Lavoro e disoccupazione: questioni di misura e di analisi

Progetto di ricerca cofinanziato dal Ministero per l'Università
e la Ricerca Scientifica e Tecnologica - Assegnazione: 1998
Coordinatore: Ugo Trivellato

Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro

Adriano Paggiaro, Nicola Torelli

Dip. di Scienze Statistiche, Univ. di Padova

Working Paper n. 15 ottobre 1999

Unità locali del progetto:

Dip. di Economia Politica, Univ. Di Modena	(coord. Michele Lalla)
Dip. di Economia "S. Cognetti De Martiis", Univ. di Torino	(coord. Bruno Contini)
Dip. Di Statistica, Univ "Ca' Foscari" di Venezia	(coord. Tommaso Di Fonzo)
Dip. di Metodi Quantitativi, Univ. di Siena	(coord. Achille Lemmi)
Dip. di Scienze Statistiche, Univ. di Padova	(coord. Ugo Trivellato)

Dip. di Scienze Statistiche
via S. Francesco 33, 35121 Padova

1. Introduzione

La possibilità di condurre analisi dinamiche dei comportamenti dell'offerta di lavoro è strettamente legata alla disponibilità di informazioni longitudinali sulla popolazione di interesse; tuttavia, non sempre i dati con periodo di riferimento differente sono disponibili in un unico archivio, ed è pertanto necessario collegare informazioni sulla medesima unità contenute in archivi differenti. Una preziosa risorsa per ottenere informazioni longitudinali, è costituita dai dati raccolti in indagini correnti sulle forze di lavoro nelle quali il disegno di rilevazione prevede la rotazione del campione. In tali casi, seppure l'obiettivo di primario interesse è quello di ottenere, a partire dai dati sezionali, misura e descrizione delle caratteristiche e delle cadenze dell'occupazione e della disoccupazione, per una parte delle unità campionarie si hanno informazioni sulla condizione rispetto al lavoro in più occasioni di indagine successive. Per una discussione più generale sulle opportunità per l'analisi dinamica del mercato del lavoro offerte da indagini che utilizzano campioni ruotati, si veda Trivellato e Torelli (1989).

Le indagini campionarie che prevedano una struttura longitudinale sono solitamente organizzate in modo che risulti immediato riconoscere quali siano i record relativi nel tempo al medesimo individuo, per cui la definizione del campione longitudinale richiede unicamente la disponibilità di algoritmi efficienti di ricerca dei record che presentano, in archivi riferiti a tempi diversi, la medesima chiave identificativa. Nel caso della Rilevazione Trimestrale delle Forze di Lavoro (RTFL) effettuata per l'Italia dall'Istat, tuttavia, un identificatore per ciascun record individuale non esiste e la chiave che identifica la medesima famiglia in occasioni di indagine diverse è, in un numero non trascurabile di casi, riportato erroneamente. Per la creazione di un archivio longitudinale, quanto più esaustivo e privo di duplicazioni ed errori, è quindi necessario ricorrere a metodi per l'abbinamento esatto di record che mirano a reperire dati relativi alla medesima unità contenuti in due o più archivi.

D'altra parte, l'impiego di tecniche di abbinamento esatto viene invocato sempre più spesso sia come parte integrante dei processi di collezione e organizzazione dei dati che nella fase di controllo delle operazioni sul campo in indagini complesse e su larga scala, e procedure di abbinamento esatto di record vengono utilizzate, ad esempio, quale operazione preliminare per condurre analisi del grado di copertura in rilevazioni censuarie oppure per integrare dati campionari e dati provenienti da fonte amministrativa.

Le crescenti capacità computazionali dei calcolatori hanno consentito negli ultimi anni lo sviluppo di tecniche che, oltre ai notevoli vantaggi in termini di velocità, efficienza e riproducibilità dei risultati, consentono di raggiungere, e talvolta superare, le caratteristiche di elasticità e accuratezza che fino a tempi recenti rendevano più affidabile il controllo manuale.

Una prima formalizzazione degli aspetti statistici legati alle procedure di abbinamento esatto si ha con 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. Dettagliate rassegne dei singoli problemi da affrontare in una procedura di abbinamento, nelle varie fasi di definizione, di abbinamento vero e proprio e di analisi dei risultati, si trovano in Kills e Alvey (1985) e Newcombe (1988). 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.

Nel presente lavoro, si presenta, con qualche dettaglio, un'applicazione di una procedura per l'abbinamento esatto per la costruzione di un archivio longitudinale per i dati della RTFL. Tale procedura incorpora alcuni dei risultati più recenti relativi alla teoria dell'abbinamento esatto e quindi costituisce una generalizzazione e un miglioramento delle procedure già proposte al medesimo scopo (Moriani, 1981; Giusti, Marliani e Torelli, 1991). In particolare, nel paragrafo 2 si introducono i metodi di abbinamento esatto, definendo le caratteristiche delle informazioni richieste e delle differenti strategie possibili, con attenzione a metodi probabilistici con algoritmi iterativi di stima. Nel paragrafo 3 si definiscono le caratteristiche specifiche di un'applicazione di tali metodi ai dati di successive rilevazioni della RTFL tenendo conto delle novità che hanno interessato l'indagine negli ultimi anni. Nel paragrafo 4, infine, si presentano i risultati dell'applicazione della procedura di abbinamento proposta, con un'analisi delle caratteristiche del campione longitudinale ottenuto e dell'efficienza degli specifici metodi utilizzati.

2. Metodi per l'abbinamento esatto di record

2.1. Definizione di un problema di abbinamento esatto

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 ai medesimi individui; 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. I mancati abbinamenti sono invece, almeno in tale situazione applicativa, considerati meno gravi. La più immediata conseguenza dei mancati abbinamenti si traduce, infatti, solo in una non eccessiva riduzione della dimensione campionaria. E' comunque da valutare con attenzione la possibilità che agli errati abbinamenti sia associato un problema di selettività del campione; ciò avviene se gli individui vengono esclusi dall'abbinamento per motivi correlati con quelli di interesse nell'analisi: si pensi ad esempio alla maggior probabilità di commettere errori nella risposta per individui anziani e/o con un basso livello di istruzione.

2.2. 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.

2.3. I pesi di abbinamento: metodi deterministici e probabilistici

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 dev'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.

2.4. Stima dei pesi con 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).

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; 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 la stima di *m*, l'ipotesi di indipendenza fra gli errori nelle singole variabili appare invece più realistica (Thibaudeau, 1993) e permette notevoli semplificazioni computazionali, pur non essendo una scelta obbligata per il metodo proposto. Per una descrizione del significato di queste assunzioni in un'applicazione empirica si veda il sottoparagrafo 4.1.

Con l'assunzione di indipendenza, i valori di *m* per le singole variabili 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:

$$\hat{m}^i=\frac{\sum_{j=1}^{N}\gamma_j^i\hat{g}_j}{\sum_{j=1}^{N}\hat{g}_j},i=1...I. \qquad (4)$$

La stima di *m* per le singole coppie, sempre per l'ipotesi di indipendenza, avviene in modo moltiplicativo, utilizzando le stime provenienti dalle (4) a seconda dei risultati del confronto presenti nei vettori γ:

$$\hat{m}_j=\prod_{i=1}^{I}\left(\hat{m}^i\right)^{\gamma_j^i}\left(1-\hat{m}^i\right)^{1-\gamma_j^i}. \qquad (5)$$

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*:

$$\hat{p} = \frac{\sum_{j=1}^{N} \hat{g}_j}{N}. \qquad (6)$$

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 (5) per tutti i vettori γ che si rivelano identici. Inoltre, anche l'utilizzo di (4) e (6) 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.

2.5. Scelta della soglia e stima degli 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 a quali no. Belin e Rubin (1995) osservano, attraverso alcune verifiche empiriche sui risultati dell'abbinamento, che l'effetto maggiore sugli errori di abbinamento è legato ad una cattiva scelta della soglia, mentre sembrano di minor rilievo la fase di definizione e stima dei pesi. La soglia deve pertanto 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 primo semplice metodo di verifica possibile è un controllo manuale effettuato su un campione di record; in particolare, se il sistema di confronti permette di discriminare con buona precisione le coppie, risulta spesso sufficiente prevedere un controllo limitato alle coppie con pesi "vicini" alla soglia, la cui assegnazione è più dubbia.

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.

Una recente alternativa al controllo manuale è fornita da metodi legati alla modellazione statistica degli errori di abbinamento. Belin e Rubin (1995) presentano metodi che si basano, seppure con interpretazioni differenti, su una stima che tenga conto della presenza della variabile latente *G*, relativa alla classificazione delle singole coppie in M o U. Si distinguono in particolare due classi di modelli: (a) un approccio diretto che prevede una regressione logistica della variabile *G* sui pesi *w*; (b) un approccio indiretto basato sulla stima di un modello di mistura che tenga conto delle due distinte distribuzione condizionate dalle quali è formata la distribuzione osservata dei pesi.

Un punto vincolante della formulazione di Belin e Rubin (1995) è che in entrambi i metodi è richiesta la disponibilità di un campione di abbinamenti certi, a partire dal quale poter stimare alcuni parametri da utilizzare nella stima degli errori. Tale assunzione, oltre a non essere spesso attuabile nella pratica per la mancanza di tale campione, pone delle restrizioni forti sulla somiglianza delle differenti situazioni di abbinamento. In particolare, per la regressione logistica l'assunzione è che i parametri che legano *G* ai pesi *w* siano gli stessi per tutte le procedure di abbinamento, e si utilizzano le stime sul campione iniziale per abbinare il campione di interesse. Per il modello di mistura, l'assunzione riguarda invece il rapporto fra le varianze delle due distribuzioni condizionate dei pesi, oltre ai parametri delle trasformazioni Box-Cox necessarie per renderle normali; in particolare la prima ipotesi appare restrittiva, alla luce delle forme differenti che assumono le distribuzioni dei pesi al variare, ad esempio, delle strategie di blocco o della proporzione di individui potenzialmente abbinabili nei due archivi.

Winkler (1995) sostiene che il metodo di Belin e Rubin, oltre a dipendere fortemente dal campione utilizzato per le stime iniziali, fornisce risultati soddisfacenti solo nel caso particolare in cui l'ipotesi di indipendenza fra gli errori non sia troppo restrittiva e le due distribuzioni condizionate dei pesi risultino ben distinte. Per una stima più precisa delle probabilità di errore nei frequenti casi in cui tali assunzioni non sono verificate, fra le altre proposte Winkler suggerisce alcune restrizioni di convessità nello spazio parametrico, che consentono di limitare le possibili soluzioni a quelle ritenute più realistiche e velocizzare la massimizzazione della verosimiglianza.

Torelli e Paggiaro (1999) propongono invece un'alternativa che permette di stimare i parametri direttamente attraverso l'algoritmo EM proposto nel sottoparagrafo 2.4. In particolare, il legame logistico fra probabilità di abbinamento e pesi *w* riscontrato nella (3) fa propendere per un metodo diretto di stima, che a differenza di quello analizzato da Belin e Rubin consente di allentare l'ipotesi di indipendenza. I risultati principali, ottenuti anche a partire da alcuni esperimenti di simulazione, confermano che la sottostima degli errori osservata da Belin e Rubin dipende essenzialmente dalla qualità della stima dei pesi. In particolare, si osserva una buona stima della quota di errati abbinamenti se si considera la dipendenza fra le variabili nella stima di *u*, mentre vi è un'evidente sottostima con l'assunzione di indipendenza.

3. Procedure per l'abbinamento esatto nella RTFL

3.1. La Rilevazione Trimestrale delle Forze di Lavoro

Al fine di ottenere periodicamente una fotografia dettagliata della situazione dell'offerta del lavoro in Italia, l'Istat conduce dal 1959 la Rilevazione Trimestrale delle Forze di Lavoro. La RTFL prevede un sistema di rotazione del campione attraverso il quale è possibile ottenere informazioni longitudinali.

Nonostante non vengano rilevate nell'indagine trimestrale informazioni su alcune variabili cruciali per la modellizzazione dei comportamenti dell'offerta di lavoro (ad esempio, dati su redditi e ricchezza), la RTFL è un punto di riferimento fondamentale per una qualunque analisi dell'offerta di lavoro in Italia. Trivellato (1991) fornisce un'analisi approfondita dei possibili utilizzi dei dati della RTFL per l'analisi del mercato del lavoro.

Nella sua forma attuale, dopo le ultime modifiche apportate nel 1992, la RTFL prevede un disegno di campionamento a due stadi: le unità di primo stadio sono i Comuni (ne vengono selezionati circa 1.300) stratificati secondo la loro dimensione demografica; le unità di secondo stadio sono le famiglie così come risultano dagli elenchi anagrafici. Complessivamente, ad ogni occasione di indagine il campione comprende approssimativamente 70.000 famiglie (corrispondenti a circa 200.000 individui). Per un resoconto delle modifiche apportate alla RTFL negli anni si veda Di Pietro (1993), mentre un'analisi approfondita dell'impatto delle novità più recenti si trova in Casavola e Sestito (1994).

L'estrazione delle unità di secondo stadio avviene in modo sistematico dagli elenchi delle anagrafi comunali: in aprile, all'inizio di ogni ciclo annuale, in ogni comune vengono formati i 9 elenchi di famiglie che compongono le sezioni di rotazione; viene inoltre formato un elenco suppletivo di famiglie per l'eventuale sostituzione di quelle unità che non sono reperibili o che rifiutano di collaborare. Il sistema di rotazione è del tipo 2-2-2: ogni famiglia viene intervistata per due trimestri consecutivi, esce dal campione per i successivi due, e rientra per altre due occasioni, per un totale di 4 interviste nell'arco di 15 mesi. Con questa struttura, in linea teorica sarebbe possibile seguire il 50% del campione per due trimestri successivi o a distanza di un anno, mentre per ogni sezione di rotazione (il 25% del campione in ogni singola rilevazione) è possibile l'abbinamento per tutti i 4 trimestri nei quali le famiglie appartenenti alla sezione stessa vengono intervistati.

La struttura di rotazione del campione della RTFL è stata sfruttata dall'Istat, dal 1979 al 1991, per la produzione di stime sui flussi all'interno del mercato del lavoro (Moriani, 1981), con pubblicazione di matrici di transizione trimestrali e annuali a livello nazionale. Altre interessanti applicazioni che prevedono l'utilizzo di procedure di abbinamento esatto sulla RTFL sono in Giusti, Marliani e Torelli (1991) e Favro-Paris, Gennari e Oneto (1996); le strategie di abbinamento adottate in questi lavori sono analizzate con qualche dettaglio nel seguito.

3.2. L'abbinamento nella RTFL: scelta delle variabili di confronto

I problemi principali per identificare i record attribuibili allo stesso individuo in occasioni successive derivano da alcune caratteristiche organizzative dell'indagine. In particolare: (i) l'unità campionaria al secondo stadio è la famiglia, e non esiste un identificatore univoco e invariante nel tempo associato agli individui presenti nella famiglia; (ii) le informazioni utilizzabili per identificare gli individui non sono completamente affidabili perché possono essere affette, a seconda dei casi, da errori nella fase di raccolta dei dati (errori nell'attribuzione dei codici identificativi delle famiglie o nella rilevazione) o alle fasi successive (trascrizione, registrazione, correzioni da parte del piano di compatibilità, ecc.). Risulta pertanto necessario l'utilizzo dei metodi di abbinamento esatto di record descritti nel paragrafo precedente, al fine di utilizzare al meglio la capacità discriminante dei dati disponibili. Le informazioni che appaiono utili per la procedura di abbinamento sono le medesime definite in Giusti, Marliani e Torelli (1991), con l'unica differenza dell'esclusione di "condizione professionale" e "precedenti lavorativi", considerate poco discriminanti sulla scorta di verifiche preliminari effettuate su alcuni sottocampioni relativi a Lombardia e Campania per i primi due trimestri del 1996. Le variabili così individuate sono classificabili in tre tipologie:

1) Codifiche Istat per individuare la famiglia da inserire nel campione:
   - regione;
   - provincia;
   - comune;
   - sezione di rotazione;
   - codice familiare.
2) Caratteristiche individuali invarianti nel tempo:
   - sesso;
   - giorno di nascita;
   - mese di nascita;
   - anno di nascita.
3) Caratteristiche che variano solo in direzioni prefissate:
   - relazione con l'intestatario;
   - stato civile;
   - titolo di studio.

La procedura di abbinamento utilizzata originariamente dall'Istat (Moriani, 1981) prevedeva una concordanza totale su un numero ridotto di variabili, con un criterio apparentemente molto restrittivo che non ammetteva alcun errore; il ridotto numero e la ridotta capacità discriminante delle variabili di confronto utilizzate (si segnala in particolare l'utilizzo della sola età in luogo della data di nascita) non consentiva però la protezione desiderata sul fronte dei falsi positivi. Favro-Paris, Gennari e Oneto (1996) suggeriscono una procedura che, mantenendo le caratteristiche di rigidità della precedente, utilizza solamente le variabili ai primi due punti della lista; il numero ridotto di variabili è compensato dal recupero dell'intera data di nascita, che consente una maggiore protezione dai falsi positivi. Come si vedrà in seguito, comunque, tale protezione non è completa, e la scelta di non ammettere errori nelle variabili comporta

un notevole numero di mancati abbinamenti a fronte di dubbi vantaggi sul rischio di falsi abbinamenti.

Per quanto riguarda la definizione dei vettori di confronto, in questo lavoro le scelte generalizzano in parte quelle di Giusti, Marliani e Torelli (1991), che utilizzano variabili dicotomiche prevedendo concordanza perfetta per tutte le variabili, tranne quelle per cui si possono codificare a priori cambiamenti ammissibili nel tempo: stato civile (discordanza solo se il potenziale individuo appare divenire "celibe o nubile" dopo essere stato sposato) e titolo di studio (ammessi vari tipi di "promozione"). La generalizzazione prevede, oltre alle codifiche di concordanza perfetta e discordanza completa, la possibilità di un codice che identifichi la concordanza parziale. Verifiche preliminari hanno consentito di individuare tipologie specifiche di risultati del confronto in coppie relative alla medesima unità per le quali non vi era perfetta concordanza. Si verifica chiaramente che lo specifico risultato del confronto contiene informazioni non trascurabili ai fini delle decisioni di abbinamento e utili per una stima efficiente dei pesi. In particolare, in questo contesto codifiche di concordanza parziale sono opportune per la data di nascita (quando vi è concordanza su una delle due cifre), per la relazione con l'intestatario (se una delle due risposte è "altro parente o affine") e per il titolo di studio (titoli non ammissibili ma "contigui").

Si segnala, infine, che sarebbero possibili ulteriori affinamenti nella definizione delle codifiche, anche alla luce del fatto che alcune informazioni sono fortemente influenzate dal piano di compatibilità che l'Istat utilizza per rendere coerenti i dati a livello sezionale (ci si riferisce, in particolare, al piano adottato fino ad aprile 1999). Dall'insieme dei risultati dell'abbinamento vi è una certa evidenza che una variabile chiave per il piano di compatibilità è l'anno di nascita ed è sulla base di questa che vengono corrette altre eventuali incoerenze; è pertanto evidente che un errore sulla data di nascita si può ripercuotere anche sulle altre variabili, causando quasi certamente l'eliminazione dell'individuo in una qualunque procedura di abbinamento, con un conseguente irrimediabile aumento dei falsi negativi. Una possibile soluzione a questi problemi potrebbe derivare dall'abbinamento dei dati grezzi raccolti nell'indagine, prima che siano analizzati dal piano di compatibilità; tuttavia, oltre alla necessità di prevedere codici specifici per dati mancanti e valori non previsti, i risultati ottenuti applicando procedure di abbinamento ai dati grezzi non conducono ad apprezzabili differenze e a sensibili miglioramenti.

3.3. Trattamento delle famiglie sostitutive

Oltre a quelle "classiche" descritte in precedenza, una possibile fonte di errore proviene dalla struttura stessa della RTFL che, non essendo un'indagine propriamente *panel*, non prevede il tentativo di seguire le famiglie o i singoli individui che per un qualunque motivo non sono reperibili. In tal caso, infatti, gli individui escono dal campione, mentre le famiglie vengono sostituite con una famiglia con caratteristiche simili estratta dall'elenco sostitutivo, allo scopo di mantenere inalterata la struttura sezionale del campione.

Prima del 1992, le famiglie subentrate mantenevano il codice delle precedenti. Giusti, Marliani e Torelli (1991) avevano osservato come in questo modo aumentasse notevolmente la probabilità di errati abbinamenti; anche al fine di eliminare tale problema, a partire dal 1992 il sistema di assegnazione dei codici familiari è stato

modificato. In particolare, le innovazioni introdotte consistevano nell'assegnare nuovi codici alle famiglie sostitutive, e nel prevedere esplicitamente nel questionario una nuova variabile "elenco di provenienza" (per tale variabile erano previsti due possibili codici: 1 = famiglia dell'elenco base; 2 = famiglia dell'elenco suppletivo). In linea teorica, le famiglie provenienti dall'elenco suppletivo avrebbero potuto essere direttamente eliminate dal campione di famiglie da porre a confronto, senza rischio di falsi negativi in quanto tali individui erano certamente assenti nelle precedenti rilevazioni.

Un'analisi della variabile suddetta, in una prima fase della costruzione della procedura di abbinamento, ne ha rivelato però l'estrema inaffidabilità. Le famiglie identificate come provenienti dall'elenco suppletivo sono circa il 3,5% del totale delle famiglie del campione in ogni rilevazione, e quasi la metà di esse appartengono alla sezione di rotazione entrante; questo indica, come prevedibile, che è meno frequente la sostituzione della famiglia una volta che questa viene intervistata per la prima volta. Sono inoltre presenti alcuni individui con un codice non previsto pari a 0 (0,5-1%) che, per tutti i casi sottoposti a verifica manuale, appare riferibile ad una mancata assegnazione del codice che identifica la famiglia come appartenente all'elenco base.

Secondo le convenzioni che ha fissato l'Istat, le famiglie estratte dall'elenco suppletivo entrano a far parte dell'elenco base a partire dalla rilevazione successiva, per cui le coppie che presentano la successione di codici 2-1 per la variabile elenco di provenienza nelle due occasioni sono da considerare di fatto equivalenti alle coppie 1-1. Quanto detto è confermato dal fatto che, effettuando un accurato controllo manuale, su 200 famiglie che a parità di altri codici presentano la coppia 2-1 circa il 95% risultano da abbinare. Le uniche coppie di codici che potrebbero possedere un potere discriminante sono pertanto 1-2 (famiglie sostituite nella seconda rilevazione) e 2-2 (famiglie sostituite in entrambe), ed in tali casi non si dovrebbe teoricamente abbinare nessuna famiglia. I risultati non sono però conformi alle attese, con 8 famiglie su 15 abbinate nel primo caso e 23 su 31 nel secondo.

È pertanto evidente che, a meno di un miglioramento nella qualità dei dati relativi all'elenco di appartenenza, l'eventuale eliminazione di tali confronti porterebbe ad un notevole numero di falsi negativi, a fronte di un dubbio incremento nella protezione dai falsi positivi. Nel seguito non si utilizza il codice relativo all'elenco di provenienza come variabile di confronto per la procedura di abbinamento.

3.4. Scelta delle variabili di blocco

Le procedure di abbinamento utilizzate in precedenza sulla RTFL (Giusti, Marliani e Torelli, 1991, Moriani, 1981, Favro-Paris *et al*., 1996) considerano come blocco "naturale" l'identificativo familiare, composto da regione, provincia, comune, sezione di rotazione e codice di famiglia. Tali variabili hanno certamente un elevato potere discriminante, poiché il numero di confronti da effettuare all'interno della famiglia è ridotto, ma un blocco troppo restrittivo potrebbe portare ad un gran numero di falsi negativi a fronte di un incerto miglioramento sui falsi positivi.

Al fine di verificare l'affidabilità di tale blocco, si effettua un'analisi attraverso un abbinamento preliminare con un blocco alternativo. In particolare, si utilizzano come variabili di blocco quelle relative alla data di nascita, effettuando tutti i confronti, anche al di fuori del blocco sul codice familiare, fra gli individui nati, a meno di errori, nello

stesso giorno. I risultati presentati nel seguito si riferiscono all'applicazione della procedura di abbinamento per i dati della RTFL per i primi due trimestri degli anni 1995 e 1996 per Lombardia e Campania, con archivi iniziali contenenti approssimativamente 40.000 individui (per una più dettagliata descrizione dei campioni utilizzati si veda Paggiaro, 1998). Nell'analisi, si considerano potenzialmente abbinabili le coppie che presentano una sola discordanza fra le variabili di confronto, con un semplice procedimento di tipo deterministico. Definite come "errori apparenti" le discordanze su una sola variabile, si effettua infine una verifica manuale sui record per individuare quali fra le coppie con un solo errore siano effettivamente da abbinare.

- Regione e Provincia: nessun errore apparente.

- Comune: 265 errori apparenti, con le coppie potenzialmente abbinabili tutte appartenenti a 5 comuni. La verifica svolta documenta però che tali apparenti falsi negativi sono in realtà dovuti ad una duplicazione del campione fra comuni differenti: nel I trimestre i comuni A e B condividono esattamente lo stesso campione, probabilmente perché le informazioni relative ad uno dei due sono venute a mancare; se, ad esempio, nel II trimestre il comune B esce dal campione, l'abbinamento senza blocco considera sia le coppie A-A che le coppie B-A, in realtà relative ai medesimi individui. Eliminate queste situazioni "patologiche", non risultano errori nella rilevazione del codice di comune.

- Sezione di Rotazione: le coppie di record con variabili di confronto concordanti tranne la sezione di rotazione sono 987, pari a più del 5% del totale di abbinamenti teoricamente possibili. La dimensione del fenomeno induce pertanto ad un'analisi approfondita della provenienza di tali discordanze: 118 sono dovute ad errori di rilevamento su singoli individui o piccole porzioni di campione, mentre assume particolare importanza il fatto che 869 coppie (quasi il 90%) risultano concentrate in soli 7 comuni.
Un'analisi più dettagliata delle informazioni ottenibili su tali comuni mostra come l'intero campione non sia soggetto a rotazione, con il risultato che il codice di sezione rimane immutato (3-3 e 4-4) invece di seguire la successione naturale (3-1 e 4-2). Le possibili spiegazioni di tale fenomeno sono due, con conseguenze differenti: (a) il campione del comune è stato replicato dal trimestre precedente per mancata rilevazione, per cui il comune deve in realtà essere eliminato dal campione longitudinale; (b) il comune effettivamente non effettua la rotazione, ma le interviste sul campione sono reali, per cui un abbinamento con blocco sulla sezione porterebbe ad eliminare tutti gli individui del comune, con considerevole perdita di numerosità campionaria ma anche possibili distorsioni da selezione del campione non probabilistica. Si segnala comunque che un simile problema di selezione potrebbe verificarsi anche inserendo nel campione tutti gli abbinamenti provenienti dal comune, in quanto in questo modo non si abbinano più solo 2 sezioni ma 4, ed il comune viene ad avere un peso doppio rispetto a quello previsto.
È infine interessante notare che la sezione di rotazione risulta l'unica variabile di confronto per la quale il peso degli errori varia notevolmente da trimestre a trimestre: ad esempio, fra il I ed il II trimestre del '95 il numero di coppie abbinate con errore è circa il 30% di quelli registrati nello stesso periodo del '96, con una notevole variazione del peso di tale variabile nella valutazione dell'efficienza delle procedure di abbinamento.

• Codice Familiare: come per la variabile precedente, anche il codice familiare presenta un numero di "errori apparenti" notevole; per i primi due trimestri del '96 vi sono 462 coppie di record relativi a 14 comuni, ed i livelli sono simili per il '95. Un'analisi più approfondita mostra che tali errori sono dovuti per la quasi totalità ad uno slittamento dei codici familiari dovuto alla sostituzione di alcune famiglie, pratica non prevista nelle procedure sul campo predisposte dall'Istat: se ad esempio la famiglia 11 esce dal campione, la famiglia 12 nella rilevazione successiva prende il suo posto e tutta la serie di codici successivi slitta di 1. Anche in questo caso, pertanto, un blocco rigido sulla variabile comporterebbe una notevole perdita nel campione, anche se con un rischio di distorsione ridotto rispetto al caso precedente.

I risultati empirici dell'abbinamento preliminare suggeriscono pertanto un blocco a livello comunale, con l'evidenza di probabilità praticamente nulle di falsi negativi dovuti al blocco stesso. Poiché il numero di confronti da effettuare rimarrebbe comunque notevole, soprattutto per i comuni di dimensione maggiore, all'interno del blocco principale appare utile mantenere la logica del "doppio blocco alternativo" fra codice familiare e data di nascita; in questo modo si mantiene anche una forte protezione rispetto ai falsi positivi, al prezzo di una limitata perdita in termini di falsi negativi. Seguendo Jabine e Scheuren (1986), la stima della copertura può avvenire mediante tecniche cattura-ricattura: in ipotesi di indipendenza fra blocco sulla famiglia e sulla data di nascita, si calcola il prodotto fra la frequenza di errori sulle variabili che caratterizzano i due blocchi; per il '96, ad esempio, gli errori sono circa l'8% sui codici ed il 3% sulla data di nascita, con una perdita di circa 0,25% dei possibili abbinati.

4. Abbinamento probabilistico con i dati della RTFL

4.1. Caratteristiche peculiari dell'algoritmo EM per la procedura di abbinamento

Individuate le variabili di blocco e di confronto, rimangono da definire le specifiche scelte relative al procedimento di stima dei pesi. Per la stima delle probabilità *m* si adotta l'ipotesi di indipendenza fra gli errori sulle singole variabili di confronto, che appare realistica se si assume tali errori siano in gran parte di rilevazione e trascrizione. Assunzioni diverse potrebbero eventualmente risultare opportune per tenere conto di errori che dipendono da particolari situazioni relative all'intervista (ad esempio quando il rispondente è un *proxy*) o da incongruenze introdotte dal piano di compatibilità. D'altra parte, se si rinuncia all'assunzione di indipendenza per la stima degli *m* il carico computazionale aumenta notevolmente mentre le stime che si ottengono non sembrano essere, almeno nel caso qui considerato, sensibilmente diverse da quelle che si ottengono data l'assunzione di dipendenza.

Per quanto riguarda la stima di *u*, essendo questa esterna all'algoritmo iterativo è possibile tener conto in dettaglio delle interazioni fra le variabili, senza particolari problemi computazionali. In questo modo si tiene conto del fatto che alcune combinazioni di valori nelle variabili di confronto si osservano con una frequenza molto diversa da quella che risulterebbe dall'assunzione di indipendenza. A titolo di esempio, si pensi all'elevata associazione che esiste fra caratteristiche invarianti come anno di nascita e sesso e altre variabili quali stato civile, relazione con l'intestatario, e istruzione, che per bambini e adolescenti possono assumere coerentemente solo un

sottoinsieme limitato di valori; quando per una coppia si osserva una concordanza congiunta di tali variabili questo deve avere un potere discriminatorio per l'abbinamento notevolmente inferiore a quello che si otterrebbe con l'ipotesi di indipendenza.

Un'analisi dettagliata, effettuata nel campione casuale utilizzato per la stima di *u*, mostra effettivamente come sia necessario tener conto delle relazioni fra le 5 variabili di confronto sopra citate. Come prevedibile, non risultano invece associate con le altre variabili giorno e mese di nascita, sezione di rotazione e codice familiare. Stime di *u* basate sulle due ipotesi alternative di indipendenza o correlazione fra le variabili, condizionatamente ad U, portano a risultati diversi che si ripercuotono sulle stime dei pesi nell'algoritmo EM e sulla stima della quota di errori commessi nell'abbinamento (Torelli e Paggiaro, 1999). Tenendo conto anche del relativo aumento nel carico computazionale, l'abbinamento definitivo viene pertanto effettuato inserendo esplicitamente le relazioni osservate fra le variabili elencate nel procedimento di stima dei pesi.

Rispetto alle scelte effettuate da Giusti, Marliani e Torelli (1991), per la stima di *u*, si ritiene che l'allargamento del blocco al livello comunale, consentendo confronti anche al di fuori del codice familiare, elimini la necessità di una stima che tenga conto del blocco. Se infatti in quel caso era necessario distinguere le probabilità per dimensione della famiglia, non pare esistano motivi per considerare stime di *u* differenziate per comune. Tale scelta potrebbe anzi essere controproducente, in quanto un campione casuale vincolato ad abbinamenti all'interno del comune aumenterebbe la probabilità di associare casualmente vettori effettivamente relativi allo stesso individuo, con una conseguente forte distorsione nelle stime di *u*.

L'algoritmo EM applicato con tali assunzioni sui dati della RTFL converge molto velocemente dopo pochi passi, ed anche con regole di arresto molto severe non si superano le 10 iterazioni; i risultati cui si perviene sono gli stessi qualunque siano i valori iniziali assegnati, e nemmeno la velocità dell'algoritmo ne è influenzata in modo apprezzabile.

I risultati presentati in seguito sono basati su una scelta per la soglia, espressa come probabilità che due record si riferiscano alla medesima unità, pari a 0,95. Tale scelta si è rivelata, in questa situazione applicativa, del tutto ragionevole essendo supportata, inoltre, dal fatto che: (a) tentativi con soglie inferiori conducono a numerosi abbinamenti multipli per lo stesso record, segnale di un aumento del rischio di falsi positivi a fronte di un minimo recupero in termini di falsi negativi; (b) si osserva una robustezza rispetto all'abbinamento su differenti campioni, con le modalità oltre la soglia prescelta che risultano essenzialmente le stesse per tutti i trimestri analizzati, se si escludono pochi valori al margine con numerosità molto ridotte.

4.2. Analisi dei risultati dell'abbinamento

Nella tabella 1 sono riassunti i risultati della procedura di abbinamento per i primi due trimestri del 1998 relativamente alle regioni Lombardia e Campania: oltre alle frequenze osservate dei vettori di confronto si presentano, esclusivamente per le coppie con un unico errore nell'abbinamento, i valori dei pesi (3) stimati attraverso l'algoritmo EM e si definiscono nel dettaglio le singole stime dei pesi *m* ed *u*.

Tabella 1. Risultati della procedura di abbinamento

Esito del confronto	**Abbinamento 98I-98II**						
	Freq. Abbin.	***Prob.***	***w***	***m***	***ux10^5***	***Errati Abbin.***	***Abbin. Mult.***
Tutte concordanze	**13293**	0,9999	11,90	0,788	0,5	**0**	68
1 errore	**2954**					**5**	83
blocco su data di nascita							
- *sezione di rotazione*	104	0,9853	5,12	0,006	4	3	4
- *codice familiare*	314	0,9882	5,34	0,020	10	2	17
blocco su codice							
- *giorno di nascita (1)*	212	0,9884	5,36	0,012	6	0	0
- *mese di nascita (1)*	142	0,9891	5,42	0,008	4		
- *anno di nascita (1)*	229	0,9928	5,83	0,015	4	0	7
- *anno di nascita (0)*	109	0,9758	4,61	0,008	8	0	6
doppio blocco							
- *relaz. Capofamiglia (1)*	220	0,9999	10,52	0,017	0,05	0	0
- *relaz. Capofamiglia (0)*	54	0,9994	8,25	0,003	0,09	0	1
- *sesso*	193	0.9994	8,30	0,015	0,4	0	35
- *stato civile*	76	0,9999	10,94	0,006	0,01	0	0
- *titolo di studio (1)*	1112	0.9998	9,72	0,068	0,4	0	11
- *titolo di studio (0)*	189	0,9996	8,82	0,012	0,2	0	2
2 errori	**130**					**0**	9
blocco su data di nascita	4					0	5
blocco su codice	25					0	1
doppio blocco	101					0	3
3 errori	**5**					**0**	0
Totale abbinamenti	**16382**					**5**	160
Abbinati potenziali	**17700**	**Totali**	**~16700**	**Osserv.**			
Percentuale abbinati	**92,5%**		**98%**				

Di particolare interesse appare confrontare i pesi dei vettori risultanti dai differenti tipi di blocco. Innanzitutto, i vettori risultanti dal doppio blocco hanno pesi vicini ad 1, con quasi 2000 individui supplementari abbinati, e si presentano pesi elevati anche per molti vettori con 2 errori su tali variabili.

Produce buoni risultati anche la rinuncia al blocco per sezione di rotazione e codice familiare: i vettori con un solo errore hanno pesi elevati, e superano la soglia anche alcune combinazioni con un ulteriore errore su variabili esterne al blocco. Inoltre, le differenti prove effettuate mostrano come l'algoritmo sia in grado di adeguarsi alle particolarità delle singole rilevazioni, con la presenza di comuni che non effettuano la rotazione del campione che si ripercuote in pesi superiori per tutti i vettori che presentano discordanza sulla sezione.

Per valutare i risultati della procedura si è anche fatto ricorso all'ispezione manuale dei record, e ciò ha consentito di identificare con ottima approssimazione quali coppie appartengono all'insieme da abbinare. I risultati di tale operazione hanno permesso di verificare come le coppie con pesi superiori alla soglia siano quasi tutte da abbinare, ed i pochi abbinamenti errati vengano poi eliminati nel controllo sugli abbinamenti multipli; notevole invece la presenza di potenziali falsi positivi nelle coppie con peso inferiore alla soglia, a conferma della bontà della scelta di una soglia elevata come protezione.

La procedura attribuisce, inoltre, pesi superiori alla soglia a tutti i vettori di confronto che documentano differenze solo per una cifra che compone la data di nascita, e solo per l'anno di nascita si ammettono errori su entrambe le cifre. Tale risultato sembra derivare soprattutto dal valore elevato della probabilità *u* di osservare tali vettori casualmente.

4.3. Valutazione dell'efficienza dell'algoritmo di abbinamento

Al fine di valutare l'efficienza delle procedure di abbinamento adottate, è necessario conoscere con buona approssimazione il numero di abbinati potenziali, individui presenti in entrambe le rilevazioni per i quali è possibile ottenere l'abbinamento dei record. Ad esempio, per quanto riguarda gli abbinamenti fra due trimestri consecutivi, vanno valutate solo le unità che appartengono alle due sezioni che partecipano ad entrambe le rilevazioni ed è inoltre possibile identificare, sulla base dell'elenco dei comuni che partecipano ad una coppia di rilevazioni, le unità residenti in comuni presenti in entrambi i trimestri. Una valutazione preliminare del numero di individui almeno potenzialmente abbinabili è quindi fornita dalla dimensione dell'archivio più piccolo fra i due da abbinare dopo avere condotto le operazioni di razionalizzazione descritte sopra. La quota di abbinati sul totale dei potenzialmente abbinabili è quindi un indicatore grezzo della bontà della procedura. Le informazioni fornite dal codice di avvenuta sostituzione potrebbero consentire di restringere ulteriormente la definizione alle famiglie presenti in entrambi i trimestri; è evidente, quindi, che il numero di abbinati potenziali viene in questo modo sovrastimato e, di conseguenza, la reale efficienza del metodo è in genere superiore a quella che si ottiene dall'indicatore grezzo.

Nell'abbinamento fra il I ed il II trimestre 1998 per le due suddette regioni, le numerosità campionarie totali sono rispettivamente 40094 e 40196. I comuni presenti in una sola occasione sono 41, e la numerosità totale dei comuni di interesse scende rispettivamente a 35389 e 35431. Dal campione vengono inoltre esclusi i 2 comuni per i quali si era verificata la mancata rotazione. Una prima approssimazione per eccesso del numero di abbinati potenziali si ottiene pertanto calcolando la metà della popolazione dei comuni presenti in entrambi i trimestri (circa 17700 unità). La percentuale di abbinamento risultante è pari a 92,5%. Una valutazione più precisa, anche se leggermente per difetto, si ottiene attraverso una verifica manuale dei falsi negativi, ottenuta abbassando la soglia fino a 0,1, livello oltre il quale si ritiene siano molto poche le coppie di record relativi al medesimo individuo. Il numero di potenziali abbinati complessivo scende così fino a circa 16700, valutazione realistica se si considerano i fisiologici fenomeni di *attrition*. Nel suo complesso, nonostante la scelta di una soglia piuttosto conservativa, il procedimento di abbinamento produce ottimi risultati sul fronte dei falsi negativi, con una quota di abbinati vicina al 98%.

Per una valutazione del rischio di falsi positivi, appare utile innanzitutto una breve analisi dei risultati dell'eliminazione delle coppie relative ad abbinamenti multipli di uno stesso record. Gran parte dei record della prima rilevazione relativi ad individui potenzialmente abbinabili risultano infatti abbinati ad almeno un record nella rilevazione successiva, e di conseguenza una buona parte delle coppie che superano erroneamente la soglia presentano record che siano già stati abbinati in modo corretto.

Il procedimento di depurazione conclusiva dagli abbinamenti multipli prevede di considerare abbinate le coppie con peso superiore, mentre in caso di pesi identici la

scelta avviene in modo casuale. Jaro (1989) propone un'interessante alternativa che consente di massimizzare il peso complessivo di abbinamento all'interno del blocco, ma il fatto di ammettere errori nei codici identificativi ne complicherebbe l'utilizzo.

Entrando nel dettaglio dei risultati, 68 record presentano abbinamenti multipli con concordanza perfetta (gemelli dello stesso sesso), confermando che anche un metodo restrittivo che non ammetta alcun errore presenta un certo rischio di falsi positivi. In aggiunta a queste, solo 92 coppie presentano abbinamenti multipli, delle quali 35 sono relative a gemelli di sesso diverso. Questo indica che il rischio di falsi positivi, anche se si mantenessero tutte le coppie che superano la soglia prescelta, non aumenta al punto da consigliare metodi più conservativi; tale indicazione vale a maggior ragione se si considera che le coppie appena descritte vengono poi in gran parte eliminate risolvendo gli abbinamenti multipli, con solamente 5 falsi positivi che entrano realmente nel campione finale. Questi casi possono essere relativi a quel numero limitato di coppie in cui entrambi i record presentano abbinamento singolo, o che casualmente presentano un peso maggiore di quelle corrette, e si ritiene che il loro peso sia limitato se confrontato con il notevole recupero in termini di falsi negativi.

Per concludere, si noti che l'efficienza del metodo proposto aumenta notevolmente nel caso di abbinamento effettuato sulle 4 occasioni consentite dalla struttura longitudinale della RTFL. In questo caso, infatti, è molto maggiore la probabilità di osservare un errore in almeno una delle 4 rilevazioni, ed il numero di abbinati con criteri restrittivi ne risente maggiormente. Alcune prove effettuate sul periodo 95I-96II mostrano infatti come, a fronte di un numero sempre limitato di falsi positivi, il numero di abbinati passi da 4400 a 6448, con un incremento del 47% rispetto ai risultati di un metodo restrittivo ed una percentuale di abbinati superiore al 90%.

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.

Casavola, P., Sestito, P. (1994). L' indagine ISTAT sulle forze di lavoro, *Lavoro e Realzioni Industriali*, 1, 179-195.

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.

Di Pietro, E. (1993). La nuova indagine ISTAT sulle forze di lavoro. *Economia & Lavoro*, 27, 1, 57-64.

Duncan, G.J., Kalton, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 1, 97-117.

Favro-Paris, M.M., Gennari. P., Oneto, G.P. (1996). La durata della disoccupazione in Italia: un'applicazione della struttura longitudinale dell'indagine sulle forze di lavoro. *Quaderni di Ricerca ISTAT*, 4, 1-79.

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

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.

Moriani, C. (1981). Forze di lavoro e flussi di popolazione. Supplemento al *Bollettino Mensile di Statistica*, Istat, 15, 5-15.

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. (1998). *Modelli di mistura per l'analisi della durata della disoccupazione: aspetti metodologici e un'applicazione ai dati dell'indagine sulle forze di lavoro*.

Tesi di Dottorato di Ricerca, Dipartimento di Scienze Statistiche, Università di Padova.

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*.

Trivellato, U. (1991) (a cura di), *Forze di Lavoro: Disegno dell'Indagine e Analisi Strutturali*. ISTAT, Annali di Statistica, 9, 11.

Trivellato, U. e Torelli, N. (1989). Analysis of labor force dynamics from rotating panel survey data. *Bulletin of the International statistical Institute*, 53, 2, 425-444.

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

Working Papers già pubblicati

1. E. Battistin, A. Gavosto e E. Rettore, *Why do subsidized firms survive longer? An evaluation of a program promoting youth entrepreneurship in Italy*, Agosto 1998.
2. N. Rosati, E. Rettore e G. Masarotto, *A lower bound on asymptotic variance of repeated cross-sections estimators in fixed-effects models*, Agosto 1998.
3. U. Trivellato, *Il monitoraggio della povertà e della sua dinamica: questioni di misura e evidenze empiriche*, Settembre 1998.
4. F. Bassi, *Un modello per la stima di flussi nel mercato del lavoro affetti da errori di classificazione in rilevazioni retrospettive*, Ottobre 1998.
5. Ginzburg, M. Scaltriti, G. Solinas e R. Zoboli, *Un nuovo autunno caldo nel Mezzogiorno? Note in margine al dibattito sui differenziali salariali territoriali*, Ottobre 1998.
6. M. Forni e S. Paba, *Industrial districts, social environment and local growth. Evidence from Italy*, Novembre 1998.
7. B. Contini, *Wage structures in Europe and in the USA: are they rigid, are they flexible?*, Gennaio 1999.
8. B. Contini, L. Pacelli e C. Villosio, *Short employment spell in Italy, Germany and Great Britain: testing the "Port-of-entry" hypothesis*, Gennaio 1999
9. B. Contini, M. Filippi, L. Pacelli e C. Villosio, *Working careers of skilled vs. unskilled workers*, Gennaio 1999
10. F. Bassi, M. Gambuzza e M. Rasera, *Il sistema informatizzato NETLABOR. Caratteristiche di una nuova fonte sul mercato del lavoro*, Maggio 1999.
11. M. Lalla e F. Pattarin, *Alcuni modelli per l'analisi delle durate complete e incomplete della disoccupazione: il caso Emilia Romagna*, Maggio 1999.
12. A. Paggiaro, *Un modello di mistura per l'analisi della disoccupazione di lunga durata*, Maggio 1999.
13. T. Di Fonzo e P. Gennari, *Le serie storiche delle forze di lavoro per il periodo 1984.1-92.3: prospettive e problemi di ricostruzione*, Giugno 1999.
14. S. Campostrini, A. Giraldo, N. Parise e U. Trivellato, *La misura della partecipazione al lavoro in Italia: presupposti e problemi metodologici di un approccio "time use"*, Ottobre 1999.
15. A. Paggiaro e N. Torelli, *Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro*, Ottobre 1999.

Le richieste di working papers vanno indirizzate alla Sig.ra Patrizia Piacentini, Dipartimento di Scienze Statistiche, via C. Battisti 241-243, 35121 Padova, oppure all'indirizzo e-mail patrizia.piacentini@unipd.it.