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

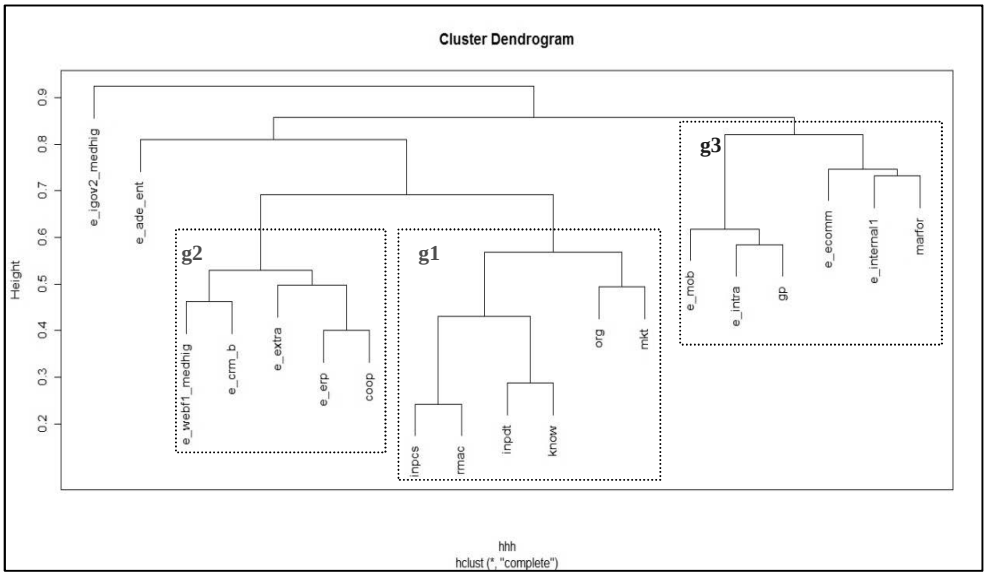
		Totale imprese			% di utilizzatori ICT per		
		Non inno	Inno	Totale	Totale	*di cui:*	
Variabili ICT: l'impresa utilizza						*non inno*	*inno*
e_intra	rete intranet	1082	2549	3631	36,7	29,8	70,2
e_extra	rete extranet	688	1853	2541	25,7	27,1	72,9
e_mob	connessione Internet mobile	952	2278	3230	32,7	29,5	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.

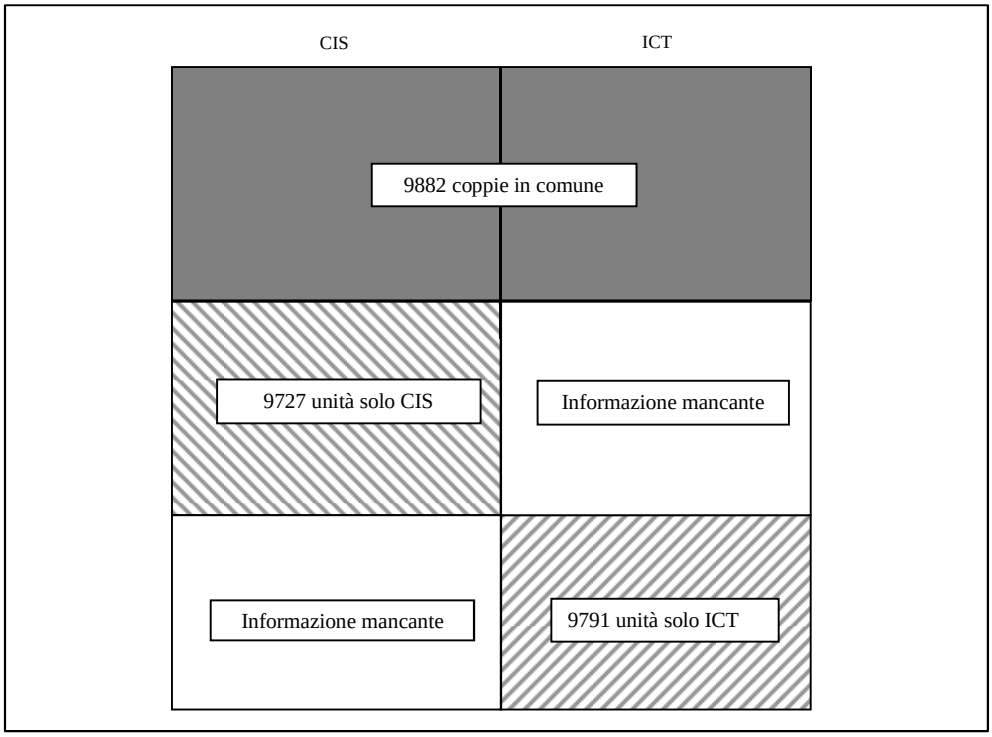
7. Sintesi dei risultati e valutazione della qualità

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.

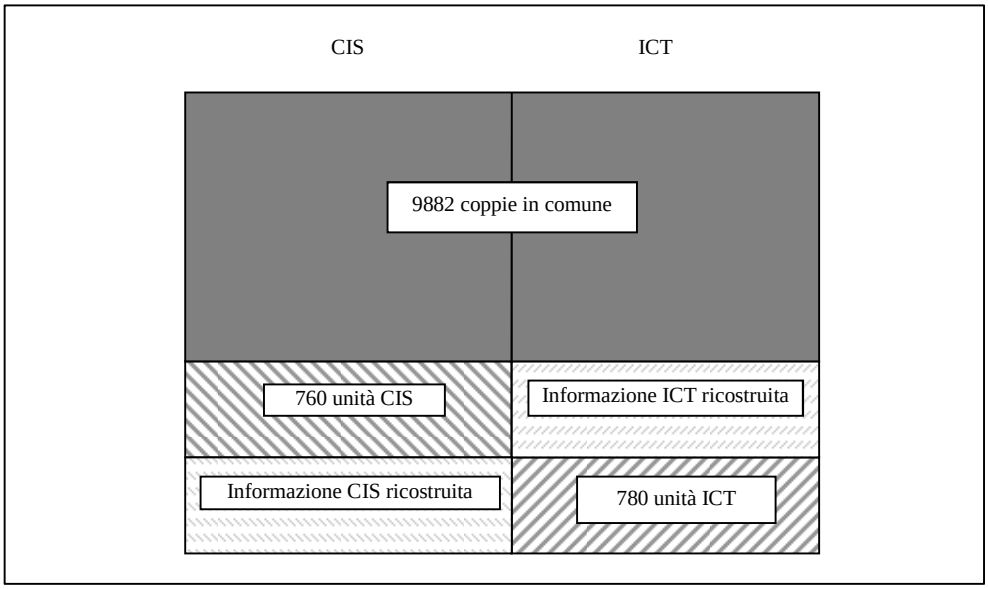
Figura 7.1. La disponibilità di dati dopo l'abbinamento deterministico per codice impresa.



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.

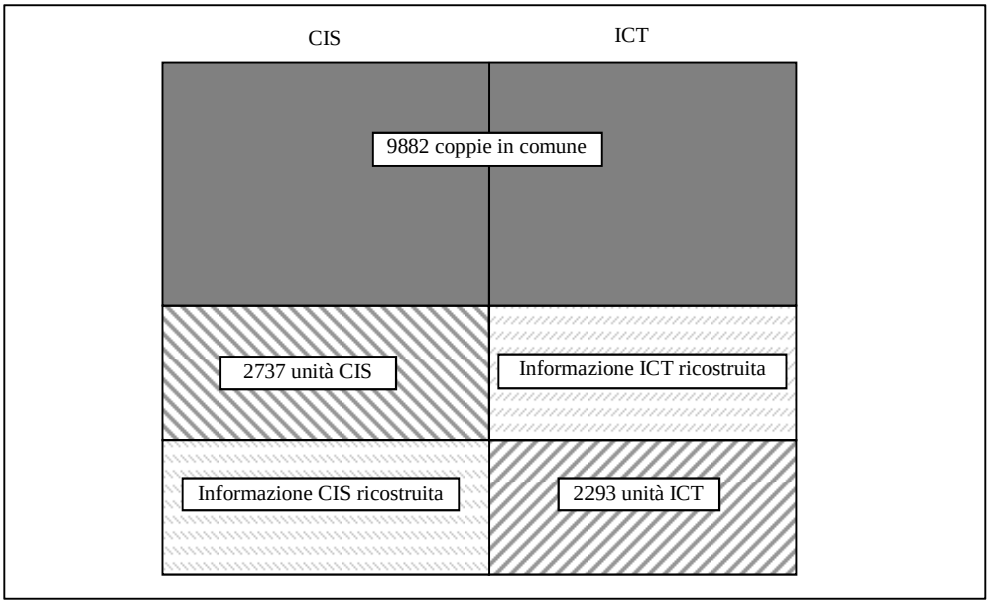
La probabilità di corretto abbinamento ed altre misure che valutano la qualità dei risultati del linkage (come ad esempio il tasso di mancato abbinamento e il tasso di falso abbinamento) devono giocare un ruolo fondamentale nelle successive analisi statistiche che sul dataset completo si intendono effettuare. 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.

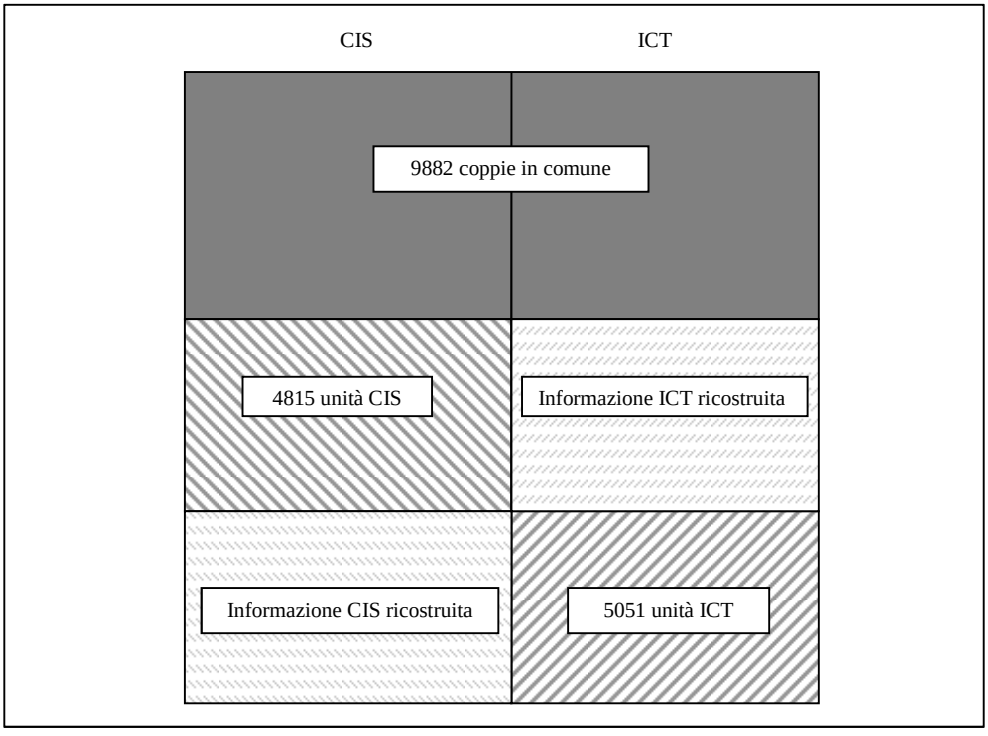


La stima della probabilità di corretto abbinamento per questo dataset è 0.96: ancora una volta, per le 9882 coppie costituite effettivamente dalla stessa unità è stata considerata una probabilità di corretto abbinamento pari a 1 mentre per le restanti 5030 coppie, ricostruite attraverso le operazioni 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.

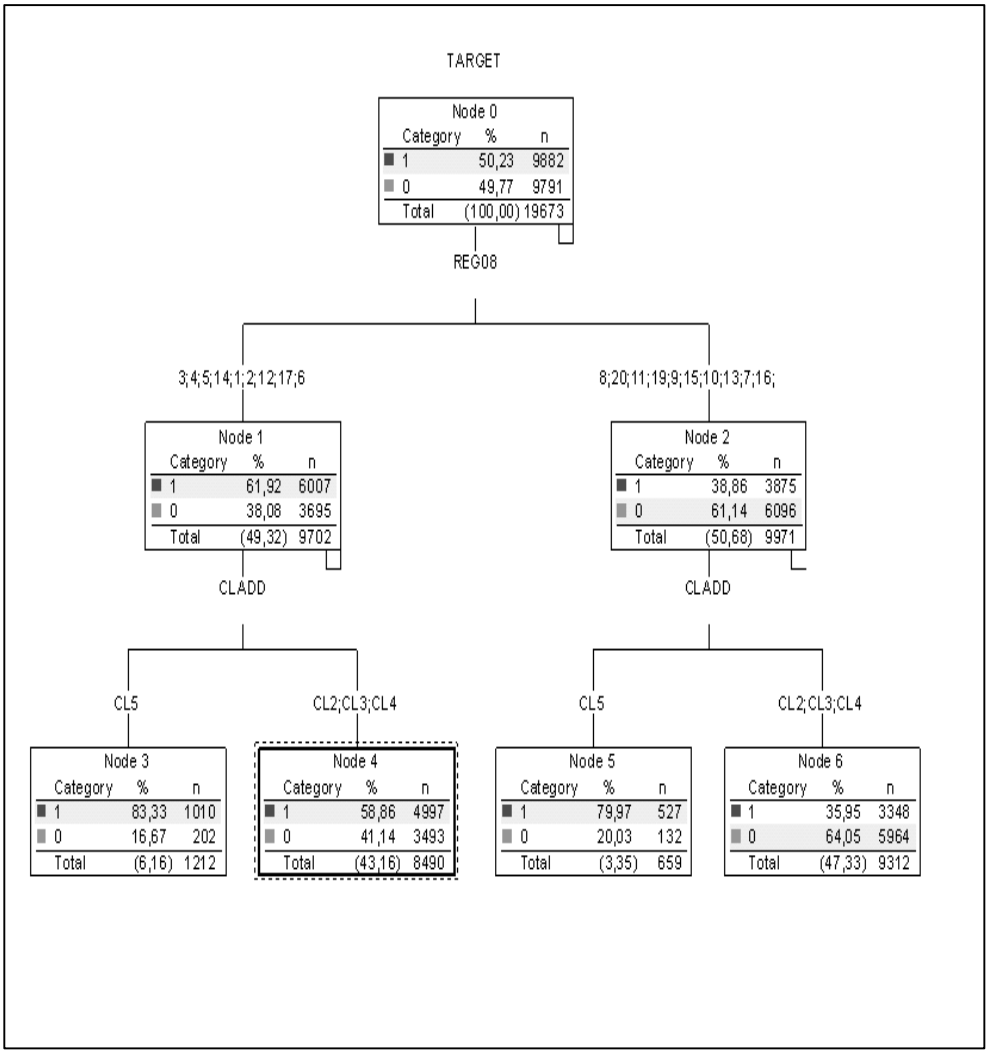
Non si apprezzano significative differenze nelle variabili esplicative per le due indagini.

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