SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

Metodi statistici per il record linkage

A cura di: Mauro Scanu

Il volume vuole essere uno strumento utile sia a chi si avvicina per la prima volta al problema del record linkage sia a chi usa tali metodologie durante la propria attività lavorativa. Per i primi si è definito in modo formale il problema del record linkage e la logica delle soluzioni proposte (capitoli 2, 3, 4 e 5). Per i secondi si è tentato di dare un resoconto il più possibile completo e aggiornato dei metodi proposti e applicati (capitoli 6 e 7). Inoltre vengono sottolineati aspetti (quali lo studio degli errori di linkage e il loro effetto sulle analisi dei dati, capitoli 8 e 9) solitamente trascurati.
In generale, viene data particolare importanza alla definizione dei metodi in un contesto statistico/probabilistico formalizzato. Questo viene fatto per mettere in luce le ipotesi implicite nei metodi proposti, le semplificazioni adottate (a volte non giustificabili) e le loro conseguenze. Per facilitare la lettura, vengono trattati in appendice due argomenti più tecnici: il metodo EM per ottenere stime di massima verosimiglianza dei parametri in presenza di dati mancanti e l'interpretazione in termini di teoria dell'informazione del metodo per il record linkage proposto da Fellegi e Sunter.

Indice

Prefazione

Questo volume raccoglie i risultati delle attività di studio e ricerca sui metodi per l'integrazione di dati provenienti da più fonti svolte dall'Unità Operativa "Metodologie e sistemi di supporto all'integrazione" a partire dal 1999. Gli ambiti di ricerca in questo contesto sono svariati. Per quanto riguarda i metodi che hanno un fondamento statistico, gli argomenti che sono stati trattati (e vengono tuttora studiati) dall'unità sono l'abbinamento esatto (che chiameremo con il termine ormai noto a livello internazionale: *record linkage*) e l'abbinamento statistico (noto come *statistical matching*, o *synthetical matching* o ancora *data fusion*). In questo volume viene discusso esclusivamente il problema dell'abbinamento esatto, rimandando a altri volumi la discussione sull'abbinamento statistico. L'attività di studio e ricerca in questi ambiti è stata portata avanti da Mauro Scanu, sotto la guida di Giovanna D'Angiolini, capo dell'Unità Operativa. Questo volume è stato scritto interamente da Mauro Scanu. Chiunque avesse commenti può scrivere al seguente indirizzo e-mail: scanu@istat.it.

Per quanto l'autore si assuma la responsabilità di quanto è esposto in questo volume, non si possono dimenticare coloro con i quali si è collaborato, studiando insieme le metodologie proposte in letteratura e discutendole con il fine di individuare delle soluzioni plausibili a problemi metodologici. Fra tutti, si ringrazia in modo particolare Marco Fortini, che ha avuto e continua ad avere un ruolo importante, dovuto alla profonda conoscenza della problematica dell'abbinamento esatto e delle metodologie statistiche connesse alla sua soluzione. In questi ultimi tre anni si è anche collaborato con successo con Brunero Liseo dell'Università di Roma "La Sapienza", e con Alessandra Nuccitelli. Grazie a loro, insieme a Marco Fortini, si è individuata una strategia bayesiana per l'abbinamento esatto. Inoltre si ringraziano Giovanna D'Angiolini, Giulio Barcaroli, Lucia Coppola e Giovanni Seri, per i loro commenti (estremamente importanti) sulle precedenti versioni di questo volume. Un ringraziamento particolare va infine al professore con il quale l'autore ha iniziato la sua attività di ricerca all'Università, Giovanni Battista Tranquilli. Non verranno dimenticati i suoi insegnamenti e il suo rigore scientifico.

Il presente lavoro ha usufruito dei finanziamenti del progetto MIUR-COFIN2002 dal titolo "Metodi statistici per l'integrazione di dati da fonti diverse".

Capitolo 1

L'abbinamento di dati provenienti da fonti diverse

Negli ultimi decenni l'attività di integrazione di fonti diverse ha avuto un grande sviluppo in moltissimi campi. Uno dei più importanti è quello relativo all'integrazione di fonti gestite dalla Pubblica Amministrazione. Come sottolineato da Fellegi (1997), i motivi di tale sviluppo sono essenzialmente tre.

1. **l'occasione** - Dopo la seconda guerra mondiale diversi stati hanno rivolto un particolare interesse alla costituzione di sistemi di "welfare state" e di tassazione sempre più complessi, in modo da venire incontro alle esigenze della popolazione. Questa attività ha contribuito a creare grandi basi di dati contenenti informazioni dettagliate sia per i singoli individui sia per le imprese.

2. **lo strumento** - La tecnologia informatica ha, nel contempo, facilitato la manutenzione di queste basi di dati, così come l'integrazione delle informazioni contenute nelle diverse fonti e il calcolo di elaborazioni di dati integrati sempre più complessi.

3. **la necessità** - I governi hanno assunto sempre più ruoli e funzioni che hanno accresciuto, a loro volta, i bisogni informativi. Spesso tali bisogni possono essere soddisfatti dall'uso congiunto di diverse fonti a disposizione, senza richiedere a individui o imprese informazioni che hanno già fornito altrove.

In questo lavoro l'attenzione è rivolta a una classe di strumenti per l'integrazione di dati di fonte diversa che va sotto il nome di *record linkage*. Secondo quanto scritto in un lavoro di Belin e Rubin (1995), una procedura di integrazione di dati provenienti da fonti diverse va sotto il nome di record linkage se soddisfa la seguente definizione.

Definizione 1.1 *Una procedura di record linkage è una tecnica algoritmica il cui scopo è identificare quali coppie di record di due basi di dati*[1] *corrispondono ad una stessa unità.*

Sinonimi di record linkage sono: *exact matching* e *computerized matching*. Dal primo sinonimo è sorta la traduzione italiana di record linkage, ovvero *abbinamento esatto*. Non bisogna farsi ingannare dalla traduzione: il record linkage è una tecnica di abbinamento esatto in quanto *l'obiettivo* è quello di collegare *esattamente* le unità che appartengono a basi di dati diverse, ma come si vedrà in seguito non sarà necessariamente esatto nel risultato. Anzi, spesso si è costretti ad accettare un certo margine di errore. La seconda definizione fa invece riferimento esplicito al mezzo con il quale le diverse basi di dati vengono integrate: il computer.

[1] D'ora in avanti ogni fonte da integrare viene chiamata "base di dati", intendendo per base di dati una matrice "unità-variabili". Un record in una base di dati è il vettore delle modalità delle variabili associato ad ogni unità.

I metodi per il record linkage hanno ormai una storia quarantennale alle spalle, considerando il lavoro di Newcombe *et al*. (1959) come il primo lavoro sull'argomento. In Italia, l'interesse per questi metodi sembra più recente, come testimoniato dalle applicazioni in ambito demografico (Pinelli, 1984, è un primo esempio; una rassegna è disponibile in Coccia, Gabrielli e Sorvillo, 1993). In particolare in ISTAT questi metodi sono stati usati per i dati delle forze lavoro (ad esempio Giusti, Marliani, Torelli, 1991) oltre che per la costruzione di ASIA (Archivio Statistico delle Imprese Attive; si veda ad esempio Garofalo, 1998). Dalla letteratura disponibile, i metodi per il record linkage si possono dividere in due grandi gruppi: metodi *ad hoc* o euristici e metodi basati su modello, o "statistici". I motivi che portano alla definizione di metodi euristici sono diversi, e possono essere sintetizzati nei seguenti punti.

1. Sono molto rare le persone che si occupano prevalentemente dei metodi per l'integrazione di fonti diverse; personale inesperto è spesso portato a definire metodi ad hoc.

2. Il problema dell'integrazione di dati di fonti diverse viene visto come un problema prettamente, se non esclusivamente, informatico. Gli statistici vengono coinvolti solo nella fase finale di analisi dei dati.

3. Oggi sono presenti software commerciali per l'integrazione di dati da fonti diverse che comprendono metodi statistici ormai consolidati in letteratura. Questi software sono prodotti essenzialmente in paesi anglosassoni (Canada, Stati Uniti e Regno Unito in primo luogo) e non sempre sono adattabili alle caratteristiche specifiche dei dati italiani.

Qui facciamo riferimento a metodi statistici per il record linkage (anche se si possono trovare accenni a metodi euristici). I motivi che portano a privilegiare metodi statistici sono:

- i metodi euristici non sono automaticamente in grado di controllare la qualità dei risultati;

- sia per valutare la qualità dei risultati, sia per la fase di abbinamento vero e proprio, i metodi euristici necessitano dell'intervento di personale specializzato per il controllo manuale dei record.

Il primo punto sottolinea il fatto che l'abbinamento viene fatto "alla cieca". Queste procedure possono quindi essere prese in considerazione solo quando si ripone molta fiducia nell'accuratezza dei dati a disposizione. Si avverte che il mancato controllo della qualità del record linkage, e quindi della possibilità di errati abbinamenti, può incidere notevolmente sulla qualità delle analisi statistiche successive (come evidenziato nel capitolo 9). Il secondo punto incide sui costi. Infatti, oltre al costo degli addetti al record linkage per il periodo di tempo necessario, bisogna includere anche il costo dell'eventuale nuovo contatto con le unità (individui o imprese) per le quali l'abbinamento risulta più problematico. Inoltre il personale specializzato, oltre ad essere costoso, spesso è difficile da reperire.

Al contrario, i metodi statistici basano le proprie scelte su procedure dove la possibilità di commettere errori di abbinamento viene *controllata*. Inoltre i costi derivanti dall'uso di personale specializzato, che spesso sono inevitabili anche in queste situazioni, possono essere analizzati in un contesto decisionale e possono essere minimizzati (in proposito si veda il capitolo 4). La letteratura sul record linkage è ormai vasta e sembra utile avere a disposizione uno strumento che descriva tutti i metodi proposti in un quadro il più possibile coerente, cercando di evidenziarne i pregi e, eventualmente, i difetti. Questo è l'obiettivo del presente lavoro.

1.1 A cosa serve il record linkage

La fase di integrazione di dati di fonti diverse può soddisfare diversi scopi. Jabine e Scheuren (1986), sottolineano l'esistenza di due gruppi di obiettivi: obiettivi *non statistici* e obiettivi *statistici*.

Gli obiettivi non statistici che possono essere risolti attraverso l'applicazione di procedure di record linkage sono orientati allo sviluppo e alla manutenzione di basi di dati per una serie di attività, utili alle amministrazioni fiscali, alle banche o alle compagnie di assicurazione, alle amministrazioni giudiziarie, solo per citarne alcune. Queste basi di dati contengono informazioni su insiemi ben definiti di individui o oggetti utili alle rispettive amministrazioni (ad esempio coloro che pagano le tasse per le amministrazioni fiscali, i singoli conti correnti per le banche, gli individui che commettono un reato per le amministrazioni giudiziarie). La fase di integrazione può essere finalizzata, ad esempio, all'aggiornamento dei dati in loro possesso. Un obiettivo non statistico molto particolare è stato affrontato dalla chiesa Mormone, con sede a Salt Lake City: ricostruire l'albero genealogico delle famiglie (NeSmith, 1997, White, 1997).

I principali obiettivi statistici possono essere raggruppati come segue.

- *Lo sviluppo di una lista di unità da usare come lista di campionamento o come lista di riferimento per un censimento*. A volte, le liste di unità da usare a fini statistici (ad esempio per estrarre campioni) sono costruite attraverso l'unione degli insiemi di unità appartenenti a diverse basi di dati. Un esempio è fornito dall'archivio ASIA (Archivio Statistico delle Imprese Attive, in proposito si veda ad esempio Garofalo e Viviano, 2000), che è costruito riunendo in un'unica lista le imprese presenti in diverse basi di dati[2]. In questo contesto, è necessario individuare le imprese che sono presenti contemporaneamente in più basi di dati, in modo da considerarle una volta sola.

- *La ricongiunzione di due o più fonti per poter disporre in un'unica base di dati di più informazioni a livello di unità*. Alcune analisi richiedono informazioni che non sono disponibili su un'unica fonte. Ad esempio, ciò avviene quando si vogliono studiare gli effetti di alcuni fattori di rischio (nell'alimentazione o nel lavoro) sulle cause di morte: è quindi inevitabile collegare i risultati delle indagini sull'alimentazione o sulle forze lavoro con i registri delle cause di morte. Altre volte molte informazioni sono già disponibili su archivi amministrativi e può convenire ristrutturare il questionario di un'indagine omettendo le domande relative a queste informazioni, che vengono poi recuperate attraverso il record linkage.

- *L'integrazione di fonti diverse come strumento per migliorare la copertura e proteggersi dagli errori di risposta nei censimenti e nelle indagini*. Questo obiettivo non è stato studiato approfonditamente (un primo tentativo, per popolazioni animali, è in Robinson-Cox, 1998), ma è indubbio che le informazioni presenti, ad esempio, negli archivi amministrativi sono una risorsa preziosa da affiancare alle usuali tecniche di imputazione ed editing.

- *Il conteggio degli individui di una popolazione attraverso metodi di "cattura-ricattura"*. L'esempio più importante è dato dalla stima della sottocopertura del censimento, effettuata abbinando i record del censimento con quelli dell'indagine post-censuaria. Uno degli

[2] Le principali sono: il registro delle imposte del Ministero delle Finanze; il registro delle imprese e delle unità locali delle Camere di Commercio, Industria Artigianato e Agricoltura; il registro dell'INPS

elementi fondamentali per questa operazione è il numero di unità rilevate sia nel censimento che nell'indagine post-censuaria (in proposito si veda, ad esempio, Wolter, 1986).

- *L'integrazione di fonti diverse come strumento per valutare la bontà di un metodo di protezione dal rischio di identificazione dei dati rilasciati*. I dati rilasciati dagli uffici di statistica sono vincolati all'obbligo di riservatezza. Diversi ricercatori stanno valutando la possibilità di usare il record linkage come strumento per descrivere in che misura le unità della base di dati rilasciata sono identificabili (si veda, ad esempio, Duncan e Lambert, 1989, e Winkler, 1998).

Due importanti riferimenti bibliografici, dove sono disponibili un gran numero di applicazioni sia statistiche sia non statistiche, sono i volumi degli atti di due conferenze tenute nel 1985 e nel 1997 negli Stati Uniti: Kills e Alvey (1985) e Alvey e Jamerson (1997). Un ulteriore riferimento bibliografico dove sono reperibili alcuni lavori che testimoniano l'interesse in Italia per questi metodi è Filippucci (2000).

1.2 I codici identificativi e gli errori

Tutti gli obiettivi precedentemente illustrati richiedono una fase di ricerca delle unità (siano esse individui, famiglie, imprese o altro) che sono presenti contemporaneamente in due o più basi di dati. Se alle unità presenti nelle due basi di dati è associato un unico codice identificativo (ad esempio il codice fiscale), e questo codice è riportato correttamente in ambedue le occasioni, il ricongiungimento delle due basi di dati è un'operazione estremamente semplice. Ma a volte questi identificativi non sono presenti, oppure non sono trascritti accuratamente.

Esempio 1.1 *Un esempio di base dati che non possiede un unico codice identificativo degli individui è dato dai campioni delle forze lavoro. La tecnica di campionamento adottata è quella del campionamento ruotato*[3] *per cui i campioni intervistati in due rilevazioni successive possiedono circa il 50% delle unità in comune. La ricongiunzione dei record delle unità rilevate in due occasioni successive è un passo necessario se, ad esempio, si vogliono condurre analisi sui flussi di individui che entrano e escono dal mercato del lavoro. Se l'unità di interesse è la famiglia, un codice affidabile è disponibile. Per riconoscere i singoli individui, invece, è necessario far riferimento a un insieme di variabili, fra le quali l'età in anni compiuti, il sesso, la relazione con il capofamiglia e il grado di istruzione (per maggiori dettagli si vedano Giusti, Marliani e Torelli, 1991, e Torelli, 1998).* □

Nei casi in cui un unico codice identificativo non esiste, oppure non è affidabile, è inevitabile fare riferimento a un insieme di variabili presenti in ambedue le basi di dati e che *congiuntamente* sono in grado di identificare un'unità: chiameremo queste variabili *variabili chiave*. Esistono un certo numero di errori che influenzano la qualità delle variabili chiave e che pregiudicano la possibilità di aggancio dei record di due basi di dati attraverso *merge*.

1. **Errori di trascrizione**. Ad esempio un individuo può immettere una data di nascita sbagliata. A questo tipo di errori si possono far risalire anche le **variazioni di codice**, come ad esempio l'uso di diverse versioni del nome (che si verifica quando il proprio

[3] Per maggiori informazioni sulla tecnica di campionamento si veda ISTAT, 2001

nome di battesimo è composto da diverse componenti, oppure quando corrisponde a una trasformazione di un nome più diffuso). La presenza di un intervistatore può ovviare a alcuni di questi errori. Un ultimo tipo di errore si verifica quando qualche unità non risponde a una o più variabili chiave (mancata risposta parziale).

2. **Errori di registrazione**. Questi errori avvengono durante la fase di registrazione dei dati. Errori in questa fase sono in qualche modo controllabili da parte dell'ente che produce i dati (come suggerito in ISTAT, 1989, pagina 115, e in Fortini, 2001, alla voce "Registrazione su supporto informatico") ma difficilmente si riesce ad eliminarli.

Chernoff (1980) dimostra in modo formale che gli errori nelle variabili chiave riducono l'efficacia dell'informazione congiunta delle variabili chiave per il ricongiungimento delle unità di due basi di dati[4]. Questi errori pregiudicano l'utilizzo del *merge* come metodo di aggancio fra unità nelle diverse basi di dati. Tipicamente, il record linkage fatto attraverso merge quando sono presenti errori nelle variabili chiave produce:

- i falsi non abbinamenti: alcuni record delle due basi di dati fanno riferimento alla stessa unità ma il *merge* non è in grado di individuarli in quanto almeno una variabile chiave è affetta da qualche errore;
- i falsi abbinamenti: alcuni record possono essere abbinati anche se in realtà fanno riferimento a unità diverse. Le variabili chiave coincidono per via degli errori prima descritti.

La caratteristica delle procedure statistiche di record linkage è quella di risolvere il problema dell'abbinamento dei record di due diverse basi di dati *tenendo sotto controllo il livello degli errori che possono essere generati*. Questo aspetto porta ad una formalizzazione del problema del record linkage e a una sua trattazione in termini decisionali e statistici.

1.3 Le fasi del record linkage

I passi necessari per abbinare i dati di due basi di dati attraverso record linkage sono essenzialmente tre, come sottolineato da Jabine e Scheuren (1986):

1. pre-elaborazione
2. applicazione dell'algoritmo di record linkage
3. analisi.

Il primo punto è cruciale, ma spesso prescinde dall'obiettivo dell'integrazione di fonti diverse. In particolare riguarda la necessità di rendere *compatibili* e *omogenee* le informazioni contenute nelle due basi di dati. Questa fase è stata studiata da diversi autori, fra gli altri Jabine e Scheuren (1986). Di seguito si elencano i passi preliminari necessari per poter applicare un metodo di record linkage.

[4]Uno degli strumenti più usati nel record linkage è il rapporto delle verosimiglianze fra due distribuzioni, che in media definisce l'informazione di Kullback-Leibler (si veda l'appendice C). Chernoff ha dimostrato che l'informazione di Kullback-Leibler associata, ad esempio, a una variabile binaria che viene riportata con errore solo nel 3% dei casi è la metà di quella che si sarebbe osservata se la variabile binaria fosse stata registrata senza errori

- **Scegliere le variabili chiave** - Nel paragrafo 1.2 abbiamo visto che a volte le due basi di dati che devono essere integrare non assegnano lo stesso codice identificativo delle unità. In questo caso è necessario identificare le unità attraverso una combinazione di identificativi parziali (le *variabili chiave*). La scelta delle variabili chiave è estremamente delicata. In linea di principio, tutte le variabili in comune fra le due basi di dati possono essere usate congiuntamente per identificare le unità, ma molte di queste non sono necessarie per l'integrazione fra le basi di dati. In genere si sceglie il numero minimo di variabili chiave che congiuntamente identificano le unità, fra le variabili in comune nelle due basi di dati che sono *universali* (ovvero tutte le unità devono rispondere a queste variabili) e *permanenti* (ovvero immodificabili nel tempo). Inoltre è opportuno selezionare le variabili chiave fra le variabili più *accurate* (anche se problemi di qualità sono spesso inevitabili) e *non sensibili* (ovvero che non violino il diritto alla riservatezza delle unità). Non sempre è possibile soddisfare tutti questi requisiti: non è raro trovare applicazioni di record linkage che, ad esempio,usano come variabili chiave il "titolo di studio" o lo "stato civile" per abbinare individui, ovvero variabili che possono modificarsi nel tempo.

- **Migliorare la qualità dei dati nelle basi di dati da integrare** - Per quanto possibile, è utile fare in modo che le basi di dati a disposizione siano estremamente accurate per evitare errori negli abbinamenti. In particolare è necessario promuovere l'accuratezza e la completezza delle variabili chiave. Si supponga di dover gestire un'indagine campionaria sulle famiglie, i cui risultati vengono poi arricchiti da una serie di informazioni desumibili da un archivio amministrativo. Difficilmente si può fare qualcosa sull'archivio. Al contrario l'indagine può essere pianificata in modo da assicurare la completezza e l'accuratezza dei dati sulle variabili chiave e soprattutto che queste vengano rilevate in un formato *compatibile* (in termini di definizione e categorizzazioni) con le corrispondenti variabili disponibili sull'archivio amministrativo.

- **Standardizzazione delle variabili** - Può risultare utile trasformare in modo opportuno le modalità delle variabili chiave in modo da rendere più semplice per i computer il riconoscimento delle differenze. Questo avviene in particolare per variabili come "nome", "cognome" e "indirizzo". Per queste variabili spesso si preferisce eliminare i titoli (come *sig.*, *dr.*, per gli individui, *srl*, *spa* per le imprese, *via*, *piazza* per gli indirizzi). In alcuni casi le modalità di queste variabili vengono trasformate in modo da limitare gli effetti derivanti da errori di digitazione o possibili differenze nella pronuncia di nomi stranieri. Negli Stati Uniti e in Canada questi metodi di *parsing* sono ampiamente applicati. Uno dei più usati è il codice SOUNDEX, che trasforma ad esempio il cognome in un codice alfanumerico a quattro caratteri: la prima lettera del cognome seguito da tre caratteri in funzione delle consonanti successive alla prima lettera. In ISTAT un metodo di standardizzazione degli indirizzi è fornito, ad esempio, da SISTER.

- **Blocking and Sorting** - Per facilitare il controllo dei record da parte dei programmi software per il record linkage, spesso è necessario ordinare (*sorting*) opportunamente i record nelle due basi di dati e dividerli in gruppi (*blocking*). Quest'ultima operazione influenza in modo notevole i risultati del record linkage, e ad essa è dedicato un paragrafo del capitolo 5.

1.4 Organizzazione dei capitoli successivi

Come detto nel paragrafo precedente, l'attività legata all'integrazione di basi di dati attraverso record linkage può dividersi in tre fasi: oltre a quella che racchiude le operazioni preliminari, le ulteriori due fasi riguardano la scelta e l'applicazione di un metodo di record linkage e l'analisi dei risultati. Queste ultime due fasi verranno descritte approfonditamente nei capitoli seguenti.

Nei capitoli 2 e 3 si definisce il problema del record linkage in modo formale. Si caratterizza il problema del record linkage come un problema decisionale, e si costruisce un modello statistico per l'insieme delle osservazioni a disposizione. Quindi si divide la procedura di record linkage in due fasi.

La prima è la fase di soluzione del *problema decisionale*, ovvero riguarda la definizione di uno strumento che consenta di affermare se due record fanno riferimento alla stessa unità oppure no. In primo luogo viene descritta la regola "ottimale" definita da Fellegi e Sunter (1969), di cui daremo conto nel capitolo 4. Successivamente, nel capitolo 7, vengono introdotti alcuni altri approcci al record linkage diversi rispetto alla procedura delineata da Fellegi e Sunter.

La seconda fase è invece una fase *statistica*, e riguarda la stima degli elementi necessari per la costruzione delle regole decisionali. Questo è uno degli argomenti più ampiamente dibattuti nella letteratura sul record linkage e verrà affrontato nel capitolo 6.

Nel capitolo 5 si evidenziano due modifiche da apportare alla procedura di Fellegi e Sunter (e a qualsiasi altra procedura di record linkage) per far sì che l'uso di tali procedure non sia pregiudicato da problemi di ordine logico o computazionale. Il problema della qualità della procedura viene studiato nel capitolo 8 e l'influenza degli errori di abbinamento sulle analisi statistiche viene discusso nel capitolo 9.

In appendice si delinea il metodo di stima di massima verosimiglianza in presenza di dati mancanti che va sotto il nome di metodo EM, ampiamente usato nel capitolo 6, e si presentano i punti di contatto esistenti fra alcuni strumenti definiti nelle procedure di record linkage e quelli definiti nella teoria dell'informazione.

Capitolo 2

Definizioni preliminari

2.1 I dati osservati

Per utilizzare i metodi di record linkage sono necessari i seguenti elementi.

1. Almeno due rilevazioni (statistiche o amministrative) che hanno in comune un insieme non vuoto di unità. Per semplicità faremo sempre riferimento a due rilevazioni A e B e denoteremo gli insiemi di unità osservati nelle due rilevazioni con $\mathcal{A}$ e $\mathcal{B}$, rispettivamente di numerosità ν_A e ν_B.

2. Un gruppo di variabili *chiave* rilevate sia in A che in B. Supponiamo che queste variabili siano k, $k \geq 1$, e rappresentiamo la variabile che osserva congiuntamente le k variabili chiave con: $(X_1, ..., X_k)$.

3. La variabile $(X_1, ..., X_k)$ è in grado di identificare le unità, ovvero ad ogni k-upla osservata di modalità delle k variabili chiave corrisponde una ed una sola unità della popolazione investigata.

Come è stato già visto nel paragrafo 1.2, date queste condizioni, il *merge* fra le liste di unità delle due rilevazioni fatto rispetto alle modalità di $(X_1, ..., X_k)$ sarebbe sufficiente per individuare quelle unità che sono state osservate in entrambe le rilevazioni A e B. In altri termini, anticipando una notazione che sarà estremamente utile in seguito, il *merge* determina le coppie (a, b), $a \in \mathcal{A},\ b \in \mathcal{B}$, dove a e b sono le etichette di una stessa unità in $\mathcal{A}$ e $\mathcal{B}$. Va sottolineato che le condizioni necessarie affinché l'operazione di *merge* porti al risultato desiderato sono:

1. che la variabile $(X_1, ..., X_k)$ venga rilevata senza errore su tutte le unità in $\mathcal{A}$ e $\mathcal{B}$,

2. che non ci siano mancate risposte parziali nelle due rilevazioni,

3. che la variabile $(X_1, ..., X_k)$ non subisca modificazioni nell'intervallo di tempo che intercorre fra le due rilevazioni A e B (ad esempio, se una variabile fra le k chiavi descrive lo "stato civile" di un individuo, può accadere che questa variabile si modifichi nel tempo compromettendo il risultato del *merge*).

Queste condizioni non si verificano affatto in contesti reali. Ne segue che la procedura di identificazione delle unità presenti sia in $\mathcal{A}$ che in $\mathcal{B}$ deve essere formalizzata come una procedura di decisione che si basa su un opportuno confronto fra le modalità di $(X_1, ..., X_k)$ assunte dalle unità nelle due liste.

2.2 Formalizzazione del problema

Il problema che il record linkage deve risolvere può essere formalizzato nel seguente modo. Si considerino tutte le coppie formate da unità rispettivamente della lista A e B:

$$\mathcal{A} \times \mathcal{B} = \Big\{(a, b) : a \in \mathcal{A}, b \in \mathcal{B}\Big\},$$

dove con il simbolo $\times$ intendiamo il prodotto cartesiano fra due insiemi. L'obiettivo che si vuole raggiungere consiste nel determinare una particolare bipartizione dell'insieme $\mathcal{A} \times \mathcal{B}$ in due sottoinsiemi disgiunti ed esaustivi $\mathcal{M}$ e $\mathcal{U}$:

$$\mathcal{M} \bigcap \mathcal{U} = \emptyset, \qquad \mathcal{M} \bigcup \mathcal{U} = \mathcal{A} \times \mathcal{B}, \tag{2.1}$$

con $\mathcal{M}$ formato dalle unità rilevate sia nell'occasione A che in B, cioè:

$$\mathcal{M} = \Big\{(a, b) \in \mathcal{A} \times \mathcal{B} : a = b\Big\},$$

e $\mathcal{U}$ formato dalle restanti coppie in $\mathcal{A} \times \mathcal{B}$:

$$\mathcal{U} = \Big\{(a, b) \in \mathcal{A} \times \mathcal{B} : a \neq b\Big\}.$$

Si vuole sottolineare fin da subito che l'interesse non è rivolto al *numero di unità rilevate nelle due occasioni* (cioè alla cardinalità di $\mathcal{M}$), ma a specificare esattamente *quali unità sono state osservate nelle due occasioni A e B* (cioè quali sono le coppie componenti $\mathcal{M}$). Anche se, da quanto appena detto, l'interesse è ristretto al solo insieme $\mathcal{M}$, si verificherà in seguito che la definizione dell'insieme $\mathcal{U}$ è comunque importante, in quanto consente l'uso di tecniche statistiche standard. Le coppie $(a, b) \in \mathcal{M}$ sono state chiamate in diversi modi da chi si è occupato di record linkage. Qui chiamiamo "match" ogni singola coppia $(a, b) \in \mathcal{M}$, e "non-match" le restanti coppie.

Uno strumento utile a definire la bipartizione (2.1) è fornito da un indicatore di appartenenza delle coppie a $\mathcal{M}$:

$$c_{a,b} = \begin{cases} 1 & \text{se } (a, b) \in \mathcal{M} \\ 0 & \text{se } (a, b) \in \mathcal{U} \end{cases} \tag{2.2}$$

$a = 1, ..., \nu_A$, $b = 1, ..., \nu_B$. Il valore assunto dalla funzione (2.2) per tutte le coppie $(a, b) \in \mathcal{A} \times \mathcal{B}$ è quindi il *parametro* di interesse nel record linkage. Questo parametro può essere organizzato in una matrice $\mathbf{c} = \{c_{a,b}\}$, con ν_A righe e ν_B colonne. Se in ogni lista i record si riferiscono ad una ed una sola unità, cioè se è sempre vero che

$$a \neq a', \qquad \forall\, a, a' \in \mathcal{A} \tag{2.3}$$

$$b \neq b', \qquad \forall\, b, b' \in \mathcal{B} \tag{2.4}$$

allora la matrice $\mathbf{c}$ deve soddisfare i vincoli:

$$c_{a,b} \in \{0, 1\}, \qquad \forall\, a \in \mathcal{A},\ \forall\, b \in \mathcal{B}, \tag{2.5}$$

$$\sum_{b=1}^{\nu_B} c_{a,b} \leq 1, \qquad \forall\, a \in \mathcal{A}, \tag{2.6}$$

$$\sum_{a=1}^{\nu_A} c_{a,b} \leq 1, \qquad \forall\, b \in \mathcal{B}. \tag{2.7}$$

Il primo vincolo è dovuto alla definizione della funzione indicatrice (2.2). Il secondo e il terzo vincolo specificano invece che ogni unità di una lista può abbinarsi con al più un'unità dell'altra lista in quanto non si ammette che una stessa unità possa essere presente più di una volta in ciascuna delle liste di unità osservate nelle rilevazioni A e B. Se ogni unità di $\mathcal{A}$ viene rilevata anche nell'occasione B allora il vincolo (2.6) si trasforma in

$$\sum_{b=1}^{\nu_B} c_{a,b} = 1, \qquad \forall\, a \in \mathcal{A}$$

(il vincolo (2.7) si modifica allo stesso modo quando ogni unità in $\mathcal{B}$ deve associarsi con un'unità di $\mathcal{A}$).

In molte occasioni le condizioni (2.3) e (2.4) non sono vere. Ad esempio questo accade quando le unità rappresentate nelle liste $\mathcal{A}$ e $\mathcal{B}$ sono le unità locali di imprese, e si vogliono riunire tutti i record che fanno riferimento alla stessa impresa. I vincoli (2.6) e (2.7) quindi decadono (infatti le somme per riga e per colonna di $\mathbf{c}$ possono fornire un valore superiore a 1), e vengono sostituiti dalla seguente regola:

$$c_{a,b} = 1;\ c_{a,b'} = 1;\ c_{a',b} = 1 \Longrightarrow c_{a',b'} = 1; \qquad \forall\, a, a' \in \mathcal{A},\ b, b' \in \mathcal{B}, \tag{2.8}$$

ovvero se a e a' si riferiscono alla stessa unità, ambedue si devono abbinare agli stessi record presenti nella lista $\mathcal{B}$. I vincoli posti definiscono un insieme di matrici $\mathcal{C}$. Questo è lo *spazio dei possibili parametri*, cioè il più piccolo insieme, in base ai vincoli conosciuti, a cui appartiene il vero, ma incognito, parametro $\mathbf{c}$, cioè la vera ma incognita bipartizione (2.1).

Esempio 2.1 *Sia $\mathcal{A} = \{a_1, a_2\}$ e $\mathcal{B} = \{b_1, b_2\}$. Supponiamo siano validi i vincoli (2.5), (2.6) e (2.7). Allora* $\mathbf{c}$ *è una matrice* 2×2*:*

$$\mathbf{c} = \begin{pmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{pmatrix}$$

e l'insieme dei possibili parametri $\mathcal{C}$ è formato dalle seguenti matrici:

$$\mathcal{C} = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}$$

Se le condizioni (2.6) e (2.7) sono sostituite dalla (2.8) allora l'insieme $\mathcal{C}$ diviene:

$$\mathcal{C} = \left\{ \begin{array}{cccc} \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} & \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \end{array} \right\}$$

□

L'obiettivo del record linkage consiste nella determinazione della matrice $\mathbf{c}$ (non osservata) attraverso l'osservazione delle variabili chiave $(X_1, ..., X_k)$ sulle ν_A unità rilevate in A e le ν_B unità rilevate in B. Sia $\mathbf{x}_A = \{x^A_{a,j}\}$ la matrice con ν_A righe e k colonne il cui generico

termine $x^A_{a,j}$ rappresenta la modalità della variabile X_j osservata sull'unità a in $\mathcal{A}$, $j = 1, ..., k$, $a = 1, ..., \nu_A$. Allo stesso modo venga definita la matrice $\mathbf{x}_B = \{x^B_{b,j}\}$, con ν_B righe e k colonne.

Lo strumento fondamentale del record linkage è il *confronto* fra i valori assunti dalle variabili chiave $(X_1, ..., X_k)$ per la coppia (a, b), con $a \in \mathcal{A}$ e $b \in \mathcal{B}$. In generale, rappresentiamo questo confronto con il simbolo:

$$\mathbf{y}_{ab} = f(x^A_{a,1}, ..., x^A_{a,k}; x^B_{b,1}, ..., x^B_{b,k}). \tag{2.9}$$

Questa funzione ha il compito di registrare quanto sono diversi i valori delle variabili chiave nelle due unità poste a confronto. In linea di principio, ci si aspetta livelli bassi di diversità per le coppie (a, b)in $\mathcal{M}$, e livelli di diversità più elevati per le restanti coppie. Questo comportamento è alla base della regola di decisione per il record linkage. Se i due gruppi di coppie $\mathcal{M}$ e $\mathcal{U}$ fossero caratterizzati da valori di $\mathbf{y}_{ab}$ che si sovrappongono, la funzione $f(.)$ sarebbe un cattivo strumento per la decisione del record linkage. La capacità di *discriminazione* della funzione $f(.)$ è quindi un elemento assai importante, che definisce in un certo senso un aspetto della qualità del meccanismo di record linkage. Fra le diverse funzioni $f(.)$, si considerino i seguenti esempi.

Esempio 2.2 *Il modo più semplice di definire la funzione (2.9) consiste nel verificare esclusivamente se le due unità presentano la stessa modalità di una variabile chiave oppure no. La funzione (2.9) si trasforma in un vettore a k componenti:*

$$\mathbf{y}_{ab} = \left(y^1_{ab}, \cdots, y^k_{ab}\right),$$

con y^h_{ab} funzione indicatrice delle "uguaglianze":

$$y^h_{ab} = \begin{cases} 1 & \textit{se } x^A_{a,h} = x^B_{b,h} \\ 0 & \textit{altrimenti.} \end{cases} \tag{2.10}$$

Con la parola "altrimenti" si intende sia il caso in cui $x^A_{a,h} \neq x^B_{b,h}$, sia il caso in cui almeno uno dei due elementi sia mancante (mancata risposta parziale). L'insieme di valori che può essere assunto dal vettore $\mathbf{y}_{ab}$, ovvero lo spazio complessivo dei confronti che rappresentiamo con il simbolo $\mathcal{D}$, è costituito da tutti i vettori di k elementi composti da 0 e 1, la cui cardinalità è $|\mathcal{D}| = 2^k$. □

Esempio 2.3 *Supponiamo ora che le variabili chiave X^h siano numeriche*[1]*. Il generico elemento y^h_{ab} di $\mathbf{y}_{ab}$ può ora essere definito nel seguente modo:*

$$y^h_{ab} = |x^A_{a,h} - x^B_{b,h}|. \tag{2.11}$$

Le informazioni contenute nella funzione (2.11) sono superiori a quelle contenute nella funzione (2.10). Infatti la (2.11) identifica tutte le coppie che presentano lo stesso valore della variabile X^h con $y^h_{ab} = 0$, e scinde l'insieme delle coppie che non coincidono nella variabile X^h a seconda della "lontananza" fra le modalità. Questa ricchezza potrebbe risultare utile nella fase

[1]Si avverte fin da ora che ciò avviene assai raramente, soprattutto per le variabili continue. In genere le variabili chiave sono variabili qualitative

di decisione del record linkage. Se indichiamo il valore minimo e massimo assunto da X^h nelle due basi dati A e B con i simboli:

$$\min(X^h) = \xi^h \qquad \max(X^h) = \phi^h$$

lo spazio dei confronti per la singola variabile X^h, $\mathcal{D}^h$, è contenuto nell'intervallo $[0, \phi^h - \xi^h]$. Ne consegue che lo spazio dei confronti complessivo, $\mathcal{D}$, è definito dal prodotto degli intervalli $\mathcal{D}^h$, $h = 1, ..., k$. □

Esempio 2.4 *In questo caso, concentriamo l'attenzione non sulle coppie che presentano modalità diverse della variabile X^h, come nell'esempio precedente, ma sulle coppie che presentano lo stesso valore. Queste ultime nei due esempi precedenti vengono visualizzate dal confronto $y^h_{ab} = 1$ quando il confronto è del tipo (2.10) e $y^h_{ab} = 0$ quando il confronto è del tipo (2.11). Al contrario si consideri il confronto:*

$$y^h_{ab} = \begin{cases} x_h & \text{se } x^A_{a,h} = x^B_{b,h} = x_h \\ \emptyset & \text{altrimenti.} \end{cases} \tag{2.12}$$

L'insieme delle coppie che coincidono nella variabile X^h viene diviso in tanti gruppi quante sono le modalità di X^h. Questo tipo di funzione di confronto viene considerato da Fellegi e Sunter (1969) e Winkler (1989a) e sarà oggetto di maggiore approfondimento nel paragrafo 6.5. Indicando con $\mathcal{X}$ l'insieme delle modalità assunte dalla variabile X^h nelle due basi dati A e B, e con $\mathcal{X}_0$ l'insieme $\mathcal{X}$ a cui è stato aggiunto il valore $\emptyset$:

$$\mathcal{X}_0 = \mathcal{X} \bigcup \{\emptyset\},$$

lo spazio dei confronti $\mathcal{D}^h$ è ora contenuto in $\mathcal{X}_0$. □

Esempio 2.5 *La funzione di confronto che mantiene tutte le informazioni è quella che registra le modalità della variabile X^h assunte dalle due unità a e b poste a confronto:*

$$y^h_{ab} = \left(x^A_{a,h}, x^B_{b,h}\right). \tag{2.13}$$

Lo spazio dei confronti $\mathcal{D}^h$ per la variabile X^h è ora contenuto nel prodotto cartesiano:

$$\mathcal{D}^h = \mathcal{X}_A \times \mathcal{X}_B$$

dove $\mathcal{X}_A$ e $\mathcal{X}_B$ sono gli insiemi delle modalità di X^h assunte rispettivamente nelle basi dati A e B. □

Gli esempi appena riportati descrivono quattro funzioni di confronto con caratteristiche molto diverse. La funzione (2.10) è la meno discriminante, le funzioni (2.11) e (2.12) hanno una capacità discriminante intermedia, mentre la funzione (2.13) mantiene tutte le informazioni possedute dai dati. Come esempio, si consideri una sola variabile chiave X con confronto Y e che le $\nu_A \times \nu_B$ coppie siano rappresentabili come i punti di un rettangolo. Le figure 2.1 e 2.2 descrivono schematicamente cosa accade quando ci si riferisce a uno dei 4 confronti introdotti. In particolare le quattro definizioni della variabile Y (2.10)-(2.13) implicano una diversa suddivisione del rettangolo. Se la variabile Y (confronto della variabile X) è definita attraverso la regola (2.10), le

Figura 2.1 - La figura a sinistra rappresenta gli insiemi di coppie determinati dall'uso dei confronti (2.10) quando si ha una sola variabile chiave. La figura a destra usa confronti del tipo (2.11).

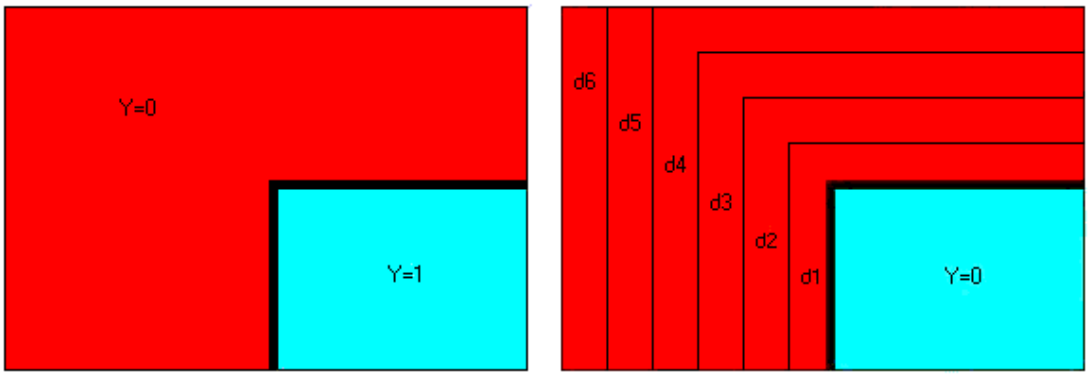
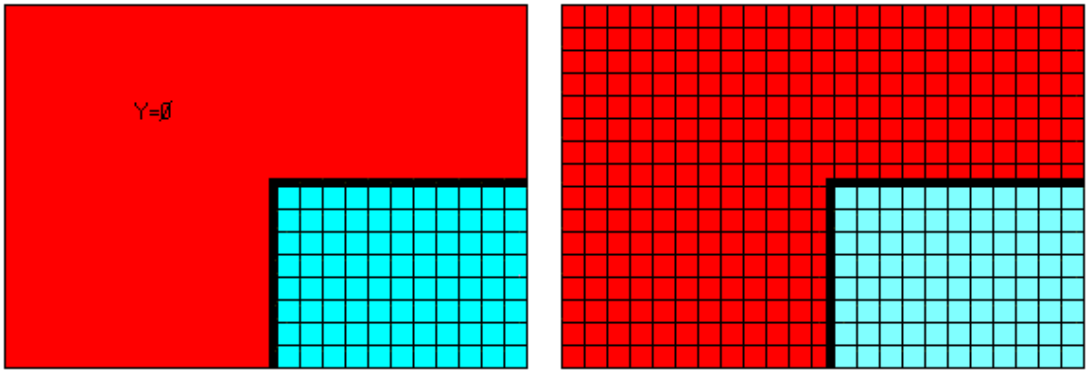


Figura 2.2 - La figura a sinistra rappresenta gli insiemi di coppie determinati dall'uso dei confronti (2.12) quando si ha una sola variabile chiave. La figura a destra usa confronti del tipo (2.13).



$\nu_A \times \nu_B$ coppie vengono divise in due gruppi: quello in cui le coppie presentano lo stesso valore di X ($Y = 1$), e quello in cui X è diverso. Se Y è definito come in (2.11), le coppie che presentano lo stesso valore di X rimangono nello stesso gruppo visto precedentemente (e identificato ora dal valore $Y = 0$), mentre le coppie che differiscono in X vengono divise in tanti gruppi quanti sono i livelli di distanza fra coppie di modalità di X (in questo caso 6). Se Y è definito come in (2.12), le coppie che presentano valori diversi di X rimangono indistinte, come quando Y viene definito da (2.10), mentre le coppie che coincidono in X vengono divise in tanti gruppi quante sono le modalità di X. Se Y viene definito dalla (2.13), le coppie che coincidono in X vengono divise come quando Y è definito dalla (2.12), mentre le coppie che differiscono in X vengono divise in $v_X(v_X - 1)$ gruppi, se v_X indica il numero di modalità di X.

L'importanza dell'informazione posseduta dalle modalità delle variabili chiave, e quindi della

maggiore finezza della partizione dello spazio dei confronti, è sottolineata dal seguente esempio.

Esempio 2.6 *Si supponga che i record da confrontare facciano riferimento a individui e che una variabile chiave X^h sia il cognome. Si considerino quindi due coppie (a_1, b_1) e (a_2, b_2) con confronto, secondo la (2.13),*

$$y^h_{a_1 b_1} = \left(x^A_{a_1,h}, x^B_{b_1,h}\right) = (\textit{Rossi, Rossi}) \tag{2.14}$$

$$y^h_{a_2 b_2} = \left(x^A_{a_2,h}, x^B_{b_2,h}\right) = (\textit{Zabrinski, Zabrinski}). \tag{2.15}$$

Il confronto del tipo (2.12) fornisce praticamente lo stesso risultato:

$$y^h_{a_1 b_1} = \textit{Rossi}, \quad y^h_{a_2 b_2} = \textit{Zabrinski.}$$

Per le stesse coppie, i confronti (2.10) e (2.11) sono equivalenti:

$$y^h_{a_1 b_1} = y^h_{a_2 b_2} = 1$$

per (2.10) e:

$$y^h_{a_1 b_1} = y^h_{a_2 b_2} = \emptyset$$

per i confronti del tipo (2.11). Ma è plausibile ritenere che un'uguaglianza nel cognome "Zabrinski" avvalori l'ipotesi $a_1 = b_1$ molto più che un'uguaglianza nel cognome "Rossi", data la diffusione dei due cognomi. La capacità discriminante delle modalità di una variabile chiave non è invece gestita dai confronti (2.10) e (2.11). □

Tenendo conto esclusivamente della capacità discriminante di $\mathbf{y}$, la funzione di confronto più opportuna è la (2.13). Nonostante ciò, la funzione di confronto (2.13) è stata presa in considerazione solo da Copas e Hilton (1990), mentre considerazioni sulla capacità discriminante sono disponibili, ad esempio, in Winkler (1989, 1995). Nel prossimo capitolo si vedrà che la funzione di confronto deve essere scelta tenendo conto anche di altri fattori, e che la (2.13) non è necessariamente la scelta più appropriata.

Commento 2.1 *Jaro (1989), Winkler (1990) e Porter e Winkler (1997) hanno definito delle funzioni particolarmente elaborate (Winkler le definisce "metriche") per il confronto fra le variabili chiave rilevate su due unità . In particolare queste metriche tengono conto del numero di caratteri presenti nei record posti a confronto, del numero di caratteri in comune nei due record e del numero di caratteri presenti in ambedue i record ma in posti diversi (traslazioni).*

Winkler e Porter (1997) considerano anche l'uso dei "bigrammi" (bigrams nella versione inglese). I bigrammi sono tutte le coppie di caratteri consecutivi presenti in una stringa. Ad esempio, la parola "bigramma" è formata dai bigrammi: bi, ig, gr, ra, am, mm, ma. Il confronto fra due stringhe basato sui bigrammi è una funzione con valori nell'intervallo $(0, 1)$ definita dal rapporto fra il numero di bigrammi in comune con il numero medio di bigrammi presenti nelle due stringhe. □

Capitolo 3

Il modello statistico per i dati nel record linkage

Nel paragrafo 2.2 è stato visto che l'obiettivo delle procedure di record linkage riguarda la individuazione della matrice $\mathbf{c}$. In altri termini, le procedure di record linkage si configurano come procedure di decisione, o meglio di classificazione delle diverse coppie (a, b) nelle coppie che sono match e non-match. Come tali, le procedure di record linkage non sono procedure statistiche in senso stretto, ovvero non hanno come obiettivo quello di descrivere delle sintesi di fenomeni osservati su una popolazione, o di inferenza su parametri relativi a uno o più meccanismi aleatori, dei quali si è osservato solo un campione. Al più, queste procedure hanno affinità con procedure statistiche consolidate, come i test (paragrafo 3.2). Ma la statistica riveste un ruolo fondamentale sia per quanto riguarda la stima degli elementi necessari alle procedure per il record linkage sia per quanto riguarda la fase finale di analisi dei dati ottenuti attraverso l'applicazione di procedure per il record linkage. È estremamente importante che le diverse fasi del record linkage (quella di stima degli elementi per le procedure di decisione, la fase di decisione vera e propria e la fase finale di analisi) siano viste in un quadro coerente, e non vengano fra loro separate. È per questo che si ritiene fondamentale anticipare in questo capitolo il modello statistico/probabilistico per i dati usati nel record linkage.

3.1 La distribuzione di probabilità per le osservazioni

Nel capitolo precedente sono stati descritti rispettivamente il contesto nel quale è definito il problema del record linkage e lo strumento principale (la funzione dei confronti) per poterlo risolvere. In particolare si è visto che l'unico elemento utile per poter decidere se una coppia (a, b) è un match è dato dal vettore dei confronti $\mathbf{y}_{ab}$, $a \in \mathcal{A}$, $b \in \mathcal{B}$. Le procedure di record linkage statisticamente fondate sfruttano l'informazione fornita dal vettore $\mathbf{y}_{ab}$ attraverso la distribuzione:

$$P(\mathbf{Y} = \mathbf{y}|c), \qquad \mathbf{y} \in \mathcal{D}, \tag{3.1}$$

dove $\mathbf{Y}$ è la variabile "vettore dei confronti", $\mathcal{D}$ è lo spazio dei possibili valori assumibili dalla variabile $\mathbf{Y}$ e c è un indicatore che vale 1 nel caso delle coppie che sono match, e zero per i non-match. Il presente capitolo è totalmente dedicato alla distribuzione (3.1), che nel caso dei match verrà chiamata $m(\mathbf{y})$ (paragrafo 3.1.1) e nel caso dei non-match $u(\mathbf{y})$ (paragrafo 3.1.2). Prima di entrare nel dettaglio, è bene chiarire il concetto (finora generico) di "distribuzione" utilizzato per la (3.1). A seconda dei vari contesti il significato sarà quello di distribuzione di probabilità[1] o di

[1] Senza perdere in generalità, adottiamo la notazione in (3.1) sia per variabili discrete che continue. Per queste ultime bisogna intendere la (3.1) come una densità di probabilità. Comunque queste ultime si incontrano raramente nel record linkage, dato che difficilmente tali variabili vengono usate come variabili chiave. Una eccezione di rilievo è data dal

distribuzione di frequenze. A seconda dell'interpretazione che si adotta, estrema attenzione deve essere rivolta alla interpretazione dei risultati (come evidenziato nel capitolo 4). La diversa natura della distribuzione (3.1) viene definita e giustificata nel commento 3.1.

Commento 3.1 *Nei capitoli successivi si farà riferimento a due specifiche interpretazioni della (3.1): in termini di probabilità (approccio di tipo superpopolazione[2]) e in termini di distribuzione di frequenza (approccio tipo popolazione finita).*

L'interpretazione della (3.1) in termini di probabilità è giustificata dal seguente modello. Si consideri una popolazione finita di unità (individui, imprese, ...). Si supponga che il valore delle variabili chiave sia generato da una variabile aleatoria $(T^1, ..., T^k)$ *in modo indipendente e identicamente distribuito per ogni unità della popolazione. Il valore generato dalla generica variabile* T^h *su un'unità corrisponde al vero valore della variabile chiave sull'unità stessa. Si considerino quindi due basi dati* A *e* B *che si riferiscono alla popolazione e in cui sono rilevate le variabili chiave. Se le basi dati registrano il valore delle variabili chiave senza errore, ovvero il valore generato da* $(T^1, ..., T^k)$ *per ogni unità della popolazione, la matrice* $\mathbf{c}$ *è univocamente e correttamente determinata. Al contrario, si considerino le variabili aleatorie* ϵ_h^A *e* ϵ_h^B, $h = 1, ..., k$, *che registrano l'errore rispetto al valore vero e che fanno osservare rispettivamente* X_h^A *e* X_h^B *al posto di* T^h. *Si suppone che le variabili generatrici di errori si applicano in modo indipendente e identicamente distribuito secondo un'opportuna distribuzione di probabilità su ogni unità della popolazione, ma tale distribuzione può essere diversa a seconda della rilevazione* A *o* B *(per via di una diversa preparazione dei rilevatori, di una diversa tecnica di rilevazione o altro). La distribuzione di probabilità (3.1) è quindi definita come trasformazione delle variabili* $(T^1, ..., T^k)$ *e degli errori. Si sottolinea che un modello così complesso è utile ai soli fini della stima della (3.1), in modo tale che le* $\nu_A \times \nu_B$ *osservazioni* $\mathbf{y}_{ab}$ *possono essere considerate come un campione casuale*[3] *sul quale poter applicare gli usuali metodi di stima (in particolare il metodo di massima verosimiglianza).*

Contrariamente a quanto appena affermato per l'approccio tipo superpopolazione, l'approccio tipo popolazione finita viene usato quando si hanno informazioni sufficienti per poter definire le due distribuzioni di frequenze, ad esempio per via di esperienze di integrazione di basi dati passate o simili (si veda il paragrafo 6.1). L'interpretazione della (3.1) in termini di distribuzione di frequenze si ottiene considerando i valori generati dalle variabili T^h, ϵ_h^A *e* ϵ_h^B, $h = 1, ..., k$, *non più come aleatori ma come dati. Si hanno quindi* $\nu_A \times \nu_B$ *vettori* $\mathbf{y}_{ab}$ *osservati sulla popolazione finita delle* $\nu_A \times \nu_B$ *coppie. Questi vettori formano una distribuzione di frequenze, rispettivamente per le coppie che sono match e per i non-match (questo condizionamento non è osservato, ma è comunque presente nell'insieme delle* $\nu_A \times \nu_B$ *coppie).*

In seguito verrà detto esplicitamente se si fa riferimento a un approccio tipo "superpopolazione" o all'approccio tipo "popolazione finita". □

---

lavoro di Belin e Rubin (1995, discusso anche nel paragrafo 8.2.1).

[2] In Särndal *et al*. (1993) è disponibile un'ampia discussione su questi modelli. Si vuole sottolineare che il concetto di "superpopolazione" ci è utile solo al fine di giustificare l'aleatorietà di $\mathbf{Y}$. Inoltre, contrariamente a quanto accade nel campionamento da popolazioni finite, il "campione" generato dal modello di superpopolazione viene osservato in tutte le sue unità (coppie): ciò di cui non si dispone è il valore assunto da $c_{a,b}$.

[3] per "campione casuale" si intende un campione di osservazioni generate in modo i.i.d. da una variabile aleatoria. Questa ipotesi permette di definire con facilità la funzione di verosimiglianza relativa a un campione e ha avuto un notevole successo nella letteratura sul record linkage. Purtroppo non è corretta. Per un commento sulla sua validità si rimanda al paragrafo 6.2.1

Commento 3.2 *I due approcci, tipo superpopolazione e tipo popolazione finita, hanno una diretta influenza sulla natura del condizionamento nella (3.1). Nel caso dell'approccio tipo superpopolazione, il condizionamento c fa riferimento a ogni singola coppia (a, b). Se la coppia (a, b) è un match ($c_{a,b} = 1$) allora il valore del vettore dei confronti sarà generato da una distribuzione diversa rispetto al caso in cui la coppia sia un non-match ($c_{a,b} = 0$). Nel caso dell'approccio tipo popolazione finita, la distribuzione di frequenza (3.1) sarà definita sulle due sottopopolazioni di coppie dei match ($c = 1$) e dei non-match ($c = 0$). Dato che nei capitoli successivi l'approccio tipo superpopolazione avrà una rilevanza maggiore, senza perdere in generalità il condizionamento verrà definito su $c_{a,b}$, indipendentemente dall'approccio che si sta discutendo.* □

Commento 3.3 *L'approccio tipo "popolazione finita" può essere facilmente adattato al caso in cui le due basi dati A e B appartengono a due popolazioni diverse, rispettivamente di N_A e N_B unità, contenenti un sottoinsieme non vuoto di unità in comune. In questo caso, si suppone che le ν_A unità in $\mathcal{A}$ sono un campione estratto dalla popolazione di N_A unità. Un significato analogo viene assegnato alle ν_B unità di B. La (3.1) esprime sia la distribuzione di frequenze di $\mathbf{Y}$ sulle $N_A \times N_B$ coppie formate dalle due popolazioni, sia la probabilità che la variabile $\mathbf{Y}$ assuma il valore $\mathbf{y}$ su una coppia (a, b) dove sia a che b sono estratti casualmente dalle corrispondenti popolazioni. Per poter utilizzare questa logica anche quando almeno una delle due liste è un archivio amministrativo, Fellegi e Sunter (1969) affermano che è sufficiente assumere che il meccanismo di selezione delle unità negli archivi (che non è casuale) sia "indipendente" dalle variabili chiave. Ad esempio, se un archivio è l'archivio dei dati fiscali delle persone fisiche (in tutto ν_P unità che possono essere considerate come un campione non casuale della popolazione italiana di N_P unità) si deve richiedere che le variabili chiave usate per l'aggancio dei record con le unità di una lista B (ad esempio variabili socio-demografiche) siano indipendenti dal fatto che un individuo presenti o no una dichiarazione fiscale. D'ora in avanti non faremo riferimento a quest'ultima situazione, anche se i commenti che verranno fatti per l'approccio tipo "popolazione finita" possono essere facilmente estesi anche a questo contesto.* □

La distribuzione di $\mathbf{Y}|c$ sarà diversa a seconda di come è definita la funzione $f(.)$ in (2.9). Consideriamo alcuni esempi relativi ai confronti (2.10), (2.11) e (2.13) descritti nel capitolo precedente.

Esempio 3.1 *Se il vettore dei confronti*

$$\mathbf{y}_{ab} = (y_{ab}^1, ..., y_{ab}^k) \tag{3.2}$$

è definito secondo la (2.10):

$$y_{ab}^h = \begin{cases} 1 & \textit{se } x_{a,h}^A = x_{b,h}^B \\ 0 & \textit{altrimenti} \end{cases} \qquad h = 1, ..., k,$$

la variabile aleatoria $\mathbf{Y}$ è multinomiale con spazio dei confronti $\mathcal{D}$, dato dall'insieme di tutti i vettori di dimensione k composti da 1 e 0, di cardinalità 2^k. I parametri della distribuzione multinomiale rappresentano le probabilità che ogni singolo vettore di $\mathcal{D}$ venga generato. Questi parametri sono dipendenti dal valore assunto da $c_{a,b}$. Si sottolinea che questa distribuzione è appropriata sia nel caso in cui si adotti un approccio tipo "superpopolazione", sia quando l'approccio è tipo "popolazione finita" (riferimenti al commento 3.1). □

Esempio 3.2 *Adottando un approccio tipo "modello di superpopolazione", se ogni componente del vettore dei confronti (3.2) è definito secondo la funzione:*

$$y_{ab}^h = |x_{a,h}^A - x_{b,h}^B|$$

con spazio dei confronti $\mathcal{D}^h = I\!R^+ \bigcup \{0\}$*, la corrispondente variabile* Y^h *può essere definita come una qualsiasi distribuzione continua: ad esempio una distribuzione esponenziale. Altrimenti si può considerare una distribuzione opportuna per la differenza non in valore assoluto:*

$$X_{a,h}^A - X_{b,h}^B,$$

ad esempio una gaussiana. Il corrispondente vettore dei confronti a k *dimensioni è allora distribuito come una variabile aleatoria multinormale con vettore medio:*

$$\mu(c_{a,b}) = \begin{bmatrix} \mu_1(c_{a,b}) \\ \mu_2(c_{a,b}) \\ . \\ . \\ \mu_k(c_{a,b}) \end{bmatrix}$$

e matrice di varianze e covarianze $\Sigma(c_{a,b})$*.* □

Esempio 3.3 *Supponiamo ora che il vettore (3.2) sia definito nel seguente modo:*

$$y_{ab}^h = \left(x_{a,h}^A, x_{b,h}^B\right).$$

La distribuzione di probabilità della variabile confronto Y^h *per la variabile chiave* X^h*:*

$$Y^h = (X_A^h, X_B^h)$$

è la distribuzione doppia delle risposte delle unità rilevate nelle due occasioni A *e* B *alla variabile* X^h*. La forma della distribuzione dipende in modo sostanziale dal condizionamento* $c_{a,b}$ *(in proposito si vedano gli esempi 3.5 e 3.7).* □

Il significato dei meccanismi probabilistici appena definiti è estremamente importante, ed è strettamente legato al parametro $c_{a,b}$ che definisce la (3.1). In particolare, $c_{a,b}$ può assumere solo i due valori 1 (la coppia (a, b) è un match) o 0 (la coppia (a, b) è un non-match). Consideriamo distintamente questi due casi. Come esempi tratteremo esclusivamente il caso dei confronti (2.10) e (2.13), che sono di gran lunga i più interessanti e i più usati.

3.1.1 La distribuzione dei confronti per i match

Per come sono stati definiti i match, consideriamo le coppie (a, b) con $c_{a,b} = 1$. Senza perdere in generalità, definiamo l'unità che si sta analizzando con il simbolo a. La funzione di confronto $\mathbf{y}_{aa}$ registra le eventuali differenze nelle risposte alle variabili chiave della stessa unità nelle due occasioni A e B. Dato che le k variabili chiave dovrebbero assumere, a meno di errori, la stessa modalità per ogni unità nelle due occasioni A e B, la densità (3.1)

$$m(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y} | c_{a,a} = 1), \qquad \mathbf{y} \in \mathcal{D}, \tag{3.3}$$

sarà concentrata nei valori $\mathbf{y} \in \mathcal{D}$ che rappresentano al più una lieve differenza fra le variabili chiave. I due esempi successivi sono validi sia quando vige un approccio tipo superpopolazione, sia quando l'approccio è tipo popolazioni finite.

Esempio 3.4 *Se y^h_{aa} è definito secondo la (2.10), la distribuzione multinomiale $m(\mathbf{y})$ per $\mathbf{Y}|c_{a,a} = 1$ è concentrata intorno ai vettori con molti elementi uguali a 1 (l'unità risponde allo stesso modo alla maggior parte delle variabili chiave) e assume vettori con molti zeri con probabilità molto vicina a zero.* □

Esempio 3.5 *Sia ora y^h_{aa} definita come in (2.13). La distribuzione di probabilità di Y^h rappresenta la probabilità congiunta delle risposte alla variabile chiave X^h per le unità che sono state osservate nelle due occasioni A e B (cioè le coppie in $\mathcal{M}$). In questo contesto, la distribuzione (3.1) ristretta alla variabile Y^h è data da:*

$$P(Y^h = y|c_{a,a} = 1) = P\Big((X^h_A, X^h_B) = (x_A, x_B)\Big|c_{a,a} = 1\Big) =$$
$$= \int_t P\left((X^h_A, X^h_B) = (x_A, x_B)\Big|c_{a,a} = 1, T^h = t\right)\beta^h(t)dt, \qquad (x_A, x_B) \in \mathcal{D}.$$

La distribuzione

$$P\Big((X^h_A, X^h_B) = (x_A, x_B)\Big|c_{a,b} = 1, T^h = t\Big)$$

corrisponde alla massima informazione sulla "qualità" delle due rilevazioni A e B analizzate congiuntamente. Nel caso dell'approccio tipo popolazioni finite, questa distribuzione è da intendersi come la distribuzione di frequenze delle risposte degli individui che sono osservati nelle due occasioni A e B e che presentano lo stesso valore (incognito) della variabile di interesse: $T^h = t$. La distribuzione $\beta^h(z)$ assume un significato analogo. Avere a disposizione tale informazione renderebbe le procedure di record linkage estremamente efficienti e precise. □

3.1.2 La distribuzione dei confronti per i non-match

La distribuzione di $\mathbf{Y}|c_{a,b} = 0$:

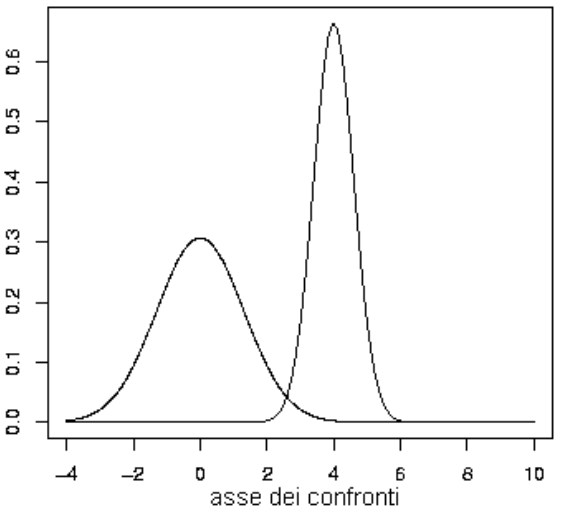
$$u(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|c_{a,b} = 0), \qquad \mathbf{y} \in \mathcal{D}, \tag{3.4}$$

ha interpretazione analoga a quanto visto nel paragrafo precedente. La differenza sostanziale è che questa distribuzione assume valori che registrano bassi livelli di diversità fra i valori assunti dalle variabili chiave nelle due occasioni con probabilità bassa. Usiamo gli stessi esempi illustrati nel paragrafo 3.1.1.

Esempio 3.6 *Se y^h_{ab} è definito secondo la (2.10), la distribuzione multinomiale $u(\mathbf{y})$ assume con bassa probabilità vettori composti da molti elementi uguali a 1. Ciò è vero sia in un approccio tipo superpopolazione che popolazioni finite.* □

Esempio 3.7 *Sia ora y^h_{ab} definita come (2.13). Gli argomenti successivi sono validi esclusivamente per l'approccio tipo superpopolazione. Abbiamo supposto che a unità diverse*

Figura 3.1 - Esempio di densità della variabile confronto per i match (a destra) e i non-match (a sinistra).



il vero valore T^h si associ in modo indipendente (si veda in proposito il commento 3.1). Inoltre gli errori si applicano sui veri valori in modo indipendente e identicamente distribuito per due diverse unità. Di conseguenza si ipotizza che il modello opportuno per la variabile $Y^h_{ab}|c_{a,b} = 0$ è:

$$P(Y^h = y|c_{a,b} = 0) = P\Big((X^h_A, X^h_B) = (x_A, x_B)\Big|c_{a,b} = 0\Big) = P\left(X^h_A = x_A\right) P\left(X^h_B = x_B\right)$$

ovvero due individui diversi rispondono in modo indipendente l'uno dall'altro. □

3.1.3 Alcune considerazioni sulle distribuzioni dei confronti

Un elemento chiave dei metodi per il record linkage è rappresentato dalla "distanza" che intercorre fra le distribuzioni $m(\mathbf{y})$ e $u(\mathbf{y})$. In particolare, quanto più queste distribuzioni sono "distanti" tanto più è semplice il compito di discriminare le diverse coppie nei due insiemi $\mathcal{M}$ e $\mathcal{U}$.

La figura 3.1.2 è un esempio di due distribuzioni per una variabile Y^h (una per i match e l'altra per i non-match) che per semplicità ipotizziamo essere distribuite secondo una normale. In questa figura i match assumono valori dei confronti sostanzialmente diversi dai non-match. Tranne per una piccola zona fra 2 e 4, queste distribuzioni forniscono informazioni precise sulle decisioni da prendere una volta osservato il valore del confronto $\mathbf{y}$ di una coppia (in proposito si veda il capitolo 4). Generalmente la distribuzione per i match è più concentrata della distribuzione per i non-match.

Nell'appendice C si vedrà che le informazioni migliori (ovvero la maggiore "distanza" fra le distribuzioni) viene assicurata da confronti del tipo (2.13). Purtroppo questo tipo di confronti induce anche una maggiore complessità della distribuzione di probabilità di $\mathbf{Y}$. Soffermiamoci

su una variabile Y^h quando la corrispondente variabile chiave (X_A^h e X_B^h rispettivamente per le due rilevazioni) è numerica e con supporto l'insieme dei numeri reali $\mathbb{R}$ (approccio tipo superpopolazione). Considerando solo gli esempi di funzioni di confronto finora definite, si vede che:

1. se y_{ab}^h è definita come (2.10), il supporto di Y^h è $\{0, 1\}$;
2. se y_{ab}^h è definita come (2.11), il supporto di Y^h è $\mathbb{R}^+$;
3. se y_{ab}^h è definita come (2.12), il supporto di Y^h è $\mathbb{R} \bigcup \{\emptyset\}$;
4. se y_{ab}^h è definita come (2.13), il supporto di Y^h è $\mathbb{R}^2$.

Nella maggior parte dei metodi di record linkage, la semplicità della variabile aleatoria definita nel caso 1 ha indotto l'utilizzo della funzione di confronto (2.10), nonostante la sua scarsa capacità discriminatoria[4].

3.2 Il problema del record linkage e alcune soluzioni

Il problema del record linkage può essere formalizzato nel seguente modo.

Definizione 3.1 *Data una coppia (a, b), $a \in \mathcal{A}$, $b \in \mathcal{B}$, si deve decidere se l'osservazione $\mathbf{y}_{ab}$ è stata generata dalla distribuzione $m(\mathbf{y})$ o $u(\mathbf{y})$.*

Se $\mathbf{y}_{ab}$ è un'osservazione generata dalla distribuzione $m(\mathbf{y})$ allora (a, b) è una coppia in $\mathcal{M}$. Se invece è generata da $u(\mathbf{y})$, che è la distribuzione di $\mathbf{Y}$ in $\mathcal{U}$, allora si può affermare che la coppia (a, b) è una coppia in $\mathcal{U}$, e che quindi le due unità a e b sono diverse[5].

Uno strumento utile alla soluzione del problema posto è il test di ipotesi. Vengono definiti di seguito alcuni esempi di test di ipotesi utili alla soluzione del problema. Questi fanno riferimento a test di tipo Neyman-Pearson per le sue caratteristiche ottimali. Distinguiamo il caso in cui la variabile di confronto sia continua (come può accadere in (2.11)) dal caso in cui sia discreta (come accade ad esempio in (2.10)).

Esempio 3.8 - y continua. *Supponiamo di dover verificare quale fra le seguenti due ipotesi:*

H_0 : *la distribuzione vera è $m(\mathbf{y})$ (ipotesi nulla)*

H_1 : *la distribuzione vera è $u(\mathbf{y})$ (ipotesi alternativa)*

sia vera, avendo osservato esclusivamente il valore $\mathbf{y}_{ab}$ del confronto sulla coppia (a, b). Le ipotesi sono semplici, ovvero sono formate da un'unica distribuzione. In questo contesto è noto che esiste una procedura test ottimale: il lemma di Neyman-Pearson (per una sua definizione si veda, ad esempio, Lehmann, 1986). La procedura di decisione si svolge nel modo seguente.

1. *Si definisca la probabilità di errore di prima specie che si è disposti a tollerare:*

$$\lambda = P(\textit{scegliere } H_1 | H_0)$$

[4] Infatti si vedrà nel capitolo 6 che questa distribuzione deve essere stimata nella grande maggioranza dei casi, e la semplicità indotta dalla (2.10) induce un minor numero di parametri da stimare

[5] Questa regola, assai intuitiva, può essere modificata nel caso in cui si consideri un approccio di tipo Bayesiano. Di ciò si parlerà nel capitolo 7

2. *Si consideri la statistica-test definita dal lemma di Neyman-Pearson (il rapporto delle verosimiglianze):*

$$t(\mathbf{y}) = \frac{m(\mathbf{y})}{u(\mathbf{y})}$$

3. *Si individui la soglia τ_λ tale che:*

$$P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} < \tau_\lambda \middle| c_{a,b} = 1\right) = \lambda. \tag{3.5}$$

In pratica l'insieme $\mathcal{D}$ delle possibili osservazioni $\mathbf{y}$ viene diviso in due regioni in base al valore assunto dalla statistica test:

$$\Phi(\lambda) = \left\{\mathbf{y} : \frac{m(\mathbf{y})}{u(\mathbf{y})} < \tau_\lambda\right\}$$

e $\mathbb{R} - \Phi(\lambda)$. $\Phi(\lambda)$ è la regione di rifiuto dell'ipotesi H_0 al livello λ, dove λ è, riscrivendo la (3.5), la probabilità che la decisione di rifiutare H_0 sia sbagliata:

$$\int_{\Phi(\lambda)} m(\mathbf{y}) d\mathbf{y} = \lambda.$$

4. *Il test rifiuta l'ipotesi nulla se l'osservazione $\mathbf{y}_{ab}$ relativa alla coppia (a, b) è in $\Phi(\lambda)$, e si può affermare che la coppia in questione è in $\mathcal{U}$. Se l'osservazione $\mathbf{y}_{ab}$ è tale che $t(\mathbf{y}_{ab})$ è in $\mathbb{R} - \Phi(\lambda)$ si accetta l'ipotesi nulla.* □

Esempio 3.9 - y continua. *Invertiamo ora l'ordine delle ipotesi. Supponiamo quindi che:*

H_0 : *la distribuzione vera è $u(\mathbf{y})$ (ipotesi nulla)*

H_1 : *la distribuzione vera è $m(\mathbf{y})$ (ipotesi alternativa).*

Come in precedenza, si deve scegliere una delle due ipotesi in base all'evidenza campionaria data dal confronto $\mathbf{y}_{ab}$ relativo alla coppia (a, b). Utilizzando di nuovo il lemma di Neyman-Pearson si ottiene la seguente procedura di decisione.

1. *Si definisca la probabilità di errore di prima specie che si è disposti a tollerare:*

$$\mu = P(\text{scegliere } H_1 | H_0).$$

2. *Si individui la soglia τ_μ tale che:*

$$P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} > \tau_\mu \middle| c_{a,b} = 0\right) = \mu. \tag{3.6}$$

In pratica la statistica test è ancora il rapporto di verosimiglianze:

$$t(\mathbf{y}) = \frac{m(\mathbf{y})}{u(\mathbf{y})}$$

e l'insieme $\mathcal{D}$ delle possibili osservazioni $\mathbf{y}$ viene diviso in due regioni in base al valore assunto dalla statistica test:

$$\Phi(\mu) = \left\{ \mathbf{y} : \frac{m(\mathbf{y})}{u(\mathbf{y})} > \tau_\mu \right\}$$

e $I\!R - \Phi(\mu)$. $\Phi(\mu)$ è la regione di rifiuto del test al livello μ, ed è tale che, riscrivendo la (3.6),

$$\int_{\Phi(\mu)} u(\mathbf{y}) d\mathbf{y} = \mu.$$

3. *Se $\mathbf{y}_{ab} \in \Phi(\mu)$ si rifiuta l'ipotesi nulla e si può affermare che la coppia (a, b) in questione è in $\mathcal{M}$. In caso contrario si accetta l'ipotesi nulla.* □

Esempio 3.10 - y discreta. *Ritorniamo al caso in cui le ipotesi poste a confronto siano:*

H_0 : *la distribuzione vera è $m(\mathbf{y})$ (ipotesi nulla)*

H_1 : *la distribuzione vera è $u(\mathbf{y})$ (ipotesi alternativa).*

Se la funzione di confronto $\mathbf{y}$ è discreta, le procedure descritte nei due esempi precedenti devono essere leggermente modificate, in quanto non è possibile determinare gli insiemi di confronti $\mathbf{y}$ dove prendere delle decisioni certe (si accetta o rifiuta H_0 come negli esempi 3.8 e 3.9) per una qualsiasi probabilità di errore di prima specie (livello di qualità del test che si desidera). È invece necessario avviare un "gioco probabilistico": avendo osservato il confronto $\mathbf{y}$ si accetta o rifiuta H_0 secondo una opportuna distribuzione di probabilità. Questi test sono chiamati "test randomizzati" (per una definizione più precisa si veda Lehmann, 1986). Il lemma di Neyman-Pearson porta alla seguente regola.

1. *Si definisca la probabilità di errore di prima specie che si è disposti a tollerare:*

$$\lambda = P(\text{scegliere } H_1 | H_0).$$

2. *Si individui la soglia τ_λ e la probabilità P_λ tali che:*

$$P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} < \tau_\lambda \middle| c_{a,b} = 1 \right) < \lambda \leq P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} \leq \tau_\lambda \middle| c_{a,b} = 1 \right),$$

$$P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} = \tau_\lambda \middle| c_{a,b} = 1 \right) P_\lambda = \lambda - P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} < \tau_\lambda \middle| c_{a,b} = 1 \right).$$

3. *La regola di decisione del test è la seguente:*

   a- si assume con certezza la decisione H_0 se si osserva un confronto $\mathbf{y}$ tale che

$$\frac{m(\mathbf{y})}{u(\mathbf{y})} > \tau_\lambda;$$

   b- se il confronto $\mathbf{y}$ è tale che

$$\frac{m(\mathbf{y})}{u(\mathbf{y})} = \tau_\lambda$$

   allora si avvia il gioco probabilistico: si rigetta l'ipotesi nulla con probabilità P_λ e si accetta con probabilità $1 - P_\lambda$;

c- se il confronto $\mathbf{y}$ *è tale che*

$$\frac{m(\mathbf{y})}{u(\mathbf{y})} < \tau_\lambda$$

allora si rifiuta l'ipotesi nulla. □

Esempio 3.11 - y discreta. *Invertiamo ora l'ordine delle ipotesi. Supponiamo quindi che:*

H_0 : *la distribuzione vera è* $u(\mathbf{y})$ *(ipotesi nulla)*

H_1 : *la distribuzione vera è* $m(\mathbf{y})$ *(ipotesi alternativa).*

La procedura di decisione derivante dal lemma di Neyman-Pearson è molto simile a quella dell'esempio 3.10.

1. *Si definisca la probabilità di errore di prima specie che si è disposti a tollerare:*

$$\mu = P(\text{scegliere } H_0 | H_1).$$

2. *Si individui la soglia* τ_μ *e la probabilità* P_μ *tali che:*

$$P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} > \tau_\mu \middle| c_{a,b} = 0\right) < \mu \leq P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} \geq \tau_\mu \middle| c_{a,b} = 0\right),$$

$$P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} = \tau_\mu \middle| c_{a,b} = 0\right) P_\mu = \mu - P\left(\frac{m(\mathbf{y})}{u(\mathbf{y})} > \tau_\mu \middle| c_{a,b} = 0\right).$$

3. *La regola di decisione del test è la seguente:*

   a- si assume con certezza la decisione H_0 *se si osserva un confronto* $\mathbf{y}$ *tale che*

$$\frac{m(\mathbf{y})}{u(\mathbf{y})} < \tau_\mu;$$

   b- se il confronto $\mathbf{y}$ *è tale che*

$$\frac{m(\mathbf{y})}{u(\mathbf{y})} = \tau_\mu$$

   allora si avvia il gioco probabilistico: si rigetta l'ipotesi nulla con probabilità P_μ *e si accetta con probabilità* $1 - P_\mu$*;*

   c- se il confronto $\mathbf{y}$ *è tale che*

$$\frac{m(\mathbf{y})}{u(\mathbf{y})} > \tau_\mu$$

   allora si rifiuta l'ipotesi nulla. □

Capitolo 4

La procedura di Fellegi e Sunter

Chiarito il contesto nel quale applichiamo i metodi per il record linkage, definiamo ora una procedura di abbinamento ottimale. Questa procedura è stata definita da Fellegi e Sunter nel loro lavoro del 1969, e costituisce il "cuore" della maggior parte delle regole di record linkage discusse da allora ad oggi. Sono state definite anche altre procedure, che non tengono conto dei risultati ottenuti da Fellegi e Sunter o ne tengono conto in modo parziale. Tali risultati verranno discussi nel capitolo 7.

4.1 L'abbinamento come problema di decisione

Per una coppia di record (a, b), chi sta conducendo il record linkage ha a disposizione tre possibili decisioni:

- A_m: si decide che la coppia è un match
- A_u: si decide che la coppia è un non-match
- $A_\emptyset$: non si hanno sufficienti informazioni per poter decidere.

Naturalmente le decisioni vengono prese in base alle uniche informazioni che si hanno a disposizione sulla coppia: il confronto $\mathbf{y}_{ab}$ osservato sulla coppia stessa. Le decisioni A_m e A_u vengono chiamate "decisioni positive". La decisione $A_\emptyset$ viene presa quando si ritiene che l'osservazione non fornisce sufficienti informazioni per scegliere una decisione positiva. Viene chiamata anche "match incerto", in quanto a questa decisione corrisponde il controllo manuale della coppia di record da parte di impiegati specializzati per verificare lo status della coppia. Quindi a $A_\emptyset$ corrisponde un costo in termini monetari e di tempo, e è opportuno scegliere $A_\emptyset$ nel numero minore possibile di casi (meno probabile).

4.1.1 Le fasi della procedura di decisione

Le fasi per la costruzione della procedura di decisione sono essenzialmente due. In primo luogo si trasforma il vettore $\mathbf{y}_{ab}$, di dimensione k, in un numero reale. Questo passo viene risolto attraverso la definizione di una funzione dei vettori di confronto, che chiamiamo $t(\mathbf{y}_{ab})$. A questa funzione viene dato il nome di "peso". Ad esempio, se i confronti sono definiti come in (2.10) il peso può essere definito da:

$$t(\mathbf{y}_{ab}) = \sum_{h=1}^{k} y_{ab}^h. \tag{4.1}$$

La procedura di decisione viene facilitata se $t(\mathbf{y}_{ab})$ contiene tutte le informazioni utili a discriminare le coppie che provengono da $\mathcal{M}$ e $\mathcal{U}$. Il secondo passo della procedura di decisione consiste nella determinazione di due intervalli disgiunti, attraverso cui identificare rispettivamente i pesi $t(\mathbf{y}_{ab})$ delle coppie (a, b) in $\mathcal{M}$ e quelli delle coppie (a, b) in $\mathcal{U}$. Ad esempio, per i pesi (4.1) si devono determinare due numeri τ_1 e τ_2, con $0 < \tau_1 < \tau_2 < k$. L'intervallo $[\tau_2, k]$ sarà quello caratterizzante le coppie in $\mathcal{M}$, dato che i confronti definiti dalla (2.10) registrano plausibilmente un maggior numero di 1 quando $(a, b) \in \mathcal{M}$: in questo intervallo si sceglie A_m. Per lo stesso motivo l'intervallo $[0, \tau_1]$ caratterizza le coppie che sono non-match, e la decisione associata ai pesi in questo intervallo è A_u.

Fra i diversi modi di definire pesi e intervalli, illustriamo nel paragrafo successivo quello proposto da Fellegi e Sunter.

4.1.2 I pesi e gli intervalli proposti da Fellegi e Sunter

Fellegi e Sunter propongono di trasformare ogni vettore di confronto $\mathbf{y} \in \mathcal{D}$ nel peso

$$t(\mathbf{y}) = \frac{m(\mathbf{y})}{u(\mathbf{y})}. \tag{4.2}$$

Se tale peso assume un valore alto, il confronto $\mathbf{y}$ si presenta con maggiore probabilità per le coppie in $\mathcal{M}$ rispetto alle coppie in $\mathcal{U}$, e quindi è più verosimile ritenere che $\mathbf{y}$ provenga da una coppia appartenente a $\mathcal{M}$; se $t(\mathbf{y})$ ha un valore è basso è più verosimile ritenere che $\mathbf{y}$ provenga da una coppia in $\mathcal{U}$.

Il significato intuitivo del peso giustifica il passo successivo: si definiscano due soglie τ_μ e τ_λ, con $\tau_\mu > \tau_\lambda$, dipendenti da valori μ e λ fissati, $0 < \lambda < 1$, $0 < \mu < 1$, il cui significato verrà spiegato in seguito. Queste due soglie identificano gli intervalli in cui vengono prese le decisioni A_m, A_u e $A_\emptyset$:

- per gli $\mathbf{y} \in \mathcal{D}$ con peso $t(\mathbf{y}) \geq \tau_\mu$ (ovvero per i pesi più alti) si prende la decisione A_m;
- per gli $\mathbf{y} \in \mathcal{D}$ con peso $t(\mathbf{y}) \leq \tau_\lambda$ (ovvero per i pesi più bassi) si prende la decisione A_u;
- per gli $\mathbf{y} \in \mathcal{D}$ con peso $\tau_\lambda < t(\mathbf{y}) < \tau_\mu$ si stabilisce che la coppia è un match incerto, adottando la decisione $A_\emptyset$.

4.1.3 Le probabilità di errore per le decisioni

Alla regola di decisione appena definita è associato un certo livello di errore. Infatti una coppia (a, b) appartenente a $\mathcal{U}$ può presentare un confronto $\mathbf{y}$ il cui peso $t(\mathbf{y})$ è non inferiore a τ_μ con probabilità $u(\mathbf{y})$. Sia μ la probabilità che la decisione A_m sia sbagliata. Per come sono stati costruiti gli intervalli:

$$\mu = \sum_{\mathbf{y}: t(\mathbf{y}) \geq \tau_\mu} u(\mathbf{y}). \tag{4.3}$$

Allo stesso modo, sia λ la probabilità che la decisione A_u sia sbagliata:

$$\lambda = \sum_{\mathbf{y}: t(\mathbf{y}) \leq \tau_\lambda} m(\mathbf{y}). \tag{4.4}$$

Fellegi e Sunter suggeriscono di fissare inizialmente i livelli di errore μ e λ a valori considerati accettabili, e passare quindi alla determinazione delle soglie τ_μ e τ_λ corrispondenti, in base alle formule (4.3) e (4.4). In questo caso può capitare che non esista un vettore di confronto $\boldsymbol{\xi} \in \mathcal{D}$ che verifichi la relazione:

$$P\Big(t(\mathbf{Y}) \geq t(\boldsymbol{\xi})\Big|c = 0\Big) = \sum_{\mathbf{y}:t(\mathbf{y})\geq t(\boldsymbol{\xi})} u(\mathbf{y}) = \mu, \tag{4.5}$$

così come può non esistere un vettore di confronto $\boldsymbol{\zeta} \in \mathcal{D}$ tale che, al livello λ fissato:

$$P\Big(t(\mathbf{Y}) \leq t(\boldsymbol{\zeta})\Big|c = 1\Big) = \sum_{\mathbf{y}:t(\mathbf{y})\leq t(\boldsymbol{\zeta})} m(\mathbf{y}) = \lambda. \tag{4.6}$$

Fellegi e Sunter suggeriscono di randomizzare gli estremi degli intervalli, in modo tale da ottenere insiemi in cui le decisioni A_m e A_u sono associate con le probabilità di errore μ e λ fissate.

Per quanto riguarda l'intervallo per la decisione A_m, si tratta di individuare quel vettore di confronto $\boldsymbol{\xi} \in \mathcal{D}$ tale che

$$P\Big(t(\mathbf{Y}) > t(\boldsymbol{\xi})\Big|c = 0\Big) < \mu$$

e

$$P\Big(t(\mathbf{Y}) \geq t(\boldsymbol{\xi})\Big|c = 0\Big) > \mu$$

e si pone $\tau_\mu = t(\boldsymbol{\xi})$. Dato che le coppie che sono non-match assumono pesi nell'intervallo $[\tau_\mu, +\infty)$ con probabilità superiore a μ, si lima la probabilità di questo intervallo dividendo la probabilità:

$$P\Big(t(\mathbf{Y}) = \tau_\mu\Big|c = 0\Big) = P\Big(t(\mathbf{Y}) \geq \tau_\mu\Big|c = 0\Big) - P\Big(t(\mathbf{Y}) > \tau_\mu\Big|c = 0\Big)$$

in due parti, che sono le componenti della randomizzazione. Sia

$$P_\mu = \frac{\mu - P\Big(t(\mathbf{Y}) > \tau_\mu\Big|c = 0\Big)}{P\Big(t(\mathbf{Y}) \geq \tau_\mu\Big|c = 0\Big) - P\Big(t(\mathbf{Y}) > \tau_\mu\Big|c = 0\Big)}. \tag{4.7}$$

la probabilità che si prenda la decisione A_m se si osserva un peso pari a τ_μ. Con probabilità $1 - P_\mu$ la coppia viene considerata un match incerto.

Lo stesso ragionamento può essere fatto quando non esiste un valore che soddisfa la (4.6). Preso il vettore di confronto $\boldsymbol{\zeta} \in \mathcal{D}$ tale che

$$P\Big(t(\mathbf{Y}) < t(\boldsymbol{\zeta})\Big|c = 1\Big) < \lambda$$

e

$$P\Big(t(\mathbf{Y}) \leq t(\boldsymbol{\zeta})\Big|c = 1\Big) > \lambda,$$

si ponga $\tau_\lambda = t(\boldsymbol{\zeta})$ e si determini la probabiltà:

$$P_\lambda = \frac{\lambda - P\Big(t(\mathbf{Y}) < \tau_\lambda\Big|c = 1\Big)}{P\Big(t(\mathbf{Y}) \leq \tau_\lambda\Big|c = 1\Big) - P\Big(t(\mathbf{Y}) < \tau_\lambda\Big|c = 1\Big)}. \tag{4.8}$$

Tabella 4.1 - Regola di decisione quando sono vere le (4.5) e (4.6).

Peso dei confronti	$P(A_m\|\mathbf{y})$	$P(A_\emptyset\|\mathbf{y})$	$P(A_u\|\mathbf{y})$
$t(\mathbf{y})\geq\tau_\mu$	1	0	0
$\tau_\lambda<t(\mathbf{y})<\tau_\mu$	0	1	0
$t(\mathbf{y})\leq\tau_\lambda$	0	0	1

Commento 4.1 *Si vuole sottolineare il fatto che, sia che le (4.5) e (4.6) siano rispettate o no, i valori di soglia τ_λ e τ_μ sono determinati in modo che:*

$$\tau_\mu = \min\Big\{t \in I\!R : P\Big(t(\mathbf{Y}) > t\Big|c = 0\Big) \leq \mu\Big\},$$

$$\tau_\lambda = \max\Big\{t \in I\!R : P\Big(t(\mathbf{Y}) < t\Big|c = 1\Big) \leq \lambda\Big\}.$$

□

4.1.4 L'ottimalità della regola di Fellegi e Sunter

Nei paragrafi precedenti è stato più volte affermato che la regola di decisione formulata da Fellegi e Sunter è, in un certo senso, "ottimale". Per chiarire questo concetto, è necessario definire dapprima rispetto a quali altre procedure è ottimale, e quindi in che senso lo è.

Fellegi e Sunter considerano un insieme di regole plausibili per poter scegliere fra le tre decisioni A_m, A_u e $A_\emptyset$ definite dal seguente "gioco probabilistico":

Definizione 4.1 *Avendo osservato il confronto $\mathbf{y}$, $\mathbf{y} \in \mathcal{D}$, si scelga fra le decisioni A_m, A_u e $A_\emptyset$ secondo la distribuzione:*

$$P(A_m|\mathbf{y}),\ P(A_u|\mathbf{y}),\ P(A_\emptyset|\mathbf{y})$$

con

$$P(A_m|\mathbf{y}) \geq 0,\ P(A_u|\mathbf{y}) \geq 0,\ P(A_\emptyset|\mathbf{y}) \geq 0$$

e

$$P(A_m|\mathbf{y}) + P(A_u|\mathbf{y}) + P(A_\emptyset|\mathbf{y}) = 1.$$

Le regole della definizione 4.1 sono tante quanti sono i modi diversi in cui si possono definire le distribuzioni $\{P(A_m|\mathbf{y}),\ P(A_u|\mathbf{y}),\ P(A_\emptyset|\mathbf{y}),\ \mathbf{y} \in \mathcal{D}\}$. Si noti che le regole di decisione definite nei paragrafi 4.1.2 e 4.1.3 rientrano nell'insieme di regole della definizione 4.1 sia quando sono valide le (4.5) e (4.6) che quando queste condizioni non sono rispettate. Le probabilità delle decisioni sono rappresentate rispettivamente nelle tabelle 4.1 e 4.2.
Per tutte queste regole è ancora possibile definire le probabilità di errore μ e λ. La probabilità che una coppia appartenente a $\mathcal{M}$ venga erroneamente assegnata a $\mathcal{U}$ con una regola della definizione 4.1 è:

$$\lambda = \sum_{\mathbf{y}\in\mathcal{D}} m(\mathbf{y})P(A_u|\mathbf{y}). \tag{4.9}$$

Tabella 4.2 - Regola di decisione quando non sono vere le (4.5) e (4.6).

Peso dei confronti	$P(A_m\|\mathbf{y})$	$P(A_\emptyset\|\mathbf{y})$	$P(A_u\|\mathbf{y})$
$t(\mathbf{y})>\tau_\mu$	1	0	0
$t(\mathbf{y})=\tau_\mu$	P_μ	$1-P_\mu$	0
$\tau_\lambda<t(\mathbf{y})<\tau_\mu$	0	1	0
$t(\mathbf{y})=\tau_\lambda$	0	$1-P_\lambda$	P_λ
$t(\mathbf{y})<\tau_\lambda$	0	0	1

Allo stesso modo, la probabilità di prendere una decisione errata per una coppia che è un non-match è:

$$\mu = \sum_{\mathbf{y}\in\mathcal{D}} u(\mathbf{y})P(A_m|\mathbf{y}). \tag{4.10}$$

Fellegi e Sunter forniscono la seguente definizione di ottimalità di una regola.

Definizione 4.2 *Si considerino tutte le regole della definizione 4.1 con probabilità di errore $\lambda^* \leq \lambda$ e $\mu^* \leq \mu$. Una regola è ottimale in questa classe se minimizza la probabilità di decidere che una coppia è un match incerto:*

$$\sum_{\mathbf{y}\in\mathcal{D}} m(\mathbf{y})P(A_\emptyset|\mathbf{y}); \quad \sum_{\mathbf{y}\in\mathcal{D}} u(\mathbf{y})P(A_\emptyset|\mathbf{y}).$$

Dato che ogni match incerto deve essere studiato da personale specializzato per stabilirne la natura, le regole ottimali hanno il pregio di minimizzare i costi. Fellegi e Sunter dimostrano formalmente che la regola descritta nelle tabelle 4.1 o 4.2 è ottimale fra tutte quelle con livello di errori minore o uguale rispettivamente a λ e μ.

Commento 4.2 *Si nota che gli intervalli per le decisioni positive individuati dalla regola di decisione di Fellegi e Sunter (nelle tabelle 4.1 e 4.2) sono esattamente gli intervalli delle zone di rifiuto dei test negli esempi 3.8 e 3.9 (quando* **y** *è continua) e 3.10 e 3.11 quando* **y** *è discreta. Le zone di rifiuto dei test sono quelle in cui la probabilità di commettere un errore è sotto il diretto controllo di chi sta conducendo l'analisi. L'ottimalità della regola di Fellegi e Sunter (cioè la minor probabilità di decidere per un match incerto) è una diretta conseguenza dell'ottimalità dei test di Neyman-Pearson usati negli esempi 3.8-3.11.*

Oltre a tener conto dei costi connessi alle operazioni di record linkage, la regola di Fellegi e Sunter ha il pregio di trattare in modo simmetrico le ipotesi che compongono le decisioni positive, mentre nei test discussi negli esempi 3.8-3.11 una delle decisioni positive si comporta come l'ipotesi principale da accettare o rifiutare. □

Commento 4.3 *Risulta evidente che chi deve abbinare due liste A e B sceglie la regola di record linkage in base a due criteri:*

1. *la qualità per mezzo delle probabilità λ e μ*

2. *il budget e/o i tempi a disposizione per condurre il record linkage.*

Questi due fattori sono in contrapposizione fra loro. Infatti, supponiamo di voler garantire la qualità del record linkage in modo perfetto, imponendo che le probabilità di commettere un errore connesso alle decisioni positive siano nulle (cioè $\lambda = 0$ e $\mu = 0$). La regola di Fellegi e Sunter richiederebbe che le regioni delle tabelle 4.1 e 4.2 corrispondenti alle decisioni positive fossero vuote, e tutti i record fossero considerati come match incerti e analizzati manualmente. In pratica, annullare le probabilità di errore corrisponde a massimizzare i costi. Viceversa, si può rendere vuoto l'insieme dei match incerti (e quindi annullare i costi) solo al prezzo di incrementare le probabilità di errore. □

Commento 4.4 *Non tutte le coppie (λ, μ) di probabilità di errore della regola di record linkage sono ammissibili. Infatti è possibile definire una soglia massima per questi valori. La funzione che esprime questa soglia è stata definita da Fellegi e Sunter quando $\mathbf{Y}$ è una variabile discreta (si veda l'appendice I di Fellegi e Sunter, 1969). Qui si dà una definizione equivalente per il caso descritto nella tabella 4.1.*

Si fissi un unico valore di soglia ($\tau = \tau_\lambda = \tau_\mu$) e si considerino le funzioni:

$$U(\tau) = \int_{\{\mathbf{y}: t(\mathbf{y}) \geq \tau\}} u(\mathbf{y}) d\mathbf{y}, \qquad \tau \geq 0,$$

$$M(\tau) = \int_{\{\mathbf{y}: t(\mathbf{y}) < \tau\}} m(\mathbf{y}) d\mathbf{y}, \qquad \tau \geq 0.$$

Si noti che, fissato $\tau_\mu = \tau$, il valore $U(\tau)$ è la probabilità μ mentre $M(\tau)$ rappresenta il valore massimo per la probabilità λ (dato $\mu = U(\tau)$, se $\lambda > M(\tau)$ le regioni per le decisioni dei match e dei non-match sarebbero sovrapposte). Si definisca un piano cartesiano avente sull'asse delle ascisse i valori μ e sull'asse delle ordinate i valori λ. Si construisca la funzione (monotona decrescente) definita dai punti di coordinate:

$$f(\mu) = \Big(U(\tau),\ M(\tau)\Big) = \Big(\mu,\ M(U^{-1}(\mu))\Big).$$

La parte di piano compresa fra gli assi cartesiani e la funzione $f(\mu)$ definisce le coppie di probabilità di errori (μ, λ) ammissibili per la regola di Fellegi-Sunter. Se si considera una coppia di errori sulla funzione $f(\mu)$ allora si rende l'insieme dei match incerti vuoto. Una coppia di probabilità di errore al di sotto della funzione fa sì che l'insieme di match incerti diventi via via più grande (più probabile). □

Commento 4.5 *Se si adotta un approccio tipo "popolazioni finita" (commento 3.1), allora le probabilità di errore λ e μ assumono un significato particolarmente interessante. In questo caso $m(\mathbf{y})$ è la distribuzione di frequenze di $\mathbf{Y}$ sulle coppie in $\mathcal{M}$ (N coppie), mentre $u(\mathbf{y})$ è la distribuzione di frequenze di $\mathbf{Y}$ sulle coppie in $\mathcal{U}$. Applicando la regola di Fellegi-Sunter su tutte le $\nu_A \times \nu_B$ coppie, ed indicando con $I_{A_u}(a, b)$ la funzione indicatrice che vale 1 se si assume la decisione A_u e zero altrimenti, si ha:*

$$\lambda = \sum_{\mathbf{y} \in \mathcal{D}} m(\mathbf{y}) P(A_u|\mathbf{y}) = E\left[\sum_{(a,b) \in \mathcal{A} \times \mathcal{B}} \frac{I_{A_u}(a, b)}{N}\right].$$

Quindi λ diventa la frequenza attesa delle coppie in $\mathcal{M}$ che vengono classificate erroneamente come non-match secondo la regola di Fellegi-Sunter. Anche μ può essere interpretato allo stesso modo. Infatti:

$$\mu = \sum_{\mathbf{y}\in\mathcal{D}} u(\mathbf{y})P(A_m|\mathbf{y}) = E\left[\sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} \frac{I_{A_m}(a,b)}{\nu_A \times \nu_B - N}\right]$$

descrive la frequenza attesa delle coppie in $\mathcal{U}$ che vengono erroneamente classificate come match. Se, infine, le procedure di decisione non sono randomizzate (come nella tabella 4.1), μ e λ non sono frequenze attese ma frequenze effettive. □

Commento 4.6 *Anche quando la variabile $\mathbf{Y}$ è definita secondo l'approccio di tipo superpopolazione (commento 3.1) può essere desiderabile che μ e λ assumano lo stesso significato descritto nel commento 4.5. Questo è ancora possibile grazie alla legge dei grandi numeri. Ad esempio, si consideri l'approccio tipo superpopolazione delineato negli esempi 3.5 e 3.7. Supponendo che il numero N di coppie che sono match sia sufficientemente elevato (questo è generalmente vero nel record linkage) la distribuzione di frequenze sulle N coppie di $\mathbf{Y}$ si approssima (a meno di errori trascurabili) alla corrispondente distribuzione del modello di superpopolazione $m(\mathbf{y})$. Lo stesso ragionamento può essere fatto anche per $u(\mathbf{y})$. Per variabili più complicate rispetto a variabili X^h qualitative, invece che alla legge dei grandi numeri si può far riferimento ad altri strumenti (come il teorema di Glivenko-Cantelli).* □

4.2 Alcune trasformazioni

A volte non vengono usati i pesi $t(\mathbf{y})$ definiti in (4.2), ma alcune trasformazioni.

Una trasformazione adottata molto spesso consiste nel considerare il logaritmo dei pesi suggeriti da Fellegi e Sunter

$$w(\mathbf{y}) = \log\Big(t(\mathbf{y})\Big). \tag{4.11}$$

Questa trasformazione viene adottata in particolare quando i confronti fra le variabili chiave Y^h, $h = 1, ..., k$, sono indipendenti sia per le coppie che sono match che per le coppie che sono non-match. Infatti la (4.11) si scompone nella somma dei pesi relativi ognuno a un confronto Y^h:

$$\begin{aligned} w(\mathbf{y}) &= \log\Big(m(\mathbf{y})\Big) - \log\Big(u(\mathbf{y})\Big) = \log\left(\prod_{h=1}^{k} m_h(y^h)\right) - \log\left(\prod_{h=1}^{k} u_h(y^h)\right) = \\ &= \sum_{h=1}^{k} \log\Big(m_h(y^h)\Big) - \log\Big(u_h(y^h)\Big) = \sum_{h=1}^{k} \log\left(\frac{m_h(y^h)}{u_h(y^h)}\right). \end{aligned}$$

Poiché il logaritmo è una funzione strettamente crescente, l'ordinamento fra i vettori di confronto $\mathbf{y} \in \mathcal{D}$ indotto dai pesi $w(\mathbf{y})$ è identico a quello indotto dai pesi $t(\mathbf{y})$.

Una diversa trasformazione viene adottata in ISTAT, ad esempio per i metodi di record linkage usati per ricongiungere i dati delle forze lavoro relativi a indagini successive. Questa trasformazione, (Torelli, 1998, Torelli e Paggiaro, 1999, Larsen e Rubin, 2001), necessita di un elemento in più rispetto a quelli necessari per il calcolo di $t(\mathbf{y})$: la probabilità p che una coppia sia un match,

$$p = P(C_{a,b} = 1).$$

Ad ogni vettore di confronto $\mathbf{y} \in \mathcal{D}$ si associa il nuovo peso:

$$t^*(\mathbf{y}) = \frac{e^{w(\mathbf{y})}}{e^{w(\mathbf{y})} + \frac{1-p}{p}}. \tag{4.12}$$

Una giustificazione per questa probabilità sarà disponibile nel paragrafo 6.2. Questo peso $t^*(\mathbf{y})$ è una trasformazione logistica del logaritmo dei pesi $w(\mathbf{y})$. Dato che anche la funzione logistica in (4.12) è una funzione strettamente crescente del peso $w(\mathbf{y})$, in quanto $(1-p)/p$ è un numero positivo, l'ordinamento fra i vettori di confronto $\mathbf{y} \in \mathcal{D}$ indotto dai pesi $t^*(\mathbf{y})$ è identico a quello indotto dai pesi (4.2). Quindi, per gli stessi errori μ e λ è possibile determinare dei valori di soglia τ^*_μ e τ^*_λ che sezionano lo spazio dei confronti $\mathcal{D}$ negli stessi insiemi indotti dalle regole nella tabelle 4.1 e 4.2.

Il vantaggio nell'uso dei pesi (4.12) deriva dal loro significato: $t^*(\mathbf{y})$ è ora la probabilità che la coppia che presenta il confronto $\mathbf{y}$ sia un match. Infatti per il teorema di Bayes:

$$\begin{aligned} P(C=1|\mathbf{Y}=\mathbf{y}) &= \frac{P(C=1)P(\mathbf{Y}=\mathbf{y}|C=1)}{P(C=1)P(\mathbf{Y}=\mathbf{y}|C=1) + P(C=0)P(\mathbf{Y}=\mathbf{y}|C=0)} = \\ &= \frac{p\,m(\mathbf{y})}{p\,m(\mathbf{y}) + (1-p)\,u(\mathbf{y})} \end{aligned} \tag{4.13}$$

e dividendo numeratore e denominatore per $p\,u(\mathbf{y})$:

$$P(C=1|\mathbf{Y}=\mathbf{y}) = \frac{t(\mathbf{y})}{t(\mathbf{y}) + \frac{1-p}{p}} = \frac{e^{w(\mathbf{y})}}{e^{w(\mathbf{y})} + \frac{1-p}{p}} = t^*(\mathbf{y}).$$

Quindi gli intervalli delle tabelle 4.1 e 4.2 assumono un nuovo significato: si assume la decisione A_m quando la probabilità che una coppia sia un match, dato che presenta il vettore di confronto $\mathbf{y}$, è elevata, e viceversa si assume la decisione A_u quando la probabilità che la coppia sia un match, dato che il suo vettore di confronto è $\mathbf{y}$, è bassa.

4.3 Come eseguire la procedura su tutte le coppie

Nel paragrafo precedente è stata definita la regola di Fellegi-Sunter facendo riferimento a una generica coppia (a, b) e alle caratteristiche di ottimalità del metodo. Di seguito si illustrano i passi per applicare la regola su tutte le $\nu = n_A \times n_B$ coppie definite dal confronto fra A e B. Si consideri che, avendo a che fare con ν coppie, i vettori di confronto osservati sono un numero N finito.

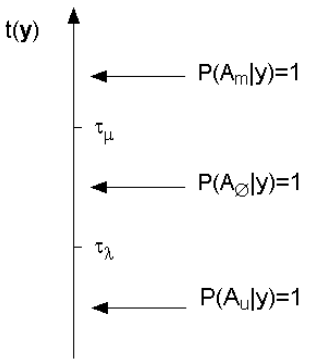
1. Si ordinino i possibili vettori di confronto $\mathbf{y} \in \mathcal{D}$ in modo decrescente rispetto al peso $t(\mathbf{y})$. Ai vettori $\mathbf{y}$ con $m(\mathbf{y}) > 0$ e $u(\mathbf{y}) = 0$ vengono assegnati i primi posti dell'ordinamento in modo arbitrario. Si etichetti ogni vettore di confronto con il corrispondente posto nell'ordinamento, diciamo j, $j = 1, ..., N$ così che, se $\mathbf{y}$ è il j-esimo vettore nell'ordinamento, allora $m(\mathbf{y}) = m_j$ e $u(\mathbf{y}) = u_j$.

2. Fissate le due probabilità di errore μ e λ, si determinino i due numeri interi $n < n'$ tali che

$$\sum_{j=1}^{n-1} u_j < \mu \leq \sum_{j=1}^{n} u_j \tag{4.14}$$

$$\sum_{j=n'}^{N} m_j \geq \lambda > \sum_{j=n'+1}^{N} m_j. \tag{4.15}$$

Figura 4.1 - Per le coppie che presentano un peso pari a τ_μ è necessario affiancare un esperimento "fittizio" che accetti con probabilità P_μ la decisione A_m e con la restante probabilità la decisione $A_\emptyset$. Un esperimento analogo è necessario per le coppie con peso pari a τ_λ. Per le altre coppie la decisione da prendere è certa.



Senza perdere in generalità, supponiamo che le disuguaglianze in (4.14) e (4.15) siano strette, ovvero nelle condizioni della tabella 4.2. Il caso per la tabella 4.1 è immediato.

3. Si determinino le probabilità P_λ secondo la (4.8) e P_μ come in (4.7).

4. Per ogni coppia nell'ordinamento le decisioni che vengono prese sono le seguenti:

   - per i vettori di confronto **y** con etichetta $j \leq n-1$ si accetti la decisione A_m
   - per i vettori di confronto **y** con etichetta $j = n$ si accetti la decisione A_m con probabilità P_μ e la decisione $A_\emptyset$ con probabilità $1 - P_\mu$
   - per i vettori di confronto **y** con etichetta $n < j \leq n' - 1$ si accetti la decisione $A_\emptyset$
   - per i vettori di confronto **y** con etichetta $j = n'$ si accetti A_u con probabilità P_λ e l'ipotesi $A_\emptyset$ con probabilità $1 - P_\lambda$
   - per i vettori di confronto **y** con etichetta $j \geq n' + 1$ si accetti la decisione A_u.

Questa procedura è esemplificata nella figura 4.1.

4.4 Sviluppi necessari

Come anticipato all'inizio di questo capitolo, la procedura proposta da Fellegi e Sunter non è direttamente applicabile. I passi per renderla applicabile verranno descritti nei capitoli successivi, dove si discutono i seguenti aspetti.

1. I problemi di ordine computazionale rendono spesso necessario restringere l'insieme delle coppie che il computer deve abbinare. Il metodo suggerito va sotto il nome di bloccaggio ed è analizzato nel paragrafo 5.1.

2. Quando sono vere le (2.3) e (2.4) si devono modificare i risultati della procedura nel paragrafo 4.3 per far sì che vengano rispettati i vincoli (2.5), (2.6) e (2.7). Nel paragrafo 5.2 si discutono le proposte per soddisfare questi requisiti.

3. In questo capitolo le distribuzioni $m(\mathbf{y})$ e $u(\mathbf{y})$ sono state considerate note, ma questo avviene raramente. Senza queste informazioni non è possibile definire i pesi $t(\mathbf{y})$ né le probabilità di errore μ e λ. Il problema della stima delle distribuzioni è di gran lunga il più dibattuto nella letteratura sul record linkage, e viene affrontato nel capitolo 6.

4. La qualità del record linkage è ben rappresentata da μ e λ, come visto nei paragrafi precedenti. Ma quando i punti 1, 2 e 3 vengono applicati, anche μ e λ devono essere stimati. I metodi di stima finora formulati vengono esposti nel capitolo 8.

Capitolo 5

Alcune modifiche alla procedura di Fellegi e Sunter

In questo capitolo, vengono definite due tecniche che modificano l'*input* e l'*output* della procedura proposta da Fellegi e Sunter. Per quanto riguarda l'input, è possibile che le coppie sottoposte alla procedura di record linkage siano un numero talmente grande da risultare ingestibile anche per i moderni elaboratori. Per risolvere questo tipo di problema è stata definita la fase di "bloccaggio" delle due basi dati A e B (paragrafo 5.1). Per quanto riguarda l'output, è possibile che le coppie considerate match non rispettino i vincoli (2.6) e (2.7). Jaro (1989) utilizza un algoritmo di ricerca operativa per ovviare a questo problema (paragrafo 5.2).

5.1 Il bloccaggio

Spesso il numero di coppie che le procedure di record linkage devono analizzare è particolarmente elevato. Come già detto, se i due gruppi di unità $\mathcal{A}$ e $\mathcal{B}$ sono rispettivamente di ν_A e ν_B unità, il numero di coppie presenti in $\mathcal{A} \times \mathcal{B}$ è di $\nu_A \times \nu_B$ unità.

Esempio 5.1 *Si supponga di dover abbinare i record di due basi dati A e B. Se i record contenuti in $\mathcal{A}$ e in $\mathcal{B}$ sono $1.000.000$, il numero di coppie da considerare sono $1.000.000.000.000$. Per basi dati più grandi, le cifre diventano ancora maggiori.* □

Questi numeri risultano ingestibili anche per gli elaboratori più potenti. Se, ad esempio, un programma di record linkage è in grado di controllare un milione di coppie al minuto, il controllo di tutte le coppie dell'esempio 5.1 richiederebbe:

$$\frac{1.000.000.000.000}{1.000.000 \times 60 \times 24} \approx 695$$

giorni.

Per ovviare a questo problema, Fellegi e Sunter (1969) propongono di eliminare dal controllo le coppie che possono essere considerate "improduttive". Infatti la maggior parte delle coppie risultanti dal prodotto cartesiano $\mathcal{A} \times \mathcal{B}$ sono non-match, e l'incidenza dei non-match sul totale delle coppie cresce al crescere delle dimensioni di $\mathcal{A}$ e $\mathcal{B}$. Il metodo proposto va sotto il nome di *blocking*, qui tradotto in *bloccaggio*.

5.1.1 Come si esegue un bloccaggio

Si consideri una variabile Z presente sia in A che in B che si suppone non sia affetta da errori o mancate risposte. Supponiamo anche che la semplice conoscenza di Z non individui univocamente le singole unità, altrimenti sarebbe sufficiente eseguire un *merge* fra le due liste attraverso Z per risolvere l'obiettivo del record linkage (capitolo 2). Supponiamo infine che Z assuma v_Z modalità distinte nelle due liste (per semplicità $z = 1, ..., v_Z$) rispettivamente con frequenza $\nu_{A;z}$ e $\nu_{B;z}$, $z = 1, ..., v_Z$. Gli insiemi $\mathcal{A}$ e $\mathcal{B}$ vengono quindi divisi in v_z gruppi $\mathcal{A}_z$ e $\mathcal{B}_z$, dove le unità $a \in \mathcal{A}_z$ e $b \in \mathcal{B}_z$ sono tali che:

$$z_a = z_b = z, \qquad \forall\ a \in \mathcal{A}_z,\ b \in \mathcal{B}_z,\ z = 1, ..., v_z.$$

Invece di studiare le $\nu_A \times \nu_B$ coppie in $\mathcal{A} \times \mathcal{B}$, Fellegi e Sunter propongono di analizzare solo le coppie nei v_z sottoinsiemi:

$$\bigcup_{z=1}^{v_z} \mathcal{A}_z \times \mathcal{B}_z, \tag{5.1}$$

ovvero solo le coppie di unità che presentano la stessa modalità di Z. Il numero complessivo di coppie che compongono l'unione in (5.1) è:

$$\sum_{z=1}^{v_Z} \nu_{A;z}\nu_{B;z}.$$

Esempio 5.2 *Continuando con l'esempio 5.1, se una variabile Z è distribuita uniformemente in $\mathcal{A}$ e in $\mathcal{B}$ e assume 1000 modalità (ad esempio Z può essere definita dalla combinazione di tre variabili con 10 modalità ciascuna), gli insiemi $\mathcal{A}_z$ e $\mathcal{B}_z$ sono composti ognuno da 1000 unità, e il tempo necessario ai calcoli è:*

$$\frac{1000 \times 1000 \times 1000}{1.000.000 \times 24 \times 60} = 0,69$$

giorni. □

Questa procedura porta ad un aumento dell'efficienza del record linkage in quanto vengono eliminate quelle coppie che non possono essere match in quanto non coincidono nella variabile Z. Se la variabile Z è affetta da errori e/o mancate risposte, è necessario porre estrema attenzione agli effetti dell'operazione del bloccaggio.

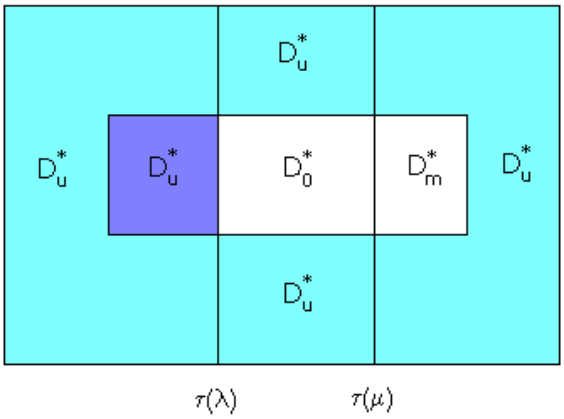
Si denoti con $\mathcal{D}^*$ l'insieme dei vettori di confronto nel sottoinsieme di unità

$$\mathcal{D}^* = \left\{ \mathbf{y}_{ab}, (a, b) \in \bigcup_{z=1}^{v_z} \mathcal{A}_z \times \mathcal{B}_z \right\}.$$

Lasciando inalterati gli elementi della procedura di Fellegi e Sunter τ_μ, τ_λ e il peso $t(\mathbf{y})$, una volta eseguito il bloccaggio la procedura controlla solamente le coppie nel rettangolo più interno della figura 5.1. La procedura viene quindi sintetizzata dalle decisioni:

$$d^*(\mathbf{y}) = \begin{cases} A_m & \text{se } t(\mathbf{y}) \geq \tau_\mu \text{ e } \mathbf{y} \in \mathcal{D}^* \\ A_\emptyset & \text{se } \tau_\lambda < t(\mathbf{y}) < \tau_\mu \text{ e } \mathbf{y} \in \mathcal{D}^* \\ A_u & \text{se } t(\mathbf{y}) \leq \tau_\lambda \text{ o } \mathbf{y} \in \mathcal{D} - \mathcal{D}^* \end{cases}$$

Figura 5.1 - Il rettangolo più grande corrisponde all'insieme delle $\nu_A \times \nu_B$ coppie. Il rettangolo interno corrisponde all'insieme delle coppie dopo avere effettuato il bloccaggio. L'insieme $\mathcal{D}_u^*$ è formato dall'unione delle sezioni in grigio chiaro e grigio scuro. L'insieme $\mathcal{D}_u$ è invece la sezione di rettangolo a sinistra di τ_λ, l'insieme $\mathcal{D}_m$ è delimitato da τ_μ mentre $\mathcal{D}_\emptyset$ è la zona compresa fra τ_λ e τ_μ.



In pratica, sempre controllando la figura 5.1, ora si prende la decisione A_m nella regione dei confronti $\mathbf{y}$:

$$\mathcal{D}_m^* = \left\{ \left\{ \mathbf{y} : t(\mathbf{y}) \geq \tau_\mu \right\} \bigcap \mathcal{D}^* \right\}$$

e la decisione $A_\emptyset$ nella regione

$$\mathcal{D}_\emptyset^* = \left\{ \left\{ \mathbf{y} : \tau_\lambda < t(\mathbf{y}) < \tau_\mu \right\} \bigcap \mathcal{D}^* \right\}$$

mentre nella regione

$$\mathcal{D}_u^* = \mathcal{D}_u \bigcup (\mathcal{D}^*)^c$$

si prende la decisione A_u.

A questa regola di decisione non corrispondono più i livelli di errore λ e μ. La probabilità che un non-match venga considerato un match (decisione A_m) è ora:

$$P\Big(d^*(\mathbf{Y}) = A_m \Big| c = 0\Big) = P\Big(t(\mathbf{Y}) \geq \tau_\mu \Big| c = 0\Big) - P\Big(\mathbf{Y} \in [\mathcal{D}_m - \mathcal{D}_m^*] \Big| c = 0\Big) < \mu \quad (5.2)$$

mentre viene presa la decisione A_u quando la coppia è un match con probabilità:

$$P\Big(d^*(\mathbf{Y}) = A_u \Big| c = 1\Big) = P\Big(t(\mathbf{Y}) \leq \tau_\lambda \Big| c = 1\Big) + P\Big(\mathbf{Y} \in [\mathcal{D}_u^* - \mathcal{D}_u] \Big| c = 1\Big) > \lambda \quad (5.3)$$

Quindi la procedura di bloccaggio presenta uno svantaggio e un vantaggio. Lo svantaggio è rappresentato dall'aumento di probabilità nel perdere dei match per via del bloccaggio:

$$P\Big(\mathbf{Y} \in [\mathcal{D}_u^* - \mathcal{D}_u] \Big| c = 1\Big).$$

Il vantaggio è legato alla diminuzione di probabilità di dichiarare dei match falsi per tutti quei confronti $\mathbf{y}$ che si trovano al di fuori di $\mathcal{D}^*$:

$$P\Big(\mathbf{Y} \in [\mathcal{D}_m - \mathcal{D}_m^*]\Big| c = 0\Big).$$

Oltre a Fellegi e Sunter, anche Kelley (1984, 1985) si è occupato delle procedure di bloccaggio. L'autore mostra in un esempio che la regola $d^*(\mathbf{y})$ è migliore rispetto ad altre regole definibili dopo il bloccaggio (Kelley, 1984) e fornisce una trattazione formale di come scegliere la procedura di bloccaggio migliore (Kelley, 1985). La procedura definita da Kelley è comunque applicabile solo quando i confronti fra variabili chiave sono fra loro indipendenti e soddisfano la definizione (2.10).

Commento 5.1 *A volte vengono eseguite più operazioni di bloccaggio. Ad esempio Armstrong e Saleh (2000) considerano 6 fasi successive di record linkage per abbinare il* National Register of Electors *e il* Canada Customs and Revenue Agency Data Base. *Ogni fase successiva analizza solo le coppie che non sono state dichiarate match precedentemente e le coppie analizzate vengono bloccate secondo combinazioni diverse di variabili.*

Torelli e Paggiaro (1999) hanno invece considerato un "doppio blocco alternativo" per abbinare le basi dati delle forze lavoro relative a due rilevazioni successive. In pratica i blocchi sono formati dalle unità che coincidono o nel codice familiare o nella data di nascita. □

Commento 5.2 *La procedura di bloccaggio è stata introdotta essenzialmente per risolvere problemi di tipo computazionale, come si è già detto. Ma queste procedure sono importanti anche per le successive applicazioni di metodi statistici. In proposito si veda il commento 6.2.* □

5.2 L'eliminazione dei risultati incongruenti, Jaro (1989)

La regola di decisione di Fellegi e Sunter (capitolo 4) può portare a risultati incongruenti quando devono essere valide le condizioni (2.6) e (2.7), ovvero quando ogni unità di A può essere abbinata al più ad un'unità di B e viceversa. Infatti la regola di Fellegi e Sunter dichiara come match tutte le coppie con peso (4.2) superiore ad una soglia fissata τ_μ. Fra le coppie che sono state dichiarate match può accadere, ad esempio, che si ritrovino contemporaneamente le coppie (a, b) e (a, b'), cioè che l'unità a della base dati A risulti abbinata sia a b che a b', contraddicendo il vincolo (2.6).

Se anche si adottasse la strategia di abbinare ogni unità $a \in \mathcal{A}$ con la corrispondente unità $b \in \mathcal{B}$ il cui peso è massimo e superiore a τ_λ, la soluzione trovata rischia di legare la stessa unità $b \in \mathcal{B}$ a più unità $a \in \mathcal{A}$.

Per ovviare a questa patologia della procedura di record linkage, Jaro propone un metodo di assegnazione basato su un algoritmo di ricerca operativa. Si considerino per ogni coppia (a, b) il peso $t(\mathbf{y}_{ab})$ e il coefficiente $c_{a,b}$ definiti nel capitolo 2. Si consideri quindi una matrice dei pesi avente come righe le unità $a \in \mathcal{A}$ e come colonne le unità del file $b \in \mathcal{B}$, e composta dai pesi $t(\mathbf{y}_{ab})$. Se la matrice non fosse quadrata (cioè se $\nu_A \neq \nu_B$), si aggiungono tante righe o colonne quante sono necessarie per renderla tale, e si assegnano valori negativi molto bassi agli elementi aggiunti della matrice per impedire che queste coppie fittizie siano scelte. Sia $\nu_{A\times B} = \max\{\nu_A, \nu_B\}$ la dimensione della matrice quadrata ora formata. Il problema della scelta delle coppie da abbinare si riduce ad un problema di programmazione lineare: assegnare a ogni

coppia (a, b) il valore $\hat{c}_{a,b}$ ("stima" di $c_{a,b}$) che, in base al vincolo (2.5), può assumere 0 (non match) o 1 (match) in modo tale che la funzione

$$\Psi = \sum_{a=1}^{\nu_{A\times B}} \sum_{b=1}^{\nu_{A\times B}} \hat{c}_{a,b} t(\mathbf{y}_{ab})$$

sia massima, sotto i vincoli:

$$\sum_{a=1}^{\nu_{A\times B}} c_{a,b} = 1, \qquad b = 1, ..., \nu_{A\times B},$$

$$\sum_{b=1}^{\nu_{A\times B}} c_{a,b} = 1, \qquad a = 1, ..., \nu_{A\times B}.$$

Questo problema è stato ampiamente dibattuto in ricerca operativa e va sotto il nome di "problema del trasporto". Soluzioni a tale problema sono disponibili, ad esempio, in Lawler (1976). Una volta individuate le $\nu_{A\times B}$ coppie candidate ad essere abbinate, le si ordinano secondo il loro peso $t(\mathbf{y}_{ab})$ e le coppie che vengono dichiarate abbinate sono quelle con peso superiore alla soglia prefissata. Queste coppie soddisfano contemporaneamente due requisiti:

- non vengono abbinate due coppie aventi un'unità in comune;
- le coppie che vengono abbinate massimizzano una funzione obiettivo interessante: le coppie candidate infatti sono quelle ammissibili di massimo peso complessivo Ψ.

Commento 5.3 *La procedura di selezione delle coppie attraverso il metodo di Jaro ora esposto disturba le proprietà della procedura di Fellegi e Sunter, ad esempio nelle probabilità di errore. In generale, l'influenza di questa procedura sui metodi di record linkage è da verificare.* □

Commento 5.4 *Winkler (1994) sottolinea che l'algoritmo di assegnazione definito in questo paragrafo, usato da Jaro (1989) per abbinare il censimento del 1980 nella zona di Tampa, Florida, con la* Post Enumeration Survey*, è stato originariamente definito da Burkard e Derigs (1980, pp. 1-11). Una modifica di questo algoritmo, che non necessita della trasformazione della matrice dei pesi in una matrice quadrata, è stata utilizzata da Winkler e Thibaudeau (1991).* □

Capitolo 6

La stima delle distribuzioni di confronto

Il problema di gran lunga più dibattuto nella letteratura sul record linkage riguarda il calcolo delle distribuzioni $m(\mathbf{y})$ e $u(\mathbf{y})$, $\mathbf{y} \in \mathcal{D}$. Le soluzioni a disposizione sono le più diverse e, a grandi linee, possono raggrupparsi in:

- metodi che fanno riferimento esclusivamente ai dati che si stanno studiando, ovvero alle ν_A osservazioni in A e ν_B osservazioni in B;
- metodi che fanno riferimento a informazioni esterne;

e ancora:

- metodi che fanno riferimento a modelli statistici semplici;
- metodi che fanno riferimento a modelli statistici complessi.

Con la prima classificazione si dividono i metodi di calcolo delle due distribuzioni a seconda dell'insieme di informazioni (dati, osservazioni, ...) di cui si è in possesso. La seconda classificazione, invece, fa riferimento al modo in cui vengono "relazionate" le informazioni ricavabili dalle diverse variabili chiave. Come modello statistico semplice consideriamo il caso in cui i vettori

$$\mathbf{y}_{ab} = (y_{ab}^1, ..., y_{ab}^k)$$

che esprimono il confronto delle k variabili chiave nelle due occasioni sono generati da k variabili aleatorie Y^h fra loro indipendenti. Quindi la (3.3):

$$m(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|c_{a,a} = 1), \qquad \mathbf{y} \in \mathcal{D}$$

diventa

$$m(\mathbf{y}) = \prod_{h=1}^{k} P(Y^h = y^h|c_{a,a} = 1) = \prod_{h=1}^{k} m_h(y^h), \qquad \mathbf{y} \in \mathcal{D} \tag{6.1}$$

e allo stesso modo la (3.4)

$$u(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|c_{a,b} = 0), \qquad \mathbf{y} \in \mathcal{D}$$

diventa:

$$u(\mathbf{y}) = \prod_{h=1}^{k} P(Y^h = y^h|c_{a,b} = 0) = \prod_{h=1}^{k} u_h(y^h), \qquad \mathbf{y} \in \mathcal{D}. \tag{6.2}$$

I modelli statistici complessi sono invece modelli che tengono conto della possibilità di dipendenze statistiche fra i diversi confronti Y^h, $h = 1, ..., k$. Naturalmente i secondi sono più difficili da studiare rispetto ai primi, ma i primi più difficilmente si verificano (Thibaudeau, 1993, Winkler, 1995).

In questo capitolo affronteremo il problema della stima di $m(\mathbf{y})$ e $u(\mathbf{y})$ attraverso diversi approcci, via via più complessi. In particolare, i diversi metodi di stima vengono legati alle ipotesi sui dati che si hanno a disposizione (in proposito si veda il commento 3.1). Abbiamo infatti affermato che la variabile $\mathbf{Y}$ può essere spiegata attraverso un "modello di superpopolazione" o, condizionatamente ai valori veri e agli errori, attraverso uno schema tipo "popolazione finita".

Quando si ragiona con un approccio tipo popolazione finita, le distribuzioni $m(.)$ e $u(.)$ sono la distribuzione di frequenze di $\mathbf{Y}$ sulle coppie formate dalle unità delle due basi dati $\mathcal{A}$ e $\mathcal{B}$ (in proposito si veda il commento 3.2). Queste due distribuzioni possono essere stimate o attraverso il controllo manuale di un "campione" opportuno delle $\nu_A \times \nu_B$ coppie, oppure da informazioni esterne[1]. A questi casi viene dedicato il paragrafo 6.1.

Quando si suppone che le $\nu_A \times \nu_B$ coppie siano generate da un meccanismo di superpopolazione, il modello che definisce la fase di generazione dei valori veri e degli errori (ovvero delle variabili $\mathbf{Y}$, come descritto nel commento 3.1) viene affiancato da un ulteriore modello, che descrive la fase di generazione delle coppie e del loro status $C_{a,b}$ ($C_{a,b} = 1$ indica un match, mentre $C_{a,b} = 0$ indica un non-match). I meccanismi di generazione dei dati precedenti definiscono una distribuzione di probabilità congiunta delle variabili $\mathbf{Y}$ e $C_{a,b}$:

$$P(\mathbf{Y} = \mathbf{y}, C = c) = P(C = c)P(\mathbf{Y} = \mathbf{y}|C = c), \qquad \mathbf{y} \in \mathcal{D},\ c = 0, 1 \tag{6.3}$$

dove $P(\mathbf{Y} = \mathbf{y}|C = c)$ rappresenta una delle distribuzioni $m(.)$ o $u(.)$ (a seconda del condizionamento) finora usate. La (6.3) costituisce il nucleo della verosimiglianza che consente di definire gli stimatori dei parametri di interesse (paragrafo 6.2). Sono stati ipotizzati diversi modelli per le distribuzioni $m(.)$ e $u(.)$. Questi sono determinati in base alla funzione di confronto $\mathbf{y}$ usata (il confronto (2.10)viene discusso nei paragrafi 6.3 e 6.4, mentre confronti del tipo (2.12) e (2.13) vengono analizzati nel paragrafo 6.5) e al tipo di dipendenza fra le variabili chiave.

6.1 L'uso delle informazioni a disposizione

Come detto precedentemente, chi sta conducendo un record linkage deve sfruttare al massimo le informazioni di cui è in possesso per poter calcolare le distribuzioni $m(\mathbf{y})$ e $u(\mathbf{y})$. Queste distribuzioni, come dichiarato nel commento 3.1, possono essere considerate le distribuzioni di frequenza dei confronti dei match e dei non-match su $N_A \times N_B$ coppie di due popolazioni da dove sono estratte le $\nu_A \times \nu_B$ coppie da abbinare (al limite $N_A = \nu_A$ e $N_B = \nu_B$). Di seguito si elencano i metodi più frequentemente usati per ricavare le distribuzioni $m(.)$ e $u(.)$. Si tenga presente che questi sono metodi "ad hoc", ovvero nulla è noto sulle loro proprietà. I primi due metodi sfruttano esclusivamente i dati $\mathcal{A}$ e $\mathcal{B}$ a disposizione per il record linkage.

[1] A prima vista potrebbe sembrare che un approccio di stima attraverso campionamento da popolazioni finite potrebbe essere preso in considerazione quando si può ipotizzare che le due basi dati A e B siano due campioni di popolazioni più grandi (come nel commento 3.3). Ma spesso questa risulta essere un'ipotesi di comodo, formulata per la prima volta da Fellegi e Sunter (1969). Infatti anche se fosse ipotizzabile che le basi dati A e B sono campioni di popolazioni più grandi, quasi mai sono campioni casuali (ad esempio ciò non è ipotizzabile per gli archivi)

1. Un primo metodo per stimare le due distribuzioni consiste nel considerare un campione (in genere non casuale[2]) di coppie di record da $\mathcal{A} \times \mathcal{B}$. Ogni coppia del campione viene controllata manualmente da impiegati, in modo da accertarne lo status. Il risultato dell'abbinamento delle coppie di record su questo campione, e quindi della divisione delle coppie di record nel campione in quelle appartenenti a $\mathcal{M}$ e $\mathcal{U}$, viene usato per "stimare" le distribuzioni $m(.)$ e $u(.)$. Queste stime possono essere definite dal semplice calcolo delle frequenze osservate dei confronti $\mathbf{y}$ per le coppie che sono match e le coppie che sono non-match, a metodi di calcolo più raffinati (come quelli indicati nel paragrafo 6.6).

Esempio 6.1 *Questo metodo è descritto, ad esempio, in Copas e Hilton (1990) (come detto nel commento 3.1, Copas e Hilton fanno riferimento al meccanismo generatore di errori, ma il metodo può essere facilmente esteso anche agli altri casi). Il loro obiettivo è l'integrazione del dataset degli stranieri entrati nel Regno Unito, per i quali alla frontiera si rilevano nome, cognome, sesso, data di nascita e nazionalità tramite la "landing card", con l'elenco degli stranieri usciti dal Regno Unito, per i quali vengono rilevate le stesse caratteristiche precedenti tramite la "embarkation card". La fase di abbinamento attraverso record linkage probabilistico è stata svolta usando le distribuzioni $m(\mathbf{y})$ e $u(\mathbf{y})$ desumibili dal confronto manuale degli stessi elenchi relativo a un periodo di sole due settimane. Per questi dati vengono presentati diversi modelli utili a rappresentare l'errore di misura, come si discute nel paragrafo 6.6.* □

2. Per quanto riguarda il calcolo della distribuzione $u(\mathbf{y})$, si può considerare il ragionamento proposto da Jaro (1989). Il numero di coppie totale $\nu = \nu_A \times \nu_B$ dei confronti è composto prevalentemente da coppie in $\mathcal{U}$ piuttosto che in $\mathcal{M}$. Ad esempio se le due liste A e B sono composte rispettivamente da 5 e 10 unità, si ha che $\nu = 50$, mentre le coppie in $\mathcal{M}$ sono al massimo 5 e le coppie in $\mathcal{U}$ sono minimo 45. Questa differenza nella numerosità di coppie in $\mathcal{M}$ e $\mathcal{U}$ diventa sempre più marcato al crescere delle dimensioni delle due liste A e B. Infatti, indicando con $[\mathcal{S}]$ la numerosità (cardinalità) di un insieme $\mathcal{S}$, si ha che il numero massimo di possibili match sul numero di coppie in $\mathcal{A} \times \mathcal{B}$ è:

$$\frac{[\mathcal{M}]}{[\mathcal{A} \times \mathcal{B}]} = \frac{\min\{\nu_A, \nu_B\}}{\nu_A \nu_B} = \frac{1}{\max\{\nu_A, \nu_B\}}.$$

Questa frazione decresce al crescere della dimensione massima di $\mathcal{A}$ e $\mathcal{B}$. Per questo motivo Jaro propone di stimare le probabilità $u(\mathbf{y})$ ignorando la presenza di coppie $(a, b) \in \mathcal{M}$. La frequenza relativa di $\mathbf{y}$ sulle ν coppie in $\mathcal{A} \times \mathcal{B}$ può essere considerata una buona approssimazione della vera distribuzione di $\mathbf{y}$ in $\mathcal{U}$. Se $\mathcal{U}$ è grande e non gestibile tramite programmi informatici, un campione di coppie fra le ν estratto in modo opportuno è sufficiente per ottenere una buona stima di $u(\mathbf{y})$.

La rarità delle coppie $(a, b) \in \mathcal{M}$ impone la scelta di metodi alternativi per ottenere delle stime attendibili dei parametri della distribuzione $m(.)$. Jaro suggerisce l'approccio delineato nel punto 1. Se per qualche motivo il metodo in 1. non fosse applicabile (ad esempio per l'indisponibilità di personale esperto per il controllo manuale delle coppie di record), Jaro propone un approccio basato sull'algoritmo EM (E*xpectation*-M*aximization*, Dempster *et al*., 1977) in grado di individuare le stime di massima verosimiglianza dei parametri incogniti in presenza di dati mancanti (in proposito si veda il paragrafo 6.3.2).

[2] Si vedrà nel punto 2. che le coppie in $\mathcal{M}$ sono estremamente rare fra le $\nu_A \times \nu_B$ coppie che si stanno confrontando, e non sempre è agevole definire un piano di campionamento che consenta la stima della distribuzione $m(.)$.

Tabella 6.1 - Distribuzioni dei confronti relativi alle variabili chiave "Cognome", "Nome" e "Anno di nascita" quando il confronto è definito da (2.10) suggeriti da Newcombe (1988) per il Regno Unito.

Variabile chiave	*Y*	*m*(*y*)	*u*(*y*)
Cognome	*y*=1	0,965	0,001
	y=0	0,035	0,999
Nome	*y*=1	0,790	0,009
	y=0	0,210	0,991
Anno di nascita	*y*=1	0,773	0,011
	y=0	0,227	0,989

Per la stima delle distribuzioni $m(.)$ e $u(.)$ si può fare riferimento anche ad informazioni esterne. Di seguito, descriviamo tre diverse possibili fonti esterne.

3. Una fonte di informazioni spesso presa in considerazione è costituita da esperienze affidabili di integrazione (fatte attraverso controllo manuale dei record, su piccoli campioni o dati provenienti da aree "tipo" delimitate) che hanno coinvolto le variabili chiave di interesse. Su queste fonti vengono calcolate le frequenze relative dei confronti $\mathbf{y}$ per i match e i non-match: le frequenze vengono poi usate per applicare la procedura Fellegi-Sunter. Naturalmente questo metodo si basa su un'ipotesi molto forte: i dataset da cui stiamo calcolando $m(\mathbf{y})$ e $u(\mathbf{y})$ devono essere della stessa qualità e devono rappresentare la stessa popolazione. In particolare esiste il rischio della così detta "site to site variability" (Winkler, 1985a, 1985b, Arellano, 1992, e Belin, 1993), ovvero $m(.)$ e $u(.)$ possono essere molto diversi da luogo a luogo e compromettere il risultato del record linkage.

4. Alcune variabili possono essere caratterizzate da distribuzioni $m(\mathbf{y})$ e $u(\mathbf{y})$ desumibili dal loro andamento nella popolazione di riferimento. Ad esempio, si supponga che $\mathbf{y}$ sia definito dalla (2.10) e si consideri il confronto Y^h definito dalla variabile chiave X^h="mese di nascita". Poiché X^h può assumere solo 12 modalità, se le nascite sono equidistribuite durante l'arco dell'anno e non ci sono errori o mancate risposte, la probabilità che una coppia in $\mathcal{U}$ contenga lo stesso mese di nascita è 1/12 ($u_h(1) = 1/12$), mentre in $\mathcal{M}$ è 1 ($m_h(1) = 1$). Tramite piccole modifiche a queste distribuzioni per tener conto della presenza di errori o mancate risposte si ottiene il risultato di interesse.

5. A volte si hanno a disposizione pubblicazioni che descrivono le uguaglianze o discordanze su alcune variabili, come in Newcombe (1988). Newcombe definisce le distribuzioni $m(.)$ e $u(.)$ per alcune variabili chiave molto usate per abbinare dati relativi a individui, come il cognome, il nome e l'anno di nascita. Queste distribuzioni sono elencate nella tabella 6.1. Un esempio di questo tipo di approccio è disponibile anche in Rogot, Sorlie e Johnson (1986).

6.2 L'uso dei modelli statistici

Un'alternativa ai casi precedenti, necessaria quando non si dispone di personale specializzato per il controllo manuale dei dati, né di informazioni esterne, consiste nel supporre che sulle $\nu_A \times \nu_B$ coppie in $\mathcal{A} \times \mathcal{B}$ la variabile aleatoria $\mathbf{Y}$ (si veda il capitolo 3) ha generato i valori del confronto $\mathbf{y}_{ab}$, $a = 1, ..., \nu_A$, $b = 1, ..., \nu_B$, in modo indipendente ed identicamente distribuito (in pratica si adotta un approccio di superpopolazione come descritto nel commento 3.1; si rimanda al paragrafo 6.2.1 per un approfondimento di questa impostazione). Inoltre si ipotizza che anche la variabile indicatrice $C_{a,b}$, che descrive l'appartenenza della coppia (a, b) a $\mathcal{M}$ o $\mathcal{U}$, sia aleatoria (per la definizione dei valori che la variabile $C_{a,b}$ assume si veda il paragrafo 2.2). In pratica, il modello di superpopolazione si arricchisce di un nuovo "meccanismo generatore di dati", che in questo caso sono le coppie. Questo meccanismo generatore di coppie, C, genera in modo indipendente e identicamente distribuito una coppia che è un match con probabilità p e una coppia che è un non-match con probabilità $1 - p$. Condizionatamente al valore di C, alla coppia viene assegnato un valore della variabile $\mathbf{Y}$. Di conseguenza una coppia (a, b) ha associata la coppia di valori $(\mathbf{y}_{ab}, c_{a,b})$ secondo la distribuzione congiunta:

$$
\begin{aligned}
&P\Big(\mathbf{Y}_{ab} = \mathbf{y}, C_{a,b} = c \Big| p, \{m(.)\}, \{u(.)\}\Big) = \\
&= P\Big(C_{a,b} = c \Big| p\Big) P\Big(\mathbf{Y}_{ab} = \mathbf{y} \Big| C_{a,b} = c, \{m(.)\}, \{u(.)\}\Big) = \\
&= \Big(p\ m(\mathbf{y})\Big)^c \Big((1-p)\ u(\mathbf{y})\Big)^{1-c}, \qquad\qquad \mathbf{y} \in \mathcal{D},\ c \in \{0, 1\}.
\end{aligned}
$$

La verosimiglianza sulle $\nu_A \times \nu_B$ coppie è quindi definita da:

$$
L\Big(p, \{m(.)\}, \{u(.)\} \Big| \{\mathbf{y}_{ab}\}, \{c_{a,b}\}\Big) = \prod_{(a,b)} \Big(p\ m(\mathbf{y}_{ab})\Big)^{c_{a,b}} \Big((1-p)\ u(\mathbf{y}_{ab})\Big)^{1-c_{a,b}}. \qquad (6.4)
$$

Nel caso in cui i confronti delle variabili chiave siano indipendenti, cioè sono valide le (6.1) e (6.2), la (6.4) si trasforma in:

$$
\begin{aligned}
&L\Big(p, \{m(.)\}, \{u(.)\} \Big| \{\mathbf{y}_{ab}\}, \{c_{a,b}\}\Big) = \\
&= \prod_{(a,b)} \left(p \prod_{h=1}^{k} m_h(y_{ab}^h)\right)^{c_{a,b}} \left((1-p) \prod_{h=1}^{k} u_h(y_{ab}^h)\right)^{1-c_{a,b}}. \qquad (6.5)
\end{aligned}
$$

Commento 6.1 *In realtà, chi si è occupato di record linkage non ha mai assunto esplicitamente una struttura dei dati così complessa, governata da "meccanismi generatori di valori" (siano essi gli indicatori di match $c_{a,b}$ o le osservazioni X^h). Ma questa ipotesi è implicita quando si vogliono usare le verosimiglianze (6.4) o (6.5). Di conseguenza, l'oggetto della procedura di stima indicata nei paragrafi successivi è la distribuzione di probabilità dei meccanismi generatori dei dati. Queste stime verranno quindi utilizzate per implementare le regole di decisione discusse nel capitolo 4. Si rimanda al commento 4.6 per fare in modo che, anche con un approccio tipo superpopolazione, la procedura di decisione possa essere interpretata con riferimento alle $\nu_A \times \nu_B$ coppie delle basi dati che si stanno confrontando.* □

Commento 6.2 *Le procedure di stima per i parametri in (6.4) e (6.5) sono efficaci solo quando i match sono un numero sufficientemente grande rispetto ai non-match. Per questo motivo le procedure di bloccaggio (capitolo 5) risultano estremamente utili in quanto rendono i match meno rari nell'insieme delle coppie che si stanno confrontando. In particolare, viene suggerito spesso di creare dei blocchi che contengono almeno il 5% di match.* □

6.2.1 Un commento ai modelli proposti

L'ipotesi cruciale per le verosimiglianze indicate precedentemente è che le $\nu_A \times \nu_B$ variabili $\mathbf{Y}_{ab}$ siano fra loro indipendenti, in modo da giustificare l'uso della produttoria rispetto alle coppie in (6.4) e (6.5). L'ipotesi di indipendenza è falsa[3] se si fa riferimento a confronti del tipo (2.10) (e in genere a qualsiasi tipo di confronto). Verificare questo è semplice. Si considerino 4 coppie, derivanti dal confronto fra le unità a e a' in $\mathcal{A}$ e b e b' in $\mathcal{B}$. Si consideri una sola variabile di confronto, con confronto Y. Se

$$Y_{ab} = 1,\ Y_{ab'} = 1,\ Y_{a'b} = 1$$

allora il confronto per la coppia (a', b') non può che essere: $Y_{a'b'} = 1$. Quindi la conoscenza di quanto accade nelle prime tre coppie, vincola a un risultato deterministico il valore assunto per la coppia (a', b').

Lo stesso ragionamento si può fare anche per le variabili $C_{a,b}$. Infatti quando sono validi i vincoli (2.6), se è noto che $C_{a,b} = 1$, le restanti variabili $C_{a,b'}$, $b' \neq b$, devono assumere il valore 0 con probabilità 1, rendendo l'ipotesi di indipendenza implausibile.

Non è ancora chiara l'influenza del mancato rispetto dell'ipotesi di indipendenza per le variabili $\mathbf{Y}_{ab}$ e $C_{a,b}$ sulle procedure di record linkage. Per adesso ipotizziamo valide le verosimiglianze (6.4) e (6.5). Si sottolinea che questi modelli di indipendenza risultano cruciali per i metodi discussi nei paragrafi 6.3, 6.4, 7.2, 7.3 e 8.2.1.

6.3 Le stime nel caso più semplice

Supponiamo che i confronti $\mathbf{y}$ siano definiti dalla (2.10), e quindi che $\mathbf{y}$ sia una variabile multinomiale di dimensione k. In pratica, se sono vere le (6.1) e (6.2), la (6.5) si semplifica ulteriormente ponendo $m_h = m_h(1)$ e $u_h = u_h(1)$ e quindi:

$$L\Big(p, \{m(.)\}, \{u(.)\} \Big| \mathbf{y}_{ab}, c_{a,b}, (a,b) \in \mathcal{A} \times \mathcal{B}\Big) =$$
$$= \prod_{(a,b)} \left(p \prod_{h=1}^{k} m_h^{y_{ab}^h} (1 - m_h)^{1 - y_{ab}^h} \right)^{c_{a,b}} \left((1-p) \prod_{h=1}^{k} u_h^{y_{ab}^h} (1 - u_h)^{1 - y_{ab}^h} \right)^{1 - c_{a,b}} . \quad (6.6)$$

Supponiamo inoltre di non possedere informazioni esterne, oltre a quelle che vengono fornite dai dati a disposizione. Se i confronti Y^h sono fra loro indipendenti (e quindi se è vero il modello (6.6)), Fellegi e Sunter suggeriscono il metodo dei momenti come metodo di stima delle distribuzioni $m(.)$, $u(.)$ e del numero di coppie che sono match N (ovvero la cardinalità di $\mathcal{M}$).

[3] Questa osservazione è descritta in un piccolo paragrafo in Kelley (1984). Da allora, nessuno si è preoccupato di questo problema. In proposito si veda Fortini *et al*. (2002).

Questo metodo si basa sulla soluzione di un sistema di equazioni nel caso in cui le variabili di confronto siano $k = 3$ (paragrafo 6.3.1). Se $k > 3$ la soluzione del sistema di equazioni si deve ricercare attraverso opportuni metodi numerici (Winkler, 1995).

Supponendo ancora valido il modello (6.6) di indipendenza fra le Y^h, Jaro (1989) suggerisce un metodo basato sulla stima di massima verosimiglianza dei parametri (paragrafo 6.3.2) usando l'algoritmo EM (appendice B). Jaro afferma che l'EM fornisce soluzioni migliori rispetto a quelle ottenibili dal sistema proposto da Fellegi e Sunter in quanto quest'ultimo è numericamente instabile in molte applicazioni. In particolare, Jaro afferma che le soluzioni del sistema di Fellegi e Sunter, quando $k > 3$, sono estremamente sensibili ai valori iniziali e, a volte, è necessario introdurre delle "penalty functions" per mantenere le stime delle probabilità fra zero e uno. Al contrario, Thibaudeau (1989) e Winkler (1989b, 1992) verificano in molte situazioni che il valore verso cui converge l'algoritmo EM non viene influenzato dai valori iniziali dei parametri.

In questo paragrafo si illustrano sia il metodo proposto da Fellegi e Sunter sia quello proposto da Jaro.

6.3.1 Il sistema di equazioni di Fellegi e Sunter (1969)

Fellegi e Sunter sono stati i primi a notare che, nel caso più semplice di indipendenza fra i confronti delle variabili chiave e di definizione dei confronti dati dalla (2.10), si possono ottenere agevolmente delle stime dei parametri incogniti del modello (6.5).

Queste stime si fondano sulla definizione di un certo numero di equazioni ($2k+1$ equazioni se k sono le variabili di confronto) che descrivono le "frequenze attese" di alcuni eventi. Le frequenze che si studiano sono le seguenti:

- Λ_r: frequenza delle coppie sul totale delle $\nu_A \times \nu_B$ coppie con $Y_{ab}^h = 1$ per $h \neq r$ e Y_{ab}^r qualsiasi, $r = 1, \ldots, k$;
- Φ_r: frequenza delle coppie sul totale delle coppie con $Y_{ab}^r = 1$ e Y_{ab}^h qualsiasi per $h \neq r$, $r = 1, \ldots, k$;
- Θ: frequenza delle coppie sul totale delle coppie con $Y_{ab}^h = 1$ per $h = 1, \ldots, k$.

In pratica Θ contiene solo le coppie (a, b) che sono identiche nelle k variabili chiave, Φ_r le coppie (a, b) che sono identiche nella r-esima variabile chiave e Λ_r le coppie (a, b) che sono identiche nelle restanti $k-1$ variabili chiave. Facendo variare r fra 1 e k si ottengono effettivamente $2k+1$ frequenze. Si sottolinea fin da subito che le $2k+1$ frequenze sono facilmente calcolabili dalle $\nu_A \times \nu_B$ coppie osservate.

Si ragioni condizionatamente ai risultati del meccanismo generatore di coppie. In pratica lo status $c_{a,b}$ delle coppie (a, b) non viene considerato aleatorio, anche se rimane incognito. Supponiamo che ci siano N match, con N anch'esso incognito.

I valori attesi rispetto a $\mathbf{Y}$ delle $2k+1$ frequenze precedenti definiscono $2k+1$ equazioni:

$$\nu_A \nu_B E(\Lambda_r) = N \prod_{h \neq r} m_h + (\nu_A \nu_B - N) \prod_{h \neq r} u_h, \qquad r = 1, \ldots, k, \tag{6.7}$$

$$\nu_A \nu_B E(\Phi_r) = N m_r + (\nu_A \nu_B - N) u_r, \qquad h = 1, \ldots, k, \tag{6.8}$$

$$\nu_A \nu_B E(\Theta) = N \prod_{h=1}^{k} m_h + (\nu_A \nu_B - N) \prod_{h=1}^{k} u_h, \tag{6.9}$$

dove m_h, u_h, $h = 1, \ldots, k$ e N (numero di *match*) sono $2k + 1$ parametri incogniti. La prima equazione vuol dire che il numero atteso di vettori di confronto composto esclusivamente da 1 in tutte le posizioni $r \neq h$, mentre per l'h-esima posizione ci può essere sia uno 0 che un 1, è dato dalla somma del numero atteso di coppie rispettivamente in $\mathcal{M}$ e in $\mathcal{U}$ che presentano quel *pattern* di confronto. Le altre equazioni hanno un significato analogo. I valori attesi di Λ_r, Φ_r e Θ dipendono dalle distribuzioni $m(.)$ e $u(.)$ incognite, e quindi non sono conosciuti. Uno dei classici metodi di stima di parametri incogniti è il metodo dei momenti. In pratica si sostituiscono i valori incogniti dei parametri (nel nostro caso le medie di Λ_r, Φ_r e Θ) con i valori osservati nel campione, ovvero con le frequenze osservate Λ_r, Φ_r e Θ. Una volta effettuata la sostituzione, se $k > 3$ il sistema di equazioni deve essere risolto attraverso opportuni metodi numerici; se invece $k = 3$, Fellegi e Sunter forniscono le soluzioni (pag. 1208-1210 in Fellegi e Sunter, 1969).

6.3.2 Le stime di massima verosimiglianza tramite EM, Jaro (1989)

Nel paragrafo (6.1), punto 2., si è accennato al metodo di stima che Jaro propone per la distribuzione $u(.)$. Per la distribuzione $m(.)$ viene proposta invece una tecnica più raffinata, che ha come obiettivo la stima di massima verosimiglianza dei parametri incogniti in presenza di dati mancanti: l'algoritmo EM (appendice B).

Jaro suppone che sia valida la verosimiglianza (6.6):

$$L\Big(p, \{m(.)\}, \{u(.)\}\Big|\mathbf{y}_{ab}, c_{a,b}, (a,b) \in \mathcal{A} \times \mathcal{B}\Big) =$$
$$= \prod_{(a,b)} \left(p \prod_{h=1}^{k} m_h^{y_{ab}^h}(1 - m_h)^{1-y_{ab}^h} \right)^{c_{a,b}} \left((1-p) \prod_{h=1}^{k} u_h^{y_{ab}^h}(1 - u_h)^{1-y_{ab}^h} \right)^{1-c_{a,b}},$$

caratterizzata da confronti fra variabili chiave indipendenti e del tipo (2.10). Inoltre ipotizza che si passi attraverso una fase preliminare di bloccaggio (come quella descritta nel paragrafo 5.1), per ridurre la complessità computazionale.

I passi dell'algoritmo EM, descritti anche nell'appendice B, sono quindi i seguenti.

1. Si costruisca la distribuzione di frequenza dei vettori di confronto $\mathbf{y}_{ab}$ su tutti i blocchi. Sia $f_{\mathbf{y}}$ la frequenza assoluta del vettore $\mathbf{y}$, per ogni $\mathbf{y} \in \mathcal{D}$.

2. Si assegnino valori iniziali $\hat{p}^{(0)}$, $\hat{m}_h^{(0)}$ e $\hat{u}_h^{(0)}$ ai parametri incogniti p, m_h e u_h, $h = 1, ..., k$. Jaro non suggerisce valori particolari, ma si raccomanda che $\hat{m}_h^{(0)} > \hat{u}_h^{(0)}$, $h = 1, ..., k$ (se ciò non si verificasse, si assumerebbero come valori iniziali probabilità difficilmente giustificabili: in particolare la probabilità che si verifichi un'uguaglianza, $Y = 1$, sarebbe più alta per i non-match che per i match).

3. All'iterazione $i + 1$, i valori assegnati ai dati mancanti $c_{a,b}$ dal passo E non dipendono dalla coppia (a, b) esaminata, ma solo dal vettore di confronto $\mathbf{y}_{ab}$ osservato. I 2^k valori diversi di $\hat{c}_{a,b}^{(i+1)}(\mathbf{y}_{ab})$ sono stimati da:

$$\hat{c}_{a,b}^{(i+1)}(\mathbf{y}_{ab}) =$$
$$= \frac{\hat{p}^{(i)} \prod_{h=1}^{k} (\hat{m}_h^{(i)})^{y_{ab}^h}(1 - \hat{m}_h^{(i)})^{1-y_{ab}^h}}{\hat{p}^{(i)} \prod_{h=1}^{k} (\hat{m}_h^{(i)})^{y_{ab}^h}(1 - \hat{m}_h^{(i)})^{1-y_{ab}^h} + (1 - \hat{p}^{(i)}) \prod_{h=1}^{k} (\hat{u}_h^{(i)})^{y_{ab}^h}(1 - \hat{u}_h^{(i)})^{1-y_{ab}^h}}.$$

4. Sempre all'iterazione $i+1$, il passo M consiste nella stima di massima verosimiglianza dei parametri incogniti p, m_h e u_h, $h = 1, ..., k$, ottenuti attraverso il metodo di massima verosimiglianza, avendo sostituito i dati mancanti $c_{a,b}$ con le stime del passo precedente: $\hat{c}_{a,b}^{(i+1)}(\mathbf{y}_{ab})$. Usando le frequenze calcolate al punto 1, le stime sono:

$$\hat{m}_h^{(i+1)} = \frac{\sum_{\mathbf{y}\in\mathcal{D}} \hat{c}^{(i+1)}(\mathbf{y}) y_h f_{\mathbf{y}}}{\sum_{\mathbf{y}\in\mathcal{D}} \hat{c}^{(i+1)}(\mathbf{y}) f_{\mathbf{y}}}$$

$$\hat{u}_h^{(i+1)} = \frac{\sum_{\mathbf{y}\in\mathcal{D}} (1 - \hat{c}^{(i+1)}(\mathbf{y})) y_h f_{\mathbf{y}}}{\sum_{\mathbf{y}\in\mathcal{D}} (1 - \hat{c}^{(i+1)}(\mathbf{y})) f_{\mathbf{y}}}$$

$$\hat{p}^{(i+1)} = \frac{\sum_{\mathbf{y}\in\mathcal{D}} \hat{c}^{(i+1)}(\mathbf{y}) f_{\mathbf{y}}}{\sum_{\mathbf{y}\in\mathcal{D}} f_{\mathbf{y}}}.$$

5. I punti 3 e 4 vengono iterati finché i parametri stimati non risultano stabili.

Jaro consiglia di assumere come output esclusivamente le stime dei parametri m_h e p. Questo perché, una volta che le coppie hanno subito la fase di bloccaggio (capitolo 5), i valori di u_h subiscono una forte distorsione. Quindi per le u_h si raccomanda l'uso delle stime descritte al punto 2. del paragrafo 6.1.

Commento 6.3 *Winkler (1989) considera anche il caso di stima dei parametri p m_h, u_h, $h = 1, ..., k$, del modello (6.6) attraverso il metodo EM quando i parametri soddisfano dei vincoli convessi. Questo aspetto si approfondisce nel paragrafo 6.4.2.* □

6.3.3 La stima del peso *t*(y)* tramite l'algoritmo EM

Nel paragrafo 4.2 si è visto che possono essere adottati dei pesi diversi da quelli suggeriti da Fellegi e Sunter, senza per questo modificare la struttura della regola di decisione. Una di queste trasformazioni consiste nell'associare a ogni vettore di confronto $\mathbf{y} \in \mathcal{D}$ il peso:

$$t^*(\mathbf{y}) = \frac{t(\mathbf{y})}{t(\mathbf{y}) + \frac{1-p}{p}}.$$

Per quanto visto nella formula (4.13), il peso $t^*(\mathbf{y})$ corrisponde esattamente al valor medio di $(C|\mathbf{Y} = \mathbf{y})$, ovvero al valore usato per sostituire i valori incogniti $c_{a,b}$ nel passo E dell'algoritmo EM usato nel paragrafo 6.3.2. Torelli (1998, 1999) suggerisce quindi di stimare i pesi $t^*(\mathbf{y})$ con i valori $\hat{c}_{a,b}(\mathbf{y})$ ottenuti nell'ultima iterazione dell'algoritmo EM. Se questi pesi vengono invece calcolati immediatamente dopo l'ultimo passo M dell'algoritmo EM (cioè quello che fornisce le stime di massima verosimiglianza dei parametri p, m_h e u_h, $h = 1, ..., k$) si ha che i pesi $t(\mathbf{y})$ stimati attraverso la procedura di Jaro e quelli ottenuti seguendo questo procedimento conducono alla stessa regola di decisione.

Commento 6.4 *Per semplicità espositiva, si è accennato a questo procedimento quando i confronti sono del tipo (2.10) e i confronti relativi alle singole variabili chiave sono indipendenti. Lo stesso procedimento può essere applicato quando i confronti o il modello utilizzati sono più complicati.* □

6.4 L'uso di modelli di dipendenza fra variabili di confronto

Fin da subito diversi autori si sono posti il problema della plausibilità dell'ipotesi di indipendenza fra i confronti Y^h delle variabili chiave. Questa ipotesi semplifica ciò che avviene in realtà. Dato che le variabili Y^h sono del tipo (2.10), i modelli loglineari appaiono i più utili per definire delle adeguate strutture di dipendenza. Fissato il modello loglineare di dipendenza fra le variabili Y^h, la stima dei parametri di questo modello viene condotta attraverso opportune modifiche del metodo EM (paragrafo 6.4.1). Strutture loglineari con diversi modelli di interazione per le due distribuzioni $m(.)$ e $u(.)$ sono state considerate, ad esempio, da Armostrong e Mayda (1993) e Thibaudeau (1993). Se non si hanno informazioni sul modello loglineare relativo alle due distribuzioni $m(.)$ e $u(.)$, la stima di queste due distribuzioni diventa complicata. In questo caso Winkler (1989b, 1992, 1995) propone di usare ancora il metodo EM per la stima dei parametri, adottando un modello loglineare non necessariamente giusto ma "sufficientemente generale" e vincoli convessi per i parametri delle distribuzioni (paragrafo 6.4.2).

Un'ulteriore generalizzazione riguarda il numero di gruppi in cui possono essere divise le $\nu_A \times \nu_B$ coppie di record. Finora si sono considerati sempre due gruppi: quello dei match ($\mathcal{M}$) e quello dei non-match ($\mathcal{U}$). Winkler (1992) afferma che spesso è appropriato considerare un numero di gruppi maggiore. Ad esempio si considerino due basi dati A e B i cui record si riferiscono a individui e dove una variabile chiave rileva l'"Indirizzo". In questo caso conviene dividere l'insieme $\mathcal{U}$ nell'insieme delle coppie non-match dove a e b risiedono allo stesso indirizzo e nell'insieme dei non-match dove a e b risiedono a indirizzi diversi. Infatti si possono ipotizzare due diverse distribuzioni $u(.)$ per questi due gruppi di non-match, dove le variabili chiave, condizionatamente all'uguaglianza nella variabile chiave "Indirizzo", sono maggiormente dipendenti fra loro rispetto a quando l'indirizzo è diverso. Ad esempio, per una coppia di individui diversi ma residenti allo stesso indirizzo è molto probabile che se c'è una coincidenza nella variabile "Cognome" allora si verificherà una coincidenza anche nel "Numero di telefono", mentre coincidenze di questo tipo sono indipendenti per coppie di unità che abitano in indirizzi diversi. L'estensione di questi metodi a un numero qualsiasi di gruppi e a un qualsiasi livello di interazione fra le variabili Y^h è stato descritto da Larsen e Rubin (2001) (rimandiamo la discussione su questo punto al paragrafo 7.2).

6.4.1 Le strutture di dipendenza fra i confronti: Thibaudeau (1993), Armstrong e Mayda (1993)

La verosimiglianza (6.6) può essere facilmente spiegata attraverso un modello a variabili latenti, come descritto in Thibaudeau (1989, 1993) e in Armstrong e Mayda (1993). In particolare, il modello (6.6) è un modello di indipendenza fra le variabili Y^h condizionatamente al gruppo di appartenenza della coppia ($C = 1$ se la coppia è in $\mathcal{M}$ e $C = 0$ altrimenti). Poiché la variabile C non è osservata, questa variabile può essere vista come una variabile latente. Indicando con $n_{c;\mathbf{y}}$ la frequenza delle coppie nel gruppo c ($c = 0, 1$) con confronto fra variabili chiave $\mathbf{y}$, ($\mathbf{y} \in \mathcal{D}$), Thibaudeau considera inizialmente il seguente modello loglineare[4]:

$$\log\left(E(n_{c;\mathbf{y}})\right) = \theta + \theta_c + \sum_{h=1}^{k} \theta^h_{y^h} + \sum_{h=1}^{k} \theta^h_{c;y^h} \tag{6.10}$$

[4]Si ricorda che il modello loglineare è definito da una relazione lineare fra il logaritmo delle frequenze attese delle celle e i parametri di interazione fra le variabili

con i vincoli:

$$\theta_1 = -\theta_0; \;\; \theta_1^h = -\theta_0^h; \;\; \theta_{c;1}^h = -\theta_{c;0}^h; \;\; \theta_{1;y^h}^h = -\theta_{0;y^h}^h, \tag{6.11}$$

per $c = 0, 1, \;\; h = 1, ..., k, \;\; y^h = 0, 1$. Questo modello include un termine per la variabile latente (θ_c), un termine per il confronto di ogni variabile chiave ($\theta_{y^h}^h$) e un termine che descrive l'interazione fra ogni variabile chiave e la variabile latente ($\theta_{c;y^h}^h$). Non compaiono interazioni fra confronti di variabili, ovvero siamo nel caso di indipendenza fra le variabili chiave Y^h come nel modello (6.6). La stima può essere condotta usando il metodo EM descritto nel paragrafo 6.3.2.

Thibaudeau, usando coppie di dati sulle quali la variabile latente è stata osservata (i dati del censimento del 1990 per la zona di Saint Luois, Missouri, e i corrispondenti dati della prova generale del censimento svolta nella stessa zona nel 1988), ha notato che i confronti fra variabili chiave possono allontanarsi di molto dal modello (6.10). Questo problema è molto più evidente una volta che i data set vengono bloccati rispetto all'unità geografica (ovvero si considerano solo le coppie di record che risiedono nella stessa unità geografica; per maggiori dettagli si consideri il paragrafo 5.1). L'operazione di bloccaggio rende i confronti fra le 4 variabili chiave

- "Cognome"
- "Numero civico"
- "Via"
- "Numero di telefono"

particolarmente dipendenti. Ad esempio si considerino due individui diversi (quindi, nonostante il bloccaggio, la coppia è in $\mathcal{U}$) che risiedono nella stessa zona (e quindi vengono studiati dal record linkage per verificare il loro status). Si supponga che questi due individui abbiano lo stesso cognome. Intuitivamente la probabilità che i due record siano identici in una delle restanti tre variabili chiave condizionatamente al fatto che presentano lo stesso cognome è più alta della corrispondente probabilità marginale. Per questo motivo, Thibaudeau afferma che è opportuno considerare un modello alternativo a (6.10), che rappresenti una forma di dipendenza fra queste 4 variabili per le coppie che sono non-match. Nell'esempio da lui considerato, le variabili chiave sono 11; per semplicità Thibaudeau suppone che le prime 4 siano cognome, numero civico, via e numero di telefono, ovvero le variabili che risultano dipendenti per le coppie che sono non-match. Il modello loglineare rappresenta i valori attesi delle frequenze nel modo seguente:

$$\begin{aligned} \log\left(E(n_{c;\mathbf{y}})\right) \;\; = \;\; & \theta + \theta_c + \sum_{h=1}^{11} \theta_{y^h}^h + \sum_{h=1}^{11} \theta_{c;y^h}^h + \\ & + (1-c)\left(\sum_{1 \le j < l \le 4} \theta_{y^j,y^l}^{j,l} + \sum_{1 \le j < l < v \le 4} \theta_{y^j,y^l,y^v}^{j,l,v} + \theta_{y^1,y^2,y^3,y^4}^{1,2,3,4} \right). \end{aligned} \tag{6.12}$$

Il coefficiente $(1 - c)$ sta ad indicare che la dipendenza fra le variabili di confronto viene considerata solo per le coppie che sono non-match, e solo per le quattro variabili chiave di cui abbiamo parlato. Oltre ai vincoli sui parametri definiti in (6.11) devono essere imposti anche gli ulteriori vincoli:

$$\sum_{y^j} \theta_{y^j,y^l}^{j,l} = \sum_{y^l} \theta_{y^j,y^l}^{j,l} = \sum_{y^j} \theta_{y^j,y^l,y^v}^{j,l,v} = \sum_{y^l} \theta_{y^j,y^l,y^v}^{j,l,v} = \sum_{y^v} \theta_{y^j,y^l,y^v}^{j,l,v} = 0,$$

$$\sum_{y^1} \theta^{1,2,3,4}_{y^1,y^2,y^3,y^4} = \sum_{y^2} \theta^{1,2,3,4}_{y^1,y^2,y^3,y^4} = \sum_{y^3} \theta^{1,2,3,4}_{y^1,y^2,y^3,y^4} = \sum_{y^4} \theta^{1,2,3,4}_{y^1,y^2,y^3,y^4} = 0.$$

La stima dei parametri per questo modello è simile al metodo EM, ma Thibaudeau afferma che il passo di massimizzazione M deve essere sostituito dall'applicazione dell'algoritmo di Newton-Raphson, dato che spesso il massimo non è ottenibile in forma chiusa.

In alternativa, si può considerare il metodo iterativo di Haberman (1975), chiamato in inglese *iterative scaling*, usato ad esempio da Armstrong e Mayda (1993). Questo metodo consente di ottenere stime di massima verosimiglianza dei parametri dei modelli loglineari con variabili latenti (per una esposizione esauriente dei metodi di massima verosimiglianza per modelli loglineari con variabili latenti si rimanda a Haberman, 1978). Per questo metodo conviene restringere l'attenzione ai soli modelli loglineari gerarchici. Inoltre supponiamo che, se un effetto interattivo fra r variabili chiave è presente nel modello, ad esempio $\theta^{1,2,3,4}_{y^1,y^2,y^3,y^4}$, allora deve essere non nullo anche l'effetto interattivo fra le stesse variabili e la variabile latente, ovvero $\theta^{1,2,3,4}_{c;y^1,y^2,y^3,y^4}$. Senza perdere in generalità, verranno considerate solo le stime dei parametri della distribuzione dei confronti per le coppie che sono match ($c = 1$). Per questa distribuzione assumiamo che il modello loglineare delle variabili di confronto $\mathbf{Y}$ sia caratterizzato da v tabelle sufficienti minimali, S_1,...,S_v, e sia $\mathbf{l}$ il generico vettore marginale di modalità di una di queste tabelle.

Il procedimento di iterative scaling può essere esemplificato nel modo seguente.

1. Si considerino le stime preliminari delle distribuzioni $m(.)$, $u(.)$ e della probabilità p: indichiamo queste stime iniziali con $m^0(.)$, $u^0(.)$ e p^0. Ad esempio, Armstrong e Mayda suggeriscono di considerare le stime che si ottengono sotto il modello di indipendenza dei confronti delle variabili chiave. Queste stime preliminari consentono di effettuare una prima suddivisione delle frequenze osservate $n_{\mathbf{y}}$ nella parte che compete ai match:

$$\phi^0_{1;\mathbf{y}} = \frac{p^0\, m^0(\mathbf{y})}{p^0\, m^0(\mathbf{y}) + (1-p^0) u^0(\mathbf{y})} n_{\mathbf{y}}$$

   e in quella che compete ai non-match

$$\phi^0_{0;\mathbf{y}} = \frac{(1-p^0)\, u^0(\mathbf{y})}{p^0\, m^0(\mathbf{y}) + (1-p^0) u^0(\mathbf{y})} n_{\mathbf{y}}.$$

   Si noti che $n_{\mathbf{y}} = \phi^0_{1;\mathbf{y}} + \phi^0_{0;\mathbf{y}}$, ma che $\phi^0_{1;\mathbf{y}}$ e $\phi^0_{0;\mathbf{y}}$ possono essere numeri non interi (in realtà sono approssimazioni delle frequenze attese per i confronti $\mathbf{y}$ per i match e i non-match).

2. La tabella dei confronti per i match $\{\phi^0_{1;\mathbf{y}}\}$ definita al punto precedente viene adattata alle tabelle sufficienti minimali del modello loglineare per i match. Si consideri la prima tabella sufficiente minimale, S_1, e sia $\mathbf{l}$ il generico vettore marginale di modalità in S_1. Si aggiornano tutte le frequenze dei confronti $\mathbf{y}$ compatibili con il vettore marginale $\mathbf{l}$:

$$n^1_{1;\mathbf{y}} = n\, p^0\, m^0(\mathbf{y}) \frac{\sum_{\mathbf{l}} \phi^0_{1;\mathbf{y}}}{\sum_{\mathbf{l}} n\, p^0\, m^0(\mathbf{y})}$$

   dove le somme precedenti definiscono le marginali su $\mathbf{l}$ rispettivamente delle tabelle $\phi^0_{1;\mathbf{y}}$ e $\{np^0 m^0(\mathbf{y})\}$. Si esegue questa operazione per ogni $\mathbf{y}$ compatibile con $\mathbf{l}$ e per ogni $\mathbf{l}$ definibile in S_1, ottenendo in questo modo una nuova tabella complessiva $\{n^1_{1;\mathbf{y}}\}$.

L'aggiornamento con le altre tabelle viene fatto allo stesso modo. Supponendo note le frequenze al passo $i-1$, $i = 2, 3, ...v$, le frequenze al passo i sono definite dall'operazione di redistribuzione rispetto alla tabella sufficiente minimale S_i (in questo caso $\mathbf{l}$ è il vettore marginale di modalità in S_i):

$$n^i_{1;\mathbf{y}} = n^{i-1}_{1;\mathbf{y}} \frac{\sum_{\mathbf{l}} \phi^0_{1;\mathbf{y}}}{\sum_{\mathbf{l}} n^{i-1}_{1;\mathbf{y}}}. \tag{6.13}$$

L'ultimo aggiornamento definisce le frequenze $n^v_{1;\mathbf{y}}$, $\mathbf{y} \in \mathcal{D}$.

3. Completato l'aggiornamento delle frequenze rispetto a tutte le tabelle marginali della configurazione sufficiente minimale per il modello loglineare definito, si aggiorna la partizione delle frequenze osservate $n_{\mathbf{y}}$ come al punto 1:

$$\phi^v_{1;\mathbf{y}} = \frac{n^v_{1;\mathbf{y}}}{n^v_{1;\mathbf{y}} + n^v_{0;\mathbf{y}}} n_{\mathbf{y}}$$

per i match e:

$$\phi^v_{0;\mathbf{y}} = \frac{n^v_{0;\mathbf{y}}}{n^v_{1;\mathbf{y}} + n^v_{0;\mathbf{y}}} n_{\mathbf{y}}$$

per i non-match.

4. Si aggiornano le frequenze in base alle tabelle sufficienti minimali per il modello loglineare per la distribuzione dei confronti per i match, usando la formula (6.13) dove al posto delle frequenze $\phi^0_{1;\mathbf{y}}$ si usano le frequenze $\phi^v_{1;\mathbf{y}}$.

L'algoritmo viene fermato quando le frequenze stimate ai passi successivi subiscono alterazioni inferiori a un valore soglia prefissato.

Si deve sottolineare che, mentre Haberman afferma che un algoritmo del genere può convergere a massimi locali, e quindi è preferibile usare diversi valori iniziali ($m^0(.)$, $u^0(.)$ e p^0) per essere sicuri del risultato, Thibaudeau, Winkler e altri hanno notato che per il record linkage questo non è vero.

Commento 6.5 *Thibaudeau ha osservato un incremento notevole dell'efficienza delle procedure di record linkage quando vengono usati i modelli del paragrafo precedente. In particolare, modelli di dipendenza fra i confronti Y^h delle variabili chiave si rivelano più discriminanti rispetto al modello (6.6). Questo maggiore potere discriminante è misurato dal numero maggiore di coppie che vengono dichiarate match dalla procedura di record linkage e dalla minore incidenza di coppie dichiarate match in modo errato (falsi match). Per una analisi generale della qualità del record linkage si rimanda al capitolo 8.* □

Metodi iterativi per stimare i parametri di modelli loglineari con variabili latenti quando alcuni parametri subiscono dei vincoli sono stati sviluppati anche da Winkler (1989, 1993), come descritto nel paragrafo 6.4.2.

6.4.2 Cosa fare se il modello loglineare non è noto: Winkler (1989, 1993)

Il problema che Winkler analizza riguarda la stima dei parametri di un generico modello loglineare a variabili latenti (il modello (6.12) ne è un esempio), ovvero:

$$L\Big(p, \{m(.)\}, \{u(.)\}\Big|\mathbf{y}_{ab}, \{c_{a,b}\}\Big) = \prod_{(a,b)} \Big(p\, m(\mathbf{y}_{ab})\Big)^{c_{a,b}} \Big((1-p)\, u(\mathbf{y}_{ab})\Big)^{1-c_{a,b}} \tag{6.14}$$

dove $m(\mathbf{y})$ e $u(\mathbf{y})$, $\mathbf{y} \in \mathcal{D}$, sono definiti da modelli loglineari distinti. Se i modelli loglineari per $m(\mathbf{y})$ e $u(\mathbf{y})$ sono noti si può prendere in considerazione uno dei metodi illustrati nel paragrafo 6.4.1. Se i modelli non sono noti, Winkler (1989b, 1992) e Rubin e Stern (1993) hanno verificato che non è opportuno scegliere il modello attraverso test del tipo chi-quadrato per valutare l'adattamento delle stime ai modelli, in quanto i risultati generalmente sono poco soddisfacenti.

Winkler (1989b, 1993) suggerisce di usare un modello di interazione sufficientemente generico, ad esempio modelli log-lineari che includano tutte le interazioni di ordine 3, anche se questo non è vero né è giustificato da esperienze precedenti o test. La fase di stima dei parametri di questi modelli viene poi fatta restringendo l'insieme dei parametri a un sottoinsieme dello spazio dei parametri fondato su conoscenze a priori. Winkler afferma che se i vincoli posti sono appropriati, le stime dei parametri e le regole di decisione sono "buone" tanto quanto quelle che si ottengono usando i modelli loglineari specifici, corretti. Per poter ottenere delle buone stime, Winkler suggerisce di usare ancora l'algoritmo EM, ma modificato opportunamente in modo da tener conto dei vincoli. Questo algoritmo viene chiamato EMH, dove H sta per Haberman che ha ottenuto i risutati principali (Haberman, 1977).

Algoritmo EMH Si consideri il modello (6.14), dove le distribuzioni $m(.)$ e $u(.)$ sono multinomiali e soddisfano opportuni modelli loglineari. I parametri da stimare sono:

$$\boldsymbol{\omega} = \Big\{p, \{m(\mathbf{y}), \mathbf{y} \in \mathcal{D}\}, \{u(\mathbf{y}), \mathbf{y} \in \mathcal{D}\}\Big\}. \tag{6.15}$$

Indichiamo con Ω lo spazio di tutti i possibili parametri $\boldsymbol{\omega}$. Haberman (1977) ha dimostrato il seguente teorema.

Teorema 6.1 *Supponiamo che i parametri da stimare siano definiti da misture di multinomiali, come in (6.15) (il teorema rimane valido quando la mistura è definita con delle distribuzioni di Poisson). Se $\boldsymbol{\omega}_i$ e $\boldsymbol{\omega}_{i+1}$ sono due stime successive ottenute attraverso l'algoritmo EM allora, per ogni $0 \leq \alpha \leq 1$, si ha:*

$$L(\boldsymbol{\omega}_i) \leq L\Big(\alpha\, \boldsymbol{\omega}_i + (1-\alpha)\, \boldsymbol{\omega}_{i+1}\Big)$$

dove $L(\boldsymbol{\omega})$ rappresenta la verosimiglianza per $\boldsymbol{\omega} \in \Omega$.

Qualsiasi parametro $\boldsymbol{\omega} \in \Omega$ nel segmento che unisce $\boldsymbol{\omega}_i$ con $\boldsymbol{\omega}_{i+1}$ è caratterizzato da una verosimiglianza non inferiore rispetto al punto di partenza $\boldsymbol{\omega}_i$. Questo teorema giustifica, ogni qual volta è possibile, la restrizione dell'insieme Ω dei parametri ad un sottoinsieme chiuso e convesso $\mathcal{R}$. Fra i vincoli suggeriti da Winkler alcuni sono ovvi, come:

$$p \leq \frac{\min\{\nu_A, \nu_B\}}{\nu_A \times \nu_B},$$

altri sono invece giustificati da informazioni esterne, come:

$$m_h \geq q_h, \quad h = 1, ..., k$$

$$u_h \leq r_h, \quad h = 1, ..., k$$

con $q_h \geq 0$ e $r_h \leq 1$, $h = 1, ..., k$. Identifichiamo, come in Meng e Rubin (1993), con g_i, $i = 1, ..., S$, i vincoli tipici del metodo IPF (*Iterative Proportional Fitting* in inglese). Questi vincoli vengono posti quando non è possibile determinare una stima diretta di massima verosimiglianza dei parametri del modello loglineare. Le iterazioni del metodo EMH sono indicate di seguito.

1. Si consideri un parametro di partenza $\boldsymbol{\omega}_0 \in \mathcal{R}$, che soddisfa quindi i vincoli imposti. Si esegua il passo E che completa il dataset dei dati mancanti.

2. Usando i vincoli imposti dai dati completi e il primo vincolo g_1, si determini la stima di massima verosimiglianza $\boldsymbol{\omega}_1 \in \Omega$. Se $\boldsymbol{\omega}_1 \in \mathcal{R}$ si lasci questa stima inalterata. Se non risiede nella regione convessa $\mathcal{R}$, si determini il valore α, $0 \leq \alpha \leq 1$, che "proietti" la stima ottenuta su $\mathcal{R}$, come indicato nel teorema 6.1. I parametri appena stimati vengono usati per riempire i dati mancanti con il passo E.

3. Condizionatamente ai vincoli imposti dai dati completati e g_2, si determinino le stime di massima verosimiglianza $\boldsymbol{\omega}_2 \in \Omega$. Se $\boldsymbol{\omega}_2 \in \mathcal{R}$ si lasci questa stima inalterata, altrimenti la si proietti sullo spazio convesso $\mathcal{R}$. Si esegua di nuovo il passo E per completare i dati rispetto ai dati mancanti con questo nuovo parametro.

4. Si continui a seguire la stessa procedura, con i vincoli rimanenti g_i, $i = 3, ..., S$.

Si iterino gli S passi descritti in precedenza finché le stime dei parametri non si stabilizzano.

6.5 Metodi basati sulle frequenze: Fellegi e Sunter (1969), Winkler (1989)

Come già detto, i confronti del tipo (2.10) sono poveri di informazioni sulle caratteristiche della coppia. Una informazione preziosa non contenuta nei confronti (2.10) è data dalla frequenza con cui si verificano le modalità delle variabili chiave. Questo discorso era già presente in Newcombe *et al*. (1959) per quanto riguarda variabili chiave come il cognome. L'idea sottostante l'uso delle frequenze con cui si osservano le modalità è stata discussa nell'esempio 2.6. Questo metodo è stato analizzato, fra gli altri, da Fellegi e Sunter (1969) e da Winkler (1989). Nei loro lavori vengono esposte diverse considerazioni di "buon senso", che comunque hanno effetti sul piano pratico migliorando la qualità del record linkage.

Si considerino due liste A e B che contengono unità distinte rispetto alle variabili chiave. Supponiamo inoltre che i confronti siano del tipo (2.12), ovvero le coppie di unità coincidenti in una variabile chiave vengono distinte in base alla modalità della variabile stessa. Supponiamo infine che sia valido il modello (6.5), ovvero che i confronti Y^h fra le variabili chiave siano fra loro indipendenti. Per via di quest'ultima ipotesi, si possono considerare distintamente le diverse variabili chiave per ricostruire le distribuzioni $m(.)$ e $u(.)$ di interesse. In particolare facciamo riferimento alla variabile X^1 che può essere, ad esempio, il "cognome".

Sia Fellegi e Sunter che Winkler suggeriscono di costruire le distribuzioni di frequenza della variabile chiave nelle due liste. Supponendo che nelle due occasioni siano stati rilevati v cognomi diversi, siano:

$$f_1, f_2, ..., f_v, \qquad \sum_{l=1}^{v} f_l = \nu_A,$$

le frequenze dei v cognomi associate alle ν_A unità osservate nell'occasione A e:

$$g_1, g_2, ..., g_v, \qquad \sum_{l=1}^{v} g_l = \nu_B,$$

le frequenze osservate sulle unità della lista B. Le coppie $(a, b) \in \mathcal{A} \times \mathcal{B}$ che presentano il cognome coincidente e pari alla v-esima modalità sono esattamente $f_v \times g_v$, e lo stesso si può dire per le restanti $v - 1$ modalità. Questi valori sono facilmente calcolabili dai dati a disposizione. Di queste coppie solo alcune sono dei match e appartengono a $\mathcal{M}$. Sia

$$h_1, h_2, ..., h_v, \qquad \sum_{l=1}^{v} h_l = N,$$

dove N è il numero di coppie in $\mathcal{M}$, la distribuzione di frequenze dei cognomi sulle coppie che sono match. Questa distribuzione non è osservabile, ma si dispone solamente della relazione:

$$h_l \leq \min\{f_l, g_l\}, \qquad l = 1, 2, ..., v.$$

Come soluzione *ad hoc*, Winkler (1989) ha usato in alcune applicazioni:

$$h_l = \begin{cases} \min\{f_l, g_l\} & \text{se } f_l > 1 \text{ o } g_l > 1 \\ 2/3 & \text{se } f_l = 1 \text{ e } g_l = 1 \end{cases} \qquad l = 1, 2, ..., v. \tag{6.16}$$

In pratica il numero di coppie che sono match coincidono con la frequenza più piccola osservata nelle due occasioni, e se una modalità viene osservata solo una volta sia in A che in B allora a questa coppia viene imposta una probabilità pari a 2/3 di essere un match e 1/3 di essere un non-match.

Fellegi e Sunter, invece, propongono di considerare il modello:

$$\frac{f_l}{\nu_A} = \frac{g_l}{\nu_B} = \frac{h_l}{N}, \qquad l = 1, ..., v \tag{6.17}$$

e di stimare dai due campioni (dalle due liste) osservati nelle occasioni A e B le frequenze di interesse tenendo conto del modello 6.17 attraverso opportuni metodi.

I modelli (6.16) e (6.17) non considerano la possibilità che il cognome venga riportato nel modo sbagliato (errori di compilazione, di digitazione) oppure venga omesso (mancata risposta parziale). A questo scopo, sia Fellegi e Sunter che successivamente Winkler affermano che è necessario considerare le seguenti probabilità di errore:

1. e_A e e_B: le probabilità che un cognome osservato contenga degli errori rispettivamente nella lista A e B;

2. e_{A0} e e_{B0}: le probabilità che un cognome non venga riportato rispettivamente nella lista A e B;

3. e_T: un individuo presenti cognomi diversi nelle due occasioni senza commettere errori (ad esempio per una donna il primo cognome è da nubile e il secondo da sposata).

Inoltre gli autori ipotizzano che:

1. gli errori descritti nei tre punti precedenti si verifichino in modo indipendente l'uno dall'altro;

2. gli errori siano indipendenti rispetto ai diversi cognomi.

Aggiornando le frequenze h_l, f_l e g_l con le probabilità ora indicate, si ottengono le probabilità rispettivamente:

- di osservare una delle v modalità del cognome, si identifichi questo evento con il simbolo x_l, $l = 1, ..., v$,
- di osservare cognomi diversi, si identifichi questo evento con il simbolo d, e
- che almeno un elemento della coppia sia mancante, $\emptyset$,

sia per le coppie che sono match (cioè per la distribuzione $m(.)$) che per le coppie che sono non-match (cioè $u(.)$). Per le ipotesi fatte in precedenza queste distribuzioni possono essere stimate da:

$$
\begin{aligned}
m(x_l) &= \frac{h_l}{N}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0}), \qquad l = 1, ..., v \\
m(d) &= [1-(1-e_A)(1-e_B)(1-e_T)]\,(1-e_{A0})(1-e_{B0}) \\
m(\emptyset) &= 1-(1-e_{A0})(1-e_{B0}) \\
u(x_l) &= \frac{f_l g_l - h_l}{\nu_A \nu_B - N}(1-e_A)(1-e_B)(1-e_T)(1-e_{A0})(1-e_{B0}), \qquad l = 1, ..., v \\
u(d) &= \left[1-(1-e_A)(1-e_B)(1-e_T)\sum_{l=1}^{v}\frac{f_l g_l - h_l}{\nu_A \nu_B - N}\right](1-e_{A0})(1-e_{B0}) \\
u(\emptyset) &= 1-(1-e_{A0})(1-e_{B0}).
\end{aligned}
$$

Per semplificare ulteriormente i calcoli, si può supporre che le probabilità di errore siano comunque molto piccole, e che quindi il loro prodotto sia trascurabile. In particolare l'evento: "due unità assumano la stessa modalità anche se tutte e due le osservazioni sono errate" viene considerato impossibile (quindi è nullo il prodotto $e_A e_B$). Sotto queste ipotesi le stime precedenti si semplificano in:

$$
\begin{aligned}
m(x_l) &\approx \frac{h_l}{N}(1-e_A-e_B-e_T-e_{A0}-e_{B0}), \qquad l = 1, ..., v && (6.18)\\
m(d) &\approx e_A+e_B+e_T && (6.19)\\
m(\emptyset) &\approx e_{A0}+e_{B0} && (6.20)\\
u(x_l) &\approx \frac{f_l g_l - h_l}{\nu_A \nu_B - N}(1-e_A-e_B-e_T-e_{A0}-e_{B0}), \qquad l = 1, ..., v && (6.21)\\
u(d) &= \left[1-(1-e_A-e_B-e_T)\sum_{l=1}^{v}\frac{f_l g_l - h_l}{\nu_A \nu_B - N}\right](1-e_{A0}-e_{B0}) && (6.22)\\
u(\emptyset) &\approx e_{A0}+e_{B0}. && (6.23)
\end{aligned}
$$

Dalle equazioni (6.18) e (6.21) è possibile dedurre i pesi per le coppie che presentano la stessa modalità x_l della variabile X^1. Infatti, tenendo conto della (4.2) e della (6.16), si ha:

$$
t(x_l) = \begin{cases} h_l(\nu_A\nu_B - N)/(f_l g_l - h_l)N & \text{se } f_l > 1 \text{ o } g_l > 1 \\ 2(\nu_A\nu_B - N)/N & \text{se } f_l = 1 \text{ e } g_l = 1 \end{cases} \qquad l = 1, 2, ..., v.
$$

Per il calcolo di questi pesi, Winkler (1989) dà alcuni suggerimenti. Ad esempio i valori h_l e N possono essere vincolati alle stime delle probabilità m_1 e u_1 del paragrafo 6.3.2 ottenute, ad esempio, tramite l'algoritmo EM. Infatti le probabilità $m(x_l)$ e $u(x_l)$, $l = 1, ..., v$, devono soddisfare i vincoli:

$$m_1 = \sum_{l=1}^{v} m(x_l), \quad u_1 = \sum_{l=1}^{v} u(x_l).$$

Per quanto riguarda le probabilità di errore, e_{A0} e e_{B0} sono facilmente desumibili dalle basi dati a disposizione, controllando l'incidenza dei valori mancanti della variabile chiave di interesse in A e B. Per le altre probabilità risultano estremamente utili le approssimazioni definite nelle formule (6.18)-(6.23). Infatti se si volessero conoscere le probabilità e_A, e_B e e_T singolarmente, si dovrebbe fare riferimento a informazioni esterne. Al contrario dalla formula (6.19) è sufficiente ricavare, sempre con l'algoritmo EM, una stima di $1-m_1$ (ovvero della probabilità che una coppia che è un match fornisca cognomi diversi) per avere una stima di

$$e_A + e_B + e_T$$

che è quanto richiesto da tutte le equazioni (6.18)-(6.23).

Per una valutazione complessiva dei metodi proposti rispettivamente da Fellegi e Sunter e da Winkler e ad alcune estensioni si rimanda a Yancey (2000).

6.6 I confronti più informativi: Copas e Hilton (1990)

Come nel paragrafo precedente, anche Copas e Hilton ritengono che l'informazione fornita dalle modalità delle variabili chiave sia estremamente importante. Per questo motivo, come già ampiamente anticipato più volte, Copas e Hilton adottano i confronti del tipo (2.13). Dato che questi confronti sono i più difficile da studiare, inizialmente si fa riferimento ad una sola variabile chiave, X^h, che si suppone assuma v modalità $x = 1, ..., v$, e al corrispondente confronto Y^h. Solo successivamente si riuniscono le informazioni provenienti da più confronti e si stima la loro distribuzione congiunta.

Si è già visto negli esempi 3.5 e 3.7 come si devono modellizzare le distribuzioni di probabilità di $Y^h = (X_A^h, X_B^h)$ quando la coppia (a, b) è un match ($C = 1$) e quando non lo è ($C = 0$). Se la coppia è un non-match, le osservazioni (X_A^h, X_B^h) provengono da unità diverse ed è quindi lecito che i due valori provengano da variabili indipendenti:

$$u(y^h) = P\left(X_A^h = x_A\right) P\left(X_B^h = x_B\right) = p_{x_A} p_{x_B},$$

dove p_{x_A}, $x_A = 1, ..., v$, è la distribuzione marginale di X^h sulla lista A e p_{x_B}, $x_B = 1, ..., v$, è la corrispondente distribuzione marginale sulla lista B. Queste distribuzioni possono facilmente essere stimate dai dati a disposizione.

Il problema riguarda invece la stima della distribuzione di Y^h quando la coppia è un match. Come descritto nell'esempio 6.1, gli autori suggeriscono di affidarsi a un campione di coppie che sono match, controllato attraverso la revisione manuale. Questi dati possono essere usati per stimare la distribuzione:

$$m(y^h) = P\left(X_A^h = x_A, X_B^h = x_B\right) = p_{x_A, x_B} \tag{6.24}$$

attraverso la frequenza congiunta osservata sul campione delle modalità (x_A, x_B), $x_A, x_B = 1, ..., v$. La distribuzione (6.24) è caratterizzata da v^2 parametri incogniti, ed è tanto più complicato stimarla quanto maggiore è il numero di modalità della variabile X^h.

Copas e Hilton suggeriscono allora di definire la (6.24) attraverso modelli che possiedono un numero di parametri incogniti relativamente ridotto. Ne propongono alcuni, che definiamo nel seguente paragrafo.

6.6.1 Il modello di errore di misura

Per ogni coppia in $\mathcal{M}$, assumiamo che la variabile[5] X^h assuma un valore, incognito, T. T indica quindi il vero ma incognito valore della variabile X^h associato all'unità che si sta studiando. Sia

$$P(T = t) = \beta_t, \qquad t = 1, ..., v,$$

la vera distribuzione della variabile di interesse nella popolazione. Nelle due occasioni A e B siamo in grado di osservare le variabili X_A^h e X_B^h, che possono essere diverse fra loro e diverse dal vero valore T che dovrebbero assumere. Supponendo che, dato il vero valore $T = t$, le osservazioni differiscano da t allo stesso modo nelle due occasioni:

$$P(X_A^h = x|T = t) = P(X_B^h = x|T = t) = \alpha_{xt}, \qquad x = 1, ..., v;\ t = 1, ..., v, \tag{6.25}$$

e che, condizionatamente a T le osservazioni X_A e X_B siano indipendenti, la distribuzione congiunta delle osservazioni è data da:

$$p_{x_A,x_B} = \sum_{t=1}^{v} \alpha_{x_A t}\alpha_{x_B t}\beta_t, \qquad x_A, x_B = 1, ..., v. \tag{6.26}$$

Commento 6.6 *Copas e Hilton affermano che l'ipotesi di indipendenza fra X_A e X_B condizionatamente a T, usata per definire la (6.26), è molto meno restrittiva di quanto possa sembrare. Se T non viene considerato come il "vero valore assunto dalle unità", è sufficiente definire T in modo che contenga gli errori sistematici.* □

Il modello (6.26) è estremamente complesso, in quanto coinvolge $v + v^2$ parametri. La prima semplificazione che Copas e Hilton considerano consiste nel ritenere che le probabilità α_{xt} siano piccole quando $x \neq t$, implicando che prodotti fra α_{xt} possono essere considerati praticamente nulli. Il modello approssimato diventa:

$$p_{x_A,x_B} = \begin{cases} \alpha_{x_B x_A}\beta_{x_A} + \alpha_{x_A x_B}\beta_{x_B} & x_A \neq x_B \\ \left(1 - 2\sum_{r\neq x_A} \alpha_{r x_A}\right)\beta_{x_A} & x_A = x_B. \end{cases} \tag{6.27}$$

Commento 6.7 *La (6.27) si ottiene con i seguenti passaggi. Per l'assunzione sulle probabilità*

[5] D'ora in avanti si fa riferimento a variabili immodificabili nel tempo, come le variabili "sesso" o "data di nascita"

α_{xt}, annullando le quantità che contengono α_{xt} più di una volta, si ottiene, per $x_A \neq x_B$:

$$
\begin{aligned}
& p_{x_A,x_B} = \sum_t \alpha_{x_A t}\alpha_{x_B t}\beta_t \approx \\
& \approx \alpha_{x_A x_A}\alpha_{x_B x_A}\beta_{x_A} + \alpha_{x_B x_B}\alpha_{x_A x_B}\beta_{x_B} = \\
& = \left(1 - \sum_{r \neq x_A} \alpha_{r x_A}\right)\alpha_{x_B x_A}\beta_{x_A} + \left(1 - \sum_{r \neq x_B} \alpha_{r x_B}\right)\alpha_{x_A x_B}\beta_{x_B} \approx \\
& \approx \alpha_{x_B x_A}\beta_{x_A} + \alpha_{x_A x_B}\beta_{x_B}.
\end{aligned}
$$

Passaggi simili portano al risultato per $x_A = x_B$. □

Il modello (6.27) rappresenta una semplificazione computazionale rispetto al modello (6.26), ma le difficoltà legate al numero di parametri rimangono. Copas e Hilton suggeriscono di imporre una struttura ai parametri α_{xt} e β_t. Gli autori considerano un esempio legato alla variabile "sesso", con modalità maschio=1, femmina=2, mentre un valore mancante viene codificato con il valore 3. Ipotizzando che:

$$\alpha_{12} = \alpha_{21} = a;\ \alpha_{31} = \alpha_{32} = b;\ \beta_3 = \alpha_{13} = \alpha_{23} = 0,$$

dove a indica la probabilità che si dichiari il sesso opposto a quello vero e b che alla variabile sesso corrisponda una mancata risposta, la matrice delle probabilità $p_{x_A x_B}$ è simmetrica con valori dati da:

$$
\begin{pmatrix}
\beta_1(1-2a-2b) & a & \beta_1 b \\
 & (1-\beta_1)(1-2a-2b) & (1-\beta_1)b \\
 & & 0
\end{pmatrix}
$$

Quindi è necessario stimare solo 3 parametri: a, b, β_1. Indicando con f_{rs} le frequenze osservate nel *training sample* di n coppie che sono match, $r = 1, 2, 3$, $s = 1, 2, 3$, ed eliminando dal campione le coppie f_{33}, cioè quelle che presentano un valore mancante nella variabile "sesso" nelle due occasioni A e B, le stime di massima verosimiglianza dei parametri vengono indicate da Copas e Hilton e sono:

$$\hat{a} = \frac{f_{12} + f_{21}}{2(n - f_{33})},$$

$$\hat{b} = \frac{f_{13} + f_{23} + f_{31} + f_{32}}{2(n - f_{33})},$$

$$\hat{\beta}_1 = \frac{f_{11} + f_{13} + f_{31}}{f_{11} + f_{13} + f_{31} + f_{22} + f_{23} + f_{32}}.$$

6.6.2 Il modello "hit-miss"

Il modello hit-miss è un modello specifico per le probabilità α_{xt}. L'ipotesi semplificatrice consiste nel ritenere che

$$p_x = \beta_x,$$

ovvero che la distribuzione marginale delle osservazioni coincida con la distribuzione della variabile T e che quindi quest'ultima sia desumibile dai data-set a disposizione. Inoltre si

considera una "prova binomiale", che "centra il bersaglio" (hit) con probabilità $1-a$, e "fallisce il bersaglio" (miss) con probabilità a. La prova binomiale serve a questo scopo: se osserviamo il valore x della variabile X^h, si centra il bersaglio se $T=x$ e si manca negli altri casi. Queste considerazioni portano a definire un modello per le probabilità α_{xt}. Avendo osservato $X^h=x$ con probabilità β_x per quanto ipotizzato precedentemente, si manca il bersaglio con probabilità:

$$\alpha_{xt}=a\beta_x, \qquad x\neq t,$$

e si centra con la probabilità restante[6]:

$$\alpha_{xt}=1-\sum_{x\neq t}a\beta_x=1-a(1-\beta_x), \qquad x=t.$$

La distribuzione congiunta delle osservazioni diventa:

$$p_{x_A,x_B}=\begin{cases} a(2-a)\beta_{x_A}\beta_{x_B} & x_A\neq x_B \\ \beta_{x_A}[1-a(2-a)(1-\beta_{x_A})] & x_A=x_B. \end{cases} \tag{6.28}$$

Commento 6.8 *La determinazione delle probabilità in (6.28) è giustificata dai seguenti passaggi. Per $x_A\neq x_B$:*

$$\begin{aligned} &p_{x_A,x_B}=\sum_t \alpha_{x_At}\alpha_{x_Bt}\beta_t= \\ &=\alpha_{x_Ax_A}\alpha_{x_Bx_A}\beta_{x_A}+\alpha_{x_Bx_B}\alpha_{x_Ax_B}\beta_{x_B}+\sum_{t\neq\{x_A,x_B\}}\alpha_{x_At}\alpha_{x_Bt}\beta_t= \\ &=[1-a(1-\beta_{x_A})]a\beta_{x_B}\beta_{x_A}+a\beta_{x_A}[1-a(1-\beta_{x_B})]\beta_{x_B}+\sum_{t\neq\{x_A,x_B\}}a^2\beta_{x_A}\beta_{x_B}\beta_t= \\ &=2a\beta_{x_A}\beta_{x_B}(1-a)+a^2\beta_{x_A}^2\beta_{x_B}+a^2\beta_{x_A}\beta_{x_B}^2+a^2\beta_{x_A}\beta_{x_B}(1-\beta_{x_A}-\beta_{x_B})= \\ &=a\beta_{x_A}\beta_{x_B}(2-a). \end{aligned}$$

Allo stesso modo si ottiene il risultato per $x_A=x_B$. □

Commento 6.9 *Il modello hit-miss per le probabilità α_{xt} è una semplificazione del teorema di Bayes. Infatti:*

$$\alpha_{xt}=P(X^h=x|T=t)=\frac{P(X^h=x)P(T=t|X^h=x)}{P(T=t)}=\beta_x a_{xt}$$

dove:

$$a_{xt}=\frac{P(T=t|X^h=x)}{P(T=t)} \qquad x=1,...,v;\ t=1,...,v.$$

Quindi il modello hit-miss semplifica il modello vero, fissando il termine a_{xt} a una costante. □

[6] e non $(1-\alpha)\beta_x$, come poteva sembrare dalle definizioni precedenti. Il nome "hit-miss" non è quindi esatto, anche se ben esemplifica il modello. La logica che ha portato alla formulazione di questo modello viene spiegata nel commento 6.9.

Inserendo la probabilità $b > 0$ che a un'osservazione corrisponda una mancata risposta, le probabilità α_{xt} definite dal modello hit-miss assumono la forma seguente:

$$\alpha_{xt} = \begin{cases} a\beta_x & 1 \leq x \neq t \leq v \\ 1 - b - a(1 - \beta_x) & 1 \leq x = t \leq v \\ b & x \text{ mancante.} \end{cases}$$

Assumendo che $\beta_b = 0$, e tenendo conto che:

$$p_{x_A,x_B} = p_{x_B,x_A}$$

la distribuzione di probabilità congiunta per le osservazioni nelle due occasioni assume ora la forma:

$$p_{x_A,x_B} = \begin{cases} a(2 - a - 2b)\beta_{x_A}\beta_{x_B} & 1 \leq x_A < x_B \leq v \\ \beta_{x_A}[(1-b)^2 - a(2 - a - 2b)(1 - \beta_{x_A})] & 1 \leq x_A = x_B \leq v \\ b(1-b)\beta_{x_A} & 1 \leq x_A \leq v;\ x_B \text{ mancante} \\ b^2 & x_A \text{ e } x_B \text{ mancanti.} \end{cases} \tag{6.29}$$

I parametri da stimare sono: $b,\ \beta_x,\ x = 1, ..., v,\ \delta = a(2 - a - 2b)$. Una stima di b si ottiene dal *training sample* delle coppie che sono match considerando la frequenza relativa dei valori mancanti. Come detto precedentemente, la distribuzione β_x è stimata dalla distribuzione di frequenze osservata nel *training sample* della variabile X^h fra i record che non presentano X^h come valore mancante. Per δ si tenga presente che:

$$P(X^h_A \neq X^h_B) = \sum_{x_A \neq x_B} p_{x_A x_B} = \delta \sum_{x_A \neq x_B} \beta_{x_A}\beta_{x_B} = \delta\Big(1 - \sum_x \beta_x^2\Big).$$

Quindi una stima di δ è fornita dal rapporto fra la frequenza di coppie nel *training sample* che possiedono valori discordanti ($x_A \neq x_B$), diviso per $1 - \sum_x \beta_x^2$, stimato già precedentemente.

Commento 6.10 *Copas e Hilton determinano anche il logaritmo del peso (4.2) da associare ad una coppia (a, b) sotto il modello hit-miss. Ricordando che si sta lavorando con una sola variabile chiave, X^h, il peso (4.2) assume la forma:*

$$t(y^h_{ab}) = \frac{m(y^h_{ab})}{u(y^h_{ab})} = \frac{p_{x_A,x_B}}{p_{x_A}p_{x_B}},$$

e il logaritmo del peso $t(y^h_{ab})$ è dato da:

$$\log\Big(t(y^h_{ab})\Big) = \begin{cases} \log(\delta) + 2\log(1-b) & 1 \leq x_A < x_B \leq v \\ \log\Big[1 - \delta(1-b)^{-2}(1 - \beta_{x_A})\Big] - \log(\beta_{x_A}) & 1 \leq x_A = x_B \leq v \\ 0 & x_A \text{ o } x_B \text{ mancante.} \end{cases}$$

Sotto il modello hit-miss, il contributo dato dalle coppie in cui almeno un componente è mancante è nullo. □

Commento 6.11 *Copas e Hilton determinano la divergenza simmetrizzata Δ (appendice C) fra il modello hit-miss e il modello che considera confronti del tipo (2.10). Da questo confronto si evince che il primo modello trattiene molta più informazione rispetto al secondo.* □

Commento 6.12 *Copas e Hilton considerano anche altri due modelli che generalizzano il modello hit-miss. Infatti nel modello (6.29) è sottintesa un'ipotesi forte di simmetria delle probabilità $p_{x_A x_B}$. Le modifiche apportate sono finalizzate a:*

- *far dipendere le probabilità di un "miss", a, dal vero valore $T = t$ (al proposito si veda il commento 6.9);*
- *alterare la distribuzione delle osservazioni quando si verifica un "miss".*

Queste modifiche intervengono nella definizione del "modello hit-miss per classi affini" e del "modello hit-miss di distanza".
I modelli che si possono considerare sono comunque innumerevoli. □

6.6.3 Come combinare i risultati ottenuti per le diverse variabili chiave

Se il modello è del tipo (6.5), le distribuzioni congiunte $m(.)$ e $u(.)$ relative a tutti i k confronti fra le variabili chiave si ottengono facilmente dalle corrispondenti marginali. Se il modello di indipendenza condizionata fra i k confronti non è adatto, Copas e Hilton suggeriscono una generalizzazione del modello hit-miss.

Nel modello hit-miss per una variabile chiave X^1, si ignori la possibilità di osservare valori mancanti ($b = 0$), e si supponga che la probabilità di un miss, a, sia costante per ogni individuo (cioè per ogni coppia di record), ma vari per le diverse coppie, in modo tale che $\gamma = a(2 - a)$ abbia media μ_γ. La distribuzione (6.28) diventa allora:

$$p_{x_A,x_B} = \begin{cases} \mu_\gamma \beta_{x_A} \beta_{x_B} & x_A \neq x_B \\ \beta_{x_A}[1 - \mu_\gamma(1 - \beta_{x_A})] & x_A = x_B. \end{cases}$$

I parametri di questo modello vengono stimati come nel modello hit-miss, ma sostituendo μ_γ a γ. Si supponga di stimare lo stesso modello anche per una seconda variabile chiave X^2, con parametri $\dot{\beta}_x$, $x = 1, ..., \dot{v}$, e $\mu_{\dot{\gamma}}$.

Il modello hit-miss correlato valido per rappresentare le due variabili chiave si basa su alcune ipotesi. Si supponga che le prove binomiali "hit-miss", rappresentate in questo caso dalle variabili γ e $\dot{\gamma}$, siano correlate con covarianza $\sigma_{\gamma\dot{\gamma}}$. Inoltre, condizionatamente a γ e $\dot{\gamma}$ le due prove binomiali sono indipendenti. Infine si supponga che i veri valori delle due variabili siano fra loro indipendenti, ovvero T è indipendente da $\dot{T}$. La distribuzione congiunta delle coppie di osservazioni $p_{x_A,x_B;\dot{x}_A,\dot{x}_B}$ per X^1 e X^2 diventa:

$$\begin{cases} \beta_{x_A}\beta_{x_B}\dot{\beta}_{\dot{x}_A}\dot{\beta}_{\dot{x}_B}(\mu_\gamma\mu_{\dot{\gamma}} + \sigma_{\gamma\dot{\gamma}}) & x_A \neq x_B,\ \dot{x}_A \neq \dot{x}_B, \\ \beta_{x_A}\dot{\beta}_{\dot{x}_A}\dot{\beta}_{\dot{x}_B}\left\{\mu_{\dot{\gamma}}[1 - \mu_\gamma(1 - \beta_{x_A})] - (1 - \beta_{x_A})\sigma_{\gamma\dot{\gamma}}\right\} & x_A = x_B,\ \dot{x}_A \neq \dot{x}_B, \\ \beta_{x_A}\beta_{x_B}\dot{\beta}_{\dot{x}_A}\left\{\mu_\gamma[1 - \mu_{\dot{\gamma}}(1 - \dot{\beta}_{\dot{x}_A})] - (1 - \dot{\beta}_{\dot{x}_A})\sigma_{\gamma\dot{\gamma}}\right\} & x_A \neq x_B,\ \dot{x}_A = \dot{x}_B, \\ \beta_{x_A}\dot{\beta}_{\dot{x}_A}\Big\{[1 - \mu_\gamma(1 - \beta_{x_A})][1 - \mu_{\dot{\gamma}}(1 - \dot{\beta}_{\dot{x}_A})] + & \\ \qquad + (1 - \beta_{x_A})(1 - \dot{\beta}_{\dot{x}_A})\sigma_{\gamma\dot{\gamma}}\Big\} & x_A = x_B,\ \dot{x}_A = \dot{x}_B. \end{cases}$$

Per stimare i parametri di questa distribuzione, si può seguire il seguente procedimento. Prima di tutto si stimano i parametri β_x, $x = 1, ..., v$, per la variabile X^1 e $\dot{\beta}_x$, $x = 1, ..., \dot{v}$, per la variabile

Tabella 6.2 - Frequenze dei confronti del tipo uguaglianza/disuguaglianza per due variabili chiave: X^1 e X^2

	X^2		
X^1	Uguaglianza	Disuguaglianza	Totale
uguaglianza	n_{11}	n_{12}	$n_{1.}$
disuguaglianza	n_{21}	n_{22}	$n_{2.}$
Totale	$n_{.1}$	$n_{.2}$	n

X^2. I parametri μ_γ e $\mu_{\dot\gamma}$ si stimano come si è stimato γ nel paragrafo sul modello hit-miss. A tal fine, la tabella 6.2 è di notevole aiuto. Infatti la stima dei parametri μ_γ e $\mu_{\dot\gamma}$ è:

$$\hat\mu_\gamma = \frac{n_{.2}}{n(1-\sum\beta_x^2)},$$

$$\hat\mu_{\dot\gamma} = \frac{n_{2.}}{n(1-\sum\dot\beta_x^2)}.$$

All'ultimo passo si può stimare la covarianza attraverso:

$$\hat\sigma_{\gamma\dot\gamma} = \hat\mu_\gamma\hat\mu_{\dot\gamma}\left(\frac{n\chi^2}{n_{.2}n_{2.}}\right)^{1/2}$$

dove χ^2 indica il coefficiente chi-quadrato della tabella (6.2) per valutare l'indipendenza fra le due variabili. Se $\chi^2 = 0$, anche $\sigma_{\gamma\dot\gamma} = 0$, e si ritorna al caso di variabili di confronto indipendenti.

Capitolo 7

Altri metodi per il record linkage

La proposta di Fellegi e Sunter, delineata nel capitolo 4, non è l'unico metodo di record linkage disponibile in letteratura. Alcuni altri metodi usano strumenti diversi per poter decidere se una coppia (a, b) è un match oppure no. In questo capitolo ne vengono descritti alcuni. Inizialmente si definiscono alcune procedure euristiche, evidenziandone le differenze con la procedura di Fellegi e Sunter. In particolare nel paragrafo 7.1 viene descritta una procedura ad hoc (Armstrong e Mayda, 1993). Questa procedura sembra infatti affine alla procedura proposta da Larsen e Rubin (2001), che però ha il vantaggio di essere "statisticamente fondata" (paragrafo 7.2). Infine si fanno alcuni cenni a una procedura Bayesiana per il record linkage scaturita da una collaborazione fra l'Istat e il Dipartimento di Studi Geoeconomici, Statistici, Storici per l'Analisi Regionale dell'Università di Roma "La Sapienza". Un ulteriore metodo, basato sulla formalizzazione della funzione dei costi legati al record linkage, è stato introdotto da Tepping (1968), ma non verrà trattato.

7.1 Metodi non statistici per il record linkage

I metodi non statistici per il record linkage si caratterizzano per due elementi:

1. scelta "soggettiva" dei pesi da affiancare ai confronti $\mathbf{y}$;

2. scelta "soggettiva" del livello di soglia al di sopra del quale una coppia viene considerata match.

Ad esempio Belin (1993) assegna pesi distintamente per le singole variabili chiave, dando peso +2 (-2) quando c'è accordo (disaccordo) sulle variabili chiave "età" o "numero telefonico" e peso +1 (-1) quando c'è accordo (disaccordo) sulle variabili chiave "sesso" e "relazione di parentela". I pesi per l'accordo/disaccordo sulle variabili chiave "età" e "numero telefonico" sono il doppio rispetto a quelli sul "sesso" e la "relazione di parentela" in quanto le prime variabili vengono ritenute più discriminanti delle seconde. Ciò non toglie che questi pesi sono stati posti arbitrariamente, e i pesi sulle variabili più discriminanti potevano essere posti pari al triplo o più di quelli non discriminanti. Inoltre non è detto che il potere discriminante delle variabili "età" e "numero telefonico" sia equivalente.

Nel secondo punto, si sottintende che il livello di soglia che divide le coppie considerate match da quelle considerate non-match rischia di essere posto senza badare ai possibili errori. Per poter valutare gli errori è necessario considerare una fase costosa di controllo manuale delle coppie, come si evidenzia nel capitolo 8.

7.1.1 Un esempio di procedura ad hoc

La procedura che descriviamo ora, definita in Armstrong e Mayda (1993), è una procedura ad hoc che imita il metodo di Fellegi e Sunter, senza sfruttarne le caratteristiche statistiche. Vengono considerati i confronti più semplici, definiti in (2.10) e, inizialmente, dei pesi che assomigliano ai pesi $t(\mathbf{y})$ definiti in (4.2) quando si ipotizza l'indipendenza fra le variabili di confronto, come nel modello (6.6):

$$m(\mathbf{y}) = \prod_{h=1}^{k} m_h^{y_{ab}^h}(1-m_h)^{1-y_{ab}^h},$$

$$u(\mathbf{y}) = \prod_{h=1}^{k} u_h^{y_{ab}^h}(1-u_h)^{1-y_{ab}^h}.$$

Il metodo richiede delle stime iniziali per le probabilità m_h e u_h. Per le probabilità m_h Armstrong e Mayda si affidano a esperienze precedenti e a valutazioni sulla qualità delle variabili chiave. Le stime delle probabilità u_h si ottengono dalla frequenza di $Y^h = 1$ nell'insieme delle $\nu_A \times \nu_B$ coppie in $\mathcal{A} \times \mathcal{B}$ (come descritto nel punto 2. nel paragrafo 6.1). Queste stime preliminari, si denotino con i simboli m_h^0 e u_h^0, $h = 1, ..., k$, vengono usate per costruire i pesi (4.2), $t^0(\mathbf{y})$, per il vettore di confronto $\mathbf{y}$, $\mathbf{y} \in \mathcal{D}$. Fissati due valori soglia $\tau_1 > \tau_2$, si costruiscano due insiemi di coppie M^0 e U^0 attraverso la regola:

- $(a, b) \in M^0$ se $t^0(\mathbf{y}_{ab}) > \tau_1$
- $(a, b) \in U^0$ se $t^0(\mathbf{y}_{ab}) < \tau_2$.

Le distribuzioni di frequenze relative dei confronti $\mathbf{y}$ per le coppie in M^0 e U^0 vengono usate come nuove stime delle probabilità m_h e u_h: si denotino con i simboli m_h^1 e u_h^1. Si itera il procedimento precedente finché le stime delle probabilità m_h e u_h subiscono variazioni inferiori a una soglia prefissata.

Gli autori affermano che le stime u_h^0 in genere non subiscono grandi variazioni, mentre la prima iterazione produce cambiamenti sostanziali per le probabilità m_h. In genere, queste ultime probabilità rimangono stabili dalla seconda iterazione in poi.

Commento 7.1 *Il procedimento appena descritto non può definirsi statistico per 2 motivi. Il primo è che i pesi, anche se non completamente soggettivi, non corrispondono a una stima sensata del peso $t(\mathbf{y})$. In pratica, i valori stimati di m_h e u_h possono essere molto distanti dai valori veri, anche quando il modello di indipendenza fra i confronti delle variabili chiave è adatto. Il peso che si ottiene non è più quindi una valida approssimazione del rapporto delle verosimiglianze. Il secondo motivo riguarda il valore soglia. I valori per m_h e u_h che si ottengono alla fine delle iterazioni, non essendo valori adatti a rappresentare i parametri m_h e u_h, non possono essere usati neanche per la stima di μ e λ, rendendo necessaria una procedura campionaria per conoscere il livello degli errori.* □

7.2 Il metodo iterativo di Larsen e Rubin (2001)

Larsen e Rubin introducono un metodo innovativo rispetto a quelli visti fino ad ora. In primo luogo, la loro procedura può definirsi come una "mistura" fra la procedura di decisione di Fellegi e

Sunter, con pesi definiti come in Torelli (paragrafo 4.2) e stima dei parametri basata su modelli, e una procedura iterativa, basata sul controllo manuale dei dati. L'obiettivo è quello di minimizzare la presenza degli errori nella regola di decisione, e nel contempo di minimizzare l'insieme delle coppie il cui status è incerto (ovvero per le quali si prende la decisione $A_\emptyset$ secondo la notazione del capitolo 4).

In secondo luogo, Larsen e Rubin generalizzano un modello studiato da Winkler (1992), definendo un modello statistico per i confronti fra le variabili chiave Y^h quando le $\nu_A \times \nu_B$ coppie vengono raggruppate in G (≥ 3) classi distinte (finora, tranne che nel paragrafo 6.4.1, è sempre stata considerata una bipartizione nelle coppie che sono match, $\mathcal{M}$, e nelle coppie che sono non-match, $\mathcal{U}$). Il motivo che ha condotto Larsen e Rubin a considerare un modello così generale è il seguente: è stato sempre notato che le relazioni di dipendenza che si vengono a instaurare fra i confronti delle variabili chiave Y^h possono essere molto complesse, e non identificabili con un unico modello di dipendenza. Quindi si suppone che il gruppo dei match e il gruppo dei non-match si possono dividere a loro volta in sottogruppi di coppie caratterizzati da diversi "gradi" di relazione fra le variabili di confronto. Il tipo di relazione fra le variabili di confronto in ogni gruppo viene descritto da un modello loglineare. La verosimiglianza (6.4), definita nel caso in cui le coppie potevano appartenere a due soli gruppi (gli abbinamenti e i non abbinamenti), viene generalizzata al caso in cui la verosimiglianza è una mistura di G distribuzioni (modelli loglineari), una per ogni gruppo. Per questi modelli è necessario considerare una notazione leggermente più sofisticata rispetto a quella adottata fino ad ora. Sia:

- p_g: frazione di coppie che appartengono alla classe g, $g = 1, ..., G$, $\sum_g p_g = 1$. Nella verosimiglianza (6.5) questi parametri erano solo 2: p per le coppie in $\mathcal{M}$ e $(1-p)$ per le coppie in $\mathcal{U}$.

- $p_{\mathbf{y};g} = P(\mathbf{Y}_{ab} = \mathbf{y}|g)$, ovvero la probabilità che il vettore di confronto per la coppia (a, b) assuma il valore $\mathbf{y}$ sapendo che la coppia appartiene al gruppo g, $g = 1, ..., G$, $\mathbf{y} \in \mathcal{D}$.

- La matrice $\mathbf{c}$ registra per ogni coppia (a, b) il gruppo di appartenenza, ovvero:

$$c_{a,b} \in \{1, 2, ..., G\}, \qquad (a, b) \in \mathcal{A} \times \mathcal{B}.$$

La verosimiglianza diventa allora:

$$L(\mathbf{p}, \mathbf{p}_{.;\mathbf{g}}, g = 1, ..., G|\mathbf{y}, \mathbf{c}) = \prod_{a=1}^{\nu_A} \prod_{b=1}^{\nu_B} \left[\prod_{g=1}^{G} p_g \left(\prod_{\mathbf{y} \in \mathcal{D}} p_{\mathbf{y};g}^{d(\mathbf{y}_{ab}, \mathbf{y})} \right)^{d(c_{a,b};g)} \right] \tag{7.1}$$

dove:

$$d(\eta; \theta) = \begin{cases} 1 & \text{se } \eta = \theta \\ 0 & \text{altrimenti.} \end{cases}$$

Il metodo di record linkage proposto da Larsen e Rubin si può dividere in due fasi. Nella prima, si "stima" il modello 7.1 fra un insieme di modelli plausibili. La seconda fase è la fase di decisione vera e propria.

7.2.1 La scelta del modello

La fase di scelta del modello si compone di 4 passi.

1. Il primo passo consiste nel decidere un insieme di modelli loglineari candidati a spiegare i dati (i confronti) a disposizione. Si vuole sottolineare che i modelli specificano due elementi.

   - Il numero di gruppi $G \geq 3$. Si suppone ancora che le coppie possano essere match o non-match, ma si ammette che i due gruppi $\mathcal{M}$ e $\mathcal{U}$ possano a loro volta dividersi in sottogruppi che hanno un comportamento omogeneo, ovvero che rispondono alla stessa distribuzione di probabilità.
   - Per ogni gruppo si suppone un modello loglineare.

   Si ipotizzi ci siano S modelli candidati.

2. Usando l'algoritmo EM (appendice B), si stimano i parametri $\{p_g^s\}$ e $\{p_{\mathbf{y};g}^s\}$ in (7.1), $s = 1, ..., S$. Siano $\{\hat{p}_g^s\}$ e $\{\hat{p}_{\mathbf{y};g}^s\}$, $s = 1, ..., S$, le stime corrispondenti.

3. Fra i diversi modelli proposti, $s = 1, ..., S$ si sceglie quello le cui probabilità stimate si avvicinano di più alle probabilità che Larsen e Rubin chiamano "semi-empiriche", e che indicano con il simbolo $\tilde{p}_g^s$ e $\tilde{p}_{\mathbf{i};g}^s$. Queste probabilità uniscono informazioni provenienti dai dati con opinioni personali, e quindi rappresentano valori plausibili delle probabilità cercate. Supponendo che $G = 3$ e che il gruppo $g = 1$ sia il gruppo dei match $\mathcal{M}$ (che non viene diviso) mentre il gruppo dei non-match $\mathcal{U}$ viene diviso in due sottogruppi, Larsen e Rubin danno i seguenti suggerimenti.

   - Un valore iniziale per $\tilde{p}_1$, (cioè per il gruppo dei match), può essere definito da una frazione del rapporto fra il numero di unità nel dataset più piccolo (cioè $N = \nu_A \wedge \nu_B$) ed il numero di coppie totali che si hanno (cioè $\nu_A \times \nu_B$). Questa scelta è dovuta al fatto che il numero massimo di abbinamenti non può superare il numero di unità presenti in ognuno dei due dataset.
   - Dato che le coppie che sono un match tendono a avere molte variabili di confronto uguali (cioè i vettori di confronto sono composti per lo più da valori 1), Larsen e Rubin consigliano di adattare la distribuzione $\tilde{p}_{\mathbf{y};1}$ per il gruppo dei match alle frequenze relative associate ai vettori di confronto con un numero elevato di 1. Le restanti due distribuzioni possono essere calibrate tenendo conto delle frequenze relative osservate sui vettori di confronto con un numero elevato di 0.

   Queste scelte si basano sui dati a disposizione sia su opinioni personali, ma non su dati provenienti da altri siti e controllati da impiegati, data la forte "site-to-site variability" (Winkler 1985a, 1985b, Arellano 1992, Belin 1993).

4. Si considera la distanza di Kullback-Leibler (si veda l'appendice C; in particolare si faccia riferimento a Bishop *et al*., 1975, p. 344-348, per l'estensione al caso delle tabelle di contingenza) fra la distribuzione semi-empirica e quella stimata al passo 2:

$$\Delta(\tilde{\mathbf{p}}, \hat{\mathbf{p}}) = \sum_{\mathbf{y} \in \mathcal{D}} \sum_{g=1}^{G} n_g \tilde{p}_g \tilde{p}_{\mathbf{y};g} \log \left(\frac{\tilde{p}_g \tilde{p}_{\mathbf{y};g}}{\hat{p}_g^s \hat{p}_{\mathbf{y};g}^s} \right), \qquad s = 1, ..., S,$$

per individuare quale fra gli S modelli candidati si avvicina di più alla situazione indicata dalle probabilità semiempiriche. Sia s^* il modello scelto.

Si sottolinea che una volta scelto il modello al passo 4, le stime dei parametri del modello (7.1) sono le stime di massima verosimiglianza ottenute al passo 2.

7.2.2 La fase decisionale: una procedura iterativa

Una volta scelto il modello per i dati a disposizione, Larsen e Rubin propongono una procedura iterativa, che combina fasi statistiche con fasi di revisione di dati attraverso impiegati. Sottostante questa procedura c'è l'ipotesi che il controllo manuale da parte degli impiegati determina senza errore lo status delle coppie.

Questa procedura può essere sintetizzata nei seguenti passaggi.

1. Si ordinano i vettori di confronto osservati $\mathbf{y}$ in senso decrescente rispetto alla stima di massima probabilità di appartenenza al gruppo degli abbinamenti (supponiamo $g = 1$)[1]:

$$\frac{\hat{p}_1^{s^*}\hat{p}_{\mathbf{y};1}^{s^*}}{\sum_{g=1}^{G}\hat{p}_g^{s^*}\hat{p}_{\mathbf{y};g}^{s^*}}. \tag{7.2}$$

2. Secondo il meccanismo di decisione di Fellegi-Sunter (si veda il capitolo 4) si decide se la coppia (a, b) è un match (A_m), un non-match (A_u) o un match incerto ($A_\emptyset$).

3. Fra le coppie il cui status è incerto, si selezionano le coppie con vettore di confronto $\mathbf{y}$ alle quali corrisponde una probabilità (7.2) vicina al livello di soglia ϕ. Queste coppie vengono sottoposte a revisione manuale con impiegati per una decisione definitiva. Larsen e Rubin suggeriscono di fissare ϕ pari al valore più piccolo fra i seguenti:

   - $80\%(\nu_A \wedge \nu_B)$;
   - la numerosità stimata attraverso il modello del gruppo di coppie che sono match $\mathcal{M}$.

   La scelta di questo livello di soglia è legata a considerazioni di ordine pratico. Se l'insieme dei match è molto più grande dell'ampiezza della lista più piccola, questo insieme conterrà molte coppie abbinate erroneamente. Se il gruppo dei match possiede un numero di coppie molto inferiore al numero di unità presenti nella lista più piccola, ci potrebbero essere molti abbinamenti che non sono stati individuati.

4. I parametri del modello s^* vengono ora di nuovo stimati tenendo conto di tutte le informazioni raccolte. In particolare alle coppie analizzate dagli impiegati viene assegnata probabilità di appartenenza al gruppo dei match pari a 1 se è stato appurato che la coppia è un match; viceversa viene assegnata probabilità 0. Per gli altri parametri si ricorre, come al solito, alla stima di massima verosimiglianza attraverso l'algoritmo EM. Il modello al quale si fa riferimento è leggermente diverso. Si indichi con $\mathcal{V}$ l'insieme delle coppie per le quali lo status è stato accertato manualmente, ovvero i valori $\{c_{a,b}, (a, b) \in \mathcal{V}\}$ sono ora conosciuti. I restanti valori $\{c_{a,b}, (a, b) \notin \mathcal{V}\}$ vengono considerati ancora incogniti (in pratica non vengono tenute in considerazione le decisioni sulle coppie prese dalla procedura

[1] Come avviene in Torelli (1998), si veda anche il paragrafo 4.2

di record linkage). La funzione di densità dei dati che si deve analizzare ora si compone di due parti:

$$\prod_{(a,b)\in\mathcal{V}}\prod_{g=1}^{G}\left(p_g\left[\prod_{\mathbf{y}\in\mathcal{D}}p_{\mathbf{y};g}^{d(\mathbf{y}_{ab},\mathbf{y})}\right]\right)^{d(c_{ab},g)}\prod_{(a,b)\notin\mathcal{V}}\prod_{g=1}^{G}\left(p_g\left[\prod_{\mathbf{y}\in\mathcal{D}}p_{\mathbf{y};g}^{d(\mathbf{y}_{ab},\mathbf{y})}\right]\right)^{d(c_{ab},g)} \tag{7.3}$$

dove la differenza sta nel fatto che i valori $c_{a,b}$ nella seconda parte di (7.3) devono essere considerati valori mancanti.

5. Si svolgono i passi precedenti in modo iterativo (adattamento del modello; decisione A_m, A_u o $A_\emptyset$ per le coppie il cui status è incerto; analisi di alcune coppie fra quelle il cui status rimane incerto da parte degli impiegati; nuovo adattamento del modello).

6. La regola di arresto che si prende in considerazione è la seguente: l'algoritmo si ferma quando il numero di match che viene determinato dall'analisi degli impiegati decresce fino a rappresentare una frazione piccola del numero totale di coppie analizzate dagli impiegati. Una volta che si è deciso di fermare la procedura iterativa, il modello s^* viene sottoposto a stima per l'ultima volta usando tutte le informazioni sullo status delle coppie che sono state analizzate manualmente. Le coppie non ancora classificate vengono ordinate in modo decrescente rispetto alla loro probabilità di essere un match. Queste coppie vengono dichiarate rispettivamente match o non-match fino a raggiungere i tassi di errore specificati. I casi che non si riesce ad assegnare al gruppo degli abbinamenti o dei non abbinamenti vengono sottoposti al controllo da parte di impiegati specializzati.

Come si vede, la metodologia proposta da Larsen e Rubin necessita solo di alcuni input (formalizzati nelle probabilità semiempiriche) e del lavoro da parte degli impiegati per un limitato numero di coppie. Non c'è bisogno di *training sample,* e questo è una garanzia contro la site-to-site variability. L'uso degli impiegati e del model fitting fatto in modo iterativo assicura che i risultati finali siano robusti rispetto a ipotesi iniziali che possono essere sbagliate.

Commento 7.2 *La differenza sostanziale con il metodo di Fellegi e Sunter riguarda l'uso della revisione manuale delle coppie il cui status è incerto fatto in modo iterativo, in modo da affinare sempre più le stime dei parametri del modello (7.1) e in modo che l'insieme delle coppie il cui status è incerto viene minimizzato. La differenza con il metodo di Armstrong e Mayda riguarda invece la stima dei parametri del modello: ora si considerano stime di massima verosimiglianza del modello, dopo aver stimato un adeguato modello di dipendenza fra i confronti Y^h delle variabili chiave, tenendo conto di tutte le coppie.* □

7.3 Un approccio Bayesiano, Fortini *et al*. (2001)

La maggior parte dei metodi per il record linkage finora descritti fanno essenzialmente riferimento all'uso della *funzione di verosimiglianza* (ovvero della conoscenza di $m(.)$ e $u(.)$) per decidere a quale gruppo $\mathcal{M}$ o $\mathcal{U}$ appartengono le diverse coppie (a, b). Come visto con i pesi (4.2), se una coppia (a, b) presenta un vettore di confronto $\mathbf{y}_{ab}$ che è più verosimile osservare fra le coppie in $\mathcal{M}$ piuttosto che fra le coppie in $\mathcal{U}$, si decide di considerare la coppia come un match (e viceversa). Il "parametro" di interesse è quindi la matrice $\mathbf{c}$ che descrive per ogni coppia (a, b) lo status di appartenenza a uno dei due gruppi $\mathcal{M}$ e $\mathcal{U}$.

Nella logica Bayesiana si riuniscono tutte le informazioni sul parametro di interesse, $\mathbf{c}$, descrivendo in questo modo una *distribuzione a priori* sui valori che $\mathbf{c}$ può assumere. In base alle osservazioni, ovvero ai vettori dei confronti $\mathbf{y}_{ab}$ osservati sulle $\nu_A \times \nu_B$ coppie, si aggiorna la distribuzione a priori su $\mathbf{c}$ attraverso il teorema di Bayes, ottenendo la *distribuzione a posteriori* per $\mathbf{c}$. Su questa distribuzione si basano le possibili inferenze sul valore più plausibile per $\mathbf{c}$.

Le motivazioni che hanno portato a formulare una procedura Bayesiana per il record linkage sono essenzialmente due:

- in una procedura bayesiana è naturale descrivere per ogni coppia (a, b) qual è la probabilità che la coppia stessa sia un match, condizionatamente ai dati osservati $\mathbf{y}_{ab}$;
- in una procedura bayesiana è naturale definire la probabilità che *congiuntamente* alcune coppie siano match oppure no.

La prima affermazione è di ordine pratico, e coincide con la possibilità di interpretare più facilmente l'output del record linkage. Si sottolinea che non è necessario lavorare in un ambito strettamente bayesiano per poter definire queste probabilità (basti pensare ai pesi usati da Torelli, paragrafo 4.2, e da Larsen e Rubin, paragrafo 7.2). Ma al contrario di quanto accade nei due riferimenti precedenti, la probabilità che una coppia sia match dato un certo pattern di confronto $\mathbf{y}$ non corrisponde alla frequenza relativa delle coppie che sono match fra tutte quelle che presentano il confronto $\mathbf{y}$, ma incorpora le informazioni a disposizione sul possibile stato di ogni coppia. La seconda caratterizzazione è invece specifica dell'approccio bayesiano e risulta estremamente utile. Infatti nel capitolo 5 si è visto che è necessario far riferimento a procedure di ricerca operativa per poter rispettare i vincoli (2.5), (2.6) e (2.7). In un contesto bayesiano, questi vincoli sono definiti *in modo naturale* all'interno della procedura stessa descrivendo opportunamente il supporto della distribuzione di probabilità congiunta che più coppie siano match o non-match.

7.3.1 Le distribuzioni a priori

Senza perdere in generalità, si considerino confronti del tipo (2.10) (ovvero il confronto fra i valori di una variabile chiave su due unità può dare due soli risultati: uguale, 1, o diverso, 0). Sia $\mathcal{D}$ l'insieme dei vettori di confronto $\mathbf{y}$ possibili, ovvero l'insieme dei 2^k vettori di k elementi composti solo da valori 0 e 1. In questo contesto, invece della verosimiglianza (6.4), che modellizza la distribuzione congiunta di osservare $\mathbf{Y}_{ab} = \mathbf{y}_{ab}$ e $C_{a,b} = c_{a,b}$ per ogni coppia (a, b), si consideri la seguente verosimiglianza:

$$
\begin{aligned}
& L\Big(\mathbf{c}, \{m(.)\}, \{u(.)\} \Big| \mathbf{y}_{ab}, (a,b) \in \mathcal{A} \times \mathcal{B}\Big) = \\
& = \prod_{(a,b)} \Big(m(\mathbf{y})^{d(\mathbf{y},\mathbf{y}_{ab})}\Big)^{c_{a,b}} \Big(u(\mathbf{y})^{d(\mathbf{y},\mathbf{y}_{ab})}\Big)^{1-c_{a,b}} = \qquad (7.4) \\
& = \prod_{\mathbf{y}\in\mathcal{D}} m(\mathbf{y})^{\sum_{(a,b)} d(\mathbf{y},\mathbf{y}_{ab}) c_{a,b}} u(\mathbf{y})^{\sum_{(a,b)} d(\mathbf{y},\mathbf{y}_{ab})(1-c_{a,b})} \qquad (7.5)
\end{aligned}
$$

dove

$$
d(\mathbf{y}, \mathbf{y}_{ab}) = \begin{cases} 1 & \text{se } \mathbf{y}_{ab} = \mathbf{y} \\ 0 & \text{altrimenti.} \end{cases}
$$

La differenza fra (6.4) e (7.5) sta nel ruolo assegnato a $\mathbf{c}$: nella seconda diventa un elemento attraverso il quale massimizzare la funzione di verosimiglianza. In pratica la verosimiglianza

(7.5) permette di verificare quali configurazioni sullo status delle coppie sono più verosimili in base alle osservazioni dei confronti $\mathbf{y}_{ab}$ sulle $\nu_A \times \nu_B$ coppie. Un problema indotto dalla (7.5) è che questa dipende anche da altri parametri, le distribuzioni $m(.)$ e $u(.)$, in genere incognite come detto nel capitolo 4. Queste distribuzioni, nel presente contesto, assumono il ruolo di *parametri di disturbo*.

Una distribuzione a priori per la matrice c

Sia $\mathbf{C}$ la variabile aleatoria che descrive l'incertezza sul valore della matrice $\mathbf{c}$. Una distribuzione a priori per $\mathbf{C}$ può essere definita in due passi. Il primo passo consiste nel definire una distribuzione a priori sul numero di match, cioè sulla cardinalità di $\mathcal{M}$. Sia

$$\pi_H(h), \qquad h = 0, 1, ..., N = \min\{\nu_A, \nu_B\}$$

tale distribuzione. Questa può spesso essere formalizzata, facendo riferimento a esperienze precedenti o alle caratteristiche statistiche delle basi dati A e B che si stanno studiando.

Il secondo passo consiste nel definire la distribuzione condizionata delle configurazioni $\mathbf{c}$ dato il numero di match $h = 0, 1, ..., N$, cioè:

$$P(\mathbf{C} = \mathbf{c}|H = h), \qquad \mathbf{c} \in \mathcal{C}; \;\; h = 0, 1, ..., N.$$

La distribuzione a priori per $\mathbf{C}$ è quindi definita dalla relazione:

$$P(\mathbf{C} = \mathbf{c}) = \pi_H(h) P(\mathbf{C} = \mathbf{c}|H = h)$$

dove l'uguaglianza vale in quanto, per h opportuno,

$$P(\mathbf{C} = \mathbf{c}) = P(\mathbf{C} = \mathbf{c}, H = h).$$

Si sezioni l'insieme delle matrici $\mathbf{c} \in \mathcal{C}$ nei sottoinsiemi contenenti matrici con lo stesso numero di match:

$$\mathcal{C}^h = \left\{ \mathbf{c} \in \mathcal{C} : \sum_{(a,b)} c_{a,b} = h \right\}, \qquad h = 0, 1, ..., N.$$

La scelta fatta da Fortini *et al.* (2001) per queste distribuzioni a priori è la seguente. Si ipotizza che la distribuzione a priori per $\mathbf{C}|(H = h)$ sia uniforme sull'insieme $\mathcal{C}^h$. Infatti in genere non si hanno informazioni sufficienti per favorire alcune matrici $\mathbf{c}$ rispetto ad altre all'interno dello stesso insieme $\mathcal{C}^h$. Al contrario, la distribuzione di H si ipotizza che sia una binomiale di parametro ξ. Questo parametro deve essere calibrato in modo che la distribuzione di H renda più plausibili il numero di match che le informazioni a priori favoriscono. Ad esempio, ξ può essere scelto come la frequenza relativa dei match osservati in occasioni simili.

Le distribuzioni a priori per i parametri di disturbo

In questo contesto, le due distribuzioni $m(.)$ e $u(.)$ sono due multinomiali con parametri:

$$m(\mathbf{y}), \qquad \sum_{\mathbf{y} \in \mathcal{D}} m(\mathbf{y}) = 1,$$

$$u(\mathbf{y}), \qquad \sum_{\mathbf{y}\in\mathcal{D}} u(\mathbf{y}) = 1.$$

Queste distribuzioni si ritengono generate da due vettori aleatori $\mathbf{M}$ e $\mathbf{U}$ a priori indipendenti da $\mathbf{C}$. Una distribuzione a priori appropriata per questi parametri incogniti è la distribuzione di Dirichlet[2]:

$$\mathbf{M} \sim D_{|\mathcal{D}|-1}(.;\boldsymbol{\alpha})$$
$$\mathbf{U} \sim D_{|\mathcal{D}|-1}(.;\boldsymbol{\beta}),$$

con $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ vettori degli iperparametri[3] delle due distribuzioni di Dirichlet. Questi vettori di iperparametri devono soddisfare la condizione:

$$\alpha_{\mathbf{y}} > 0,\ \beta_{\mathbf{y}} > 0, \qquad \mathbf{y} \in \mathcal{D}.$$

La calibrazione degli iperparametri è estremamente importante. Fortini *et al*. (2001) hanno considerato nelle loro sperimentazioni i seguenti valori:

$$\alpha_{\mathbf{y}} = \theta^{\sum_{h=1}^{k} y_h - \phi}, \qquad \mathbf{y} \in \mathcal{D},\ \theta > 0,\ \phi \in I\!R$$

per il vettore $\boldsymbol{\alpha}$, e

$$\beta_{\mathbf{y}} = \theta^{\phi - \sum_{h=1}^{k} y_h}, \quad \mathbf{y} \in \mathcal{D},\ \theta > 0,\ \phi \in I\!R,$$

per il vettore $\boldsymbol{\beta}$. Questa scelta, da un lato riduce il numero di iperparametri a 2, ovvero a θ e ϕ. Inoltre è in grado di ben rappresentare l'informazione discriminante contenuta nei confronti fra le variabili chiave. Ad esempio, gli iperparametri $\alpha_{\mathbf{y}}$ ordinano gerarchicamente le osservazioni possibili $\mathbf{y} \in \mathcal{D}$ in modo che la distribuzione a priori per $\mathbf{M}$ pone maggior probabilità su valori alti per i parametri $m_{\mathbf{y}}$ caratterizzati da vettori $\mathbf{y}$ con una elevata presenza di 1. L'argomento opposto è valido per la distribuzione a priori $\mathbf{U}$.

7.3.2 L'analisi a posteriori

L'analisi a posteriori richiede il calcolo della distribuzione a posteriori per la variabile aleatoria $\mathbf{C}$. Il primo passo per ottenere questa distribuzione consiste nel marginalizzare la verosimiglianza (7.5) rispetto ai parametri di disturbo. Questa operazione porta alla verosimiglianza:

$$L\Big(\mathbf{c}\Big|\mathbf{y}_{ab},\ (a,b) \in \mathcal{A}\times\mathcal{B}\Big) \propto$$
$$\propto \frac{\prod_{\mathbf{y}\in\mathcal{D}} \Gamma\left(\sum_{(a,b)}[d(\mathbf{y},\mathbf{y}_{ab})c_{a,b}] + \alpha_{\mathbf{y}}\right)\Gamma\left(\sum_{(a,b)}[d(\mathbf{y},\mathbf{y}_{ab})(1-c_{a,b})] + \beta_{\mathbf{y}}\right)}{\Gamma\left(h + \sum_{\mathbf{y}\in\mathcal{D}}\alpha_{\mathbf{y}}\right)\Gamma\left(N - h + \sum_{\mathbf{y}\in\mathcal{D}}\beta_{\mathbf{y}}\right)} \tag{7.6}$$

dove h è il numero di match indicati dalla matrice $\mathbf{c}$ ovvero:

$$h = \sum_{(a,b)} c_{a,b}.$$

[2]Questa è infatti la distribuzione *coniugata* per le distribuzioni multinomiali, in proposito si veda Bernardo e Smith (1994) . La scelta di una distribuzione coniugata per i parametri incogniti semplifica i calcoli per determinare la distribuzione a posteriori. Per una definizione della distribuzione di Dirichlet si veda Kotz, Balakrishnan e Johnson (2000).

[3]I parametri delle distribuzioni a priori prendono il nome di iperparametri

L'applicazione del teorema di Bayes porta alla distribuzione a posteriori per $\mathbf{C}$, che risulta essere proporzionale a:

$$P\Big(\mathbf{C}=\mathbf{c}\Big|\mathbf{y}_{ab},\ (a,b)\in\mathcal{A}\times\mathcal{B}\Big)\propto\pi_H(h)P(\mathbf{C}=\mathbf{c}|H=h)L\Big(\mathbf{c}\Big|\mathbf{y}_{ab},\ (a,b)\in\mathcal{A}\times\mathcal{B}\Big).$$

La funzione precedente è il risultato principale della procedura bayesiana di record linkage. La determinazione di questo output non è semplice, in quanto è necessario usare delle procedure numeriche per calcolare i parametri della (7.6)[4]. Da questo output si possono considerare diverse sintesi, utili a definire una "stima puntuale" per $\mathbf{c}$. Un esempio è dato dalla configurazione $\mathbf{c}^*$ che rende massima la distribuzione a posteriori.

Commento 7.3 *Una procedura bayesiana può apparire in contrasto con gli obiettivi della statistica ufficiale, in quanto comporta l'uso di elementi "soggettivi" (le distribuzioni a priori) nelle analisi. Al contrario, le distribuzioni a priori devono essere viste come l'occasione per poter immettere nelle procedure statistiche (come quelle di record linkage) le conoscenze di un Istituto di Statistica, derivanti ad esempio dalle esperienze passate.* □

[4]In proposito si rimanda ai lavori di Fortini *et al*. (2000) e alla tesi di dottorato di Nuccitelli (2001)

Capitolo 8

La qualità del record linkage

Come visto nelle sezioni precedenti, le procedure di record linkage producono due tipi di errore: i falsi match e i falsi non-match. La valutazione dell'incidenza di questi errori nell'output del procedimento di record linkage consente di misurarne la qualità. Fin dal primo capitolo si è sottolineato che se si adotta un procedimento statistico, e in particolare un modello per la stima dei parametri (ovvero quelli del paragrafo 6.2 e successivi), e il modello statistico è adeguato, si ottiene facilmente un'indicazione della "qualità attesa". Inoltre sono stati definiti dei metodi adatti per procedure ad hoc di record linkage. Nei prossimi paragrafi si analizzano ambedue gli approcci, iniziando con le procedure ad hoc.

8.1 Il tasso di errato abbinamento: FMR

Si supponga di essere nel caso in cui le $\nu_A \times \nu_B$ osservazioni non fanno riferimento a nessun modello di generazione dei dati ovvero, nelle interpretazioni fornite nel commento 3.1, quando $\nu_A = N_A$ e $\nu_B = N_B$. Si è già detto nel commento 4.5 che i livelli di errore μ e λ che caratterizzano il record linkage assumono il significato di "frequenze di errore" attese nell'insieme delle coppie abbinate e non abbinate. Per poter calcolare queste frequenze è però necessario avere informazioni sulle distribuzioni di frequenze $m(.)$ e $u(.)$ sulle $\nu_A \times \nu_B$ coppie (non note quando si adottano procedure ad hoc).

Un modo equivalente per valutare l'incidenza degli errori generati dalla procedura di record linkage è dato da due semplici rapporti che non fanno riferimento esplicito alle distribuzioni di frequenze dei confronti: il tasso di errato abbinamento, spesso abbreviato in FMR (*False Match Rate* in inglese) e il tasso di errato non abbinamento (per semplicità FNR). La loro definizione è semplice. Indicando con $\hat{c}_{a,b}$ l'indicatore che vale uno se la procedura di record linkage stabilisce che la coppia (a, b) è un match e zero quando stabilisce che la coppia è un non-match (in contrapposizione a $c_{a,b}$ che indica se effettivamente la coppia è un match, ovvero se $(a, b) \in \mathcal{M}$), il tasso di errato abbinamento è definito dal rapporto:

$$FMR = \frac{\sum_{(a,b)\in\mathcal{U}} \hat{c}_{a,b}}{\sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} \hat{c}_{a,b}} \tag{8.1}$$

ovvero è il numero di falsi match sul totale dei match dichiarati dalla procedura di record linkage. Allo stesso modo il tasso FNR è definito da:

$$FNR = \frac{\sum_{(a,b)\in\mathcal{M}} (1 - \hat{c}_{a,b})}{\nu_A \times \nu_B - \sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} \hat{c}_{a,b}}. \tag{8.2}$$

Dato che questi tassi dipendono dalla vera ma incognita bipartizione di $\mathcal{A} \times \mathcal{B}$ in $\mathcal{M}$ e $\mathcal{U}$, si è costretti a confrontare i risultati della procedura di record linkage adottata e del record linkage manuale su un campione di coppie, supponendo che l'abbinamento manuale individui senza errori lo status di ogni coppia. Da questo confronto si possono ottenere delle stime di FMR e FNR. Un esempio di una procedura di questo tipo, utile a calibrare un valore di soglia in una procedura di record linkage "ad hoc" (come quella del paragrafo 7.1) viene descritta nel paragrafo seguente.

Commento 8.1 *I tassi FMR e FNR descrivono l'incidenza degli errori sull'output del record linkage. A volte può essere utile definire questi tassi in modo leggermente diverso, calcolando l'incidenza degli errori rispetto alla configurazione* **c** *vera. In particolare i tassi:*

$$\widetilde{FMR} = \frac{\sum_{(a,b)\in\mathcal{U}} \hat{c}_{a,b}}{\sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} c_{a,b}}$$

$$\widetilde{FNR} = \frac{\sum_{(a,b)\in\mathcal{M}} (1 - \hat{c}_{a,b})}{\nu_A \times \nu_B - \sum_{(a,b)\in\mathcal{A}\times\mathcal{B}} c_{a,b}}$$

valutano rispettivamente il numero di falsi match (falsi non-match) sul totale dei veri match (veri non-match). Quando vengono definiti modelli per la generazione dei dati, condizionatamente al numero dei match e dei non match, questi tassi sono aleatori solo nel numeratore, al contrario di quanto accade per FMR e FNR. □

Commento 8.2 *Come già detto all'inizio del paragrafo, esiste una relazione diretta fra i tassi μ e λ e i tassi FMR e FNR (ovvero $\widetilde{FMR}$ e $\widetilde{FNR}$). La definizione del tasso μ è:*

$$\mu = \sum_{\mathbf{y}\in\mathcal{D}} u(\mathbf{y}) P(A_m|\mathbf{y})$$

e quando non si adottano procedure di randomizzazione per la decisione (come nel caso della tabella 4.1) il numeratore di μ coincide con quello di FMR e di $\widetilde{FMR}$. Quindi questi tassi sono legati dalla relazione:

$$\mu = \frac{\sum_{(a,b)} \hat{c}_{a,b}}{\sum_{(a,b)} (1 - c_{a,b})} FMR = \frac{\sum_{(a,b)} c_{a,b}}{\sum_{(a,b)} (1 - c_{a,b})} \widetilde{FMR}.$$

Lo stesso vale per λ:

$$\lambda = \frac{\nu_A \times \nu_B - \sum_{(a,b)} \hat{c}_{a,b}}{\sum_{(a,b)} c_{a,b}} FNR = \frac{\nu_A \times \nu_B - \sum_{(a,b)} c_{a,b}}{\sum_{(a,b)} c_{a,b}} \widetilde{FNR}.$$

□

8.1.1 Un esempio di stima di FMR e FNR attraverso controllo manuale

Si consideri una procedura di record linkage caratterizzata da un unico valore di soglia $\tau_\lambda = \tau_\mu = \tau$, in modo tale che vengano esclusi i match incerti (i match derivanti dalla decisione $A_\emptyset$, si veda il capitolo 4). In questo caso il numero di falsi positivi e il numero di falsi negativi sono legati da una relazione inversa al variare della soglia τ. Infatti, al crescere di τ diminuisce

Tabella 8.1 - Numero di coppie classificate secondo lo status accertato con l'abbinamento manuale e secondo l'abbinamento proposto dalla procedura di record linkage computerizzato

	Record Linkage Manuale	
Record Linkage Computerizzato	Match	Non-match
Abbinati	n_{11}	n_{12}
Non-abbinati	n_{21}	n_{22}

il numero di falsi positivi e aumenta il numero di falsi negativi. Siano $\widetilde{FMR}(\tau)$ e $\widetilde{FNR}(\tau)$ le funzioni dei tassi $\widetilde{FMR}$ e $\widetilde{FNR}$ al variare della soglia τ. Naturalmente la funzione $\widetilde{FMR}(\tau)$ è una funzione non crescente, mentre $\widetilde{FNR}(\tau)$ è una funzione non decrescente al crescere di τ.

Un discorso analogo non è immediatamente valido per $FMR(\tau)$ e $FNR(\tau)$, dato che queste funzioni al variare di τ si modificano sia al numeratore che al denominatore. Nonostante questa considerazione, è plausibile ritenere che anche $FMR(\tau)$ e $FNR(\tau)$ abbiano un andamento simile, come ipotizzano ad esempio Bartlett *et al.* (1993). Questi autori suggeriscono di scegliere la soglia τ^* che minimizza contemporaneamente FMR e FNR, ovvero il valore τ^* determinato dall'incrocio fra $FMR(\tau)$ e $FNR(\tau)$. La stima di questi tassi al variare di τ diventa quindi essenziale.

Si selezionino n coppie fra le $\nu_A \times \nu_B$ in $\mathcal{A} \times \mathcal{B}$. Questo campione viene analizzato sia manualmente che attraverso una procedura computerizzata di record linkage. Osservando "congiuntamente" i risultati della procedura di record linkage computerizzata e di quella manuale si può ricostruire la tabella 8.1.

Questa tabella consente il calcolo delle seguenti stime di FMR e FNR:

$$\widehat{FMR} = \frac{n_{12}}{n_{11} + n_{12}}$$

$$\widehat{FNR} = \frac{n_{21}}{n_{21} + n_{22}}.$$

In modo equivalente si possono costruire delle stime per i tassi $\widetilde{FMR}$ e $\widetilde{FNR}$.

Esempio 8.1 *Bartlett* et al. *(1993) descrivono alcuni esempi di record linkage dove il calcolo della soglia e dei livelli di errore ad essa associata sono calcolati secondo la procedura appena illustrata. In uno di questi esempi, si era posto il problema di abbinare attraverso record linkage i dati del Farm Operator Cohort Data Base (FOCDB), contenente variabili socio-demografiche degli agricoltori, con il Canadian Mortality Data Base (CMDB), contenente i record di tutti i decessi registrati. Di tutte le coppie di record ottenibili dal confronto delle due basi di dati ne sono state analizzate 21300. Tramite il controllo manuale su queste coppie si è appurato che 455 sono veri match mentre 20845 no. Sulle stesse coppie è stata applicata una procedura di record linkage probabilistico il cui software è stato sviluppato presso Statistics Canada. Supponendo che i risultati dell'abbinamento manuale siano veri, il confronto fra i risultati delle due procedure di abbinamento permette di costruire due distribuzioni di frequenze: quella del numero di falsi positivi e quella dei falsi negativi al variare della soglia τ. Queste due curve hanno andamenti opposti: crescente per il numero di falsi negativi e decrescente per il numero di falsi positivi. Il punto in cui le due curve si incontrano indica il valore soglia assunto dagli autori. Infatti se*

Tabella 8.2 - Numero di coppie classificate secondo lo status accertato con l'abbinamento manuale (RL manuale) e secondo l'abbinamento proposto dalla procedura di record linkage computarizzato (CRL) per l'esempio FOCDB/CMDB (Bartlett *et al*., 1993).

	Record Linkage Manuale		
Record Linkage Computerizzato	Match	Non-match	**Totale**
Abbinati	417	36	**453**
Non-abbinati	38	20809	**20847**
Totale	**455**	**20845**	**21300**

viene considerato un valore più piccolo, bisogna sopportare un numero superiore di falsi positivi, mentre un valore più grande di τ genera un numero superiore di falsi negativi sulle 21300 coppie sotto studio. In base alla soglia τ prescelta viene poi costruita la tabella 8.2. Dalla tabella si evince che i tassi di errore sono rispettivamente il 7,9% per FMR e lo 0,2 % per FNR. Bartlett et al. *(1993) verificano nei loro esempi che aumentando le informazioni per il record linkage, ovvero usando più variabili chiave, le stime dei due tassi di errore si riducono. Non è detto che questa affermazione sia sempre valida.* □

Commento 8.3 *Sarebbe preferibile poter estrarre un campione di coppie in $\mathcal{A}\times\mathcal{B}$ utile al calcolo delle stime FMR e FNR secondo un appropriato piano di campionamento. Questo fatto non è mai stato sottolineato dai diversi studi sul record linkage, tranne che da Paggiaro e Torelli (1999). Nel loro lavoro si sottolinea che fra le coppie dichiarate match sarebbe preferibile estrarre quelle con peso più basso, in quanto queste sono maggiormente candidate a essere dei falsi match. In pratica, è come se si proponesse di usare il peso $t(\mathbf{y})$ (o equivalentemente il peso $t^*(\mathbf{y})$) come variabile ausiliaria.* □

8.2 Stima dei tassi di errore basata su modello

Si supponga di avere ν_A unità in $\mathcal{A}$ e ν_B unità in $\mathcal{B}$ tali che N di questi siano match (ovvero si ragiona condizionatamente al meccanismo generatore della coppia, secondo quanto detto nel paragrafo 6.2). Supponiamo inoltre che sulle ν_A unità di $\mathcal{A}$ e ν_B unità di $\mathcal{B}$ si applichi un "meccanismo generatore degli errori", come evidenziato nel commento 3.1. In questo contesto il parametro μ rappresenta la probabilità con cui una coppia che è un non-match viene dichiarata match. Nel caso di assenza di randomizzazione nelle decisioni si ha:

$$\mu = P\Big(t(\mathbf{Y}) \geq \tau_\mu \Big| c = 0\Big).$$

Un significato equivalente viene assunto da λ:

$$\lambda = P\Big(t(\mathbf{Y}) \leq \tau_\lambda \Big| c = 1\Big).$$

Siano $\hat{m}(\mathbf{y})$ e $\hat{u}(\mathbf{y})$, $\mathbf{y} \in \mathcal{D}$, stime opportune delle distribuzioni $m(\mathbf{y})$ e $u(\mathbf{y})$, $\mathbf{y} \in \mathcal{D}$, ottenute ad esempio attraverso uno dei metodi del capitolo 6. Sostituendo queste stime alle definizioni di μ e λ si calcolano le stime seguenti:

$$\hat{\mu} = \sum_{\mathbf{y}\in\mathcal{D}} \hat{u}(\mathbf{y}) P(A_m|\mathbf{y})$$

$$\hat{\lambda} = \sum_{\mathbf{y} \in \mathcal{D}} \hat{m}(\mathbf{y}) P(A_u|\mathbf{y}).$$

Commento 8.4 *L'uso del modello per il calcolo dei livelli di errore è favorito da diversi autori, ad esempio da Armstrong e Mayda (1993), in quanto il campionamento di coppie da $\mathcal{A} \times \mathcal{B}$ e il successivo controllo manuale è costoso e lento. Armstrong a Mayda verificano su dati simulati e su dati reali la bontà dei livelli di errore relativamente a tre diverse situazioni. Le prime due sono basate su modello: il metodo proposto da Fellegi e Sunter (paragrafo 6.3.1), basato sull'indipendenza fra i confronti Y^h delle variabili chiave; il modello loglineare con variabili latenti discusso nel paragrafo (6.4.1). Il terzo è relativo alla procedura ad hoc descritta nel paragrafo 7.1. Il risultato è che le stime basate su modello possono essere particolarmente buone. Inoltre incorporare le dipendenze attraverso un modello loglineare rende più affidabili le stime degli errori, ottenute sia su casi simulati che su casi reali. Naturalmente se il modello considerato è errato (soprattutto per quanto riguarda la possibilità di dipendenza statistica fra i confronti Y^h delle variabili chiave) anche le stime dei tassi di errore μ e λ risultano poco affidabili.* □

Queste considerazioni possono essere generalizzate anche al caso in cui si suppone che esista un meccanismo generatore delle coppie C: in pratica non consideriamo più il caso in cui si hanno ν_A unità in $\mathcal{A}$ e ν_B unità in $\mathcal{B}$ con N unità in comune, come nel paragrafo 6.2.

Torelli e Paggiaro (1999) suggeriscono un parametro (Q) che può essere visto in relazione al tasso FMR:

$$Q = \sum_{t^*(\mathbf{y}_{ab}) > \tau_\mu^*} \Big(1 - t^*(\mathbf{y}_{ab})\Big),$$

dove:

$$1 - t^*(\mathbf{y}_{ab}) = P(C_{a,b} = 0 | \mathbf{Y}_{ab} = \mathbf{y}_{ab}).$$

Il numeratore di FMR descrive il numero di non-match (ovvero il numero di coppie in $\mathcal{U}$) dichiarati match dalla procedura di record linkage (numero di match falsi). Q descrive il numero atteso di coppie che sono non-match fra quelle dichiarate match:

$$Q = E\left[\sum (1 - C_{a,b}) \middle| \mathbf{Y}_{ab} = \mathbf{y}_{ab},\ \forall\ (a,b)\right],$$

dove la somma è estesa alle coppie i cui vettori di confronto $t^*(\mathbf{y})$ assumono un valore superiore alla soglia τ_μ^*.

Belin e Rubin (1995) definiscono una misura leggermente diversa. Per loro la misura della qualità del record linkage è data dalla probabilità che vengano generate coppie che sono non-match quando queste coppie presentano un peso superiore a τ_μ (ovvero un peso al quale è associata la decisione A_m):

$$P\Big(C = 0 \Big| t(\mathbf{Y}) \geq \tau_\mu\Big). \tag{8.3}$$

In questo modo, sapendo come funzionano i meccanismi generatori delle coppie e degli errori, si ha la possibilità di calibrare opportunamente il livello di soglia τ_μ. Belin e Rubin propongono un modo per stimare le distribuzioni utili al calcolo della probabilità (8.3), avendo a disposizione dei campioni di prova[1] (in proposito si veda il paragrafo 8.2.1).

[1] Winkler e Thibaudeau (1991) affermano che il procedimento proposto da Belin e Rubin è adeguato per il caso in cui le distribuzioni di $\mathbf{Y}|C = 1$ e $\mathbf{Y}|C = 0$ sono molto separate. In altri casi non porta ai risultati sperati.

Commento 8.5 *Come nel commento 4.6, i parametri (8.3) e Q possono assumere un significato alternativo, più vicino alle $\nu_A \times \nu_B$ coppie che si stanno analizzando. In questi casi il parametro (8.3) indica la frequenza delle coppie che sono non-match fra quelle che vengono dichiarate match. Per questo motivo la stima della (8.3) viene chiamata anche "stima di FMR". Il parametro Q è invece pari alla somma delle frequenze relative dei falsi match condizionatamente a ogni peso superiore alla soglia τ.* □

8.2.1 La calibrazione della qualità del record linkage: Belin e Rubin (1995)

Belin (1989) ha mostrato che il modo di stimare il tasso FMR definito da Fellegi e Sunter funziona in modo poco soddisfacente in quanto la loro procedura si basa in modo essenziale sull'ipotesi di indipendenza fra le variabili di confronto, ipotesi contestata da molti (capitolo 6).

Belin e Rubin propongono quindi un metodo per la scelta delle decisioni da prendere sulle diverse coppie, in modo da rispettare opportuni livelli di qualità del record linkage. Ciò viene fatto assegnando a ogni livello di soglia (i valori τ_λ e τ_μ) un livello di errore stimato. Il metodo da loro proposto può essere visto come un passaggio successivo da aggiungere a quello della determinazione dei pesi (rapporti di verosimiglianza) associati a ogni coppia. Quindi Belin e Rubin suppongono che questi pesi $t(\mathbf{y}_{ab})$ siano già stati assegnati a ogni coppia $(a, b) \in \mathcal{A} \times \mathcal{B}$. Inoltre fanno le seguenti ipotesi.

- Le osservazioni $t(.)$ sono determinazioni di una v.a. T che è una mistura di 2 distribuzioni di probabilità assolutamente continue: una relativa alle coppie che sono abbinamenti e l'altra relativa alle coppie composte da unità diverse. La probabilità di appartenere all'uno o l'altro gruppo viene posta uguale a p.

- Attraverso due opportune trasformazioni (una per le coppie che sono abbinamenti, l'altra per le coppie che non lo sono) dei pesi $t(\mathbf{y}_{ab})$, si ottengono dei nuovi valori w_{ab} provenienti da distribuzioni normali. Le trasformazioni che Belin e Rubin adottano e consigliano sono quelle appartenenti alla classe di funzioni definita da Box e Cox. Queste classi di funzioni sono definite sull'insieme $\mathcal{T}$ dei valori assumibili dal peso $t(.)$, caratterizzate da due parametri, $\gamma \in I\!R$ e $\omega > 0$:

$$w = \psi(t; \gamma, \omega) = \begin{cases} (t^\gamma - 1)/(\gamma \omega^{\gamma - 1}) & \text{se } \gamma \neq 0 \\ \omega \, \log(t) & \text{se } \gamma = 0 \end{cases} \tag{8.4}$$

  (per una definizione più generale di questa classe di funzioni si consideri Box *et al.*, 1964). È da notare che le funzioni $\psi(t; \gamma, \omega)$ sono monotone crescenti rispetto a t, e quindi conservano l'ordinamento delle coppie indotto dai pesi $t(\mathbf{y}_{ab})$.

Belin e Rubin suppongono anche che sia noto un campione di prova di coppie per le quali l'appartenenza a $\mathcal{M}$ o $\mathcal{U}$ è conosciuta. I passi da seguire in questo caso sono i seguenti.

- Si utilizzi il campione di prova per ottenere i parametri della trasformazione in (8.4). In pratica il campione di prova fornisce i parametri γ_m, ω_m per la trasformazione della distribuzione dei confronti delle coppie che sono un match, e i parametri γ_u e ω_u per l'altra distribuzione. Sia $f_m(w|\theta_m, \sigma_m^2)$ la distribuzione normale delle coppie che sono abbinamenti, e $f_u(w|\theta_u, \sigma_u^2)$ la distribuzione delle trasformazioni dei pesi per le altre coppie.

- La funzione di verosimiglianza dei parametri $\boldsymbol{\xi} = (p, \theta_m, \theta_u, \sigma_m^2, \sigma_u^2)$ date le osservazioni congiunte $(w_{ab}, c_{a,b})$ è:

$$L(\boldsymbol{\xi}|\mathbf{y}, \mathbf{c}) = \prod_{a=1}^{\nu_A} \prod_{b=1}^{\nu_B} \Big[p f_m(w_{ab}|\theta_m, \sigma_m^2) \Big]^{c_{a,b}} \Big[(1-p) f_u(w_{ab}|\theta_u, \sigma_u^2) \Big]^{(1-c_{a,b})}.$$

  La logica sottostante la determinazione di questa verosimiglianza è la stessa che ha caratterizzato la verosimiglianza (6.4).

- La funzione di verosimiglianza precedente, come nel caso già visto nel paragrafo 6.3.2, può essere massimizzata nei suoi parametri incogniti $\boldsymbol{\xi}$ attraverso l'algoritmo EM. Considerati noti i risultati dell'iterazione i, l'iterazione $i+1$ darà i seguenti risultati:

passo E

$$\hat{c}_{a,b}^{(i+1)} = \frac{\hat{p}^{(i)} f_m(w_{ab}|\hat{\theta}_m^{(i)}, \hat{\sigma}_m^{2^{(i)}})}{\hat{p}^{(i)} f_m(w_{ab}|\hat{\theta}_m^{(i)}, \hat{\sigma}_m^{2^{(i)}}) + (1-\hat{p}^{(i)}) f_u(w_{ab}|\hat{\theta}_u^{(i)}, \hat{\sigma}_u^{2^{(i)}})}. \tag{8.5}$$

passo M

$$\hat{\theta}_m^{(i+1)} = \frac{\sum_{ab} \hat{c}_{a,b}^{(i+1)} w_{ab}}{\sum_{ab} \hat{c}_{a,b}^{(i+1)}}, \quad \hat{\theta}_u^{(i+1)} = \frac{\sum_{ab} (1-\hat{c}_{a,b}^{(i+1)}) w_{ab}}{\sum_{ab} (1-\hat{c}_{a,b}^{(i+1)})},$$

$$\hat{\sigma}_m^{2^{(i+1)}} = \frac{\sum_{ab} \hat{c}_{a,b}^{(i+1)} (w_{ab} - \hat{\theta}_m^{(i+1)})^2}{\sum_{ab} \hat{c}_{a,b}^{(i+1)}}, \quad \hat{\sigma}_u^{2^{(i+1)}} = \frac{\sum_{ab} (1-\hat{c}_{a,b}^{(i+1)}) (w_{ab} - \hat{\theta}_u^{(i+1)})^2}{\sum_{ab} (1-\hat{c}_{a,b}^{(i+1)})},$$

$$\hat{p}^{(i+1)} = \frac{1}{\nu_A \nu_B} \sum_{ab} \hat{c}_{a,b}^{(i+1)}.$$

- Per far sì che l'algoritmo EM non converga nelle varie iterazioni verso i confini dello spazio dei valori assumibili dai parametri (ad esempio quando una delle varianze è nulla) si è soliti porre un ulteriore vincolo sul rapporto fra le varianze σ_m^2/σ_u^2.

Stima di FMR

Come detto, nel metodo di record linkage proposto da Fellegi e Sunter, una volta ordinate le coppie candidate secondo il loro peso, si decide di considerare come match tutte le coppie che hanno un peso maggiore alla soglia τ prefissata. Il modello statistico probabilistico sui pesi appena introdotto permette di calcolare in modo molto naturale la probabilità di commettere l'errore di considerare come match coppie che sono non-match. In particolare, la probabilità (8.3) diventa:

$$P(C=0|T>\tau) = \frac{(1-p)\Big[1-\Phi\Big(\frac{w_U(\tau;\gamma_U,\omega_U)-\theta_U}{\sigma_U}\Big)\Big]}{(1-p)\Big[1-\Phi\Big(\frac{w_U(\tau;\gamma_U,\omega_U)-\theta_U}{\sigma_U}\Big)\Big] + p\Big[1-\Phi\Big(\frac{w_M(\tau;\gamma_M,\omega_M)-\theta_M}{\sigma_M}\Big)\Big]} \tag{8.6}$$

dove $\Phi(x)$ è la funzione di ripartizione della distribuzione normale standardizzata calcolata nel punto $x \in I\!R$.

I parametri incogniti della (8.6) vengono sostituiti con le stime di massima verosimiglianza ottenute con il metodo EM nel paragrafo precedente, e questo consente di stimare la probabilità

(8.3) associata a ogni soglia prefissata τ. Una stima della variabilità degli stimatori dei parametri incogniti definiti dall'algoritmo EM può essere ottenuta attraverso il metodo SEM (Meng e Rubin, 1991).

8.3 Da quali fattori dipende il tasso FMR?

Belin (1989, 1993) si propone di studiare l'effetto sul tasso FMR di una serie di scelte adottate da chi conduce l'abbinamento di basi dati attraverso record linkage. L'autore non tiene conto delle metodologie ideate successivamente al 1990. Nonostante ciò, il metodo usato per valutare differenti procedure di record linkage è particolarmente interessante e si illustra di seguito.

Da quanto visto nei capitoli precedenti, esistono diversi passi dove può intervenire il libero arbitrio di chi sta conducendo il record linkage:

- la scelta delle variabili chiave;
- la scelta delle variabili di blocco;
- il modo in cui assegnare i pesi per l'accordo/disaccordo dei confronti (al posto dei pesi 4.2);
- il modo in cui vengono trattate le coppie che non concordano su tutte le variabili chiave ma sono vicine al pieno accordo, se si adotta una definizione dei confronti del tipo (2.10);
- il modo in cui vengono trattati i dati mancanti;
- l'algoritmo per determinare i match incerti;
- la scelta della soglia al di sopra della quale dichiarare una coppia match;
- il luogo da dove si reperiscono i dati.

Di tutti questi fattori, solamente l'ultimo non è sotto il diretto controllo di chi sta conducendo l'integrazione fra basi dati.

Belin ha usato i dati della prova generale del censimento svolta nel 1988 e la corrispondente indagine di qualità PES (*Post Enumeration Survey*). Per queste basi dati, usate anche da Winkler e Thibaudeau (1992) e da Belin e Rubin (1995), sono noti l'insieme $\mathcal{M}$ dei match e $\mathcal{U}$ dei non-match, determinati dal Bureau of the Census attraverso revisione manuale. Le variabili che possono essere usate come variabili chiave sono:

- nome e cognome
- indirizzo
- età
- razza
- sesso
- numero di telefono
- stato civile

- relazione con il capo famiglia.

Belin ha confrontato diverse tipologie di record linkage. Dato che l'analisi condotta da Belin è stata pubblicata nel 1993, ha potuto usare solo pochi metodi fra quelli delineati nei capitoli precedenti, e ha fatto largo uso di metodi "ad hoc". Inoltre ha considerato solo verosimiglianze del tipo (6.6). Ha quindi implementato un'analisi fattoriale sui risultati che lega una funzione monotona del FMR ai fattori che definiscono ogni record linkage. Per stabilizzare la varianza della variabile risposta, Belin utilizza al posto del tasso FMR la funzione:

$$\arcsin\left(\sqrt{FMR}\right).$$

Fra gli altri, alcuni fattori che caratterizzano i diversi record linkage analizzati da Belin sono:

A assegnazione dei pesi alle variabili "nome" e "cognome";

B assegnazione dei pesi per le variabili diverse dal "nome" e "cognome";

C aggiustamenti per tener conto della dipendenza statistica fra i confronti Y^h delle variabili di confronto;

D inclusione delle variabili "stato civile" e "relazione con il capo famiglia" fra le variabili chiave;

E troncamento a 4 o 7 numeri la variabile "numero di telefono";

F luogo di riferimento dei dati. I dati censuari e della PES fanno riferimento alla prova generale effettuata in tre luoghi: la regione orientale dello stato di Washington; Columbia, Missouri; St. Louis, Missouri;

G numero di record della PES che vengono dichiarati match (in pratica la soglia τ che discrimina le coppie dichiarate match da quelle dichiarate non-match non viene scelta in funzione dell'errore atteso, ma in funzione del numero di coppie dichiarate match; su queste si verifica quale è il tasso FMR).

I pesi in A e B possono essere costruiti come in (4.2) e stimati attraverso il metodo indicato nel paragrafo 6.3.2, oppure attraverso i metodi che usano le frequenze delle modalità del paragrafo 6.5 oppure alcuni metodi *ad hoc*, che assegnano a priori dei pesi ai risultati dei confronti. La differenza fra i pesi assegnati in A e B consiste nei diversi pesi *ad hoc* considerati nei due casi e nel diverso modo in cui possono essere gestite coppie di record che non presentano la stessa modalità delle variabili chiave, ma modalità molto vicine. Belin stesso ammette che il fattore C viene eseguito attraverso una procedura *ad hoc*, non tenendo conto quindi delle considerazioni nel paragrafo 6.4. Ad esempio, se "nome", "età" e "sesso" non coincidono, nel caso di indipendenza fra i confronti Y^h deve essere considerato il prodotto dei pesi (4.2) relativi a ogni singola variabile, conducendo a un peso più grande di quello "vero". In questo caso, Belin sottrae un numero sufficientemente grande al peso risultante. Il fattore D è estremamente interessante, e permette di valutare se l'aggiunta di altre variabili chiave può essere utile a ridurre il FMR.

L'analisi della varianza dei risultati ottenuti ha evidenziato alcuni fattori che hanno un basso effetto sul tasso FMR. In particolare una combinazione di fattori che risulta favorita da questa analisi considererebbe: l'uso di una procedura di assegnazione dei pesi *ad hoc* per le variabili "nome" e "cognome" (con pesi pari a ± 4 a seconda che $Y^h = 1$ o a 0); l'uso della procedura

formalizzata da Fellegi e Sunter (capitolo 4) per assegnare un peso per il confronto di ogni variabile diversa da “nome” e “cognome”; l’aggiustamento dei pesi per tener conto della dipendenza fra i confronti Y^h; l’esclusione delle variabili “stato civile” e “relazione con il capo famiglia”, il troncamento a 7 numeri la variabile “numero di telefono”. Come già detto all’inizio di questo paragrafo, non è particolarmente importante la combinazione di fattori che Belin ritiene favorita, ma il modo in cui è arrivato a determinarla.

Capitolo 9

Gli effetti degli errori di abbinamento sulle analisi statistiche

Si è visto nel capitolo 2 che l'obiettivo del record linkage è l'individuazione delle coppie che compongono $\mathcal{M}$. Questo risultato può essere usato in diversi modi. In questo capitolo se ne discute uno: il calcolo di un parametro per una o più variabili (paragrafo 9.1). Questo argomento, abbondantemente studiato nella letteratura statistica, deve essere trattato con cautela quando si usano basi dati costruite attraverso record linkage: infatti l'insieme di coppie dichiarate match dal record linkage è un *insieme stimato*. L'errore di abbinamento, studiato nel capitolo 8, deve essere tenuto in considerazione quando si applicano metodi statistici. In questo capitolo si vuole sottolineare quindi che la fase di record linkage non deve essere vista come una fase a sé stante rispetto a tutte le altre fasi di trattamento del dato, ma in congiunzione ad esse. Non si conosce ancora a fondo la materia, e si può far riferimento solo a pochi articoli riportati nei prossimi due paragrafi. Naturalmente il calcolo di parametri non è l'unico argomento che viene influenzato dall'errore di abbinamento. Un altro argomento particolarmente importante riguarda gli effetti dell'errore di abbinamento sui metodi cattura-ricattura e in particolare sulla stima della sottocopertura del censimento (in proposito si rimanda a Biemer, 1988, e Ding e Fienberg, 1994).

9.1 Gli effetti degli errori di abbinamento sui parametri di una popolazione

A volte le analisi statistiche richiedono informazioni che non sono reperibili in un'unica rilevazione o fonte. Esempi di questo tipo si hanno in biostatistica, dove si può essere interessati a analizzare congiuntamente le informazioni ricavate in diversi studi medici (Beebe, 1985, Newcombe, 1988, Armstrong e Saleh, 2000, Bartlett *et al.*, 1993). L'interesse per queste analisi non è trascurabile nella statistica ufficiale: basti pensare all'importanza che va assumendo l'uso congiunto di fonti amministrative e statistiche.

I primi a trattare l'argomento degli effetti dell'errore di abbinamento sul calcolo dei parametri di una popolazione per una o più variabili disponibili su una base dati ottenuta attraverso record linkage sono Neter, Maynes e Ramanathan (1965). Presentano un modello particolarmente semplice, non necessariamente adatto a descrivere casi reali, ma dal quale si desume che, quando la base dati è ottenuta attraverso record linkage, il calcolo di alcuni parametri attraverso gli stimatori usuali può risultare inefficiente.

Il modello che ipotizzano è il seguente. Si supponga di avere una base dati A composta da N unità. Inoltre si ipotizza che ogni unità $a \in \mathcal{A}$ sia osservata anche in una base dati B, ovvero esiste

$b \in \mathcal{B}$ tale che $a = b$. Sia $\mathcal{B}$ composto anch'esso da N unità: il vincolo (2.6) diventa:

$$\sum_{b=1}^{N} c_{a,b} = 1, \qquad \forall\, a \in \mathcal{A}. \tag{9.1}$$

In questo modo si sa che

1. ogni unità $a \in \mathcal{A}$ definisce un match con un'unità $b \in \mathcal{B}$, e l'insieme $\mathcal{M}$ ha numerosità N.

Sia $\mathbf{X}$ il vettore delle k variabili chiave in comune ai due campioni, e sia U una variabile rilevata nella base dati B, ma non disponibile in A. Attraverso un metodo di record linkage, si associa ad ogni unità $a \in \mathcal{A}$ un'unità $b \in \mathcal{B}$.

Commento 9.1 *È da notare che la base dati $\mathcal{B}$ viene vista come un "donatore" di record, e quindi l'ovvio vincolo*

$$\sum_{a=1}^{N} c_{a,b} = 1, \qquad \forall\, b \in \mathcal{A}$$

non viene imposto. Come conseguenza si ha che il meccanismo di abbinamento di un record da $\mathcal{B}$ a $\mathcal{A}$ si ripete in modo identico e indipendente per tutte le unità $a \in \mathcal{A}$. □

Per via dell'errore di abbinamento, non possiamo essere certi che ad ogni unità $a \in \mathcal{A}$ sia associato il valore corretto u_a. Si indichi con il simbolo Υ il risultato dell'abbinamento attraverso record linkage. Neter, Maynes e Ramanathan, oltre all'ipotesi al punto 1, fanno le seguenti ipotesi:

2. ogni unità $a \in \mathcal{A}$ ha la stessa probabilità p di essere abbinata correttamente
3. ogni unità $a \in \mathcal{A}$ ha la stessa probabilità q di essere abbinata a un'unità sbagliata di $b \in \mathcal{B}$, in modo tale che:
$$p + (N-1)q = 1.$$

Sotto le ipotesi ai punti 1, 2 e 3, la variabile Υ è definita nel modo seguente:

$$\Upsilon_a = \begin{cases} u_a & \text{con probabilità } p \\ u_b & \text{con probabilità } q \quad (a \neq b). \end{cases} \tag{9.2}$$

Neter, Maynes e Ramanathan verificano l'effetto delle differenze fra i valori $\Upsilon_a = \upsilon_a$ (ovvero i valori di U ottenuti dopo aver applicato una procedura di record linkage), e i valori u_a (i veri valori di U) sul calcolo di alcuni parametri.

Commento 9.2 *In particolare ipotizzano di estrarre con ripetizione un campione di n elementi dagli N elementi della popolazione, e determinano che la media campionaria di Υ è uno stimatore corretto della media di U sulle N unità, mentre la varianza campionaria di Υ sovrastima la varianza di U sulle N unità. Inoltre, il coefficiente di correlazione fra U e una variabile Z (quest'ultima non affetta da errore di abbinamento) viene sottostimato in valore assoluto.* □

Scheuren e Winkler (1993) sostengono che difficilmente l'ipotesi 3 si possa ritenere valida: ogni record $a \in \mathcal{A}$ non ha la stessa probabilità q di essere abbinato a una qualsiasi unità $b \in \mathcal{B}$. Ad esempio si supponga di avere come variabile chiave il codice fiscale, e per via di un errore tipografico il codice fiscale di un'unità a viene riportato erroneamente in un carattere. Allora

non è vero che una qualsiasi unità $b \in \mathcal{B}$ (cioè qualsiasi altro codice fiscale disponibile in $\mathcal{B}$) è candidata ad essere abbinata con a con la stessa probabilità, ma alcune unità b avranno *chance* maggiori di altre.

Il modello che gli autori propongono, adottato anche da Larsen (1999), considera l'ipotesi 1 di Neter *et al*. (1965) e sostituisce le ipotesi 2 e 3 con la più plausibile:

4. la variabile Υ, errore di abbinamento, è definita da:

$$\Upsilon_a = \begin{cases} U_a & \text{con probabilità } q_{aa} \\ U_b & \text{con probabilità } q_{ab}, \quad (a \neq b), \end{cases} \tag{9.3}$$

con

$$q_{aa} + \sum_{b \neq a} q_{ab} = 1, \qquad \forall\, a \in \mathcal{A}.$$

Sotto questo modello si può dimostrare che la media di Υ sulle N unità:

$$\bar{\Upsilon} = \frac{1}{N} \sum_{a=1}^{N} \Upsilon_a$$

è uno stimatore distorto (rispetto all'errore di abbinamento) della media di U sulle stesse N unità:

$$\bar{U} = \frac{1}{N} \sum_{a=1}^{N} U_a.$$

Infatti si ha:

$$E(\Upsilon_a) = U_a + (q_{aa} - 1)U_a + \sum_{b \neq a} q_{ab} U_b = U_a + B_a,$$

dove

$$B_a = (q_{aa} - 1)U_a + \sum_{b \neq a} q_{ab} U_b \tag{9.4}$$

rappresenta la distorsione legata all'osservazione a. Di conseguenza il valore atteso della media aritmetica di Υ, $\bar{\Upsilon}$, calcolato secondo la (9.3) è:

$$E(\bar{\Upsilon}) = \frac{1}{N} \sum_{a=1}^{N} E(\Upsilon_a) = \bar{U} + \frac{1}{N} \sum_{a=1}^{N} B_a.$$

9.1.1 La stima dei parametri in un modello di regressione lineare

Un argomento estremamente interessante con notevoli implicazioni pratiche è lo studio dell'effetto degli errori di abbinamento sulla stima dei parametri di un modello di regressione lineare fra h variabili Z_1,...,Z_h, disponibili nella base dati A e non affette quindi da errore di abbinamento, e una variabile U, disponibile nella base dati B e ricostruita in A attraverso record linkage. Al posto di U si osserva quindi Υ, definita come in (9.3).

Sia $\mathbf{Z}$ una matrice N per h la cui a-esima riga è:

$$\mathbf{z}_a^T - \bar{\mathbf{Z}}^T = (z_{a,1} - \bar{z}_1, ..., z_{a,h} - \bar{z}_h), \qquad a = 1, ..., N.$$

Allo stesso modo si definiscano i vettori (di dimensione N) $\boldsymbol{\Upsilon}$ e $\mathbf{B}$:

$$\boldsymbol{\Upsilon} = \begin{pmatrix} \Upsilon_1 \\ \Upsilon_2 \\ \vdots \\ \Upsilon_N \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_N \end{pmatrix},$$

dove $\mathbf{B}$ è il vettore delle distorsioni (9.4). Sia $\boldsymbol{\beta}$ il vettore dei coefficienti di regressione fra U e $(Z_1, ..., Z_h)$ ottenuto attraverso il metodo dei minimi quadrati sulle N osservazioni. Dato che non si è in grado di disporre degli N valori U_a, $a = 1, ..., N$, il metodo naïve per la stima del vettore di parametri $\boldsymbol{\beta}$ consiste nell'applicare il metodo dei minimi quadrati sulle osservazioni $\boldsymbol{\Upsilon}$ e $\mathbf{Z}$, ottenendo lo stimatore:

$$\beta^* = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Upsilon}.$$

Il valore medio di questo stimatore rispetto all'errore di abbinamento descritto in (9.3) è:

$$E(\beta^*) = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T E(\boldsymbol{\Upsilon}) = \beta + (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{B}. \tag{9.5}$$

Di conseguenza, è sufficiente riuscire a stimare la distorsione

$$(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{B}$$

per poter ottenere uno stimatore corretto di $\boldsymbol{\beta}$. A questo scopo, si suggerisce di adottare la distribuzione di probabilità:

$$q_{ab} = \frac{P(C_{a,b} = 1|\mathbf{y}_{ab})}{\sum_{b=1}^{N} P(C_{a,b} = 1|\mathbf{y}_{ab})}, \qquad b \in \mathcal{B},$$

per ogni $a \in \mathcal{A}$, dove $P(C_{a,b} = 1|\mathbf{y}_{ab})$ può essere stimato attraverso controllo manuale di un campione di coppie o l'utilizzo di opportuni modelli (in proposito si veda il capitolo 6). Si sottolinea che $P(C_{a,b} = 1|\mathbf{y}_{ab})$ coincide anche con il peso $t^*(\mathbf{y})$ definito nei paragrafi 4.2 e 6.3.3.

Scheuren e Winkler (1993) suggeriscono di considerare per ogni unità $a \in \mathcal{A}$ le due unità b_1 e b_2 in $\mathcal{B}$ alle quali corrispondono i valori più alti delle probabilità q_{ab}:

$$q_{ab_1} \geq q_{ab_2} \geq q_{ab}, \qquad \forall\, b \in \mathcal{B}.$$

Quindi considerano una stima di B_a data da:

$$\hat{B}_a = (q_{ab_1} - 1)u_{b_1} + q_{ab_2}u_{b_2}.$$

Larsen (1999) e Lahiri e Larsen (2000) non hanno ristretto l'analisi a due soli valori, e definiscono lo stimatore:

$$\hat{\beta}_U = (\mathbf{W}^t\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{\Upsilon}, \tag{9.6}$$

con:

$$W_a = \sum_{b=1}^{N} q_{ab}z_b.$$

Lahiri e Larsen (2000) hanno dimostrato che questo stimatore è non distorto.

Larsen (1999) conduce diverse simulazioni per studiare questi stimatori e verificare il guadagno in efficienza rispetto allo stimatore naïve (ovvero uno stimatore che non tiene conto dell'errore di abbinamento dovuto al record linkage).

Le simulazioni sono state condotte secondo il seguente schema. Sono stati simulati due gruppi di unità $\mathcal{A}$ e $\mathcal{B}$ di uguale numerosità N, con N generato da una distribuzione uniforme fra 500 e 4000. Sono stati generati N valori da una variabile (U, Z) normale doppia con coefficiente di correlazione r generato da una distribuzione uniforme fra 0,3 e 0,9. I valori relativi a Z sono stati assegnati al gruppo A e i valori U sono stati assegnati alle N unità in B. L'aggancio fra A e B viene fornito da k variabili chiave X^h, $h = 1, ..., k$, con k generato da una uniforme fra 5 e 10. Ogni variabile X^h può assumere un numero di modalità generato da una distribuzione uniforme fra 2 e 10, in modo tale che:

- u_h è un numero compreso fra 0,1 (quando le modalità sono 10) e 0,5 quando le modalità sono 2;
- m_h è generato da una uniforme fra 0,6 e 1.

Inoltre i confronti Y^h vengono assunti fra loro indipendenti, come nel modello (6.6). Viene infine generata una informazione utile per il bloccaggio delle N unità: i blocchi hanno ampiezza generata da una distribuzione uniforme fra 5 e 20 unità. Il bloccaggio è tale da non creare falsi non-match.

I risultati vengono sintetizzati attraverso la somma dei quadrati degli errori (differenza fra i parametri veri e quelli stimati rispettivamente con il metodo naïve e con il metodo nella formula (9.6)), e mostrano che nel secondo caso la somma dei quadrati degli errori è pari ad un terzo di quella in cui le stime sono ottenute con il metodo naïve. Anche la percentuale di intervalli di confidenza al 95% che ricoprono il vero parametro è molto superiore nel caso (9.6) piuttosto che nel metodo naïve.

Ulteriori simulazioni sono disponibili su Lahiri e Larsen (2000).

Commento 9.3 *I modelli di errore di abbinamento analizzati in questo capitolo possono essere particolarmente utili quando si utilizza l'integrazione fra due basi dati A e B come metodo per imputare dati mancanti in A. Questo approccio, viene discusso in Robinson-Cox (1998) per un esempio specifico, e viene utilizzato da Winkler e Scheuren (1996) e Scheuren e Winkler (1997) in un approccio iterativo che combina le fasi di record linkage, imputazione e analisi dei dati in una procedura iterativa.* □

Bibliografia

Alvey W., Jamerson B. (1997). *Record Linkage Techniques - 1997, Proceedings of an International Workshop and Exposition*, Federal Committee on Statistical Methodology, Office of Management of the Budget. National Academy Press, Washington D.C.

Arellano M.G. (1992). Commento a Newcombe H.B., Fair M.E., Lalonde P. "The use of names for linking personal records". *Journal of the American Statistical Association*, **87**, 1204-1206.

Armstrong J., Mayda J.E. (1993). "Model-based estimation of record linkage error rates". *Survey Methodology*, **19**, 137-147.

Armstrong J., Saleh M. (2000). "Weight estimation for large scale record linkage applications". *Proceedings of the section on Survey Research Methods, American Statistical Association*, 1-10.

Bartlett S., Krewski D., Wang Y., Zielinski J.M. (1993). "Evaluation of error rates in large scale computerized record linkage studies". *Survey Methodology*, **19**, 3-12.

Beebe G.W. (1985). "Why are epidemiologists interested in matching algorithms?", in Kills B., Alvey W. (editori) *Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies*, 139-143.

Belin T.R. (1989). "A proposed improvement in computer matching techniques". *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 784-789.

Belin T.R. (1993). "Evaluation of sources of variation in record linkage trough a factorial experiment". *Survey Methodology*, **19**, 137-147.

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

Bernardo J.M., Smith A.F.M. (1994). *Bayesian Theory*. Wiley, New York.

Biemer P.P. (1988). "Modeling matching error and its effect on estimates of census coverage error". *Survey Methodology*, **14**, 117-134.

Bishop Y.M.M., Fienberg S.E., Holland P.W. (1975). *Discrete Multivariate Analyses: Theory and Practice*. MIT Press, Cambridge.

Box G.E.P., Cox D.R. (1964). "An analysis of transformations (with discussion)". *Journal of the Royal Statistical Society, Series B*, **26**, 211-246.

Burkard R.E., Derigs U. (1980). *Assignment and Matching Problems: Solution Methods with FORTRAN-Programs*. Springer Verlag, New York.

Chernoff H. (1980). “The identification of an element of a large population in the presence of noise”. *Annals of Statistics*, **8**, 1179-1197.

Coccia G., Gabrielli D., Sorvillo M.P. (1993). “Prospettive di utilizzazione delle tecniche di linkage in ambito demografico”. *Quaderni di Ricerca - Metodologia e Informatica*, **7**, ISTAT, Roma.

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

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

Ding Y., Fienberg S.E. (1994). “Dual system estimation of census undercount in the presence of matching error”. *Survey Methodology*, **20**, 149-158.

Duncan G., Lambert D. (1989). “The risk of disclosure for microdata”. *Journal of Business & Economic Statistics*, **7**, 207-217.

Fellegi I.P. (1997). “Record linkage and public policy - A dynamic evolution”, in Alvey W., Jamerson B. (editori) *Record Linkage Techniques - 1997, Proceedings of an International Workshop and Exposition*, 3-12.

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

Filippucci C. (a cura di) (2000). *Tecnologie Informatiche e Fonti Amministrative nella Produzione dei Dati*. Franco Angeli, Milano.

Fortini M. (2001). *Linee guida metodologiche per rilevazioni statistiche*, disponibile on-line all’indirizzo: `http://www.istat.it/metadati/lineeg/lg.html`.

Fortini M., Liseo B., Nuccitelli A., Scanu M. (2000). “Bayesian approaches to record linkage”. Università degli Studi di Roma “La Sapienza”, Dipartimento di studi geoeconomici, statistici, storici per l’analisi regionale, working paper n. 15.

Fortini M., Liseo B., Nuccitelli A., Scanu M. (2001). “On Bayesian record linkage”. *Research in Official Statistics*, **4**, 185-198. Pubblicato anche in E. George (editore), *Monographs of Official Statistics, Bayesian Methods*, EUROSTAT, 155-164.

Fortini M., Nuccitelli A., Liseo B., Scanu M. (2002). “Modelling issues in record linkage: a Bayesian perspective”. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, in corso di stampa.

Garofalo G. (1998). “The ASIA Project (setting up of the Italian business register) synthesis of the methodological manual”. *Proceedings of the* 12^{th} *Meeting of the International Roundtables on Business Survey Frames*, Helsinki, Settembre 1998.

Garofalo G., Viviano C. (2000). “The problem of links between legal units: statistical techniques for enterprise identifications and the analysis of continuity”. *Rivista di Statistica ufficiale*, **1/2000**, 5-27.

Giusti A., Marliani G., Torelli N. (1991). “Procedure per l’abbinamento dei dati individuali delle forze di lavoro”, in *Forze di Lavoro: Disegno dell’Indagine e Analisi Strutturali, Annali di Statistica*, serie IX, A 120, ISTAT, Roma, 121-148.

Haberman S. (1974). *The Analysis of Frequency Data*. The University of Chicago Press, Chicago.

Haberman S. (1975). “Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation”. *Proceedings of the Section on Statistical Computing, American Statistical Association*, 45-50.

Haberman S. (1977). “Product models for frequency tables involving indirect observation”. *Annals of Statistics*, **5**, 1124-1147.

Haberman S. (1978). *Analysis of Qualitative Data*. Academic Press, New York.

ISTAT (1989). *Manuali di Tecniche di Indagine: Il Sistema di Controllo della Qualità dei Dati, volume 6*, ISTAT, collana *Metodi e Norme*, Roma.

ISTAT (2001). *Forze di Lavoro - Media 2000*, Annuario n. 6, ISTAT, Roma.

Jabine T.B., Scheuren F.J. (1986). “Record linkages for statistical purposes: methodological issues”. *Journal of Official Statistics*, **2**, 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*, **84**, 414-420.

Kelley R.B. (1984). “Blocking considerations for record linkage under conditions of uncertainty”. *Proceedings of the Social Statistics Section, American Statistical Association*, 602-605.

Kelley R.B. (1985). “Advances in record linkage methodology: a method for determining the best blocking strategy”, in Kills B., Alvey W. (editori) *Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies*, 199-203.

Kills B., Alvey W. (1985). *Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, 9-15 maggio 1985. Washington: Internal Revenue Service.

Kirkendall N.J. (1985). “Weights in computer matching: applications and an information theoretic point of view”, in Kills B., Alvey W. (editori) *Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methodologies*, 189-197.

Kotz S., Balakrishnan N., Johnson N.L. (2000). *Continuous Multivariate Distributions, 2nd Edition*. Wiley, New York.

Kullback S. (1959). *Information Theory and Statistics*. Wiley, New York.

Lahiri P., Larsen M.D. (2000). “Model-based analysis of records linked using mixture models”. *Proceedings of the section on Survey Research Methods, American Statistical Association*, 11-19.

Larsen M.D. (1999). “Multiple imputation analysis of records linked using mixture models”. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 65-71.

Larsen M.D., Rubin D.B. (2001). “Iterative automated record linkage using mixture models”. *Journal of the American Statistical Association*, **96**, 32-41.

Lawler E. L. (1976). *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart & Winston, New York.

Lehmann E.L. (1986). *Testing Statistical Hypotheses, Second Edition*. Springer Verlag, New York.

Little R.J.A., Rubin D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Meng X.L., Rubin D.B. (1991). “Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm”. *Journal of the American Statistical Association*, **86**, 899-909.

Meng X.L., Rubin D.B. (1993). “Maximum likelihood via the ECM algorithm: a general framework”. *Biometrika*, **80**, 267-278.

NeSmith N.P. (1997). “Record linkage and genealogical files”, in Alvey W., Jamerson B. (editori) *Record Linkage Techniques - 1997, Proceedings of an International Workshop and Exposition*, 358-361.

Neter J., Maynes S., Ramanathan R. (1965). “The effect of mismatching on the measurement of response errors”. *Journal of the American Statistical Association*, **60**, 1005-1027.

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

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

Nuccitelli A. (2001). *Integrazione di dati mediante tecniche di abbinamento esatto: una rassegna critica ed una proposta in ambito bayesiano*. Tesi di dottorato di ricerca in Metodi Statistici per l’Economia e l’Impresa, XIII ciclo.

Paggiaro A., Torelli N. (1999). “Una procedura per l’abbinamento di record nella rilevazione trimestrale delle forze di lavoro”. ISTAT, documentazione tecnica del file standard panel della rilevazione trimestrale delle forze lavoro.

Pinelli A. (1984). “The record linkage in the study of infant mortality: some aspects concerning data quality”. *Statistica*, **44**, 675-686.

Porter E.H., Winkler W.E. (1997). “Approximate string comparison and its effect on advanced record linkage system”. *Bureau of the Census, Statistical Research Division, Statistical Research Report Series*, n. RR97/02.

Robinson-Cox J.F. (1998). “A record linkage approach to imputation of missing data: analyzing tag retention in a tag-recapture experiment”. *Journal of Agricultural, Biological and Environmental Statistics*, **3**, 48-61.

Rogot E., Sorlie P., Johnson N.J. (1986). “Probabilistic methods in matching census samples to the national death index”. *Journal of Chronic Deseases*, **39**, 719-734.

Rubin D.B., Stern H.S. (1993). “Testing in latent class models using a posterior predictive check distribution”, in von Eye A., Clogg C. (editori), *Latent Variables Analysis: Applications for Developmental Research*, 420-438. Sage, Thousand Oaks.

Särndal C., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Scheuren F., Winkler W.E. (1993). “Regression analysis of data files that are computer matched”. *Survey Methodology*, **19**, 39-58.

Scheuren F., Winkler W.E. (1997). “Regression analysis of data files that are computer matched - part II”. *Survey Methodology*, **23**, 157-165.

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

Thibaudeau Y. (1989). “Fitting log-linear models when some dichotomous variables are unobservable”. *Proceedings of the Section on statistical computing, American Statistical Association*, 283-288.

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 della Società Italiana di Statistica*, Sorrento, 14-17 aprile 1998.

Torelli N., Paggiaro A. (1999). “Problemi di stima in una procedura di abbinamento esatto”, in *Atti del convegno “Verso i censimenti del 2000”*, 7-9 giugno 1999, Udine.

White D. (1997). “A review of the statistics of record linkage for genealogical research”, in Alvey W., Jamerson B. (editori) *Record Linkage Techniques - 1997, Proceedings of an International Workshop and Exposition*, 362-373.

Winkler W.E. (1985). “Exact matching lists of businesses: blocking, subfield identification and information theory”. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 438-443. Pubblicato in versione più estesa in Alvey W., Kills B. (editori) *Record Linkage Techniques - 1985, Proceedings of the Workshop on Exact Matching Methods*, 227-241.

Winkler W.E. (1988). “Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage”. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671. Disponibile in versione più estesa in: *Bureau of the Census, Statistical Research Division, Statistical Research Report Series*, n. RR00/05.

Winkler W.E. (1989a). “Frequency-based matching in the Fellegi-Sunter model of record linkage”. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 778-783. Disponibile in una versione più estesa in: *Bureau of the Census, Statistical Research Division, Statistical Research Report Series*, n. RR00/06.

Winkler W.E. (1989b). “Near automatic weight computation in the Fellegi-Sunter model of record linkage”. *Proceedings of the Annual Research Conference, Washington D.C., U.S. Bureau of the Census*, 145-155.

Winkler W. E. (1990). “String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage”. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.

Winkler W.E. (1992). “Comparative analysis of record linkage decision rules”. *Proceedings of the Section on Survey Research Methods , American Statistical Association*, 829-834.

Winkler W.E. (1993). “Improved decision rules in the Fellegi-Sunter model of record linkage”. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279.

Winkler W.E. (1994). “Advanced methods for record linkage”. *Bureau of the Census, Statistical Research Division, Statistical Research Report Series*, n. RR94/05.

Winkler W.E. (1995). “Matching and Record Linkage”, in Cox B.G., Binder D.A., Chinnappa B.N., Christianson A., Colledge M., Kott P.S. (editori) *Business Survey Methods*, 355-384. Wiley, New York.

Winkler W.E. (1998). “Re-identification methods for evaluating the confidentiality of analytically valid microdata”. *Research in Official Statistics*, **2**, 87-104

Winkler W.E. (2000). “Machine learning, information retrieval and record linkage”. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-29.

Winkler W.E., Scheuren F. (1996). “Recursive analysis of linked data files”. *Proceedings of the 1996 Census Bureau Annual Research Conference*, 920-935.

Winkler W.E., Thibaudeau Y. (1991). “An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial Census”. *Bureau of the Census, Statistical Research Division, Statistical Research Report Series*, n. RR91/09.

Wolter K.M. (1986). “Some coverage error models for census data”. *Journal of the American Statistical Association*, **81**, 338-346.

Yancey W.E. (2000). “Frequency-dependent probability measures for record linkage”. *Bureau of the Census, Statistical Research Division, Statistical Research Report Series*, n. RR00/07.

Appendice A

Elenco delle notazioni

- A: base dati della prima rilevazione.
- B: base dati della seconda rilevazione, da integrare con A.
- a: generica unità rilevata in A.
- b: generica unità rilevata in B.
- ν_A: numero di unità osservate in A.
- ν_B: numero di unità osservate in B.
- $\mathcal{A}$: insieme delle ν_A unità in A.
- $\mathcal{B}$: insieme delle ν_B unità in B.
- $\mathcal{A} \times \mathcal{B}$: insieme delle coppie (a, b), $a \in \mathcal{A}$, $b \in \mathcal{B}$, sottoponibili a confronto.
- ν: numero totale delle coppie di unità confrontabili: $\nu = \nu_A \times \nu_B$.
- $\mathcal{M}$: insieme delle coppie $(a, b) \in \mathcal{A} \times \mathcal{B}$ che sono match.
- $\mathcal{U}$: insieme delle coppie $(a, b) \in \mathcal{A} \times \mathcal{B}$ che sono non-match.
- N: numero di coppie che sono match, ovvero cardinalità di $\mathcal{M}$, con

$$N \leq \min\{\nu_A, \nu_B\}.$$

- $c_{a,b}$: indicatore di appartenenza della coppia (a, b) a $\mathcal{M}$ o $\mathcal{U}$.
- $\mathbf{x}_a^A$: vettore delle osservazioni $x_{a,h}^A$, $h = 1, ..., k$, delle k variabili chiave sull'unità $a \in \mathcal{A}$.
- $\mathbf{x}_b^B$: vettore delle osservazioni $x_{b,h}^B$, $h = 1, ..., k$, delle k variabili chiave sull'unità $b \in \mathcal{B}$.
- $\mathbf{y}_{ab}$: funzione dei confronti fra le osservazioni $\mathbf{x}_a^A$ e $\mathbf{x}_b^B$

$$\mathbf{y}_{ab} = f(x_{a,1}^A, ..., x_{a,k}^A; x_{b,1}^B, ..., x_{b,k}^B).$$

- y_{ab}^h: confronto fra le osservazioni della variabile h, h-esimo componente del vettore $\mathbf{y}_{ab}$.
- $\mathcal{D}$: insieme dei possibili vettori di confronto $\mathbf{y}$.

- $m(\mathbf{y})$: distribuzione della variabile dei confronti per le coppie che sono match.
- $u(\mathbf{y})$: distribuzione della variabile dei confronti per le coppie che sono non-match.
- $t(\mathbf{y})$: peso assegnato al confronto $\mathbf{y} \in \mathcal{D}$ dalla procedura di record linkage di Fellegi e Sunter:
$$t(\mathbf{y}) = \frac{m(\mathbf{y})}{u(\mathbf{y})}.$$
- $w(\mathbf{y})$: trasformazione logaritmica del peso $t(\mathbf{y})$.
- $t^*(\mathbf{y})$: trasformazione logistica del peso $w(\mathbf{y})$.
- λ: probabilità che la procedura di record linkage decida che una coppia è non-match quando è un match.
- μ: probabilità che la procedura di record linkage decida che una coppia è un match quando è un non-match.
- τ_λ: soglia per il confronto dei pesi nella procedura di Fellegi e Sunter legata alla probabilità di errore λ.
- τ_μ: soglia per il confronto dei pesi nella procedura di Fellegi e Sunter legata alla probabilità di errore μ.
- $\mathbf{Y}$: generica variabile aleatoria che genera il vettore di confronto $\mathbf{y}$.
- $\mathbf{Y}_{ab}$: variabile aleatoria che genera il vettore di confronto per la coppia (a, b). In genere le variabili aleatorie $\mathbf{Y}_{ab}$, $a = 1, ..., \nu_A$, $b = 1, ..., \nu_B$, sono fra loro indipendenti e identicamente distribuite a $\mathbf{Y}$.
- C: generica variabile aleatoria dicotomica che esprime con quale probabilità è possibile generare (estrarre nel caso di popolazioni finite) una coppia che è un match. La variabile C assume i valori 0 (non-match) e 1 (match) secondo la distribuzione di probabilità:
$$C = \begin{cases} 1 & \text{con probabilità } p \\ 0 & \text{con probabilità } 1 - p \end{cases}$$
- $C_{a,b}$: variabile aleatoria che esprime per la coppia (a, b) lo status della coppia, ovvero se è match o non-match.
- $\mathbf{C}$: matrice aleatoria con ν_A righe e ν_B colonne il cui generico elemento è $C_{a,b}$.
- $\mathcal{C}$: insieme delle possibili matrici che descrivono lo status di tutte le $\nu_A \times \nu_B$ coppie, tenendo conto dei possibili vincoli sui valori $c_{a,b}$: ad esempio si considerino i vincoli (2.5), (2.6) e (2.7).
- $\mathcal{C}^h$: insieme delle matrici $\mathbf{c} \in \mathcal{C}$ con esattamente h match, $h = 0, 1, ..., N$.
- f_l: frequenza delle unità $a \in \mathcal{A}$ che presentano la modalità l della variabile chiave X, $l = 1, .., v$.

- g_l: frequenza delle unità $b \in \mathcal{B}$ che presentano la modalità l della variabile chiave X, $l = 1, .., v$.
- h_l: frequenza delle coppie $(a, b) \in \mathcal{A} \times \mathcal{B}$ le cui unità presentano la stessa modalità l della variabile chiave X, $l = 1, .., v$,
$$h_l \leq \min\{f_l, g_l\}.$$
- e_A: probabilità che la variabile X venga riportata con errore nella lista A.
- e_B: probabilità che la variabile X venga riportata con errore nella lista B.
- e_{A0}: probabilità che la variabile X non venga riportata nella lista A (mancata risposta).
- e_{B0}: probabilità che la variabile X non venga riportata nella lista B (mancata risposta).
- e_T: probabilità che una unità riporti due modalità diverse nelle due occasioni di rilevazione.
- G: numero di classi in cui si dividono la $\nu_A \times \nu_B$ coppie (in genere $G = 2$, l'insieme dei match $\mathcal{M}$ e l'insieme dei non-match $\mathcal{U}$).
- $d(.,.)$: funzione di distanza fra due oggetti:
$$d(\eta, \theta) = \begin{cases} 1 & \text{se } \eta = \theta \\ 0 & \text{altrimenti.} \end{cases}$$
- $\Delta(\mathbf{p}, \boldsymbol{\pi})$: distanza di Kullback-Leibler fra le distribuzioni $\mathbf{p}$ e $\boldsymbol{\pi}$:
$$\Delta(\mathbf{p}, \boldsymbol{\pi}) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{\pi_i}\right).$$

Appendice B

Il metodo EM

L'algoritmo EM è stato definito da Dempster, Laird e Rubin (1977) e ha come obiettivo la stima di massima verosimiglianza dei parametri di una distribuzione, date n osservazioni campionarie generate in modo indipendente e identicamente distribuito dalla distribuzione stessa. La particolarità di questo algoritmo risiede nel fatto che le stime vengono ottenute nel caso in cui alcuni dati sono mancanti. In questa situazione non è possibile ottenere stime dirette di massima verosimiglianza, ma è necessario ricorrere a procedure iterative. Si consideri un campione di n elementi generati in modo indipendente attraverso la stessa legge di probabilità $p(x|\theta)$. Quando alcuni elementi del campione sono mancanti, la funzione di verosimiglianza è funzione di[1]:

- i parametri incogniti θ
- il vettore dei dati osservati $\mathbf{x}_{obs}$
- il vettore dei dati mancanti $\mathbf{x}_{mis}$.

Supponendo che i dati mancanti siano stati generati dalla stessa legge di probabilità dei dati osservati (per definizioni più esatte del principio di *ignorabilità* si rimanda a Little e Rubin, 1987), le stime di massima verosimiglianza si ottengono iterando questi due passi:

1. il passo E (*expectation*) in cui, dato il vettore dei parametri θ ottenuto al passo precedente, si sostituisce il vettore di dati mancanti $\mathbf{x}_{mis}$ con il corrispondente valore atteso $E(\mathbf{x}_{mis}|\theta)$;
2. il passo M (*maximization*) in cui si ricerca il vettore di parametri θ che massimizza la funzione di verosimiglianza avendo sostituito le osservazioni mancanti con i valori $E(\mathbf{x}_{mis}|\theta)$ trovati al passo E.

Applichiamo quanto appena detto nel contesto specifico dei modelli usati nel record linkage. In particolare gli autori che ne hanno fatto uso (ad esempio Jaro, 1989, Winkler, 1988, Torelli e Paggiaro, 1999, Larsen e Rubin, 2001) hanno sempre ipotizzato confronti del tipo (2.10). Come detto nel capitolo 6 la funzione di verosimiglianza più semplice per questo tipo di confronti è la (6.6):

$$L\Big(p, \{m(.)\}, \{u(.)\}\Big|\mathbf{y}_{ab}, c_{a,b}, (a,b) \in \mathcal{A} \times \mathcal{B}\Big) =$$
$$= \prod_{(a,b)} \left(p \prod_{h=1}^{k} m_h^{y_{ab}^h}(1-m_h)^{1-y_{ab}^h} \right)^{c_{a,b}} \left((1-p) \prod_{h=1}^{k} u_h^{y_{ab}^h}(1-u_h)^{1-y_{ab}^h} \right)^{1-c_{a,b}} . \quad \text{(B.1)}$$

[1] a rigore la funzione di verosimiglianza è una funzione di θ dato il vettore osservato.

I parametri da stimare sono p, m_h e u_h, $h = 1, ..., k$, i dati osservati sono i valori y^h_{ab}, $a = 1, ..., \nu_A$, $b = 1, ..., \nu_B$, $h = 1, ..., k$, e i dati mancanti sono i valori della matrice $\mathbf{c}$, che indicano l'appartenenza delle coppie (a, b) a $\mathcal{M}$ o $\mathcal{U}$.

I passi dell'EM sono i seguenti.

1. Si fissino dei valori iniziali dei parametri incogniti: $\hat{p}, \hat{\mathbf{m}}, \hat{\mathbf{u}}$. Si suppongano ora noti i risultati dell'iterazione i-esima.

2. All'iterazione $i + 1$, il passo E consiste nel rimpiazzare i dati mancanti $c_{a,b}$ con i loro valori attesi, ovvero:

$$\hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab}) = \\ = \frac{\hat{p}^{(i)} \prod_{h=1}^{k} (\hat{m}^{(i)}_h)^{y^h_{ab}} (1 - \hat{m}^{(i)}_h)^{1-y^h_{ab}}}{\hat{p}^{(i)} \prod_{h=1}^{k} (\hat{m}^{(i)}_h)^{y^h_{ab}} (1 - \hat{m}^{(i)}_h)^{1-y^h_{ab}} + (1 - \hat{p}^{(i)}) \prod_{h=1}^{k} (\hat{u}^{(i)}_h)^{y^h_{ab}} (1 - \hat{u}^{(i)}_h)^{1-y^h_{ab}}},$$

   dove si vuole sottolineare che le "stime" di $c_{a,b}$ ottenute a questo passo sono identiche per tutte le coppie (a, b) che presentano lo stesso vettore di confronti $\mathbf{y}_{ab}$. La quantità $\hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab})$ è il valore atteso di C in quanto C è una variabile bernoulliana di parametro p. Quindi il suo valore atteso, avendo osservato il confronto $\mathbf{y}_{ab}$, è:

$$E(C|\mathbf{y}_{ab}) = P(C = 1|\mathbf{y}_{ab}) = \\ = \frac{P(C = 1)P(\mathbf{Y} = \mathbf{y}_{ab}|C = 1)}{P(C = 0)P(\mathbf{Y} = \mathbf{y}_{ab}|C = 0) + P(C = 1)P(\mathbf{Y} = \mathbf{y}_{ab}|C = 1)},$$

   dove l'ultimo passaggio è giustificato dal teorema di Bayes. Sostituendo, si ottiene la formula per $\hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab})$.

3. Sempre all'iterazione $i+1$, il passo M consiste nel completare la funzione di verosimiglianza in (B.1) con i valori $\hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab})$ trovati al passo E. Si determinano quindi i valori dei parametri m_h, u_h e p che rendono massima tale funzione. Supponendo di essere in casi sufficientemente regolari, la massimizzazione può essere fatta nel modo usuale: si risolvono le equazioni ottenute derivando la funzione di verosimiglianza rispetto ai parametri incogniti e uguagliando a zero tali derivate, ottenendo:

$$\hat{m}^{(i+1)}_h = \frac{\sum_{(a,b)} \hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab}) y^h_{ab}}{\sum_{(a,b)} \hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab})}, \qquad h = 1, ..., k$$

$$\hat{u}^{(i+1)}_h = \frac{\sum_{(a,b)} (1 - \hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab})) y^h_{ab}}{\sum_{(a,b)} (1 - \hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab}))}, \qquad h = 1, ..., k$$

$$\hat{p}^{(i+1)} = \frac{\sum_{(a,b)} \hat{c}^{(i+1)}_{a,b}(\mathbf{y}_{ab})}{\nu_A \times \nu_B}.$$

4. Si iterano i passi M e E finché la differenza fra i parametri stimati in due iterazioni successive è inferiore a una soglia prefissata.

I passi ora definiti possono essere estesi dal modello (B.1) a modelli più complessi. Può accadere che per questi modelli non esistano stime di massima verosimiglianza dirette dei parametri incogniti, come quelle del punto 3. Ad esempio ciò accade in Larsen e Rubin (2001, si veda anche il paragrafo 7.2) dove si ipotizzano modelli loglineari generali. In questo caso, quando necessario, si devono sostituire le stime di massima verosimiglianza dirette dei parametri con quelle ottenibili attraverso metodi iterativi, come l'IPF (*Iterative Proportional Fitting*, si veda Haberman, 1974).

Appendice C

Il record linkage e la teoria dell'informazione

Nei capitoli precedenti è stato usato spesso il rapporto di verosimiglianze come strumento per il record linkage di due basi dati. Il rapporto di verosimiglianze è uno strumento fondamentale nella *teoria dell'informazione* (Kullback, 1959). Qui si interpretano alcuni risultati del record linkage secondo quest'ottica. In particolare si fa riferimento al lavoro di Copas e Hilton (1990) per valutare le funzioni di confronto più opportune, e a quello di Kirkendall (1985) per una giustificazione della regola di decisione di Fellegi e Sunter.

La capacità discriminatoria dei confronti

Tutte le funzioni di confronto $\mathbf{y}_{ab}$ sono funzione di quella definita dalla (2.13). Di conseguenza, la (2.13) possiede la maggiore ricchezza di informazioni. Copas e Hilton (2001) dimostrano in modo formale che i confronti del tipo (2.13) sono effettivamente i migliori, e a qualsiasi tipo di confronto diverso corrisponde una "perdita di informazione". Questa perdita viene misurata attraverso la distanza di Kullback-Leibler (Kullback, 1959). Per semplicità, consideriamo le seguenti notazioni per la distribuzione della variabile X^h:

$$
\begin{aligned}
p_{x_A,x_B} &= P\Big((X_A^h, X_B^h) = (x_A, x_B)\Big| c_{a,b} = 1\Big), \\
p_{x_A} &= P\left(X_A^h = x_A\right), \\
p_{x_B} &= P\left(X_B^h = x_B\right).
\end{aligned}
$$

Negli esempi 3.5 e 3.7 si è descritta la differenza fra le distribuzioni del confronto (2.13) quando la coppia di record è un match e quando è un non-match: infatti quando la coppia è un match la distribuzione è p_{x_A,x_B}, mentre quando è un non-match, la distribuzione assume la forma:

$$
P\Big((X_A^h, X_B^h) = (x_A, x_B)\Big| c_{a,b} = 0\Big) = p_{x_A} p_{x_B}.
$$

È estremamente importante che le due distribuzioni siano fra loro il più possibile lontane (come in figura 3.1.2). Infatti questo vorrebbe dire che il generico confronto $Y^h = (X_A^h, X_B^h)$ si distribuisce in modo totalmente diverso nelle due popolazioni $\mathcal{M}$ e $\mathcal{U}$, facilitando il compito della corretta assegnazione delle coppie (a, b) alle due popolazioni. Per valutare la distanza fra le due distribuzioni, torna utile la definizione di *distanza di Kullback-Leibler*. La distanza di Kullback-Leibler Δ_1 per questa coppia di distribuzioni prende ad oggetto il logaritmo del rapporto di verosimiglianza

$$
D = \log\left(\frac{p_{x_A,x_B}}{p_{x_A} p_{x_B}}\right),
$$

e è definito come il valore medio di D quando è vera la distribuzione per i match:

$$\Delta_1 = \sum_{(x_a,x_b)} p_{x_A,x_B} \log\left(\frac{p_{x_A,x_B}}{p_{x_A}p_{x_B}}\right).$$

Questa media è non negativa per la disuguaglianza di Jensen, e nulla se e solo se le due distribuzioni poste a confronto coincidono. Allo stesso modo, se è vera la distribuzione per i non-match, il valore medio di D è:

$$\Delta_2 = \sum_{(x_a,x_b)} p_{x_A}p_{x_B} \log\left(\frac{p_{x_A,x_B}}{p_{x_A}p_{x_B}}\right).$$

La Δ_2 è non positiva, sempre per la disuguaglianza di Jensen. La loro differenza:

$$\Delta = \Delta_1 - \Delta_2 = \sum_{(x_a,x_b)} (p_{x_A,x_B} - p_{x_A}p_{x_B}) \log\left(\frac{p_{x_A,x_B}}{p_{x_A}p_{x_B}}\right)$$

è nota nella teoria dell'informazione come la *divergenza simmetrizzata*. Questa è una misura del potere discriminatorio dei dati nel decidere fra le due distribuzioni. Una definizione diversa di confronto rispetto alla (2.13) fornisce una divergenza simmetrizzata Δ_c fra le corrispondenti distribuzioni $m(.)$ e $u(.)$ con:

$$\Delta_c \leq \Delta.$$

Ciò significa che per confronti diversi da (2.13) è più complicato (o nei casi migliori ugualmente complicato) discriminare le coppie di unità nelle due popolazioni dei match e dei non-match.

Una giustificazione per la procedura di Fellegi-Sunter

Fellegi e Sunter forniscono una dimostrazione dell'ottimalità della regola, nel senso della definizione 4.2 (si veda Fellegi e Sunter, 1969). Kirkendall (1985) fornisce una giustificazione dal punto di vista della teoria dell'informazione, estremamente utile per mettere in luce alcuni aspetti statistici rilevanti.

Kirkendall afferma che il peso (4.2) è strettamente legato a ciò che Kullback (1959) chiama "numero informativo" (*information number* nella versione originale). Si indichi con:

$$\Big\{P(C = c),\ c = 0, 1\Big\}$$

la distribuzione di probabilità che una coppia sia un match oppure no. Per il teorema di Bayes si ha:

$$P(C = 1|\mathbf{y}_{ab} = \mathbf{y}) = \frac{P(C = 1)m(\mathbf{y})}{P(C = 1)m(\mathbf{y}) + P(C = 0)u(\mathbf{y})},$$
$$P(C = 0|\mathbf{y}_{ab} = \mathbf{y}) = \frac{P(C = 0)u(\mathbf{y})}{P(C = 1)m(\mathbf{y}) + P(C = 0)u(\mathbf{y})}.$$

Il rapporto fra le due quantità precedenti conduce al rapporto:

$$\frac{P(C = 1|\mathbf{y}_{ab} = \mathbf{y})}{P(C = 0|\mathbf{y}_{ab} = \mathbf{y})} = \frac{P(C = 1)m(\mathbf{y})}{P(C = 0)u(\mathbf{y})},$$

il cui logaritmo dà:

$$\log\left(\frac{P(C=1|\mathbf{y}_{ab}=\mathbf{y})}{P(C=0|\mathbf{y}_{ab}=\mathbf{y})}\right)=\log\left(\frac{P(C=1)}{P(C=0)}\right)+\log\left(\frac{m(\mathbf{y})}{u(\mathbf{y})}\right).$$

Quindi:

$$\log\left(\frac{m(\mathbf{y})}{u(\mathbf{y})}\right)=\log\left(\frac{P(C=1)}{P(C=0)}\right)-\log\left(\frac{P(C=1|\mathbf{y}_{ab}=\mathbf{y})}{P(C=0|\mathbf{y}_{ab}=\mathbf{y})}\right). \tag{C.1}$$

Il rapporto delle logverosimiglianze in (C.1) fornisce una "misura" dell'informazione usata per discriminare le coppie una volta che si è osservato il vettore di confronto $\mathbf{y}$, e va sotto il nome di "numero informativo":

$$I(m,u;\mathbf{y})=\log\left(\frac{m(\mathbf{y})}{u(\mathbf{y})}\right).$$

L'informazione *media* per discriminare l'ipotesi:

- H_0: la distribuzione vera è $m(.)$

dall'ipotesi:

- H_1: la distribuzione vera è $u(.)$

è, sotto H_0:

$$\Delta(m,u)=\int_{\mathbf{y}\in\mathcal{D}} m(\mathbf{y})I(m,u;\mathbf{y})\,d\mathbf{y},$$

la distanza di Kullback-Leibler fra le distribuzioni $m(.)$ e $u(.)$. Le proprietà della distanza $\Delta(m,u)$, utili ai nostri scopi, sono:

- $\Delta(m,u)\geq 0$ qualunque siano le distribuzioni $m(.)$ e $u(.)$ poste a confronto;
- $\Delta(m,u)=0$ se e solo se le due distribuzioni coincidono.

Queste proprietà consentono di definire uno strumento di decisione fra le ipotesi H_0 (si considera la coppia un match) o H_1 (considero la coppia un non-match) una volta che si è osservato il risultato campionario $\mathbf{y}$ (osservazione di un vettore di confronto su una coppia). Kirkendall suggerisce di calcolare la distribuzione campionaria delle osservazioni, $\hat{p}(\mathbf{y})$, $\mathbf{y}\in\mathcal{D}$, che nel caso di una sola osservazione campionaria, ad esempio $\mathbf{y}_{ab}$, è:

$$\hat{p}(\mathbf{y})=\begin{cases}1 & \text{se } \mathbf{y}=\mathbf{y}_{ab}\\ 0 & \text{altrimenti.}\end{cases}$$

Quindi, si calcolano le informazioni medie $\Delta(\hat{p},m)$ e $\Delta(\hat{p},u)$ che misurano la "distanza" fra ognuna delle due ipotesi H_0 e H_1 e l'osservazione campionaria. Queste informazioni medie sono definite da:

$$\Delta(\hat{p},m)=\log\left(\frac{1}{m(\mathbf{y})}\right)$$

e

$$\Delta(\hat{p},u)=\log\left(\frac{1}{u(\mathbf{y})}\right).$$

La distribuzione che dista il meno possibile dall'osservazione (ovvero da $\hat{p}(.)$) è l'ipotesi (H_0 o H_1) con informazione media minore ed è quindi la decisione migliore da prendere. Per valutare quale fra le due informazioni medie è minore, si può considerare la differenza:

$$w(\mathbf{y}) = \Delta(\hat{p}, u) - \Delta(\hat{p}, m) = \log\left(\frac{m(\mathbf{y})}{u(\mathbf{y})}\right).$$

Quindi $w(\mathbf{y})$ coincide, tranne che per una trasformazione monotona (in proposito si veda il paragrafo 4.2), con il peso (4.2). La logica sottostante questa statistica test è sempre la stessa: più è piccolo $T(\mathbf{y})$ più l'osservazione $\mathbf{y}$ informa se la coppia che si sta analizzando è composta da unità diverse.