di un'impresa, divengono una chiave di lettura per classificare le imprese nazionali sulla base del loro orientamento strategico. In particolare, attraverso l'analisi dei gruppi (*cluster analysis*) sono state individuate cinque distinte tipologie di imprese, trasversali rispetto alle consuete classificazioni strutturali.

La partizione, le cui caratteristiche fondamentali sono riportate nella Tavola 2, restituisce un risultato ricco di informazioni, che non si esaurisce su un unico asse di lettura, poiché non emerge una sola leva strategica che domina ciascun cluster, ma strategie complesse rappresentate da diverse combinazioni dei profili precedentemente individuati.

Tavola 2 − Caratteristiche e profili strategici delle imprese manifatturiere con almeno tre addetti - 2011

CLUSTER	Imprese	% di imprese	Addetti medi	Produttività (a)	Profili strategici (b)			Strategie prevalenti	Mercato geografico di riferimento
					Dinamismo	Proiezione estera	Complessità organizzativa		
Piccolo cabotaggio	150.075	63,5	7,7	35,5	17,1	9,4	5,1	Tutela della quota di mercato	Locale/ Nazionale (c)
Internazionali tascabili	50.451	21,4	21,1	49,9	45,3	59,4	9,8	Accesso ai mercati; Tutela della quota di mercato	Internazionale
Dinamiche spinte	16.345	6,9	20,7	48,4	80,8	23,4	9,0	Ampliamento gamma di prodotti/servizi;	Nazionale/ Internazionale
Conservatrici	16.037	6,9	26,6	62,4	26,0	15,1	31,4	Tutela della quota di mercato	Nazionale

Tavola 2 segue − Caratteristiche e profili strategici delle imprese manifatturiere con almeno tre addetti - 2011

CLUSTER	Imprese	% di imprese	Addetti medi	Produttività (a)	Profili strategici (b)			Strategie prevalenti	Mercato geografico di riferimento
					Dinamismo	Proiezione estera	Complessità organizzativa		
Unità complesse	3.390	1,5	167,4	77,3	54,5	45,2	68,7	Tutela della quota di mercato; Ampliamento gamma dei prodotti/servizi	Internazionale
Totale	**236.398**	**100,0**	**15,0**	**44,3**	**28,7**	**21,9**	**9,0**		

Fonte: Istat, Rapporto annuale sulla situazione economica del Paese 2012. Elaborazioni su dati provvisori del Censimento Industria e Servizi 2011.

(a) Valore aggiunto per addetto (media in migliaia di euro).Il dato si riferisce al 2010.

(b) Indici normalizzati a 100. I valori variano tra un minimo di 10 e un massimo di 100, a seconda dell'intensità con cui il profilo strategico caratterizza il singolo gruppo di imprese.

(c) Locale= area di mercato comunale o regionale.

Il primo gruppo ("Piccolo cabotaggio") include la maggioranza delle imprese manifatturiere (poco più del 63% del totale): le unità produttive che vi appartengono sono caratterizzate da un basso dinamismo, si rivolgono prevalentemente a un mercato locale (comunale o regionale) e presentano un'organizzazione aziendale molto semplificata. In questo gruppo di imprese la proiezione internazionale è sostanzialmente assente.

A un estremo idealmente opposto si trova il quinto gruppo (le "Unità complesse"), che comprende solo l'1,5% delle imprese. Le unità di questo cluster sono accomunate dalla presenza di un dinamismo e una proiezione sui mercati internazionali relativamente elevati, e soprattutto da una elevata complessità organizzativo-gestionale.

Oltre un quinto delle imprese ricade nel gruppo delle "Internazionali tascabili", popolato da imprese a bassa complessità organizzativa, ma dinamiche e con una forte vocazione internazionale, non solo in termini commerciali e produttivi, ma anche di relazioni interaziendali.

Infine, l'insieme delle "Dinamiche" comprende circa il 7% delle imprese manifatturiere e include le unità con il profilo strategico più vivace: si trovano qui le aziende che fanno più ricorso a innovazioni di prodotto, di processo, organizzative e di marketing, che hanno una proiezione internazionale relativamente elevata, che puntano sull'arricchimento dell'offerta e sull'accesso a nuovi mercati.

L'ultimo cluster ha un peso simile a quello delle "Dinamiche" ed è composto da imprese "Conservatrici", che si segnalano per un'organizzazione aziendale complessa, a fronte di un ruolo relativamente contenuto dei profili legati al dinamismo e alla proiezione internazionale.

In questo quadro, il profilo strategico legato alla proiezione internazionale delle imprese emerge come tratto distintivo di almeno due ampi segmenti del comparto manifatturiero. Si tratta, da un lato, delle "unità complesse", dall'altro – con maggiore forza – delle "Internazionali tascabili", nel quale si collocano circa 50mila imprese manifatturiere, che assorbono un milione di addetti

Le notevoli differenze tra questi due gruppi, in termini di dimensione d'impresa e di

copertura settoriale, attirano l'attenzione su due importanti caratteristiche del profilo stesso: a) la sua pervasività tra i diversi segmenti industriali, a testimonianza di una vocazione internazionale "diffusa" delle imprese manifatturiere italiane; b) la possibilità che le strategie d'impresa possano risultare più rilevanti, ai fini della performance, rispetto ai consueti fattori strutturali legati a dimensione e settore. Quest'ultimo punto, in particolare, merita di essere ulteriormente approfondito, perché chiama in causa le diverse modalità con cui le imprese partecipano alla competizione internazionale.

3. I modelli di internazionalizzazione delle imprese manifatturiere italiane

Un banco di prova di rilievo per l'analisi delle strategie di internazionalizzazione delle imprese è rappresentato dall'esame delle diverse forme di presenza sui mercati internazionali a cavallo degli anni più difficili della crisi, tra il 2007 e il 2010.

Ciò richiede in primo luogo un complesso lavoro di costruzione di basi di dati adeguate ad affrontare i diversi aspetti in cui il fenomeno si presenta. A tale scopo è stata costruita, per gli anni 2007 e 2010, una innovativa base dati derivante dall'integrazione di un ampio numero di indagini statistiche e dati amministrativi, comprensiva di osservazioni per circa 30.500 imprese manifatturiere attive in entrambi gli anni, che nel 2010 impiegavano circa 1,7 milioni di addetti.

A partire da questa base di dati e dalla copiosa letteratura sul tema, è stata costruita una tassonomia delle forme di internazionalizzazione delle imprese manifatturiere italiane a controllo nazionale, formata da quattro classi mutualmente esclusive: *a)* Solo esportatori; *b)* Esportatori-importatori; *c)* Globali; *d)* Multinazionali. Con un percorso che idealmente procede da modelli più elementari a strutture via via più complesse, le prime tre classi individuano altrettante forme di internazionalizzazione commerciale, mentre l'ultima si riferisce all'internazionalizzazione produttiva. In tal modo, la classe più elementare, relativa ai "Solo esportatori", è costituita da imprese che non importano ma svolgono un'attività di esportazione verso i paesi Ue e/o verso un massimo di quattro aree geografiche extra-Ue;[8] nella seconda classe e terza classe sono state incluse le imprese che effettuano attività sia di esportazione che di importazione ("Esportatori-importatori"); nella terza quelle che vendono in almeno cinque aree extra-europee ("Globali") e nell'ultima le imprese italiane che hanno controllate estere ("Multinazionali"). Ogni azienda è assegnata, per ciascun anno di riferimento, a una sola classe; qualora un'impresa presenti più caratteristiche di classi diverse, essa è attribuita alla classe più elevata. Ciò significa, ad esempio, che un'impresa che svolge contemporaneamente attività di import ed export e vende in almeno sei aree al di fuori dell'Unione europea non è classificata tra gli "Esportatori-importatori", ma tra le imprese "Globali".

Dall'analisi dei dati emerge in primo luogo che le imprese manifatturiere che nel 2010 attuavano forme più complesse di internazionalizzazione presentavano dimensioni maggiori e più elevati livelli di efficienza (Tavola 3), oltre che una più accentuata diversificazione

[8] Il mercato mondiale è stato ripartito in undici aree: Unione europea-27; Paesi europei non Ue; Africa settentrionale; Altri paesi africani; America settentrionale; America centro-meridionale; Medio oriente; Asia centrale; Asia orientale; Oceania; Altri territori e destinazioni.

produttiva, misurata in termini di varietà di beni esportati (Tavola 4).

Tavola 3 - Caratteristiche strutturali delle imprese manifatturiere per forme di internazionalizzazione – Anno 2010

FORME DI INTERNAZIO-NALIZZAZIONE	Imprese		Addetti		Addetti medi	Fatturato medio *(migliaia di euro)*	Produttività media *(val. aggiunto per addetto)*	Profittabilità media *(MOL/ val.agg.)*	Grado medio di apertura *(esport. /fatt. totale)*
	Numero	%	Numero	%					
Multinazionali	2.230	5,9	366.156	21,0	164,2	50.322,0	70,9	32,1	47,8
Globali	8.358	22,3	714.052	40,9	85,4	28.102,1	62,8	32,9	49,7
Esportatori-importatori	14.754	39,3	496.159	28,4	33,6	9.065,1	55,4	34,8	25,7
Solo esportatori	12.173	32,4	170.913	9,8	14,0	2.293,6	43,1	34,3	17,7
Totale	**37.515**	**100,0**	**1.747.279**	**100,0**	**46,6**	**13.561,6**	**54,0**	**34,1**	**29,8**

Fonte: ISTAT, elaborazioni su base dati micro COE–FATS.

La forma di internazionalizzazione prevalente tra le imprese manifatturiere italiane presenti sui mercati internazionali è quella degli Esportatori-importatori (39,3%). Poco meno di un terzo (il 32,4%) è composto da imprese solo esportatrici; il 22% circa è "Globale", mentre solo il 5,9% sono multinazionali, sebbene questa percentuale sia superiore a quella relativa al complesso del sistema economico (3,4%).

In termini di occupazione, tuttavia sono le imprese Globali a pesare di più (per circa il 41%), mentre le Esportatrici-importatrici spiegano il 28,4% degli addetti complessivi e le Multinazionali il 21%. Le imprese Solo esportatrici, relativamente numerose, occupano meno del 10% degli addetti totali. Del resto, la dimensione media aziendale cresce vistosamente all'aumentare della complessità del modello di internazionalizzazione adottato: dagli oltre 164 addetti medi delle multinazionali italiane si passa agli 85 delle imprese Globali, fino ai 14 addetti medi delle imprese Solo esportatrici.

Forme di internazionalizzazione più complesse, infine, anche coerentemente con la maggiore dimensione media d'impresa, si accompagnano a livelli più elevati di produttività (misurata in termini di valore aggiunto per addetto), ma soprattutto a una presenza più estesa e articolata sui mercati esteri: le multinazionali sono quelle che più delle altre diversificano in termini di numero di prodotti esportati (37,6 a fronte degli 11,9 della media totale), numero di settori presidiati con l'export (6,6 contro una media pari a 2,9) e numero di paesi di destinazione delle esportazioni (34,5 a fronte del 13,2 della media complessiva, si veda la Tavola 4).

Tavola 4 - Forme di internazionalizzazione e diversificazione produttiva nelle imprese manifatturiere - Anno 2010

FORME DI INTERNAZIONALIZZAZIONE	Diversificazione produttiva		
	Numero di prodotti esportati	Numero di settori in cui si esporta	Numero di paesi in cui si esporta
Multinazionali	37,6	6,6	34,5
Globali	23,0	4,4	30,7
Esportatori-importatori	8,2	2,5	7,6
Solo esportatori	4,0	1,6	3,9
Totale	**11,9**	**2,9**	**13,2**

Fonte: ISTAT, elaborazioni su base dati micro COE–FATS.

È ragionevole ritenere che, nel tentativo di contenere gli effetti reali della crisi, tra il 2007 e il 2010 le scelte degli imprenditori sulle modalità di presenza sui mercati internazionali siano cambiate, con conseguente impatto sulla performance d'impresa. Per valutare la consistenza e gli effetti di tali cambiamenti, occorre analizzare anzitutto come sia mutata la distribuzione delle imprese tra le forme di internazionalizzazione. Ebbene, tra il 2007 e il 2010 quasi tre quarti (74,6%) delle imprese a controllo italiano presenti nel campione in entrambi gli anni hanno mantenuto invariate le proprie modalità di internazionalizzazione. Nello stesso periodo, tuttavia, un consistente numero di imprese cambia posizione nella scala dell'internazionalizzazione: l'11,1% del campione è transitato verso tipologie meno evolute, mentre il 14,6% si è spostato verso forme più complesse di internazionalizzazione.

Lo spostamento netto verso modalità più evolute di partecipazione ai mercati esteri è spinto in particolare dalle imprese che nel 2007 erano "Solo esportatrici" e successivamente hanno affiancato all'attività di export anche quella di importazione, solo parzialmente controbilanciata dal ritrarsi, verso la stessa classe di "Esportatori-importatori", di imprese precedentemente "Globali" (4,1% di casi).

Diviene allora opportuno analizzare l'effetto di questi spostamenti – e delle persistenze nelle diverse classi – sulla performance delle imprese. A questo proposito, semplici esercizi di stima[9] condotti sulla base dati così modificata rivelano che le transizioni verso tipologie più evolute di internazionalizzazione hanno un effetto positivo e significativo sulla performance delle imprese, in termini di variazione del valore aggiunto e andamento degli addetti impiegati in Italia: ad esempio, tra il 2007 e il 2010 ampliare su scala mondiale la propria attività attraverso un salto da "Solo esportatore" a "Globale" si è accompagnato a un aumento del 10,9% del valore aggiunto, praticamente identico all'incremento di valore aggiunto registrato da chi è passato dalla condizione di "Globale" a quella di impresa multinazionale (+10,8%). Vistoso è anche l'effetto positivo associato alla transizione da "Esportatore-importatore" a "Globale" (+8,7%). Al contrario, per chi ha fatto il percorso inverso, e ha visto ridimensionarsi l'estensione della propria presenza sui mercati esteri passando a forme più elementari di internazionalizzazione, si rileva un sensibile

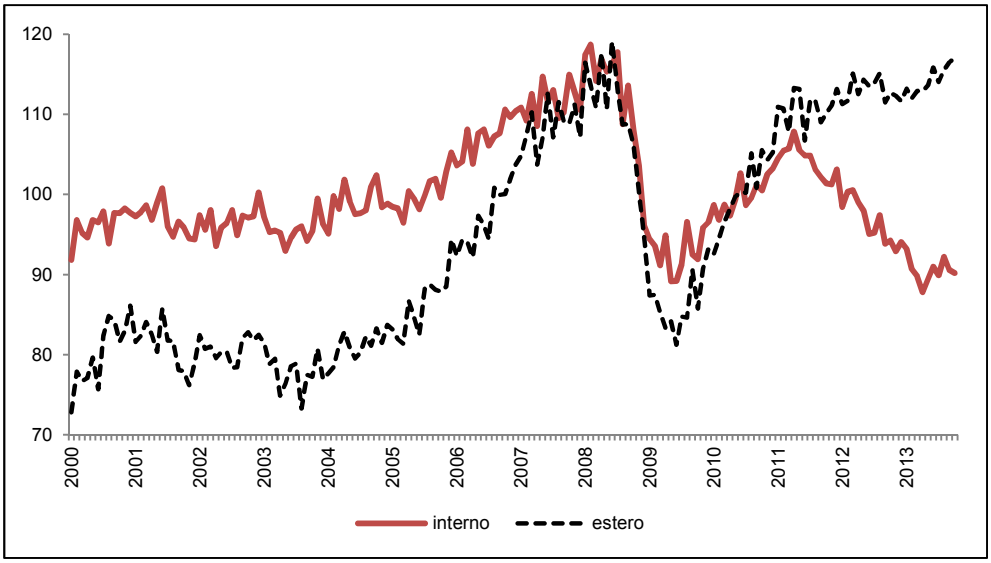
[9] Gli effetti sono stati stimati attraverso modelli lineari (OLS), controllando per la dimensione d'impresa e la ripartizione geografica di localizzazione delle imprese.

peggioramento di performance: ripiegare dalla condizione di "Globale" a quella di "Esportatore-importatore" si associa a una contrazione di valore aggiunto pari al 4,8%, mentre più ampia è la diminuzione legata alla transizione da "Esportatore-importatore" a "Solo esportatore" (-16,1%).

4. Le dinamiche delle imprese manifatturiere esportatrici nel 2010-2013

Nel corso della nuova recessione avviatasi nel secondo trimestre del 2011 e durata nove trimestri consecutivi, le imprese manifatturiere italiane hanno sperimentato una fase di accelerata e intensa riallocazione delle vendite dal mercato interno ed estero, che non trova riscontro nell'esperienza dell'ultimo decennio, e trova riscontro, in ambito europeo, solo in Spagna (Figura 2).

Figura 2 - Indici del fatturato dell'industria italiana per mercato di destinazione – dati mensili destagionalizzati – 2010=100



Fonte: Istat

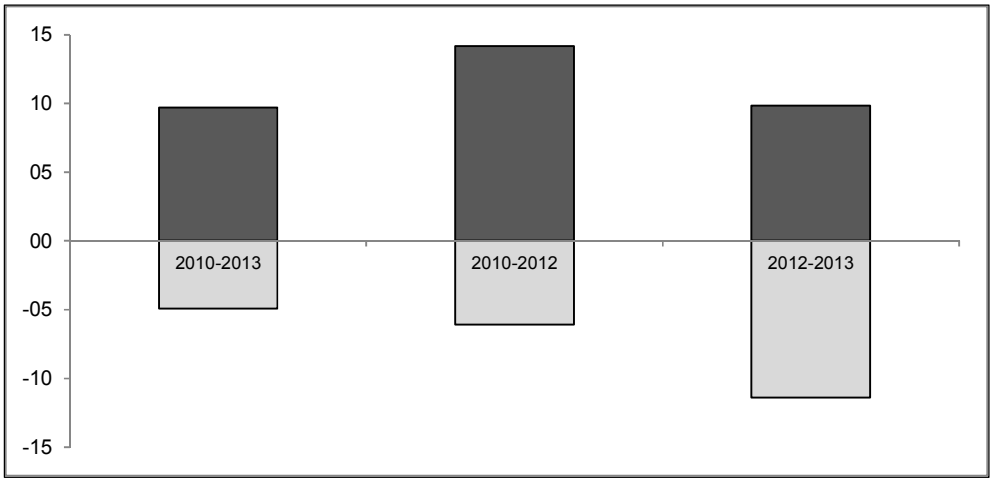
Questa divaricazione ha stimolato le imprese esportatrici a sviluppare ulteriormente la propria presenza sui mercati esteri, in modo da compensare le forti perdite registrate su quello interno. L'intensificazione della proiezione estera ha aumentato la competizione tra imprese, determinando percorsi individuali significativamente differenziati.

Il grado di eterogeneità nelle dinamiche individuali delle imprese esportatrici può essere misurato attraverso basi di dati ad alta rappresentatività, che consentono di scomporre le dinamiche aggregate in flussi riconducibili ai diversi segmenti di imprese, distinte per dimensione, settore, performance economica, caratteristiche strutturali, comportamenti strategici e orientamenti di mercato.

Con queste finalità, è stato costruito un panel di 29mila imprese manifatturiere persistentemente esportatrici nel periodo 2010-2013 (gennaio-maggio di ogni anno), dalla copertura molto elevata: 241 miliardi di export nell'intero 2010 (91% del totale) e 108 miliardi nei primi cinque mesi del 2013. La struttura delle imprese del panel per classe dimensionale vede presenti circa 6mila microimprese (con meno di 10 addetti), 16mila piccole imprese (10-49 addetti), 6mila medie imprese (50-249 addetti) e 1.200 grandi imprese. In proposito, tuttavia, va sottolineata la modificazione del peso relativo delle diverse classi dimensionali in termini di export: tra il 2010 e il 2013 è aumentato il peso delle micro e piccole imprese (dal 14,6% al 16,4%) e delle medie imprese (dal 30,3% al 32%), mentre l'incidenza delle grandi imprese è passata dal 55,1% al 51,6%. Questo pattern dimensionale è confermato anche se si considerano separatamente i flussi Ue ed extra-Ue.

Un primo spunto di analisi è quello relativo alla scomposizione della dinamica complessiva dell'export nei contributi dovuti alle imprese in espansione e a quelle in flessione. La figura 3 riporta, per tre periodi (2010-2013, 2010-2012 e 2012-2013) la scomposizione della variazione media annua dell'export nei contributi dovuti alle imprese in crescita ed a quelle in contrazione.

Figura 3 - Contributi positivi e negativi alla crescita media annua dell'export delle imprese in espansione e di quelle in flessione



Fonte: Istat, Rilevazione sul commercio estero

Rispetto alla tendenza degli ultimi tre anni, nel 2013 emerge un forte peggioramento del contributo negativo delle imprese in flessione di export, associato ad un ridimensionamento dell'intensità di crescita di quelle in espansione.

Concentrando l'attenzione alle dinamiche registrate tra 2012 e 2013, è possibile rilevare diverse specificità. Un aspetto di fondo dominante è dato dalla notevole differenza tra la crescita (+2,6%) dell'export verso l'area extra-Ue e il calo (-4,6%) verso l'area Ue. Dal punto di vista dimensionale, la performance esportativa dell'ultimo anno è correlata inversamente alla dimensione aziendale: +8% per le micro; +3,8% per le piccole; -0,3% per le medie; -4% per le grandi, che si riduce a -2% al netto delle imprese che operano nel settore della raffinazione. Se si valutano queste dinamiche alla luce delle performance dei diversi segmenti dimensionali registrate negli anni passati, nel 2012-2013 emerge

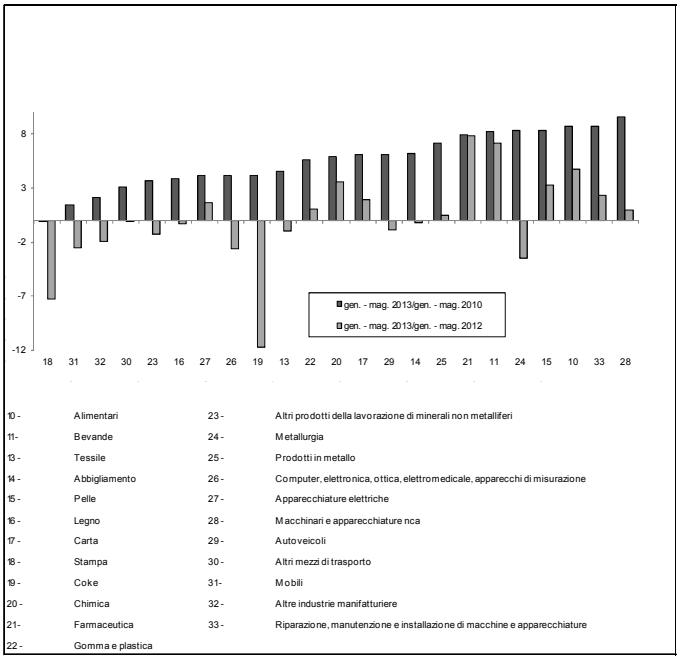
chiaramente un rilevante peggioramento per le medie e soprattutto grandi imprese.

Un ulteriore aspetto, legato a quello dimensionale, è che il sostegno all'export proviene unicamente dalla crescita, seppure lieve (+0,6%) delle imprese con bassa propensione all'export (quelle che esportano meno del 20% del fatturato). Le imprese con media esposizione (fatturato esportato compreso tra il 20% e il 60% del fatturato totale) mostrano una caduta del 3,4%, che si ridimensiona allo 0,2% per le imprese fortemente esposte.

La persistenza della recessione sta quindi stimolando una maggiore presenza all'estero da parte di imprese, soprattutto di piccola dimensione, finora prevalentemente orientate al mercato interno; d'altra parte, quelle con maggiore esposizione mostrano segnali di difficoltà a mantenere adeguati tassi di crescita.

Dal punto di vista settoriale, è possibile analizzare la performance dei diversi comparti in termini di intensità e diffusione – interna al settore – delle spinte alla crescita dell'export e valutare come la fase più recente abbia modificato il quadro di medio periodo. La Figura 4 mostra i settori industriali ordinati in base alla variazione mediana dell'export delle imprese registrata (in media annua) tra il 2010 e il 2013. A questa misura è stata associata quella relativa all'ultimo anno.

Figura 4 - Imprese manifatturiere persistentemente esportatrici nel 2010-2013. Variazioni mediane annue dell'export per settore (2010-13 e 2012-13)



Fonte: Istat, Rilevazione sul commercio estero

Come si vede, nella parte destra della figura emergono i settori delle macchine e apparecchiature, alimentare, pelle, metalli, bevande e farmaceutica. Per ognuno di questi settori la metà delle imprese ha sperimentato un tasso di crescita medio annuo dell'export pari o superiore all'8%.

In questo quadro tendenziale, solo i settori farmaceutico, delle bevande, alimentare, hanno mantenuto un profilo di "crescita diffusa" relativamente elevato anche tra il 2012 e il 2013. Il comparto delle macchine, che è quello che ha registrato i migliori risultati nel periodo 2010-2013, ha nettamente peggiorato la propria performance "diffusa", mentre la caduta relativamente più intensa sembra interessare il comparto dei metalli.

Come si è visto, la fase successiva alla crisi del 2008-2009 ha visto rilevanti modificazioni interne al sistema delle imprese esportatrici, che hanno reagito al quadro macroeconomico interno e internazionale in modo differenziato per settore, dimensione, area di sbocco prevalente.

Allo scopo di individuare i segmenti di imprese persistentemente competitivi, quelli in declino strutturale, e quelli con dinamiche differenziate nelle diverse fasi, le imprese del panel sono state suddivise in 4 gruppi:

1. imprese con export in crescita in entrambi i periodi 2010-2012 e 2012-2013;
2. imprese con export in flessione in entrambi i periodi 2010-2012 e 2012-2013;
3. imprese con export in flessione nel 2010-2012 e in crescita nel 2012-2013;
4. imprese con export in crescita nel 2010-2012 e in flessione nel 2012-2013.

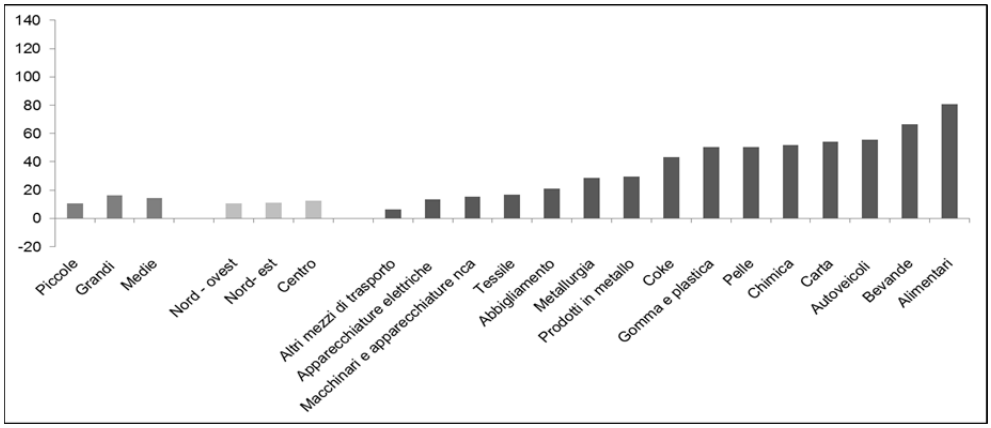
Le imprese manifatturiere che hanno aumentato le esportazioni sia nel 2010-2012 sia nel 2012-2013 sono circa 8.900; si tratta del 31% del totale delle imprese sempre esportatrici nel periodo 2010-2013 (gennaio-maggio di ogni anno), e spiegano oltre il 45% del valore complessivo dell'export dei primi cinque mesi del 2013.

Come nel paragrafo precedente, anche in questo caso, attraverso la stima di modelli probit, sono state analizzate le determinanti della probabilità di appartenere a ciascun gruppo di imprese. In particolare, una crescita continua delle esportazioni si associa a un profilo strategico che vede le aziende operare su scala globale, ma più estensivamente che intensivamente: si tratta infatti di imprese che già nel 2010 esportavano in almeno cinque aree extra-Ue (erano "Globali", secondo la tassonomia precedentemente descritta), e che nel triennio 2010-2013 hanno presidiato un numero crescente di mercati e offerto un numero crescente di prodotti. Per queste imprese l'eventuale aumento della quota di export nell'extra-Ue rappresenta però un fattore di spinta solo se accompagnato da un aumento dei mercati serviti: al contrario, il solo concentrarsi intensivamente sull'extra-Ue, senza accrescere i mercati di sbocco, diviene un fattore di rischio per la dinamica recente delle esportazioni. Per le imprese dall'export sempre crescente, inoltre, le strategie risultano più importanti dei caratteri strutturali, alcuni dei quali hanno tuttavia un ruolo significativo: queste unità produttive sono per lo più medio-grandi, con sede nelle regioni del centro-nord, e appartengono prevalentemente ai settori alimentare, delle bevande e degli autoveicoli. Al contrario, sono rare, in questo gruppo "virtuoso", le microimprese, le aziende delle regioni meridionali e insulari, e quelle attive nei settori tradizionali, in particolare mobili.

Un quadro speculare presentano le imprese le cui esportazioni sono diminuite in entrambi i periodi. Si tratta di poco più di 4.500 unità, che spiegano il 7,5% delle esportazioni dei primi cinque mesi del 2013. Soprattutto, coerentemente con quanto appena

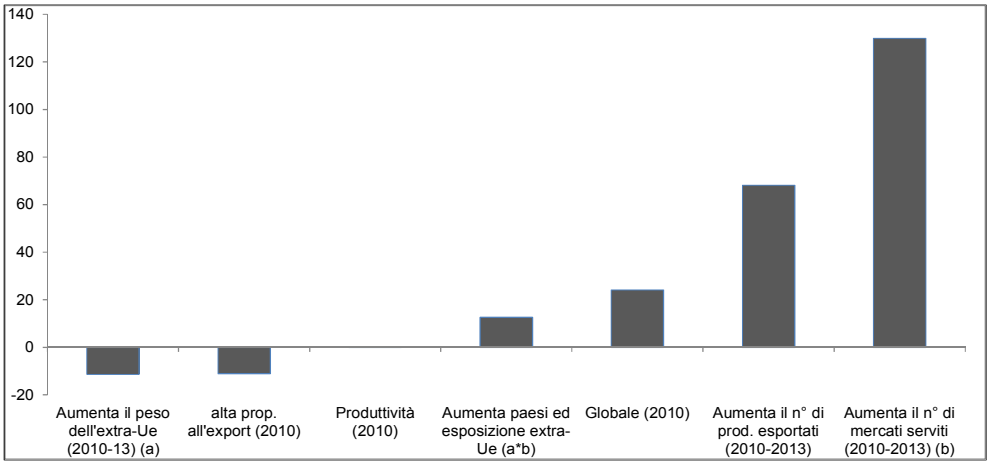
visto a proposito delle imprese dall'export sempre crescente, questo insieme di imprese ha visto declinare le proprie esportazioni con tanta maggiore probabilità quanto più intensa – ma meno estesa in termini di mercati serviti – è stata nel 2010 la quota di fatturato da esportazioni e la focalizzazione su (pochi) mercati extra-Ue. Il profilo strutturale dell'impresa con export "sempre decrescente" è quello di un'azienda di dimensioni medio-piccole, operante nei settori tradizionali o a bassa intensità di scala quali tessile, abbigliamento, legno, mobili, ma senza una marcata caratterizzazione territoriale.

Figura 5 - Imprese manifatturiere persistentemente esportatrici nel 2010-2013. Determinanti della probabilità di crescita persistente dell'export (sia nel 2010-12, sia nel 2012-13) - Effetti territoriali, dimensionali e settoriali



Fonte: Elaborazioni su dati Istat

Figura 6 - Imprese manifatturiere persistentemente esportatrici nel 2010-2013. Determinanti della probabilità di crescita persistente dell'export (sia nel 2010-12, sia nel 2012-13) - Effetti imputabili a tipologia, strategie e profilo economico dell'impresa



Fonte: Elaborazioni su dati Istat

Vi è poi un gruppo di imprese manifatturiere esportatrici che nell'ultimo anno ha visto tornare ad aumentare le esportazioni dopo aver vissuto un declino dell'export nel biennio 2010-2012. Si tratta di poco meno di 5.900 unità, rappresentative di circa il 13% dell'export totale manifatturiero del periodo gennaio-maggio 2013. Anche in questo caso l'andamento si è accompagnato a un aumento dei mercati serviti, soprattutto in paesi extra-Ue, da parte di imprese attive prevalentemente nei settori dell'elettronica, delle apparecchiature elettriche e dei mezzi di trasporto diversi dagli autoveicoli.

Infine, un terzo delle imprese considerate (circa 9.600 unità, rappresentative di circa il 35% delle esportazioni complessive del comparto) ha subìto un calo dell'export nel 2012-2013 dopo un precedente biennio di crescita. Da un lato queste imprese si caratterizzano per una strategia orientata a una intensa propensione all'esportazione, dall'altro lato spiccano una connotazione territoriale legata alle regioni settentrionali e una caratterizzazione settoriale fortemente orientata alle attività tipiche del Made in Italy, come la meccanica, la metallurgia, i metalli, la produzione di autoveicoli e di derivati della lavorazione di minerali non metalliferi.

5. Conclusioni

Nel contesto ciclico del 2011-2013, caratterizzato da una profonda crisi della domanda interna, ma anche nello scenario atteso per il prossimo biennio, la capacità delle imprese manifatturiere italiane di mantenere ed espandere la propria posizione sui mercati internazionali appare un fattore sempre più cruciale per la loro competitività e più in generale per la crescita dell'intera economia.

Le analisi presentate, basate su ampie basi di dati microeconomici, hanno voluto dare un contributo informativo al problema, mostrando da un lato come la capacità di intercettare la domanda estera sia stata notevolmente diversa tra le imprese, con ampi segmenti che non hanno saputo cogliere le opportunità offerte da una domanda estera comunque crescente, seppure con forti differenziazioni geografiche e settoriali; dall'altro che, soprattutto nei periodi di maggiore debolezza del ciclo, ai fini della tenuta economica e di una buona performance diviene rilevante investire in forme più complesse di internazionalizzazione.

Riferimenti bibliografici

Altomonte C., T. Aquilante e G. Ottaviano. 2012. *The Triggers of Competitiveness: the EFIGE Cross Country Report*, Bruegel Blueprint Series, Volume 17.

Barba Navaretti G. e A.J. Venables. 2004. *Multinational firms in the world economy*. Princeton University Press, Princeton

Chaney T. 2008. Distorted gravity: the intensive and extensive margins of international trade. *American Economic Review.* 98: 1707-1721.

Istat. 2013.1. *Rapporto sulla competitività dei settori produttivi*. Febbraio.

Istat. 2013.2. *Rapporto annuale: la situazione del Paese nel 2012*. Maggio.

Istat 2013.3. *9° Censimento dell'industria e dei servizi e Censimento delle istituzioni non profit. Primi risultati*. Luglio.

Melitz M.J. 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*. 71(6): 1695-1725.

Melitz M.J. e G.I.P. Ottaviano. 2008. Market size, trade and productivity. *Review of Economic Studies.* 75(1): 295–316.

Wagner J. 2011. International trade and firm performance: a survey of empirical studies since 2006. *Institute for the Study of Labor (IZA) Discussion Paper*. n. 5916, august.

La performance delle piccole e medie imprese italiane: un'analisi empirica[1]

Ernesto Cassetta[2], Marina Schenkel[3]

Sommario

Sulla base dei dati ISTAT provenienti da un panel di imprese manifatturiere nel periodo 2001-2008, il contributo si propone di indagare se e in che misura le differenze settoriali, regionali e dimensionali incidano sulla performance delle imprese italiane di dimensione inferiore ai 250 addetti. I risultati ottenuti sembrano confermare la sempre minore incidenza delle variabili settoriali e territoriali (localizzazione e appartenenza a sistemi distrettuali) come fattori esplicativi delle dinamiche di performance delle imprese.

Abstract

On the basis of a panel of manufacturing enterprises reconstructed by Istat in the period 2001-2008, this work aims at examining if and to what extent sectoral, regional and dimensional differences affect the performance of Italian enterprises with less than 250 employees. The results obtained seem to indicate an increasingly lower importance of sectoral and territorial variables (localisation and inclusion within industrial districts) in determining enterprises' performance.

Parole chiave: performance, piccole e medie imprese, settore manifatturiero

Keywords: performance, small and medium-sized enterprises, manufacturing industries

[1] Gli autori ringraziano Stefania Rossetti, Adriano Paggiaro, Sandrine Labory e i partecipanti alla XXXII Conferenza scientifica annuale dell'ASSOCIAZIONE ITALIANA DI SCIENZE REGIONALI (AISRe) "Il ruolo delle città nell'economia della conoscenza", Torino, 15-17 settembre 2011, dove è stata presentata una versione preliminare del lavoro. I dati utilizzati nel presente lavoro sono di fonte ISTAT, e relativi al Panel di bilanci delle piccole e medie imprese. Le elaborazioni sono state condotte presso il Laboratorio per l'Analisi dei Dati ELEmentari dell'Istat e nel rispetto della normativa in materia di tutela del segreto statistico e di protezione dei dati personali. I risultati e le opinioni espresse sono di esclusiva responsabilità degli autori e non costituiscono statistica ufficiale. Si precisa che le analisi sono state condotte senza utilizzare i pesi di riporto all'universo.

[2] Ricercatore (Università degli Studi di Udine), e-mail: ernesto.cassetta@uniud.it.

[3] Professore Ordinario (Università degli Studi di Udine), e-mail: marina.schenkel@uniud.it.

1. Premessa

Nel dibattito sull'evoluzione del sistema produttivo italiano e dell'industria manifatturiera sembra emergere la rilevanza di caratteristiche e comportamenti specifici a livello di impresa quali fattori determinanti della performance. La crescente integrazione a livello mondiale dei mercati reali e finanziari, il perfezionamento del mercato unico europeo e l'introduzione dell'euro, nonché la riorganizzazione su base globale delle filiere, hanno posto in evidenza debolezze del nostro sistema produttivo connesse alla capacità delle singole imprese di adattare, sotto una pluralità di aspetti, la propria offerta e le proprie scelte produttive e organizzative alla pressione concorrenziale a livello internazionale. Vari interrogativi, anche in una prospettiva di policy, derivano da questa tendenza: l'emergere di singolarità strategiche delle imprese ha reso difficilmente riconoscibili le particolarità territoriali dei processi in corso? Se sì, la riorganizzazione produttiva e l'espansione delle reti a cui partecipano le imprese ha reso insignificante il riferimento al territorio originario? Inoltre, il tramonto di modelli di successo del passato implica il declino anche delle imprese, la cui identità coincideva con tali modelli? O non è invece l'emergere dell'unicità dell'impresa nell'eterogeneità del tessuto industriale che rende possibile il mantenimento dei livelli di benessere e di sviluppo finora raggiunti?

Sotto questo profilo, l'analisi dell'evoluzione del settore manifatturiero ha tracciato una distinzione fra imprese che sono in grado di modificare i propri comportamenti in maniera coerente alla mutata estensione del mercato e condizioni di produzione, e aziende incapaci di adeguarsi a tali mutamenti. Queste ultime realizzano performance sistematicamente inferiori alle prime, sebbene non necessariamente tali da determinare la loro espulsione dal mercato (Barba Navaretti et al., 2007; Oropallo e Rossetti, 2007; Arrighetti e Traù, 2012) Dal lato delle politiche, in presenza di forti eterogeneità e varietà nei comportamenti, identificare i fattori esplicativi delle superiori performance può permettere un incremento della percentuale di imprese con maggiori capacità di adattamento, e quindi un miglioramento dei risultati del sistema produttivo nel suo complesso (Secchi e Tamagni, 2009). Tale direttrice di indagine, di carattere empirico, è facilitata dalla possibilità di accedere e trattare, in misura maggiore che in passato, panel longitudinali di dati a livello di impresa.

Il presente contributo si colloca all'interno di questo filone di ricerca, e si propone di fornire ulteriori spunti al dibattito, focalizzandosi sulla relazione fra caratteristiche specifiche delle imprese desumibili da dati di bilancio e da altre fonti, e la loro performance, espressa attraverso gli usuali indicatori di redditività. Si verificherà inoltre se e in quale misura incidano su tali relazioni le differenze settoriali, regionali e dimensionali. Più specificamente, l'analisi si propone di mettere in luce eventuali differenze territoriali e settoriali nelle evidenze finora emerse in lavori che, basandosi sugli stessi dati, hanno preso in esame la redditività delle imprese nell'intero sistema produttivo nazionale. Si verificherà se sono confermati i risultati ottenuti da Monducci et al. (2010) e Istat (2010), i quali, accanto all'importanza della dimensione e della produttività, individuano l'effetto positivo dell'attività esportativa.

L'analisi sarà riferita a un panel di piccole e medie imprese manifatturiere sempre attive nel periodo 2001-2008. Il periodo esaminato consente di analizzare la dinamica del sistema produttivo italiano in un periodo di intensa evoluzione prima della crisi economica che comincia a dispiegare i propri effetti negativi proprio a partire dal secondo trimestre del 2008. Si ricorda che la crisi economica interviene in un quadro complesso che aveva visto

crescere il Pil pro-capite del nostro Paese in media dello 0,7% annuo nel periodo 2001-2007 (Monducci et al., 2010; Istat, 2010). Contestualmente, l'incremento medio annuo del valore aggiunto del settore manifatturiero era stati pari all'1,1% l'anno, superiore a quello di Spagna e Regno Unito, in linea con quello della Francia, ma di molto inferiore a quello ottenuto dalla Germania (+3,2%).

Oltre alla presente introduzione, il lavoro è strutturato come segue. Nel paragrafo successivo, dopo un richiamo al dibattito teorico sulla dinamica industriale, si riportano i principali risultati di alcuni recenti studi empirici condotti in Italia su micro-dati di impresa. Il paragrafo 3 descrive il data base utilizzato nella verifica empirica, evidenziando alcune dinamiche delle diverse misure di redditività aziendale. I paragrafi 4 e 5 presentano i risultati di due differenti stime dei parametri che legano le varie misure di performance ad altre caratteristiche individuali delle piccole e medie imprese manifatturiere italiane incluse nel panel. L'ultimo paragrafo discute i risultati e ne analizza limiti, indicando ulteriori obbiettivi della ricerca.

2. Eterogeneità e determinanti della performance delle imprese

La persistenza di profitti eterogenei, e sistematicamente superiori a quelli che si dovrebbero realizzare in mercati concorrenziali, messa in rilievo dai dati longitudinali di impresa disponibili a partire dai primi anni '90 (Bartelsmann, 2000), non è direttamente spiegabile se si presuppone che le imprese caratterizzate da livelli di *performance* inferiori sono eliminate dai meccanismi di selezione dei mercati Tuttavia una varietà di approcci teorici, da quelli di ispirazione classica, come la funzione di produttività di Sylos Labini (1984)[4], a modelli di selezione più complessi, che si originano dal lavoro di Alchian, 1950)[5] o, più recentemente, aderenti alla teoria evolutiva dell'impresa (Dosi e Grazzi, 2006; Dosi, 2008, Dosi et al. 2011), permettono di analizzare i fattori che determinano il relativo successo nel tempo delle imprese e la loro sopravvivenza.

Alla base dell'approccio evolutivo vi è l'idea che le opportunità di crescita e di permanenza sul mercato delle imprese dipendono da loro caratteristiche specifiche, esprimibili attraverso una serie di indicatori, o misure di fitness (Secchi e Tamagni, 2009), che ne misurano la capacità di adattamento.

Il divario di performance sarebbe originato dal differenziale di efficienza, e sarebbe dunque il riflesso dell'operare dei meccanismi concorrenziali, piuttosto che il segnale di eventuali malfunzionamenti del mercato[6]. Allo stesso modo, la presenza di medesimi livelli di performance risulta compatibile con differenze significative e persistenti in molte caratteristiche dell'impresa (addetti totali, fatturato, produttività, internazionalizzazione, livello di investimenti, ricerca e sviluppo, ecc.).

Nel quadro teorico appena sintetizzato, quale variabile può essere utilizzata per misurare e confrontare la performance, e dunque la capacità di adattamento relativa delle imprese? Il

[4] Ringraziamo un anonimo referee per la segnalazione del contributo di Sylos Labini (Tronti, 2009 e 2010; Corsi e Guarini, 2007), utile in particolare nell'ambito del presente lavoro per esaminare gli effetti delle variazioni del costo del lavoro (v. oltre).

[5] Modelli teorici di selezione comprendono Jovanovic (1982); Hopenhayn (1992); Ericson e Pakes, (1995).

[6] Per un approfondimento si veda Gobbo (1997).

più comune indicatore della performance delle imprese è il livello di produttività[7]. In un contesto di crescente competizione a livello internazionale, la produttività più elevata accresce la probabilità di sopravvivenza, dando la possibilità alle imprese di praticare prezzi più bassi, di crescere in misura relativamente maggiore e di acquisire quote di mercato più ampie. Allo stesso tempo, alla maggiore produttività dovrebbe corrispondere un livello di profitto economico più elevato, che dovrebbe a sua volta consentire maggiori capacità di finanziamento degli investimenti materiali e immateriali[8], e in particolare gli investimenti in ricerca e sviluppo[9]. Inoltre, in presenza di costi fissi di entrata nei mercati esteri, i differenziali di produttività possono spiegare l'autoselezione delle imprese nei mercati dell'esportazione, riflettendosi in una maggiore propensione all'internazionalizzazione delle imprese più produttive (Melitz, 2003; Melitz e Ottaviano, 2008). L'impiego diretto di indicatori di profitto, quali ad esempio margine operativo lordo, redditività lorda, ROI o ROE, come misura di performance delle imprese, può essere criticato per la scarsa corrispondenza di tali grandezze con la relativa variabile teorica considerata (Mueller, 1990). Tuttavia, come osservato da Syverson (2010), anche la produttività può riflettere invece che l'efficienza delle singole imprese, fattori di domanda, in particolare quelli relativi al potere di mercato.

2.1 Alcuni recenti studi sul sistema manifatturiero italiano

Nei contributi che hanno indagato le recenti dinamiche del sistema manifatturiero italiano, emerge una varietà di posizioni sulle relazioni esistenti fra le variabili oggetto di indagine. In generale, le divergenze sono in larga parte riconducibili alla diversità dei dati e dei periodi esaminati.

La maggior parte delle analisi prende in considerazione misure della *performance* delle imprese diverse dalla profittabilità: il successo esportativo (Barba Navaretti et al., 2007; Guelpa, Foresti e Trenti, 2007), la crescita della produttività (Dosi et al., 2011), la crescita del fatturato e delle ore lavorate (Accetturo, Giunta e Rossi, 2011).

L'eterogeneità della *performance* delle imprese all'interno dello stesso settore (anche definito in termini stretti)[10] sembra configurare una nuova forma di dualismo, trasversale rispetto alle distinzioni settoriali e territoriali (Barba Navaretti et al., 2007; Oropallo e Rossetti, 2007; Arrighetti e Traù, 2012). Altre conclusioni però non ricevono lo stesso consenso, soprattutto per quanto riguarda l'influenza della dimensione e dell'internazionalizzazione. Con riferimento al rapporto fra dimensione e produttività, sembrano ottenere un migliore risultato le imprese di piccola e media dimensione (10-250) rispetto a quelle micro e alle grandi, secondo alcuni studi condotti a livello nazionale (Monducci et al., 2010), ma non nei mercati esteri (Brandolini e Bugamelli, 2009), nei quali le imprese risultano di dimensione maggiore

In effetti, anche se normalmente le piccole imprese sono considerate incapaci di

[7] Per una rassegna recente sugli studi empirici relativi alla relazione fra produttività e crescita si veda Syverson (2010).

[8] È da ricordare che il modello di Sylos si differenzia dagli altri perché gli investimenti dipendono crucialmente soprattutto da effetti di domanda (Sylos Labini, 1984).

[9] Dobbiamo a un anonimo *referee* l'osservazione che imprese ad alta profittabilità e bassa produttività non hanno bisogno di investire per tutelare la loro produttività. Invece imprese che sono interessate da cadute di profitti investono per migliorare la propria produttività.

[10] L'eterogeneità delle imprese porta a rilevanti effetti di composizione (Monducci et al., 2010; Istat, 2010).

internazionalizzarsi, un numero crescente di analisi empiriche sembra mettere in questione questo "fatto stilizzato", o per lo meno suggerire l'esistenza di eterogeneità e di effetti statici e dinamici, e di lungo e di breve periodo (Sterlacchini, 2001; Brancati, 2010; Monducci et al., 2010). L'internazionalizzazione delle imprese, d'altra parte, pure spesso invocata come la via maestra per l'uscita dalla crisi, viene anche considerata pericolosa per la maggiore esposizione alla concorrenza e al ciclo internazionale (Accetturo, Giunta e Rossi, 2011). Risulta peraltro ampiamente confermata la relazione fra produttività e capacità esportativa (Brandolini e Bugamelli, 2009; Dosi et al., 2011).

Le maggiori divergenze fra previsioni della teoria e risultati empirici si riscontrano relativamente al legame tra dimensione e dinamica della produttività. Anche se le imprese più grandi dimostrano un valore aggiunto per addetto costantemente maggiore (Monducci et al., 2010; Istat, 2010), il ristagno della produttività accomuna le grandi e piccole imprese italiane (Dosi et al., 2011).

Varie analisi recenti prendono in considerazione anche la relazione fra profitti e produttività, su basi dati simili a quella da noi usata (Dosi, 2008; Secchi e Tamagni, 2009; Bottazzi et al. 2010; Dosi et al., 2011). L'obiettivo è quello di indagare empiricamente la possibile presenza di un circolo virtuoso, nel quale la maggiore produttività si traduce in maggiori profitti, che a loro volta dovrebbero costituire la base per la crescita dell'impresa. Il legame positivo fra produttività e profitti viene confermato, ma non quello fra queste variabili e la crescita. Gli autori si limitano a rilevare questa evidenza empirica, riservandosi di indagare ulteriormente su un'adeguata spiegazione teorica.

Nel presente lavoro si mira a verificare se altre variabili influiscano, oltre alla produttività, sulla redditività delle imprese, sia in una relazione di medio-lungo periodo (stima cross-section) sia di breve periodo (stima panel), lasciando a successivi studi il compito di approfondire se e come la redditività sia o meno legata alla crescita dell'impresa.

3. Descrizione dei dati

L'analisi empirica effettuata si fonda su una banca dati resa accessibile dall'Istituto Nazionale di Statistica presso il Laboratorio per l'Analisi dei Dati ELEmentari (ADELE).

Il dataset raccoglie nel suo complesso dati di bilancio relativi alle 76.464 imprese, società di capitali, sempre attive nel periodo 2001-2008, integrandoli con ulteriori informazioni provenienti dal registro statistico delle imprese attive (ASIA), dalle statistiche mensili del commercio con l'estero (COE) e dalla rilevazione trimestrale su occupazione, retribuzioni e oneri sociali (OROS). La banca dati esclude le imprese di dimensione inferiore ai 10 addetti e quelle coinvolte, nel periodo di riferimento, in eventi di acquisizioni, scorpori o fusioni. Per questo motivo il database non risulta rappresentativo delle imprese di dimensione superiore ai 500 addetti, ben poche delle quali non sono state interessate dalle citate trasformazioni. E' da sottolineare che il data set include soltanto le imprese sempre attive, e non rileva i processi di entrata/uscita avvenuti nel periodo. Nel complesso, il dataset rappresenta il 32% delle imprese con almeno 10 addetti attive nel 2008, con una copertura di poco inferiore al 30% in termini

di addetti (Oropallo e Rossetti, 2012)[11]

Lo studio dei bilanci aziendali consente di analizzare la performance delle imprese in funzione di alcuni indici caratteristici della gestione aziendale e di una serie di caratteristiche, quali intensità tecnologica delle produzioni, localizzazione geografica, dimensione aziendale e appartenenza a gruppi di impresa.

Tabella 1 - Piccole e medie imprese manifatturiere per dimensione, settore di attività e localizzazione geografica

	2001	2005	2008	2001	2005	2008
Dimensione						
DIM 10-19	11.491	11.319	11.461	47,6%	46,9%	47,5%
DIM 20-49	8.755	8.790	8.563	36,3%	36,4%	35,5%
DIM 50-249	3.905	4.042	4.127	16,2%	16,7%	17,1%
Totale	**24.151**	**24.151**	**24.151**	**100,0%**	**100,0%**	**100,0%**
Settore di attività						
CA_Alimentari, Bevande, Tabacco	1.662	1.662	1.662	6,9%	6,9%	6,9%
CB_Tessili, Abbigliamento, Pelle	3.481	3.481	3.481	14,4%	14,4%	14,4%
CC_Legno, carta e stampa	2.198	2.198	2.198	9,1%	9,1%	9,1%
CD_Coke	69	69	69	0,3%	0,3%	0,3%
CE_Chimica	757	757	757	3,1%	3,1%	3,1%
CF_Farmaceutica	120	120	120	0,5%	0,5%	0,5%
CG_Gomma e materiali non metalliferi	3.172	3.172	3.172	13,1%	13,1%	13,1%
CH_Metallurgia e metallo	4.828	4.828	4.828	20,0%	20,0%	20,0%
CI_Computer, elettronica e ottica	681	681	681	2,8%	2,8%	2,8%
CJ_Apparecchi elettrici	991	991	991	4,1%	4,1%	4,1%
CK_Apparecchi nca	3.479	3.479	3.479	14,4%	14,4%	14,4%
CL_Autoveicoli e mezzi di trasporto	609	609	609	2,5%	2,5%	2,5%
CM_Altra manifattura	2.104	2.104	2.104	8,7%	8,7%	8,7%
Totale	**24.151**	**24.151**	**24.151**	**100,0%**	**100,0%**	**100,0%**
Localizzazione						
Nord Ovest	9.892	9.877	9.868	41,0%	40,9%	40,9%
Nord Est	7.572	7.577	7.580	31,4%	31,4%	31,4%
Centro	4.441	4.439	4.449	18,4%	18,4%	18,4%
Mezzogiorno	2.246	2.258	2.254	9,3%	9,3%	9,3%
Totale	**24.151**	**24.151**	**24.151**	100,0%	100,0%	100,0%
Appartenenza a un distretto o a un gruppo multinazionale a controllo italiano						
Imprese appartenenti a un distretto	11.648	11.688	11.721	48,2%	48,4%	48,5%
Imprese appartenenti a gruppo multinazionale a controllo italiano	7	28	3.154	0,0%	0,1%	13,1%

[11] Un anonimo referee calcola sulla base del Censimento dell'Industria e servizi del 2011, che gli addetti a queste imprese rappresenterebbero circa il 15.8 % dell'occupazione totale

La presente analisi considera un panel di imprese di dimensione compresa fra i 10 e i 250 addetti operanti nell'industria in senso stretto, corrispondenti a circa il 42% del totale delle imprese presenti nel dataset che include anche le imprese di servizi[12].

La tabella 1 descrive la composizione e l'evoluzione nel periodo considerato del dataset sulla base di alcune caratteristiche rilevanti, quali classe dimensionale, area geografica, settore di attività e appartenenza ad un distretto e ad un gruppo multinazionale a controllo italiano. Dal confronto degli stock, il tasso di turbolenza risulterebbe pressoché nullo. In particolare, sembrerebbe evidente la quasi totale assenza di passaggi interni per categorie dimensionali, a testimonianza dell'assenza di crescita interna. L'unico lieve aumento ha riguardato la classe dimensionale 50-249 addetti. Si rilevano infine spostamenti di circoscrizione geografica, ma non di settore. Peculiare è infine il dato sul numero di imprese appartenenti a un gruppo multinazionale (a controllo italiano) che crescono notevolmente nel periodo considerato. Poiché come si è accennato il database include le sole imprese attive non coinvolte in eventi di trasformazione societaria, tale evoluzione riflette verosimilmente una crescente propensione all'internazionalizzazione in un contesto di difficile congiuntura[13].

Nel seguito si riporta l'andamento delle principali misure di profittabilità nel periodo considerato (Figura 1), nonché dei valori assunti dalle stesse nel 2001 e nel 2008 ulteriormente suddivisi per dimensione di impresa e area geografica (Figure 2 e 3).

[12] Sempre sulla base del Censimento delll'Industria e Servizi del 2011, le imprese incluse nel panel rappresenterebbero circa il 6.6 % dell'occupazione totale

[13] Per un'analisi delle caratteristiche dell'internazionalizzazione in una circoscrizione (il Nord Est) nello stesso periodo, v. Corò, Schenkel e Volpe (2012).

Figura 1 - Andamento delle principali misure di profittabilità (valori mediani)

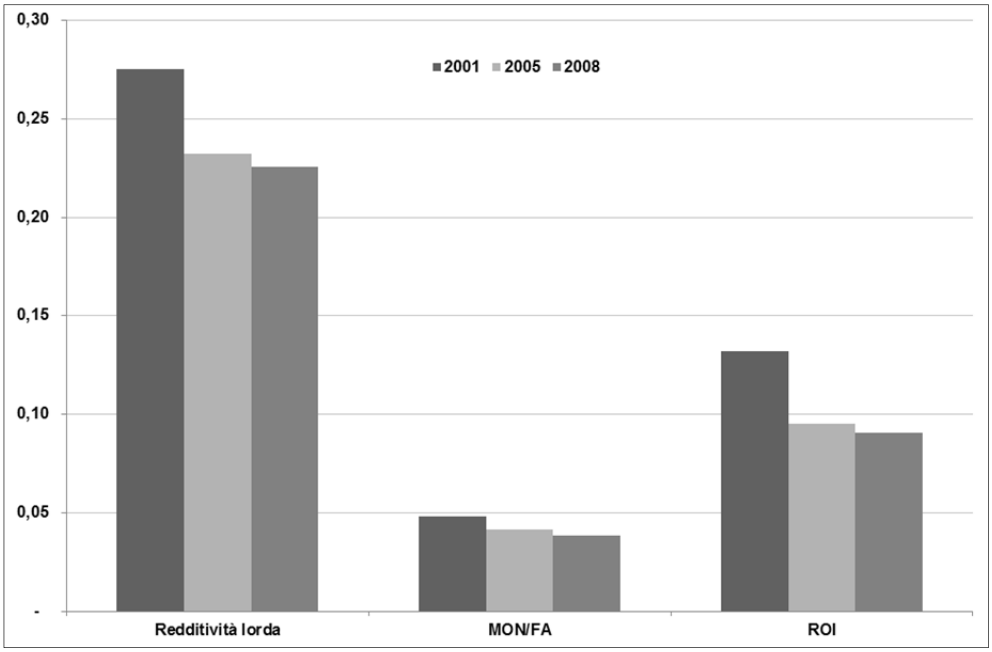


Figura 2 - Andamento delle principali misure di profittabilità per dimensione (valori mediani)

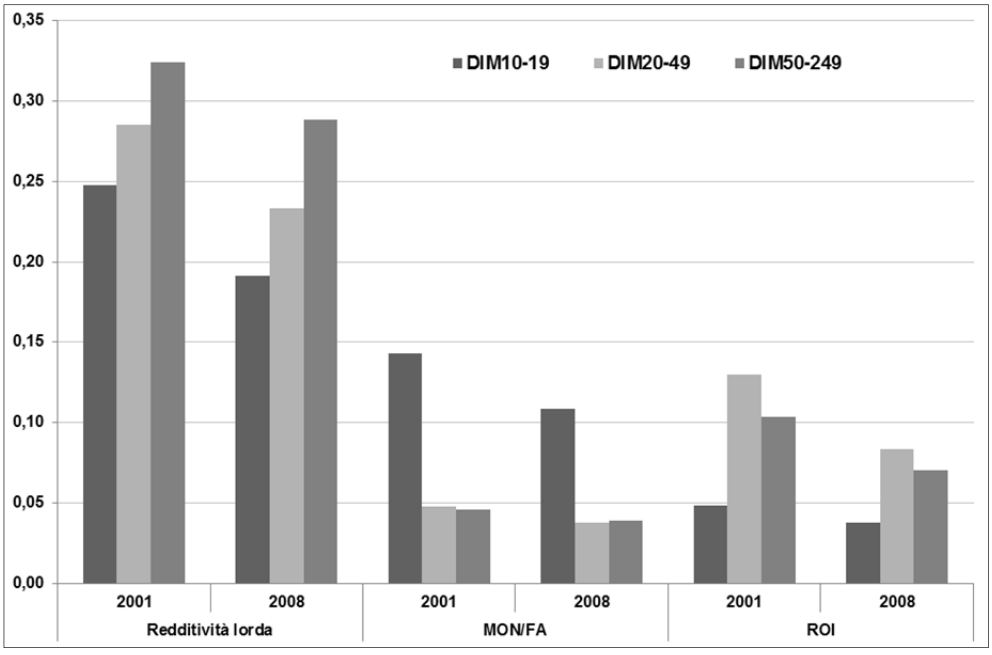
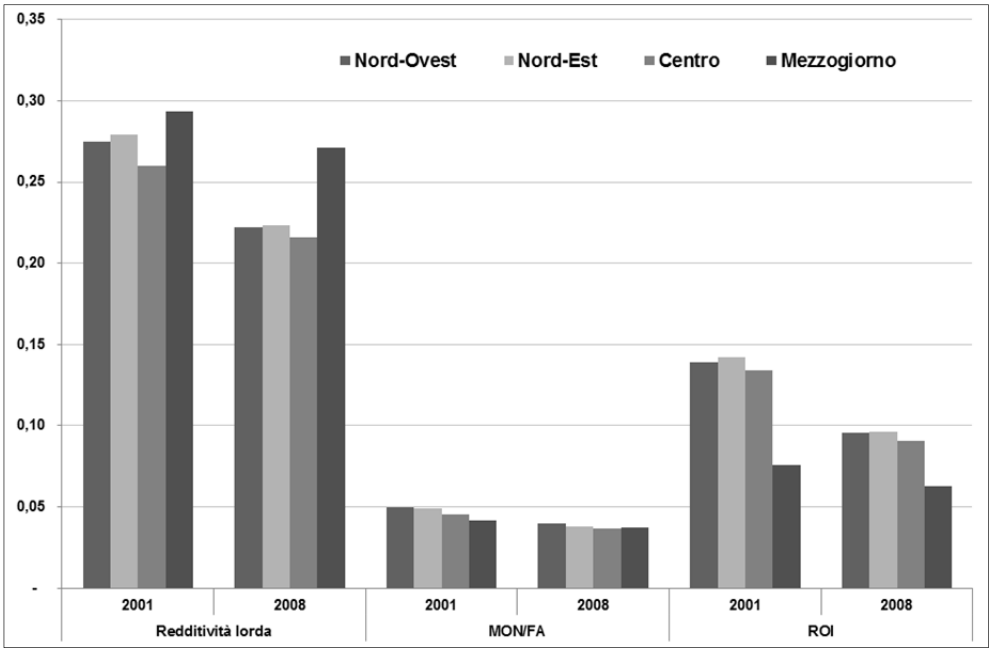


Figura 3 - Andamento delle principali misure di profittabilità per area geografica (valori mediani)



Le Figure 4 e 5 riportano la dinamica della sola redditività lorda per area geografica e classe dimensionale delle imprese considerate. La scelta di prestare particolare attenzione alla redditività lorda dipende soprattutto dall'opportunità di cogliere il reddito generato dall'impresa nella sola gestione caratteristica e industriale, depurando l'analisi dagli effetti relativi alle politiche di ammortamento, nonché alle scelte finanziarie e fiscali.

Ciò vale a maggior ragione nel caso in cui la base di dati, come in questo caso, sia costituita da informazioni provenienti da bilanci aziendali. In secondo luogo la redditività lorda consente di tenere conto nel costo del lavoro dell'input fornito dal titolare dell'impresa. Si tratta di una correzione che ha notevole rilevanza nel caso delle PMI. È da rilevare che i valori mediani della redditività lorda risultano sempre in calo nel periodo considerato, tranne nel biennio 2006-2007, con una riduzione complessiva di poco inferiore al 21,5%. Le imprese che registrano una contrazione minore sono quelle meridionali (-13,6%) e quelle di maggiore dimensione (-11,2%). Il calo complessivo della redditività lorda è stato invece piuttosto uniforme nelle altre aree del Paese, mentre ha colpito in modo particolare le imprese con dimensione inferiore ai 20 addetti.

Figura 4 - Evoluzione della redditività lorda per classe dimensionale (valori mediani, 2001=100)

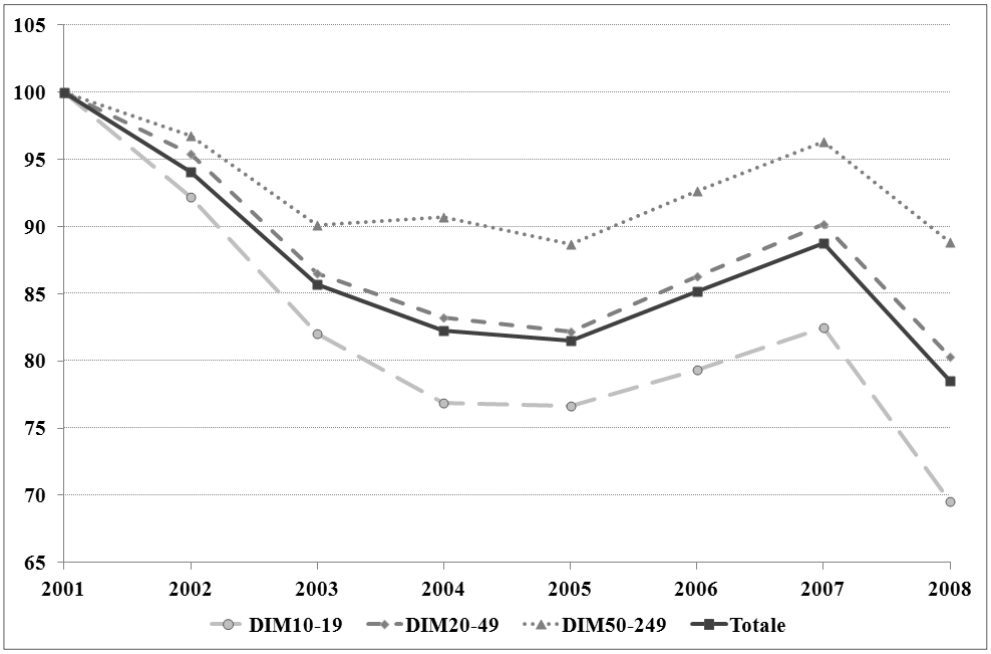
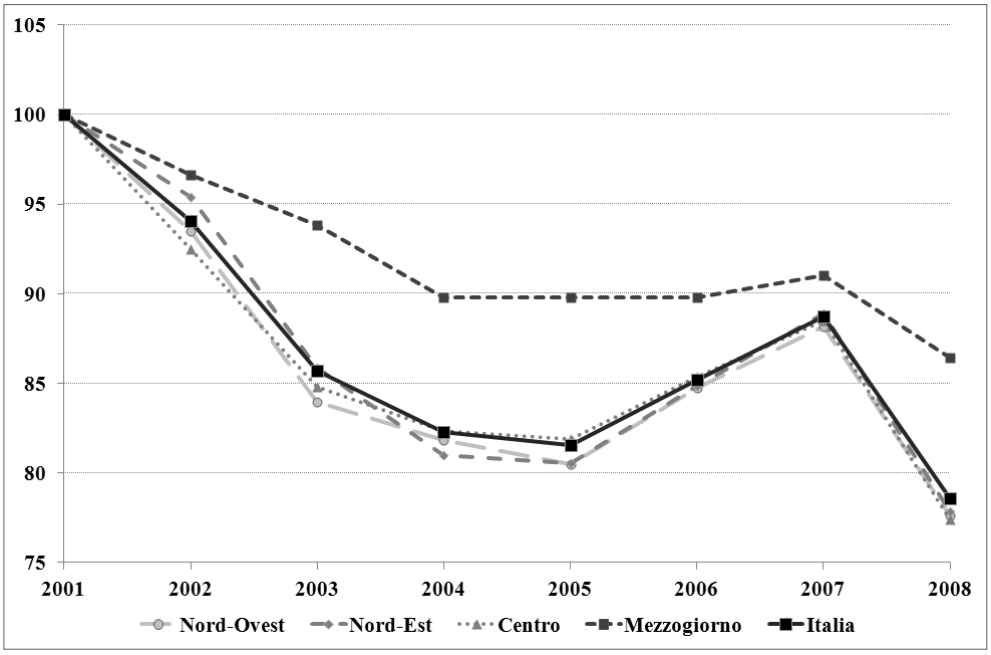


Figura 5 - Evoluzione della redditività lorda per area geografica (valori mediani, 2001=100)



4. L'analisi econometrica

4.1 La stima cross-section

La regressione cross-section è stata effettuata sulla base della seguente specificazione empirica:

$$\Delta\pi_{i,2001\text{–}08} = \alpha + \beta\pi_{i,2001} + \beta X_i + \mu_D + \gamma_S + \delta_G + \vartheta_{Distr} \qquad (1.1)$$

La specificazione del modello cross-section considera come variabile dipendente il tasso di crescita (logaritmico) della redditività dell'impresa i nel periodo 2001-2008, mentre le variabili esplicative sono il valore iniziale della variabile dipendente in logaritmo, il logaritmo del livello medio delle variabili specifiche alla singola impresa e alcune dummy, descritte di seguito[14].

Le variabili esplicative X_i, che corrispondono a quelle considerate negli studi sulla performance delle imprese italiane sopra ricordati, sono:

- la produttività del lavoro (PROD), misurata dal rapporto fra valore aggiunto e addetti totali[15];
- il grado di internazionalizzazione (INTER), misurato dal rapporto fra il totale delle esportazioni e il fatturato complessivo realizzato;
- il livello di integrazione verticale (INTVER), misurato dal rapporto fra valore aggiunto e fatturato (cd. indice di Adelman);
- gli investimenti totali lordi (INVTOTLORDI), misurati dalla variazione sull'anno precedente del totale delle immobilizzazioni materiali e immateriali;
- il costo del lavoro per addetto (COSTOLAVORO_ADDETTO);
- il fatturato (FATTURATO);
- il mark-up (MARK_UP), definito dal rapporto fra i ricavi totali e i costi totali di produzione.

Si sono inoltre inserite dummy relative alla classe dimensionale, alla circoscrizione geografica e al settore di attività dell'impresa. In relazione a quest'ultima variabile, le imprese sono state aggregate in macro-settori omogenei sulla base della classificazione Ateco a 2 cifre. Si è infine considerata l'appartenenza o meno dell'impresa ad un distretto mentre è stata omessa quella ad un gruppo multinazionale a controllo italiano in ragione della peculiare crescita del valore della variabile concentrata nell'ultimo biennio.

Nella seguente tabella (Tavola 2) si riportano i risultati della stima.

[14] La specificazione adottata si ispira alle regressioni "à la Barro" (Barro, 1991), originate nell'ambito della teoria della crescita, e mira a considerare aspetti di lungo periodo.

[15] Tale variabile è correlata ad altre misure di produttività, come la Produttività Totale dei Fattori (Aiello et al., 2012), ed è comunque preferita anche da altri autori (Dosi, 2008).

Tavola 2 - Redditività delle imprese italiane: risultati delle regressione cross-section

	Variabile dipendente		
	Red_lorda	MON_FA	ROI
LRed_lorda01	-0.117***	-	-
	(0.00159)		
LMON_FA01	-	-0.123***	-
		(0.00139)	
LROI01	-		-0.0990***
			(0.00148)
LnPROD	0.0747***	0.162***	0.0430***
	(0.00542)	(0.00590)	(0.00719)
LnINTER	0.00155***	0.00107*	0.000352
	(0.000400)	(0.000427)	(0.000552)
LnINTVER	-0.0611***	0.0782***	0.0344***
	(0.00362)	(0.00421)	(0.00507)
LnINVTOTLORDI	0.00782***	-0.0000200	-0.0283***
	(0.000627)	(0.000671)	(0.000875)
LnCOSTOLAVORO_ADDETTO	-0.0804***	-0.158***	-0.0531***
	(0.00636)	(0.00640)	(0.00793)
LnFATTURATO	0.00134	-0.00948***	0.0251***
	(0.00256)	(0.00269)	(0.00338)
LnMARK_UP	0.311***	0.123***	0.0643*
	(0.0205)	(0.0214)	(0.0270)
DIM1019	-0.00248	-0.000341	0.0436***
	(0.00497)	(0.00543)	(0.00688)
DIM2049	-0.00376	-0.00225	0.0159***
	(0.00338)	(0.00370)	(0.00473)
DISTRETTO	-0.00216	-0.00253	-0.00442
	(0.00181)	(0.00193)	(0.00249)
NORDEST	-0.0000285	-0.00187	-0.000652
	(0.00205)	(0.00219)	(0.00283)
CENTRO	0.00150	0.00380	-0.000755
	(0.00259)	(0.00273)	(0.00353)
MEZZ_ISOLE	0.00739*	-0.00448	-0.0227***
	(0.00373)	(0.00408)	(0.00532)
Dummy settoriali	Sì	Sì	Sì
Costante	-0.350***	-0.299***	-0.264***
	(0.0478)	(0.0446)	(0.0575)
N	11.547	10.958	11.029
adj. R^2	0.329	0.420	0.321

Nota: Errori standard robusti in parentesi con *, ** e *** che indicano rispettivamente un livello di significatività a 1%, 5% e 10%.

Il segno negativo del logaritmo del valore iniziale della variabile, sintomo di regressione verso la media, è del tutto coerente con le aspettative (Dosi et al., 2011). Analogamente non sorprendono il segno positivo della produttività e del mark-up e quello negativo del costo del lavoro. La quota delle esportazioni sul Valore aggiunto è significativa e positiva, tranne quando la variabile dipendente è il ROI, caso in cui non è significativa. Analogamente gli investimenti totali lordi sono significativi e con segno positivo, ma se la variabile dipendente è il ROI il segno è negativo. Infine il segno della variabile che riflette le scelte di delocalizzazione o integrazione verticale (Valore Aggiunto/Fatturato) è positivo se la variabile dipendente è la Redditività Lorda/Valore aggiunto, ma è negativo negli altri due casi (variabile dipendente Margine Netto/Fatturato o ROI). La dummy per classe dimensionale è significativa, con segno positivo (variabile omessa classe 49-250), soltanto quando la variabile dipendente è il ROI. Le dummies che indicano l'appartenenza a un distretto non sono mai significative. Alcune dummies settoriali (variabile omessa Coke e combustibili) sono significative nei riguardi di Margine Operativo Netto/Fatturato e ROI, mai invece su Redditività lorda/Valore Aggiunto. La differenza dei punti di vista dai quali viene valutata la profittabilità delle imprese a seconda della variabile dipendente prescelta (Traù, 2013) è la spiegazione immediata della difformità di questi risultati, che andrà ulteriormente indagata nel proseguimento della ricerca

5. La stima panel

La regressione è stata effettuata sulla base della seguente specificazione empirica:

$$\pi_{i,t} = \mu_t + \beta X_{i,t} + \gamma z_i + \alpha_i + \varepsilon_{i,t} \qquad (1.2)$$

con i = 1, …, n; t = 1, …, T.

Nell'equazione 1.2 π_i indica la variabile di redditività dell'impresa i al tempo t, e X_i è il vettore delle variabili esplicative individuali di impresa al tempo t sopra descritte, con l'aggiunta della dummy MULT relativa all'appartenenza dell'impresa ad un gruppo multinazionale italiano. Nel presente paragrafo ci si focalizzerà in particolare sulla redditività lorda che, come in precedenza accennato, consente di cogliere in misura relativamente migliore il reddito generato dall'impresa nella sola gestione caratteristica e industriale, depurando l'analisi dagli effetti relativi alle politiche di ammortamento, nonché alle scelte finanziarie e fiscali. I risultati relativi alle altre misure di redditività sono comunque riportati in appendice.

Nella tabella 3, sono confrontati i parametri stimati nelle regressione panel attraverso il modello a effetti fissi (FE) e a effetti causali RE). Dato il risultato del test di Hausman, è da notare che le stime RE non sono consistenti. A fini di confronto si riportano anche le stime ottenute attraverso il modello dei minimi quadrati (OLS).

Tavola 3 - Performance delle imprese italiane: risultati delle regressione panel. Variabile dipendente: Redditività Lorda

VARIABILE DIPENDENTE	FE	RE	RE (con dummy settoriali)	OLS
	LnRed_lorda	LnRed_lorda	LnRed_lorda	LnRed_lorda
LnPROD	1.641***	1.695***	1.695***	1.680***
	(0.0419)	(0.0306)	(0.0306)	(0.0239)
LnINTER	-0.000190	-0.00202	-0.00000171	-0.000826
	(0.00270)	(0.00141)	(0.00146)	(0.00143)
LnINTVERT	0.339***	0.166***	0.166***	0.115***
	(0.0306)	(0.0158)	(0.0160)	(0.0115)
LnINVTOTLORDI	-0.00446**	0.00136	0.000790	0.0122***
	(0.00143)	(0.00129)	(0.00129)	(0.00152)
LnCOSTOLAVORO_ADDETTO	0.337***	0.239***	0.238***	0.186***
	(0.0222)	(0.00991)	(0.00987)	(0.00906)
LnFATTURATO	-1.831***	-1.819***	-1.817***	-1.779***
	(0.0509)	(0.0350)	(0.0354)	(0.0258)
LnMARK_UP	0.415***	0.447***	0.446***	0.416***
	(0.0496)	(0.0865)	(0.0863)	(0.0970)
DIM1019	0.0928***	0.0857***	0.0841***	0.0350
	(0.0213)	(0.0183)	(0.0183)	(0.0179)
DIM2049	0.0625***	0.0880***	0.0863***	0.0678***
	(0.0139)	(0.0110)	(0.0110)	(0.0108)
DISTRETTO	0.0299	-0.00257	-0.000166	0.00366
	(0.0302)	(0.00651)	(0.00657)	(0.00600)
MULT	0.000698	-0.00610	-0.00578	-0.0114
	(0.0124)	(0.0114)	(0.0114)	(0.0125)
NORDEST	-0.0234	0.0221**	0.0216**	0.0134*
	(0.0705)	(0.00734)	(0.00732)	(0.00668)
CENTRO	-0.0851	0.0256**	0.0264**	0.0212*
	(0.137)	(0.00979)	(0.00978)	(0.00867)
MEZZ_ISOLE	-0.132	0.0642***	0.0611***	0.0503***
	(0.0916)	(0.0138)	(0.0139)	(0.0128)
Dummy Settoriali			SI	SI
_cons	-4.974***	-4.520***	-4.586***	-4.130***
	(0.377)	(0.174)	(0.201)	(0.167)
N	51.574	51.574	51.574	51.574
adj. R^2				0.587
R^2 within	0.4519	0.4496	0.4497	
R^2 between	0.5743	0.5982	0.5996	
R^2 overall	0.5685	0.5889	0.5898	

Nota: Errori standard robusti in parentesi con *, ** e *** che indicano rispettivamente un livello di significatività a 1%, 5% e 10%.

Il test di Hausman rifiuta l'ipotesi di non correlazione fra le variabili esplicative e le variabili omesse.

La produttività del lavoro esercita un'influenza positiva sulla redditività lorda di impresa, e il costo del lavoro un'influenza negativa. Il segno negativo della variabile internazionalizzazione potrebbe segnalare l'effetto della maggiore esposizione alla competizione globale. La variabile mark-up presenta segno positivo, come pure la variabile fatturato, indicando la possibile presenza di potere di mercato e economie di scala. Emerge l'influenza evidente delle scelte di integrazione verticale. Il coefficiente della variabile valore aggiunto/fatturato risulta infatti significativo e con segno positivo. Il segno della relazione fra la redditività lorda e l'ammontare annuo di investimenti è negativo nella stima a effetti fissi[16].

Le dummy di circoscrizione sono quasi sempre significative e con segno positivo (variabile omessa Nord Ovest)[17], mentre L'appartenenza dell'impresa ad un gruppo multinazionale a controllo nazionale, e l'appartenenza a un distretto non sono significative[18].

Le dummies dimensionali (classe di addetti 10-19 e 20-49) risultano significative e di segno positivo (variabile omessa classe 50-249). Questo risultato non può non sorprendere, dato che sia il livello che la dinamica della profittabilità sembrano penalizzare la piccola dimensione (si vedano le Figure 3 e 4). Anche se il punto richiede di essere ulteriormente approfondito, può essere interpretato come riflesso dei limiti alla crescita delle PMI derivanti da minore tassazione, minori vincoli normativi e altre "diseconomie di scala" istituzionali, al netto dell'effetto del minor costo del lavoro e delle altre variabili esplicative introdotte nella stima.

6. Conclusioni

Per una corretta interpretazione dei risultati ottenuti, è opportuno ricordare alcuni limiti dei dati e delle metodologie utilizzate. Innanzitutto il panel considerato esclude le imprese soggette a fenomeni di fusione e acquisizione e quelle che hanno cessato l'attività. Dato che le imprese caratterizzate da performance inferiori alla media sono presumibilmente maggiormente interessate da tali fenomeni, ne potrebbe derivare *selection bias* a favore delle imprese di maggiore redditività. Ciò vale in particolar modo per le imprese di ridotta dimensione che hanno tassi di mortalità relativamente superiori e sono maggiormente coinvolte in operazioni di trasformazione. Uno dei risultati riportati nei paragrafi che precedono, e cioè la maggiore redditività delle piccole imprese a parità di altre condizioni, potrebbe dunque essere in gran parte attribuito al fatto che risultano incluse nel campione soltanto le "migliori", o almeno le più stabili, fra queste. D'altra parte potrebbe essere proprio la maggiore redditività (e produttività) a consentire fenomeni di crescita esterna, e

[16] Questi risultati si confermano anche quando la variabile dipendente è costituita da altri indicatori di profittabilità, quali il ROI e il Margine Operativo Netto/Fatturato (i risultati di queste elaborazioni sono riportati in appendice). È comunque da tener presente che le stime a effetti Random sono inconsistenti, secondo il test di Hausman.

[17] Risultano non significative le dummies settoriali, incluse nella regressione a effetti random (variabile omessa Coke e prodotti petroliferi). Ricordiamo ancora che secondo il test di Hausman queste stime sono inconsistenti.

[18] I risultati divergono se cambia la misura di profittabilità. In particolare se la variabile dipendente è il ROI, la dummy Nord Est è significativa e positiva, e le dummies Mezzogiorno e appartenenza a un gruppo multinazionale a controllo italiano sono significative e negative. Se la variabile dipendente è il Margine Operativo Netto/Fatturato sono significative le dummies Nord Est e Centro, e appartenenza a un distretto.

dunque il panel potrebbe escludere anche le imprese a maggiore performance. In ogni caso il panel per costruzione non consente di studiare l'entrata e l'uscita delle imprese

Il secondo *caveat* è relativo al fatto che non sarebbe lecito sulla base di questi risultati sostenere un nesso causale preciso fra le diverse misure di redditività e le variabili esplicative utilizzate nell'analisi, al di là della correlazione. Rimane quindi da spiegare il senso dell'associazione fra *performance* e scelte strategiche d'impresa (investimenti materiali e immateriali, esportazioni, integrazione verticale) che già si rilevava passando in rassegna i risultati ottenuti da altri autori. Per esempio, le imprese più efficienti sono quelle che esportano o, viceversa, le imprese esportatrici guadagnano in efficienza perché esposte alla concorrenza internazionale (Oropallo e Rossetti, 2012; Ferrante et al., 2011; Castellani e Giovannetti, 2011)? Questa ambiguità dovrebbe suggerire una certa cautela nell'esprimere ricette di policy troppo unilaterali.

Un problema sostanzialmente irrisolto, che sarà necessario approfondire a un ulteriore stadio dell'analisi, si evidenzia quando si prendono in considerazione le difformità di risultati che si ottengono usando diverse misure della profittabilità. Nell'interpretazione inoltre bisogna tener presente che i coefficienti dell'analisi cross section non sono direttamente confrontabili con quelli delle analisi panel, dato che nell'analisi cross-section, la variabile dipendente è costituita dal tasso di incremento della redditività, di cui si stima l'elasticità rispetto al suo valore iniziale e ai valori medi delle variabili esplicative, mentre nelle seconde è il livello della redditività che viene regredito sul livello delle variabili esplicative.

Tenendo presenti queste cautele, alcuni risultati sembrano comunque comuni ai due tipi di analisi.

In particolare, la non significatività dell'appartenenza ai distretti mette in luce la frammentazione delle filiere produttive e la diffusione di reti internazionali di imprese, che sembrano essere tratti caratteristici dell'attuale contesto economico globale. Tali fattori rendono assai più complessa che in passato l'analisi del legame fra impresa e territorio.

In generale è prematuro affermare che l'influenza del territorio è stata cancellata dai processi di globalizzazione, lasciando come unica protagonista la singola impresa, portatrice di eterogeneità avulse dalla sua localizzazione. Tuttavia questa influenza si esplica in forme diverse dal passato, come testimonia una crescente evidenza empirica[19].

In secondo luogo, due variabili sembrano manifestare un legame robusto con la profittabilità (comunque misurata): la produttività e il costo del lavoro. In particolare il coefficiente della produttività, sia considerata come tasso di variazione nella stima *cross-section*, che come livello in quella *panel*, risulta sempre significativo e positivo, in armonia con i risultati teorici e empirici citati nel par. 2 e 3. Il coefficiente del costo del lavoro, pur significativo e con il previsto segno negativo, è tuttavia molto inferiore all'unità. Dato che effetti di domanda sono difficilmente misurabili sulla base di dati micro, e che la variabile investimenti materiali e immateriali è inserita nella stima, una spiegazione di questo risultato può esser identificata nella riorganizzazione della produzione a cui le imprese ricorrono per contrastare la perdita di profitto derivante dall'aumento dei salari[20].

Incerti risultano invece i risultati per quanto riguarda le variabili "strategiche"

[19] Per una sintesi v. a es. Corò, Schenkel e Volpe (2012).

[20] Ciò significa che a parità di produttività le imprese con costo del lavoro maggiore hanno una redditività minore, ma tuttavia riescono a contrastare in una certa parte lo svantaggio con politiche di recupero dei margini distributivi. Si tratta di una parte dell'"effetto Ricardo" nella funzione di produttività di Sylos Labini.

fondamentali: investimenti, internazionalizzazione, integrazione verticale. L'instabilità dei risultati, a seconda delle variabili dipendenti e dei metodi di stima, impongono di studiare una specificazione dinamica del modello che approfondisca le complesse relazioni fra le variabili e i nessi di causalità. Si potranno inoltre fornire ulteriori elementi empirici per lo studio del nesso fra profittabilità/produttività e crescita.

Appendice

Tavola 4 - Performance delle imprese italiane: risultati delle regressione panel. Variabile dipendente: ROI

	FE	RE	RE (con dummy settoriali)	OLS
LnPROD	1.914***	1.604***	1.633***	1.223***
	(0.0809)	(0.0537)	(0.0542)	(0.0438)
LnINTER	0.00394	0.00437	-0.00551*	-0.00269
	(0.00420)	(0.00243)	(0.00250)	(0.00262)
LnINTVERT	0.694***	0.525***	0.493***	0.441***
	(0.0676)	(0.0316)	(0.0321)	(0.0281)
LnINVTOTLORDI	-0.0587***	-0.0987***	-0.0965***	-0.156***
	(0.00234)	(0.00229)	(0.00228)	(0.00294)
LnCostolavoro_addetto	0.581***	0.0733***	0.0837***	0.0998***
	(0.0347)	(0.0162)	(0.0162)	(0.0167)
LnFatturato	-1.664***	-1.404***	-1.443***	-1.094***
	(0.0912)	(0.0609)	(0.0611)	(0.0485)
LnMark_up	0.939***	1.021***	1.005***	1.084***
	(0.108)	(0.113)	(0.117)	(0.159)
DIM1019	0.0859*	0.174***	0.183***	0.156***
	(0.0338)	(0.0294)	(0.0294)	(0.0323)
DIM2049	0.0628**	0.105***	0.112***	0.0779***
	(0.0241)	(0.0193)	(0.0193)	(0.0209)
DISTRETTO	0.124	0.00643	-0.00995	-0.0102
	(0.0722)	(0.0122)	(0.0121)	(0.0113)
MULT	-0.0889***	-0.116***	-0.119***	-0.154***
	(0.0225)	(0.0217)	(0.0217)	(0.0255)
NORDEST	-0.108	0.0364**	0.0398**	0.0412***
	(0.256)	(0.0138)	(0.0136)	(0.0125)
CENTRO	-0.177	0.0218	0.0387*	0.0132
	(0.208)	(0.0175)	(0.0173)	(0.0159)
MEZZ_ISOLE	0.0611	-0.400***	-0.325***	-0.294***
	(0.217)	(0.0281)	(0.0281)	(0.0276)
Dummy Settoriali			SI	SI
Costante	-13.00***	-4.381***	-4.879***	-3.551***
	(0.539)	(0.248)	(0.289)	(0.263)
N	50479	50479	50479	50479
adj. R^2				0.293
R^2 within	0.3849	0.3492	0.3518	
R^2 between	0.0609	0.2125	0.2320	
R^2 overall	0.1063	0.2559	0.2733	

Errori standard fra parentesi. * p < 0.05, ** p < 0.01, *** p < 0.001

Il test di Hausman rifiuta l'ipotesi di non correlazione fra le variabili esplicative e le variabili omesse.

Tavola 5 - Performance delle imprese italiane: risultati delle regressione panel. Variabile dipendente: Margine Operativo Netto/Fatturato

	FE	RE	RE (con dummy settoriali)	OLS
LnPROD	1.990***	1.892***	1.909***	1.737***
	(0.0495)	(0.0340)	(0.0345)	(0.0371)
LnINTER	0.00499	0.00575***	-0.000167	0.000101
	(0.00323)	(0.00160)	(0.00167)	(0.00169)
LnINTVERT	1.187***	0.990***	0.980***	0.925***
	(0.0358)	(0.0180)	(0.0185)	(0.0211)
LnINVTOTLORDI	-0.00778***	-0.0248***	-0.0230***	-0.0381***
	(0.00173)	(0.00159)	(0.00159)	(0.00183)
LnCostolavoro_addetto	0.199***	0.0218	0.0272*	0.0272*
	(0.0261)	(0.0113)	(0.0113)	(0.0111)
LnFatturato	-1.844***	-1.796***	-1.806***	-1.667***
	(0.0580)	(0.0384)	(0.0392)	(0.0398)
LnMark_up	0.530***	0.472***	0.467***	0.578***
	(0.0604)	(0.102)	(0.104)	(0.147)
DIM1019	0.0719**	0.102***	0.106***	0.0787***
	(0.0274)	(0.0224)	(0.0224)	(0.0223)
DIM2049	0.0398*	0.0537***	0.0571***	0.0361*
	(0.0203)	(0.0144)	(0.0144)	(0.0143)
DISTRETTO	0.117*	0.0268***	0.0158*	0.0113
	(0.0577)	(0.00768)	(0.00769)	(0.00718)
MULT	-0.0170	-0.0187	-0.0201	-0.00743
	(0.0176)	(0.0165)	(0.0165)	(0.0183)
NORDEST	-0.101	0.0242**	0.0233**	0.0198*
	(0.186)	(0.00853)	(0.00849)	(0.00795)
CENTRO	0.00784	0.0503***	0.0505***	0.0404***
	(0.212)	(0.0109)	(0.0109)	(0.0101)
MEZZ_ISOLE	0.252	-0.0821***	-0.0529**	-0.0539**
	(0.172)	(0.0179)	(0.0180)	(0.0173)
Dummy Settoriali			SI	SI
Costante	-6.994***	-3.806***	-4.182***	-3.633***
	(0.442)	(0.183)	(0.218)	(0.190)
N	50456	50456	50456	50456
adj. R^2				0.566
R^2 within	0.4761	0.4701	0.4707	
R^2 between	0.4920	0.5580	0.5645	
R^2 overall	0.5073	0.5593	0.5646	

Errori standard fra parentesi. * p < 0.05, ** p < 0.01, *** p < 0.001

Il test di Hausman rifiuta l'ipotesi di non correlazione fra le variabili esplicative e le variabili omesse.

Riferimenti bibliografici

Accetturo A., A. Giunta e S. Rossi. 2011. Le imprese italiana tra crisi e nuova globalizzazione. *l'Industria. Rivista di Economia e Politica Industriale*. 1: 145-164.

Alchian A.A. 1950. Uncertainty, Evolution and Economic Theory. *Journal of Political Economy*. 58: 211–221.

Arrighetti A. e F. Traù. 2012. Far from the madding crowd. Sviluppo delle competenze e nuovi percorsi evolutivi delle imprese italiane. *l'Industria. Rivista di Economia e Politica Industriale*. 1: 7-59.

Barba Navaretti G., M. Bugamelli, R. Faini, F. Schivardi e A. Tucci. 2007. *Le imprese e la specializzazione produttiva dell'Italia. Dal macrodeclino alla microcrescita?* Rapporto preparato per conto della Fondazione Rodolfo Debenedetti per il Convegno: I vantaggi dell'Italia, Roma 22 marzo.

Barro R.. 1991. Economic Growth in a Cross-Section of Countries. *Quarterly Journal of Economics*. 106(2): 407-443.

Bartelsman E. J., e M. Domas (2000) "Understanding Productivity; Lessons from Longitudinal Microdata", *Journal of economic literature*, 38: 569-594.,

Bianchi P. e C. Pozzi. 2010. Crisi economica e politica industriale. In Bianchi P. e C. Pozzi, (a cura di). *Le politiche industriali alla prova del futuro. Analisi per una strategia nazionale*. Bologna: Il Mulino.

Bottazzi G., G. Dosi, N. Jacoby, A. Secchi e F. Tamagni. 2010. Corporate performances and market selection: some comparative evidence. *Industrial and Corporate Change*, 19: 1953-1996

Brancati R. 2010. *Fatti in cerca di idee*. Roma, Donzelli.

Brandolini A. e M. Bugamelli, a cura di. 2009. Rapporto sulle tendenze del sistema produttivo italiano. *Questioni di Economia e Finanza (Occasional papers)*. N. 45. Aprile. Banca d'Italia. Roma.

Cassetta E. e M. Schenkel. 2011. *Redditività nel sistema manifatturiero italiano: analisi di un panel di imprese (2001-2008)*. XXXII Conferenza scientifica annuale, Associazione Italiana di Scienze Regionali (AISRe). Il ruolo delle città nell'economia della conoscenza. Torino, 15-17 settembre.

Castellani D. e G. Giovannetti. 2011. Productivity and the International Firm: Dissecting Heterogeneity. *Journal of Economic Policy Reform*. 13: 25-42.

Corò G., M. Schenkel e M. Volpe (2012) "Apertura internazionale e cambiamento strutturale nel sistema manifatturiero del Nord Est. *L'Industria*, 33: 193-204,

Corsi M. e G. Guarini (2007) "La fonction de productivité de Sylos Labini: aspects theoriques et empiriques", *Revue d'economie industrielle 118: 55-78.*

Dosi G. 2008. Regolarità statistiche nell'evoluzione dei settori industriali: l'evidenza empirica e le sfide per la teoria. *l'Industria. Rivista di Economia e Politica Industriale., XXIX*: 185-219.

Dosi G. e M. Grazzi (2006)"Technologies as problem-solving procedures and technologies as input-output relations: some perspectives on the theory of production", *Industrial and Corporate Change, 15: 73-102.*

Dosi G., M. Grazzi, C. Tomasi e A. Zeli. 2011. L'industria manifatturiera negli ultimi due decenni prima della crisi: le micro-dinamiche sottostanti ai trend aggregati. *Economia e Politica Industriale*. 38(1): 63-95.

Ferrante M.R., M. Freo e A. Viviani. 2011. Is there a heterogeneous post-entry effect of exporting on firm productivity? Evidence from a panel of Italian manufacturing firms. Relazione presentata al Convegno: *L'analisi dei dati di impresa per la conoscenza del sistema produttivo italiano: il ruolo della statistica ufficiale*. Istat, Roma novembre.

Foresti G., F. Guelpa e S. Trenti. 2008. I distretti industriali alla prova della palingenesi. *l'Industria. Rivista di economia e politica industriale*, 3: 547-570.

Geroski P. e A. Jacquemin. 1988. The Persistence of Profits: A European Perspective. *Economic Journal*. 98: 375-89.

Gobbo F. 1997. *Il mercato e la tutela della concorrenza*. Bologna: Il Mulino.

Hopenhayn H. 1992. Entry, exit, and firm dynamics in long run equilibrium. *Econometrica*. 60: 1127–1150.

Istat. 2010. *Rapporto Annuale 2009. La situazione economica del paese*. Roma 26 Maggio.

Jovanovic B. 1982. Selection and the Evolution of Industry, *Econometrica.* 50(3): 649-670.

Melitz M. J. (2003) "The Impact of Trade on Intra-Industry Reallocations and Agggregate Industry Productivity" *Econometrica*, 71: 1695-1725.

Melitz M. J. E Ottaviano G. I. P.(2008) "Market Size, Trade, and Productivity" *Review of Economic Studies*, 75: 295-316.

Monducci R., P. Anitori, F. Oropallo e C. Pascucci. 2010. Crisi e ripresa del sistema industriale italiano: tendenze aggregate ed eterogeneità delle imprese. *Economia e Politica Industriale*. 37(3): 93-116.

Mueller D. C. 1990. *The Dynamics of Company Profits*. Cambridge: Cambridge University Press.

Oropallo F. e S. Rossetti. 2007. Entrepreneurs' Behaviour and performance: An Empirical Analysis on Italian Firms. *Rivista di Politica Economica*. 97(3): 1-18.

Oropallo F. e S. Rossetti. 2012. *Data integration and Productivity Estimation at a Firm Level*. Relazione presentata alla 46° Riunione Scientifica della SIS. Roma, giugno 2012.

Oropallo F. e S. Rossetti. 2011. Esportazioni e produttività, un'analisi panel sulle imprese manifatturiere italiane. Relazione presentata al Convegno: *L'analisi dei dati di impresa per la conoscenza del sistema produttivo italiano: il ruolo della statistica ufficiale*. Istat. Roma Novembre 2011.

Secchi A. e F. Tamagni. 2009. Un'analisi empirica delle relazioni tra crescita d'impresa, produttività e profittabilità, In Rondi L. e F. Silva, a cura di. *Produttività e cambiamento nell'industria italiana. Indagini quantitative*. Bologna: Il Mulino. pp.39-65.

Sylos Labini P. (1984) *Le forze dello sviluppo e del declino*, Bari-Roma, Laterza,

Syverson C. 2010. What determines productivity? *NBER Working Paper,* No. 15712.

Traù F. (2013) "L'eterogeneità dei risultati economici delle imprese negli anni della globalizzazione e della crisi" *QA-Rivista dell'Associazione Rossi-Doria*, n. 4: 7-42.

Tronti L (2009) "La crisi di produttività dell'economia italiana: scambio politico ed estensione del mercato" *Economia & Lavoro*, XLIII: 139-157.

Tronti L (2010) "La crisi di produttività dell'economia italiana: modello contrattuale e incentivi ai fattori", *Economia & Lavoro*, XLIV: 47-70.

Managing census complexity through highly integrated web systems[1]

Federico Benassi,[2] Mauro Bruno[3], Maura Giacummo,[4] Marco Silipo[5], Giulia Vaste,[6] Donatella Zindato[7],

Abstract

During 2011 General Population and Housing Census methodological innovations towards the planning of a register-based census were introduced. Such innovations allowed to reach a satisfactory cost-benefit balance but increased the survey complexity and the risk of errors.
Istat focus on web technologies was aimed at performing an innovative census, both in terms of methodology and costs and of data dissemination timeliness.
This paper shows how a highly integrated web information system was used to manage a census workflow involving several actors and multiple phases and integrating different data sources.
The success of 2011 Census led Istat to adopt such technological infrastructure as the core of the 'Continuous General Population and Housing Census'.

Keywords: register-supported census, multimode data collection, survey management system.

1. Introduction

The 2011 census round saw the introduction of important methodological and technical innovations in the Italian population census. Among the most important ones were: the use of municipality population registers (LAC) as enumeration lists; the questionnaires mail out; a multimode data collection system; the on line crosscheck of census data and population register's records; an enumeration strategy differentiated according to the size of the municipality. Such innovations were designed to reduce problems related to conventional censuses (operational burden on municipal census offices, long delays between data collection and data dissemination, respondents' burden) and represent a first step towards the planning of a register-based census.

[1] The authors of the article would like to thank Francesca Bruno, Ph.D. student at Cornell University, for her significant help revising the text. The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.
[2] Ricercatore (Istat), e-mail: benassi@istat.it.
[3] Tecnologo (Istat), e-mail: mbruno@istat.it.
[4] Tecnologo (Istat), e-mail: magiacum@istat.it.
[5] Tecnologo (Istat), e-mail: silipo@istat.it.
[6] Tecnologo (Istat), e-mail: vaste@istat.it.
[7] Primo Tecnologo (Istat), e-mail: zindato@istat.it.

Moreover, the way to cope with new types of errors had to be found: errors due to the use of registers such as coverage errors or errors due to time misalignments; errors due to the new data collection techniques, which allow greater flexibility but also determine a greater possibility of enumeration duplications.

Namely, an answer to the following question had to be provided: how can we minimize any possible errors? Or, in other words, how can we perform a *quality* census, to be at the same time (i) innovative in methodology as well as able to improve efficiency both in costs and in data dissemination timeliness; (ii) easy to be managed by census operators; and (iii) monitored in all its complex phases?

The Italian Statistical Institute's answer to these issues was a census focused on web technologies. Istat developed a highly integrated web information system that supported all of the different phases of the enumeration process. Such system, becoming itself one of the main innovations of the new strategy, was composed of three web applications: a) a web based management and monitoring system (SGR) accessible to all census operators, supporting every enumeration process activity; b) the online questionnaire (QPOP); c) an online documentation system, including different kinds of contents such as wiki, faq, manuals, legislation, etc.

Aim of this paper is to describe the impact that methodological innovations had from the technology point of view, to present the character and the main features of the technological architecture implemented by Istat and to analyze in details SGR functional logic. This last represented in fact, in the authors views, the joining link between planning expectations and empirical problems inherent to the different phases of the census.

The paper is organized as follows: section 2 provides a description of the main features of the new census strategy and of the main functions of the IT system designed for managing it; section 3 outlines the system architecture, focusing on the technological innovations; section 4 describes in detail the main functionalities of SGR, pointing out how it supported the complex census process, allowing cooperative operators' work and data integration; section 5 presents some concluding remarks.

2. New methods and techniques for the 2011 Italian census

2.1 The 2001 experience and the need of a new strategy

A decennial population census has been taken in Italy since 1861, based on the conventional methodology of complete field enumeration (so called "door-to-door" enumeration). Census forms were delivered and collected by enumerators and self-filled in by respondents. All information was collected and processed on a complete basis (without making use of any sampling technique) while the same economic, human and organisational resources were allocated to every household.

Obvious main goals of the census are the determination of the legal population and the collection of information on its main demographic and socio-economic characteristics. Furthermore, there is a third main goal to be achieved by the Italian census: the update of Municipal Population Registers (*Anagrafi*) on the basis of census results, as provided by the *Law on Population Registers* (*Regolamento Anagrafico*). This entails performing a crosscheck of census data and population register's records (the so-called *confronto*

censimento-anagrafe) concurrently to field-enumeration (Mastroluca and Zindato 2009).

Some figures will help us to give an idea of what a complex and demanding organization was required by the conventional census. In 2001, 22 million private households, accounting for a total of more than 57 million persons, were enumerated by 100.000 enumerators and 10.000 co-ordinators, organised in a network of 8.101 Municipal Census Offices (MCO's) and 103 Provincial Census Offices (PCO's).

The two main actors of this huge operation were: Istat, who has the responsibility for designing and coordinating census activities and processing and disseminating census data; and Municipal Census Offices, who are entrusted the responsibility of fieldwork, of revising completed questionnaires and of crosschecking census data and municipality records (and who, in most cases, contribute to the census with their own financial resources to the state allocated budget).

As to the 2011 census, a number of factors raised questions about the appropriateness of continuing to rely on conventional methodology. Namely, among these were: a) the huge organizational effort imposed on municipalities, exposed to a sudden and time-concentrated increase of workload; b) the need of improving of data dissemination timeliness; c) the increasing difficulty by enumerators of finding people at home, due to changes in both population life-style and structure (e.g. growing percentage of one-person households or of the so-called *dink* - double income no kids - couples), especially in larger municipalities; d) an increasing feeling of both dislike towards the census and public concern for confidentiality.

In order to identify the main critical points of fieldwork organization, a number of studies were conducted (Fortini et. al. 2007), pointing out the following as the main critical issues of 2001 census:

a) need to establish, co-ordinate and upkeep a massive network of enumerators and coordinators, with high difficulties of finding adequately skilled resources, high turn over rates and subsequent training problems;
b) high number of actors, highly differing from one another as to size, capacities, resources;
c) strong delays concerning the collection phase (the delivery of questionnaires was begun on time only by the 28% of municipalities while a mere 2% of them ended the collection according to the schedule).

Furthermore, the organizational impact of census operations turned out to be strongly dependent on the municipality population size: information from the 2001 census monitoring system showed that the biggest municipalities had the largest difficulties in meeting the field operations' deadlines while small municipalities struggled to cope with financial problems (Fortini et. al. 2007).

The results of the Pilot Survey held in 2009, designed in order to test several alternative enumeration strategies (Cassata and Tamburrano 2011), further proved the need of a modular and flexible census strategy, aimed at minimizing the aforementioned criticalities and taking into account the demographical and sociological changes occurred in the Italian society since the 2001 census.

2.2 Main features of 2011 census strategy

Combining the study of census experiences of other countries with a more effective use of administrative data held by municipality population registers, a completely new strategy

was designed (Kotzamanis et al. 2004; Abbatini et al. 2007; Ferruzza et. al. 2007). Such a strategy relies on a number of methodological and technical innovations, and on the crucial role of a census web management system, being the backbone of every phase of the enumeration process (Istat 2009).

Main features of the 2011 census have continued to be the completeness and simultaneity of the population count, but the fieldwork was guided by registers and supported by the use of new data collection techniques and new territorial instruments designed to improve coverage and quality of the enumeration.

The "door-to-door" census became a *register-supported* census, implemented by means of questionnaires' *mail out* to households registered into municipal population registers. Namely, the 8092 Municipality Population Registers were used as lists of households (and addresses) to which census questionnaires were mailed out. Self-completed questionnaires were collected by a multimode system which included Internet, return at any post office in Italy, return to Municipal Collection Centres and, finally, targeted recovery of non-response by enumerators.

Thanks to the use of questionnaires *mail out* (instead of enumerators' delivery) and of a multimode data collection system (where enumerators are just one of the possible return modes, and hierarchically the last one), the front-office staff recorded a dramatic reduction (about 40%) and a great flexibility was allowed to respondents (Picci and Sindoni 2012).

A major issue concerning the new strategy was represented by over-coverage and undercoverage list errors typically affecting a *register-supported* enumeration (i.e. a number of households included in the municipality population register might be no longer residing in the municipality and, conversely, a number of households actually residing on the municipality territory might not be included in the population register). While over-coverage was 'automatically' corrected by a field enumeration that relied on questionnaire *mail out* to units included in the list (by assuming that households residing no more in the municipality would not receive therefore not return the questionnaire), specific measures were required in order to manage potential undercoverage. To this aim, data provided by different sources (such as the revenues agency or foreigners permits to stay) were used to set up a list of persons not included in registers but potentially residing in the municipality. An additional list was based on the pre-census Address Numbers' Survey, containing information on potentially inhabited housing units for which there was no corresponding entry in the municipality records. Enumerators have been sent out to look for and deliver questionnaires at these addresses. Another basic feature of the new strategy was its modular nature. The necessity to differentiate census organization according to the needs and capacities of the different actors required a strategy consisting of a set of modules to be applied flexibly according to the size of the municipality. More precisely, municipalities were divided in two main size categories, and a different combination of modules was planned for each of them (Zindato 2012) (see Table 1).

A major change concerning only largest municipalities (i.e. those with at least 20.000 inhabitants and all province seats) was the shift towards the production of estimations concerning the socio-economic set of census variables. These estimations were produced by using a *long form* on a sample of households. Data produced in this way are significant at a census area (grouping of contiguous and homogeneous enumeration areas) level.

In the same subset of municipalities, a pre-census addresses' survey was carried out to the aim of producing a field-checked geo-coded list of addresses with the related number of housing units, in order to produce auxiliary information to be used to limit undercounting. The additional costs required by the setting up of an address list for the smallest municipalities compared to the corresponding advantages in terms of accuracy and quality of the count restricted the planning of this operation to the largest municipalities (Picci and Sindoni 2012).

Table 1 - Census strategy main modules by municipality type

MODULE	Municipality type	
	20.000 inhabitants or more	<20.000 inhabitants
Pre-census Address Numbers' Survey	X	
Setting up of census areas	X	
Use of pre-census lists derived from Municipality Population Registers	X	X
Use of sampling for collecting socio-economic information (short form/long form strategy)	X	
Crosscheck of census data and population register's records	X	X

Finally, a last but much-important change concerning all municipalities was the actual performing of the crosscheck of census data and population register's records at the same time of the enumeration and through a standardized and public (i.e. visible to Istat) instrument.

2.3 The new strategy adopted and IT management required

The strategy designed for the 2011 census aimed at reducing both municipalities workload and respondents' burden, at holding down costs and at enhancing data dissemination timeliness. The basic idea was to differentiate organization according to the needs and capacities of the different field-work actors and to reduce the burden on respondents by giving them the possibility to choose the return mode that would better satisfy their needs. The three main pillars of such a strategy were as follows:

a) lesser use of front office staff but reinforcement of municipalities back office activities;
b) flexibility of fieldwork organization within Municipal Census Offices;
c) flexibility of collection techniques.

Indeed, the management of a modular and flexible strategy represented a big challenge. On one hand it helped solving problems that traditionally had a great impact on census process, negatively affecting the timeliness of data dissemination. On the other hand, the introduction of this new strategy implied a higher level of complexity and a multiplication of risk factors (Benassi et. al. 2013).

The management of such a complex and diverse enumeration strategy entailed the need of a very flexible web management system. As already mentioned, such a web management system was in fact crucial to the performing and success of the entire census, being a

complete instrument that guided and supported census operators during all the survey phases. It replaced previous monitoring systems, which offered a dashboard with quantitative indicators for tracking questionnaires without providing any management or support tools to census operators (Istat 2012). Basically, it was designed to provide the different users of the system with: (i) up-to-date information at different aggregation levels, including single questionnaire level; (ii) a tool for cooperative working, guided through a forced workflow of questionnaire life-cycle.

The coexistence of different return modes i.e. of information coming from different sources (on-line questionnaire, Post Offices monitoring system, MCOs) required a web IT system constantly updated enabling census staff to follow the status of every questionnaire over time.

The system, accessible online to all of the different levels of census staff, enabled the status of every individual questionnaire to be followed in almost real time, thus allowing the targeted recovery of missing questionnaires: the availability of constantly updated information on the status of each questionnaire enabled enumerators to be directed only to households to which the questionnaire was sent but not yet returned (crosscheck of census and municipality records)**.**

As already mentioned, auxiliary lists enabled the targeted and systematic recovery of individuals not registered in municipal records. Such lists were integrated and loaded into the system, thus allowing enumerators to systematically check undercoverage, as long as they were on the field for the recovery of non-response.

Furthermore, the system was designed to automate back-office work and to guarantee flexibility to fieldwork organization within each Municipal Census Office. Municipal Census Offices managers had to assign an organisational role and a system profile to every user and allocate enumeration areas to enumerators. Each census office could thus freely decide how to distribute work in terms of assignment of enumeration areas to enumerators and back office work to operators. A hierarchical organisation could also be defined by setting dependency relationships between staff with a coordinator role and other staff and of enumerators to co-ordinators.

The system also included an important function to be used for performing the on-line crosscheck between census data and population registers and for the production of the related accounts. By entering in SGR the identification data of enumerated people, as long as questionnaires were returned, and comparing them with municipality records updated to 8 October 2011 (available in the system) Municipality officers performed a real-time crosscheck which enabled the earlier conclusion of the census and the earlier dissemination of the results. In fact, the first and very important final data (i.e. data on the municipalities' legal population by sex, age and citizenship) were timely released before the end of questionnaires data capture, on the basis of data entered in SGR.

Finally, being as well a monitoring system, SGR also allowed to produce census progress reports.

The census web based management and monitoring system was part of a general strategy aiming at minimising errors, reducing organizational workload and holding down costs. The other fundamental components of the overall strategy were the online questionnaire (QPOP) and the online documentation system.

The broad use of QPOP (33,4% of returned questionnaires) resulted not only in a significant reduction of municipalities front-office work (reduction in the number of

enumerators) but also of back-office work in that the quality revision to be performed by MCOs on paper questionnaires need not be done on electronic questionnaires (Picci and Sindoni 2012). In fact, QPOP guided respondents in the correct compilation of their questionnaire through consistency rules and error checking. Moreover, QPOP presented to the respondent only the correct set and sequence of questions to be filled in, so that the online compilation turned out to be easier, faster and less error prone than the paper form (Virgillito and Tininini 2012). The reduction in the number of pages scanned also resulted in some costs' reduction, even though a remarkable reduction could have been achieved only by avoiding sending all paper questionnaires.

Finally, a not negligible role was played by the online documentation system, which functioned as the reference site for the network of SGR operators, who used it to access documentation materials, such as manuals and legislative references, but also as an up-to-date information site for contents such as wiki, faq and communications. It has to be underlined the importance for such a complex and vast organization of an open space for sharing standardized information, contributing to reduce problems due to informational asymmetry.

3. System architecture

One of the main complexities of setting up such an important and business critical system is the integration of multiple data sources and the ability to guarantee the availability of the whole system on a 24/7 basis, in particular during peak hours.

These goals can be achieved only by designing and implementing a flexible architecture, which is based on balanced and replicated systems and uses a consolidated middleware infrastructure where components can be plugged in easily and without massive code reworks.

As shown in Figure 1, SGR was a core part of such architecture, integrating its modules with the questionnaire application that was used by respondents for data entry but also from MCOs who chose to enter data from the paper questionnaires.

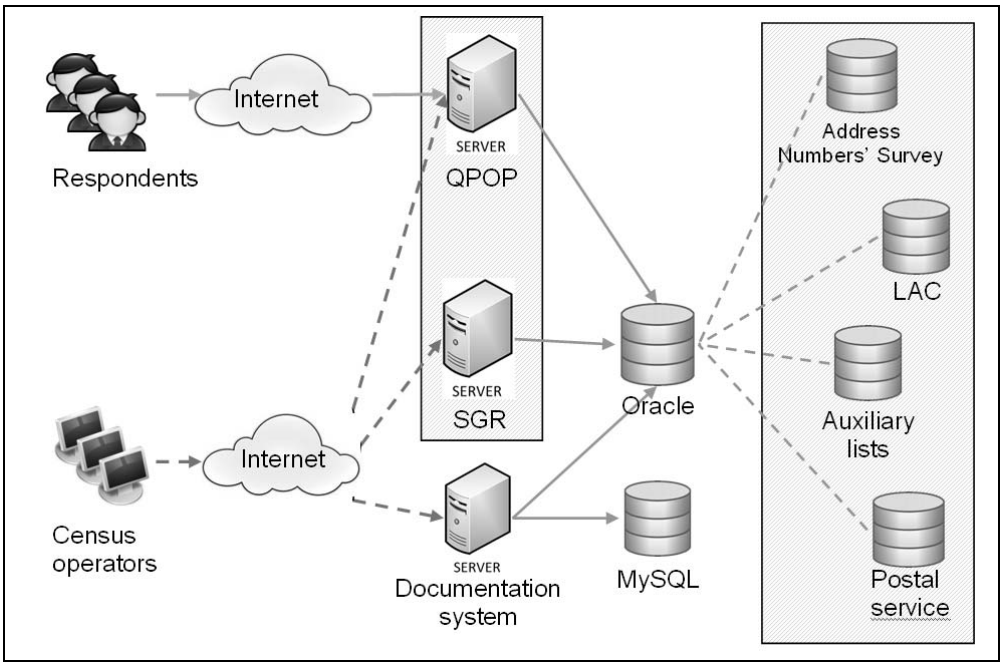
Data on which SGR was based relied on a very complex ORACLE database which had to be accurately tuned for data access optimization from a very high amount of users, including respondents and operators.

3.1 Software technologies

The core software infrastructure of SGR and QPOP followed the Model-View-Controller design pattern. The main technology on which SGR was based is the Java Enterprise Edition (JEE) platform. Java was the core language used for the development of the last Agriculture Census management system. The team that was in charge of the development of SGR for the Population census took advantage of the previous experience, improving the software in terms of security and general quality, reusing many parts of the infrastructure but switching to the state-of-the-art versions of the frameworks chosen as building blocks. The heavy use of frameworks proved to provide a significantly positive impact, producing a cleaner code that was easy to write, test and maintain, resulting in a more robust application. The implementation exploited in particular three widespread open source frameworks: Struts2, Spring and Hibernate.

Hibernate proved to be fundamental for developers, as it acts as a software layer which eases working with database tables through a simple model based on Java objects, called "beans". This speeds up development time and delegates the SQL interaction with the DBMS to the framework: programmers only have to deal with the generally more familiar Java syntax.

Figure 1 – System architecture



Another technology widely used in SGR was AJAX: this Javascript-based technology acts as glue between the GUI and the server-side components of the system, allowing a tight interaction among the controls available on the end-user interface (buttons, lists, collapsible data sections) and consequent real-time actions happening on the server. This technology brought real improvements to the end-user, who was actually guided in his work since the data SGR presented to him were always fetched from the database using small and focused queries. This also brought important performance benefits: for example, when the user selected a province from a dropdown list and was called to select a municipality to work with, the municipalities he could choose from were a subset which depended on the previous choice of the province.

Another component of the architecture was the online documentation portal developed by means of a widely popular open source CMS (Typo3), entirely written in PHP. This led to an additional integration problem due to the fact that Typo3's database management system is MySQL while SGR data was Oracle-based: since any SGR user had to be authenticated on both systems, user data had to be constantly synchronized, through a custom script.

The components of the system architecture are shown in Figure 1. QPOP, SGR and the Documentation system were three integrated web applications. The integration mainly concerned the authentication system and the management of questionnaires returns. With respect to the authentication system, on the one hand, a single-sign-on mechanism was provided for SGR and QPOP, allowing an SGR operator to access QPOP for the data entry of his assigned questionnaires. On the other hand, SGR and the Documentation system authentication tables were synchronized allowing users to access both applications with the same credentials. Concerning the questionnaire returns, QPOP and SGR shared information on the questionnaire status: when a citizen completed the online compilation of his questionnaire this information was available in SGR; when a questionnaire was returned and registered in SGR the online compilation was disabled.

The database was loaded with normalized data from different sources, e.g. Address Numbers' Survey database, LAC, auxiliary lists, Postal service. A detailed description of the integration of such sources is provided in Section 4.

3.2 Security issues

Security is the major concern in a system where the number of users involved reaches a potential of many millions. When the lights of the media focus on such an event like the population census, malicious and talented programmers in search of popularity often undertake a new challenge with themselves. Thus, the organization of the defence has to be really accurate and must take into account most of the state-of-the-art security techniques (OWASP 2013, SAFECode 2013).

The first mechanisms that must be put in place are the most widely consolidated security best practices: authentication based system, strong encryption algorithms, role-based authorization levels and strong user-profiled URL protection.

Since SGR and QPOP relied on a cluster of Tomcat Application Servers, it was straightforward to exploit the capabilities of this middleware infrastructure: in order to implement Authentication using Single Sign On, Tomcat realms and the *j_security_check* mechanisms were used, storing passwords as SHA-256 hashes on the systems. As a consequence a breach of security would not have disclosed users' credentials but only their encrypted representation.

Authorization was implemented using the core Java Enterprise Edition patterns i.e. a complex deployment descriptor with a fine-grained set of operator profiles that came out from a territorial and functional perspective. Any SGR or QPOP user, once logged in the system, could navigate only between the set of functionalities of his competence (through the JEE security-constraints and their correspondent auth-constraints) and could browse only data related to his territory or personally assigned to him (questionnaires in particular).

This last defence against the so-called "Privilege Escalation attack" in particular was achieved through a further check on the backend of SGR and QPOP. Suitable business logic was implemented in order to verify the user authorization to access the requested information.

3.3 Testing phase

The system underwent both functional and infrastructural tests (SWEBOK 2004, OWASP 2013). Functional tests were performed by an internal team in order to verify the

consistency against the software requirements. An additional set of tests on the intermediate releases of the system were demanded to an external team of testers chosen among the census staff across the national territory.

The infrastructural test phases involved a partner company that was asked to perform several vulnerability assessments against SGR and QPOP in order to evaluate the stability and the security of the applications. Tests were performed both using the Black Box approach (that is, without knowing any detail of the systems to be violated) and the White Box approach (using valid credentials and then trying to get more information than allowed to your user's profile). The analysis was split in three phases:

- Vulnerability Assessment: the system was attacked from an external user trying to access protected data exploiting application vulnerabilities by means of techniques like Information Gathering, Data Tampering, Buffer Overflow, XSS and SQL injection. In particular the Data Tampering technique was used in order to check the system against the risk of Privilege Escalation (authenticated users that can increase their authorization level using particular tools);
- Code Review: using static analysis tools it was possible to determine if the applications were compliant to coding best-practices, if they were subject to memory leaks or DB connection leaks, if there was unreached or unused code or if there were code snippets that could cause performance decrements;
- Penetration Test: the system infrastructure was tested against Denial of Service attacks and a port scanning was performed.

All tests were conducted on both a specific test environment and the final production infrastructure, to make sure results of the test were consistent and reliable. The final results proved that both SGR and QPOP were not vulnerable to attacks, and the success of the overall infrastructure during the survey confirmed its quality.

4 The web based management and monitoring system

SGR was designed as a collection of functions, each related to a process phase and customized according to the user profile that accesses it. About eighty functions, grouped in menus and sub-menus, were developed to support all the census activities. Therefore SGR resulted in a modular, flexible, and scalable system which allowed an agile development process. Although more than twenty developers were involved in the realization of SGR, the plug-and-play design of SGR allowed a strong cooperation and a rapid development of all the functionalities.

The system was dynamically customized according to the profile of the logged user: both the menus and sub-menus were personalized displaying only the functions the operator was authorized to use.

To illustrate the functionalities offered by SGR we focus on the following macro-areas: a) integration of multiple data sources (registers, Address Numbers' Survey, etc.); b) interaction of different actors of the survey process (enumerators, supervisors, Istat personnel, postal service); c) implementation of the workflow for the questionnaire life-cycle management; and d) up-to-date monitoring of the survey progress.

4.1 Integration of multiple data sources

As we said before, an important innovation of the 2011 census was the use of municipality population registers (LAC): questionnaires were personalized with information concerning the householder and mailed out by the Italian postal service. In particular, the SGR database was loaded with normalized data, almost sixty millions of individuals, coming from LAC dating to 31/12/2010. Further, to take into account population flows in the period between 31/12/2010 and the census date, i.e. 9/10/2011, a second data loading was performed for municipalities with more than twenty thousands respondents. Small municipalities were supported in such operations by SGR through suitable functionalities.

As already mentioned, LAC can be affected by coverage errors. So, data provided by other sources, such as auxiliary lists and Address Numbers' Survey, were integrated in the system to help detect undercoverage. In particular, the Address Numbers' Survey allowed the detection of buildings with no corresponding LAC individuals: almost nine millions of possible undercoverage signals were loaded in the system and checked by the enumerators.

SGR managed also the different questionnaire return modes offered to respondents, i.e. web compilation, return to postal office, to Municipal Collection Centers and to enumerators. On the one hand the integration between QPOP and SGR allowed a real-time monitoring of the online compilations; on the other hand the integration with services provided by the Italian postal service was necessary in order to load information about both deliveries of the personalized questionnaires to the respondents and returns to the postal offices.

4.2 Interaction of different actors of the enumeration process

SGR included several user profiles, each characterized by a different territorial visibility (national, regional, provincial, municipal) and a number of available functions (Table 2).

User profiles with national visibility had monitoring functionalities that enabled them to monitor the survey progress on the whole territory and to take strategic decisions during the survey. User profiles with regional visibility were provided to Istat regional census staff. The regional employees' duty was to organise and coordinate the survey on their territory and to support the municipal operators. The municipal operator was the key profile in SGR, being in fact responsible for all operative phases of the census. Such users could: (i) define the local survey network, i.e. supervisors and related enumerators; (ii) assign enumeration areas with related questionnaires to the enumerators and (iii) monitor the progress in the questionnaires life-cycle. Enumerators and supervisors carried out field work and back-office activities, such as registration of the questionnaire returns to the Municipal Collection Centers, check of data provided by the Address Numbers' Survey in order to detect undercoverage, etc.

Actually, SGR gave the possibility to create autonomously the survey network and more than eighty five thousand operators' accounts were created. This was a significant result since the creation of the survey network was not managed as a centralized task but it was distributed on the whole territory.

Table 2 - User profiles

USER PROFILE	Functionalities	Territorial visibility
Istat	Monitoring functionalities at all territorial level, i.e. national, regional, provincial, municipal and enumeration areas level	National
URC	Monitoring and support functionalities at regional, provincial, municipal, and enumeration areas level	Regional
UPC	Monitoring and support functionalities at provincial, municipal, and enumeration areas level	Provincial
UCC	Monitoring and support functionalities at municipal and enumeration areas level. Functionalities for creating the municipal operators network. Functionalities for assigning, coordinating and supervising both enumerators and supervisors work. Operative functionalities for field work and back-office activities.	Municipal
CoC	Functionalities for assigning, coordinating and supervising enumerators work. Operative functionalities for field work and back-office activities.	Assigned enumeration areas
Ril	Operative functionalities for field work and back-office activities	Assigned enumeration areas

4.3 Implementation of the workflow for the questionnaire life-cycle management

SGR guided the operators to conduct the survey correctly, offering a fixed path through the questionnaire working phases. Each questionnaire working phase was linked to a state. The transition between two different states was realized through SGR functions or through external operations, such as the completion of the online compilation. Each function had both pre-condition states, i.e. states that allowed the function use, and post-condition states, i.e. states assigned by the function to the questionnaire. In this way SGR defined a flow of questionnaire states, which guided and forced the questionnaire life-cycle. Such questionnaire life-cycle management allowed also cooperative operators work. For example we can consider the following scenario: 1) a back-office operator registers the paper questionnaire return; 2) SGR updates the questionnaire state; 3) consequently the enumerator, responsible for that questionnaire management, is informed of the return and can proceed with following working phases.

Thus SGR was a distributed workflow system, in which, on the one hand, each operator worked autonomously and, on the other, a centralized monitoring of the overall census activities was provided. As a result, using SGR as a survey tool allowed for cost-effectiveness, real-time management, support for cooperative work and on-going monitoring.

The most important function in SGR was the “diary”: a control panel that showed to the survey operator an up-to-date list of his assigned questionnaires, built by the different sources described in section 4.1. Each element of the list displayed the respondent’s name and address, the questionnaire state, and the operations already performed on the questionnaire. The operator could modify the questionnaire state: on the basis of the current questionnaire state the diary showed the possible next states. Further the diary displayed the

information to be edited according to the state transition. In such a way the diary allowed to manage the significant aspects of the process, such as the coexistence of different questionnaire returns offered to respondents. Each return put the questionnaire in a defined state, through internal or external functions, i.e. web compilation, loading of postal office information in the system database and census operator registration. Such information, which was visible in the diary, allowed the complete monitoring of the returns. As a result, the enumerators were able to check only the respondents that had not returned the questionnaire. The enumerator's work was thus more efficient and the quality of the process increased.

Through the diary it was also possible to enter information concerning questionnaire summary data. Such information was mandatory in order to reach a questionnaire final state. The availability of summary data in SGR has been a key element for a rapid dissemination of provisional data.

The diary resulted in a complete instrument for all the operators involved in the survey process: the enumerators used it as a control panel that showed them an up-to-date list of assigned questionnaires and guided their field-work; the supervisors used it as an instrument to monitor the work of their assigned enumerators; the municipal operators referred to the diary to monitor the survey progress in enumeration areas.

4.4 Up-to-date monitoring of the enumeration progress

SGR allowed a comprehensive monitoring of the survey process. On the one hand, SGR offered a detailed supervising of all the operations performed on each questionnaire, displaying author and date of each operation. On the other hand, several reports were provided in order to show the progress of each survey phase: delivery of the questionnaires by the postal service; returns of the questionnaires; survey progress on the basis of questionnaire states; activity of the operators (with a high level of detail).

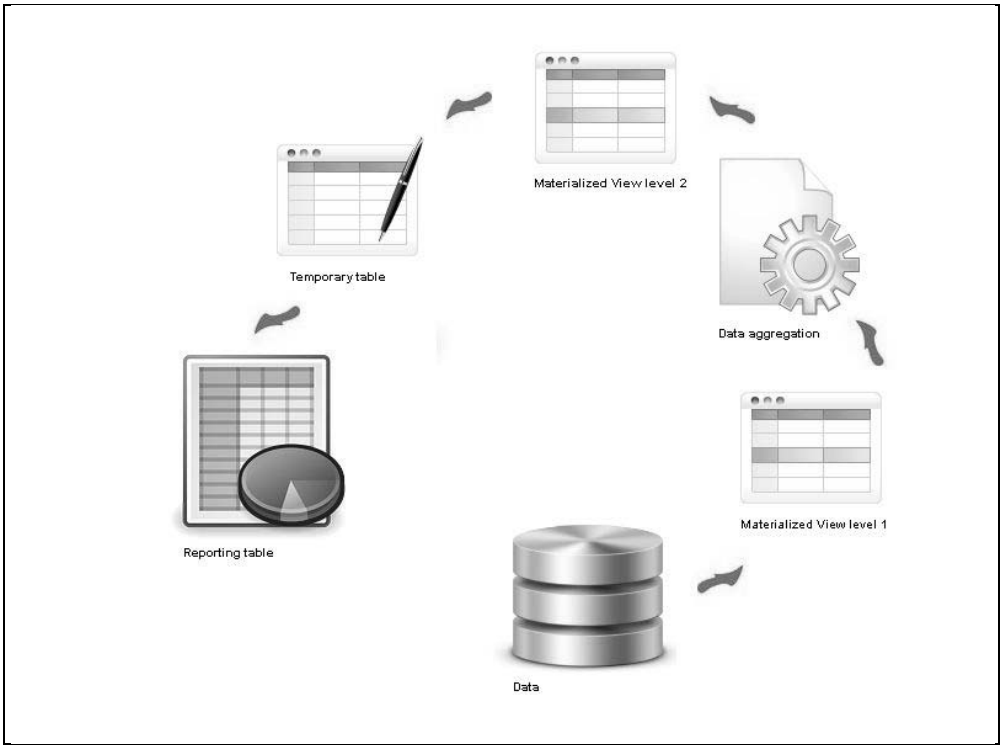
In order to provide a continuous monitoring, without affecting the performance of the application, the reports were updated by the system at regular intervals of one hour. All upgrade procedures were handled within the database, delegating to the application server only the visualization of already executed summaries.

In the database were created materialized views and temporary tables that, with different levels of territorial aggregation, produced the required results (Figure 2). The materialized views, being updated incrementally, allowed saving only those variations that occurred over a period of time, with significant savings in execution times.

Reports were available at different data aggregation level. In particular, it was possible to display data concerning a given operator or data at different territorial levels, i.e. national, regional, provincial, and municipal.

Every report could be exported in the most common formats like XLS and PDF, but also in a CSV plain text file, allowing users to import data in any custom tool.

Figure 2 – Monitoring reports production



5. Concluding remarks

SGR turned out to be a very useful instrument that supported the 2011 Italian population census: it was daily accessed by more than one hundred thousands users during the survey; it supported a complex mixed-mode collection system, with about 8.500.000 online questionnaires (33,4% of the total returns), 7.826.000 questionnaires returned at Municipal Collection Centers, over three million questionnaires returned to enumerators and the remaining 20% returned to post offices.

Even though measurements on the quality of the survey results have not been performed yet, the management of the survey pointed out significant improvements towards a cost-effective and *quality* census. SGR greatly contributed to achieving such results.

First, it supported a register-supported census, allowing a reduction in the related errors. Such results were achieved through the use of multiple data sources as well as a constant and complete monitoring of all survey phases.

Second, SGR offered a predefined and forced workflow in the questionnaire life-cycle management; this enables enumerators' work to be less prone to errors and more cooperative.

Third, SGR was used by operators with different responsibilities, providing suitable

functionalities for each profile: (i) the “diary” effectively supported the enumerators’ field work and (ii) reports allowed monitoring to users with different territorial visibility, also supporting the strategic decisions of Istat census managers.

SGR was originally adopted in the 2010 Agricultural Census, and it was recently used in the 2011 Industry and Services Census. In this last survey, SGR was enriched with new functionalities, such as micro data check on the base of predefined rules. Due to SGR the data production process was agile and efficient.

Since SGR software architecture is mainly framework-based and standard compliant, it revealed itself much suitable for constituting the foundation of a set of additional web systems that have been recently implemented in order to support the subsequent stages of the census process. As few but significant examples worth mentioning are the Post Enumeration Survey (PES) and the System for the Review of the Municipal Population Registers (Sirea).

Istat technological innovation plans include an SGR generalization with the aim to adopt it as a system for the management of many other surveys. In particular, it will be a central component in the design of the ‘Continuous Population Census’, which is due to replace the ‘one shot’ census, with its first cycle beginning on 2016.

References

Abbatini D., L. Cassata, F. Martire, A. Reale, G. Ruocco and D. Zindato. 2007. *La progettazione dei Censimenti Generali 2010-2011. 2 – Analisi comparativa di esperienze censuarie estere e valutazione di applicabilità di metodi e tecniche ai censimenti italiani.* Documenti ISTAT, n. 9. http://www.istat.it/it/files/2011/04/2007_9.pdf (April 2013).

Benassi F., L. Cassata, G. Sindoni and D. Zindato. 2013. "Tales from the 2011 Italian Population Census. The use of a multi-mode data collection system: lessons learnt and future challenges". Relazione presentata alla Conferenza: 5th Conference of the European Survey Research Association (ESRA). Ljubljana 15-19 July.

Cassata, L. and M.T. Tamburrano. 2011. *The 15th Population Census Pilot Survey: how the register driven census changes the enumerators role* in: S. Migani and M. Costa (Eds.) "Statistics in the 150 years from Italian unification", Serie Ricerche n. 10, Università di Bologna, Dipartimento di Scienze Statistiche "Paolo Fortunati", Bologna. http://amsacta.unibo.it/3202/1/Quaderni_2011_10_SIS2011_BookofShortPaper.pdf (April 2013).

Fortini M., G. Gallo, E. Paluzzi, A. Reale and A. Silvestrini. 2007. *La progettazione dei Censimenti Generali 2010-2012. 3 – Criticità di processo e di prodotto nel 14° Censimento generale della popolazione e delle abitazioni: aspetti rilevanti per la progettazione del 15° Censimento*. Documenti ISTAT, n. 10. http://www.istat.it/it/files/2011/04/2007_10.pdf (April 2013).

Ferruzza A., S. Mastroluca and D. Zindato. 2007. "I censimenti esteri: modelli a confronto alla luce dei regolamenti internazionali". Relazione presentata alla Conferenza: Censimenti generali 2010-2011. Criticità e innovazioni. Roma 21-22 November. http://www3.istat.it/istat/eventi/2007/interconferenza/interventi/Ferruzza_Mastroluca_Zindato.pdf (July 2013).

Istat. 2009. *A new strategy for the 2011 Italian Population Census. Product innovations and the compliance with CES Recommendations,* UNECE/CES Group of Experts on Population and Housing Censuses, Twelfth Meeting, Geneva, 28-30 October. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2009/5.e.pdf (April 2013).

Istat. 2012. *A new strategy for the 2011 Lessons learned from use of registers and geocoded databases in population and housing census,* Sixtieth plenary session, Paris, 6-8 June. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2012/22-IP_Italy.pdf (April 2013).

Kotzamanis B., G. Cantisani, A. Dekker, D. Logiadu Didika, M. N. Duquenne and a. Castori. 2004. *Documentation of the 2000 Round of Population and Housing Census in the EU, EFTA and Candidates Countries*, Luxemburg: Office for Official Publications of the European Communities. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-CC-04-003/EN/KS-CC-04-003-EN.PDF (April 2013).

Mastroluca, S. and D. Zindato. 2009. *Censimento della popolazione e delle Abitazioni*, in: Egidi, V. and A. Ferruzza (eds.) Navigando tra le fonti demografiche e sociali, Istituto Nazionale di Statistica, Roma. http://www3.istat.it/dati/catalogo/20100325_01/Navigando_tra_le_fonti_demografiche_sociali.pdf (July, 2013).

OWASP, Open Web Application Security Project https://www.owasp.org/index.php/Main_Page (July 2013)

Picci, M. and Sindoni G. 2012. "Nuovi metodi e tecniche per il censimento della popolazione. I primi numeri del Censimento 2011". Relazione presentata alla Conferenza: Interrogare le fonti 2: un confronto interdisciplinare sull'uso delle fonti statistiche. Napoli, 14-15 Giugno 2012.

SAFECode, *Fundamental Practices for Secure Software Development* http://www.safecode.org/publications/SAFECode_Dev_Practices0211.pdf (July 2013)

SWEBOK, *Guide to the Software Engineering Body of Knowledge 2004 version*, chapter 5. http://www.computer.org/portal/web/swebok/htmlformat (July 2013)

Virgillito, A. and L. Tininini. 2012. *The Web-based Data Collection in the Italian Population and Housing Census Meeting on the Management of Statistical Information Systems*. MSIS 2012. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2012/18_Italy.pdf (April 2013).

Zindato D. 2012. "Nuove prospettive per il Censimento della popolazione. Verso il censimento continuo". Relazione presentata alla Conferenza: Interrogare le fonti 2: un confronto interdisciplinare sull'uso delle fonti statistiche. Napoli, 14-15 Giugno 2012.

Methodology for the production assessment of tourism industries[1]

Sandra Maresca,[2] Massimo Anzalone,[3] Ilaria Piscitelli, [4]

Abstract

This paper deals with the illustration of the methodology developed within Istat National Accounts Directorate in order to assess the production account of tourism industries and other industries, and it is part of a wider process of implementation for the first Italian Tourism Satellite Account (ITSA). The approach adopted for the compilation of the Italian TSA Table 5 (IT5), this latter being one of a complete set of 10 TSA tables, processes data starting from the format provided by the Italian Supply and Use tables (ISUT). Such procedure has been gradually developed by means of a set of worksheets hierarchically organized in a modular and additive structure finalized to the automatic compilation of IT5.

Keywords: National Accounts (NA); Supply and Use Tables (SUT); Tourism Satellite Account (TSA); Italian Table 5 (IT5); class of economic activity; tourism industry; worksheets.

Executive summary: the first Italian TSA (ITSA) is the result of an inter-institutional collaboration attended by the most important Italian institutions involved in tourism field and research[5], conducted under the technical coordination of the Department of Italian National Accounts.

This work refers exclusively to the production accounts of Italian tourism industries and other industries, (Table 5 of a TSA), and is articulated as follows:

- in paragraph 1: a brief reminder about the principal purposes served by a TSA;
- in paragraph 2: key concepts of TSA;
- in paragraph and 3 and 4: notes on the structure of the ITSA and the IT5;
- in paragraph 5: and its subsections: sources used and methodology developed to assess production account of Italian tourism industries;
- in paragraph 6: and its subsections: special issues; intermediate consumption; value added;

---

[1] This paper is partially based on the paper "Versus the first Italian Tourism Satellite Accounts: the production approach", presented at the 20th International Input-Output Conference, 29th June 2012, Bratislava, whose authors are: Sandra Maresca (Istat); Ilaria Piscitelli (Istat); Massimo Anzalone (Ministry of Economy and Finance). The article engages exclusively the authors, the views expressed do not imply any liability on the part of Istat.

[2] Senior Researcher, Italian National Institute of Statistics - Istat, e-mail: maresca@istat.it.

[3] Official, Italian Ministry of Economy and Finance, massimo.anzalone@tesoro.it.

[4] Researcher, Italian National Institute of Statistics - Istat, e-mail: piscitelli@istat.it.

[5] Italian National Institute of Statistics; Italian Central Bank; University of Messina; International Center of Studies on the Tourism Economy; National Observatory of Tourism.

- in paragraph 7: description of the integrated set of worksheets for compilation of IT5.

Introduction

The central framework of National Accounts (NA) provides a suitable structure for an adequate depiction of the several aspects which characterize economic systems, but in some cases working within the central framework is not sufficient in order to make apparent and to describe in more depth aspects that are hidden or surface only in a limited number of points. It is necessary to develop satellite accounts in which to accommodate widened concepts and detailed accounts that would have overburdened the central ones.

There are two general types of satellite accounts: those that basically expand and or organize the detail in the core accounts and those that go outside of the core altogether by expanding their conceptual boundary[6] .

The first type is particularly suitable for analyzing specific economic sectors as tourism. In it the basic intention is not to use alternative economic concepts, but simply to focus on a certain field or aspect of economic and social life in the context of National Accounts. In this context Tourism Satellite Account (TSA) is an extension of the NA system. As such, it highlights the economic transactions that are recorded in NA, but which are related specifically to tourism. In particular, the TSA identifies and emphasizes the transactions between visitors in an economy and the industries that serve them[7].

The TSA shares the basic concepts and definitions concerning the different aspects of tourism provided by the 2008 *International Recommendations for Tourism Statistics* (*IRTS2008*): different forms of tourism; the different main purposes of a tourism trip; the concept of tourism expenditure and its different categories related to the different forms of tourism; the different classifications that can be used in the analysis of tourism.

The structure of the TSA has been conceived aligned and integrated with the international macroeconomic frameworks provided by SNA93 and ESA95, and in particular with the structure of the "Supply and Use" tables (SUT) of National Accounts.

The result of such tight integration is the international manual for compilation of a TSA, *2008 Tourism Satellite Account (TSA): Recommended Methodological Framework* – TSA:RMF (*TSA:RMF2008)*, in which the setting up of a tourism satellite account is guided step by step.

The work presented in this paper is part of the implementation process for the first Italian TSA (ITSA) and is specifically referred to the compilation of the Italian Table 5 (IT5), regarding the production account of tourism industries and other industries. In this context the focus of the work is thus the measurement of tourism sector using the supply side approach, namely identify the economic activities which serve the tourism demand and estimate their production.

The content of this paper will be organized in the explanation of the methodology developed to estimate the production account of Italian tourism sector, and in the illustration of the set of integrated worksheets prepared for a smooth and automatic compilation of IT5.

---

[6] More details in System of National Accounts 1993 (SNA93); chapter XXI, par.21.45- and 21.46. pg. 614.

[7] The concept of visitors and industries will be explained in the following.

The acronyms used in this paper are listed below:

BOP:	Balance of Payments;
CPC	Central Product Classification
ESA	European System of Accounts
IC and P2:	Intermediate Consumptions;
ISIC:	International Standard Industrial Classification of All Economic Activities;
GDP:	Gross Domestic Product
GVA	Gross Value Added
NA:	National Accounts;
NACE	Nomenclature d'activité de la Communauté Européenne
P1:	Production;
PT:	Package tours;
SBS:	Structural Business Surveys;
SNA	System of National Accounts
SUT:	Supply and Use Tables;
T5:	Table 5 of TSA: production accounts of tourism industries and other industries;
TA:	Travel Agencies;
TO:	Tour Operators;
TSA:	Tourism Satellite Account;
TVA:	Tourism Value Added;
VATI:	Value Added of Tourism Industries;

1. Role of the Tourism Satellite Account

The implementation of a TSA implies a "*detailed analysis of all aspects of the demand for good and services which might be associated with tourism; the establishment of an actual interface with the supply of such goods and services within the economy of reference; the highlight of how this supply interacts with other economic activities, using the SUT as reference*"[8].

A TSA serves a number of purposes. First and foremost, it provides a coherent framework within which to integrate, reconcile, organize and analyze the variety of economic statistics relevant to tourism, both on the supply side (i.e., production and costs of tourism industries; level of use of labour; investments in productive capital; role of Public Administration) and on the demand side (i.e., different types of tourisms; type of goods and services on demand). This is the more important the more tourism is not an explicitly identified industry within the statistical system as it cross-cuts several industries[9]. In a TSA

[8] *TSA:RMF2008*, pg.2.

[9] As already touched on when T7 and T10 has been introduced, the concept of tourism supply embraced by a TSA is other, and broader, than that typically understood in tourism statistics. In TSA the dimension of supply (the productive activities which provide products bought by tourists) takes shape after and according to the statement of the demand structure, which is made up by means of tourists expenditure's behavior.

tourism's various components are pulled together and articulated into analytical tables; as such, it explicitly defines the tourism industry within the statistical system.

A complete TSA is composed by 10 tables, but the full compilation of the first 6 is considered enough to obtain a measure of the magnitude of tourism sector within the whole national economy. Each table is useful as individual evaluation element, as they collect detailed information on a specific aspect of tourism.

The whole concept of tourism internal demand is surrounded in Tables 1-4 and tables 8-9. The first 3 tables estimate tourism consumption, respectively inbound (T1), domestic (T2) and outbound (T3). Internal tourism consumption (T4) is the sum of T1 and T2 and represents the quantitative basis for comparison with the supply side. Tables 8 and 9 contribute to define a broader concept of total tourism internal demand, and refer to tourism gross fixed capital formation of tourism industries and other industries (T8) and tourism collective consumptions (T9).

Tables 5 (T5) and table 7 (T7) depict the TSA's concept of tourism supply. T5 describes the production account of tourism industries and other industries and constitutes the preliminar step for the calculation of tourism value added (TVA), one of the most important tourism indicator produced in the context of a TSA. T7 shows, for each tourism industry, the number of establishments and the level of employment, this latter in different measures: jobs, hours works, full-time equivalent jobs.

Nevertheless, the TSA would be an underutilized tool without an analysis of the interrelation between tourism supply and demand. In order to permit such dialogue[10] a common room is necessary: Table 6, which provides the suitable conceptual and methodological environment by reproducing the structure of the "Supply and Use" tables of NA.

Table 10 provides a set of non monetary indicators (number of trips, types of accommodation, modes of transport, number of establishments[11]) that are relevant to specify the characteristics of the economic variables. In this context A TSA offers a link bridge between economic data and non-monetary information on tourism.

Being rooted in the System of National Economic Accounts a TSA provides an economic measure of the importance of tourism in terms of expenditures, gross domestic product, value added and employment which are comparable with similar measures referred to the overall economy of reference, to one or more productive sectors, to different regions, countries or international areas of interest.

In general terms, a TSA serves to define what are considered to be the tourism products

[10] Both in the restricted, but "core", concepts of internal tourism consumption (T4) and of production of tourism industries and other industries (T5).

[11] Both T7 and T10 require the number of establishments for tourism industries. For Italian T7 data source has been the Statistical Register of Active Enterprises (ASIA), which provided the number of establishments. Instead, T10 has gathered data from the monthly census survey *"Occupancy of tourist accommodation establishments "*, conducted by Istat. For this latter survey is in force the Regulation on tourism statistics (EU) 692/2011, based on which the statistical units for collecting data from the supply side is the Local-Kind-of-Activity-Units (LKAU). The difference is substantial. An establishment is classified according to its main economic activity. Its eventual secondary activities surface only by means of LKAU. In the case of only one economic activity, the concept (and the number) of establishment and LKAU correspond; otherwise the same establishment can be composed by two or several LKAU according to the degree of diversification of its production. As a result, the number of establishments in IT7 and IT10 is quite different. A recent work presented by Italy at 12th Global Forum of Prague deals just about this issue. For details http://www.tsf2014prague.cz/assets/downloads/Paper%202.1_Francesca%20Petrei,%20Maria%20Teresa%20Santoro_IT.pdf.

and the tourism industries, and consequently helps to shape the development of tourism statistics.

Last, but not least, it must be stressed that a TSA "*as such only makes possible to measure the direct effects of consumptions on output and value added of tourism industries and other industries serving them*[12]". This means that tourism's impact on the economy is not fully reflected in the TSA tables as any measurement of the indirect and induced effects are taken into consideration by a TSA[13].

2. Key concepts of the TSA

The key elements crucial to define tourism sector are in some cases shared with other related frameworks, such as SNA and the Balance of Payments (BOP), in other are specific to tourism statistics.

For both SNA and BOP the concept of residency has been used to define a traveller, in the international context, as someone travelling outside his or her country of residence. This concept is also used in tourism statistics and TSA in order to distinguish the different forms of tourism (international and domestic).

However since tourism statistics are concerned with domestic as well as international tourism flows, the concept of "usual environment" is specific to tourism scope, where it is used as a defining condition, additional to that of residence, whether someone is a visitor to a place or not.

The usual environment is defined "*as the geographical area (though not necessarily a contiguous one) within which an individual conducts his/her regular life routines*"[14]. This concept plays a major role in tourism statistics and relates to the place where the individual lives and works or studies and includes any other places frequented. This notion is not precisely defined in the international standard, thereby allowing a country to apply the tourism concept on its own specifications[15]

In the TSA people who are engaged in tourism are called visitors. By definition, not all travel is tourism, and not all travellers are visitors.

According to *IRTS2008*[16] *"a visitor is a traveller taking a trip to a main destination outside his/her usual environment, for less than a year, for any main purpose (business, leisure or other personal purpose) other than to be employed by a resident entity in the country or place visited. These trips taken by visitors qualify as tourism trips."*

Visitors are a subset of a wider category of travellers, these latter being "*who moves between different geographic locations, for any purpose and many duration*[17]". Visitors are

---

[12] *TSA:RMF2008, Annex 6 pg.95*.

[13] Indirect effects regard the additional inputs required by tourism industries to other productive sectors in order to serve visitors. In addition, the increase of income distributed to the labor force and to the owners of productive capital resulting from incremental visitor demand also generates increased demand for goods and services through a rise in household consumption (induced effects). *TSA:RMF2008, Annex 6 pg.95*.

[14] *IRTS2008*, par. 2.21.

[15] *Tourism Satellite Accounts in the European Union, Volume 1, Report on the implementation of TSA in 27 EU. Member States* – Luxembourg, 2009

[16] *IRTS2008*, pgg 9-10.

[17] Ibidem.

further split in tourists and same-day visitors. The former are those who stay one or more nights away from home, while the latter are those who spend no nights away from home.

What makes tourism sector special "*is the temporary situation in which an individual in the capacity of consumer finds himself*"[18]. Being a visitor is a transient situation, related to a specific trip. Once the trip is over, the individual loses his/her condition of being a visitor and goes back to being a mere consumer.

This feature affects all aspects concerning tourism supply's measurements in the context of TSA.

Indeed, in a TSA perspective tourism is "*a demand side phenomenon and refers to the activities of visitors, and their role in the acquisition of goods and services*[19]". The notion of activity encompasses all that visitors do for a trip or while a trip, as long as he/she stays in the condition of visitors. As a result the definition of tourism expenditure and tourism consumption does not depend on the mere purchase of specific tourism good and services, but on the fact that the purchase has been made by a visitor, and not by a consumer.

In this context the supply side is consequently settled up through the identification of goods and services (products) bought by visitors and of the productive activities that provide those products, hence, not only accommodation services, as typically occurs in tourism statistics from the supply side, but several others.

The dependence of tourism supply structure on the behavior of the demand witnesses its main feature: being a crosscut sector. This characteristic distinguishes the TSA both at a conceptual and methodological level, and has impact on the implementation of T5.

3. The content of the production account of tourism industries (T5)

The supply side in tourism statistics is mainly represented by the accommodation industry. In this restricted depiction, tourism supply is entirely surrounded by the correspondent economic activities. According to the Italian classification of economic activities (ATECO2007), accommodation activities are contained in Division 55[20]. However, for the purpose of a TSA, the concept of supply is much broader than that adopted in tourism statistics. Indeed, it includes all productive activities that provide the goods and services that visitors acquire.

Such supply takes place in TSA's table 5[21].

"*TSA Table 5 (T5) compiles the production accounts of tourism industries and other industries in accordance with the TSA-RMF classifications of industries and products*"[22].

The structure of T5 is suitable to make explicit the relationship between productive

[18] *TSA:RMF2008*; chapter. 1, pg.2.

[19] *TSA:RMF2008, pg.1.*

[20] *ATECO2007* is the national application of the European classification of economic activities (NACE), in its turn directly derived from international one (ISIC). In it accommodation services included in tourism statistics are individuated at Group level, in particular 55.1, 55.2 and 55.3.

[21] As already mentioned, within the complete set of 10 tables of a TSA, T5 represents a crucial, but not the only one, element of the analysis of tourism from the supply side. Besides the T5, useful to estimate the production accounts of tourism industries and other industries, the other table supply-oriented is T7 concerning *employment in the tourism.* ITSA includes Table 7, not yet completed at the time this document was written.

[22] *TSA in the European Union*, Vol.3, par.3.5, pg.41

activities and the supply of products, as it is mostly based on the format of matrix of production provided by SUT. The main purpose of T5 is to prepare and compile data on gross value added (GVA) for tourism industries and other industries by transforming the national production account into a TSA production account.

T5 is divided into three main sections and copies the format of national accounting matrixes, that is to say it refers both to economic activities (industries) and products (good and services).

The middle pane of T5 is the national matrix of domestic production derived from SUT. In this section production refers to all productive activities of the country, grouped according to the criteria of the TSA, and articulated according the international standard of tourism products[23]. In T5 the total output corresponds to the total national production derived from NA, except for the differences attributable to the net assets of tour operators, which will be discussed in detail in section 6.1.

Below the matrix production, places the vector of intermediate consumptions, by industry but not by product. Even for intermediate consumption are valid the aforesaid equivalence and the exception with national economy.

The difference between production and intermediate consumptions yields the value added, that in T5 is equivalent to that of national economy at total level.

In accordance with the format established in SNA93, the production is required at basic prices; intermediate consumption at purchaser's prices; gross value added at basic prices.

In T5, as in NA matrix of production, the overall national economy is represented: what does make the difference is the perspective pursued. T5 constitutes a tool for a tourism analysis, in it industries and products are highlighted and embedded so as to bring out the tourism sector.

This difference affects not only the criterion for grouping and displaying industries and products in the frame of T5, but implies the need to develop a specific procedure to evaluate production (P1), intermediate consumptions (P2) and value added (VA) for each industry of T5[24].

For the analysis of production and production processes, the establishment is the most suitable unit from which to gather data for SNA as well as the TSA[25]. In NA industries are classified according to the international classification of economic activities (ISIC[26]). A industry is defined as a group of establishments that engages in the same of a similar kind of economic activity.

However, tourism in not an industry in this sense. Rather, tourism cuts across industries identified in ISIC because it is dependent on the consumer's purchases as a visitor. Moreover, because visitors purchase goods and services from many different industries, the TSA must identify and separate out the tourism output of each of them, regardless that the economic activities isolated at ISIC level is tourism related or not.

---

[23] See in the next section.

[24] See in the next.

[25] According to the Regulation on tourism statistics, the statistical unit for collecting data on the supply side is the local kind-of-activity unit. In particular, *a LKAU is the part of a KA5 which corresponds to a local unit*. See "*Methodological manual for tourism statistics, v. 1.2*", par. 2.1.2, pg. 56, Eurostat, 2012, - Luxembourg.

[26] In European Countries the classification adopted is the NACE, compatible with international one. For details see in the follow.

From a methodological point of view the compilation of T5 requires that two different approaches will be met. Firstly, in a supply perspective, production must be evaluated with reference to a selected tourism economic activities, identified at the recommended ISIC level of detail, regardless that their main characteristic output (tourist by definition) will be totally absorbed by visitors, and regardless that they may have non-tourism secondary activities which, however, will be included in their total production. This approach aims to isolate the only tourism economic activities in order to compose each tourism industry and, consequently, tourism sector as a whole.

Instead, in a demand perspective, what is necessary to consider is the act of acquisition of a product by visitors. This act marks as tourist a generic product, and it can be produced by tourism industries as well as by those non-tourism economic activities eliminated in the above industry-oriented perspective.

In order to take into account both needs, these two point of view were merged in the compilation process of IT5, and will be explained in detail in the next dedicated sections.

Within T5 the value added generated by the only tourism industries (VATI) is the most important indicator produced. For each tourism industry VATI refers to its total production, which may includes tourism or not-tourism products.

It should be stressed that none consideration on its actual tourist destination is yet included in the VATI. It is still a gross value added of tourism industries, and only a part of it will be allocated to satisfy the demand of visitors. Furthermore, being limited to the tourism characteristic industries, VATI doesn't emphasize the productive role performed by the rest of economic activities, whose total output may certainly comprise products which will directly serve tourists.

In contrast with the central framework of NA, where the supply of a commodity always equals its demand, in a TSA the supply of a tourism commodity usually exceeds tourism demand. This is because tourism supply includes the total production of a tourism commodity whether it is purchased by a visitor or not.

4. Framework of the ITSA from the supply side (IT5)

In a supply perspective, one of the most important issues is the description and measurement of the role of tourism in the supply of goods. Inasmuch tourism supply is understood as the direct provision to visitors of the goods and services that make up tourism expenditure, the first object is to define tourism products, and then the productive activities that provide them.

Although visitors can make expenses in any product category, in order to facilitate international comparisons and to construct the accounting tourism systems, the international methodologies emphasize those product categories that are more closely related to tourism, i.e. accommodation, transport, travel agency services, and so on.

In accordance to the criteria established in the SNA93[27], international manuals[28] identify two kind of tourism specific products:

- *characteristics products:* those whose tourism expenditure on the product should represent a significant share of total tourism expenditure/supply of the product. That is to say; without tourism they would cease to exist at a significantly level. Characteristics products are made of two subcategories, according to their level of international comparison:
  - *internationally comparable products*
  - *Country-specific tourism characteristic products*
- *connected products:* those whose tourism expenses are significance and recognized, although their link to tourism is limited worldwide. Consequently, a list of such products will be Country-specific.

Besides tourism specific products, international standards foresee *non tourism-related consumption products*, namely, all other consumption of goods and services that do no belong to the previous categories[29].

Indeed, bearing in mind the role of visitor's behavior in the categorization of tourism products, and taking into account that what makes tourism product is its acquisition by a visitors, all kind of them are potentially tourism products, and they actually become as a such when purchased by visitors.

A first classification of products, according to their categorization as internationally comparable tourism characteristic, led to a long list provided by *IRTS:2008*[30]. Identified in CPC[31] they have been grouped at 4 digits level in a further, recommended list of tourism characteristic products[32], grouping by main categories, as shown in Table 1.

This second list defines the core of tourism characteristic demand and represents the preparatory step for the subsequent identification of economic activities involved in the production of such products (namely, to outline the boundary of the supply).

As occurred with the list of characteristic products, a list of tourism characteristic activities is defined[33], that is to say, those that typically produce tourism characteristic products. According to the correspondence between the international classifications of products (CPC) and productive activities (ISIC), and starting from the aforementioned list of products, *IRTS:2008* has listed tourism characteristic industries, that too appear in Table 1, grouping in main categories.

[27] Chapter XXI, pg.615.
[28] *IRTS:2008*, pg. 47 and *TSA:RMF2008*, pg. 28.
[29] *Ibidem.*
[30] *IRTS:2008, Annex 2*, pgg.115-118.
[31] Central Product Classification.
[32] *IRTS:2008, Annex 4*, pgg.128-139.
[33] *IRTS:2008, Annex 3*, pgg.119-127.

Table 1 – List of categories of tourism characteristic consumption products and tourism characteristic activities (tourism industries).

PRODUCTS	ACTIVITIES/INDUSTRIES
1. Accommodation services for visitors	**1.** Accommodation for visitors
2. Food and beverage serving services	**2.** Food and beverage serving activities
3. Railway passenger transport services	**3.** Railway passenger transport
4. Road passenger transport services	**4.** Road passenger transport
5. Water passenger transport services	**5.** Water passenger transport
6. Air passenger transport services	**6.** Air passenger transport
7. Transport equipment rental services	**7.** Transport equipment rental
8. Travel agencies and other reservation services	**8.** Travel agencies and other reservation services activities
9. Cultural services	**9.** Cultural activities
10. Sports and recreational services	**10.** Sports and recreational activities
11. Country-specific tourism characteristic goods	**11.** Retail trade of country-specific tourism characteristic goods
12. Country-specific tourism characteristic services	**12.** Country-specific tourism characteristic activities

This ensemble of activities limits the TSA from the perspective of the supply and defines the tourism industry as a whole.

In principle, the mere observation of the aforementioned list and criteria for specialization can define the characteristic activities. Nevertheless, it is necessary to point out a few aspects, already touched on in the previous paragraph.

In supply side statistics, establishment, (the statistical unit used to compose the branch of economic activity in SNA as well as in TSA) are classified according to their main activity, which in turn is determined by the activity that generates the most value added.

As a consequence, the grouping of all establishments with the same main activity which is one of the tourism characteristic activities constitutes a tourism industry.

Nevertheless, an establishment can develop a main activity and perform one or several secondary activities. Therefore, on the one hand, characteristic activities that specialize on tourist products develop activities that are not exclusively or mainly tourist-based.

However, on the other hand, characteristic establishments may have as their main clients consumers other than visitor[34]. Nevertheless, these activities will be included when referring to the tourism industry, in other word, when considering the supply in the TSA's perspective.

Furthermore, it must be careful when transpose international standards within the national economic frame. Indeed, for Italian Food and beverage tourism industry there is no perfect overlapping between the recommended industry and the Italian one in terms of class of economic activities included – See par. 5.

[34] A typical example are restaurant and other food and beverage industries.

All these aspects have affected and sometimes made difficult the implementation of the methodology for compilation of IT5.

With reference to Table 1, both for products and industries, categories 1 to 10 comprise the core for international comparison. Instead, the two latter are Country-specific, where category 11 covers tourism characteristic goods and the corresponding retail trade activities; category 12 refers to Country-specific tourism characteristic services and Country-specific tourism characteristic activities[35].

ITSA is articulated in 11 tourism characteristic categories, both for products and for industries. The first 10 follows those recommended, so as to permit international comparison. Instead, for Country-specific categories ITSA deviates from standard requirements.

With reference to the products, the following Table 2 shows the correspondence between TSA and ITSA for T5. In IT5 the two latter tourism characteristic categories are grouped in a single item, also inclusive of connected products. Indeed, "*Other consumption and non consumption products*" is a residual category, required in order to complete the national production. For IT5 it includes both consumption and non consumption residual products.

As far as economic activities are concerned, the correspondence in terms of main characteristic categories is almost perfect[36] – see Table 3. Even in this case, the same applies as for the categories 11and 12 of products.

[35] *IRTS:2008*, pgg. 29-30.

[36] The alignment with international standards in terms of main categories of characteristic tourism industries doesn't mean that the same equivalence is reproduced in terms of ISIC at 4 digits level of economic activities , as suggested in Annex 3 of IRTS:2008. For Italy this discrepancy concerns the tourism industry of Food and beverage.

Table 2 - Tourism characteristic products and other products in international and Italian T5

INTERNATIONAL T5	ITALIAN T5
A. Consumption products	-
A.1 Tourism characteristic products	**A.1 Tourism characteristic products**
1 Accommodation service for visitors	**1** Accommodation service for visitors
2 Food and beverage serving services	**2** Food and beverage serving services
3. Railway passenger transport services	**3**. Railway passenger transport services
4. Road passenger transport services	**4**. Road passenger transport services
5 Water passenger transport services	**5** Water passenger transport services
6 Air passenger transport services	**6** Air passenger transport services
7 Transport equipment rental services	**7** Transport equipment rental services
8 Travel agencies and other reservation services	**8** Travel agencies and other reservation services
9 Cultural services	**9** Cultural services
10 Sports and recreational services	**10** Sports and recreational services
11 Country-specific tourism characteristic goods	**11** Country-specific tourism characteristic products and connected products
12 Country-specific tourism characteristic services	-
A.2 Other consumption products	
B. **Non consumption products**	**A.2 Other consumption and non consumption products**
B.1 Valuables	
B.2 Other non consumption products	
TOTAL OUTPUT	**TOTAL OUTPUT**

Table 3 – Tourism characteristic industries and other industries in international and Italian T5

INTERNATIONAL T5	ITALIAN T5
Tourism characteristic industries	**Tourism characteristic industries**
1 Accommodation for visitors	**1** Accommodation for visitors
2 Food and beverage serving industry	**2** Food and beverage serving industry
3 Railway passenger transport	**3** Railway passenger transport
4 Road passenger transport	**4** Road passenger transport
5 Water passenger transport	**5** Water passenger transport
6 Air passenger transport	**6** Air passenger transport
7 Transport equipment rental	**7** Transport equipment rental
8 Travel agencies and other reservation services industry	**8** Travel agencies and other reservation services industry
9 Cultural industry	**9** Cultural industry
10 Sports and recreational industry	**10** Sports and recreational industry
11 Retail trade of country-specific tourism characteristic goods	**11** Retail trade of country-specific tourism characteristic goods and other non specific goods
12 Country specific tourism industries	-
Total	**Total**
Other Industries	**Other Industries**
Output of domestic producers	**Output of domestic producers**

5. Methodology for implementation of Italian T5

The whole process of collecting, preparing and processing the information required (and available) to support compilation of IT5 was developed step by step as follows:

- recognition of international requirements – mainly IRTS:2008 and TSA:RMF2008 – and their concrete application within the Italian accounting context;
- analysis of available sources (ISUT and other INA data; basic statistics deriving from SBS; administrative data);
- assessment of Italian tourism industries' production;
- treatment of special issues;
- assessment of intermediate consumptions and value added;
- automated fulfillment of IT5 by means of a set of integrated worksheets.

The first step has been examined in the previous paragraph with reference to the determination of tourism products and tourism industries, according to the international recommendations. In the following the remaining will be discussed.

5.1 Data sources

ISUT are the primary data source for the implementation of Italian T5. They consist in matrices breakdown by branch of economic activity and by product. Therein, a detailed picture is displayed about: the supply of goods and services, both of internal and of imported origin; the use of goods and services for intermediate and final consumptions; the components constituting the value added generated by the branches. At a national level of analysis they highlight the connection between economic activities and products, by means of the description of internal production's processes and products' related transactions.

As shown in Figure 1, production and intermediate consumptions matrices provide data by branch and by product. Accordingly, an accurate outlining of Italian tourism sector strictly depends on the level of detail of the available Italian SUT.

For the first IT5 both products and activities have been traced starting from the highest level of detail provided by the Italian SUT, namely 106 branches of economic activities and 266 items of products (format not disseminated), in which both principal and secondary activities were distinguished .

First of all, Italian tourism industries have been identified in NA's production matrix after excluding from the scope of analysis those activity groupings which clearly did not include any tourism industry. For each tourism industry so individuated, total production, inclusive of both main and secondary products, has been estimated by merging the two perspective industry/product[37], that is to say by conducting a cross analysis of the production's matrix content.

It is worth remembering that that selection does not lead to an equivalence with tourism consumption, in other words, with the demand estimates. What will guide to a pure tourism production will be its comparison with the level of tourism demand.

Figure 1– Simplified scheme of Italian production matrix

PRODUCTS	Branches of economic activity						Total
	1	2	…	i	…	N	
1	$p_{1.1}$	$p_{1.2}$	…	$p_{1.i}$	…	$p_{1.N}$	$P_{1.}$
2	$p_{2.1}$	$p_{2.2}$	…	$p_{2.i}$	…	$p_{2.N}$	$P_{2.}$
…	…	…	…	…	…	…	…
j	$p_{j.1}$	$p_{j.2}$	…	$p_{j.i}$	…	$p_{j.N}$	$P_{j.}$
…	…	…	…	…	…	…	
M	$P_{M.1}$	$P_{M.2}$	…	$P_{M.i}$	…	$P_{M.N}$	$P_{M.}$
Total	$\mathbf{P_{.1}}$	$\mathbf{P_{.2}}$	**…**	$\mathbf{P_{.i}}$	**…**	$\mathbf{P_{.N}}$	**P..**

[37] See par. 3, pg. 8.

Besides ISUT, a precise cut of the boundary of tourism production and intermediate consumptions (IC) required additional basic data sources. Namely:

- basic information on market component's production[38] and IC, available by *class* of economic activity (4 digits level), deriving from SBS;
- NA's employment data, available up to 5 digits level, corresponding to the *category* of Italian classification of economic activities;
- information base on *non market* component's production[39] (General Government), available up to 5 digits level, corresponding to the *category* of Italian classification;
- data concerning household final consumptions;
- administrative source.

5.2 Selection of tourism industries and products

The approach followed for the selection of tourism industries and products for IT5 is a *bottom-up*: they has been traced starting from the highest level of detail provided by Italian SUT structure and by SBS statistics.

As mentioned, Italian SUT has provided a good level of detail both for products (branches of homogeneous production) and for branches of economic activities. Such articulation allowed to extract with relative ease those branches and product relevant for tourism.

However, the branch of National Accounts, even if inclusive of tourism activities, can hardly be entirely overlapped to the corresponding tourism industry as understood in TSA, as the scope of a TSA is different and consists in grouping economic activities in a tourism logic. Therefore, after a first screening of the NA's branches, their analysis has been developed in three different steps using basic data provided by SBS surveys. For each selected branch from ISUT relevant for IT5, the process has been as follows:

- breakdown of each branch in its *classes* of economic activity – see column 2, Table 4;
- exclusion of all those *classes* in accordance with the NA's grouping concept but not relevant for the IT5's industry concept – see column 3 Table 4;
- composition of tourism industry by means of a new criterion of grouping, based on the selection of the tourism *classes* – see column 4 Table 4.

Table 4 shows, firstly, as the overlapping between *industry* (TSA) and *branch* (NA), in terms of number of *classes*, is not common. Furthermore, the different grouping criteria used in National Accounts and in T5 are based on different logics, so to make possible to split a *branch* into more than one tourism *industry* – e.g. branch n. 84 "*Renting and operating leasing activities*", involved in two different tourism industries – 7 and 10 (Table 4).

[38] Market production is defined as that referring to goods and services for sale, mostly inclusive of products sold at economically significant prices.

[39] Non market production is defined as that referring to goods and services not for sale, mostly inclusive of products free of charge or sold at non economically significant prices.

The content of Italian tourism industries in terms of class of economic activities is almost totally aligned with international standards. However, specific considerations about the Food and beverage's industry have led to a slight deviation

The class of economic activity 55.29 is excluded in IT5 but included in international T5. During the work of IT5's implementation, an in-depth analysis of the correspondence between products/activities has been carried out. Crossing information from the supply and demand side it appears clear that IRTS2008 –*Annex 2* - has specified that the inclusion of product CPC 63399 "*Other food serving services*" is justified when it is attributable to activities relating to a *"food provided by refreshment stands, fish-and-chips stands, fast-food outlets without seating, take-away facilities, etc., ice-cream parlors and cake serving places, vending machines, motorized or non-motorized carts, etc."* [40]. In the Italian statistic context, these services are related to activities included in the *class* 56.10 of Italian classification (ATECO), already included in the valuation of tourism industry, as well as in the *class* 47.81, concerning the "*Retail sale via stalls and markets of food, beverages and tobacco products"*. The United Nations defines the product 63399 as related to activities carried out in class of economic activity 56.10, as in the Italian case, and in *class* 5629, excluded in IT5.

On the basis of these analysis it was considered that the tourist part relating the product 63399 could have been attributed to the *class* 56.10.

Unlike that for activities identification's process, that for products has suffered from the lack of basic information. This has meant that it has been possible to work only using the detailed articulation provided by Italian SUT.

This lack of basic data for products has compromised the level of fineness of the production estimate process in the context of IT5, but didn't affect the measurement of the role of tourism sector as a whole within the national economic system.

Indeed, in general terms, any estimating process hides the risk of over or under estimation. So, for some tourism industries, as well as for some products, total output could be overestimated due to the "gross" content of the basic information. However, regardless the risk of over or under estimation, in TSA the total supply of a commodity usually exceeds tourism demand. In T5 the total domestic supply by product is estimated, but part of it will be excluded from tourism consumption and will be destined to the satisfaction of non-tourism demand. Indeed, as already stated, visitors are a subset of consumers, and a very few products are totally tourism-oriented[41].

The gross content of production provided by T5 will be adjusted in a tourism use only when compared with internal tourism consumption, namely in T6.

In the following Table 5 are listed the items of products derived from INA which led to the corresponding tourism products.

[40] *IRTS:2008*, par. 5.27, pg. 50.

[41] Perhaps, even travel agency's services can be bought by a consumer, and not by a visitor, albeit the tourism share of this product can be 100%. In case of change of usual environment, for example, the transport service purchased in a TA should not be considered tourism expenditure, since the purpose of the trip is not tourist.

Table 4 – Derivation of tourism industries (IT5) from branches of economic activities (ISUT*)

INA'S SELECTED BRANCHES OF ECONOMIC ACTIVITIES	ISIC. REV. 4 CODE – CLASSES INCLUDED IN THE SELECTED BRANCHES	ISIC. REV. 4 CODE – CLASSES INCLUDED IN TOURISM INDUSTRIES	IT5 TOURISM INDUSTRIES
61- Accommodation	**5510-5520-5530-5590**	**5510-5520-5530-5590**	
75- Buying and selling of real estate and real estate activities for third parties	**6810-6831-6832**	**6810-6831-6832**	**1- Accommodation for visitors**
76- Rental and management of properties owned or leased	**6820**	**6820**	
62- Food and beverage serving activities	**5610-5621-5629-5630**	**5610-5630**	**2- Food and beverage serving activities**
54- Railway transport	**4910-4920**	**4910**	**3- Railway passenger transport**
55- Other land passenger transport	**4931-4932-4939**	**4932-4939**	**4- Road passenger transport**
57- Maritime and inland water transport	**5010-5020-5030-5040**	**5010-5030**	**5- Water passenger transport**
58- Air transport	**5110-5121-5122**	**5110**	**6- Air passenger transport**
84- Renting and operating leasing activities	**7711-7712-7721-7722-7729-7731-7732-7733-7734-7735-7739-7740**	**7711**	**7- Transport equipment rental**
86- Services activities of Travel Agencies, Tour Operators and related reservation services and activities	**7911-7912-7990**	**7911-7912-7990**	**8- Travel agencies and other reservation services activities**
97- Creative, arts and entertainment activities	**9001-9002-9003-9004**	**9001-9002-9003-9004**	**9- Cultural activities**
98- Libraries, archives, museums and other cultural activities	**9101-9102-9103-9104**	**9102-9103-9104**	
84- Renting and operating leasing activities	**7711-7712-7721-7722-7729-7731-7732-7733-7734-7735-7739-7740**	**7721**	
99- Lotteries, betting and casinos related activities	**9200**	**9200**	**10- Sports and recreational activities**
100 – Sports, amusement and recreation activities	**9311-9312-9313-9319-9321-9329**	**9311-9319-9321-9329**	

*according to the articulation in 106 branches.

Table 5 – Derivation of tourism products (IT5) from branches of homogeneous production (ISUT*)

SELECTED PRODUCTS RELEVANT FOR TOURISM	IT5 PRODUCTS
194 - Hotel services and similar	1 - Accommodation services for visitors
195 - Other accommodation services	
215 - Sale of real estate services with own property made services	
216 - Real estate services for third parties	
217 - Administration services and property management for third parties	
218 - Real residential housing services	
219 - Imputed residential housing services	
196 - Food and beverage sales services	2 - Food and beverage serving services
175 - Interurban railway passengers transport services	3 - Railway passenger transport services
178 - Other land passenger transport services	4 - Road passenger transport services
184 - Shipping, cabotage and inland water passengers transport services	5 - Water passenger transport services
187 - Air transport services of passengers	6 - Air passenger transport services
233 - Services of renting and leasing of consumer goods for recreation and leisure	7 - Transport equipment rental services
237 - Travel agencies services	8 - Travel agencies and other reservation services
238 - Tour Operators services	
239 - Other reservation service and related services	
254 - Library, archives, museums and other cultural services	9 - Cultural services
255 - Creative, art and entertainment services	10 - Sports and recreational services
256 - Services relating to gambling	
257 - Management services of sports arenas and sports facilities	
258 - Sports entertainment and recreation services	

*according to the articulation in 266 products

5.3 Evaluation of production of tourism industries

As previously mentioned, by highlighting the tourism perspective IT5 represents a different way to display the national economy, both in its *market* and *non market* components.

The procedure of assessing the production of tourism industries required a different way to proceed, depending on which concerned the *market* component or *non market* component.

As will be discussed in the following paragraphs, *market* component has been reconstructed starting from basic data at the highest level of detail available (*class* of

economic activity) and through appropriate methodologies to bring coherence to the basic data with the national constraint.

Instead, for *non-market* component an overall evaluation has been carried out in order to establish the share of this production – related to the Government institutional units – within the total economy.

5.3.1 Evaluation of market production

In most cases, availability of basic data by *class* of economic activity allowed to identify those relevant for tourism industries, and consequently to exclude non-tourism *classes*. Whereas higher level of detail needed, employment's basic data, provided at 5 digits level of the Italian's classification of economic activity, was used as splitting indicator of *class*'s production. In one case the lack of basic data up to the 6 digits of Italian classification resulted in the inability to exclude that part of production related to non-tourism activity[42].

The methodology below explained refers exclusively to the market component of Italian production.

With reference to the above Figure 1, each branch of economic activity derived from Italian NA and selected for tourism sector has been further widened in terms of *classes* in it included. In the following Figure 2 two examples are highlighted: the first concerns a branch entirely tourism-related in terms of *classes* of economic activities (e.g. "*Services activities of travel agents, tour operators and related reservation services and activities"*); the second a branch partially tourism-related (e.g. "*Renting and operating leasing activities*").

For each class of economic activity (*class*) included in the selected branches of NA, SBS survey provided basic information about value of production.

Total production, by branch, obtained as the sum by *class* through basic data is not yet the value shown in SUT matrix of production, as this latter is the final result of the accounting balancing procedure with the consumptions – *Use* –, whereas the former is still a pre-balanced data. Nevertheless, production by *class* has allowed to calculate a weighting structure by *class* within its branch of reference.

[42] The reference is to the Cultural services Industry.

Figure 2- Simplified scheme of Italian production matrix, by class of economic activities

CN products	NA branches of economic activities											Total
TSA products	NA selected branches for T5's scope											
	1	2				i	n				N	
		C_1	C_k	C_S	$\sum$		C_1	C_k	C_S	$\sum$		
1	p_{11}	$_{(b2)}p_{11}$	$_{(b2)}p_{1k}$	$_{(b2)}p_{1S}$	p_{12}	p_{1i}	$_{(bn)}p_{11}$	$_{(bn)}p_{1k}$	$_{(bn)}p_{1S}$	p_{1n}	p_{1N}	$\mathbf{P_{1.}}$
2	p_{21}	$_{(b2)}p_{21}$	$_{(b2)}p_{2k}$	$_{(b2)}p_{2S}$	p_{22}	p_{2i}	$_{(bn)}p_{21}$	$_{(bn)}p_{2k}$	$_{(bn)}p_{2S}$	p_{2n}	p_{2N}	$\mathbf{P_{2.}}$
…	…	…	…	…	…	…	…	…	…	…	…	**…**
r	p_{r1}	$_{(b2)}p_{r1}$	$_{(b2)}p_{rk}$	$_{(b2)}p_{rS}$	p_{r2}	p_{ri}	$_{(bn)}p_{r1}$	$_{(bn)}p_{rk}$	$_{(bn)}p_{rS}$	p_{rn}	p_{rN}	$\mathbf{P_{r.}}$
j	p_{j1}	$_{(b2)}p_{j1}$	$_{(b2)}p_{jk}$	$_{(b2)}p_{jS}$	p_{j2}	p_{ji}	$_{(bn)}p_{j1}$	$_{(bn)}p_{jk}$	$_{(bn)}p_{jS}$	p_{jn}	p_{jN}	$\mathbf{P_{j.}}$
m	p_{m1}	$_{(b2)}p_{m1}$	$_{(b2)}p_{mk}$	$_{(b2)}p_{mS}$	p_{m2}	p_{mi}	$_{(bn)}p_{m1}$	$_{(bn)}p_{mk}$	$_{(bn)}p_{mS}$	p_{mj}	p_{mN}	$\mathbf{P_{m.}}$
M	p_{M1}	$_{(b2)}p_{M1}$	$_{(b2)}p_{Mk}$	$_{(b2)}p_{MS}$	p_{M2}	p_{Mi}	$_{(bn)}p_{M1}$	$_{(bn)}p_{Mk}$	$_{(bn)}p_{MS}$	p_{Mn}	p_{MN}	$\mathbf{P_{M.}}$
Total	$\mathbf{P_{.1}}$	$_{(b2)}\mathbf{p_{.1}}$	$_{(b2)}\mathbf{p_{.k}}$	$_{(b2)}\mathbf{p_{.S}}$	$\mathbf{P_{.2}}$	$\mathbf{P_{.i}}$	$_{(bn)}\mathbf{p_{.1}}$	$_{(bn)}\mathbf{p_{.k}}$	$_{(bn)}\mathbf{p_{.S}}$	$\mathbf{P_{.n}}$	$\mathbf{P_{.N}}$	$\mathbf{P_{..}}$

C= class of economic activity

It can be written as:

P.., the overall matrix production of Italian SUT, 266 products*106 branches ;

$P_{.i}$ the total balanced production value for the generic *i*-th branch;

$\sum_{k=1}^{S} \widehat{p}_k = \widehat{P}_{.i}$ the pre-balanced production value for the generic *i*-th branch, for k:1,…..S number of *classes* included in the branch.

The weight of the production by *class* within its branch of reference is given as:

$$x_k = \frac{\widehat{p}_k}{\widehat{P}_{.i}}$$

where *k*-th is the generic *class* and where

$$\sum_{k=1}^{S} x_k = 1$$

Such weighting structure, applied to the total balanced production of NA by branch, led to a new balanced level of production by *class*, given as:

$$p_k = x_k P_{.i}$$

now being

$$\sum_{k=1}^{S} p_k = P_{.i}$$

The result reflects, on the one hand, the production ratio by *class*, calculated starting from SBS basic (pre-balanced) data – x_k -; on the other hand, it benefits of the balancing procedure used in NA to guarantee the equality accounting achievement, as a result of the weighting of basic data by *class* with total amount of the production provided by NA – $P_{.i}$

In order to quantify the value of production for each tourism industry, according to its composition in terms of *classes* –Table 4 -, a selection of only tourism *classes* has been carried out. Appropriately grouped and summarized, such values define the new total of production for each tourism industry:

$$\sum_{k=1}^{s} p_{.k} = {}_{(I)}P_{.i}$$

for k:1,…..s the number of the subset of tourism *classes* included in the *i*-th selected branch, now become a tourism industry ($_I$)

Production by *class* provided by SBS data is not articulated by *product*. So, in order to meet such crucial requirement, a second weighting structured appeared necessary. As previously said, the lack of basic data by *product* led to the need of implementing a weighting structured using the articulation by product furnished by ISUT.

More specifically, given the value of production for the generic *j*-th product for the *i*-th branch, the weight by product can be written as:

$$\lambda_{ji} = \frac{p_{ji}}{P.i}$$

finally allowing to estimate production by *class* and by *product* as follows:

$$\lambda_{ji} p_{.k} = p_{jk}$$

Now, for each *class*:

$$\sum_{j=1}^{M} \lambda_{ji}\ p_{jk} = p_{.k}$$

and *industry:*

$$\sum_{j=1}^{M}\sum_{k=1}^{s} p_{jk} = {}_{(I)}P_{.i}$$

it is possible to articulate total production by products, for j:1,…..M, the number of products provided by the highest level of detail available[43]. In doing so it is assumed that for each product the incidence on production is the same than for all the *classes* included in the branch of reference, regardless their kind of activity.

[43] 266 items of products.

5.3.2 Evaluation of non market production

"*The public sector plays an important role in the development of tourism activities in many countries. It establishes the legal framework for the tourism activity. It establishes certain controls on the production of services, and in some cases guarantees the quality of the service that is provided through the provision of licenses and the development of codes of conduct*"[44].

The value of these different activities developed by the public administration can be established along the same parameters of measurement as any other collective non-market services, that is, through their cost of production. For public sector the value of consumption is, by convention, equal to the value of production.

Despite the availability of basic information up to 5 digits level of detail, estimates has been carried out in a different way that for market production. More specifically, the overall non-market matrix of production of NA has been analyzed in a perspective of comparative incidence on the total economy.

This choice is the result of various considerations. Evaluation criteria of basic data for non-market production are different then those for market producers. This would have required the development of an estimation process completely different from that described as above. At this stage of development of the implementation process of the first ITSA it was considered sufficient make an overall assessment of the role of public administration in the tourism sector as a whole.

Future improvements and advances in development and compilation works for the whole ITSA will regard the research of a more detailed estimation of the non-market's component of production for IT5. An additionale next step, in this specific area, could be the estimates of tourism collective consumption[45].

5.4 Relationship between tourism industries, other industries and products.

The ensemble of tourism activities listed in Table 4 defines the tourism industry as a whole, namely, the concept of tourism supply in the frame of TSA.

"*Tourism supply is understood as the direct provision to visitors of the goods and services that make up tourism expenditure. The analysis of tourism supply consists, first, in showing how the conditions are created that enable producers to provide goods and services to visitors, and, second, in describing the processes, the production costs and the economic performance of the suppliers in the tourism industries*"[46].

As just shown, in TSA tourism sector emerges from the national productive system of NA. In this regard it has already pointed out as in National Accounts a branch of economic activity is defined as "*a group of establishments engaged on the same, or similar, kinds of activity. At the most detailed level of classification, an industry consists of all the establishments falling within a single class of ISIC and which are therefore all*

[44] *IRTS:2008*, pg.60.
[45] Table 9 in the TSA.
[46] *IRTS:2008*, pg. 54

engaged on the same activity as defined in the ISIC"[47].

As in the SNA, as well as in the TSA, the establishment is used for the analysis of production and production processes. Establishments are classified according to their main activity[48], as a consequence, "*the grouping of all establishments with the same main activity which serves visitors directly and that is one of the tourism-characteristic activities constitutes a tourism industry"*[49]. However, an establishment that cater to visitors may often have more than one productive activity. Establishments,

Tourism sector as understood in TSA from the supply side is not an industry in the sense of branch: it cuts across several branches in the economic activities classification. In other words, there is no correspondence one-to-one between tourism sector as a whole and a branch of economic activity, because the former depends on the demand's purchases behavior.

Additionally, some industries are included in tourism sector even though the majority of their output can be attributed to non-tourism products. The Food and beverage services and Recreation and entertainment industries are pregnant examples. Such industries are included because without tourism their level of activity would be significantly reduced. Nevertheless, as the purchases of these goods and services is not entirely tourism, the TSA must identify and separate out the production's tourism components from each of industry.

On the other hand, tourism industries that specialize on tourism products can develop a main tourist activity and perform one or several secondary activities. It means that characteristic tourism activities that specialize on tourist products develop activities that are not exclusively tourist-based, or may have other tourism characteristic secondary activities.

By the same token "*establishments having a particular tourism-characteristic activity as a secondary activity should not be included in the tourism industry that is characterized by this activity*"[50].

As a result, the total output of any product, tourism or not, is the sum of the output generated from all industries, tourism or not, regardless that it is bought by visitors and non-visitors (consumers).

Such situation directly impact on the structure of table 5, and is illustrated in the below Figure 3. The main output of tourism industries is by definition tourism characteristic products, but they may also produce tourism connected products and other products. The main output of other industries might be any thing other than tourism characteristic products. The total output of any product is the sum of the output of this product from the total industries in the economy.

[47] *SNA93*, Chapter V, pg. 144. For tourism statistics from the supply side, according to the Regulation on tourism statistics, the statistical unit is the local kind of-activity unit. Instead, in the context of a TSA, the statistical unit of productive activities, that is to say the representation of supply side, is the *establishment*. As already mentioned, this difference is emerged in the phase of compilation of IT7 and IT10.

[48] Determined by the activity that generates the most value added.

[49] *IRTS:2008*, par.6.16, pg.55.

[50] *IRTS:2008*, pg 55.

Figure 3 – Flows of main and secondary outputs by industry

Products (P)	Industries (I)										total output by product	
	Tourism Industries (T)					Other industries (O)						
	1		…		n	1		…		n		
Characteristic products												
C_1	X		x		x	x		x		x	Σ	C_1
…	x		X		x	x		x		x	Σ	..
C_n	x		x		X	x		x		x	Σ	C_n
Connected products	x		x		x	XX		XX		XX		
CO_1	x		x		x	XX		XX		XX	Σ	CO_1
…	x		x		x	XX		XX		XX	Σ	..
CO_n	x		x		x	XX		XX		XX	Σ	CO_n
Other products	x		x		x	XX		XX		XX		
O_1	x		x		x	XX		XX		XX	Σ	O_1
…	x		x		x	XX		XX		XX	Σ	..
O_n	x		x		x	XX		XX		XX	Σ	O_n
Total output of the induxtries	Σ	T_1	Σ	…	Σ	T_n	Σ	O_1	Σ	…	Σ	O_n	$\sum$	

X= main output of the industry; x= secondary possible output of the industry; XX= main possible output of the industry
Sources: IRTS2008, Fig. 6.1, pg.56.

With reference to the Italian tourism industries the assessment of production by product has required their preliminary partition into three subset of products: tourism characteristic, Italian specific tourism characteristic, residual.

More specifically, articulation by product made available by matrix of production of Italian National Account has been marked by specific indicators:

1. "c" for characteristic products falling into the first Italian 10 categories listed in Table 2;
2. "b" for Italian specific products (among those identified), falling into the Italian 11th category of products listed in Table 2;
3. "r" for Italian residual products, headed "*Other consumption and non consumption products*" in Table 2.

These indicators make possible to articulate total production of each tourism industry into three subset of products.

Given as:

c:1,…..m = the number of tourism characteristic products;
b:1,……n = the number of Italian specific tourism characteristic products;
r:1,……z = the number of residual products.

The total production by industry is articulated as follows:

$$\sum_{k=1}^{s} \left(\sum_{c=1}^{m} p_c + \sum_{b=1}^{n} p_b + \sum_{r=1}^{z} p_r \right) = {}_{(I)}P_{.\,i}$$

From the activities side, the residual Italian productive units, "*Other industries*", complete the display of Italian economy in a tourism perspective. The assessment of its

production has required a different and a more simplified approach from that for tourism industries. For this latter the whole evaluation process is based on 4 digits level of basic information, whereas for "*Other Industries*" of IT5 the total production by product is obtained as difference between the NA constraint by product and the total imputed to all tourism industries.

With reference to the *j*-th generic item of product, the value of production imputed to *Other Industries* is simply calculated as:

$$_{(R)}P_{j.} = {}_{(NA)}P_{j.} - {}_{(TI)}P_{j.}$$

where the subscript on the left refers to the residual macro sector of *Other Industries* ($_R$); to data of matrix of production of National Accounts ($_{NA}$) and to all tourism industries ($_{TI}$).

The last step is to group such estimates according to the structural requirements of T5. The below Figure 4 displays in a schematic way how data has been organized in order to fill the production account of Italian tourism industries and other industries.

Figure 4 - Simplified scheme of Italian T5

Products	Tourism industries					Other industries	Output of domestic producers
	1	…	10	11	Total		
C. Characteristic products							
1	X	x	x	x		x	$\sum_{c=1}^{m} p_c$
…	x	X	x	x		x	
m	x	x	X	x		x	
B. Country-specific tourism characteristic products and connected products	x	x	x	X		x	$\sum_{b=1}^{n} p_b$
R. Other products	x	x	x	x		XX	$\sum_{r=1}^{z} p_r$
Total output							
Total intermediate consumption							
Total value added							

X= main output of the industry; x= secondary possible output of the industry; XX= main possible output of the industry

6. Special issues

In principle, satellite accounts share the structure of the central ones of NA, but in same cases their specific perspective requires to take the distances with the general criteria adopted by the national accounts framework.

As far as the TSA is concerned, from the supply side the pursuit of a tourism logic has led to face the following special issues:

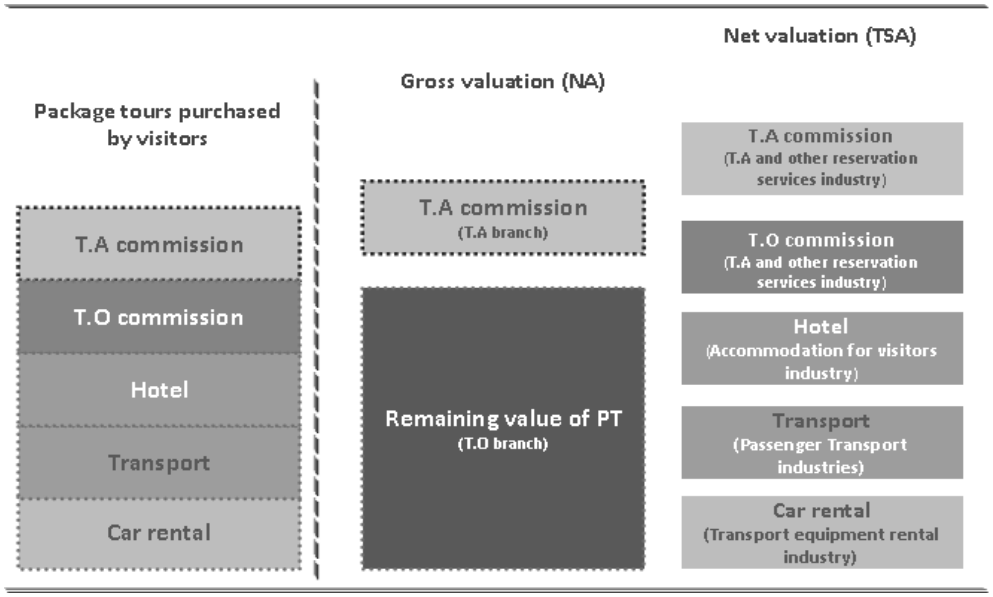
- net valuation of package tours;
- usage of second homes for tourism purposes.

6.1 From gross to net valuation of package tours (PT)

Total production estimated in IT5 as described in par.5 is not yet the whole Italian production in a tourism perspective. In fact, according to the TSA logic, the tour operating activity, which represents the main activity of Tour Operators (TO) and whose value is included in the final price of a PT, "*must be treated separately from the rest of services composing a PT and purchased through their intermediation*"[51]. In respect of such separation principle, the value of each service bundled in a package[52] must be imputed, both in terms of production and of intermediate consumptions, to the involved industries.

In Italian NA the value of PT is twice imputed (gross valuation): once to the TO branch, once more implicitly included in the total value of production of the involved branches. This duplication – both for production and for intermediate consumptions – vanishes when value added is calculated, defference between intermediate consumptions and production value. Otherwise, from a TSA point of view the value of PT is required in a net valuation.

Figure 5 - Accounting recording of a PT in NA and in TSA



In order to meet this requirement it has resorted to two different sources, at first separately treated:

a) administrative source;
b) statistical source (SUT).

[51] *IRTS:2008*, pg. 87

[52] Typical services included in a PT are: transport, accommodation, excursions and guided tours, car rental, TA's commissions.

The administrative source has provided the total amount of revenues and of costs specifically related to tour operating activity[53]. The approach followed was *top-down*: the difference between revenues and costs, and its share on total revenues, has conducted to a first estimate of the net valuation of PT.

Statistical source, in this case the matrix of intermediate consumptions derived from *Use* tables, has provided costs by product, allowing to pursuit a *bottom-up* approach. First of all, exploiting the level of detail of Italian SUT an adequate selection of products that may be included in a PT has been made. For each of these input of (tourism) services included in a PT the cost was estimated in two different ways.

For some services – e.g. transport[54], insurance services, cultural services – the calculation of a ratio between the main production of TO – PT by definition - and the total production of the accounting branch in which is enclosed appeared necessary:

$$\lambda_{jto} = \frac{p_{jto}}{P_{.to}}$$

where the subscript $_{TO}$ indicates both product and branch of Tour Operator, according to the Italian articulation 266x106. This ratio has been applied to the value of cost for each item of this first subset of services.

For the other elements of cost – e.g. commissions for travel agents, air transport, accommodation – the value provided by matrix of intermediate consumptions has been considered entirely relevant[55].

Although based on different approaches, the two assessment processes showed very similar results. A reconciliation between the two sources has been operated, so to lead to the net valuation of PT as required in *IRTS:2008*.

6.2. Usage of second homes for tourism purposes

"*The ownership of a vacation home on own account is peculiar, from a statistical perspective, because it generates both a tourism characteristic service and an equivalent tourism consumption. In the SNA 1993, a housing service on own account is associated with the ownership of a dwelling occupied by its owner, both as a production activity and as the output and consumption of a specific service. This situation covers both the principal dwelling and all other dwellings owned by a household for its own use. It covers in particular owner-occupied vacation homes*"[56].

The NA activity and product associated with the ownership of a dwelling are totally included in a TSA in the correspondent tourism industry and product – see Table 4 and Table 5.

---

[53] The use of administrative data poses problems of consistency with NA requirements. So this source has been used in order to estimate a weighting structure for costs and revenues to apply at TO production data derived from NA.

[54] Except for air transport.

[55] For these services the whole value of intermediate consumptions has been imputed to the activity of tour operating, regardless that part of them might be costs for business travels, which mainly involve accommodation and transport services.

[56] *TSA:RFM2008*, pg. 31.

In principle T5 reflects the total output of tourism and non-tourism industries, regardless the share directly demanded by visitors. However, the treatment of this specific issue required special attention during compilation of IT5. In NA's production matrix the value for imputed residential housing services is quite significant because refers to all Italian's households owners, regardless the purpose of use of dwelling. In the frame of T5 this value of production would have been totally contained because referring to tourism class and tourism product. In our opinion it would have introduced a distortion about the role performed by this industry within the whole tourism sector.

For this reason, in the context of IT5, the treatment of second homes industry represent an exception, as the total output of the industry is not a *gross* production, but yet a *tourist* part of it[57]. In particular, the total production of the Italian NA branch has been weighed with the national share of use of second homes for tourism purposes.

For this purpose the resort to two sources needed, in particular:

- final consumption expenses of households;
- census of housing.

In summary, given γ as the national incidence of a tourism use of second homes, production of this tourism industry is calculated as follows:

$$\sum_{j=1}^{M}\sum_{k=1}^{s} p_{jk}\gamma$$

6.3. Intermediate Consumptions (IC) and value added (VA)

In T5 IC are shown for industry's vector. Unlike that for production, the analytical purposes of this aggregate does not require, within a TSA's scheme, its disarticulation by product[58].

Nevertheless, at basic information level, the process of reconstruction of IC of Italian tourism industries duplicated the methodology for the production estimate, although selection and distinction of products has not been carried out.

Finally, as for the total production, also intermediate consumptions are not perfectly comparable to those related to the whole Italian economy, due to the net valuation (v) of package tours.

For the whole economy VA can be written as:

$$VA = P1 - P2$$

[57] This type of evaluation is responsible for Table 6, where the total supply by product is compared with its total tourism demand, so as to achieve a tourism share.

[58] *TSA in the European Union*, Vol.3, par.4.17, pg.42 "*it is simply suggested, if possible, to consider a breakdown of IC into the following categories of products: agriculture; ores and minerals; food and beverage; machinery and equipment; construction; distributive trade services; financial and related services; business and production services; community, social and personal services*".

differing from the tourism perspective of T5, where it is given:

$$VA= [(P1\text{-}v) – (P2\text{-}v)]$$

7. Compilation of Italian T5 by means of a structure of worksheets

The whole methodological process for compilation of IT5 is developed through a structure of different worksheets for the redistribution of the production and IC by the different NA industries and products.

First of all, a key structure from SBS at 4 digits level of Italian classification of economic activities (ATECO) has been used to decompose the total production of NA branches into the different TSA industries. The structure of IT5, based on a logic cross-industries/products, essentially reproduces the format adopted in NA for production and IC matrices, which constitutes the direct starting point for the implementation of IT5[59]. According to the highest level of detail of Italian national account balanced datasets available, efforts aimed at enlarging the internal compilation level of Italian SUT according to the needs of T5 industries and products classification.

The structure of worksheets for compilation of IT5 was built through the following phases:

1. analysis of matrices of production (P1) and of intermediate consumptions (P2) derived from Italian SUT in order to identify those branches and products suitable for the needs of IT5;
2. treatment of basic economic statistics deriving from SBS and of employment data from NA datasets, both useful to build a weighting structure by *class* to which to bind the SUT balanced data;
3. calculation of a new matrix of synthesis (both for P1 and for P2) enlarged by *class* of economic activity, for each tourism branch selected with regards to the needs of IT5;
4. dumping of all matrices of synthesis in an overall matrix suitable for the calculation of the value added of tourism industries and for comparison within the rest of economy;
5. automatic compilation of IT5, aggregating data by *class* and by product into the different IT5 industries and products.

In the immediate following these steps will be highlighted in a synthetic way.

7.1 Phase 1: analisys of P1 and P2 matrices of Italian SUT.

Starting point of IT5 compilation are the matrices of production and of IC from Italian SUT at the highest level of detail. Therein, the tourism involved branches has been identified and marked, whereas all products has been differently indicated according to their degree of tourism characteristicity – see Table 6.

[59] As well as for the subsequent T6, which compares internal consumption and total supply.

Table 6 - Simplified scheme of Italian P1 and P2 matrices for calculation of IT5 aggregates

RODUCTS	PRODUCT INDICATOR*	Branches of economic activity							Total
		1	2	…	i	…	…	N	
		r	c	…	b	…	…	r	
1	r	p_{11}	p_{12}	…	p_{1i}	…	…	$p_{1.N}$	$p_{1.}$
2	c	p_{21}	p_{22}	…	p_{2i}	…	…	$p_{2.N}$	$p_{2.}$
…	r	…	…	…	…	…	…	…	…
j	b	p_{j1}	p_{j2}	…	p_{ji}	…	…	$p_{j.N}$	$p_{j.}$
…	r	…	…	…	…	…	…	…	…
M	c	$P_{M.1}$	$P_{M.2}$	…	$P_{M.i}$	…	…	$P_{M.N}$	$P_{M.}$
Total		$P_{.1}$	$P_{.2}$	…	$P_{.i}$	…	…	$P_{.N}$	$P_{..}$

*symbols explained in par. 5.4.

The below Table 7 illustrates the example for *Food and beverage serving service* tourism industry. In it the value of production, by product, derived from Italian SUT, is shown.

Although Italian articulation is significantly enough in order to isolate the whole tourism industry, it does reveal not sufficient to split out tourism activities and no-tourism activities which compose the NA branch. The report to basic information has been crucial in order to articulate the total output of branch in its *classe's* contribution.

Table 7 – Output of Food and beverage serving service branch, by product. Year 2010 – million of euro

N*	Products	
153	Repair services	4
166	Specialized construction services	37
173	Retail trade, except of motor vehicles and motorcycles	3.837
196	Food and beverage services	67.098
197	Other food services	7.890
206	Software services for own account	4
218	Real residential housing services	19
220	Imputed residential housing services	58
227	Scientific research and development	4
235	Leasing of intellectual property and similar products, except copyrighted works	13
256	Gambling and betting services	975
TOTAL		79.939

* According to the articulation in 266 products - Source: Italian NA: SUT – matrix of production at basic price. Differences in sums are due to rounding

7.2 Phases 2 and 3: processing basic data and matrix of synthesis for tourism industries.

For each selected branch it was estimated the total value of output starting from the pre-

balanced basic information provided by SBS at 4 digits level of detail, then processed in order to establish the incidence of each *class* within its branch of reference - x_k. An additional weighting structure by *product* - $\lambda_{ji.}$ – obtained with data of NA matrix of production allows to arrive at a new balanced value of output for each *class* of economic activity, and articulated in 266 products, using the format of NA for production matrix – so called matrix of synthesis.

Table 8 shows the same data of Table 7, now articulated by each *class* of economic activity which compose the whole accounting branch of *Food and beverage serving service*. The lack of basic data even for products only allowed to proceed at the INA articulation level. Nevertheless the detailed used for IT5 makes quite easy to identify tourism characteristic products (indicated by the product indicator "c" in Table 8), whereas the category of Italian characteristic specific products results to be less precise (indicated by the product indicator "b" in Table 8).

As for products, tourism and non-tourism *classes* of economic activity has been differently marked in order to leave the accounting branch concept and to meet the tourism industry concept.

Table 8 – Output of Food and beverage serving service branch, by class and product. Year 2010 millions of euro

N*.	Products	Product indicator**	ISIC-REV.4 5610	ISIC-REV.4 5621	ISIC-REV.4 5629	ISIC-REV.4 5630	Total output by product
			c	r	r	c	
153	Repair services	r	2	0	0	1	4
166	Specialized construction services	r	19	1	3	13	37
173	Retail trade, except of motor vehicles and motorcycles	b	1.978	127	349	1.383	3.837
196	Food and beverage services	c	34.581	2.226	6.100	24.191	67.098
197	Other food services	r	4.066	262	717	2.845	7.890
206	Software services for own account	r	2	0	0	1	4
218	Real residential housing services	c	10	1	2	7	19
220	Imputed residential housing services	r	30	2	5	21	58
227	Scientific research and development	r	2	0	0	1	4
235	Leasing of intellectual property and similar products, except copyrighted works	r	7	0	1	5	13
256	Gambling and betting services	c	502	32	89	351	975
Total			41.198	2.652	7.268	28.820	79.939

*According to the articulation in 266 products – **Symbols explained in par. 4.4 - Estimates on SUT data and SBS data. Differences in sums are due to rounding

The construction of the matrix of synthesis for *Food and beverage serving service* branch begins by selecting the only tourism *classes* and tourism products – see Table 9, thus clearing the content of the previous Table 8 for those parts which are not strictly

tourist. Total output by product of the industry may be equal to the total output by product of the accounting branch, according to the number of tourism *classes* comprised in tourism industry – all *classes* or a part of it.

Tourist output of tourism industry is the sum of characteristic and specific products ("c" and "b" in Table 9). Most of the production value of the *Food and beverage* tourism industry is obviously represented by its main output, that is *Food and beverage service*. Nevertheless, it is quite spread the case of secondary productions, which may be still tourism output. For *Food and beverage* tourism industry tourist outputs (main and secondary) are listed in Table 9.

Table 9 – Output of Food and beverage serving service industry, by tourism characteristic *class* and tourism characteristic product. Year 2010 – millions of euro

N*.	Products	Product indicator**	ISIC-REV.4 5610	ISIC-REV.4 5621	ISIC-REV.4 5629	ISIC-REV.4 5630	Total output by product
			c			c	
173	Retail trade, except of motor vehicles and motorcycles	b	1.978			1.383	3.361
196	Food and beverage services	c	34.581			24.191	58.772
218	Real residential housing services	c	10			7	17
256	Gambling and betting services	c	502			351	854
Total			37.071			25.932	63.004

* According to the articulation in 266 products – **Symbols explained in par. 4.4 - Estimates on SUT data and SBS data. Differences in sums are due to rounding

Besides tourism products (characteristic and specific) *Food and beverage serving service* industry may produce other outputs, classified as non-tourism - "r" – see Table 10 – and which correspond to a residual category of outputs. As mentioned, for IT5's scope this production must be included in the total output by industry.

As far as the non-tourism *classes*, they are excluded from the concept of tourism industry, but will constitute the content of the residual *"Other industries"* category, as to equal the aggregates of production for the whole Italian economy.

Matrices of synthesis of intermediate consumptions slightly differ from the format use for production. In particular, they do not report the articulation of output by "c", "b" and "r" products - columns 7, 8 and 9 in Table 10 - since such articulation is not required and relevant for costs.

Table 10 – Output of Food and beverage serving service industry, by product Year 2010 – millions of euro

N*.	Products	Product indicator**	ISIC-REV.4 5610	ISIC-REV.4 5630	Total output by product:	Industry output by product			Other industries (ISIC-REV.4 5621/5629)
			c	c		c	b	r	
153	Repair services	r	2	1	4			*4*	*1*
166	Specialized construction services	r	19	13	32			*32*	*5*
173	Retail trade, except of motor vehicles and motorcycles	b	1.978	1.383	3.361		*3.361*		*476*
196	Food and beverage services	c	34.581	24.191	58.771	*58.771*			*8.327*
197	Other food services	r	4.066	2.845	6.911			*6.911*	*979*
206	Software services for own account	r	2	1	4			*4*	*0*
218	Real residential housing services	c	10	7	17	*17*			*2*
220	Imputed residential housing services	r	30	21	51			*51*	*7*
227	Scientific research and development	r	2	1	4			*3*	*0*
235	Leasing of intellectual property and similar products, except copyrighted works	r	7	5	11			*11*	*2*
256	Gambling and betting services	c	502	351	854	*854*			*121*
Total			41.198	28.820	70.018	***59.642***	***3.361***	***7.015***	***9.920***

* According to the articulation in 266 products – **Symbols explained in par. 4.4 - Estimates on SUT data and SBS data. Differences in sums are due to rounding

7.3 Phase 4: overall matrix for the calculation of the value added.

All matrices of synthesis of the production of tourism industries in their integral format constitute the structure of a new, overall matrix. In order to allow calculation of value added of tourism industries, such matrix is integrated with total values of intermediate consumptions of industry[60]. Now it is possible to obtain the most important indicator of T5: the value added of tourism industries (VATI). In this macro matrix all aggregates are available, in their total value, at different level of detail: by *class*, by *branch* and by *industry*.

7.4 Phase 5: automatic compilation of IT5.

Appropriately grouped to meet the concept of tourism industry and tourism product, data previously processed are automatically transferred in IT5 through a unique link key.

In standard format of T5 –Figure 4 - is not immediately apparent the contribution, in terms of share of production, intermediate consumptions and value added, of each industry

[60] However, due to the disposal of basic information, P2 data are processed and hence available at class level.

on the total of economy. This setting should be read according to the instrumental optical in which T5 is inserted in the structure of a TSA: "*the main purpose of this table is to prepare and compile data on gross value added for various industries by transforming the national production account into a TSA production account*"[61].

Nevertheless, the complex structure of hierarchically ordered worksheets developed for the compilation of IT5, exposed at earlier stages in very few words, allows the calculation of indicators and ratios of specific interest, as well as the construction of graphs for an in-depth analysis of Italian tourism sector on the supply side.

Conclusions

This work, that is part of the implementation process of the first Italian TSA, has focused on a description of the methodology developed in order to measure the production accounts of Italian tourism industries and other industries, in a consistently way with Italian NA procedures.

From a methodological point of view the close link between the core system of National Accounts and that envisaged by the TSA, particularly T5, represents a mutual strength. On the one hand, the compilation of IT5 strictly depends on the detail of information provided by Italian SUT. On the other hand, the implementation of a satellite account represents a valuable opportunity to test the accuracy of core one, due to its demand of a thorough and detailed data.

The assessment process for the whole Italian tourism sector is based on data at 4 digits level of the international classification for economic activity. As far as the products, the lack of basic data has been overcome thanks to the great detail provided by Italian SUT for the year 2010, compiled for 106 branches of economic activities and 266 products (for internal use only). This articulation significantly impacted on the successful compilation of IT5.

The key findings highlighted in this work are twofold: on the one hand, the implementation of a methodology consistency with the national accounting procedures; on the other hand, innovation and development in working methods.

The first strictly derives from the particular perspective of a TSA. Both the *bottom-up* approach, applied for treatment of basic information, and the *top-down* approach that has characterized the treatment of SUT data, has provided compilers an important element for reflection: the needs of aggregation of NA must always ensure the binding of consistency with the information base, from which they derive.

The second is based on operational aspects. The gradual compilation of the production account of IT5 led to the building of a set of integrated worksheets, each of which designed for an automatic calculation of its output. Their modular structure allows to decouple the process of compiling for a particular reference year, which ranks as a tool for a time series analysis.

IT5 represents the first experience of compilation at national level. However, the province of Bolzano has compiled territorial TSA, and thus T5, since 2005. Despite some

[61] *TSA in the European Union*, Vol.3, par.3.4, pg.41.

differences in concepts and in contents[62], this sub-national T5 provided an "experienced" final result based on which evaluate the level of finesse of the construction process of IT5.

At international level, the long experience of Spain in compiling TSA, especially as far as net valuation of TO is concerned, has been repeatedly consulted[63].

Furthermore, the deep knowledge developed by Canada led us, during the compilation process of IT5, to enrich the reference material [64].

Summing up, it is now possible to state that this first compilation of IT5 doesn't make apparent a significant gap when compared with other compilations. However, some future improvements has emerged. Among these, the special treatment of goods and the retail trade services[65].

[62] The international standards used for 2005 and 2008 Bolzano's TSA was version 2001, both for IRTS and TSA:RMF. They proposed a different definition of tourism consumption and a different articulation of products/industries. Publication on TSA for Bolzano available http://www.provinz.bz.it/astat/it/256.asp?News_action=4&News_article_id=389255.

[63] National Statistics Institute - National Accounts General Department "Spanish Tourism Satellite Account: Methodological note", Madrid, 2004.

[64] Canadian Tourism Commission, "*Study of the Canadian Tourism Satellite Account (CTSA) – Comparison of TSA-RMF and CSTA"*, Ottawa, 2004.

[65] *TSA:RMF2008, Annex 4,* pgg 88-90. The processing of goods and retail trade services has been the object of a study following the compilation of the first T5.

References:

ASTAT [2009] *Tourism Satellite Account for Alto Adige*, Bolzano

Canadian Tourism Commission [2004] *Study of the Canadian Tourism Satellite Account (CTSA) – Comparison of TSA-RMF and CSTA*, Ottawa

Council and Parliament of the European Union [2011], Regulation (UE) N. 692/2011 of the European Parliament and of the Council of 6 July 2011 concerning *European statistics on tourism* and repealing Council Directive 95/57/EC, Strasbourg.

Eurostat [1995] *European System of Accounts*, Luxembourg.

Eurostat [2009] *Tourism Satellite Accounts in the European Union. Volume 1: Report on the implementation of TSA in 27 EU Member States* - Luxembourg, 2009.

Eurostat [2009], Tourism Satellite Accounts in the European Union. Volume 2: *Comparison of methodology and empirical results*, Luxembourg

Eurostat [2009], *Tourism Satellite Accounts in the European Union. Volume 3: Practical Guide for the Compilation of a TSA: Directory of Good Practices*, Luxembourg.

Eurostat [2009], Tourism Satellite Accounts in the European Union. Volume 4: *Possibilities to obtain more up-to-date TSA key figures*, Luxembourg

Eurostat [2012], *Methodological manual for tourism statistics*, v. 1.2 - Luxembourg.

ISTAT- [2009], *Classification of economic activities - Ateco 2007*, Rome.

National Statistics Institute – National Accounts General Department [2004] *Spanish Tourism Satellite Account: Methodological note*, Madrid

Statistics Canada [2002] *Canadian Tourism Satellite Account*, Ottawa.

United Nation [1993], *System of National Account*, New York.

United Nation [2008], *System of National Account*, New York

United Nation Statistics Division: *CPC Ver.2 Detailed structure and correspondences of CPC Ver.2 subclasses to ISIC Rev.4 and HS 2007*.

United Nations and World Tourism Organization [2007], *2008 International Recommendations for Tourism Statistics*, New York, Madrid.

United Nations, EUROSTAT, OECD, World Tourism Organization [2008], *2008 Tourism Satellite Account: Recommended Methodological Framework*, Brussels/Luxembourg, New York, Paris.

United Nations and World Tourism Organization [2010], *TSA data around the World*, Madrid.

L'integrazione dei risultati delle indagini sulla tecnologia e l'innovazione nelle imprese: una sperimentazione[1]

Tiziana Tuoto[2], Laura Corallo[3], Nicoletta Cibella[4], Daniela Ichim[5], Valeria Mastrostefano[6], Alessandra Nurra[7], Mariagrazia Rinaldi[8]

Sommario

L'articolo propone l'integrazione tra i micro-dati di due indagini sulle imprese: "Information and Communication Technologies" e "Indagine comunitaria sull'innovazione". Le complesse e multiple relazioni tra l'uso della tecnologia, i modelli di innovazione delle imprese e le performance economiche sono temi di crescente importanza nella letteratura empirica sull'innovazione. Tuttavia, la bidirezionaltà delle relazioni e i nessi di causalità possono essere colti appieno solo attraverso l'integrazione dei micro-dati. Nell'articolo sono illustrate tre strategie di integrazione sperimentate sui dati di indagine del 2008 e i relativi risultati. L'utilizzo di metodologie per il record linkage in questo contesto costituisce una sperimentazione interessante anche alla luce della comparazione con altri metodi di integrazione.

Parole chiave: integrazione di indagini campionarie, record linkage, innovazione.

Abstract

This article describes the linkage of microdata stemming from two business surveys: "Information and Communication Technologies" and "Community Innovation Survey". The complex and multiple relations between the use of technology, enterprise innovation patterns and economic performances are topics of increasing importance in the empirical

[1] Il presente documento è frutto dell'opera di tutti gli autori ed è stato curato da Tiziana Tuoto. In particolare, il paragrafo 2 è da attribuire a Tiziana Tuoto, il paragrafo 3.1 è da attribuire a Valeria Mastrostefano, il paragrafo 3.2 è da attribuire a Mariagrazia Rinaldi, il paragrafo 4.1 è da attribuire a Alessandra Nurra, il paragrafo 4.2 è da attribuire a Mariagrazia Rinaldi, il paragrafo 5 è da attribuire a Daniela Ichim, Valeria Mastrostefano e Alessandra Nurra, il paragrafo 6 è da attribuire a Tiziana Tuoto, il paragrafo 6.1 è da attribuire a Nicoletta Cibella, il paragrafo 6.2 è da attribuire a Laura Corallo e Tiziana Tuoto, il paragrafo 6.3 e il paragrafo 6.4 sono da attribuire a Laura Corallo e Daniela Ichim; il paragrafo 7 è da attribuire a Tiziana Tuoto, il paragrafo 8 è da attribuire a Nicoletta Cibella, il paragrafo 9 è da attribuire a Tiziana Tuoto. Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

[2] Ricercatore (Istat), e-mail: tuoto@istat.it.
[3] Collaboratore tecnico (Istat), e-mail: corallo@istat.it.
[4] Collaboratore tecnico (Istat), e-mail: cibella@istat.it.
[5] Ricercatore (Istat), e-mail: ichim@istat.it.
[6] Ricercatore (Istat), e-mail: mastrost@istat.it.
[7] Ricercatore (Istat), e-mail: nurra@istat.it.
[8] Ricercatore (Istat), e-mail: mrmarina@istat.it.

literature on innovation. However, bidirectional relationships and causality issues can be addressed in a comprehensive manner only by integrating microdata. In the present paper, three integrating strategies tested on 2008 survey and the related results are illustrated. The use of methodologies for record linkage in this context constitutes an innovative test, also in the light of the comparison with other integration methods.

Keywords: Information and Communication Technologies, Community Innovation Survey, record linkage.

1. Introduzione

Le complesse e molteplici relazioni tra l'uso di tecnologia di tipo informatico (IT), i modelli di innovazione e le performance economiche delle imprese sono temi di crescente importanza e interesse in gran parte della letteratura empirica sull'innovazione. Una serie di studi si sono concentrati sugli aspetti complementari di IT e di innovazione. Le imprese che innovano e investono anche in IT hanno maggiori vantaggi rispetto a quelle attive solo lungo una dimensione. Le tecnologie informatiche possono aumentare l'innovazione, accelerando la diffusione delle conoscenze, facilitando reti tra le imprese, riducendo le limitazioni geografiche e aumentando l'efficienza nella condivisione delle conoscenze. Pertanto, l'inclusione delle variabili IT in modelli di innovazione spiegano maggiormente le differenze nella propensione delle imprese ad innovare e le diverse modalità di innovazione (OCSE, 2010). D'altro lato, una parte della ricerca ha esplorato la relazione causale inversa tra innovazione e uso delle IT. Poiché l'innovazione è diventata più rivolta all'informazione, alla cooperazione e basata sulla rete, le imprese innovative sono gli utenti ad alta intensità di IT: la necessità di sfruttare le esternalità delle conoscenze di rete nei processi di innovazione spinge le imprese a investire di più in certi tipi di IT (van Leeuwen G, 2008). Inoltre, un vasto filone di letteratura conduce analisi congiunte delle variabili di IT e di innovazione per indagare meglio i contributi diretti e indiretti di innovazione e di tecnologia sulla produttività delle imprese.

Gli aspetti sopra descritti non si possono studiare facilmente analizzando in modo combinato i dati aggregati, a livello nazionale o per settore industriale, sulle tecnologie e sull'innovazione. I dati aggregati in effetti suppongono che le imprese siano le stesse e abbiano comportamenti uniformi all'interno di un paese e/o di un settore. La variabilità interna dei sistemi produttivi e delle industrie, l'eterogeneità e la varietà delle imprese possono essere rilevati solo guardando i dati a livello di impresa. Le relazioni bidirezionali tra tecnologia e innovazione e i nessi di causalità possono essere affrontati in modo completo solo integrando informazioni diverse a livello micro. Le analisi basate sui microdati possono effettivamente aiutare a valutare la diversità delle imprese e monitorare i loro diversi comportamenti all'interno di un settore, per indagare se le performance delle imprese sono simili o diverse tra le industrie, all'interno di uno stesso gruppo industriale o tra le imprese di determinate dimensioni. Inoltre, i microdati permettono di valutare l'importanza relativa delle varie caratteristiche di innovazione e di tecnologia e la loro interazione nelle diverse aziende.

A questo proposito, la ricerca che si incentra sull'ipotesi di IT come fattore chiave di innovazione ha recentemente fornito evidenze empiriche a sostegno di questa idea,

combinando a livello di impresa i dati delle indagini Istat "Information and Communication Technologies" (ICT) and Community Innovation Survey (CIS) (Oecd, 2010; Eurostat, 2008) . Ma questi esercizi di abbinamento dei dati delle due indagini soffrono di alcuni inconvenienti. Una delle questioni principali non affrontate è il problema della selettività dovuta al campionamento iniziale delle piccole imprese. Inoltre, il coordinamento negativo di campioni necessari per controllare l'onere statistico sulle imprese, diminuisce la rappresentatività dei dati comuni alle due indagini ICT-CIS. Le analisi eseguite su aziende che avevano risposto contemporaneamente ad entrambe le indagini è sicuramente sbilanciata verso le unità più grandi.

L'integrazione di micro dati derivanti da due indagini che non sono state progettati per questa integrazione, la dimensione ridotta delle unità congiunte e la successiva forte distorsione di selezione richiede un trattamento statistico specifico per permettere un migliore utilizzo dei microdati. Partendo dalla ricerca in corso nel campo delle analisi sopra descritte basate su dati micro, questo lavoro riporta lo studio delle metodologie più opportune per l'integrazione dei micro-dati delle indagini ICT e CIS al fine di costruire l'insieme di dati più ampio possibile per l'analisi delle relazioni tra le principali variabili delle due indagini relative alla dotazione e utilizzo delle nuove tecnologie, modalità di innovazione, impatto sulle performance di impresa. Nel lavoro sono illustrati i risultati conseguiti relativamente ai dati delle indagini che hanno come anno di riferimento il 2008, in quanto in tale occasione la selezione campionaria delle imprese da coinvolgere è stata effettuata senza adottare il coordinamento negativo. L'utilizzo di metodologie per l'integrazione di micro-dati in questo contesto costituisce una sperimentazione: infatti, nel caso di indagini campionarie basate su coordinamento negativo dei campioni è pratica comune ricorrere a integrazioni a livello macro, e, qualora fosse desiderabile costruire il dataset completo, ci si avvale di metodologie che rientrano nella categoria dei metodi di imputazione. L'uso quindi di metodologie di record linkage in questo particolare ambito rappresenta una sperimentazione interessante anche alla luce della comparazione con altri metodi usualmente adottati in tali contesti.

Nell'articolo vengono analizzate e sperimentate metodologie di integrazione di micro-dati con l'obiettivo di proporre soluzioni replicabili nelle successive occasioni di indagine, secondo tre diversi scenari: il primo scenario tiene conto della situazione reale dei dati del 2008 caratterizzati dal mancato coordinamento dei campioni; il secondo scenario è compatibile con situazioni più generali, sempre con riferimento al contesto usuale del record linkage. Il terzo scenario invece è finalizzato alla ricostruzione del dataset integrato completo, e con riferimento all'utilizzo di metodi di record linkage rappresenta una frontiera, poiché nella situazione in esame è noto a priori che non tutte le unità sono riferite alle stesse imprese.

Il documento è organizzato come segue: il paragrafo 2 descrive gli obiettivi dell'integrazione e gli scenari proposti come output; il paragrafo 3 riassume le principali caratteristiche dell'indagine CIS - Community Innovation Survey e il successivo paragrafo 4 riporta le peculiarità dell'indagine ICT- Information and Communication Technology Survey. Il paragrafo 5 delinea le relazioni fondamentali tra le grandezze rilevate separatamente nelle due indagini, rappresentando quindi il benchmark per le analisi successive sui dati integrati. Nel paragrafo 6 sono riportati i metodi e gli strumenti di integrazione adottati, le caratteristiche salienti delle metodologie di record linkage probabilistico, lo strumento RELAIS che è servito ad implementarle, i modelli che hanno

prodotto i risultati migliori rispetto agli scenari definiti nel paragrafo 2. Il successivo paragrafo 7 riassume la situazione informativa prodotta e fornisce una valutazione della qualità dei diversi risultati considerati. Il paragrafo 8 illustra una metodologia alternativa, legata alla classificazione ad albero, per conseguire risultati utili negli scenari considerati. Infine il paragrafo 9 riporta le principali conclusioni e delinea alcuni necessari sviluppi futuri, al fine di utilizzare nella pratica i metodi qui proposti e sperimentati.

2. Gli obiettivi dell'integrazione

Le procedure di integrazione dei microdati delle indagini ICT e CIS per l'anno 2008 sono state progettate e realizzate con l'ottica di soddisfare tre diverse esigenze.

La prima esigenza è quella di ricreare, attraverso strumenti probabilistici, le condizioni particolarmente favorevoli legate all'alta sovrapposizione dei campioni verificatasi nel 2008, anche in occasioni standard in cui venga applicato il coordinamento negativo tra i campioni. Questo obiettivo viene perseguito attraverso lo studio e la sperimentazione di modelli di abbinamento probabilistico che garantiscano al massimo l'identificazione delle unità comuni alle due indagini nell'edizione 2008, che si abbinano deterministicamente per identificativo univoco (codice impresa, nel seguito) dell'Archivio Statistico delle Imprese Attive (ASIA, nel seguito). Il risultato di questo studio sarà una procedura di abbinamento probabilistico che permetterà di costruire, anche in occasioni di indagine successive e meno favorevoli, un dataset integrato di microdati analogo a quello dell'edizione 2008. Di conseguenza, il file di microdati, costruito secondo il modello individuato in base ai criteri dettati da questo primo obiettivo, non sarà numericamente molto più grande di quello ottenuto nel 2008 attraverso l'aggancio deterministico per codice impresa, in quanto il vantaggio di utilizzare il modello selezionato da questa strategia è evidente soprattutto in vista di future occasioni di indagini in cui il coordinamento negativo dei campioni ridurrà fortemente la sovrapposizione deterministica realizzata nel 2008 e con un aggancio per codice impresa, quindi, si abbineranno un numero contenuto di unità. Nel seguito dell'articolo, l'espressione **prima strategia** di linkage sarà quella volta al conseguimento di questo obiettivo.

La seconda strategia di integrazione invece si propone di incrementare la base di dati del 2008, cercando di abbinare il maggior numero possibile di unità rispetto a quelle che si agganciano secondo il codice impresa. In questo caso, lo strumento probabilistico servirà a riconoscere le stesse unità che pure non presentano identico codice impresa, evento che può verificarsi per numerosi motivi, legati alla diversa tempistica di estrazione dei campioni e conseguente diverso aggiornamento delle liste di selezione, ai fenomeni di demografia di impresa che intervengono in tali lag temporali, e così via. I modelli di abbinamento probabilistico studiati e sperimentati in questo scenario, usati congiuntamente all'aggancio deterministico, sono volti ad incrementare il più possibile la base di dati per le successive analisi preservando allo stesso tempo un valore elevato delle probabilità di corretto abbinamento, così da garantire la qualità in termini di accuratezza del risultato conseguito. Nel seguito del documento, l'espressione **seconda strategia** di linkage denoterà quella volta al conseguimento di questo obiettivo.

Infine, la terza strategia di integrazione è volta alla costruzione del file di microdati completo per tutte le unità rispondenti alle due indagini. In questo caso, le informazioni

congiunte sulle indagini CIS e ICT saranno riferite esattamente alla stessa unità per il sottoinsieme di unità per cui il codice impresa coincide, mentre per le restanti unità il legame sarà di tipo probabilistico. Quindi, è noto per costruzione che la coppia individuata non rappresenta di fatto la stessa entità, ma le metodologie di record linkage verranno impiegate per riconoscere unità che siano il più possibile "simili"; la procedura inoltre permetterà di misurare questa similitudine attraverso la probabilità di corretto abbinamento. In questo terzo scenario l'uso di tecniche di record linkage rappresenta ad oggi una sperimentazione, interessante soprattutto alla luce della comparazione con altri metodi di integrazione dei dati propriamente detti e con metodi di imputazione in generale, sia per la validazione delle ipotesi sottostanti i vari metodi sia per quanto riguarda la robustezza delle stime basate sul dataset integrato risultante. Nel paragrafo successivo verranno illustrati appunto alcuni di questi aspetti. Nel seguito del documento, l'espressione **terza strategia** di linkage sarà quella volta al conseguimento di questo obiettivo.

2.1 Le metodologie per l'integrazione: record linkage e statistical matching

L'integrazione dei risultati delle indagini CIS e ICT in questo lavoro ha lo scopo di rendere possibili tutta una serie di analisi statistiche che non possono essere condotte sfruttando in maniera disgiunta i dati di ciascuna indagine. Come evidenziato nel paragrafo precedente in relazione alla terza strategia, il problema dell'integrazione delle due indagini CIS e ICT è solo in parte risolvibile tramite le procedure di record linkage, ossia procedure il cui scopo sia quello di individuare i record afferenti alla stessa unità provenienti da due data set diversi. Infatti alcune imprese sono solo disponibili nell'indagine ICT (e non nella CIS) e viceversa. L'uso delle procedure di record linkage per riconoscere unità il più possibile "simili" è stato proposto in alcuni lavori (ad esempio Okner, 1972) e subito contestati da alcuni commentatori (Sims, 1972). In effetti, in questo contesto, l'applicazione di tecniche di record linkage è giustificato sotto l'ipotesi di indipendenza condizionata, ossia quando le variabili osservate solo su CIS ma non su ICT e le variabili osservate solo su ICT ma non su CIS sono indipendenti condizionatamente alle variabili in comune fra ICT e CIS usate come variabili di matching. Questa ipotesi non è verificabile, poiché non esiste un insieme di dati completo su cui accertare la relazione tra i gruppi di variabili.

D'altro canto non sono parimente verificabili altre assunzioni che potrebbero suggerire l'applicazione di diverse metodologie di integrazione che vanno sotto il nome di statistical matching o data fusion. Questi metodi si propongono l'analisi congiunta di due o più variabili osservate distintamente in campioni estratti dalla stessa popolazione di riferimento ma contenenti unità diverse. In particolare, il problema in esame permetterebbe di applicare metodi di statistical matching che rimuovono l'ipotesi di indipendenza condizionata sfruttando informazione ausiliaria. Infatti, mancando il coordinamento negativo dei campioni, la sovrapposizione delle unità comuni rispondenti alle due indagini è straordinariamente alta, come verrà specificato nel paragrafo 5. In questo caso, essendo note e disponibili in entrambe le indagini le variabili usate nei disegni campionari, sotto l'ipotesi che la relazione tra le variabili osservate solo in ICT e quelle osservate solo in CIS sia stimabile correttamente nel sotto-campione di unità comuni alle due indagini, è possibile rafforzare le procedure di statistical matching utilizzando metodi che sfruttano l'informazione ausiliaria (Paass 1986, Singh et al. 1993, Rassler 2002, Moriarity and Scheuren 2003). Con questo approccio però, l'ipotesi non verificabile di indipendenza condizionata viene sostituita dall'ipotesi, altrettanto non verificabile, che le relazione tra le

variabili dell'indagine CIS e quelle dell'indagine ICT osservate nel sotto-insieme di unità comuni siano valide anche per tutte le unità su cui le variabili non sono osservabili congiuntamente. Questa assunzione è comunque molto forte dato che, come verrà messo in luce nel paragrafo 5, la sovrapposizione dei campioni riguarda per la gran parte unità di grandi dimensioni in termini di addetti ed andrebbe in primis verificato che le relazioni tra variabili osservate per unità con queste caratteristiche sono valide anche per imprese con caratteristiche profondamente diverse.

Infine, a parità di difficoltà nella validazione delle ipotesi alla base delle diverse metodologie, si è rinunciato al ricorso a tecniche di statistical matching per non dover definire e quindi limitare dal principio l'insieme delle variabili su cui effettuare le successive analisi, solo alcune delle quali sono accennate nel successivo paragrafo 5. In ogni caso, approfondimenti sull'applicabilità di metodi di statistical matching, in particolare basandosi su approcci che si riconducono all'analisi dell'incertezza (D'Orazio et al 2006a e D'Orazio et al 2006b).

3. L'indagine CIS - Community Innovation Survey

3.1 Principali caratteristiche

La rilevazione CIS, sviluppata congiuntamente da Eurostat e dagli Istituti statistici dei Paesi Ue, è finalizzata a raccogliere informazioni sui processi di innovazione delle imprese europee. In particolare, fornisce un set di indicatori volti ad analizzare le strategie, i comportamenti e le performance innovative delle imprese, i fattori di ostacolo e di supporto all'innovazione e le complesse interazioni sistemiche che si attivano tra gli attori del processo innovativo. Raccoglie, infine, una serie di informazioni di carattere generale sull'appartenenza a gruppi di imprese e sul fatturato delle imprese, oltre a fornire informazioni di carattere strutturale, come l'attività economica prevalente, il numero di addetti e la regione di residenza.

A partire dal 2004, la rilevazione viene svolta con cadenza biennale ed è inserita in un quadro normativo europeo (Regolamento Ce n. 1450/2004) che ne stabilisce l'obbligatorietà per gli stati membri. L'adozione di criteri definitori e metodologie di rilevazione comuni a tutti i paesi europei (che riprendono quelli stabiliti dall'Ocse nel Manuale di Oslo) garantisce nel complesso un buon livello di comparabilità internazionale dei dati sull'innovazione.

Il periodo di riferimento dell'indagine considerata in questo lavoro è il triennio 2006-2008. La popolazione oggetto della rilevazione è costituita da 208637 imprese con almeno 10 addetti medi annui, attive nei settori dell'industria, costruzioni e servizi (2008). La rilevazione è campionaria per le imprese da 10 a 249 addetti e censuaria per quelle con almeno 250 addetti. Il disegno di campionamento è ad uno stadio stratificato con selezione delle unità a uguale probabilità. La popolazione è stata suddivisa in strati (ossia, sottoinsiemi tra loro non sovrapposti definiti sulla base di alcune caratteristiche strutturali delle unità statistiche e all'interno dei quali le unità sono fra loro omogenee riguardo alle variabili oggetto di studio). Gli strati sono definiti dalla concatenazione delle modalità identificative dei settori di attività economica (divisione Nace Rev.1.1), delle classi di addetti (10-49 addetti, 50-249 addetti, 250 addetti e oltre) e delle regioni di localizzazione

delle imprese (livello 2 della classificazione europea Nuts, disciplinata dal Regolamento Ce n.1059/2003). La raccolta dati è avvenuta principalmente tramite l'auto-compilazione di un questionario elettronico attraverso l'accesso personalizzato al sito web dell'Istat dedicato all'indagine: https://indata.istat.it. I risultati della CIS2008 si basano su 19688 risposte validate, pari al 52.1 per cento del campione teorico.

3.2 La strategia campionaria

La popolazione obiettivo della rilevazione è costituita dalle imprese con almeno 10 addetti medi[9] operanti, nel periodo di riferimento dell'indagine, nei seguenti settori di attività della Classificazione Nace Rev. 1.1: sezioni da B a J (esclusa la divisione 60), K, L e divisioni 71, 72 e 77. La popolazione utilizzata per la selezione delle unità campionarie comprende le 194825 unità appartenenti al campo di osservazione secondo le informazioni desunte dall'archivio ASIA con anno di riferimento 2006.

Il piano di campionamento utilizzato è casuale stratificato ad uno stadio, con selezione delle unità senza reimmissione e con probabilità costante all'interno di ciascuno strato. La stratificazione adottata, corrispondente alla partizione minima della popolazione che consente di ottenere i domini di stima pianificati, è stata ottenuta concatenando le modalità delle seguenti variabili:

- 56 settori di attività economica (sezioni *F* ed *I* e divisione Nace Rev.1.1 per le altre attività);
- 3 classi di addetti medi: 10-49, 50-249, 250 e oltre;
- 19 regioni amministrative e le 2 province autonome del Trentino Alto Adige.

Il numero teorico degli strati così costruiti è risultato pari a 3528, di cui 2363 contenenti almeno un'unità della popolazione da cui è stato selezionato il campione. Si è stabilito a priori di censire gli strati contenenti le imprese con almeno 250 addetti medi; il calcolo dell'allocazione è stato eseguito in modo tale da assicurare simultaneamente, per ciascuno dei domini di stima pianificati, predefiniti livelli di accuratezza della stima delle variabili: numero di addetti, fatturato e spesa totale per innovazione (cfr. tavola 3.1), compatibilmente con l'indicazione della responsabile di indagine di selezionare un campione teorico di numerosità prossima a 45000 unità. I domini di stima pianificati sono i seguenti:

- attività economica;
- attività economica × classe di addetti;
- appartenenza o no a settori *core*[10] × classe di addetti × regione amministrativa.

Il primo tipo di dominio corrisponde alla divisione Nace Rev. 1.1, ad eccezione delle imprese delle Sezioni F ed I per le quali il dettaglio è costituito dalla sezione; il secondo è definito dalla combinazione delle modalità delle variabili: Sezione Nace e classe di addetti

[9] Per coerenza con altre indagini strutturali sulle imprese industriali e dei servizi, è stata adottata la convenzione di includere nel campo di osservazione tutte le imprese con almeno 9,5 addetti medi nell'anno di riferimento.

[10] Settori Core (Nace Rev. 1.1): B, C, D, E, H, K, 46, 58, 61, 62, 63, 71.

medi (10-49, 50-249, 250 e oltre11); il terzo è definito dal concatenamento tra l'appartenenza o no alle c.d. attività *core*, la regione/provincia autonoma di residenza e la classe di addetti (10-249, 250 e oltre).

Il problema di allocazione multivariata e multidominio è stato risolto secondo la metodologia correntemente utilizzata nelle rilevazioni Istat, la quale fa riferimento ad un approccio basato sull'algoritmo proposto da Bethel (1989). La stima delle medie e varianze di strato delle variabili di interesse, necessaria per impostare il calcolo dell'allocazione ottima, è stata effettuata mediante i dati disponibili dall'edizione precedente dell'indagine (CIS 2002-2004) e l'allocazione ha infine dato luogo ad una numerosità campionaria totale di 44780 unità.

Tavola 3.1 - Errori relativi percentuali pianificati nel calcolo dell'allocazione

DOMINIO	Errori relativi percentuali pianificati		
	Numero di addetti medi	Fatturato	Spesa totale per innovazione
DOM1	0,02	0,03	0,05
DOM2	0,02	0,03	0,06
DOM3	0,02	0,04	0,07

Dopo aver determinato l'allocazione, si è utilizzata una procedura di selezione coordinata finalizzata a minimizzare la sovrapposizione tra campioni provenienti dallo stesso archivio di estrazione e relativi ad indagini differenti.

Il dataset dei rispondenti è costituito dalle 19688 imprese – pari al 52.1 per cento del campione teorico – che hanno restituito questionari validi e che, secondo le informazioni disponibili dall'archivio, esercitano attività economiche comprese nel campo di osservazione dell'indagine. L'universo di riporto, desunto da ASIA 2008 – e quindi allineato con il periodo di riferimento dell'indagine – è costituito da 208636 imprese. Il calcolo dei pesi finali è stato effettuato secondo la teoria dello stimatore di calibrazione (Deville e Särndal, 1992), in modo da garantire la convergenza delle stime delle variabili ausiliarie (numero di imprese e numero di addetti medi) ai corrispondenti totali noti, a livello dei seguenti domini (c.d. domini di calibrazione):

- divisione Nace per tutti i settori tranne quelli delle costruzioni (sezione F) e dei servizi di alloggio e ristorazione (sezione I), in cui le stime sono calcolabili per sezione;
- classe di addetti (10-49; 50-249; 250 e oltre) × sezione ;
- appartenenza o no a settori *core*[12] × regione/provincia autonoma;

[11] Le classi di addetti sono state definite adottando la stessa convenzione dell'indagine PMI, cioè –ad esempio– includendo nella classe 10-49 tutte le imprese con un numero di addetti compresto tra 9.5 (incluso) e 49.5 (escluso).

[12] Cfr. nota precedente per la definizione delle attività definite "core" per l'indagine CIS.

- macrosettore di attività economica[13] × *core* × classe di addetti (10-49; 50-99; 100-249; 250 e oltre) .

Per il calcolo dello stimatore di calibrazione è stata utilizzata una funzione di distanza logaritmica.

4. L'indagine ICT- Information and Communication Technology Survey

4.1 Principali caratteristiche

L'indagine ICT viene svolta dal 2001 con cadenza annuale e dal 2004 applica criteri definitori e metodologie di rilevazione comuni a tutti i Paesi dell'Ue sulla base di Regolamenti Comunitari che hanno definito il quadro di riferimento delle statistiche sulla società dell'informazione (Reg. Ce 808/2004 e 1006/2009) . Ogni anno regolamenti attuativi specificano gli indicatori e i focus che devono essere sviluppati dal questionario comunitario. Gli indicatori sono discussi a livello europeo da specifici gruppi di lavoro che partecipano alla definizione del frame work concettuale per la raccolta di indicatori statistici che rientrano nella Agenda digitale europea (eEurope, i2010 e 2011-2015 benchmarking).

L'obiettivo di analisi della rilevazione è quindi quello di misurare l'adozione e l'utilizzo di tecnologie dell'informazione e comunicazione nelle imprese definendone l'impatto sull'organizzazione interna e nei rapporti con l'esterno (grado di informatizzazione dei processi di acquisto e vendita, integrazione e condivisione delle informazioni con clienti, fornitori, banche, Pubblica Amministrazione, lavoratori, funzioni aziendali). I principali fenomeni osservati sono l'adozione di Internet, la tipologia di connessione utilizzata (banda larga o stretta), servizi offerti sul sito web, reti interne (intranet) ed esterne (extranet), livello di interazione on-line con la P.A., scambio automatico ed elettronico di informazioni tra sistemi informativi, condivisione elettronica di informazioni all'interno dell'impresa, utilizzo del commercio elettronico).

Il periodo di riferimento dell'indagine è gennaio 2009 per le variabili di tipo qualitativo mentre i dati economici (acquisti, ricavi, media addetti e commercio elettronico) si riferiscono all'anno precedente.

La rilevazione si svolge tra marzo e agosto dello stesso anno di riferimento dei dati e prevede due solleciti alle imprese che non risultano rispondenti nel corso dell'indagine. Una serie di documentazione a supporto della risposta viene messa a disposizione on-line insieme ad un questionario web che viene auto compilato dalle imprese sfruttando la possibilità di salvare parzialmente i dati e di modificarli fino all'invio definitivo possibile solo dopo aver superato una serie di controlli di coerenza previsti per alcuni quesiti fondamentali.

La popolazione di riferimento dell'indagine è rappresentata dalle imprese con almeno 10 addetti attive nei settori manifatturiero, energia, costruzioni e servizi non finanziari.

[13] Industria in senso stretto, costruzioni e servizi.

4.2 La strategia campionaria

La popolazione utilizzata per la selezione delle unità campionarie comprende circa 210000 unità appartenenti al campo di osservazione secondo le informazioni desunte dall'archivio ASIA 2007. Il piano di campionamento utilizzato è casuale stratificato ad uno stadio, con selezione delle unità senza reimmissione, con probabilità costante all'interno di ciascuno strato. La stratificazione adottata, corrispondente alla partizione minima della popolazione che consente di ottenere i domini di stima pianificati, è stata ottenuta concatenando le modalità delle seguenti variabili:

- 31 settori di attività economica;
- 4 classi di addetti medi: 10-49, 50-99, 100-249, 250 e oltre;
- 19 regioni amministrative e le 2 province autonome del Trentino Alto Adige.

Il numero teorico degli strati così costruiti è risultato pari a 2604, di cui 2134 contenenti almeno un'unità della popolazione da cui è stato selezionato il campione. Si è stabilito a priori di censire gli strati contenenti le imprese con almeno 250 addetti medi ed il calcolo dell'allocazione è stato eseguito in modo tale da assicurare simultaneamente, per ciascuno dei domini di stima pianificati, predefiniti livelli di accuratezza della stima delle variabili: numero di addetti, fatturato e acquisto di beni e servizi, compatibilmente con l'indicazione della responsabile di indagine di contenere la numerosità campionaria entro le 40000 unità. I domini di stima, pianificati in modo da soddisfare le richieste derivanti dai regolamenti europei e le esigenze di pubblicazione Istat, sono i seguenti:

- attività economica, secondo un'articolazione in 31 settori[14]
- macrosettore di attività economica (manifattura, costruzioni e servizi) ×
- × classe di addetti medi (10-49, 50-99, 100-249, 250 o più);
- macrosettore di attività economica × regione amministrativa o provincia autonoma;
- un ulteriore dominio articolato nelle seguenti tre modalità:
- imprese con 250 o più addetti medi;
- imprese con meno di 250 addetti medi, residenti nelle regioni: Campania, Puglia, Basilicata, Calabria e Sicilia – c.d. *obiettivo 1*;
- imprese con meno di 250 addetti medi, residenti nelle regioni al di fuori del c.d. *obiettivo 1*.

Il problema di allocazione multivariata e multidominio è stato risolto secondo la metodologia usuale nelle rilevazioni Istat (Bethel, 1989). La stima delle medie e varianze di strato delle variabili di interesse, necessaria per impostare il calcolo dell'allocazione ottima, è stata effettuata mediante i dati disponibili dall'edizione precedente dell'indagine (ICT 2007-2008) e l'allocazione ha infine dato luogo ad una numerosità campionaria totale di circa 37500 imprese.

Dopo aver determinato l'allocazione, si è utilizzata una procedura di selezione coordinata delle unità analoga a quella già descritta nel paragrafo 3.2.

Il dataset dei rispondenti utilizzato per la stima finale delle variabili di interesse è costituito da 19781 imprese che hanno restituito questionari validi e che secondo le informazioni disponibili nell'archivio ASIA 2007 appartengono al campo di osservazione

[14] L'articolazione in 31 settori è la stessa utilizzata per la stratificazione

dell'indagine, che consta di circa 220000 imprese. Il calcolo dei pesi finali è stato effettuato secondo la teoria dello stimatore di calibrazione, in modo da garantire la convergenza delle stime delle variabili ausiliarie (numero di imprese e numero di addetti medi) ai corrispondenti totali noti calcolati dall'universo di riporto nei domini di stima suindicati.

5. I legami tra le variabili CIS e ICT

Scopo di questa analisi preliminare, condotta sui dati ottenuti dal matching esatto dei rispondenti alle due indagini, è l'individuazione di relazioni tra le variabili CIS e ICT utili a definire i criteri di correzione necessari per l'imputazione dei missing contenuti nel dataset finale di dati integrati CIS-ICT ottenuto dal record-linkage. I rispondenti comuni ad entrambe le indagini sono pari a 9882 imprese e corrispondono circa alla metà delle unità presenti nei campioni finali ICT e CIS. A livello settoriale, non si riscontrano marcate divergenze in termini di incidenza dei rispondenti comuni sul totale dei rispondenti a ciascuna indagine, ad eccezione dei servizi finanziari dove l'adozione di un differente disegno di campionamento (censuario per l'indagine ICT e campionario per la CIS) ha determinato differenze molto pronunciate (44% in ICT e 93% in CIS). A livello dimensionale, le grandi imprese (quelle con almeno 250 addetti), come previsto (sono censite in entrambe le indagini), sono le più rappresentate nel dataset integrato: infatti, coprono oltre l'80% delle grandi imprese presenti nei campioni finali di CIS e ICT. Infine, confrontando le distribuzioni settoriali e dimensionali dei rispondenti congiunti con quelle dei rispondenti alle singole indagini, non si evidenziano bias significativi né a settoriale né a livello dimensionale, anche se le imprese industriali e quelle di dimensione medio-grande (cioè, con almeno 50 addetti) contribuiscono maggiormente a determinare il totale dei rispondenti congiunti.

Tavola 5.1 – Imprese rispondenti alle indagini ICT e CIS per macro-settore e classe di addetti. Anno 2008

INDAGINE ICT	Industria	Costruzioni	Servizi finanziari	Altri servizi	10-49	50-249	250+	Totale
Rispondenti finali	6163	5229	1661	6728	14344	3568	1869	19781
Imprese presenti in CIS	3725	2412	729	3016	6357	1988	1537	9882
%rispondenti congiunti (sul totale ICT)	60.44	46.13	43.89	44.83	44.32	55.72	82.24	49.96
Composizione % dei rispondenti congiunti	37.69	24.41	7.38	30.52	64.33	20.12	15.55	100.00
Composizione % dei rispondenti totali	31.16	26.43	8.40	34.01	72.51	18.04	9.45	100.00
Indagine CIS								
Rispondenti finali	7156	4378	804	7350	14430	3484	1774	19688
Imprese presenti in ICT	3734	2389	729	3030	6394	1963	1525	9882
%rispondenti congiunti (sul totale CIS)	52.18	54.57	90.67	41.22	44.31	56.34	85.96	50.19
Composizione % dei rispondenti congiunti	37.79	24.18	7.38	30.66	64.70	19.86	15.43	100.00
Composizione % dei rispondenti totali	36.35	22.24	4.08	37.33	73.29	17.70	9.01	100.00

Questa prima analisi sulle relazioni tra le variabili CIS e ICT è stata condotta a partire da un set limitato di variabili CIS e ICT selezionato sulla base della capacità esplicativa delle variabili rispetto ai principali fenomeni dell'ICT e dell'innovazione e della loro stabilità nel tempo in modo da consentire comparazioni temporali. Tra gli indicatori ICT sono stati scelti i principali indicatori chiave di benchmarking di interesse per la CE come l'uso di extranet e Intranet (e_extra, e_intra), l'interazione con PA (e_igov2_medhig), la connessione mobile a Internet (e_mob), la vendita on-line e gli acquisti (e_ecomm); inoltre per l'integrazione dei dati sono state incluse variabili che esprimono relazioni significative tra ICT e CIS e che, in futuro, potrebbero essere utilizzate per l'imputazione di variabili non direttamente osservabili come l'uso di software per effettuare analisi delle informazioni raccolte sui clienti a fini di marketing (e_crm_b) e l'utilizzo di un pacchetto software per condividere informazioni sulle vendite e / o acquisti con altri aree funzionali interne (e_erp). Infine, esperienze simili di analisi dei dati effettuate dall'Istat con l'OECD suggeriscono che l'uso delle ICT espresso in termini di servizi Web offerti dalle imprese che utilizzano una o più pagine su Internet (e_webf1_medhig) e i collegamenti automatici tra sistemi informativi (e_internal1) e tra sistemi informativi dell'impresa rispondente con altri soggetti esterni (e_ade_ent) sono importanti dimensioni da considerare come fattori drivers di adozione di innovazione da parte delle imprese.

Le variabili CIS, scelte sulla base del potere esplicativo che queste hanno nelle analisi sui processi innovativi delle imprese evidenziato da molta letteratura empirica sull'argomento, possono essere raggruppate in quattro macro-categorie: gli indicatori di input che comprendono gli investimenti materiali in macchinari e attrezzature (RMAC), le spese per R&S (RED), le attività creative meno formalizzate quali il design (RDES) e altre attività immateriali come i brevetti (KNOW); gli indicatori di output innovativo che distinguono le innovazioni di prodotto (INPDT) e processo (INPCS) e le innovazioni organizzative (ORG) e di marketing (MKT). Un terzo indicatore è dato dalla cooperazione per l'innovazione con altre imprese, fornitori, clienti, università o centri di ricerca (COOP), una misura dell'apertura verso l'esterno delle imprese innovatrici. L'ultimo gruppo di indicatori comprende due variabili strutturali non strettamente legate ai processi di innovazione, ma rilevate dall'indagine CIS: l'appartenenza dell'impresa ad un gruppo industriale (GP) e la sua presenza sui mercati esteri (MARFOR). Un primo ed evidente risultato del confronto tra variabili CIS e ICT(Tavola 5.2) è la maggiore propensione all'uso di tecnologie ICT da parte delle imprese innovative (identificate sulla base di un indicatore sintetico della propensione ad innovare).

Tavola 5.2 – Imprese che utilizzano ICT per dimensione innovativa

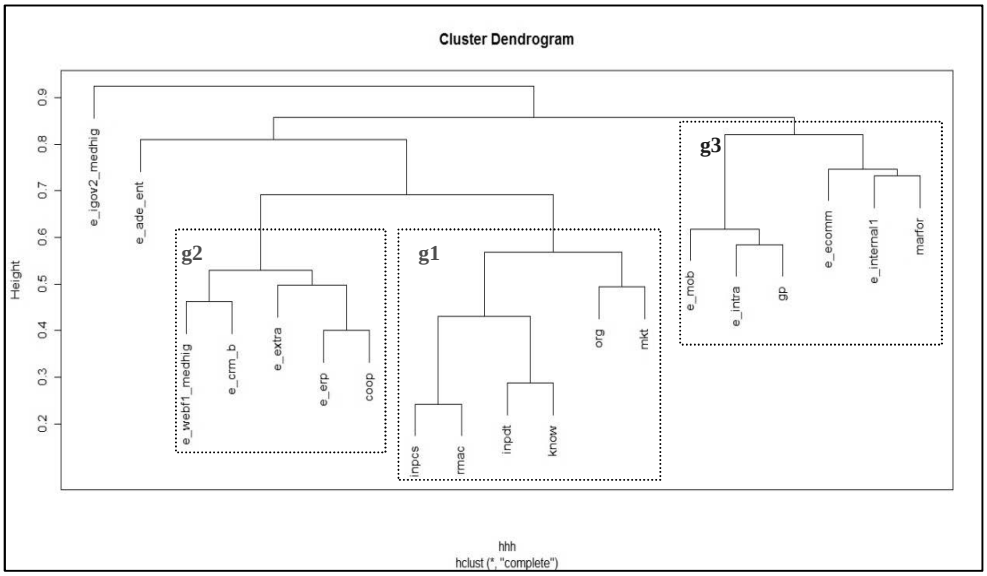
		Totale imprese			% di utilizzatori ICT per		
		Non inno	Inno	Totale	Totale	*di cui:*	
Variabili ICT: l'impresa utilizza						*non inno*	*inno*
e_intra	rete intranet	1082	2549	3631	36,7	29,8	70,2
e_extra	rete extranet	688	1853	2541	25,7	27,1	72,9
e_mob	connessione Internet mobile	952	2278	3230	32,7	29,5	70,5
e_igov2_medhig	servizi on-line offerti dalla PA almeno per rinviare moduli compilati	1713	2389	4102	41,5	41,8	58,2
e_webf1_medhig	un sito web attravero cui offre da 2 a 5 servizi non esclusivamente informativi	472	1246	1718	17,4	27,5	72,5
e_ade_ent	scambio automatico di dati tra sistemi informativi interni e quelli di altri all'esterno	1444	2572	4016	40,6	36,0	64,0
e_internal1	condivisione elettronica di informazioni con almneo due funzioni aziendali	1458	3043	4501	45,5	32,4	67,6
e_erp	software ERP per condividere info su ordini di acquisto/vendita con altre funzioni aziendali	436	1649	2085	21,1	20,9	79,1
e_crm_b	software CRM per analizzare dati sulla clientela per finalità di marketing	508	1515	2023	20,5	25,1	74,9
e_ecomm [(a)]	commercio elettronico (in vendita o acquisto)	1259	2389	3648	36,9	34,5	65,5
	(a) variabile non osservata per il settore K di intermediazione finanziaria						

Per individuare eventuali relazioni tra variabili ICT e CIS, partendo dal dataset di rispondenti comuni è stata condotta un'analisi cluster con metodo di classificazione gerarchico. I risultati, illustrati dal dendrogramma riportato nella Figura 5.1, evidenziano tre gruppi omogenei di variabili, due dei quali connettono alcuni dei fenomeni misurati nelle due differenti indagini:

- il primo gruppo interessa solo le variabili CIS e conferma le forti relazioni esistenti tra specifici input e output di innovazione e l'interdipendenza tra innovazione tecnologica e non;
- il secondo gruppo include variabili che esprimono un utilizzo di tecnologie più finalizzato e funzionale alle attività individuando imprese che utilizzano software specifici di condivisione di informazioni (ERP, CRM), che offrono servizi non solo informativi sul proprio sito web, utilizzano reti esterne per scambiare informazioni con altri attori della filiera produttiva e che allo stesso tempo sono più propense a sviluppare forme di cooperazione per l'innovazione;
- il terzo gruppo evidenzia il legame di variabili ICT maggiormente legate alla complessità organizzativa/dimensionale che spinge verso l'adozione di connessioni mobili, l'utilizzo di reti interne, il commercio elettronico, con variabili quali l'appartenenza ad un gruppo e la presenza sui mercati esteri; in questo caso il rapporto tra variabili CIS e le altre variabili ICT sembra essere meno forte e più legato a variabili strutturali che da un lato aumentano il bisogno di comunicare tra le diverse sedi del gruppo tramite una rete interna e esprimono la necessità di lavorare connessi e in mobilità e dall'altro amplificano le opportunità offerte dal commercio elettronico attraverso la presenza sui mercati esteri e lo scambio oltre i confini dei mercati rilevanti.

La vicinanza dei gruppi 1 e 2 suggerisce l'esistenza di alcune relazioni tra un utilizzo di ICT con forte impatto sulle relazioni organizzative tra imprese e clienti/fornitori e l'introduzione di innovazioni di carattere organizzativo e di marketing.

Figura 5.1 - Dendrogramma cluster di variabili ICT e CIS



6. Il record linkage probabilistico

Come noto, il record linkage (o abbinamento esatto) indica un processo di abbinamento di record che ha come obiettivo l'identificazione della stessa unità statistica, rilevata in archivi diversi o presente più volte nella stessa lista, anche in assenza di identificatori univoci o quando questi sono affetti da errori. L'identificazione dell'unità in archivi di diversa natura avviene attraverso chiavi comuni, presenti nei vari file; le chiavi possono essere anche non perfettamente corrispondenti. La complessità del record linkage dipende da molteplici aspetti, principalmente legati all'assenza di identificatori univoci o alla presenza di errori negli identificatori stessi.

Nella statistica ufficiale, l'uso di tecniche di record linkage nei vari processi di produzione è ormai diffuso da diversi anni e numerosi sono i campi di applicazione:

- individuazione dei duplicati in un file di dati individuali,
- studio dell'associazione tra variabili raccolte da fonti differenti;
- identificazione dei casi multipli in un archivio attribuibili ad un singolo individuo (ad esempio ricoveri, parti, …);
- creazione e aggiornamento di liste per la conduzione di indagini,
- re-identificazione per tutela riservatezza di micro-dati rilasciati per uso pubblico;
- determinazione della numerosità di una popolazione con il metodo cattura-ricattura;
- analisi di dati panel, etc.

Se negli archivi da abbinare sono presenti identificatori univoci non affetti da errore allora il problema non ha una grande complessità; in generale però, per analizzare dati privi di identificatori univoci o con identificatori univoci affetti da errore, sono richieste sofisticate procedure statistiche.

Formalmente, l'obiettivo del linkage è identificare un'unità che può essere rappresentata in maniera differente in due diverse fonti dati A e B. In generale, le coppie che si intende classificare come abbinamenti (cioè a e b sono la stessa unità), non abbinamenti (a e b sono due differenti unità) e possibili abbinamenti sono quelle dell'insieme Ω, prodotto cartesiano di A e B. Tale insieme ha cardinalità $n_A \times n_B$ ed è costituito da tutte le possibili coppie (a,b| a ∈ A, b ∈ B). Per individuare le coppie che si riferiscono alla stessa unità, gli abbinamenti, si ricorre al confronto tra k variabili, "variabili di match", comuni alle due fonti di dati. Tali variabili identificano in maniera univoca le unità, a meno, ovviamente, di errori o valori mancanti nelle variabili stesse; proprio a causa delle imperfezioni nelle variabili di match, l'abbinamento non può essere risolto attraverso l'utilizzo di un semplice "join" fra le due liste in esame. Il confronto tra le variabili viene effettuato per mezzo di un'opportuna funzione, scelta in base al tipo di variabile e alla sua qualità (in termini di completezza e correttezza). Per ogni coppia (a,b) ∈ Ω , si definisce un vettore γ, detto "vettore dei confronti", i cui k elementi sono il risultato del confronto tra le k variabili di match. Nel modello probabilistico per l'individuazione degli abbinamenti, si ipotizza che la distribuzione del vettore dei confronti sia una mistura di due distribuzioni, una generata dalle coppie (a,b) che effettivamente rappresentano la stessa unità, distribuzione *m*, e una generata dalle coppie (a,b) che rappresentano unità diverse, distribuzione *u*. A partire dalla stima di tali distribuzioni, è possibile costruire il peso composto di abbinamento (Fellegi and Sunter, 1969), dato dal rapporto delle verosimiglianze

$$r = \frac{m(\gamma)}{u(\gamma)} = \frac{\Pr(\gamma \mid M)}{\Pr(\gamma \mid U)},$$

dove *M* è l'insieme delle coppie che rappresentano gli abbinamenti e *U* è l'insieme delle coppie che rappresentano i non-abbinamenti, con $M \cup U = \Omega$ e $M \cap U = \varnothing$. In generale, la stima dei parametri delle distribuzioni viene ottenuta per mezzo dell'applicazione dell'algoritmo EM (Jaro 1989).

Sulla base del rapporto r, le coppie sono ordinate e sottoposte ad un processo di classificazione negli insiemi M ed U:

- se il peso r è maggiore di una certa soglia T_m allora la coppia viene classificata come match;
- quando il suo peso è inferiore alla soglia T_u la coppia viene classificata come non match;
- per le coppie il cui peso cade nell'intervallo $I=(T_u , T_m)$ non è possibile stabilire lo stato di abbinamento ma è necessario procedere ad un'ispezione manuale o comunque ad ulteriori analisi.

Secondo lo schema di decisione proposto da Fellegi e Sunter le due soglie, T_u e $T_{m,}$, sono fissate in modo che siano minimizzati sia gli errori di classificazione che la dimensione dell'area tra le soglie per cui non viene presa una decisione positiva.

In numerose applicazioni, attraverso il linkage si mira ad individuare tra le coppie solo legami del tipo 1 a 1, in cui, cioè, una unità del file A viene abbinata con una sola unità del

file B; in questi casi è necessario introdurre metodi di ottimizzazione che consentano di selezionare, tra tutte le coppie che coinvolgono le stesse unità della lista A e della lista B, quelle che rispettano il vincolo 1:1 e massimizzano la somma dei pesi r.

Infine, gli errori di classificazione nel modello di decisione proposto da Fellegi e Sunter sono di due tipi: gli abbinamenti errati, quando vengono abbinate unità che corrispondono a entità differenti (false matches) e gli abbinamenti mancati (false non-matches), quando record corrispondenti ad una stessa entità non vengono abbinati. In generale, gli abbinamenti errati si suddividono, a loro volta, in: accoppiamenti tra due unità che non dovrebbero essere abbinate tra loro ma con altri record e accoppiamenti tra unità che non dovrebbero essere abbinate affatto. Gli abbinamenti mancati sono spesso considerati con maggiore preoccupazione rispetto agli abbinamenti errati, in quanto più comuni e più complessi da rivedere (un abbinamento può essere verificato più agilmente rispetto ad un non abbinamento). Gli errori di abbinamento, sia abbinamenti errati che abbinamenti mancati, giocano un ruolo fondamentale per la valutazione della bontà dei risultati delle procedure di linkage e devono essere tenuti nella massima considerazione nelle successive analisi sui dati linkati, in quanto possono influire significativamente su di esse.

Misure sintetiche della qualità del linkage, basate su tali errori, sono i tassi di mancato abbinamento e di falso abbinamento, definiti, il primo, come il rapporto tra numero stimato di mancati abbinamenti e il totale stimato di veri abbinamenti e, il secondo come il rapporto tra il numero stimato di falsi abbinamenti e il totale di abbinamenti individuati.

6.1 Lo strumento RELAIS

RELAIS (REcord Linkage At IStat) è un toolkit sviluppato in Istat che mette a disposizione un insieme di tecniche per affrontare e risolvere problemi di record linkage. RELAIS si basa sull'idea che un processo di record linkage in quanto molto complesso può essere visto come costituito da diverse fasi per ognuna delle quali possono essere adottate diverse tecniche risolutive afferenti a diverse aree di conoscenza. La scelta della tecnica più appropriata da applicare dipende dal dominio di applicazione.

RELAIS fornisce diverse tecniche per le diverse fasi di un processo di RL, consentendo di combinare tali tecniche in modo da ottenere il processo lavorativo ottimale per la specifica applicazione.

RELAIS è stato sviluppato come progetto open source in modo tale che diverse soluzioni già disponibili nella comunità scientifica possono facilmente essere riutilizzate. E' stato rilasciato con licenza EUPL (European Union Public License). Dalla versione 2.0 RELAIS ha un'architettura basata su una base di dati relazionale. In particolare è stato scelto l'ambiente mySql per rispecchiare la filosofia open source. Per quanto riguarda l'ambiente di programmazione si è scelto di implementare RELAIS utilizzando due linguaggi aventi un paradigma di base diverso: Java, linguaggio object-oriented e R, linguaggio funzionale. Questa scelta è maturata a seguito della riflessione per cui il processo di record linkage necessita sia di tecniche prevalentemente orientate alla gestione dei dati, per le quali Java si rivela più appropriato, sia di tecniche orientate al calcolo, per le quali è più appropriato il linguaggio R. Infine la scelta è ricaduta sui linguaggi Java e R in quanto rispecchiano la filosofia open source propria del progetto RELAIS.

RELAIS intende mettere a diposizione tecniche di record linkage anche ad utenti non esperti. Per tale motivo è stata curata l'interfaccia grafica che consente di costruire facilmente progetti di RL con una buona flessibilità controllando però che vengano

rispettate le regole di precedenza delle fasi.

In generale, le fasi principali di un progetto di RL individuate in RELAIS sono:

- Preprocessamento dei dati – preparazione dei file di input;
- Creazione/riduzione dello spazio di ricerca. Le tecniche di riduzione dello spazio messe a diposizione sono: (i) bloccaggio, (ii) sorted neighborhood, (iii) nested blockingg iustapposizione delle due precedenti tecniche. Relais mette a disposzione dei metadati per la scelta delle variabili di bloccaggio più idonee;
- Scelta delle variabili identificative comuni (variabili di match); nel toolkit sono a disposizione dei metadati per supportare l'utente nella scelta delle migliori variabili di matching;
- Scelta delle funzioni di confronto. Le funzioni disponibili sono: (i) equality; (ii) numeric n comparison; (iii) 3Grams; (iv) Dice; (v) Jaro; (vi) JaroWinkler; (vii) Levenshtein; (viii) Soundex.
- Scelta del modello decisionale. Sono disponibili il modello deterministico, in particolare il matching esatto e il matching basato su regole definite dall'utente e il metodo probabilistico che implementa il modello di Fellegi-Sunter (1969).
- Selezione di match unici (linkage 1:1): si possono applicare due diverse tecniche per passare da abbinamenti n:m ad abbinamenti 1:1 in particolare si può applicare una soluzione ottimale o una soluzione greedy (applicabile anche quando la prima tecnica non porta ad un risultato a causa dell'eccessiva dimensione del problema);
- Valutazione della qualità dei risultati abbinati; nel caso dell'uso dell'approccio probabilistico i risultati della qualità dei risultati del linkage sono forniti in termini di probabilità di corretto abbinamento e mancato abbinamento.

Per ciascuna delle fasi individuate sono note e largamente utilizzate tecniche diverse. In funzione della particolare applicazione e dei dati in esame, può essere opportuno iterare e/o omettere alcune fasi, così come preferire in ciascuna fase alcune tecniche rispetto ad altre. RELAIS, già nella sua prima versione rilasciato nel 2008, mirava a rendere fruibili le tecniche di RL ad una platea più ampia dei soli esperti del settore.

Per le applicazioni relative a questo progetto e descritte nel documento è stata usata la versione 2.2 di RELAIS. Le caratteristiche specifiche del sistema possono essere trovate nel manuale utente, disponibile all' indirizzo http://www.istat.it/it/strumenti/metodi-e-software/software/relais

6.2 La preparazione dei dati

La dimensione dei file considerati per l'abbinamento è rispettivamente 19709 unità per l'indagine CIS e 19673 unità per l'indagine ICT; sono state escluse alcune unità che presentavano dati mancanti nelle variabili rilevanti per la strategia di abbinamento ed inclusi poche unità che hanno risposto tardivamente, a seguito di solleciti.

Le variabili comuni ai due dataset da utilizzare per la riduzione dello spazio di ricerca e come variabili di abbinamento sono: gli addetti nell'anno 2008 riportati nell'archivio Asia e il valore rilevato dalle indagini (sia classificati secondo i domini di stima, sia in valore assoluto, sia rielaborati secondo le trasformazioni descritte in appendice 1), il valore totale dei ricavi, in euro, realizzato dall'impresa nel corso dell'esercizio 2008, sia risultante dall'archivio ASIA che rilevato alle indagini (per quanto riguarda le imprese rispondenti alla rilevazione ICT, questa variabile non è stata richiesta alle imprese di intermediazione finanziaria - settore *K*), il codice di attività economica dell'impresa, secondo le 5 cifre della

codifica NACE, la regione e la provincia per la localizzazione geografica dell'impresa.

Per poter applicare in maniera ottimale le tecniche di abbinamento probabilistico, è stato ritenuto opportuno applicare alcune trasformazioni alle variabili continue relative al fatturato e agli addetti. Di fatto sono state considerate diverse trasformazioni, in particolare, la trasformazione logaritmica in base 10, arrotondata a 0, 1, 2, 3 cifre decimali e la distribuzione in classi della trasformazione logaritmica secondo le classi individuate dai percentili. Inoltre, quando dai dati di indagine risultavano valori mancanti per le variabili in questione, sono stati presi in considerazione i corrispondenti valori riportati nel registro ASIA. In appendice 1 si riportano in maniera dettagliata le trasformazioni applicate alle variabili continue e esperimentate nelle applicazioni descritte nel seguito.

Il primo passo per l'integrazione dei microdati delle indagini ICT e CIS per l'anno 2008 è stato eseguito un abbinamento esatto secondo la chiave di aggancio certa data dal "codice impresa" riportato nell'archivio ASIA. Tale abbinamento costituisce in qualche modo il gold standard degli abbinamenti probabilistici testati nel seguito, poiché le coppie abbinate secondo il codice impresa possono essere considerati veri abbinamenti anche nelle successive sperimentazioni probabilistiche e perché confrontando il numero di abbinamenti risultanti dalle procedure probabilistiche con il numero di abbinamenti ottenuti attraverso il codice impresa si può valutare il guadagno in termini di nuovi abbinamenti trovati dalle procedure probabilistiche.

L'abbinamento basato sull'identità del codice impresa tra i due dataset individua 9882 match.

L'abbinamento tra il dataset ICT e CIS attraverso tecniche probabilistiche applicate all'insieme di tutte le coppie generate dal confronto tra tutti i record ICT con tutti i record CIS (prodotto cartesiano) coinvolgerebbe un numero di coppie pari a

$$19709 \times 19673 = 387735157$$

Si tratta di un ordine di grandezza elevato, sia sotto il profilo della memoria di massa idonea a conservare i risultati delle elaborazioni, sia dal punto di vista dei tempi necessari a completare i calcoli.

Una questione rilevante per poter applicare metodi di abbinamento statistici è stata, dunque, di limitare il numero di confronti tra le osservazioni dei due insiemi di dati, pur non conoscendo a priori quali osservazioni costituiscano un abbinamento esatto. La contraddizione tra il proposito di limitare il numero dei confronti e l'ignoranza sull'abbinamento delle unità statistiche nei due data set è soltanto apparente: tra le informazioni disponibili ve ne possono essere alcune che permettono di classificare come poco probabili o impossibili gli abbinamenti di determinate coppie di record.

Dunque, si rileva di cruciale importanza limitare il numero dei confronti tra le osservazioni dei due insiemi di dati. Per questo motivo sono stati applicati vari metodi di riduzione: Bloccaggio, Sorted Neighborhood ed NestedBlocking predisposti nel software open source RELAIS 2.2 impiegato per il record linkage.

Sugli insiemi di coppie generati dai metodi di riduzione sopra indicati, sono stati applicati diversi modelli per il record linkage probabilistico. Nei dati da abbinare sono stati considerati anche i record che si abbinano secondo l'identificativo esatto (codice impresa).

Tre i diversi modelli implementati attraverso il software RELAIS 2.2, quelli che hanno soddisfatto gli obiettivi delle tre strategie di integrazione descritte nel paragrafo 1.1. sono descritti di seguito.

6.3 Le strategie di linkage sperimentate

6.3.1 Prima strategia: il modello probabilistico che massimizza gli abbinamenti veri

Come definito nel paragrafo 1.1, il primo obiettivo che si vuole perseguire con le metodologie di integrazione probabilistica è la massima identificazione delle unità comuni alle due indagini, secondo il codice impresa.

Tra quelli sperimentati, il modello migliore secondo la prima strategia è quello che adotta come metodo di riduzione del numero di coppie candidate il Sorted Neighbohoord e la variabile "volume d'affari 2008 di fonte ASIA" come variabile d'ordine con una finestra dei confronti pari a w=110 . Le variabili di matching utilizzate nel modello che fornisce i risultati migliori sono: la regione in cui è localizzata l'impresa, il codice di attività economica NACE a due cifre; gli addetti medi nell'anno riportati nell'archivio Asia 2008. Come soglie per l'attribuzione all'insieme dei matches e dei un-matches sono state scelte: Tm=0.8 e Tu=0.7.

Il record linkage, in questo modo, individua 12111 matches, che coinvolgono 10642 imprese CIS e 10662 imprese ICT, tra cui tutte le 9882 coppie di imprese comuni. Le 9882 vere coppie coinvolgono 11212 coppie tra le 12111 individuate come matches. Di fatto rimangono 899 coppie che coinvolgono 760 imprese in CIS che non hanno una corrispondenza esatta con le imprese in ICT.

Questa strategia, permette quindi di creare un dataset completo integrato di 11422 imprese.

Se, invece, sulle 12111 coppie individuate come matches, si esegue un'operazione di ottimizzazione affinché risultino esclusivamente abbinamenti univoci (un record dell'indagine ICT può essere abbinato con un solo record dell'indagine CIS e viceversa), le coppie da considerare valide si riducono a 9405 di cui 9115 vere coppie (con codice impresa uguale) e 290 ulteriori coppie univoche. Osserviamo che in questo caso l'operazione di riduzione 1 a 1 degli abbinamenti porta a scartare 767 coppie che sono vere secondo il codice impresa. In questa circostanza, considerare solo abbinamenti univoci significa imporre che, per le coppie che non sono uguali secondo il codice impresa, le informazioni rilevate per una singola impresa in una delle due indagini siano abbinate ad una sola impresa rilevata all'altra indagine. Volendo fare un parallelo con le metodologie di imputazione per valori mancanti, ciò significa che un'impresa CIS viene usata come donatore per completare i campi di una singola impresa ICT, una sola volta, e viceversa.

Nell'appendice 2 sono descritte nel dettaglio le analisi preliminari condotte sulle variabili di abbinamento per la messa a punto della attuale strategia migliore e i valori stimati dei parametri del modello probabilistico di abbinamento.

6.3.2 Seconda strategia: il modello probabilistico che massimizza agli abbinamenti in più rispetto ai veri

Il secondo obiettivo che si vuole perseguire con le strategie di integrazione, come riportato nel paragrafo 1.1, è quello di incrementare la base di dati per le analisi congiunte, cercando di abbinare più unità di quelle che si agganciano secondo codice impresa.

Tra quelli sperimentati, il modello che permette di individuare il maggior numero di abbinamenti in più rispetto a quelli per codice impresa è quello che riduce lo spazio di ricerca delle coppie candidate all'abbinamento applicando congiuntamente il tradizionale

bloccaggio e il metodo del Sorted Neighbohoord all'interno di ciascun blocco. Come variabile di blocco è stata utilizzata la concatenazione tra il macro-settore di attività dell'impresa e il numero di addetti codificato secondo 4 classi; in ciascun blocco il Sorted Neighbohoord ha impiegato la variabile d'ordine "volume d'affari 2008 di fonte ASIA" con una finestra dei confronti pari a w= 80. Le variabili di matching utilizzate nel modello che fornisce i risultati migliori sono: gli addetti medi e il volume d'affari 2008 riportati nell'archivio Asia confrontati secondo la funzione "*NumericComparison*" con soglia *0.8* e il codice di attività economica NACE a due cifre con funzione di confronto "*Equality*". Come soglie per l'attribuzione all'insieme dei matches e dei un-matches sono state scelte: Tm=0.8 e Tu=0.7.

Il record linkage, in questo modo, individua 103985 coppie che coinvolgono 6204 imprese CIS, di cui 3426 sono vere coppie. Di fatto, oltre alle 3426 vere coppie individuate, vengono abbinate 2737 imprese CIS e 2293 imprese ICT, che non hanno una corrispondenza esatta con le imprese dell'altra indagine ma per cui è possibile ricostruire l'informazione grazie al valore elevato delle probabilità di corretto abbinamento, che garantisce la qualità in termini di accuratezza del risultato conseguito,

Il file integrato che si può creare con questa strategia è di 8456 records, dove più di 5000 sono le coppie abbinate che non coincidono per codice impresa. Se poi si considerano tutte le coppie individuate dalla corrispondenza esatta del codice impresa, e non solo quelle abbinate dalla procedura di linkage probabilistico, è possibile costruire un file completo di 14912 imprese.

Se, invece, sulle coppie individuate come matches, si esegue un'operazione di ottimizzazione affinché risultino esclusivamente abbinamenti univoci (un record dell'indagine ICT può essere abbinato con un solo record dell'indagine CIS e viceversa), le coppie da considerare valide si riducono a 2170 di cui 1090 vere coppie (con codice impresa uguale) e 1080 ulteriori coppie univoche. Osserviamo che, ancora una volta, l'operazione di riduzione 1 a 1 degli abbinamenti porta a scartare 2336 coppie che sono vere secondo il codice impresa.

6.3.3 Terza strategia: creazione del file completo di dati integrati

La terza strategia di integrazione è volta alla costruzione di un file di microdati completo per tutte le unità rispondenti alle due indagini. A tal fine si è scelto di applicare diversi modelli di integrazione in passi successivi, riconoscendo in ogni passo le unità più simili non ancora individuate dal modello applicato al passo precedente.

In questo documento si riportano i nove modelli di integrazione che, applicati di seguito, permettono di costruire un file di microdati completo per tutte le variabili e per tutte le unità rispondenti alle due indagini.

I nove modelli di linkage e le relative scelte sono riassunti nella tabella seguente.

Tavola 6.1 Sintesi dei modelli applicati per la terza strategia.

METODO DI RIDUZIONE	Variabili di Blocking	Variabili di Sorting	Variabili di Matching
Sorted neighborhood		- volume d'affari di ASIA (finestra 110)	- addetti medi - regione - NACE a due cifre
Sorted Neighborhood		- logaritmo degli addetti (finestra 150)	- volume d'affari di ASIA - regione - NACE a due cifre
Nested Blocking	- macro-settore - numero di addetti in 4 classi	- volume d'affari di ASIA (finestra 80)	- addetti medi - volume d'affari di ASIA - NACE a due cifre
Nested Blocking	- macro-settore - regione	- volume d'affari di ASIA (finestra 10)	- addetti medi - volume d'affari di ASIA - NACE a due cifre
Nested Blocking	- codice di sezione (1 cifra NACE) - regione	- volume d'affari di ASIA (finestra 10)	- addetti medi - volume d'affari di ASIA - NACE a due cifre
Nested Blocking	- NACE a 5 cifre	- volume d'affari di ASIA (finestra 20)	- addetti medi - volume d'affari di ASIA - regione
Nested Blocking	- macro-settore - numero di addetti in 4 classi	- volume d'affari da indagine (finestra 10)	- addetti medi - volume d'affari da indagine - NACE a due cifre
Nested Blocking	- macro-settore - provincia	- volume d'affari di ASIA (finestra 30)	- addetti medi - volume d'affari di ASIA - NACE a due cifre
Nested Blocking	- codice di sezione (1 cifra NACE) - provincia	- volume d'affari di ASIA (finestra 30)	- addetti medi - volume d'affari di ASIA - NACE a due cifre

Le 143269 coppie individuate coinvolgono 14697 imprese CIS, di cui 9882 sono vere coppie. Le 9882 coppie coinvolgono 83402 coppie tra quelle proposte come matches, coinvolgendo 9882 imprese CIS. Di fatto rimangono 59867 coppie che coinvolgono 4815 imprese CIS che non hanno una corrispondenza esatta con le imprese ICT.

Il file integrato ottenuto dall'insieme delle 9 metodologie è di 19748 record.

Con la riduzione 1 a 1, le imprese con lo stesso codice impresa individuate sono 9792 e si aggiungono 4180 coppie univoche con codice impresa diverso.

6.4 Alcune considerazioni sulle procedure di abbinamento sperimentate

Da un'analisi complessiva di tutte le strategie condotte si è notato che:

- i metodi che hanno ridotto lo spazio di ricerca con un aumento dell'efficienza del record linkage sono stati il NestedBlocking e il Sorted Neighborhood;
- sia l'efficacia e l'efficienza dei due metodi di blocco sono strettamente collegate alle caratteristiche statistiche e qualitative del blocking key, infatti la variabile di blocco deve avere un alto potere discriminate e quanto più possibile priva di errori e valori mancanti ed inoltre deve avere un numero considerevole di modalità equi distribuite tra le unità. La variabile di blocco che ha presentato queste caratteristiche è stata la variabile volume d'affari 2008 di fonte Asia 2008, *vaf08daasia*;
- non si è riusciti a individuare un modello di abbinamento che raggiunga risultati soddisfacenti in termini di "veri match" considerando come variabili di matching le

due variabili numeriche: addetti e fatturato (prima strategia); al contrario si è individuato un numero maggiore di match considerando le variabili *add08mDaAsia e vaff08DaAsia* (seconda strategia);

- le variabili di matching che risultano avere un alto potere discriminante e che permettono di identificare le unità sono state il codice regione dell'impresa, *reg08*, e il codice Ateco d'impresa a due digit, *nace2*;
- per il metodo di riduzione Sorted Neighborhood si è notato che si raggiungono risultati soddisfacenti considerando come dimensione della finestra un range di 100-200;
- la variabile mac, sez, pro08, reg08, ate08 e cladd utilizzate come variabile di blocco nel metodo di riduzione Nested Blocking risultano essere più discriminanti, raggiungendo risultati più soddisfacenti in termini di quantità (numero match trovati);
- è stato scelto un range da 20 a 200 nel fissare l'ampiezza della finestra nelle strategie adottate, poiché con valori più bassi o più alti le diverse metodologie o non individuavano alcun risultato o individuavano un numero di match troppo basso ;

Inoltre, confrontando la distribuzione dei 9882 veri match nelle variabili *mac* e *cladd* nelle tre strategie di integrazione sopra descritte si sono ottenute le seguenti percentuali:

Tavola 6.2 - Distribuzione dei "veri" match nella varibili mac e cladd nelle tre strategie di Integrazione

		DETERMINISTICO	QUALITATIVO	QUANTITATIVO	9 PROVE MIGLIORI
	M1-CIS	37.80%	36.28%		30%
MAC	M2-CIS	24.17%	25.89%	66.4%	29.14%
	M3-CIS	38.03%	37.82%	33.61%	41%
	CL2-CIS	64.13%	66.39%	66.40%	71%
CLADD	CL3-CIS	11.55%	10.95%	12.22%	11%
	CL4-CIS	8.60%	8.03%	7.8%	7%
	CL5-CIS	15.71%	14.63%	13.6%	11.42%
	M1-ICT	37.78%	35.69%		28.45%
MAC	M2-ICT	24.17%	25.51%	66.8%	31.47%
	M3-ICT	38.04%	38.80%	33.2%	40.1%
	CL2-ICT	64.33%	66.53%	66.8%	70.25%
CLADD	CL3-ICT	11.48%	10.88%	12%	10.7%
	CL4-ICT	8.60%	8.10%	8.9%	8%
	CL5-ICT	15.55%	14.48%	12.25%	11.15%

Dalla Tabella 6.2 è possibile notare come la distribuzione dei 9882 veri match nei tre dataset integrati ottenuti dalle tre strategie sopra descritte nelle modalità delle variabili *mac* e *cladd* è simile, quindi le tre strategie hanno potere identificativo analogo nei confronti delle coppie che sono sicuramente la stessa unità.

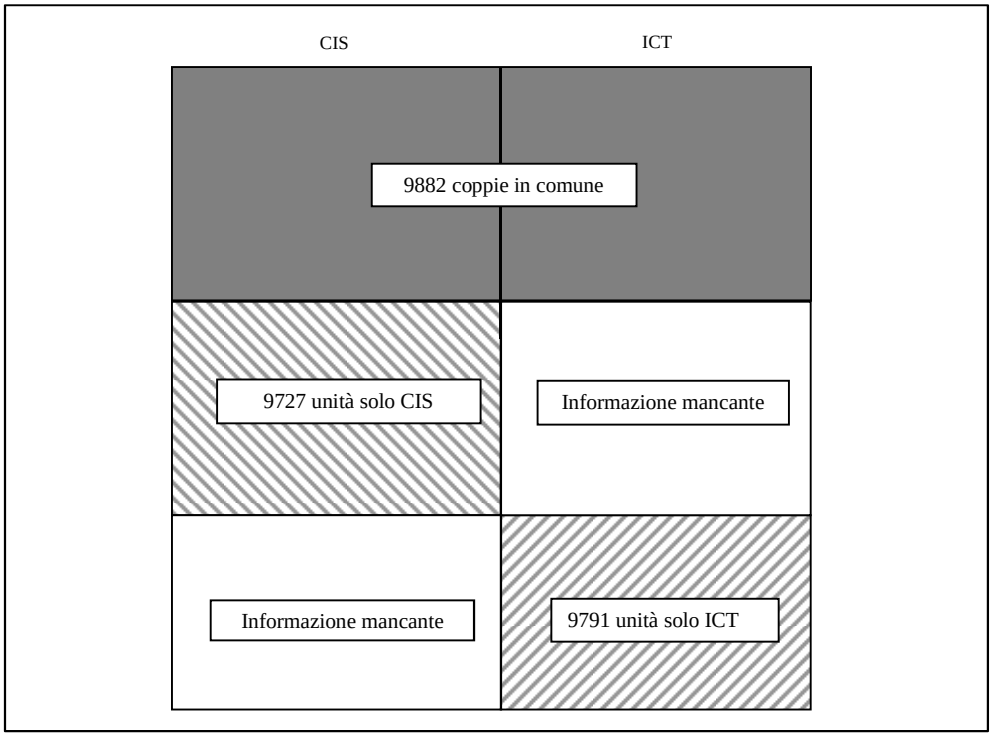
7. Sintesi dei risultati e valutazione della qualità

Rispetto all'obiettivo di creare un unico insieme completo di dati provenienti dalle indagini CIS e ICT, i risultati delle procedure di linkage delineate nei paragrafi precedenti possono essere sintetizzati attraverso le seguenti rappresentazioni.

La figura 7.1 rappresenta la disponibilità di dati dopo l'abbinamento deterministico per codice impresa. E' quindi possibile riconoscere le 9882 unità in comune per le due indagini (in colore blu), per cui l'informazione è completa, e le restanti unità, distinte e provenienti dai due campioni di rispondenti, per cui sono disponibili solo le informazione relative ad una delle due indagini (in arancio le unità provenienti dall'indagine CIS e in verde le unità provenienti dall'indagine ICT).

In questo caso, la dimensione del dataset completo a disposizione per le analisi successive è di 9882 record; la qualità di questo dataset è massima (sono effettivamente le stesse unità secondo il codice impresa) ma si perde completamente l'informazione rilevata sulle restanti 9727 unità dell'indagine CIS e sulle 9791 ulteriori unità rispondenti all'indagine ICT.

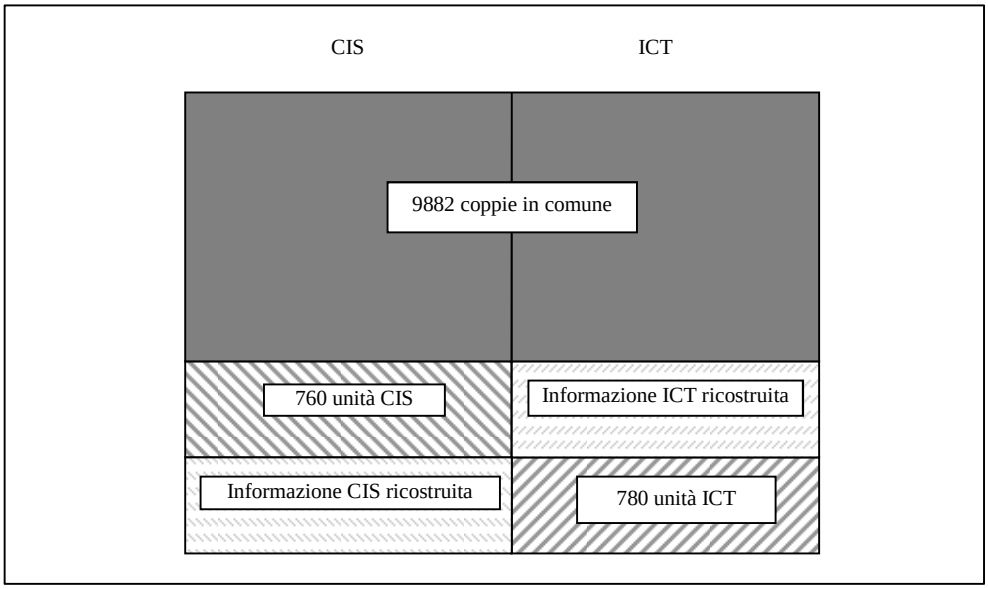
Figura 7.1. La disponibilità di dati dopo l'abbinamento deterministico per codice impresa.



La figura 7.2 invece rappresenta la disponibilità di dati dopo l'abbinamento probabilistico che massimizza l'identificazione delle vere unità comuni alle due indagini, cioè quelle con codice impresa coincidente (obiettivo individuato come prima strategia nel paragrafo 1.1).

In questo caso, la dimensione del dataset completo a disposizione per le analisi successive è di 11422 record: alle 9882 unità in comune per le due indagini (ancora in colore blu), si aggiungono 760 unità rispondenti all'indagine CIS, per cui il metodo di abbinamento permette di ricostruire le informazioni relative all'indagine ICT utilizzando unità effettivamente rispondenti a tale indagine (la parte di dataset in verde chiaro) e 780 unità rispondenti all'indagine ICT, per cui il metodo di abbinamento permette di ricostruire le informazioni relative all'indagine CIS utilizzando unità effettivamente rispondenti a quest'ultima (la parte di dataset in arancione chiaro).

Figura 7.2 - La disponibilità di dati dopo l'abbinamento col modello probabilistico individuato secondo la prima strategia.



La qualità di questo dataset è misurabile attraverso la stima della probabilità di corretto abbinamento, che è uno degli output del modello probabilistico di abbinamento fornito dal software RELAIS. Le 1540 coppie hanno tutte probabilità stimata di corretto abbinamento pari a 0.88, mentre per le 9882 coppie costituite effettivamente dalla stessa unità si può considerare una probabilità di corretto abbinamento pari a 1. In tal modo la probabilità di corretto abbinamento complessiva per questo dataset di 11422 record è stimata a 0.98.

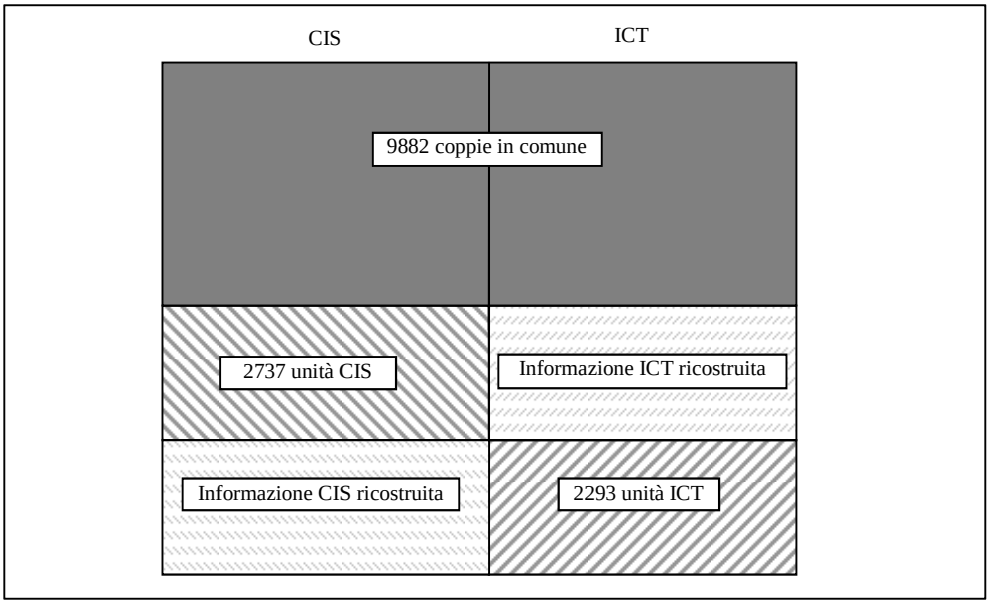
La probabilità di corretto abbinamento ed altre misure che valutano la qualità dei risultati del linkage (come ad esempio il tasso di mancato abbinamento e il tasso di falso abbinamento) devono giocare un ruolo fondamentale nelle successive analisi statistiche che sul dataset completo si intendono effettuare. Infatti, in presenza di dati provenienti da operazioni di linkage, le tradizionali metodologie statistiche possono portare a risultati fortemente distorti se non si tiene nella debita considerazione il processo di integrazione che ha generato i dati e il fatto che il linkage, come in genere tutti i processi di produzione del dato statistico, non è privo di errori. Le applicazioni di record linkage devono quindi essere

corredate da opportune informazioni sulla qualità del linkage da utilizzare con apposite metodologie di stima, volte ad assicurare la qualità delle analisi condotte sui dati abbinati. In questi termini si apprezza appieno il vantaggio di utilizzare tecniche di record linkage probabilistico, che, sotto opportune condizioni di validità dei modelli applicati, sono provvisti per definizione di indicatori in grado di misurare la qualità dei risultati ottenuti.

La figura 7.3 rappresenta la disponibilità di dati dopo l'abbinamento probabilistico che incrementa la base di dati per le analisi congiunte, cercando di abbinare un numero maggiore di unità rispetto a quelle che si agganciano secondo il codice impresa (obiettivo individuato come strategia 2 nel paragrafo 1.1).

In questo caso, la dimensione del dataset completo a disposizione per le analisi successive è di 14912 record: alle 9882 unità in comune per le due indagini (ancora in colore blu), si aggiungono 2737 unità rispondenti all'indagine CIS, per cui il metodo di abbinamento permette di ricostruire le informazioni relative all'indagine ICT utilizzando unità effettivamente rispondenti a tale indagine (la parte di dataset in verde chiaro) e 2293 unità rispondenti all'indagine ICT, per cui il metodo di abbinamento permette di ricostruire le informazioni relative all'indagine CIS utilizzando unità effettivamente rispondenti a quest'ultima (la parte di dataset in arancione chiaro). Volendo essere rigorosi, questa strategia individua solo 3426 coppie tra le 9882 che si abbinano per codice impresa, e quindi formalmente il dataset completo prodotto dalla procedura di abbinamento in senso stretto sarebbe composto da 8456 record. Tuttavia non ci sono motivi ragionevoli per escludere dalle successive analisi statistiche sul dataset completo la parte di coppie non individuate dalla procedura probabilistica ma facilmente rintracciabili attraverso la corrispondenza del codice impresa.

Figura 7.3 - La disponibilità di dati dopo l'abbinamento col modello probabilistico individuato secondo la seconda strategia.

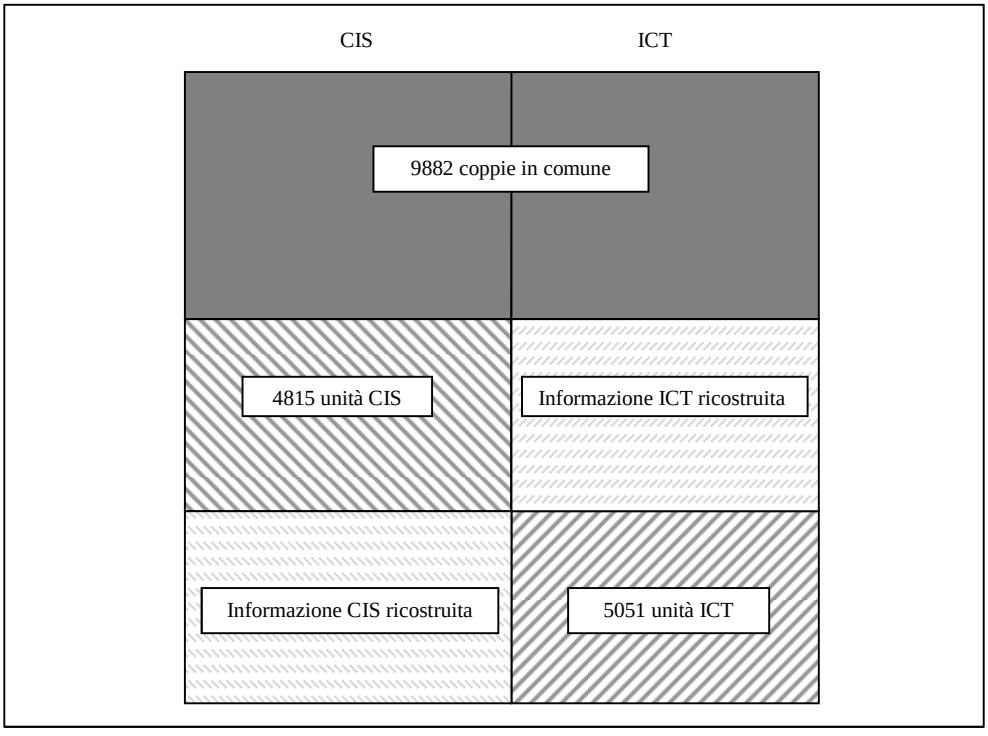


La stima della probabilità di corretto abbinamento per questo dataset è 0.96: ancora una volta, per le 9882 coppie costituite effettivamente dalla stessa unità è stata considerata una probabilità di corretto abbinamento pari a 1 mentre per le restanti 5030 coppie, ricostruite attraverso le operazioni di linkage, il modello di abbinamento fornisce una probabilità media di corretto abbinamento pari a 0.89. La probabilità di corretto abbinamento delle coppie riconosciute attraverso il modello probabilistico varia tra un minimo di 0.81 ed un massimo di 0.92.

La figura 7.4 rappresenta la disponibilità di dati dopo l'abbinamento probabilistico che tenta la costruzione del dataset completo di microdati attraverso l'applicazione di diversi modelli di integrazione in passi successivi, riconoscendo in ogni passo le unità più simili non ancora individuate dal modello selezionato al passo precedente (obiettivo individuato come strategia 3 nel paragrafo 1.1).

In questo caso, la dimensione del dataset completo a disposizione per le analisi successive è di 19748 record: alle 9882 unità in comune per le due indagini (ancora in colore blu), si aggiungono 4815 unità rispondenti all'indagine CIS, per cui il metodo di abbinamento permette di ricostruire le informazioni relative all'indagine ICT utilizzando unità effettivamente rispondenti a tale indagine (la parte di dataset in verde chiaro) e 5051 unità rispondenti all'indagine ICT, per cui il metodo di abbinamento permette di ricostruire le informazioni relative all'indagine CIS utilizzando unità effettivamente rispondenti a quest'ultima (la parte di dataset in arancione chiaro).

Figura 7.4 - La disponibilità di dati dopo l'abbinamento col modello probabilistico individuato secondo la terza strategia.



La stima della probabilità di corretto abbinamento per questo dataset è 0.94: ancora una volta, per le 9882 coppie costituite effettivamente dalla stessa unità è stata considerata una probabilità di corretto abbinamento pari a 1 mentre per le restanti 9866 coppie, ricostruite attraverso le operazioni di linkage, il modello di abbinamento fornisce una probabilità media di corretto abbinamento pari a 0.88. La probabilità di corretto abbinamento delle coppie riconosciute attraverso il modello probabilistico varia tra un minimo di 0.81 ed un massimo di 1.

E' interessante notare come, ampliando la dimensione del dataset completo ricostruito con tecniche di linkage, si abbassi il valore stimato della probabilità di corretto abbinamento. Ad ogni modo, anche nel caso della strategia 3, che permette di considerare un dataset di dimensione comparabile a quelle dei rispondenti delle due indagini, la probabilità complessiva di corretto abbinamento si conserva su valori molto elevati e ciò garantisce, in combinazione con le opportune metodologie statistiche, la qualità delle analisi successive su questi dati.

Si rimanda a sviluppi successivi una comparazione dei risultati delle diverse strategie di integrazione in termini di impatto che le stime delle probabilità di corretto abbinamento avranno sullo studio dell'analisi congiunta di variabili osservate in campioni diversi: infatti, mentre la valutazione degli effetti della probabilità di linkage (e quindi degli errori di linkage) è stata recentemente sviluppata in letteratura per lo studio di relazioni basate su modelli lineari generalizzati (Chambers 2009, Chipperfield et al. 2011), le stesse tematiche non sono state approfondite, almeno per quanto di nostra conoscenza, per i modelli utilizzati nel paragrafo 5, quelli che mirano all'individuazione di relazioni tra le variabili CIS e ICT, utili a definire i criteri di correzione necessari per l'imputazione dei missing nel data set integrato CIS-ICT.

8. Alberi di regressione

Per identificare le imprese rilevate ad entrambe le indagini, CIS ed ICT, al fine di perseguire, con metodologie diverse rispetto al record linkage probabilistico, gli obiettivi di massimizzazione degli abbinamenti e dell'utilizzo al meglio dell'informazione raccolta alle due indagini, si è fatto ricorso anche al software Answer Tree che implementa la metodologia degli alberi di classificazione e di regressione (Breiman et al,1984).

In generale, l'uso degli alberi di classificazione può essere finalizzato sia a produrre un'accurata partizione della popolazione rispetto alla variabile target e, quindi, a ricostruire l'informazione sulle unità che appartengono allo stesso nodo di quelle per cui essa è nota sia a rivelare legami nascosti tra la variabile target e altre variabili esplicative.

L'albero è costruito a partire dal nodo padre, X, a cui appartengono tutte le unità della popolazione di interesse che viene suddiviso, con successivi splits, in due nodi figli. I nodi finali, "terminali", formano una partizione del nodo padre X e, ad ognuno di essi, è associato un valore della variabile target, quello prevalente nel nodo. In generale, la partizione che corrisponde alla regola di classificazione è ottenuta mettendo insieme tutti i nodi terminali con lo stesso valore della variabile target.

In questo particolare contesto gli alberi di classificazione sono stati adottati per vedere quali variabili presenti in entrambe le rilevazioni potessero essere i migliori predittori per individuare l'appartenenza al sottoinsieme delle sole imprese rilevate ad una sola delle due

indagini; tanto più forte è il potere esplicativo delle variabili tanto più esse possono essere usate per ricostruire l'informazione sulle unità su cui è mancante, specialmente in contesti in cui la sovrapposizione tra le due indagini è contenuta. Le analisi, quindi, sono state condotte separatamente per le due indagini, dato come è costruita la variabile target. La metodologia proposta ha anche il valore aggiunto di mirare a ricostruire l'informazione per le unità che appartengono allo stesso nodo se la classificazione delle unità risulta essere corretta e, quindi, contenuta la componente di errore dovuta alla misclassificazione.

La variabile target è una variabile binaria che assume valore 1 quando l'impresa, rilevata in CIS (ICT) appartiene anche ICT (CIS) e 0 altrove; nell'analisi condotta si parte da una delle due indagini CIS e ICT che ammontano rispettivamente a 19709 e a 19673 unità e la variabile target assumerà valore 1 nel caso delle imprese comuni, 9882, rilevate ad entrambe le indagini.

Le variabili usate come predittori sono state le stesse adottate nel record linkage probabilistico (si veda appendice 1) e, quindi:

- la classe di addetti e gli addetti ;
- il codice ateco di impresa, nace, a due cifre e a tre cifre;
- le variabili di localizzazione dell'impresa, regione e provincia;
- il macrosettore di attività economica;
- il fatturato, ove presente.

A differenza del record linkage probabilistico, negli alberi di classificazione la stessa variabile può concorrere più volte, con modalità diverse, a determinare i diversi splits dell'albero e la classificazione finale di unità simili, rispetto alla modalità della variabile assunta come target.

Sono state fatte diverse prove usando i predittori tutti insieme o sottoinsiemi degli stessi, sulle due indagini, cercando di individuare le variabili migliori per delineare l'appartenenza al sottoinsieme di riferimento. Avendo una variabile target binaria si sono scelte due differenti criteri per individuare la partizione ottimale: l'entropia basata sulla funzione di verosimiglianza (che tende a separare perfettamente le unità rispetto alla variabile target) e la distanza di Gini che misura l'impurità del nodo t, I_t , che è massima se nella classe c'e' l'equidistribuzione tra le n unità del nodo t della variabile target, Y, è data dalla seguente formula:

$$I_t = 1 - \sum_{Yt=0,1} \left(\frac{n_{Yt}}{n_t} \right)^2 \quad \text{con t=1,2…m (numero totale di nodi)}$$

L'albero ottimo è quello che ha il minor tasso di errata classificazione delle unità.

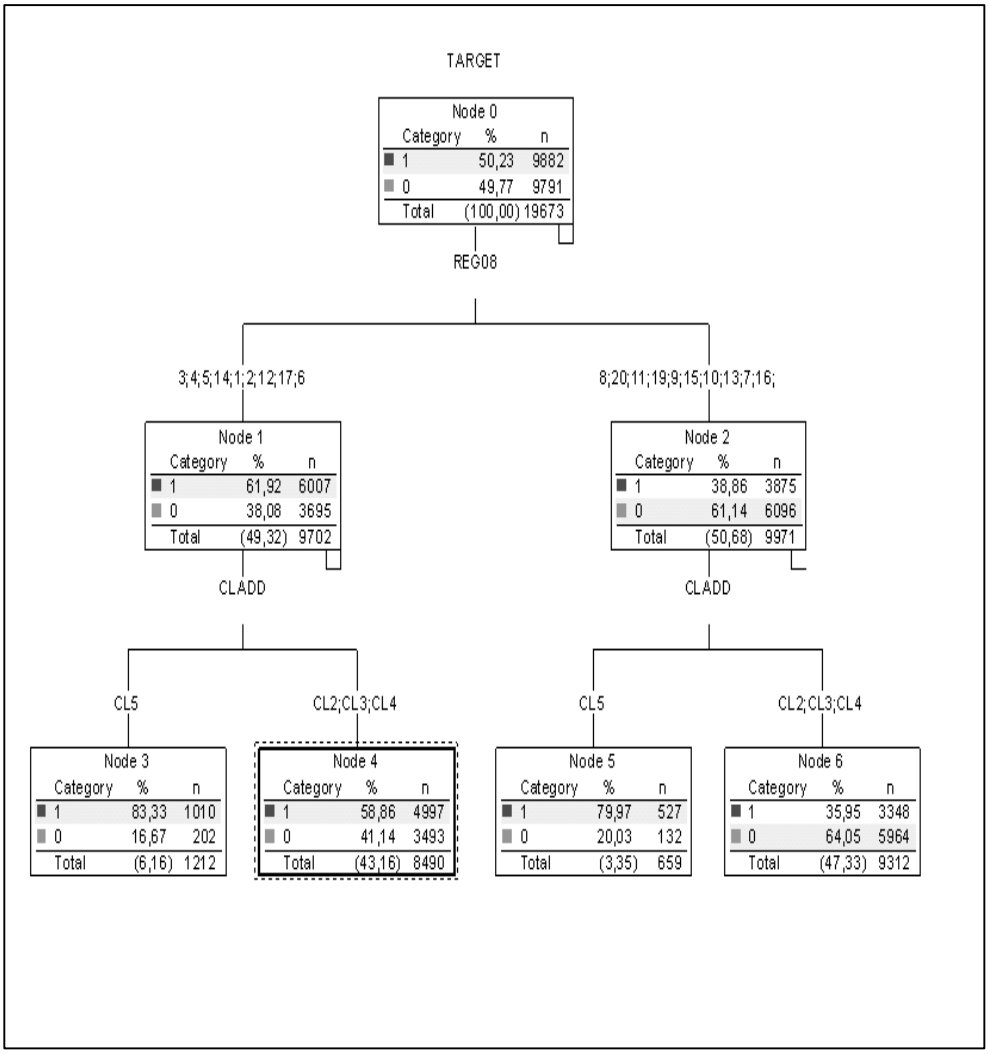
Non si apprezzano significative differenze nelle variabili esplicative per le due indagini.

Risultano essere particolarmente esplicative tra tutte le variabili inserite negli alberi di classificazione:

- la classe di addetti dell'impresa che concorre a definire l'albero migliore con diverse modalità;
- la regione di appartenenza dell'impresa;
- il macrosettore di attività economica;

In molti alberi, che tendevano a creare più alberi figli, entrava come buon predittore anche la nace (codice ateco a 2 e a 3 cifre).

Figura 8.1. Albero di classificazione finale. Imprese appartenenti all'indagine ICT.



L'albero riportato in figura 8.1 a titolo esemplificativo per mostrare quale variabili e con quali modalità esse concorrevano a determinare la partizione della popolazione di riferimento in classi omogenee secondo la variabile target, presenta 4 nodi terminali ottenuti sulla base della regione di appartenenza delle imprese e della classe di addetti (cl2:meno di 49 addetti; cl3: tra 50 e 99 addetti; cl4:tra 100 e 249 addetti; cl5: più di 250 addetti), e un indice di corretta classificazione delle unità C, pari a 0.70.

9. Conclusioni e prospettive future

Le analisi condotte a partire dal contesto reale dei dati del 2008 producono risultati incoraggianti che mostrano lo stretto legame tra le indagini CIS e ICT e mettono in evidenza l'effettiva utilità di strategie di record linkage con l'obiettivo di creare un dataset integrato, secondo i diversi scenari descritti nel documento. Resta da misurare la validità dei risultati dell'integrazione rispetto alle principali stime obiettivo delle due indagini e devono essere definite le metodologie per garantire la coerenza tra le stime ufficiali delle due indagini e quelle ottenibili del dataset integrato.

Per quanto riguarda l'obiettivo principale dell'integrazione, ossia lo studio delle relazioni tra le variabili rilevate separatamente alle due indagini, sviluppi futuri potrebbero indagare come tener conto di evidenze note rispetto a tali relazioni nella fase di linkage. Inoltre, le metodologie per la produzione di stime a partire dai dati abbinati devono tenere conto del processo statistico di integrazione che ha prodotto i dati e valutare quindi l'impatto delle operazioni di linkage e degli errori ad esse connessi sulle relazioni tra variabili oggetto di interesse.

Un ulteriore sviluppo è legato al confronto tra i risultati forniti dalle metodologie impiegate e quelli ottenibili con altri metodi. Solo a titolo di esempio, a livello macro, si potrebbero confrontare le stime ottenute dal dataset integrato con quelle fornite dall'applicazione di metodologie di statistical matching; mentre a livello micro, sarebbe interessante confrontare i risultati del record linkage con quelli derivanti dall'uso di tecniche di imputazione.

Infine, dato il notevole interesse sia a livello accademico che a livello Eurostat per i risultati delle indagini in esame, potrebbe essere studiato e implementato un piano di rilascio, per finalità di ricerca scientifica, delle informazioni contenute nel dataset integrato.

Riferimenti Bibliografici

Bethel J. (1989). Sample allocation in multivariate surveys, Survey methodology, 15, pp. 47-57

Breiman, L., J. H. Friedman, R. A. Olshen, and C. G. Stone. (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California, USA.

Cibella N., Fortini M., Spina R., Scannapieco M., Tosco L., Tuoto T. (2007). "Relais: An open source toolkit for record linkage", Rivista di Statistica Ufficiale n. 2-3/2007, pp.55-68

Chambers, R. (2009). Regression Analysis Of Probability-Linked Data. Official Statistics Research Series 4.

Chipperfield, J. O., Bishop, G. R . and Campbell P. (2011). Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data Survey Methodology, June 2011 13 Vol. 37, No. 1, pp. 13-24

Deville, J.-C., Sarndal, C.-E. (1992). Calibration estimators in survey sampling, Journal of the American Statistical Association 87: 376–382.

D'Orazio, M., Di Zio, M. e Scanu, M. (2006)a, "Statistical Matching for Categorical Data: displaying uncertainty and using logical constraints", Journal of Official Statistics, vol. 22, n. 1, pp. 1-12.

D'Orazio, M., Di Zio, M. e Scanu, M. (2006)b Statistical Matching: Theory and Practice, Wiley.

Fellegi, I.P., Sunter, A.B. (1969). "A Theory for Record Linkage", Journal of the American Statistical Association, 64, pp. 1183-1210.

Jaro, M. A. (1989). "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida", Journal of the American Statistical Society, 84 (406), pp.414–20.

Eurostat (2008), Information society: ICT impact assessment by linking data from different sources (Final report).

Leewen, van G. (2008), ICT, innovation and productivity, in: Eurostat Information society: ICT impact assessment by linking data from different sources (Final report).

Moriarity C. Scheuren F. (2003) "A Note on Rubin's Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputation", Journal of Business and Economic Statistics, 21, 65–73

Oecd (2010), Are ICT users more innovative? An Analysis of ICT-enabled innovation in Oecd firms, Oecd, Paris.

Okner B.A., (1972) "Constructing a new data base from existing microdata sets: The 1996 merge file", Annals of economic and social movements, 1, 325-362

Paass G. (1986) "Statistical match: evaluation of existing procedures and improvements by using additional information.", in Microanalytic Simulation Models to Support Social and Financial Policy, editors Orcutt G H e Quinke H, Elsevier Science, 401-422

Rassler S. (2002) Statistical Matching: a frequentist theory, practical applications and alternative Bayesian approaches, Springer

Singh A.C., Mantel H., Kinack M., Rowe G. (1993) "Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption", Survey Methodology, 19, 59–79

Norme redazionali

La Rivista di statistica ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche corredati da una nota informativa dell'autore contenente attività, qualifica, indirizzo, recapiti e autorizzazione alla pubblicazione. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di due referenti scelti tra gli esperti dei diversi temi affrontati.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file RSU stili o nella classe LaTex, entrambi disponibili on line. La lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 35 pagine. Una volta che il lavoro abbia superato il vaglio per la pubblicazione, gli autori sono tenuti ad allegare in formato originale tavole e grafici presenti nel contributo, al fine di facilitare l'iter di impaginazione e stampa. Per gli standard da adottare nella stesura della bibliografia si rimanda alle indicazioni presenti nel file on line.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 120 parole); quelli in italiano dovranno prevedere anche un abstract in inglese.

Nel testo dovrà essere di norma utilizzato il corsivo per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

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

La proprietà letteraria degli articoli pubblicati spetta alla Rivista di statistica ufficiale. È vietata a norma di legge la riproduzione anche parziale senza autorizzazione e senza citarne la fonte.

Per contattare la redazione o per inviare lavori: rivista@istat.it. Oppure scrivere a:
Segreteria del Comitato di redazione delle pubblicazioni scientifiche
all'attenzione di Gilda Sonetti
Istat
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012014003000000

ISSN 1828-1982

€ 15,00