Gruppo di lavoro per la sperimentazione di un prototipo di una procedura di abbinamento esatto per l'analisi della continuità (delibera 287/P/2000)

Relazione finale

Demografia d'impresa: l'utilizzo di tecniche di abbinamento per l'analisi della continuità

Patrizia Cella (Istat – Servizio ARC)
Giuseppe Garofalo (Istat – Servizio ARC)
Adriano Paggiaro (Università di Padova)
Nicola Torelli (Università di Trieste)
Caterina Viviano (Istat – Servizio ARC)

Roma Dicembre 2001

Indice

1. Introduzione

Il presente rapporto contiene una sintesi dei risultati prodotti dal "Gruppo di lavoro per la sperimentazione di un prototipo di una procedura di abbinamento esatto per l'analisi della continuità (delibera 287/P/2000)".

Obiettivo del gruppo di lavoro è stato quello di condurre analisi a carattere preliminare e esplorativo, di definire i meriti e le opportunità di utilizzare procedure di abbinamento esatto o di *Record Linkage* (RL) non deterministiche per condurre studi sulla dinamica delle popolazioni di imprese definite nell'archivio ASIA.

La necessità di utilizzare tecniche di abbinamento esatto è legata alla esigenza di individuare legami fra soggetti giuridici, formalmente distinti, che in realtà rappresentano una unica organizzazione economica ovvero che definiscono flussi spuri di entrate/uscite. Infatti molto spesso le modificazioni che si osservano in archivi amministrativi o giuridici non derivano da cambiamenti reali ma dall'adeguarsi dei soggetti giuridici a leggi e a regole amministrative. Gli eventi che determinano modifiche nell'esistenza delle unità possono essere inquinati dalla presenza di flussi non reali determinati da modifiche esclusivamente amministrative. E' di fondamentale importanza quindi identificare quando l'osservazione di un evento modificativo determina un evento demografico (*discontinuità statistica*) ovvero quando il verificarsi dell'evento non cambia l'identità essenziale di una unità (*continuità statistica*).

L'individuazione di eventi demografici non reali è operazione non semplice. Da un lato per l'individuazione dei legami non si possono utilizzare codici univoci (ad es. il codice fiscale che è proprio di una unità giuridica) e quindi è necessario sviluppare tecniche che abbinino caratteri identificativi delle unità giuridiche quali denominazione, indirizzo, numero di telefono, ecc… o che sfruttino particolari informazioni residenti negli archivi amministrativi quali la struttura dei soci e delle cariche che ricoprono nelle società. D'altro lato la valutazione se un abbinamento fra due differenti unità giuridiche (di cui una cessa e l'altra è una "nuova" attività) identifica un falso flusso demografico è operazione complessa e delicata che deve evitare – o mantenere all'interno di determinati limiti – l'errore di classificare come uguali due unità quando esse rappresentano realmente organizzazioni economiche distinte.

Le tecniche di abbinamento esatto di informazioni contenute su due o più *file* che si riferiscono ad una stessa popolazione, sono state sviluppate a partire dalla fine degli anni 70' sulla base della formalizzazione di Fellegi e Sunter (1969). Tale tecnica statistica, poco diffusa a livello europeo, è ritenuta di grande interesse per trattare problemi di integrazione dei dati (soprattutto quelli di natura amministrativa) o per la ricostruzione di basi di dati longitudinali, o quando , come

nel caso specifico, si cercano legami tra unità che si basano su variabili identificative. Applicazioni sono state realizzate, con un certo successo, per l'accoppiamento di record di "individui" nel campo delle statistiche sanitarie e sociali, soprattutto in USA e Canada (dove per altro sono stati sviluppati a partire dai primi anni 90' alcuni software generalizzati). Solo agli inizi degli anni 90' tali procedure sono state utilizzate per accoppiamenti di record di impresa. In questo campo la maggiore difficoltà è determinata dalle caratteristiche dei caratteri identificativi usati per il matching; rispetto agli individui le imprese presentano variabili più complesse sia nella struttura che nel significato (si pensi ad esempio alla complessità della "ragione sociale" rispetto al "nome e cognome").

Pur nell'ambito di un approccio puramente sperimentale, il lavoro effettuato aveva il compito anche di sperimentare tecniche di RL su liste di dati d'impresa di una certa consistenza numerica e di valutare i limiti e i vantaggi nell'applicazione di queste tecniche.

Il rapporto è organizzato come segue. Il paragrafo 2 è dedicato alla definizione del problema; in esso vengono richiamati i principali concetti tenendo conto anche dei contesti internazionali di riferimento; si forniscono opportune definizioni di continuità per un'impresa e si danno alcune indicazioni che risulteranno utili nel valutare l'opportunità di ricorrere a tecniche di abbinamento esatto o di "record linkage" (RL). Il paragrafo 3 contiene una succinta descrizione delle problematiche legate all'adozione di procedure probabilistiche di record linkage e fornite indicazioni sulla possibilità di implementare concretamente tali procedure. Il paragrafo 4 descrive le fasi del processo di RL sperimentato in una regione italiana (Sicilia) evidenziando inoltre le problematiche connesse all'applicazione concreta della procedura descritta nel paragrafo precedente nonché alcune soluzioni adottate. I primi risultati dell'applicazione sono descritti nel paragrafo 5. Nel paragrafo 6 viene esaminata la possibilità di ricorrere ad altri metodi, sviluppati utilizzando informazioni registrate in archivi amministrativi (anagrafe tributaria e archivio persone delle CCIAA), sia per valutare i risultati delle procedure proposte sia per integrarne i risultati. Infine, nello stesso paragrafo, viene analizzato l'impatto dei legami identificati sulle poste demografiche.

2. La demografia d'impresa: concetti, tecniche e fonti disponibili

2.1. L'archivio statistico di imprese: obiettivi

La conoscenza della struttura delle unità economiche operanti nell'ambito del territorio nazionale, e del loro evolversi nel tempo, è una informazione rilevante per gli operatori economici, per i decisori, per i ricercatori. La rilevazione di tipo censuario, come "unico" strumento di conoscenza dell'universo delle unità economiche, è ormai inadatta a causa del rapido evolversi delle strutture produttive che rendono rapidamente obsolete le informazioni statistiche. Il sistema produttivo Italiano, similmente a quello degli altri paesi dell'Unione Europea, è stato caratterizzato negli ultimi anni da un costante cambiamento nell'organizzazione, nei processi, nei prodotti, caratterizzandosi per una alta movimentazione delle piccole imprese e da una ristrutturazione "continua" delle imprese di grande dimensione. Da qui la necessità di individuare uno strumento flessibile, completo e continuamente aggiornato, il Registro Statistico (RS), come punto di riferimento per il complesso delle rilevazioni e delle analisi sul sistema economico nazionale (ed europeo).

A facilitare questo importante processo innovativo (dalle rilevazioni censuarie ai registri statistici) ha contribuito in maniera determinante il <<Regolamento n. 2186/93 del Consiglio dell'Unione Europea, relativo al coordinamento comunitario dello sviluppo dei registri di imprese utilizzati a fini statistici>>, che impone a tutti i Paesi dell'Unione Europea di realizzare un registro statistico delle imprese.

Sulla base di questo regolamento l'Istat ha realizzato, a partire dalla seconda metà degli anni 90', l'archivio ASIA (Archivio Statistico delle Imprese Attive) attraverso la valorizzazione, in senso statistico, di tutte le potenzialità informative esistenti nei differenti archivi giuridici, amministrativi e di esazione gestiti da amministrazioni pubbliche o da società di servizi a carattere nazionale. Asia infatti è stato costruito come integrazione logica e fisica delle informazioni residenti in basi di dati amministrative[1] (trattate con metodologie statistiche) e delle informazioni acquisite direttamente dalle indagini statistiche, che di solito coinvolgono le imprese di dimensione medio-grande.

Gli obiettivi principali dell'archivio sono quelli di:

- *Essere centro connettore dell'insieme delle statistiche sulle imprese*, e quindi strumento intorno a cui è possibile definire un più completo sistema informativo delle statistiche economiche.

[1] I principali archivi utilizzati sono: l'Anagrafe tributaria (circa 10 milioni di record), il Registro delle Imprese gestito dalle CCIAA (oltre 6 milioni di record), gli archivi Inps (Archivio delle dichiarazioni mensili delle imprese - DM10 - circa 1,7 milioni di record, e gli archivi degli artigiani e commercianti, oltre 4 milioni di record), l'archivio delle Soc. SEAT-Pagine Gialle (oltre 3 milioni di record).

All’interno di questo sistema l’archivio è il cardine fisico e logico. Fisico in quando alle singole unità elementari si collegano le varie tipologie di informazioni rilevate o acquisite, logico poiché in esso risiedono le meta informazioni minime necessarie (definizioni delle unità, classificazioni) per la consistenza del complesso del sistema.

- *Essere la base per la costruzione delle liste* di partenza per le indagini (*target frame*) e per l'estrazione dei campioni (*sampling frame*). Deve quindi:
  - fornire un elenco dal quale possano essere estratte le liste degli indirizzi per l'avvio dei questionari di indagini;
  - fornire una popolazione di unità di riferimento per preparare piani di campionamento efficaci e seguire le traiettorie individuali delle unità (coorti);
  - fornire la base per la proiezione dei risultati delle indagini campionarie sull'insieme della popolazione;
  - consentire di evitare duplicazioni ed omissioni nella raccolta delle informazioni;
  - garantire la convergenza tra i risultati delle diverse indagini.
- *Essere strumento per l’analisi di stock e di flusso delle imprese*.

2.2. La dimensione temporale

Il fine di un registro statistico è quello di registrare correttamente le unità statistiche, le loro connessioni e le loro caratteristiche. "Correttamente" significa che tali unità, connessioni e caratteristiche devono essere il più possibile "vicine" alle unità, connessioni e caratteristiche che esistono nel mondo reale in un determinato istante temporale.

Le unità statistiche registrate si modificano nel corso del tempo, possono cambiare i loro caratteri identificativi (modificano la ragione sociale o trasferiscono le loro attività da un luogo ad un altro), le loro caratteristiche strutturali (modificano la dimensione o l'attività prevalente), sono soggette ad eventi che le coinvolgono con altre unità, si fondono o si scorporano, cessano l'attività.
Una corretta registrazione di tutte (o le più importanti) modifiche che avvengono nella vita di una unità implica una referenziazione di questi eventi con il "tempo" in cui avvengono. La dimensione temporale, e la sua corretta gestione nell’ambito delle procedure di aggiornamento di un archivio, assume una valenza fondamentale in tutti i campi di utilizzazione dell’archivio stesso:

- Esiste una grande domanda di informazioni sulla demografia d'impresa, ad esempio il numero delle nuove imprese, il numero delle fusioni, ecc. e del loro impatto nella struttura economica (espansione/recessione, concentrazione/deconcentrazione).

- Molte statistiche economiche hanno una dimensione temporale. Il trattamento delle modificazioni delle unità e dei caratteri in un registro statistico ha chiaramente un impatto su tali statistiche e sulla loro confrontabilità nel tempo.
- Il trattamento delle modificazioni coinvolge la consistenza e la comparabilità delle statistiche. Ad esempio esso influenza la relazione che esiste fra le statistiche congiunturali e quelle strutturali. Se le liste campionarie di differenti statistiche, con lo stesso periodo di riferimento, sono realizzate a date differenti, il trattamento delle modificazioni registrate in questo periodo può avere un effetto sulle statistiche prodotte.
- Gran parte delle indagini congiunturali sono strutturate come indagini per coorti. La non corretta registrazione di un evento di fusione o scorporo può determinare incrementi/decrementi spuri degli indici prodotti.

I problemi che si pongono nella registrazione in RS degli eventi di modificazioni che interessano le unità sono molteplici:

1. *Quali modificazioni registrare?* - La gestione delle modificazioni è una attività onerosa. Non tutte le modificazioni hanno la stesse caratteristiche e la stessa valenza informativa. La mancata registrazione della modifica di ragione sociale non impatta sulla struttura dell'universo e quindi sull'efficienza dei campioni, ma sicuramente pone seri problemi nella realizzazione delle indagini. La mancata registrazione di modificazioni che hanno un impatto sull'esistenza delle unità ha rilevanza sia sulla struttura dell'universo che nella produzione delle liste. In presenza di risorse limitate una scelta di priorità è necessaria. Scelte differenti possono essere fatte, non solo sulle differenti tipologie di modificazioni, ma anche per sottoinsiemi di unità: l'aggiornamento delle imprese di dimensione medio-grande è sicuramente prioritario rispetto all'aggiornamento delle imprese di dimensione minore.
2. *Come classificare le modificazioni?* – Le modificazioni hanno significati differenti, classificarle significa associare ad esse i differenti significati informativi e individuare le differenti metodologie di trattamento. In termini del tutto generali si individuano quattro classi di eventi che possono produrre modifiche in una unità statistica: a) modifiche nei caratteri anagrafici (ad. es. cambi nella ragione sociale), b) modifiche nei caratteri di stratificazione (cambi di dimensione, di attività), c) modifiche di esistenza (cessazioni, nascite), d) modifiche per trasformazione (fusioni, scorpori).
3. *A quale data registrare la modificazione?* - Eventi quali le nascite e le cessazioni, raramente si presentano ad un preciso istante temporale, spesso sono determinati da un insieme di eventi elementari. Gli atti che si possono associare ad una nascita sono ad es.: l'acquisizione del codice fiscale, la registrazione al Registro delle Imprese, l'acquisto di beni, l'assunzione del primo

dipendente, la prima fatturazione, ecc... . Tali eventi elementari possono avvenire anche a notevole distanza di tempo. Le fusioni e scorpori sono atti registrati e quindi sono relativi ad un ben determinato istante temporale, ma molto spesso assumono validità retroattiva o successiva al momento della registrazione.

4. *Modifiche o correzioni ?* – La gestione della dimensione temporale in un archivio implica la necessità di non confondere le modifiche che avvengono su di una unità con le "correzioni" di errori presenti in tale unità. Le correzioni devono essere classificate separatamente per evitare che i saldi, che da esse derivano, fra strutture di dati a tempi differenti siano interpretati in maniera erronea.
5. *Modifiche reali o spurie ?* – Molto spesso le modificazioni che sono osservate non derivano da cambiamenti reali ma dall'adeguarsi dei soggetti legali a leggi e a regole amministrative. Questo è vero per le modificazioni di alcuni caratteri, ad es. una modifica di attività dichiarata in un registro amministrativo può essere determinata dalla necessità di acquisire benefici in termini fiscali. Ma sono soprattutto gli eventi demografici, che determinano modifiche nell'esistenza delle unità, che possono essere inquinati dalla presenza di flussi "spuri" determinati da modifiche esclusivamente amministrative.

2.3. La demografia d'impresa

2.3.1. Le definizioni di nascita e di cessazione di una impresa

La crescente domanda di informazioni sui meccanismi che alimentano la crescita economica ha reso più pressante il bisogno di statistiche sulla demografia delle imprese. L'analisi della demografia industriale occupa tradizionalmente un ruolo di primo piano nella analisi della concorrenza e della produzione (formazione dell'offerta e dei relativi prezzi, rinnovamento della struttura produttiva,..) e negli studi sulla creazione/distruzione di posti di lavoro e della mobilità dei lavoratori. Più recentemente la demografia d'impresa è stata uno strumento per l'analisi del successo e della sopravvivenza nelle nuove imprese e delle caratteristiche della nuova imprenditorialità.

Per demografia d'impresa si intende l'analisi statistica della dimensione e della composizione di una popolazione di unità statistiche (le imprese) ad un dato istante temporale e il suo sviluppo in un dato periodo con riferimento ad un dato ambito territoriale (o settoriale o dimensionale).

Per fare analisi demografica fonti naturali e preferenziali, così come anche indicato dai regolamenti europei, sono gli archivi di impresa ed in particolare gli archivi statistici. L'Archivio Statistico delle Imprese Attive (ASIA) è quindi la fonte di informazione per l'analisi statistica della popolazione di imprese e dei flussi demografici che la interessano.

Il legame tra stock e flussi è stabilito da una equazione base che deve essere soddisfatta e in cui devono essere esaminate le componenti demografiche:

$$Popolazione(t)=Popolazione(t-1)+Entrate(t)-Uscite(t)$$

I concetti di nascita di una impresa e di cessazione devono necessariamente tenere conto della definizione di impresa.

Nell'ambito del sistema statistico europeo l'impresa è definita come: "*la più piccola combinazione di unità legali che corrisponde ad una unità organizzativa per la produzione di beni e servizi, la quale beneficia di un certo grado di autonomia decisionale specialmente nella allocazione delle sue risorse correnti. Una impresa svolge una o più attività in uno o più luoghi. Una impresa può corrispondere a una sola unità legale*"[2]. I tre elementi fondamentali che caratterizzano in senso statistico l'impresa sono quindi: 1) avere un (*o più*) supporto giuridico, 2) essere una organizzazione che gode di una autonomia, 3) produrre beni o servizi.

L'esistenza di una unità legale è quindi condizione necessaria ma non sufficiente per identificare una impresa. La necessità di distinguere il concetto statistico di "impresa" da quello giuridico di "unità legale" è legata alla esistenza di unità, che se pur correntemente e legalmente registrate negli archivi amministrativi o giuridici, non rappresentano unità che organizzano un insieme di fattori al fine di realizzare attività di produzione di beni o di servizi.

Se non tutte le unità legali sono imprese, alla nascita (cessazione) di una unità legale può non corrispondere una nascita (cessazione) di una impresa e, quindi, non tutte le iscrizioni (cancellazioni) osservabili negli archivi amministrativi possono essere considerati flussi di entrata (uscita) nel Registro statistico

Sulla base di queste considerazioni la definizione europea[3] di nascita di una impresa è la seguente: "*una nascita corrisponde alla creazione di una combinazione di fattori di produzione con la restrizione che nessun'altra impresa è coinvolta nell'evento*"; similmente la definizione di cessazione di una impresa è: "*una cessazione corrisponde alla dissoluzione di una combinazione di fattori di produzione con la restrizione che nessun'altra impresa sia coinvolta nell'evento*".

[2] Regolamento del Consiglio (CEE) n° 696/93 sulle unità statistiche per l'osservazione e l'analisi del sistema produttivo della comunità

2.3.2. Il concetto di continuità di una impresa

Le definizioni richiamate si basano sul concetto di identità di una unità e quindi, nell'ambito del dinamismo delle popolazioni di imprese, è di fondamentale importanza identificare quando l'osservazione di un evento modificativo determina un evento demografico (discontinuità) ovvero quando il verificarsi dell'evento non cambia l'identità essenziale di una unità (continuità).

"*Si ha continuità quando una impresa si modifica senza effetto per la sua identità essenziale determinata dai propri fattori di produzione*"; i fattori di produzione sono l'insieme dei mezzi (management, lavoro, terra, capitale, impianti, processi, edifici…) che l'impresa utilizza per la sua azione produttiva. Il concetto di continuità, sviluppato nel 1993 dagli statistici olandesi Struijs e Willeboords, ha sostanzialmente modificato l'approccio classico alla demografia d'impresa individuando un criterio teorico sulla base del quale distinguere i flussi reali da quelli spuri e quindi operare una corretta registrazione degli eventi demografici in un archivio di imprese.

E' evidente che da un punto di vista pratico analizzare tutti (o la maggior parte) dei fattori di produzione e pesarli è estremamente difficoltoso e costoso. Per questa ragione Eurostat suggerisce, come criterio pratico, di utilizzare specifici caratteri, disponibili nell'archivio di imprese, che sono correlati ai più importanti fattori di produzione. L'ipotesi di base è quella che ad un cambiamento di questi caratteri corrisponde un cambiamento nei fattori di produzione. I caratteri presi in considerazione sono tre:

- L'unità legale che controlla l'impresa: la continuità nell'unità legale è assunta come positivamente correlata con la continuità del *management* di una impresa.
- L'attività economica svolta dall'impresa: la continuità dell'attività prevalentemente svolta può essere assunta come positivamente correlata alla continuità di una serie di fattori quali l'occupazione, gli impianti, ecc.. .
- La localizzazione dove l'attività viene svolta: la continuità nella localizzazione è ovviamente collegata alla continuità degli edifici e della terra usati dall'impresa.

La regola empirica proposta è che un' impresa è considerata una nuova unità (discontinuità) se almeno due su tre dei precedenti caratteri si modificano.

Le regole di continuità riflettono una nozione di identità che trae fondamento dal considerare l' impresa come un insieme specifico di risorse, procedure e relazioni con l'ambiente. Eurostat suggerisce opportunamente di introdurre discontinuità solo quando i mutamenti sono di "grande portata" e rapidi.

---

[3] Regolamento europeo n° 2700/98

2.3.3. Il progetto europeo sulla demografia d'impresa

Eurostat ha avviato, a partire dal 2000, un progetto che ha lo scopo di ottenere statistiche confrontabili sulla *Business Demography* all'interno dei Paesi dell'Unione Europea. A causa della numerosità e complessità degli argomenti coinvolti e della non facile attività di armonizzazione, il progetto è stato suddiviso in più passi. Il piano di lavoro proposto, discusso all'interno di un apposito *working group*, è stato organizzato individuando gli obiettivi di breve, medio e lungo termine.

Sono stati considerati obiettivi di breve termine, da realizzarsi entro il mese di Agosto del 2002, la produzione di statistiche sullo stock di imprese attive e sulle nascite di imprese reali, nonché il calcolo dei tassi di sopravvivenza e di misure sulla crescita delle nuove imprese. Obbiettivi di medio termine sono il calcolo di indicatori sulle cessazioni di impresa e l'analisi di eventi quali le fusioni e gli scorpori. Infine, è stato considerato obbiettivo di lungo termine l'analisi dei fattori che determinano il successo/insuccesso della nuova imprenditorialità.

Prima di raccogliere i dati armonizzati sulle variabili relative alle priorità di breve termine, è previsto lo svolgimento di uno studio di fattibilità da concludersi entro fine 2001. Lo scopo di questo studio è quello di testare la metodologia proposta e di evidenziare le aree problematiche.

Uno degli obiettivi principali del progetto Eurostat è quello di identificare le nascite reali di impresa. Nella definizione di nascita di una impresa è sottolineato che la creazione di una nuova impresa si verifica se nessun'altra impresa è coinvolta dall'evento.

La metodologia Eurostat è centrata sulla identificazione delle nuove imprese (le nascite reali sono una sottopopolazione delle nuove imprese) e contemporaneamente sulla identificazione delle altre imprese coinvolte nell'evento. La base informativa su cui impiantare la metodologia fa uso degli stock di imprese attive N_t (t=anno) e dei flussi di nuove imprese in t definite come quel sottoinsieme di unità in N_t che hanno intrapreso l'attività tra 01.01. e 31.12 dell'anno t. In pratica, le nuove imprese (E) sono quel sottoinsieme di unità che, confrontando lo stock t con quello di t-1, sono presenti (e attive) solo in t.

Per identificare le nascite reali all'interno delle nuove imprese la metodologia propone di confrontare alcuni caratteri delle nuove imprese con quelli delle imprese attive a N_{t-1}. Ad esempio se una nuova impresa si stabilisce nello stesso luogo e svolge la stessa attività di una impresa non più attiva, allora la nuova impresa non viene considerata una nascita reale.

L'approccio metodologico proposto è coerente con quanto delineato dalle regole di continuità di una impresa, regole che utilizzano appunto combinazioni dei caratteri di impresa per identificare

i legami di continuità e quindi depurare i flussi spuri (le nuove imprese collegate per date combinazioni di caratteri a imprese attive già esistenti) e ricavare solo i flussi reali.
In modo schematico, il processo proposto è sintetizzato nella seguente tabella:

Progetto Eurostat -Processo di identificazione delle nascite reali

Popolazione	**Informazioni Utilizzate**	**Numero di imprese**
Imprese attive nell'anno t	Fatturato/dipendenti	N_t
Imprese attive nell'anno t-1	Fatturato/dipendenti	N_{t-1}
Nuove imprese nell'anno t – Entrate	Confronto di N(t) con N(t-1)	E_t
Sottopopolazioni per il matching	Localizzazione e Settore	X_1
	Nome e localizzazione	X_2
	Nome e Settore	X_3
	Altre informazioni	X_4
	Controlli manuali sulle grandi imprese	X_z
Nascite reali		**$RE= E-f(X_1…X_z)$**

2.4. La demografia d'impresa: l'approccio italiano proposto

L'impianto definitorio descritto non può dare da solo soluzione al problema dell'individuazione dei legami fra unità formalmente diverse ma statisticamente continue. E' necessario, da una parte, far discendere convenzioni appropriate (a quale livello della classificazione di Ateco una modifica di attività produce discontinuità, in quale ambito territoriale minimo è discontinua una modifica di localizzazione, ecc.), dall'altra individuare un percorso metodologico che porti dagli eventi osservabili (registrazioni amministrative) agli eventi statistici definiti.

2.4.1. Alcuni elementi critici sulle regole di continuità

Le regole empiriche proposte dall'Eurostat per l'analisi della continuità hanno sicuramente il pregio di essere facilmente applicabili ai dati registrati in un archivio statistico, ma d'altro canto esse sembrano troppo "generali" per essere applicate ad una realtà complessa come quella delle imprese. In particolare i criteri proposti sembrano dare lo stesso peso alle modifiche che avvengono

nei tre caratteri considerati; è necessario sottolineare, al contrario, che ciascuna modificazione deve essere valutata nello specifico contesto (sottopopolazioni di unità) in cui avviene. Ad esempio, viene sottolineato che una modifica nell'unità legale non è sufficiente da sola a determinare discontinuità. Questo è vero sia per una grande società sia per una piccola impresa, gestita da un imprenditore individuale? Inoltre, con riferimento alla localizzazione, una modificazione ha lo stesso significato per una impresa plurilocalizzata e per una monolocalizzata?

Con riferimento a questa considerazione generale, di seguito, si descrivono alcune considerazioni critiche e si ridefiniscono alcune regole empiriche.

Unità legale. E' la variabile più importante per l'identificazione della discontinuità. Infatti i cambiamenti di natura giuridica o i passaggi di proprietà avvengono molto frequentemente nel corso della vita di una impresa. I principali aspetti che si evidenziano sono due:

1) In alcuni ambiti di studio, quali le nascite e le cessazioni delle piccolissime unità, sembra non realistico scindere il soggetto giuridico (imprenditore) da quello statistico (impresa): nelle imprese unipersonali esiste una identificazione fra imprenditore e impresa. Per questa ragione una modifica di unità legale (nuovo imprenditore) deve essere identificata come un fattore di discontinuità dell'impresa anche in presenza di continuità negli altri due caratteri.
2) Un'altra forte caratteristica di discontinuità esiste in presenza di un cambiamento di unità legale da impresa individuale a società di capitali. In questo caso l'introduzione di un nuovo fattore di produzione (il capitale) produce forti elementi di discontinuità nell'organizzazione di una impresa.

Attività economica. La convenzione proposta di considerare un elemento di discontinuità ogni modificazione a livello di 4 digit di NACE sembra essere troppo generale. Da un lato esistono attività economiche consistenti fra di loro, altre volte modifiche nei beni o servizi realizzati si presentano senza una corrispondente modificazione del processo produttivo. D'altro lato è importante sottolineare che alcune modifiche di attività (ad esempio dalla produzione al commercio) da sole possono rappresentare una forte discontinuità per l'impresa.

Localizzazione. Eurostat suggerisce di considerare la modifica nella localizzazione delle attività produttive di una impresa un elemento di discontinuità solo se avviene nell'ambito di grandi distanze. Questa raccomandazione da un lato non specifica a quali ambiti territoriali fare riferimento (uno spostamento fra comuni deve essere visto come "grande distanza"? ma questo equivale a considerare allo stesso livello piccoli territori comunali e le grandi aree metropolitane), dall'altro

non considera che il concetto di continuità nella localizzazione di una impresa è strettamente correlato con il tipo di attività che l'impresa svolge. Infatti mentre per alcune attività economiche, ad es. agricoltura, estrazione di minerali, la produzione è strettamente collegata con il luogo in cui avviene, per altre attività, caratterizzate o da un alta mobilità nel territorio (intermediari del commercio, piccoli trasportatori) o da una stretta corrispondenza fra luogo di residenza e luogo di lavoro (liberi professionisti), la localizzazione dell'attività ha poco potere nell'identificare una impresa. Infine, una modifica nella localizzazione per le imprese che organizzano la produzione in più unità fisiche non interrompe la continuità. Per questa tipologia di imprese, comunemente denominate plurilocalizzate, la sede principale è statisticamente individuata dal luogo dove si organizza la produzione. Un trasferimento della sede principale coincide, quindi, con il trasferimento delle sole attività amministrative-gestionali.

2.4.2. Le tecniche statistiche utilizzabili.

In repertori statistici, quale ad esempio ASIA, che traggono alimento da fonti amministrative, la definizione di continuità nel tempo tra entità statistiche presuppone che siano identificate le relazioni dinamiche esistenti tra unità amministrative apparentemente diverse. L'individuazione di tecniche statistiche applicabili è la condizione necessaria perché i criteri di continuità definiti possano essere identificati e registrati. La metodologia proposta dall'Eurostat, se garantisce una facile applicabilità delle regole di continuità in tutti i paesi europei, non garantisce la mancanza di duplicazioni nei legami, un livello accettabile di errore nei falsi-match e propone una attività manuale (e quindi costosa) nell'individuazione dei legami per le imprese con più di 20 addetti. D'altro canto la tecnica proposta (similitudine nei caratteri) non è l'unica possibile; altre tecniche possono essere utilizzate (e lo sono da differenti uffici statistici) in alternativa o in combinazione anche con riferimento alle peculiari caratteristiche (amministrative) di ciascun paese e alla disponibilità di informazioni.

1) *Identificazione delle trasformazioni societarie mediante movimenti dei lavoratori*. Questa tecnica deduce l'esistenza di trasformazioni societarie non direttamente visibili dall'entità dei flussi di lavoratori tra due (o più) imprese. Alla sua base vi è l'ovvia considerazione che se l'impresa "b" acquista l'impresa "a" si osserverà un movimento istantaneo di (quasi) tutti i lavoratori di "a" verso "b". L'applicazione di questa tecnica, utilizzata in Canada e in Danimarca e che ha avuto le prime sperimentazioni in Italia nei primi anni novanta sulla base di dati INPS, è oggi possibile (e generalizzabile) grazie alle informazioni disponibili da fonte

INAIL nell'ambito della procedura DNA (Denuncia Nominativa degli Assicurati)[4]. Tramite questa procedura le imprese devono dichiarare all'INAIL, entro 24 ore, le caratteristiche anagrafiche dei lavoratori assunti e di quelli che cessano l'attività lavorativa. Avere a disposizione, in tempo reale, informazioni congiunte sull'impresa e sul lavoratore, ambedue correttamente identificabili grazie al Codice Fiscale, permette un'applicazione relativamente semplice di questa tecnica. I limiti nell'utilizzo della tecnica dei movimenti dei lavoratori per l'identificazione di legami fra imprese, risiedono nella sua inapplicabilità per le imprese di minore dimensione per le quali, a causa del limitato numero di dipendenti, è impossibile discriminare fra i "movimenti fisiologici", derivanti da scelte dei lavoratori, e quelli "spuri", indotti da transazioni tra imprese.

2) *Le informazioni disponibili nell'Anagrafe Tributaria*. Ai fini della liquidazione dei debiti/crediti fiscali l'Anagrafe Tributaria registra tutta una serie di informazioni relative ai processi di fusione/scorporo, alle modifiche di natura giuridica (da impresa individuale a società e viceversa), alle successioni ereditarie. La copertura di questi eventi è limitata a chi richiede una continuità di alcune competenze economiche (debiti e crediti fiscali), ma proprio perché sono legati a transazioni di tipo monetario gli eventi registrati rappresentano legami certi e quindi individuano una continuità statistica fra due unità giuridiche.
3) *La copresenza dei titolari.* Le Camere di Commercio, nell'ambito dei registri delle imprese da loro gestiti, acquisiscono le informazioni relative ai soci delle imprese e alle loro cariche sociali. Quando la presenza di uno stesso soggetto fisico, in più di una impresa, è accompagnata da altri indicatori, quali la sua carica di amministratore, o quando gli stessi soci (o la maggioranza di essi) sono presenti in imprese che svolgono la stessa attività economica, si è in presenza di un forte segnale di "duplicazione" (più unità legali a cui corrisponde un'unica unità organizzativa) o, nel caso di cancellazione/iscrizione delle unità, di continuità. Questa tecnica permette anche di individuare le modifiche di natura giuridica quando ad una impresa individuale che cessa corrisponde la nascita di una società in cui il "socio di riferimento" è il vecchio imprenditore.
4) *La somiglianza degli attributi.* Questa tecnica costruisce legami fra unità sulla base della somiglianza nei valori di alcuni caratteri posseduti dalle imprese collegate. I caratteri maggiormente utilizzati sono la ragione sociale, l'indirizzo, il codice di attività economica, la dimensione, il numero di telefono. La concreta applicazione di questa tecnica, suggerita dall'Eurostat in quanto più vicina ai criteri operativi per l'analisi della continuità, presuppone lo sviluppo di metodologie di *linkage* che utilizzano come variabili di confronto stringhe di

[4] D. Lgs n. 38/2000

caratteri alfanumerici. La complessità della sua applicazione risulta evidente soprattutto in presenza di strutture complesse nei caratteri quali la ragione sociale di una impresa.

2.4.3. La registrazione della continuità in un archivio statistico

Il concetto di continuità di una impresa nasce dalla necessità di definire un evento demografico, nascita/cessazione di una unità statistica, indipendentemente dall'evento osservato, nascita/cessazione di una unità legale. A partire da questa esigenza teorica vengono definite una serie di linee guida più o meno coerenti per la sua applicazione, e metodologie, più o meno semplici da applicare, per una sua identificazione.

La funzione di un archivio statistico è sicuramente quella di descrivere e analizzare l'evolversi delle popolazioni di unità statistiche, ma tale funzione non deve entrare in conflitto con la necessità di poter dare informazioni differenti per esigenze differenti.

Decisori, economisti, analisti di impresa e del lavoro esprimono esigenze conoscitive differenti e a volte contrastanti. Forzare queste esigenze in un unico quadro definitorio, e quindi informativo, è una operazione non utile. Uno scorporo di attività, il subentro nella gestione di una impresa, la modifica di una attività economica possono o non possono essere visti come fattori di discontinuità a seconda dell'ottica in cui ci si pone e delle esigenze informative connesse: evoluzione della concentrazione, analisi della nuova imprenditorialità, analisi dei processi di verticalizzazione delle attività.

In una qualche misura l'archivio deve essere "neutrale" rispetto al complesso dei bisogni conoscitivi, deve essere capace cioè di classificare le unità, le relazioni, i caratteri in modo tale che qualsiasi utente possa trarre l'informazione di cui necessita.

Per queste ragioni la scelta effettuata, nell'ambito dell'archivio ASIA, non è stata quella di ricostruire in maniera fittizia unità "teoriche" ma quella di classificare e registrare i legami individuati tra unità. Le informazioni disponibili sono il tipo di legame (ad es. fusione/scorporo, modifica di natura giuridica, subentro, duplicazione, ecc.), i codici delle imprese coinvolte nei legami, la data di riferimento del legame, la data di individuazione del legame, la tecnica usata per individuare il legame (ad es. somiglianza nei caratteri, copresenza dei titolari,ecc..), la fonte da cui è stato acquisito il legame (ad es. INAIL, Camere di Commercio, Indagine statistica, ecc..)

3. Le tecniche di abbinamento esatto

Nel presente paragrafo si fornirà una sintetica rassegna di alcune delle problematiche essenziali riguardo le tecniche di abbinamento esatto di record di due o più archivi. Si forniranno, inoltre, una prima descrizione di una procedura informatica che costituisce il nucleo della complessa procedura considerata al successivo paragrafo.

3.1 Aspetti generali

Le tecniche di abbinamento di informazioni individuali riferite agli elementi di una stessa popolazione contenute in due o più archivi consentono di sfruttare in modo più penetrante informazioni correntemente raccolte in indagini statistiche o presenti in archivi amministrativi, anche al fine di ottenere letture integrate dei fenomeni e della loro evoluzione.

In particolare, l'impiego di tecniche di abbinamento esatto viene invocato sempre più spesso come parte integrante del processo di collezione e organizzazione dei dati e assume rilievo nella fase di controllo delle operazioni sul campo in indagini complesse. L'abbinamento esatto di record è, ad esempio, un passaggio irrinunciabile per la costruzione di archivi longitudinali oppure è operazione preliminare per le analisi del grado di copertura in indagini totali. L'integrazione di più archivi amministrativi è quindi uno dei più interessanti ambiti ove tali tecniche costituiscono uno strumento essenziale.

Molto spesso, il problema di abbinamento esatto può essere risolto banalmente attraverso l'uso di efficienti procedure informatiche volte a ricercare record che nei due archivi presentino lo stesso codice identificativo. Di maggiore rilievo è però il caso in cui non esiste un identificativo univoco e misurato senza errore per ciascun record; in tal caso il problema di abbinamento esatto diviene un problema di decisione che va risolto con l'utilizzo di adeguate procedure statistiche.

Una prima formalizzazione degli aspetti statistici legati alle procedure di abbinamento esatto si ha nel lavoro di Fellegi e Sunter (1969) che raccolgono le precedenti idee di Newcombe *et al.* (1959) e Tepping (1968) e le riportano ad una struttura coerente legata alla teoria classica della verifica di ipotesi. Winkler (1995) riassume infine i più recenti sviluppi delle tecniche di abbinamento esatto, individuando le molteplici direzioni in cui è auspicabile che si concentrino le future ricerche e delineando i problemi relativi al loro impiego per la costruzione e la gestione di archivi di imprese.

3.2. Definizione del problema

Per una trattazione generale del problema dell'abbinamento esatto conviene partire dalla formalizzazione proposta da Fellegi e Sunter (1969). Due archivi *A* e *B*, di dimensione N_A e N_B, contengono rispettivamente record *a* e *b*, una parte dei quali sono relativi alle medesime unità. Se i due archivi coincidono, si è in presenza di un problema di deduplicazione, con l'obiettivo di individuare quali fra i record dell'archivio siano riconducibili ad un'unica unità.

Lo spazio prodotto $A \times B = \{(a,b); a \in A, b \in B\}$, che include tutte le $N = N_A \times N_B$ possibili coppie di record originate dal confronto, è pertanto l'unione di due insiemi disgiunti:

- l'insieme M delle coppie relative allo stesso individuo;
- l'insieme U delle coppie con record relativi a due individui differenti.

I procedimenti di abbinamento esatto di record sono essenzialmente metodi di decisione per classificare ogni coppia come appartenente ad uno dei due insiemi M ed U. Se esiste un identificatore che permette di individuare con certezza i record relativi ad ogni individuo, il procedimento si riduce ad un algoritmo di ricerca di coloro che presentano la medesima chiave di identificazione; nel caso, invece, in cui una chiave identificativa, unica per ogni record, non esista oppure sia osservabile con errore, si tratta di impostare il problema di abbinamento come un problema di decisione che auspicabilmente conduca a rendere minimo il numero di errori di classificazione. In questo contesto, le decisioni errate possono essere di 2 tipi:

- errati abbinamenti (falsi positivi): si classifica la coppia in M, essendo in realtà *a* e *b* relativi ad individui differenti;
- mancati abbinamenti (falsi negativi): si classifica la coppia in U, essendo in realtà *a* e *b* relativi al medesimo individuo.

Poiché M ed U sono insiemi mutuamente esclusivi, non è possibile minimizzare contemporaneamente il numero di mancati abbinamenti ed il numero di errati abbinamenti: ad un elevato numero di errate decisioni per una delle due classi corrisponde una diminuzione degli errori nell'altra direzione. La scelta di un metodo di abbinamento è legata pertanto alla valutazione della gravità relativa che si attribuisce ai due tipi di errore conseguenti al processo di decisione.

Nella gran parte delle situazioni applicative si ritiene che meriti un maggiore impegno l'obiettivo di evitare gli errati abbinamenti. Ciò è vero specialmente nel caso di abbinamenti di record per condurre analisi dinamiche: in tal caso, errati abbinamenti porterebbero ad associare informazioni che sono in realtà relative ad individui differenti, dando luogo a mobilità spuria.

3.2.1. Variabili di confronto e strategie di blocco

Una procedura di abbinamento conduce a stimare, per ognuna delle *N* possibili coppie di record, il valore ignoto di una variabile indicatrice *G*, che vale 1 per le coppie in M e 0 per quelle in U. A tal fine è possibile utilizzare i valori assunti nei record da alcune variabili di confronto; in particolare, Newcombe (1988) osserva che tali variabili devono avere la capacità di discriminare al meglio gli individui presenti nei due archivi, in caso di discordanza, di concordanza o, nella migliore delle ipotesi, in entrambi i casi. La variabile "sesso", ad esempio, da poche informazioni se è concordante, mentre fornisce una forte indicazione negativa sull'abbinamento se si osserva una discordanza.

Appare pertanto chiaro che, oltre alla scelta delle variabili di confronto, assume estrema importanza la definizione di concordanza che viene assegnata ad ogni possibile confronto fra le variabili. Al fine di sfruttare al meglio le informazioni provenienti dal confronto, sarebbe necessario tenere in considerazione tutte le possibili combinazioni di modalità che possono ottenersi quando si confronti la stessa variabile presente nei due archivi; è evidente che ciò è tanto più difficile quanto maggiore è il numero di modalità. Sono pertanto necessarie delle modifiche nella definizione di concordanza che permettano di aggregare quei risultati che forniscono informazioni simili sull'abbinamento; la dimensione di tale processo di aggregazione dipende essenzialmente dalla parsimonia richiesta al modello e dalle numerosità campionarie di cui si dispone. Copas e Hilton (1990) mostrano come sia possibile calcolare la perdita di informazione derivante da definizioni più restrittive, e propongono un modello di misura in forma parametrica che permette di sfruttare al meglio i risultati del confronto pur mantenendo una scelta parsimoniosa.

Fra i metodi più utilizzati, il confronto può dare semplicemente un risultato dicotomico, con valori 1 in caso di concordanza e 0 con discordanza fra le variabili; per una specificazione più dettagliata, possono essere previsti diversi livelli di concordanza (ad esempio per l'età, tenendo conto della differenza in anni), o diverse capacità discriminanti per specifici valori delle variabili (ad esempio nel caso di cognomi più o meno comuni, per cui la concordanza di cognomi diffusi fornisce minori informazioni).

Una volta scelte le variabili di confronto e le definizioni di concordanza, l'informazione ottenibile per la j-esima coppia può essere riassunta in un vettore, che ha come singolo elemento il risultato del confronto fra le i-esime variabili:

$$\gamma_j = \left[\gamma_j^1, \gamma_j^2, ..., \gamma_j^i, ..., \gamma_j^I\right], \qquad j = 1...N .$$

Il confronto effettuato su tutte le coppie di record appartenenti ai due archivi può comunque portare ad un carico computazionale molto elevato per archivi di grandi dimensioni. Se è possibile osservare variabili di confronto con elevata affidabilità ed alto potere discriminante, una buona strategia consiste nel ridurre lo spazio dei confronti all'interno di un blocco di record che presentano concordanza perfetta su tali variabili.

Il vantaggio di tale strategia è di ridurre, spesso drasticamente, il numero di confronti ammissibili, ottenendo contemporaneamente una notevole riduzione del carico computazionale e una forte protezione contro i falsi positivi; le dimensioni di tali effetti dipendono ovviamente dalla capacità discriminante delle variabili di blocco. Di contro, tale procedura può condurre ad un aumento di falsi negativi se non è elevata l'affidabilità delle variabili di blocco prescelte. La scelta dipende pertanto essenzialmente dal peso che si vuole dare ai due tipi di errori, oltre alla disponibilità di tempo e mezzi dal punto di vista computazionale; Kelley (1985) suggerisce alcuni metodi per definire una strategia di blocco che permetta di minimizzare i costi complessivi, sia computazionali che in termini di errore negli abbinamenti.

3.2.2. I pesi di abbinamento e la loro stima

Definiti i vettori di confronto γ, rimane da stabilire come questi possano essere utilizzati per la decisione sulla classificazione delle coppie in M o U. Una possibilità è l'assegnazione ad ogni vettore di un peso *w*, sul valore del quale si basa il seguente processo decisionale per la j-esima coppia:

- $w_j \geq K_u \Rightarrow (a_j, b_j) \in \mathrm{M}$ la coppia viene abbinata;
- $K_l \leq w_j < K_u$ la decisione viene rinviata; (1)
- $w_j < K_l \Rightarrow (a_j, b_j) \in \mathrm{U}$ la coppia non viene abbinata.

La stima dei pesi *w* e la scelta delle soglie *K* sono ovviamente cruciali nella definizione del procedimento. Nel caso più semplice, si può utilizzare un criterio deterministico, dove implicitamente i valori dei pesi e delle soglie sono fissati a priori in funzione degli specifici obiettivi dell'abbinamento; in tal caso, la scelta deve essere definita in base alla predisposizione verso gli errori di abbinamento, con soglie elevate che proteggono dai falsi positivi ma sono spesso associate ad un numero elevato di falsi negativi. L'intervallo tra le due soglie non deve essere inoltre troppo ampio, in quanto la scelta di rinviare la decisione, associata spesso ad un controllo manuale delle coppie, presenta solitamente costi elevati.

Un semplice esempio di criterio deterministico consiste nell'associare i pesi *w* al numero di concordanze osservate; la scelta di abbinare può avvenire ad esempio per tutte le coppie con al massimo una discordanza, con una soglia unica implicita pari ad *I* -1 nel processo decisionale (1). In alternativa, si possono utilizzare pesi differenti per le singole variabili, ammettendo ad esempio 2 errori su quelle ritenute meno discriminanti.

Nella formulazione di Fellegi e Sunter (1969) i pesi vengono invece stimati in modo probabilistico, attraverso il rapporto fra le due verosimiglianze del vettore di confronti, rispettivamente nel caso di coppie relative allo stesso individuo (M) e coppie abbinate casualmente (U):

$$w_j = \ln \frac{P(\gamma_j | \mathrm{M})}{P(\gamma_j | \mathrm{U})} = \ln \frac{m_j}{u_j} . \qquad (2)$$

Essendo *w* un rapporto di verosimiglianza, statistica sufficiente per il problema di decisione, Fellegi e Sunter dimostrano che utilizzando la (2) la regola di decisione (1) è ottimale per ogni coppia di soglie (K_l, K_u); l'ottimalità assume qui il significato di minimizzazione della regione di indecisione, ed ha come conseguenza, ad esempio, la possibilità di fissare a priori i livelli di errore desiderati, sia per quanto riguarda i falsi positivi che i falsi negativi, rendendo minimo il numero di coppie da abbinare manualmente.

Kirkendall (1985), oltre a proporre alcuni esempi pratici per il calcolo dei pesi in (2) con differenti variabili di confronto, ne propone un'ulteriore possibile interpretazione in termini di teoria dell'informazione: nel caso i logaritmi siano espressi in base 2, i pesi sono esprimibili come *odds ratios* che permettono di aggiornare l'informazione a priori attraverso i risultati del confronto.

3.2.3. La stima mediante l'algoritmo EM

Il problema principale della procedura proposta da Fellegi e Sunter è la stima delle probabilità *m* ed *u* definite in (2), la cui accuratezza condiziona fortemente la proprietà di ottimalità. Come osserva tra gli altri Winkler (1995), risulta infatti spesso irragionevole l'assunzione che esistano campioni per i quali sia certa l'appartenenza delle coppie ad U e, soprattutto, a M. Inoltre, anche se tali campioni fossero disponibili, le stime risultanti per un'applicazione potrebbero non adattarsi ai veri, ma ignoti, valori relativi al campione che si vuole effettivamente abbinare. A tal fine, sarebbe invece necessario conoscere esattamente il valore della variabile *G* per tutte le coppie da abbinare, il che non è ovviamente possibile.

Tepping (1968) propone di effettuare una partizione preliminare delle coppie negli insiemi M ed U, stimando all'interno di questi campioni le probabilità necessarie; in questo modo si potrebbe tra l'altro evitare di ricorrere, nella stima di *m* ed *u*, alle ipotesi spesso poco realistiche di indipendenza fra gli errori nelle singole variabili di confronto, necessarie per i metodi di stima proposti da Fellegi e Sunter. Seguendo Jaro (1989), è possibile effettuare una partizione simile a quella proposta da Tepping in modo iterativo, imputando ad ogni passo il valore di *G* per tutte le coppie, e ristimando le probabilità seguendo la logica dell'algoritmo EM (Dempster *et al*., 1977). In questo modo è inoltre possibile allentare l'ipotesi di indipendenza fra gli errori, in modo da tener conto delle eventuali correlazioni fra le diverse variabili (si pensi ad esempio alla stretta relazione fra "nome proprio" e "sesso").

Per poter applicare l'algoritmo, è necessario definire la funzione di verosimiglianza dei parametri *m* ed *u*, congiuntamente a *p*, la frazione di coppie da abbinare:

$$L(m,u;p)=\prod_{j=1}^{N}\left[P(\mathrm{M})P(\gamma_j|\mathrm{M})\right]^{g_j}\left[P(\mathrm{U})P(\gamma_j|\mathrm{U})\right]^{1-g_j}=\prod_{j=1}^{N}\left[pm_j\right]^{g_j}\left[(1-p)u_j\right]^{1-g_j}.$$

Se si fissano i valori dei parametri *m*, *u* e *p*, al passo *E* dell'algoritmo EM è possibile stimare il valore atteso della variabile indicatrice *G*:

$$\hat{g}_j=E\left(g_j\middle|m_j,u_j,p\right)=\frac{pm_j}{pm_j+(1-p)u_j}=\frac{m_j/u_j}{m_j/u_j+(1-p)/p}=\frac{e^{w_j}}{e^{w_j}+(1-p)/p}. \qquad (3)$$

Si noti come il valore atteso di *G* abbia un legame diretto (*logit*) con i pesi *w* di Fellegi e Sunter, che vengono così riportati in una scala 0-1 e resi più interpretabili rispetto ai valori originali.

Il passo *M* consiste nel massimizzare la verosimiglianza per i parametri *m* e *p*, condizionatamente al valore assunto da *G*. Seguendo Jaro (1989), si ritiene invece migliore una stima degli *u* effettuata al di fuori dell'algoritmo, su un campione di coppie abbinate casualmente senza tenere conto del blocco.

I valori di *m* possono essere stimati su un campione "virtuale" di coppie appartenenti a M, formato pesando ogni singola coppia con il valore atteso di *G* calcolato in (3); la stima avviene attraverso le frequenze con cui i singoli risultati del confronto si presentano nel campione pesato. Assumendo ad esempio, per semplicità espositiva, l'indipendenza fra le probabilità di concordanza per le singole variabili, la stima di *m* per la j-esima coppia è data da una combinazione dei valori di *m* stimati per le singole variabili, a seconda dei risultati del confronto presenti nei vettori γ:

$$\hat{m}_j = \prod_{i=1}^{I} \left(\hat{m}^i\right)^{\gamma^i_j} \left(1-\hat{m}^i\right)^{1-\gamma^i_j} \qquad \text{con} \qquad \hat{m}^i = \frac{\sum_{j=1}^{N} \gamma^i_j \hat{g}_j}{\sum_{j=1}^{N} \hat{g}_j}, i = 1..I \,. \tag{4}$$

Infine, la stima di *p* è data semplicemente dalla numerosità relativa del campione "virtuale", ottenuta attraverso la media dei valori assunti dalla variabile indicatrice *G*:

$$\hat{p} = \frac{\sum_{j=1}^{N} \hat{g}_j}{N} \,. \tag{5}$$

Poiché il metodo dipende esclusivamente dai risultati del confronto, è possibile ottenere una rappresentazione più compatta dei vettori attraverso la distribuzione delle frequenze di tutti i possibili risultati ammissibili, compatibilmente con la codifica delle concordanze e la procedura di blocco. Se, ad esempio, si definiscono i vettori γ in modo dicotomico, si ottiene la seguente distribuzione:

$$\gamma_{(k)} = [1,0,1...,1,0] \text{ con frequenza } f_{(k)}, k = 1...K, K \le 2^I \,.$$

Con questa nuova caratterizzazione, il metodo viene reso notevolmente più veloce, poiché è sufficiente un'unica stima della (4) per tutti i vettori γ che si rivelano identici. Inoltre, anche l'utilizzo di (5) viene semplificato, con l'immediata estensione al caso in cui le medie calcolate vengono ponderate attraverso le frequenze con cui ogni singolo tipo di vettore viene osservato.

3.2.4. Integrazione delle stime con metodi basati sulle frequenze

Un *agreement*, ad esempio, sul cognome può avere un peso diverso a seconda che lo stesso cognome sia più o meno raro nelle coppie di record esaminate; cognomi più rari hanno un potere discriminante maggiore rispetto a quelli più comuni. E' possibile aggiustare i pesi *m* ed *u* ottenuti via EM utilizzando distribuzioni di frequenza delle componenti che costituiscono il vettore di confronto al fine di tenere conto della loro importanza relativa.

Seguendo Winkler, dati i due file iniziali *A* e *B*, per una specifica stringa, ad esempio il cognome, si calcola la distribuzione di frequenza. Indicando con f_z e g_z rispettivamente le frequenze della generica occorrenza della stringa in *A* e *B*, si ha:

$f_1, f_2, \ldots, f_m; \ \sum_z f_z = N_A;$

$g_1, g_2, \ldots, g_m; \ \sum_z g_z = N_B.$

Se l' *m*-esima stringa, ad esempio "Rossi", si presenta f_m volte nel file *A* e g_m nel file *B*, allora le frequenze relative agli accoppiamenti per la parola "Rossi" nell'insieme *AxB* saranno:

$$h_1,\ h_2,\ldots,\ h_m;\ \sum_z h_z = N_{AB}.$$

Da notare che $h_z <= \min(f_z , g_z)$. In particolare, nelle applicazioni empiriche si assume:

$h_z = \min(f_z , g_z)$ se $f_z > 1$ o $g_z > 1$;

$h_z = 2/3$ negli altri casi.

Siano *m* e *u* le stime ottenute dall'applicazione dell'algoritmo EM. Si indichi inoltre con:

$\alpha_1 = m_i \times N_{AB} / (\sum_z h_z)$ e

$\alpha_2 = u_i \times (N_A \times N_B - N_{AB}) / (\sum_z b_z)$ dove: $b_z = f_z \times g_z - h_z$ se $f_z > 1$ o $g_z > 1$;

$b_z = 1/3$ altrimenti.

Si approssima:

$m_{iz} = Pr(\text{ z-esima stringa si accoppi / M}) = \alpha_1 \times h_z / N_{AB}$;

$u_{iz} = Pr(\text{ z-esima stringa si accoppi / U}) = \alpha_2 \times b_z / (N_A \times N_B - N_{AB})$.

Risulterà:

$\sum_z m_{iz} = m_i$ e $\sum_z u_{iz} = u_i$.

Sulla base di queste nuove stime m_{iz} e u_{iz} si possono calcolare i nuovi pesi w^j

3.2.5. La scelta della soglia e stima della quota di errori

Al fine di valutare l'efficienza di un qualunque metodo di abbinamento di record è cruciale disporre di stime del numero di errati abbinamenti e mancati abbinamenti che conseguono alla sua applicazione; tali stime, fra l'altro, consentono di affrontare razionalmente il problema della scelta della soglia che consente di decidere quali coppie abbinare e quali no. La soglia deve essere determinata in un'ottica di minimizzazione degli errori, che devono essere stimati con precisione. Belin e Rubin mostrano che invece gran parte dei metodi usualmente utilizzati in precedenza si rivelano eccessivamente ottimisti, con una notevole sottostima degli errori; ciò vale in particolare per i metodi che prevedono l'ipotesi di indipendenza fra gli errori nelle diverse variabili di confronto.

Un metodo spesso utilizzato per una prima approssimazione della soglia migliore, o per definire quali siano le coppie da verificare manualmente, consiste in un'analisi grafica della distribuzione dei pesi sull'intero campione. Questa è la mistura di due distribuzioni che, se la strategia utilizzata è sufficientemente discriminante, sono concentrate in punti distanti fra loro; la distribuzione osservata dovrebbe pertanto presentare un'accentuata bimodalità, con l'altezza relativa delle due mode che

dipende essenzialmente dal numero totale di confronti effettuati e dalle strategie di blocco. La maggiore incertezza rimane nella zona in cui le code delle due distribuzioni condizionate si intersecano, e gli errori dipendono dal punto esatto in cui si posiziona la soglia, con una conferma della relazione inversa fra le proporzioni di falsi positivi e falsi negativi.

3.3 L'implementazione di procedure d'abbinamento: aspetti generali e la procedura generalizzata PARI

Le applicazioni pratiche dei metodi di abbinamento esatto di record presentano spesso delle caratteristiche specifiche dei singoli casi trattati e vengono pertanto trattate con metodi *ad hoc*. Negli ultimi anni, tuttavia, la crescente richiesta di utilizzo di tali metodi ha portato all'implementazione di *software* il più possibile generali e flessibili per l'abbinamento. Fra gli esempi più noti GRLS, prodotto da Statistics Canada, ed i *software* prodotti da esperti di record linkage come Winkler o Jaro.

Una procedura di abbinamento esatto di record automatizzata deve prevedere la possibilità di gestire le varie fasi dell'abbinamento descritte nelle sezioni precedenti, con opzioni di default ma con la possibilità di inserire da utente opzioni esterne in qualunque fase.

L'applicazione agli archivi ASIA descritta nel seguito è stata effettuata tramite la procedura PARI (Procedura per l'Abbinamento di Record Individuali) sviluppata da Paggiaro e Torelli (1998). Tale procedura, sviluppata in ambiente SAS, rappresenta la generalizzazione di quella utilizzata per l'abbinamento longitudinale della Rilevazione Trimestrale delle Forze di Lavoro dell'Istat. In particolare, le opzioni previste dalla procedura (alcune delle quali ancora in fase di sviluppo) sono le seguenti:

1. scelta del tipo di procedura
- ✓ abbinamento in senso stretto
- ✓ deduplicazione

2. analisi preliminari sulla struttura dei file da abbinare
- ✓ tipo di file e metodi per l'importazione dei dati
- ✓ variabili disponibili
- ✓ distribuzioni di frequenza marginali e congiunte

3. scelta delle variabili di confronto
- ✓ numeriche
- ✓ alfanumeriche

4. scelta dei criteri di blocco
- ✓ variabili di blocco
- ✓ blocchi alternativi

5. codifiche di confronto
- ✓ dicotomiche
- ✓ multiple

- ✓ specifiche per singole modalità
- ✓ utilizzo di algoritmi esterni di confronto (ad es. per comparatori di stringhe)

6. stima dei pesi di abbinamento
   - ✓ metodi deterministici
   - ✓ metodi probabilistici (ad es. algoritmo EM)
   - ✓ integrazione fra metodi differenti
7. scelta delle soglie per l'abbinamento
   - ✓ scelta manuale
   - ✓ scelta automatizzata
8. analisi degli abbinamenti multipli
   - ✓ definizione di cluster di record relativi allo stesso individuo
   - ✓ abbinamento one-to-one con scelta della coppia di record "migliore"
9. output
   - ✓ file di dati abbinati
   - ✓ analisi descrittive dei risultati dell'abbinamento

Fra le caratteristiche peculiari della procedura si segnalano:

- La possibilità di gestire in modo semplice procedure di abbinamento probabilistiche, deterministiche, o miste. In particolare, è previsto l'utilizzo dell'algoritmo EM per la stima dei pesi di abbinamento, anche in presenza di dipendenza fra le singole variabili di confronto.
- La possibilità di gestire da utente tutte le fasi di stima, introducendo eventuali informazioni esterne su singole variabili o combinazioni.
- La gestione della definizione di blocco, che prevede anche blocchi alternativi (si confrontano solo le coppie con concordanza in almeno una delle variabili di blocco) e la possibilità di eliminare alcune configurazioni poco interessanti.
- La stile di programmazione SAS che fa largo uso di *macro*, con la possibilità di inserire programmi *ad hoc* definiti da utente nelle varie fasi della procedura.

4. Le fasi del processo di RL

4.1. La struttura dei dati

I dati utilizzati per l'applicazione del processo del RL provengono dal Registro Statistico Asia (strutture 1998 e 1997) e sono riferiti alla regione Sicilia. Ai fini della sperimentazione è stato ricostruito un unico file di dati ***F*** che costituisce l'unione dei due stock di imprese (senza selezionare per stato di attività), in cui è possibile distinguere fra unità appartenenti a stock e unità appartenenti a flussi. La possibilità di identificare se una unità fa parte di un flusso (o appartiene ad entrambi gli stock) è garantita dall'assegnazione di una variabile che confronta lo stato di attività di Asia98/Sicilia e di Asia97/Sicilia e che assume le seguenti modalità rispetto all'anno di riferimento 1998:

A = impresa attiva in Asia98 e Asia97

E = impresa attiva in Asia98 e non attiva in Asia97 – Entrate (in Asia98)

U = impresa non attiva in Asia98 e attiva in Asia97 – Uscite (da Asia98)

Altro = imprese nate dopo il 1998,imprese in Asia97 e non classificate come imprese in Asia98

Nel complesso è stato ottenuto un file di 303.122 record composto da:

A=205.339

E=43.279

U=37.814

Altro=16.690

Da cui è possibile ricostruire per somma nelle componenti:

Asia98/Sicilia, imprese attive [A+E]	248.618
Asia97/Sicilia, imprese attive [A+U]	243.153

Un'organizzazione dei dati così fatta, operativamente, consente di applicare il RL sia per la ricerca dei duplicati di unità che per l'analisi della continuità, secondo un'ottica che sarà più chiara nel seguito.

L'equazione demografica che lega stock e flussi è la seguente[5]: (1)

ASIA98		**ASIA97**		**ENTRATE**		**USCITE**
Imprese attive nel 1998		*Imprese attive nel 1997*		*x stato di attività*		*x stato di attività*
248.618	**=**	243.153	**+**	43.279	**-**	37.814

[5] Le unità classificate come altro non fanno parte dell'equazione e se identificate nei legami verranno opportunamente escluse dall'analisi demografica

4.2. Le fasi del processo del RL

Il processo di Record Linkage ha l'obiettivo di individuare quali record, all'interno di un unico file o tra due file, si riferiscono alla stessa entità utilizzando variabili identificative associate a ciascuna unità. In particolare il RL probabilistico usa probabilità o pesi da assegnare a ciascuna coppia di record al fine di classificarla in uno dei due insiemi: i Link e i Non link.

L'intera procedura di linkage è articolata in 3 macrofasi logiche.

La prima è relativa al trattamento delle variabili di matching e alle scelte a priori per la riduzione dello spazio dei confronti. Ciascuna variabile identificativa necessita di un trattamento prima di essere utilizzata per il matching, specialmente quando si tratta di variabili espresse come stringhe di nomi all'interno del record. Lo sviluppo e l'applicazione di tecniche di standardizzazione e di parsing[6] vengono descritte nel § 4.2.1.1. Il passo successivo consiste nello specificare regole che indicano come confrontare ciascuna variabile di matching, generalmente chiamate regole di agreement/disagreement tra le variabili (§ 4.2.1.2). I risultati dell'applicazione delle regole a ciascuna variabile definisce il vettore dei risultati del confronto $\boldsymbol{\gamma}$.

Per quanto riguarda le variabili di blocco e le ulteriori riduzioni sullo spazio dei confronti (§4.2.2) alcune considerazioni sono necessarie. Sia $\boldsymbol{F}$ l'insieme di N record contenuti in un unico file, l'insieme prodotto $\boldsymbol{F}$x$\boldsymbol{F}$ contiene tutte le possibili NxN coppie da confrontare. E' evidente che all'aumentare della dimensione di $\boldsymbol{F}$ cresce fortemente quella del suo prodotto cartesiano, producendo un numero di coppie da confrontare che risulta poco gestibile. Nella pratica ma anche e soprattutto da un punto di vista concettuale, non è conveniente lavorare sull'intero spazio dei confronti per varie ragioni. In primo luogo trattandosi del caso di un unico file, lo spazio dei confronti diventerebbe Nx(N-1)/2, ossia si escludono i confronti tra gli stessi record e i casi simmetrici; inoltre la maggior parte delle coppie sarebbe sicuramente classificata tra i non-matched; infine molti confronti non si ritengono utili ai fini dell'analisi demografica per la continuità. Il concetto del bloccaggio (ma anche l'adozione di restrizioni a priori) fornisce un metodo per limitare il numero dei confronti e quindi le coppie da esaminare. In pratica il bloccaggio consiste nel partizionare il file in sottoinsiemi ad intersezione nulla (blocchi) in modo che le coppie di record da confrontare sono prese dai record che appartengono allo stesso blocco.

La seconda fase consiste nell'applicazione della metodologia di RL probabilistico (la teoria base è quella di *Fellegi-Sunter).* L'algoritmo utilizza il metodo iterativo EM per la stima dei parametri per l'assegnazione di una probabilità o peso a ciascun vettore $\boldsymbol{\gamma}$ (§4.2.3), dopo che sia

stata scelta la relazione di dipendenza/indipendenza tra le variabili di matching (§4.2.3.2). Viene introdotto l'aggiustamento per le frequenze relative per alcune variabili (§4.3.3.3) ed infine viene assegnata una regola per la scelta della soglia da utilizzare al fine di classificare le coppie ottenute in Link, Possibili Link e Non-Link (§4.3.3.4).

Le coppie classificate come Link entrano nel processo di analisi della terza macrofase. Qui si procede al raggruppamento delle coppie in cluster, alla ricerca dei legami migliori al fine di ridurre i match multipli (quelli tra più di due unità), ai controlli manuali di verifica della bontà dell'accoppiamento, alle verifiche di contenuto del legame ottenuto. Sono invece escluse dal processo di ricerca dei legami tutte le coppie classificate tra i possibili link e i non link.

4.2.1. Le variabili di matching

Per confrontare le unità, sono stati scelti, in conformità alle definizioni di continuità, i seguenti 3 caratteri identificativi: la Ragione Sociale dell'impresa; l'Indirizzo; il codice di Attività Economica. Altre due variabili quali il Codice Fiscale e il Codice di Forma Giuridica sono state utilizzate come supporto per delineare le regole di agreement/disagreement della Ragione Sociale. Al fine di poter disporre di informazioni che fossero il più possibile omogenee e trattabili in modo automatico, sono state necessarie operazioni di standardizzazione e di parsing sulle variabili Ragione Sociale e Indirizzo, di seguito descritte.

4.2.2.1. Il trattamento delle variabili di matching

Le operazioni di standardizzazione e normalizzazione consistono nella scomposizione di una stringa in parole[7], nel "ripulire" tali parole da eventuali segni di punteggiatura, simboli e, per alcuni vocaboli particolari (quali ad esempio le particelle toponomastiche dell'Indirizzo) nella sostituzione di eventuali abbreviazioni della parola con il corrispondente vocabolo (ad esempio: le parole "CTR.", "CONTR", "CNT" vengono sostituite con la parola "CONTRADA").

Dopo tali operazioni si procede ad una seconda fase che, attraverso criteri di somiglianza semantica, attribuisce ad ogni unità di testo individuata un significato specifico.

[6]Per parsing si intende la divisione di un campo a formato libero (un nome) in un set di componenti interpretabili che possono essere automaticamente confrontate

[7] Per parola si intende una concatenazione di caratteri delimitata da due separatori (blank, segni di punteggiatura). In realtà, nell'ambito dell'analisi delle Ragioni Sociali, l'interesse è rivolto all'adozione simultanea di vari tipi di unità di testo non sempre coincidenti con le semplici parole, e che possono essere riunite sotto la categoria chiamata forma testuale comprendente unità semplici (Rossi), unità composte (Maria Teresa) e unità complesse (società per azioni)

In particolare per la Ragione Sociale è stato costruito un vocabolario, costituito da 127.758 parole tra Cognomi, Nomi, parole attinenti alle Forme Giuridiche, parole attinenti alle Istituzioni, parole che riguardano l'Attività Economica, nomi di Comuni, parole non significative, parole di Fantasia[8]. Attraverso un semplice accoppiamento con tale vocabolario, è possibile assegnare ad ogni parola della Ragione Sociale un *flag* di identificazione del tipo di vocabolo (significato).

Poiché il vocabolario non è esaustivo e inoltre può accadere che ad una parola venga attribuito più di un significato, si procede ad una analisi sintattica della Ragione Sociale che si differenzia a seconda che si tratti della Ragione Sociale di una impresa individuale, di una società di persone o di capitale. Questo tipo di analisi riguarda procedimenti mediante i quali le parole assumono un significato univoco secondo il loro ordinamento all'interno della Ragione Sociale.

Due esempi possono essere utili a chiarire meglio questo concetto.

1) Si supponga di analizzare la Ragione Sociale della società di capitale: "SIMED SRL" e si supponga che l'uso del vocabolario consenta di identificare solo la seconda parola come parola di forma giuridica. Una delle regole sintattiche applicate consiste nell'assegnare il *flag* di vocabolo di fantasia ad ogni parola che precede quella di forma giuridica e che appartiene ad una Ragione Sociale di una società di capitali costituita da due sole parole.

2) Si supponga di analizzare la Ragione Sociale della società di persone: "SICILEGNO DI BRUNO VINCENZO & BRUNO ALFIO". In tal caso il vocabolario attribuisce alla parola BRUNO sia il significato di cognome che quello di nome. Poiché tale parola è seguita per due volte da un nome, un'altra regola sintattica consiste nell'assegnare univocamente il *flag* relativo al cognome se la parola è preceduta o seguita da un'altra a cui è stato attribuito univocamente il significato di nome.

Queste sono solo due delle regole applicate; procedimenti più complessi sono stati seguiti al fine di classificare nel modo migliore tutte le parole delle Ragioni Sociali.

Tale operazione è fondamentale per identificare le parole significative utili per stabilire le regole di agreement/disagreement che verranno esposte in seguito.

Per quanto riguarda l'Indirizzo avviene un'operazione simile a quanto fatto per la Ragione Sociale. Anche in questo caso infatti all'interno di tale variabile viene individuata la particella toponomastica precedentemente standardizzata, nonché uno o più numeri civici che solitamente seguono la stringa dell'Indirizzo. Sono così individuate 3 sotto-componenti: la particella toponomastica; la denominazione della strada, scomposta in parole; il numero civico.

[8] Tali parole possono essere vocaboli inventati, acronimi ecc…

4.2.1.2. Le regole di agreement/disagreement

Quando i record vengono confrontati si producono degli *outcome* che sono i risultati del confronto tra le variabili identificative di ciascuna unità. Le modalità con cui si effettuano i confronti sono stabilite da alcune regole che risultano più o meno complesse a seconda della variabile che si sta confrontando.

Per il confronto della ***Ragione Sociale*** sono stati adottati diversi criteri a seconda della tipologia di Forma Giuridica delle imprese costituenti la coppia.

Per le coppie del tipo "Impr.Individuale - Impr.Individuale", "Impr.Individuale - Soc.di Persone", "Soc.di Persone - Soc.di Persone" vengono individuati, attraverso l'analisi prima descritta, due campi, rispettivamente Cognome e Nome del titolare dell'impresa[9]. Si stabiliscono così delle regole di agreement/disagreement per la componente Ragione Sociale sulla base del confronto tra questi due campi. In particolare si stabilisce di considerare come chiave primaria di accoppiamento il Cognome. Se esiste una coincidenza del Cognome allora si considera come seconda variabile di matching il Nome.

Per le restanti tipologie di coppia ("Soc.di Persone - Soc.di Capitale", "Soc.di Capitale - Soc.di Capitale") il confronto della Ragione Sociale è risultato più complesso a causa della struttura sintattica delle Ragioni Sociali. Per queste coppie si pone il problema della scelta del tipo di parola da considerare come chiave primaria di accoppiamento. Si è deciso di scegliere come parole significative i cognomi e le parole di fantasia (queste ultime caratterizzano le Ragioni Sociali delle società di capitale). In questi casi la procedura di accoppiamento ha riguardato il calcolo di due indici: il primo dato dal rapporto tra il numero di parole significative che si sono accoppiate tra le due Ragioni Sociali poste a confronto e il numero delle parole accoppiate; il secondo dato dal rapporto tra il numero di parole accoppiate delle due Ragioni Sociali e il numero delle parole della Ragione Sociale meno lunga. La necessità di calcolare questo secondo indice deriva dal fatto che, in alcuni casi riguardanti essenzialmente le società di capitale, i consorzi e le cooperative, le Ragioni Sociali sono costituite da parole non significative se analizzate singolarmente, ma significative se analizzate nel complesso. Un esempio può essere utile a chiarire meglio questo concetto. La Ragione Sociale "SOCIETA' SICULA EDILIMPIANTI" è costituita da parole tutte non significative, per cui, se nella scelta delle coppie ci si basasse esclusivamente sul primo indice, questa Ragione Sociale non comparirebbe mai tra le coppie scelte; considerando anche il secondo indice, è possibile

[9] Si tenga presente che per le Società di Persone possono essere presenti più di un Cognome e più di un Nome. In tal caso il confronto riguarda ogni Cognome e ogni Nome individuato.

che questa Ragione Sociale si accoppi con un'altra e che quindi rientri tra gli agreement o tra gli partial-agreement[10].

Sulla base del confronto tra i Cognomi e i Nomi e/o del calcolo dei due indici si costruiscono tre variabili, sotto-componenti della Ragione Sociale. La prima relativa al confronto tra Cognomi assume modalità di **A**greement o **D**isagreement a seconda della coincidenza o meno del Cognome. La seconda relativa al confronto tra Nomi assumerà in generale modalità di **A**greement solo se l'*outcome* sul Cognome è esso stesso **A**greement e se esiste una coincidenza tra Nomi[11]. La terza relativa ai due indici calcolati, assumerà modalità di **A**greement solo se esiste una parola coincidente e significativa (1° indice diverso da 0) e contemporaneamente tutte le parole di una delle due Ragioni Sociali poste a confronto sono interamente contenute nell'altra (2° indice uguale a 1); assumerà modalità di **P**artial-**A**greement se, in generale, il numero delle parole significative sul totale delle parole accoppiate è maggiore del 50% (1° indice > 0.5) e contemporaneamente il 2° indice è maggiore di 0.8 o se il 1° indice ha valore pari a 1 e contemporaneamente il 2° indice è maggiore di 0.5; assumerà modalità di **D**isagreement negli altri casi.

Sinteticamente, volendo costruire un'unica variabile risultante dal confronto delle Ragioni Sociali (tot_**RS**), si avrà:

- ✓ **A**greement se, per le tipologie "Impr.Individuale - Impr.Individuale", "Impr.Individuale - Soc.di Persone", "Soc.di Persone - Soc.di Persone", si ha contemporaneamente coincidenza del Cognome e del Nome o se, per le restanti tipologie di coppia, la terza sottocomponente relativa ai due indici ha modalità di **A**greement;
- ✓ **P**artial-**A**greement se, per le tipologie "Impr.Individuale - Imp.Individuale", "Imp. Individuale - Soc.di Persone", "Soc.di Persone - Soc.di Persone", si ha solo una coincidenza sul Cognome o se, per le restanti tipologie di coppia, la terza sottocomponente relativa ai due indici ha modalità di **P**artial-**A**greement;
- ✓ **D**isagreement negli altri casi.

Per il confronto del carattere ***Indirizzo*** si decide di considerare come chiavi di accoppiamento la particella toponomastica, la denominazione della strada e il numero civico. Naturalmente questi tre campi non concorrono nella stessa misura al risultato complessivo del confronto. La priorità è data alla denominazione della strada: se la percentuale di parole che si accoppiano tra le due

[10] Questi due indici in effetti sono calcolati anche per le coppie "Soc. di Persone – Soc. di Persone", in quanto, anche per questi casi, potrebbero essere presenti parole di Fantasia che, solo attraverso l'analisi del Cognome e del Nome, non verrebbero prese in considerazione.

[11] Si ha **A**greement sul Nome anche nel caso in cui in una delle due Ragioni Sociali esiste una parola che precede o segue il Cognome costituita da una sola lettera puntata, la quale coincide con la prima lettera del Nome presente nell'altra Ragione Sociale

stringhe poste a confronto è almeno pari al 60% rispetto alla stringa meno lunga dei due Indirizzi allora si avrà un **A**greement per tale sotto-componente, **D**isagreement negli altri casi.

Per quanto riguarda il confronto tra i numeri civici, sono stati predisposti quattro campi per ogni Indirizzo: 1° # civico, 2° # civico, 3° # civico e 4° # civico. Si avrà un **A**greement se i numeri civici sono presenti e coincidenti in entrambi gli Indirizzi; un **P**artial-**A**greement se almeno uno dei due Indirizzi posti a confronto non ha numero civico; un **D**isagreement se i numeri civici sono presenti in entrambi gli Indirizzi, ma diversi. L'ultimo confronto è quello tra le particelle toponomastiche; si avrà un **A**greement o un **D**isagreement a seconda della coincidenza o meno tra le particelle toponomastiche.

In maniera analoga a quanto è stato fatto per la Ragione Sociale, anche per l'Indirizzo è stata creata una variabile complessiva (tot_**IND**) che costituisce il secondo elemento del vettore dei confronti e che assume le seguenti modalità:

- ✓ **A**greement se tutte e tre le sotto-componenti hanno come risultato un **A**greement.
- ✓ **P**artial-**A**greement se le sotto-componenti particella toponomastica e denominazione della strada presentano un **A**greement, mentre dal confronto dei numeri civici è risultato come *outcome* **P**artial-**A**greement o **D**isagreement;
- ✓ **D**isagreement negli altri casi.

L'ultima variabile di confronto è il codice di <u>*Attività Economica*</u>. Per tale componente si ha:

- ✓ **A**greement se i codici di Attività Economica sono risultati coincidenti o compatibili[12];
- ✓ **D**isagreement per codici di Attività Economica diversi.

Con l'applicazione di queste regole due unità che ad esempio presentano un **A**greement sulla Ragione Sociale, un **P**artial-**A**greement sull'Indirizzo e un **D**isagreement sul codice di Attività economica, presenteranno la seguente configurazione del vettore dei confronti γ=(A,PA,D).

4.2.2. Le variabili di blocco e le restrizioni imposte

Alcuni dei caratteri registrati nei due archivi Asia'97/'98 relativamente alla regione Sicilia risultano non utilizzabili perché incompleti; infatti per entrambi gli anni circa il 2% dell'informazione presenta un problema di incompletezza.

[12] Per le Ateco compatibili è stata costruita una matrice così strutturata: nella 1° colonna sono riportate una lista di codici Ateco; ad ogni codice sono state associate una o più Ateco compatibili. Si tenga presente che per tale matrice non viene rispettata la proprietà di transitività: se l'ateco *X* è compatibile con l'ateco *Y* e con l'ateco *Z*, allora *Y* è sicuramente compatibile con *X* così come *Z* è compatibile con *X*, ma non necessariamente *Y* e *Z* risulteranno tra loro compatibili.

Le motivazioni principali riguardano:

1) l'incompletezza del codice di Attività Economica a livello di 5 cifre;
2) la mancanza del C.A.P. (che da sola costituisce il 78% circa dell'errore totale).

E' stato necessario escludere tutte le unità con caratteri incompleti; ciò ha comportato una riduzione dell'insieme *F* di input, da 303.122 a 294.065 unità.

I possibili confronti che si possono formare a partire da tale insieme, sono 43.236.965.080 pari alle combinazioni senza ripetizione. E' evidente che tale dimensione risulta poco gestibile ai fini della stima delle probabilità di Link e Non-Link. E' risultata quindi utile sia l'applicazione di una tecnica di bloccaggio, sia l'applicazione di alcune tecniche di restrizione.

4.2.2.1. Le variabili di blocco

Nella scelta delle variabili da utilizzare come blocchi, si deve tener conto del rischio di possibili errati abbinamenti e dei mancati abbinamenti che potrebbero altrimenti essere recuperati in assenza di blocchi. Alla luce di tali considerazioni si è deciso di utilizzare come prima variabile di blocco il Comune, all'interno del quale utilizzare due blocchi alternativi: il C.A.P. e il codice di Attività economica. Questo doppio bloccaggio sta a significare che il mancato abbinamento tra due unità che presentano C.A.P. diversi viene recuperato se tali unità presentano uno stesso codice di Attività economica e viceversa.

4.2.2.2. Atre restrizioni sullo spazio dei confronti

Volendo ulteriormente ridurre lo spazio dei confronti, oltre a considerare solo le coppie che presentano una concordanza perfetta su almeno una delle variabili di blocco alternativo, sono state applicate le seguenti restrizioni:

1. vengono escluse tutte le coppie del tipo "Impr.Individuale - Soc.di Capitale"; tale esclusione deriva essenzialmente dalla minima possibilità di trovare transizioni di questo tipo[13]; altro motivo riguarda la struttura sintattica della Ragione Sociale delle società di capitale, che, nella maggior parte dei casi, non prevede la presenza di parole come Cognome e Nome, parole che invece costituiscono la Ragione Sociale delle imprese individuali.
2. per le coppie formate tra imprese Individuali, si escludono i casi di omonimia
3. di tutti i possibili confronti che si possono effettuare tra:

- cooperative e consorzi – altre società di capitale;
- cooperative e consorzi – società di persone ;
- cooperative e consorzi

vengono considerati solo quelli che per la variabile Ragione Sociale presentano almeno un **P**artial-**A**greement e contemporaneamente presentano **A**greement per le altre due variabili di confronto.

4. Infine, delle restanti coppie che presentano **D**isagreement di Ragione Sociale, entrano nella procedura di stima solo quelle che hanno **A**greement per le altre due variabili di confronto Indirizzo e codice di Attività economica.

Di seguito si riporta la distribuzione per provincia e per i comuni di Palermo e Catania, del numero di unità di input (prima e dopo l'esclusione delle unità aventi caratteri non normalizzati) e del numero delle coppie considerate dopo le restrizioni descritte.

Tav.4.1: *Distribuzione di frequenza per provincia*

PROVINCE	N° rkd di input	N° rkd dopo norm.	Spazio dei confronti (coppie)
Trapani	27.580	27.043	13.189
Palermo (Comune)	*42.726*	*41.300*	19.405
Palermo (prov. Escluso com. di Palermo)	27.476	26.951	6.249
Totale prov. di Palermo	**70.202**	**68.251**	**25.654**
Messina	43.776	41.681	12.018
Agrigento	25.882	25.401	8.273
Caltanissetta	15.725	15.434	5.155
Enna	9.903	9.795	2.302
Catania (Comune)	*26.347*	*24.311*	15.082
Catania (prov. Escluso com. di Catania)	42.578	41.702	13.975
Totale prov. Di Catania	**68.925**	**66.013**	**29.057**
Ragusa	19.386	19.077	24.430
Siracusa	21.743	21.370	7.547
Totale Sicilia	**303.122**	**294.065**	**127.625**

4.2.3. La stima dei pesi e l'individuazione dei Link

Le diverse combinazioni di output prodotti dall'applicazione delle regole di agreement/disagreement su ciascuna variabile di matching rappresentano il risultato del vettore dei confronti $\boldsymbol{\gamma}$ per ogni coppia di record. A ciascuna configurazione del vettore viene assegnato un peso $\boldsymbol{w}$ dato dal logaritmo neperiano del rapporto delle due probabilità $\boldsymbol{m}$ ed $\boldsymbol{u}$ stimate dall'algoritmo. Più precisamente, mentre $\boldsymbol{m}$ viene stimata usando le informazioni dell'insieme *FxF*

[13] Il passaggio tra ditta individuale a società di capitali si presenta principalmente con un tipo di società: le società a responsabilità limitata unipersonale.

vincolato dalle restrizioni imposte, la stima delle ***u*** è ottenuta sulla base di un campione casuale di coppie di record estratte dall'insieme *FxF* svincolato da restrizioni.

4.2.3.1. Stima delle ***u***: problematiche connesse alla dimensione del campione

Un problema che è emerso durante l'applicazione dell'algoritmo è stato quello di decidere la numerosità del campione, utilizzato per la stima dei pesi ***u***.

Per ragioni attribuibili essenzialmente alla disponibilità di memoria, si è deciso di considerare un campione di 3.000.000 di coppie di unità. Resta comunque da stabilire una numerosità adeguata di tale campione in modo che quest'ultimo risulti rappresentativo dell'universo costituito da tutte le configurazioni possibili del vettore dei confronti **γ**. Quando infatti la numerosità di tale campione è troppo esigua, è possibile che alcune particolari configurazioni del vettore dei confronti non siano presenti, nonostante siano presenti nell'insieme *FxF* vincolato, su cui si basa la stima delle ***m***. Ne segue che, a tali configurazioni verrà assegnata una stima delle ***u*** molto bassa (1/numerosità del campione), da cui ne deriva un peso ***w*** relativamente molto alto.

Per ragioni legate alla dimensione degli archivi, l'algoritmo descritto è stato applicato separatamente per ogni provincia della Sicilia e per i comuni di Palermo e Catania.

4.2.3.2. Scelta delle relazioni di dipendenza/indipendenza tra le variabili di matching

Per la stima delle ***m*** e delle ***u***, è possibile stabilire delle relazioni di dipendenza/indipendenza tra le variabili di matching.

Per entrambi i pesi, si è deciso di imporre le seguenti relazioni:

- ✓ Una relazione di dipendenza tra le singole sotto-componenti della variabile Ragione Sociale (Cognome, Nome e i due Indici) e la tipologia della coppia (ricavabile attraverso il confronto tra le forme Giuridiche). La ragione di tale dipendenza risiede nel fatto che i risultati del confronto di queste componenti concorrono in misura diversa al risultato della variabile Ragione Sociale a seconda della tipologia di coppia.
- ✓ Una relazione di dipendenza tra le variabili Indirizzo e Attività Economica. Ciò è motivato dal fatto che per alcune attività economiche la localizzazione è un elemento caratterizzante; per altre attività[14] non lo è.

[14] Es: attività dell'edilizia , dei servizi, dei professionisti.

4.2.3.3. L'aggiustamento per le frequenze relative

L'algoritmo EM consente di stimare le probabilità di matching relativamente ad ogni componente di confronto utilizzata. Si noti comunque che, facendo ad esempio riferimento all'**A**greement sul Cognome, un Cognome che è relativamente raro nelle coppie di unità esaminate dovrebbe avere un potere discriminante maggiore rispetto ad uno più comune. E' possibile utilizzare le distribuzioni di frequenza delle componenti che costituiscono il vettore di confronto al fine di "correggere" le probabilità ***m*** ed ***u***.

Questa tecnica di correzione per le frequenze relative è stata applicata per i seguenti casi, tenendo conto delle relazioni di dipendenza imposte per la stima dei parametri:

- ✓ Per la *sotto*-componente Cognome presente nella Ragione Sociale.
- ✓ Per le *sotto*-componente Denominazione della strada presente nell'Indirizzo.

Di seguito si riporta, a titolo di esempio, parte della distribuzione di frequenza delle configurazioni del vettore dei confronti delle coppie del comune di Palermo, sia con i pesi ***w*** risultanti dall'applicazione dell'algoritmo EM (*w_old*), sia con quelli ottenuti dopo l'aggiustamento per le frequenze (*w_new*).

Tav.4.2: *Esempio tratto dai risultati sul comune di Palermo*[15]

Cluster	Conf_FG	tot_RS	tot_IND	tot_ATE	w_old	w_new	atecouno	atecodue	freq
42	3	A	A	A	10.94	10.43	*51310*	*51310*	1
42	3	A	PA	A	6.19	6.17	*51310*	*51310*	2
42	**4**	**D**	**A**	**A**	**4.06**	**1.46**	***51310***	***51310***	**18**
42	2	A	D	A	3.31	3.26	*51310*	*51310*	1
42	**3**	**D**	**A**	**A**	**3.28**	**1.29**	***51310***	***51310***	**28**
42	3	A	D	A	2.95	2.95	*51310*	*51310*	1
42	2	A	PA	D	2.86	2.86	*51310*	*45230*	1
TOTALE COPPIE									**52**

Da check manuale è emerso che le unità coinvolte in questo cluster (a meno di una) costituiscono il mercato ortofrutticolo del comune di Palermo (Ateco=51310). Come si può notare non si riscontrano differenze significative tra *w_old* e *w_new* a meno dei casi caratterizzati dall'avere **D**isagreement sulla Ragione Sociale e **A**greement per l'Indirizzo e l'Attività Economica.

L'aggiustamento per le frequenze consente di abbassare il valore di ***w*** per alcune particolari configurazioni e quindi di classificare alcune coppie come Non-Link quando sono effettivamente dei falsi matches.

[15] Legenda: la variabile conf_FG si riferisce alla tipologia di coppia e, in questo esempio, assume le seguenti modalità: 2=confronto tra soc.Individuale e soc.di Persone; 3=confronto tra soc. di Persone; 4=confronto tra soc.di Persone e soc.di Capitale. Le altre variabili tot_RS, tot_IND e tot_ATE si riferiscono rispettivamente agli outcome delle tre componenti del vettore dei confronti Ragione Sociale, Indirizzo e Ateco.

4.2.3.4. La scelta della soglia: un confronto tra diversi metodi

Una volta ottenute le probabilità di link per ogni coppia di unità, è necessario identificare quali legami classificare nell'insieme *M* dei Link e quali classificare nell'insieme *U* dei Non-Link.

Nella presente fase, che ha carattere ancora sperimentale, si è deciso di valutare i risultati che si ottengono utilizzando due differenti strategie per individuare i valori soglia:

(a) Soglia empirica – il valore soglia viene individuato tenendo conto di due particolari configurazioni del vettore dei confronti γ e si considerano Non-Link tutte le coppie che presentano una probabilità inferiore a quella assegnata alla configurazione: tot_RS=**PA**, tot_IND=**D** e tot_ATE=**A** oppure alla configurazione: tot_RS=**PA**, tot_IND=**PA** e tot_ATE=**D**.

(b) Soglia automatica – un metodo automatico per l'individuazione del valore soglia, basato sui valori assunti dal peso ***w*** e schematizzato come segue: il primo passo consiste nell'ordinare in senso crescente tutte le configurazioni che assume il vettore dei confronti secondo il valore assunto dal peso ***w***. Dopo aver ordinato per tali pesi, si calcola la cumulata delle probabilità ***m*** e il complementare alla cumulata delle probabilità ***u***. Il valore massimo (*max_m*) per una decisione di non-match corrisponde al peso ***w*** della particolare configurazione, dove la cumulata delle ***m*** è inferiore o uguale ad un prefissato livello di errore che si è disposti a commettere nel classificare una coppia match come non-match. Tale errore è stato fissato ad un livello del 5%. Il valore minimo (*min_u*) per una decisione di match corrisponde al peso ***w*** della configurazione dove il complementare alla cumulata delle ***u*** è inferiore o uguale ad un prefissato livello di errore che si è disposti a commettere nel classificare una coppia non-match come match. Tale errore è stato fissato ad un livello dell'1%. Pertanto le coppie aventi configurazioni con valore di ***w*** maggiore o uguale a *min_u* sono classificate come Link; quelle aventi un valore di ***w*** inferiore o uguale a *max_m* vengono classificate come Non-Link; infine quelle aventi un valore di ***w*** intermedio ai due valori suddetti, vengono classificate come Possibili-Link.

Per quanto riguarda l'applicazione alla regione Sicilia si è deciso di considerare solo il valore *min_u* e quindi di classificare le coppie solo nei due insiemi dei Link e dei Non-Link.

Ulteriori approfondimenti e check manuale sarebbero necessari qualora si volessero analizzare le coppie classificate come Possibili-Link.

Va inoltre aggiunto che la soglia definita sulla base del valore *min_u*, fatta eccezione del comune di Catania e della provincia di Ragusa, non si discosta molto da quella individuata con il primo metodo empirico. Di seguito si riportano, per ogni provincia e per i due comuni di Catania e Palermo i valori soglia calcolati con i due metodi.

Tav.4.3: Valori soglia per provincia calcolati con i due metodi

PROVINCE	Soglia automatica	Soglia empirica
Trapani	0.91847	0.93629
Palermo (Comune)	0.94808	0.94467
Palermo (prov. Escluso com. di Palermo)	0.97257	0.98770
Messina	0.97322	0.97322
Agrigento	0.93769	0.96376
Caltanissetta	0.84650	0.84646
Enna	0.93551	0.92924
Catania (Comune)	0.96753	0.89418
Catania (prov. Escluso com. di Catania)	0.96488	0.96422
Ragusa	0.77599	0.87937
Siracusa	0.82257	0.86343

Si riporta anche una tavola riassuntiva delle coppie e delle unità coinvolte nell'algoritmo EM, nonché la loro classificazione nei due sottoinsiemi dei Link e dei Non-Link ottenuti con l'applicazione della soglia automatica.

Tav.4.4*: Quadro riassuntivo*

PROVINCE	Coppie EM	LINK	% LINK	NON-LINK	% NON-LINK
Trapani	13.189	954	*7,2*	12.235	*92,8*
Palermo (Comune)	19.405	2.777	*14,3*	16.628	*85,7*
Palermo (prov. Escluso com. di Palermo)	6.249	950	*15,2*	5.299	*84,8*
Messina	12.018	2.196	*18,3*	9.822	*81,7*
Agrigento	8.273	1.210	*14,6*	7.063	*85,4*
Caltanissetta	5.155	767	*14,9*	4.388	*85,1*
Enna	2.302	406	*17,6*	1.896	*82,4*
Catania (Comune)	15.082	1.330	*8,8*	13.752	*91,2*
Catania (prov. Escluso com. di Catania)	13.975	1.870	*13,4*	12.105	*86,6*
Ragusa	24.430	837	*3,4*	23.593	*96,6*
Siracusa	7.547	1.238	*16,4*	6.309	*83,6*
Totale Sicilia	**127.625**	**14.535**	***11,4***	**113.090**	***88,6***

4.2.4. Criteri di scelta dei Link definitivi: costruzione dei cluster e risoluzione dei match multipli

Con la scelta dei valori soglia si separano i Link (coppie con peso w> del valore soglia) dai Non-Link; soltanto per le coppie che ricadono nell'insieme dei Link si procede alla costruzione dei cluster, ovvero gruppi di coppie riguardanti le stesse unità. Ad esempio se le unità A, B e C sono collegate, il cluster potrebbe essere costituito dalle seguenti 3 coppie: A-B, A-C, B-C.

La formazione dei cluster è necessaria in quanto l'algoritmo EM prima descritto abbina le unità a coppie; nella costruzione del legame completo fra le unità è necessario raggruppare le unità in comune tra le coppie per stabilire se i match sono singoli (1 unità è legata a 1 unità) o multipli (molte a molte). I match multipli sono rappresentati dai cluster costituiti da più di una coppia.

Per i Link ottenuti (14.535 coppie) viene presentata in Tav.4.5 la distribuzione per provincia e per numero di coppie all'interno dei cluster, del numero di cluster costruiti e delle unità in essi coinvolte.

Tav.4.5: *I Link*

PROVINCE	Num_coppie 1			num_coppie 2			num_coppie 3			num_coppie >3			num_coppie totale		
	Num_ Cluster	tot_ Coppie	Tot_rk	num_ cluster	tot_ coppie	tot_rk	num_ cluster	tot_ coppie	tot_rk	num_ cluster	tot_ coppie	tot_rk	num_ cluster	tot_ coppie	tot_rk
Agrigento	818	818	1.636	53	106	159	48	144	154	24	142	106	943	1.210	2.055
Caltanissetta	494	494	988	25	50	75	33	99	102	19	124	89	571	767	1.254
Catania (Prov)	1.295	1.295	2.590	70	140	210	87	261	267	27	174	130	1.479	1.870	3.197
Catania (Com)	726	726	1.452	56	112	168	85	255	270	36	237	192	903	1.330	2.082
Enna	289	289	578	21	42	63	20	60	64	3	15	12	333	406	717
Messina	1.370	1.370	2.740	88	176	264	107	321	335	51	329	263	1.616	2.196	3.602
Palermo (Prov)	683	683	1.366	40	80	120	42	126	132	11	61	51	776	950	1.669
Palermo (Com)	1.332	1.332	2.664	97	194	291	106	318	333	89	933	544	1.624	2.777	3.832
Ragusa	603	603	1.206	31	62	93	25	75	81	16	97	81	675	837	1.461
Siracusa	628	628	1.256	43	86	129	46	138	147	38	386	256	755	1.238	1.788
Trapani	716	716	1.432	31	62	93	39	117	125	10	59	47	796	954	1.697
Totale Sicilia	8.954	8.954	17.908	555	1.110	1.665	638	1.914	2.010	324	2.557	1.771	10.471	14.535	23.354

Più dell'85% dei cluster consiste di match singoli (1 coppia), solo il 3% dei cluster è costituito da più di tre coppie; in totale i cluster con 1,2,3 coppie rappresentano i legami tra il 92% delle unità.

I cluster multipli sono quelli più complessi da interpretare in termini di legami tra le unità e quindi andrebbero analizzati manualmente (improponibile se troppi); un obiettivo è quello di assegnare regole per "scegliere i legami migliori" o ridurre la dimensione dei cluster "escludendo i legami rischiosi". La costruzione di regole con questi scopi non solo è necessaria per risolvere eventuali errori della procedura (non corretta normalizzazione dei caratteri, non precisa definizione di qualche regola di agreement, etc.) ma è collegata più strettamente all'interpretazione logico-economica dei legami trovati.

In generale è concretamente possibile gestire in modo automatico solo i cluster con al massimo 3 coppie (ai fini dell'interpretazione dei legami) e invece trattare in maniera manuale quelli con più coppie.

- La prima regola (**Reg_det.1**) riguarda l'attività economica svolta dall'impresa. Da controlli empirici, è stato riscontrato che le imprese che svolgono alcune attività economiche possono essere localizzate tutte allo stesso indirizzo per motivi di convenienza (ad esempio i mercati ortofrutticoli, i centri commerciali, gli studi di professionisti, etc); i legami che coinvolgono tali attività sono ad alto rischio di errore. Si è allora costruita una lista di attività economiche che in presenza di una precisa configurazione del vettore dei confronti identificano i legami non

buoni. La regola di esclusione incide sulla configurazione del vettore dei confronti γ = (D,A,A) e sulle attività elencate nella seguente tabella:

Reg_det.1 – Attività economiche particolari

Ateco	Descrizione
70	Attività immobiliari
45	Costruzioni
65-66-67	Intermediazione monetaria e finanziaria
74.1	Attività legali, contabilità
74.2	Studi architettura , ingegneria,..
74.4	Pubblicità
85.1	Attività dei servizi sanitari
55.1	Alberghi
63.3	Attività delle agenzie viaggio
63.4	Spedizionieri
52.62	Commercio al dettaglio ambulante
52.63	Commercio al dettaglio fuori dei negozi
74.84	Altre attività dei servizi n.c.a.
60.250	Trasporto merci su strada

La caratteristica principale di queste attività consiste nel fatto che non è possibile, per diverse ragioni, stabilire con la localizzazione un carattere che discrimini tra le unità. Per chiarire il concetto, due ambulanti o due spedizionieri sono per definizione non localizzabili; due studi diversi di architettura possono essere localizzati esattamente nello stesso edificio, e così anche due o più alberghi di piccole dimensioni (pensioni), specialmente in alcuni ambiti territoriali.

- La seconda regola (**Reg_det.2**) riguarda la riduzione dei cluster costituiti da più di tre coppie attraverso una funzione di scelta dei legami migliori (scelta del dominante). La logica sottostante è che più alto è il peso attribuito alla coppia (***w***) maggiore è la probabilità che essa appartenga all'insieme dei matched.

Al fine di ridurre in modo automatico il numero di coppie all'interno di cluster con match multipli si fa uso di indici di variabilità e indici posizionali calcolati sulle distribuzioni dei pesi *w* all'interno di ogni cluster.

Siano *cv* = coefficiente di variazione

q3 = il terzo quartile della distribuzione

allora se:

$w > q3$, $cv > 20\%$ $\Rightarrow$ $w^* = w$ rappresenta il peso dominante

La ricerca dei dominanti consiste nella identificazione di quelle coppie che hanno pesi ***w*** significativamente distanti dagli altri valori all'interno della distribuzione .

Dai cluster con dominante vengono escluse le coppie non scelte dalla regola; invece cluster senza dominanti vengono mantenuti inalterati: dopo l'applicazione della regola del dominante, i

cluster vengono nuovamente ricostruiti. Quale effetto (voluto) dell'applicazione di questa regola si "recuperano" cluster che inizialmente hanno più di 3 coppie, e che dopo la regola si riducono.

I cluster che rimangono con più di 3 coppie possono solo essere sottoposti a controllo manuale.

La Tav.4.6 presenta i *Link definitivi* in una distribuzione analoga a quella della precedente Tav.4.5; in essa si evidenzia quale effetto dell'applicazione delle due regole una riduzione di circa il 7% del numero di cluster.

Tav.4.6: *I Link definitivi*

PROVINCE	num_coppie 1			num_coppie 2			num_coppie 3			num_coppie totale		
	num_ cluster	tot_ coppie	tot_rk	num_ cluster	tot_ coppie	tot_rk	num_ cluster	tot_ Coppie	tot_rk	num_ cluster	tot_ coppie	tot_rk
Agrigento	799	799	1.598	57	114	171	47	141	150	903	1.054	1.919
Caltanissetta	484	484	968	26	52	78	32	96	98	542	632	1.144
Catania (Prov)	1.261	1.261	2.522	72	144	216	81	243	251	1.414	1.648	2.989
Catania (Com)	631	631	1.262	36	72	108	43	129	129	710	832	1.499
Enna	290	290	580	21	42	63	20	60	64	331	392	707
Messina	1.360	1.360	2.720	92	184	276	97	291	303	1.549	1.835	3.299
Palermo (Prov)	689	689	1.378	41	82	123	44	132	138	774	903	1.639
Palermo (Com)	1.209	1.209	2.418	101	202	303	79	237	244	1.389	1.648	2.965
Ragusa	599	599	1.198	35	70	105	25	75	81	659	744	1.384
Siracusa	602	602	1.204	45	90	135	35	105	113	682	797	1.452
Trapani	720	720	1.440	31	62	93	39	117	125	790	899	1.658
Totale Sicilia	8.644	8.644	17.288	557	1.114	1.671	542	1.626	1.696	9.743	11.384	20.655

5. Analisi dei risultati

5.1. La struttura dei legami identificati

Le coppie appartenenti ai Link definitivi costituiscono 9.743 cluster e coinvolgono 20.655 unità, collegate tra loro da legami di diverso tipo. La tav.4.6 mette in evidenza come la percentuale maggiore (circa l'85%) dei legami sono match 1 a 1.

La Tav.5.1 invece descrive i tipi di legami tra le unità, distinguendo tra i cluster che coinvolgono solo imprese individuali, quelli solo tra società e i legami tra individuo e società.

Tav.5.1: *Cluster per tipo di legame e numero di unità coinvolte*

	Numero unità				
Legami tra unità	2	3	4	Totale	% sul totale
tra individui					
2I	5.526	-	-	5.526	
3I	-	504	-	504	
4I	-	-	26	26	
sub-totale	5.526	504	26	6.056	62,2
tra società					
2S	1.035	-	-	1.035	
3S	-	132	-	132	
4S	-	-	11	11	
sub-totale	1.035	132	11	1.178	12,1
tra Individui e società					
1I-1S	2.083	-	-	2.083	
1I-2S	-	103	-	103	
1I-3S	-	-	11	11	
2I-1S	-	290	-	290	
2I-2S	-	-	11	11	
3I-1S	-	-	11	11	
sub-totale	2.083	393	33	2.509	25,8
Totale	8.644	1.029	70	9.743	

L'avere individuato i legami tra le unità è stato l'obiettivo principale del RL. L'operazione successiva è quella di utilizzare i risultati ottenuti nell'ottica della demografia di impresa, al fine di correggere gli stock dai duplicati e depurare i flussi dalle unità che continuano.

E' utile classificare i legami tra le coppie appartenenti ai link definitivi in sottopopolazioni di matching sulla base dei risultati del vettore dei confronti γ , che descrivono le combinazioni degli agreement/disagreement tra i caratteri. Tali sottopopolazioni mettono in evidenza quanti e quali caratteri sono "simili" tra le unità collegate consentendo di identificare i sottoinsiemi cui applicare le regole di continuità/duplicazione. Lo schema che segue chiarisce le modalità di classificazione e risulta coerente con lo schema elaborato e proposto dalla metodologia Eurostat (§2.3.3.).

Sottopopolazione	Tipo_legame	γ
Name & Location & Sector	N+L+S	A,A,A
Name & Location	N+L	A,A,D
Location & Sector	L+S	D,A,A – P,A,A
Name & Sector	N+S	A,D,A – A,P,A
Location & sector	l+s	P,P,A
Name & sector	n+s	P,D,A
Name & location	n+l	P,A,D – A,P,D – P,P,D
Altro	xxx	A,D,D – D,D,A – D,A,D

5.2. I legami per la continuità

I legami che identificano la continuità sono quelli che collegano unità attive nell'anno a quelle che entrano nello stock dello stesso anno (le Entrate), unità attive a imprese che escono (le Uscite), unità che entrano a unità che escono (Entrate con Uscite). Ad esempio se il collegamento riguarda 2 unità , una classificata come Entrata nell'anno t e l'altra come Uscita dall'anno t, se le regole che soddisfano la continuità (i caratteri collegati) sono verificate allora il flusso è spurio e si tratta di una sola unità attiva in t che ha subito eventi di modificazione dei caratteri.

Dai risultati della procedura di RL, il sottoinsieme dei Link definitivi, che contiene le casistiche valutabili ai fini della continuità, è costituito da 6.015 coppie, distribuite per sottopopolazione come segue.

Sottopopolazione	Numero coppie
L+S	1.608
N+L	252
N+L+S	691
N+S	573
l+s	1.102
n+l	1.760
n+s	2
Xxx	27
Totale	6.015

5.3. I legami per duplicazione e i risultati definitivi

Seguendo un altro criterio di scelta, i legami di duplicazione sono quelli che invece legano due o più unità classificate con lo stesso stato di attività e che soddisfano regole di similitudine nei caratteri più restrittive. In particolare le unità devono presentare una "similitudine" in tutti i caratteri, caratteristica che viene soddisfatta solo dalle coppie che presentano un γ =(A,A,A;

P,A,A). Ad esempio due unità attive in t, collegate perché tutti i caratteri sono simili inducono a ritenere le due unità una duplicazione. La definizione di duplicazione è molto restrittiva ma è quella che con un maggior livello di certezza ci consente di apportare delle modificazioni negli stock di imprese, riducendoli.

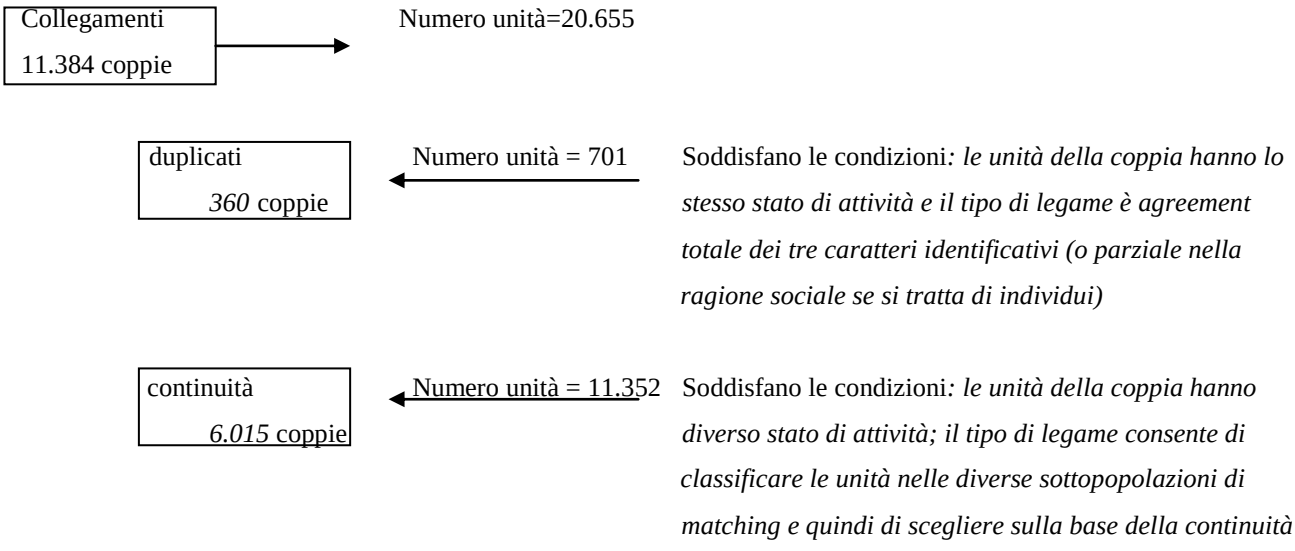
La distribuzione presentata in Tav.5.2 sintetizza i legami, in termini di cluster, coppie e unità, sia per continuità e per duplicazione.

Tav.5.2: *I legami per continuità e duplicazione*

PROVINCE	num_coppie 1			num_coppie 2			num_coppie 3			num_coppie totale		
	num_ cluster	tot_ coppie	tot_rk	num_ Cluster	tot_ coppie	tot_rk	num_ cluster	tot_ coppie	tot_rk	num_ cluster	tot_ coppie	tot_rk
Agrigento	472	472	944	43	86	129	4	12	13	494	545	1.036
Caltanissetta	284	284	568	21	42	63	7	21	22	312	347	653
Catania (Prov)	813	813	1.626	65	130	195	15	45	48	893	988	1.869
Catania (Com)	390	390	780	42	84	126	5	15	15	437	489	921
Enna	165	165	330	21	42	63	3	9	12	189	216	405
Messina	754	754	1.508	73	146	219	16	48	52	843	948	1.779
Palermo (Prov)	404	404	808	31	62	93	10	30	33	445	496	934
Palermo (Com)	721	721	1.442	71	142	213	12	36	37	804	899	1.692
Ragusa	367	367	734	27	54	81	3	9	10	397	430	825
Siracusa	391	391	782	36	72	108	6	18	21	433	481	911
Trapani	432	432	864	32	64	96	5	15	18	469	511	978
Totale Sicilia	5.193	5.193	10.386	462	924	1.386	86	258	281	5.741	6.375	12.053

In sintesi, i risultati dell'individuazione dei legami, per la continuità e per la duplicazione sono illustrati nel seguente schema e sono quelli che vengono analizzati per la demografia.

Schema **A**



6. L'integrazione con altre tecniche

6.1. Le informazioni sulle trasformazioni presenti nell'Anagrafe tributaria

Generalmente le entità economiche nella realtà si modificano nel tempo. Tali trasformazioni prendono il nome di "eventi"; ad esempio un evento di nascita è una modificazione di esistenza in quanto identifica una unità economica che prima di quell'evento non esisteva.

Informazioni sugli eventi di trasformazione giuridica delle imprese sono registrate presso l'archivio "Anagrafe tributaria" del Ministero delle Finanze e nell'ambito della realizzazione dell'archivio statistico Asia, esse sono utilizzabili dopo un opportuno trattamento.

Tra i tipi di evento quelli classificati come ***Cambio di forma giuridica e successione ereditaria*** sono di seguito scelti nel processo di validazione dei risultati del RL e in quello di integrazione per la demografia. Tali tipologie rispecchiano i concetti di continuità e in tal senso alcune descrizioni possono essere utili.

Generalmente una modifica di forma giuridica è caratterizzata da un passaggio (una continuità) di unità i cui codici fiscali cambiano: si tratta di persone fisiche che diventano persone giuridiche (ad esempio un individuo che costituisce una società di persone) e viceversa; meno frequentemente sono passaggi tra società. La caratteristica di questo evento è che la cancellazione/apertura del nuovo codice avviene contemporaneamente e, a meno di errori di registrazione, alla data dell'evento le data cessazione della prima unità e la data inizio della seconda unità coincidono. Anche la ***Successione ereditaria*** presenta stesse modalità di registrazione con la caratteristica che il passaggio avviene principalmente tra codici fiscali di persone fisiche. La struttura dell'informazione è ad esempio la seguente:

Unità1	Tipo evento	Data evento	Unità2
CF1	Successione ereditaria	1998-06	CF2

Con riferimento alla Sicilia, il file acquisito dall'Anagrafe tributaria contiene ***1.002*** coppie con tipologia di evento di nostro interesse, "*Cambio di forma giuridica e successione ereditaria**".***

I passi successivi consistono nel porre a confronto i legami certi delle coppie "eventi" con i legami trovati applicando la procedura di RL.

Step 1 – Non entrano a far parte della procedura di RL 400 coppie degli eventi.

Il motivo della loro assenza è imputabile a 3 ragioni: a) caratteri non normalizzati (21%); b) esclusione per il blocco (19%); c) esclusione per regole di restrizione (60%). In particolare questo ultimo sottoinsieme c) è costituito da legami in cui circa il 41% ha solo l'Ateco uguale o

compatibile; il 15% nessun carattere in comune e per il resto la struttura del vettore delle variabili di confronto è caratterizzata da un solo agreement o partial-agreement.

Step 2 – Entrano nella procedura di RL 602 coppie che vengono classificate in Link (333) e Non-Link (269). Per valutare la qualità di tali risultati, nelle due tabelle che seguono, sono presentate le strutture delle coppie di eventi abbinati con i Link e con i Non link; in particolare per i Non link è possibile identificare le sottopopolazioni di matching di appartenenza e il tipo di legame tra le unità.

In quest'ultimo sottoinsieme, le sottopopolazioni N+S e L+S hanno il peso predominante. Laddove è accettabile l'errore per avere escluso la sottopopolazione L+S (la procedura RL classifica circa l'80% delle coppie tra i Non link), più problematica è invece l'esclusione delle coppie N+S per le quali invece le percentuali di Link e Non link ottenuti dalla procedura RL sono circa uguali.

Tav.6.1: *Struttura delle coppie Eventi/Non Link –numero coppie*

	Legame tra le unità			
Sottopopolazione	1I-1S	2I	2I-1S	Totale
L+S	58	46	0	104
N+S	78	0	2	80
l+s	0	1	0	1
n+l	1	0	0	1
Xxx	44	39	0	83
Totale	181	86	2	269

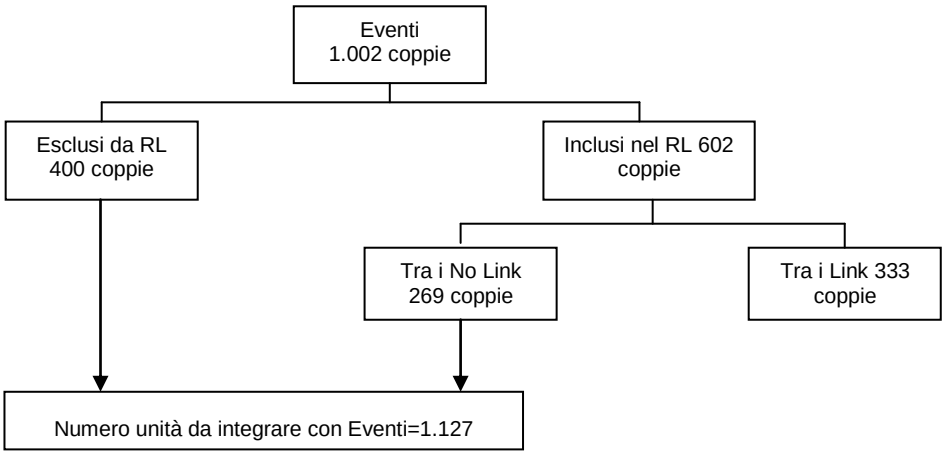
Tav.6.2: *Struttura delle coppie Eventi/Link –numero coppie*

Sottopopolazione	Totale
N+L+S	114
L+S	61
N+L	11
N+S	79
l+s	49
n+l	19
Totale	333

Per quanto riguarda le coppie abbinate con i Link si sottolinea che esse appartengono anche ai Link definitivi, ossia e quelle scelte per la continuità

Il processo di abbinamento tra i risultati del RL e gli eventi è illustrato in sintesi nello schema B.

Schema **B**

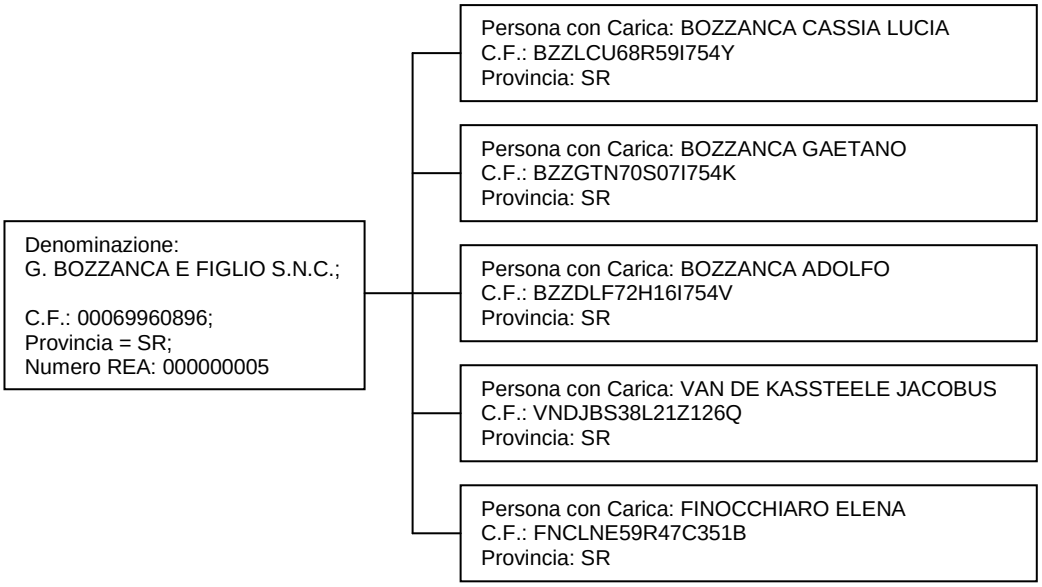


6.2. La struttura delle informazioni dell'archivio "Persona d'Impresa"

L'archivio "Persona d'Impresa" (P.I.), gestito da CCIAA, è una fonte d'informazione sui legami tra le società di Persone e i soci delle stesse. Tale archivio è organizzato per *posizione REA*, la quale identifica un'impresa limitatamente ad una CCIAA. Oltre alle informazioni relative all'impresa, sono indicate anche le Persone con Carica nell'impresa stessa, siano esse fisiche o Società parificate a persona.

Per quanto concerne la regione Sicilia, esso è costituito da **52.409** *posizioni REA*, ad ognuna delle quali sono associate una o più Persone con Carica per un totale di 133.662 records.

L'esempio seguente riporta la struttura dell'archivio:



La suddetta impresa dà luogo ad un'unica *posizione REA*, (identificata dalla sigla Provincia e dal numero REA) ed un solo Codice Fiscale Impresa, a cui corrispondono cinque diverse Persone con Carica, identificate dai rispettivi Codici Fiscali.

Poiché l'interesse è rivolto alla ricerca di legami tra le imprese presenti in Asia'97 e Asia'98, delle **52.409** *posizioni REA* ne sono state selezionate **20.993**, aventi lo stesso riferimento temporale, ovvero una data d'iscrizione inferiore all'anno 1999. A queste 20.993 imprese sono collegati 81.954 soggetti. Di questi ultimi vengono presi in considerazione solo quelli che in Asia'97 e/o in Asia'98 sono registrati come imprese.

Si possono presentare diverse situazioni:

1. L'impresa non ha nessun socio che figura come impresa individuale;
2. L'impresa può avere più di un socio che figura come impresa individuale;
3. Una Persona con Carica, sia essa stessa un'impresa o meno, può essere socio di più di un'impresa. In tal caso è stato necessario ricostruire il collegamento tra le due società di persone che hanno uno o più soci in comune.

Sulla base di queste considerazioni sono state selezionate e ricostruite **12.487** coppie di unità collegate di cui:

a) **11.023** del tipo "Società di persone - Impresa Individuale";

b) **1.464** del tipo "Società di persone - Società di persone".

Il problema che si è posto è stato come identificare i legami tra unità che fossero coerenti con il concetto di legame utile ai fini demografici. In tal senso alle **12.487** coppie sono state applicate delle regole di selezione che possono essere sintetizzate nei seguenti punti:

1. Per le coppie "Società di Persone – Impresa Individuale" valgono i seguenti criteri di scelta del legame:

1.a L'individuo deve ricoprire all'interno della società una delle seguenti cariche:

- Amministratore delegato;
- Amministratore;
- Amministratore e responsabile tecnico;
- Titolare dell'impresa artigiana;
- Amministratore unico;
- Amministratore unico e preposto;
- Socio comproprietario;
- Proprietario;
- Socio accomandatario e preposto;
- Socio amministratore;
- Socio accomandatario;
- Socio contitolare;
- Socio unico;
- Titolare;
- Titolare e responsabile tecnico.

1.b Se l'individuo non ricopre una delle cariche suddette, allora la coppia viene scelta solo se il Cognome dell'Individuo è coincidente a quello di almeno il 60% degli altri soci dell'impresa.

2. Per le coppie "Società di Persone – Società di Persone" vale il seguente criterio di scelta del legame:

2.a Tra due società devono essere presenti non meno del 50% dei soci in comune.

Una volta individuati i legami (8.002 coppie), si è proceduto alla costruzione dei cluster, analogamente ha quanto fatto nella procedura di Record Linkage. Di seguito si riporta la composizione dei cluster per numero di coppie e numero di unità coinvolte:

Tav.6.3: *Composizione dei cluster per numero di coppie e numero di unità*

n° cluster	*N° coppie*	*Tot_coppie*	*Tot_unità*
5.456	*1*	5.456	10.912
818	*2*	1.636	2.454
172	*3*	516	622
79	*>3*	394	373
6.525		**8.002**	**14.361**

I passi successivi consistono nel porre a confronto i legami certi delle 8.002 coppie provenienti da P.I. con i legami trovati applicando la procedura di RL.

Step1. - Dal totale delle 8.002 coppie se ne escludono 227, in quanto già individuati nella fonte delle Finanze: si passa così da 8.002 a **7.775** coppie.

Step2. - **4.944** coppie non entrano a far parte delle procedura di RL per 3 diverse ragioni:

a) una o entrambe le due unità costituenti la coppia non hanno tutti i caratteri normalizzati (11,5%);

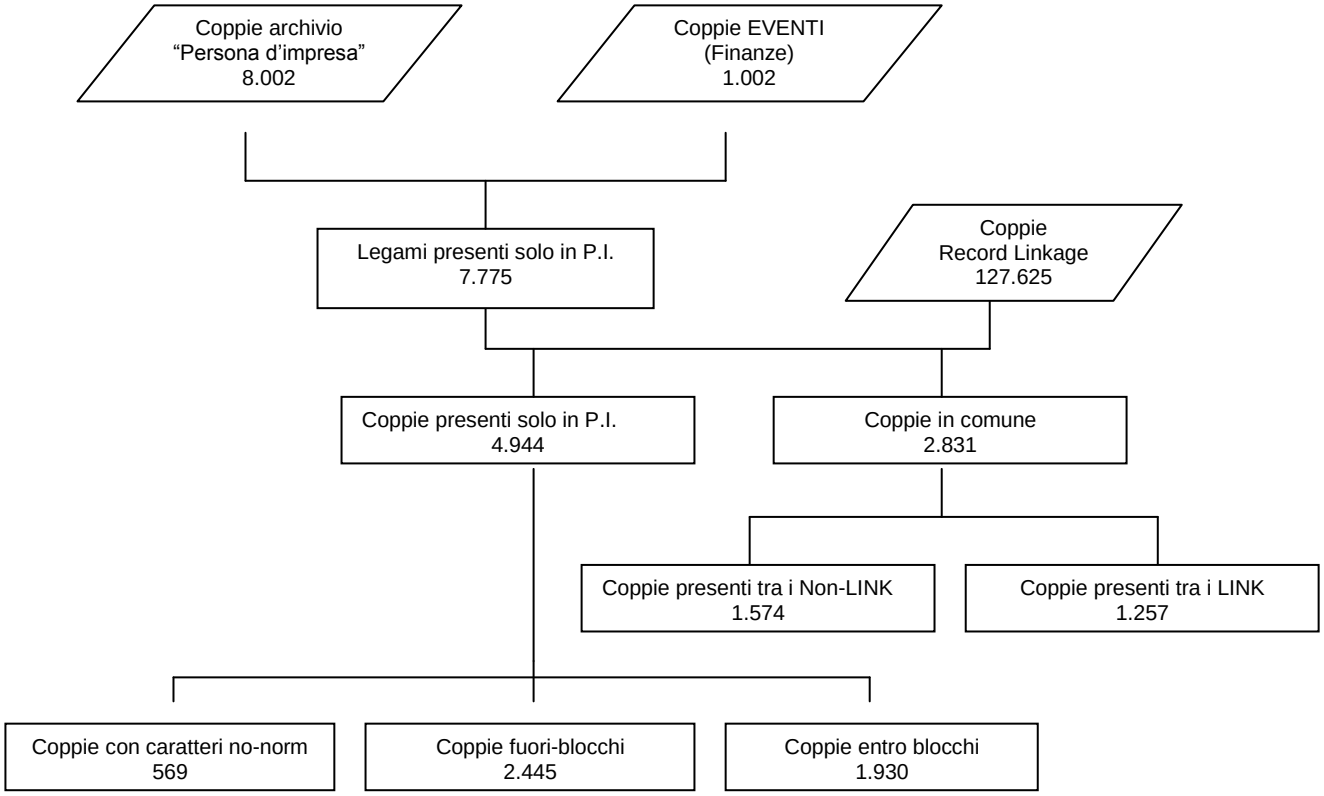
b) le unità non appartengono ad uno stesso blocco (49,5%)[16] (§ 4.2.2.1);

c) infine circa il 39% delle coppie riguardano unità che, pur appartenendo ad uno stesso blocco, non vengono prese in considerazione nella procedura di Record Linkage, perché rientrano tra gli abbinamenti esclusi dalle regole di restrizione dello spazio dei confronti (§ 4.2.2.2).

Step3. - **2.831** coppie entrano nella procedura di RL, di cui il 55,6% appartengono all'insieme dei Non-Link e il 44,4% all'insieme dei Link.

Di seguito si illustrano i risultati e gli esiti degli abbinamenti attraverso lo schema C:

*Schema **C***



Dalle 6.518 coppie non trovate nei Link del RL (4.944 presenti solo nell'archivio P.I. + 1.574 presenti tra i Non-Link), si escludono i cluster più grandi (con un numero di coppie >3) e tutte quelle coppie per le quali una delle unità è coinvolta nelle coppie presenti nei Link del RL[17]. Queste ultime vengono inserite nei cluster del RL e successivamente analizzate.

In tal modo il numero si riduce da 6.518 a **5.831** coppie.

Nella Tav.6.4 si riporta la composizione dei cluster per numero di coppie e numero di unità.

Tav.6.4: *Composizione dei cluster per numero di coppie e numero di unità*

n° cluster	N° coppie	Tot_coppie	Tot_unità
4.304	*1*	4.304	8.608
567	*2*	1.134	1.701
131	*3*	393	473
5.002		**5.831**	**10.782**

[16] In particolare delle 2.445 coppie (pari al 49,5% delle 4.994) il 38% sono legami tra unità appartenenti a province diverse, il 36% riguarda abbinamenti tra unità di diversi comuni e il restante 26% sono legami tra unità che presentano codici di attività economica e C.a.p. diversi.

Volendo effettuare un'analisi più approfondita, utile per identificare i legami di continuità e di duplicazione, è possibile classificare queste 5.831 coppie in sottopopolazioni di matching. Tali sottopopolazioni vengono individuate seguendo gli stessi criteri adottati per la costruzione delle sottopopolazioni in cui sono state classificate le coppie del Record Linkage (§5.1). L'unica eccezione riguarda il confronto sulla Ragione Sociale, a cui, trattandosi di legami certi, è stato imposto per tutte le coppie un agreement (individuato con il carattere **A**).

La Tav.6.5 riporta la distribuzione delle 5.381 coppie classificate secondo le sottopopolazioni di matching individuate con P.I. e con RL.

Tav.6.5*: Distribuzione per sottopopolazioni di matching*

Persona d'Impresa	***Record Linkage***							
	Missing	**L+S**	**N+L+S**	**N+S**	**l+s**	**n+l**	**xxx**	**Totale**
A+L	114	0	0	0	0	4	293	411
A+L+S	94	145	1	0	0	0	0	240
A+S	600	0	0	289	2	0	852	1743
A+l	45	0	0	0	0	14	150	209
Xxx	1908	0	0	0	0	0	1320	3228
Totale	**2761**	**145**	**1**	**289**	**2**	**18**	**2615**	**5831**

Facendo riferimento alle sole coppie non classificate missing per il RL (si ricorda che queste ultime appartengono ai sottoinsiemi delle "Fuoriblocco" e "Non Normalizzate"), è possibile evidenziare quanto segue.

Tav.6.6: *Distribuzione per sottopopolazioni di matching e sottoinsieme di provenienza*

	ENTROB	**NOLINK**	**Totale**
L+S	0	145	**145**
N+L+S	0	1	**1**
N+S	0	289	**289**
l+s	0	2	**2**
n+l	0	18	**18**
Xxx	1664	951	**2615**
Totale	**1664**	**1406**	**3070**

Ad esclusione delle 2.615 coppie escluse dal RL con prevalenza di Disagreement (almeno 2 caratteri su 3), problemi si riscontrano per le sottopopolazioni L+S e N+S. Analogamente a quanto è emerso nell'analisi degli eventi delle Finanze, laddove è accettabile l'errore per l'esclusione della

[17] Ad esempio, si supponga che la coppia A-B è presente nei Link (1.257); si supponga anche che esista la coppia A-C nel sottoinsieme dei "Fuoriblocco" (2.445). Quest'ultima coppia viene analizzata insieme alla coppia A-B e quindi

L+S (la procedura di Record Linkage classifica circa l'80% delle coppie come Non-Link), appare più problematica l'esclusione della composizione N+S per la quale la procedura partiziona esattamente la popolazione in Link (50%) e Non-Link (50%).

6.3. I risultati sulla demografia: l'impatto dei legami trovati sulle poste demografiche

La struttura della popolazione di imprese attive Asia98/Sicilia subisce modificazioni per effetto dell'inclusione dei legami tra le unità ottenuti attraverso le diverse tecniche di integrazione: il processo di RL; gli eventi delle Finanze e i legami della CCIAA.

Con riferimento ai risultati del RL, parte delle unità attive (A) si riduce per effetto dell'inclusione dei legami di duplicazione o perché collegate per match multipli; tutte le unità classificate come Entrate/Uscite (E/U) ma collegate con unità Attive vengono escluse dall'insieme delle rispettive poste iniziali; inoltre i collegamenti tra una E con una U comportano la riduzione delle rispettive poste iniziali e l'inclusione di una unità nello stock delle A.

La relazione demografica che lega lo stock di imprese attive nel 1998 e quello nel 1997 è quella espressa nella equazione (1); tale equazione subisce variazioni in ciascuna delle poste quando si includono i legami.

1) Legami individuati dal processo di RL

Come descritto nello schema A, risultano collegate, rispettivamente per continuità e per duplicazione, 11.352 e 701 unità. Tra di esse alcune unità hanno uno stato di attività che non incide nè sulle poste demografiche né sullo stock (le unità classificate come altro, §.4.1.) di conseguenza le unità da analizzare diventano in totale **10.849**. Nella tav.6.7 viene evidenziata la struttura delle unità distinguendo tra le unità classificate nella popolazione iniziale (Asia98), le unità collegate per continuità/duplicazione da escludere e da includere, e le unità nella popolazione finale (Asia98*).

Tav.6.7: *Classificazione delle unità con legami scelti*

	A	E	U	Totale
Unità in Asia98	5.837	2.728	2.284	10.849
Unità da escludere	896	2.500	2.190	5.586
Unità da includere	462			462
Unità in Asia98*	5.403	228	94	5.725

Se riscriviamo l'equazione demografica (1) in termini di stock e flussi riclassificati a seconda dello stato di attività (si ricorda che A corrisponde alle imprese attive nei due anni '98 e '97) essa diventa:

esclusa dalle 6.518.

Asia98=Asia97 + Entrate – Uscite $\Rightarrow$ [A+E] = [A+U] + E – U (1bis)

Sostituendo alle singole poste quelle modificate per effetto dell'inclusione dei legami, l'equazione (1bis) si modifica secondo lo schema 1.

*Schema **1***

	A+E	A+U	E	U
Equazione iniziale	248.618 =	243.153 +	43.279 -	37.814
	Equazione sulle unità con legami scelti da RL			
Iniziale	8.565	8.121	2.728	2.284
Finale	5.631	5.497	228	94
Equazione finale	A+E	A+U	E	U
	245.684 =	240.529 +	40.779 -	35.624
Differenza %	***- 1,18***	***- 1,08***	***- 5,78***	***- 5,79***

Per effetto dell'inclusione dei legami scelti con il RL la popolazione iniziale (Asia98) si riduce del 1,18%, le componenti demografiche di Entrate e Uscite di circa il 6%.

a) Legami individuati sulla base delle informazioni Finanze

Come risultato dell'integrazione delle coppie "eventi" (schema B), le unità collegate per aver subito un evento che non appartengono al sottoinsieme di quelle collegate per RL, e che per stato di attività hanno un impatto sulle poste demografiche, corrispondono a 1.127 unità della popolazione iniziale

Tav.6.8: *Classificazione delle unità con legami da eventi*

	A	E	U	Totale
Unità Asia98/eventi	603	402	122	1.127
Unità da escludere	46	359	118	523
Unità da includere	75			75
Unità in Asia98*	632	43	4	679

L'equazione demografica, per l'inclusione dei legami da eventi, si modifica secondo il seguente schema 2.

*Schema **2***

	A+E	A+U	E	U
Equazione iniziale	248.618 =	243.153 +	43.279 -	37.814
	Equazione sulle unità con legami scelti dal RL			
Iniziale	8.444	8.029	2.687	2.272
Finale	5.568	5.437	225	94
	Equazione sulle unità con legami da Finanze			
Iniziale	1.005	725	402	122
Finale	675	636	43	4
Equazione finale	A+E	A+U	E	U
	245.412 =	240.472 +	40.458 -	35.518
Differenza %	***- 1,29***	***- 1,10***	***- 6,52***	***- 6,07***

b) Legami individuati sulla base delle informazioni CCIAA

Come risultato dell'integrazione delle coppie CCIAA (schema C), le unità collegate (10.782) e che hanno un impatto sulle poste demografiche sono 9.611.

L'equazione demografica, per l'inclusione dei legami da CCIAA, si modifica secondo il seguente schema 3.

*Schema **3***

	A+E	A+U	E	U
Equazione iniziale	248.618 =	243.153 +	43.279 -	37.814
	Equazione sulle unità con legami scelti dal RL			
Iniziale	8.444	8.029	2.687	2.272
Finale	5.568	5.437	225	94
	Equazione sulle unità con legami da Finanze			
Iniziale	1.005	725	402	122
Finale	675	636	43	4
	Equazione sulle unità con legami da CCIAA			
Iniziale	8.432	7.886	1.725	1.179
Finale	4.820	4.688	282	150
	A+E	A+U	E	U
Equazione finale	241.800 =	237.274 +	39.015 -	34.489
Differenza %	***-2,74***	***-2,42***	***-9,85***	***-8,79***

Per effetto dell'inclusione dei legami individuati con l'integrazione di tutte le metodologie adottate i dati la popolazione iniziale (Asia98) si riduce del 2,7%, le Entrate del 9,8% e le Uscite dell'8,8%.

7. Conclusioni e sviluppi futuri

Come già precedentemente sottolineato il carattere dell'applicazione sviluppata è del tutto preliminare e l'obiettivo è per il momento quello di fornire materiale organizzato che consenta di definire l'apporto di procedure di linkage probabilistico alla risoluzione del problema in studio: l'individuazione di eventi spuri di demografia d'impresa. Le considerazioni generali emerse dal lavoro svolto possono essere riassunte nei seguenti tre punti:

- Avere come riferimento un quadro concettuale e definitorio chiaro, in particolare con riferimento al tema della demografia d'impresa, è condizione necessaria per produrre statistiche coerenti e confrontabili almeno a livello dell'Unione Europea. Non è però condizione sufficiente, è necessario stabilire anche un insieme non ambiguo di criteri operativi e di tecniche statistiche utilizzabili. Differenti criteri e tecniche producono risultati a volte fortemente dissimili.
- La tecnica delle similitudine dei caratteri identificativi, applicata utilizzando metodologie probabilistiche di RL, utilizzando un numero limitato di variabili identificative (ragione sociale, localizzazione, attività economica) da sola non è sufficiente ad individuare la continuità fra due unità legali. La sperimentazione ha dimostrato la necessità di integrare tale tecnica con altre, in primo luogo, quella che utilizza la "conpresenza dei titolari".
- Per la prima volta, in Italia e forse in Europa, una procedura di RL probabilistica è stata applicata ad un numero notevole di record di unità economiche. La difficoltà di applicare tale metodologia per i record di impresa è stata più volte sottolineata in letteratura, a causa dello scarso potere discriminante delle variabili che possono essere utilizzate. La sperimentazione effettuata ha dimostrato la possibilità di utilizzare, con costi risorse e tempi accessibili, un approccio probabilistico all'accoppiamento di record di impresa con la condizione di sviluppare da un lato precise procedure di parsing e di normalizzazione delle variabili di matching (in primo luogo della denominazione dell'impresa) e dall'altro una attenta analisi dei risultati basata su evidenze di tipo economico.

Con particolare riferimento alla tecnica di abbinamento esatto e alla procedura utilizzata è possibile individuare alcuni aspetti specifici che richiedono un maggiore approfondimento:

1. Il problema della scelta delle variabili di blocco. La scelta di variabili geografiche come il cap ha influenzato notevolmente i risultati; infatti la qualità dei dati usati in questa sperimentazione non era garantita per quanto riguarda i cap. La strategia di miglioramento di questo problema è

nella direzione di applicare delle procedure preliminari che consentano di elevare la qualità di queste variabili (esempio uso di stradari, cappari, provenienti da altre fonti amministrative quali le Poste Italiane). La strategia di blocco, e cioè quella di trovare legami di continuità all'interno dello stesso comune, rimarrà invariata nelle future implemetazioni; trasferimenti di unità di imprese tra comuni non potranno essere un obiettivo perseguibile da questa procedura.

2. Al momento la procedura non produce stime sugli errori di prima e di seconda specie. Lo sviluppo di questo step è di fondamentale importanza per valutare la qualità delle stime e di conseguenza la qualità dei risultati che si possono ottenere.
3. Nella fase di sperimentazione si è riscontrato che la stima delle u, basata su un campione casuale e quindi esterna al processo iterativo di stima usato invece per le m, ha evidenziato alcuni problemi di rappresentatività per i singoli blocchi; infatti l'estrazione casuale di coppie su cui stimare le u avviene senza blocchi . Un problema che è sorto è stato quello di scegliere l'universo di riferimento di tutte le coppie possibili su cui campionare. Tale scelta è ricaduta su un campionamento per province, ma rimane aperta la problematica più generale che si pone quando la procedura deve essere utilizzata a livello nazionale (103 campioni uno per ogni provincia?). In generale è stato notato che all'aumentare della numerosità della popolazione dell'area geografica si poneva il problema di aumentare in maniera adeguata la numerosità campionaria. Un'idea è quella di utilizzare campioni casuali stratificati tenendo conto dei blocchi.

Nel prossimo futuro è necessario sviluppare l'ingenerizzazione e generalizzazione della procedura che permetta la sua applicazione ad altri ambiti di studio, quali individuazione delle duplicazioni di record o la lettura integrata di informazioni acquisibili in differenti basi informative, anche attraverso l'utilizzo di solo parti di essa (ad esempio utilizzare la fase di parsing e normalizzazione per lo sviluppo di uno stradario, o per la correzione ed imputazione automatica dei cap). La generalizzazione deve prevedere la possibilità di operare scelte alternative nella individuazione delle variabili di *matching* e/o di blocco e nella definizione delle regole di *agreement/disagreement*.

Riferimenti bibliografici

Belin, T.R., Rubin, D.B. (1995). A method for calibrating false-match rate in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Copas, J.B., Hilton, F.J. (1990). Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society A*, 153, 3, 287-320.

Dempster, A.P., Laird, N.H., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B, 39, 1-38.

Fellegi, I.P., Sunter, A.B. (1969). A Theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Garofalo, G., Viaviano, C .(1999) Continuity rules: re-delineation in the Italian context. In *13th Internation Roundtable on Business Survey Frames, Paris.*

Garofalo, G., Viaviano, C .(2000) The problem of links between legal units: statistical techniques for the enterprise identification and the analysis of continuity. *In Quaderni di Ricerca – Istat N° 1/2000*

Garofalo, G., Paggiaro, A., Torelli, N., Viviano, C. (2001). A record linkage procedure for the management and the analysis of the Italian statistical business register. In *ICES-II, Proceedings of the Second International Conference on Establishment Surveys (Survey Methods for Businesses, Farms, and Institutions, June 17-21, 2000, Buffalo, New York)*, American Statistical Association, Alexandria, Virginia, pp. 1612-1617.

Giusti, A., Marliani, G., Torelli, N. (1991). Procedure per l'abbinamento dei dati individuali delle forze di lavoro. In Trivellato, U. (a cura di), *Forze di Lavoro: Disegno dell'Indagine e Analisi Strutturali*. ISTAT, Annali di Statistica, 9, 11.

Jabine, T.B., Scheuren, F.J. (1986). Record linkage for statistical purpose: methodological issues. *Journal of Official Statistics*, 2, 3, 255-277.

Jaro, M.A. (1989). Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 89, 414-420.

Kelley, R.P. (1985). Advances in record linkage methodology: a method for determining the best blocking strategy. In Kills, B. e Alvey, W. (eds.), *Record Linkage Techniques-1985, Proceedings of the Workshop on Exact Matching Methodologies*, Statistics of Income Division, U.S. Internal Revenue Service, 1299, 2-86.

Kills, B., Alvey, W. (eds.) (1985). *Record Linkage Techniques-1985, Proceedings of the Workshop on Exact Matching Methodologies*. Statistics of Income Division, U.S. Internal Revenue Service, 1299, 2-86.

Kirkendall, N.J. (1985). Weights in computer matching: applications and an information theoretic point of view. In Kills, B. e Alvey, W. (eds.), *Record Linkage Techniques-1985, Proceedings of the Workshop on Exact Matching Methodologies*, Statistics of Income Division, U.S. Internal Revenue Service, 1299, 2-86.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. Oxford University Press.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

Paggiaro, A., Torelli, N. (1999). Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro, *working paper n. 15, ottobre 1999, progetto di ricerca MURST "Lavoro e disoccupazione: questioni di misura e di analisi*.

Tepping, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. *Survey Methodology*, 19, 31-38.

Torelli, N. (1998). Integrazione di dati mediante tecniche di abbinamento esatto: sviluppi metodologici e aspetti applicativi. *Atti della XXXIX riunione scientifica SIS*.

Torelli, N., Paggiaro, A. (1999). La Stima della Quota di Errori in Procedure di Abbinamento Esatto. *Atti del Convegno SIS 99: Verso i censimenti del 2000*.

Winkler, W.E. (1995). Matching and record linkage. In Cox, B.G. et al. (eds.), *Business Survey Methods*, New York, J. Wiley.

Winkler, W.E. (????) Frequency-based matching in the Fellegi-Sunter model of record linkage,U.S. Bureau of Census.