L'applicazione *RECLINK* per il *record linkage*: metodologia implementata e linee guida per la sua utilizzazione

Alessandra Nuccitelli (*), *Francesco Bosio* (**),
Luciano Fioriti (***)

(*) Direzione Centrale delle Statistiche sui Prezzi e il Commercio con l'Estero
(**) Direzione Centrale del Censimento della Popolazione, Territorio e Ambiente
(***) FINSIEL

Sommario

L'applicazione *RECLINK* è stata messa a punto per effettuare l'abbinamento automatico di *records* relativi ad una stessa unità statistica e contenuti in due diverse tabelle di un *database Oracle* residente su *server UNIX*.
In questo documento viene descritta la metodologia statistica implementata, basata su una logica bayesiana, e vengono forniti i dettagli necessari per l'utilizzazione dell'applicazione.

Abstract

The RECLINK software was specifically designed for matching records - related to the same statistical unit - from two tables of an Oracle database resident on UNIX server.
In this paper the implemented methodology, based on a Bayesian rationale, is described and directions for using the software are provided.

1. Introduzione[1]

Il termine 'abbinamento esatto' si riferisce all'uso di tecniche algoritmiche per identificare *records* relativi ad una stessa unità statistica e contenuti in archivi diversi. In questo documento, il termine *record linkage* verrà utilizzato come sinonimo di 'abbinamento probabilistico', ovvero con riferimento all'uso particolare di tecniche algoritmiche di tipo probabilistico.

L'applicazione *RECLINK* è stata messa a punto per poter effettuare il riconoscimento automatico, basato su un metodo probabilistico, di *records* relativi ad una stessa unità statistica contenuti in due diverse tabelle di un *database Oracle*.

Le finalità dell'abbinamento esatto sono brevemente illustrate nel paragrafo successivo; nel terzo viene introdotta la terminologia essenziale per una migliore comprensione dell'argomento e nel quarto si accenna all'importanza di elaborare opportunamente i dati prima di procedere con l'abbinamento. Le idee alla base della metodologia probabilistica implementata in *RECLINK* sono esposte nel paragrafo 5, mentre nel paragrafo 6 ci si sofferma sui dettagli necessari per l'utilizzazione del *software*. Infine, nel paragrafo 7 viene riportato un esempio di interfaccia utente predisposta in *Oracle Forms Developer* (Release 6i) per agevolare l'immissione dei valori di *input*, necessari per l'esecuzione del programma, e l'ispezione dell'*output* generato.

2. Finalità delle tecniche di abbinamento esatto

L'esigenza di collegare registrazioni relative ad una stessa unità statistica (individuo, famiglia, impresa, ecc.) provenienti da fonti diverse (censimenti, indagini campionarie, archivi amministrativi) è sempre più spesso avvertita sia nelle fasi di collezione, organizzazione e controllo di dati statistici che in fase di analisi dei dati. Fatta salva la

---

([1]) Francesco Bosio e Luciano Fioriti hanno sviluppato i sorgenti dell'applicazione *RECLINK* in linguaggio *C*; il paragrafo 6 è stato scritto congiuntamente da Francesco Bosio, Luciano Fioriti e Alessandra Nuccitelli; i rimanenti paragrafi sono stati scritti da Alessandra Nuccitelli.

necessità di garantire la riservatezza dei dati[2], dal momento che si opera su informazioni a livello individuale, l'impiego di tecniche di abbinamento esatto consente di sfruttare meglio il patrimonio informativo disponibile.

Gli obiettivi statistici possono essere differenti e riguardare, ad esempio:

- il contenimento dei costi di un'indagine e/o la riduzione dell'onere per il rispondente mediante omissione di domande relative ad informazioni reperibili da archivi già esistenti;
- la costituzione di una lista di unità da utilizzare come riferimento per un censimento o per l'estrazione di campioni;
- il supporto all'applicazione di tecniche di imputazione ed *editing*;
- la possibilità di svolgere analisi altrimenti precluse (ad esempio, quando si vogliano studiare gli effetti di alcuni fattori di rischio, nell'alimentazione o nel lavoro, sulle cause di morte, può essere necessario collegare i dati di un'indagine sull'alimentazione o sul lavoro con quelli desunti dai registri dei decessi);
- il conteggio delle unità presenti in uno oppure in più archivi (ad esempio, quando i microdati di un'indagine post-censuaria di controllo vengono confrontati con quelli censuari al fine di stimare il grado di copertura di un censimento).

Tra i progetti di rilievo sulla teoria e la pratica dell'abbinamento esatto di *records* realizzati in Italia meritano attenzione la costruzione di un archivio longitudinale per i dati dell'indagine trimestrale sulle forze di lavoro (Ceccarelli, Discenza, Rosati, Paggiaro, Torelli, 2002) e la costruzione di A.S.I.A., Archivio Statistico delle Imprese Attive (Cella, Garofalo, Paggiaro, Torelli, Viviano, 2003).

La mente umana è capace di abbinare *records* in maniera molto efficace: quando il valore del campo di un *record* è mancante o scritto in maniera errata, la soluzione viene trovata in modo intelligente tenendo conto di tutte le informazioni supplementari disponibili (e non necessariamente su supporto informatico). Tuttavia, anche quando la dimensione degli archivi coinvolti non è molto elevata, è utile disporre di una procedura che abbini automaticamente la

---

([2]) I problemi che esistono per lo sviluppo di un processo integrativo delle fonti sono molteplici e, oltre al problema della tutela della riservatezza, si ricordano quelli concettuali e definitori. La trattazione di tali argomenti, tuttavia, esula dallo scopo di questo documento.

maggior parte dei *records*, in modo che soltanto pochi casi, i più ambigui, possano essere risolti mediante ispezione manuale, con grande risparmio di tempo e con notevole guadagno in termini di accuratezza e riproducibilità dei risultati.

L'applicazione *RECLINK* è stata messa a punto allo scopo di poter effettuare l'abbinamento probabilistico di *records* relativi ad una stessa unità statistica e contenuti in due diverse tabelle di un *database Oracle*.

È possibile distinguere diversi tipi di abbinamento esatto - *uno a uno*, *molti a uno* oppure *molti a molti* - a seconda del numero di *records* che si riferiscono ad ogni unità statistica:

- *uno a uno*: ad ogni unità statistica corrisponde un solo *record* in ciascuno degli archivi da abbinare; in questo caso, per ogni unità statistica rappresentata in entrambi gli archivi è necessario riconoscere la coppia di *records* ad essa corrispondenti;
- *molti a uno*: ad ogni unità statistica possono corrispondere più *records* in uno degli archivi da abbinare;
- *molti a molti*: ad ogni unità statistica possono corrispondere più *records* in entrambi gli archivi da abbinare.

Sia nel caso *uno a molti* che nel caso *molti a molti*, per ogni unità statistica rappresentata in entrambi gli archivi è necessario riconoscere tutte le coppie di *records* ad essa corrispondenti.

La metodologia implementata in *RECLINK* effettua l'abbinamento del tipo *uno a uno*.

3. Terminologia essenziale

Denotiamo con $\mathcal{A}$ e $\mathcal{B}$ due archivi costituiti rispettivamente da $v_{\mathcal{A}}$ e $v_{\mathcal{B}}$ *records*. Ogni *record* si riferisce ad una singola unità della popolazione di riferimento ed è costituito da più campi o variabili; per un individuo potrebbero essere, ad esempio, *nome*, *cognome*, *sesso*, *anno di nascita*, *luogo di nascita*, ecc., per un'impresa, *codice di attività economica*, *ragione sociale*, *numero di addetti*, ecc.. Tali variabili, non solo godono spesso di un potere identificativo diverso nei confronti delle unità, ma possono risultare affette da errori in misura differente tra loro: ad esempio, un campo come il *sesso*, presentando due modalità,

contribuisce poco all'individuazione univoca di un individuo a differenza di un campo quale il *cognome* che, tuttavia, può essere spesso trascritto in maniera errata. Inoltre, alcune variabili nel corso del tempo possono assumere modalità diverse (ad esempio, si pensi allo *stato civile* di un individuo), rendendo più arduo il riconoscimento di *records* riferibili ad una stessa unità.

Si assume che un sottoinsieme di *k* variabili, dette variabili di abbinamento, sia comune ai due archivi. In presenza di variabili di abbinamento la cui combinazione costituisca una chiave certa ed univoca, il problema di riconoscere in maniera esatta *records* relativi ad una stessa unità si risolve mediante una semplice operazione automatica di abbinamento deterministico, vale a dire verificando se i valori assunti dalla chiave coincidono perfettamente nei due archivi. Tuttavia, spesso nella pratica, anche quando i *records* nei due archivi si riferiscono ad una stessa unità, tale coincidenza non si verifica a causa di valori mancanti o errori nelle variabili di abbinamento. In questi casi, è generalmente opportuno ricorrere a una procedura di abbinamento di tipo probabilistico.

Siano X_1, X_2, ..., X_k le variabili di abbinamento nei due archivi $\mathcal{A}$ e $\mathcal{B}$.

Archivio $\mathcal{A}$

UNITÀ	VARIABILI				
i	variabili di abbinamento				
1	$x_{11}^{\mathcal{A}}$	$x_{12}^{\mathcal{A}}$	.	$x_{1k}^{\mathcal{A}}$	. $x_{1n}^{\mathcal{A}}$
2	$x_{21}^{\mathcal{A}}$	$x_{22}^{\mathcal{A}}$	.	$x_{2k}^{\mathcal{A}}$	. $x_{2n}^{\mathcal{A}}$
.	.	.	.	.	. .
a	$x_{a1}^{\mathcal{A}}$	$x_{a2}^{\mathcal{A}}$	.	$x_{ak}^{\mathcal{A}}$	. $x_{an}^{\mathcal{A}}$
.	.	.	.	.	. .
$v_{\mathcal{A}}$	$x_{v_{\mathcal{A}}1}^{\mathcal{A}}$	$x_{v_{\mathcal{A}}2}^{\mathcal{A}}$	.	$x_{v_{\mathcal{A}}k}^{\mathcal{A}}$	. $x_{v_{\mathcal{A}}n}^{\mathcal{A}}$

Archivio $\mathcal{B}$

UNITÀ	VARIABILI				
i	variabili di abbinamento				
1	$x_{11}^{\mathcal{B}}$	$x_{12}^{\mathcal{B}}$	.	$x_{1k}^{\mathcal{B}}$	. $x_{1m}^{\mathcal{B}}$
2	$x_{21}^{\mathcal{B}}$	$x_{22}^{\mathcal{B}}$	.	$x_{2k}^{\mathcal{B}}$	. $x_{2m}^{\mathcal{B}}$
.	.	.	.	.	. .
.	.	.	.	.	. .
b	$x_{b1}^{\mathcal{B}}$	$x_{b2}^{\mathcal{B}}$	.	$x_{bk}^{\mathcal{B}}$	. $x_{bm}^{\mathcal{B}}$
.	.	.	.	.	. .
$v_{\mathcal{B}}$	$x_{v_{\mathcal{B}}1}^{\mathcal{B}}$	$x_{v_{\mathcal{B}}2}^{\mathcal{B}}$	.	$x_{v_{\mathcal{B}}k}^{\mathcal{B}}$	. $x_{v_{\mathcal{B}}m}^{\mathcal{B}}$

Il generico elemento $x_{aj}^{\mathcal{A}}$ (o $x_{bj}^{\mathcal{B}}$) rappresenta il valore osservato della variabile X_j sull'unità $a \in \mathcal{A}$ (o $b \in \mathcal{B}$).

Al fine di valutare se a e b siano o meno la stessa unità, occorre specificare come operare il confronto tra le realizzazioni delle *k* variabili di abbinamento sulla coppia (a,b).

Il modo che viene utilizzato in *RECLINK* per esprimere tale confronto è il seguente:

$$y_{ab}^{j} = \begin{cases} 1 & \text{se } x_{aj}^{\mathcal{A}} = x_{bj}^{\mathcal{B}} \\ 0 & \text{se } x_{aj}^{\mathcal{A}} \neq x_{bj}^{\mathcal{B}} \end{cases} \qquad j = 1, \quad 2, \quad ..., \quad k . \qquad [1]$$

Tale definizione si applica anche nel caso di valori mancanti, assumendo che y_{ab}^{j} sia uguale a 1 se e solo se i valori $x_{aj}^{\mathcal{A}}$ e $x_{bj}^{\mathcal{B}}$ siano simultaneamente uguali e non mancanti.

L'insieme delle coppie ordinate di unità

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

di numerosità $N = n_{\mathcal{A}} \times n_{\mathcal{B}}$, è l'unione di due insiemi disgiunti: l'insieme $\mathcal{M}$ delle coppie i cui elementi si riferiscono alla stessa unità (insieme dei *matches*)

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

e l'insieme $\mathcal{U}$ delle coppie i cui elementi si riferiscono ad unità differenti (insieme dei *non-matches*)

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

Un vettore di confronto $\mathbf{y}_{ab} = \{y_{ab}^{j}\}, \ j = 1, \quad 2, \quad ..., \quad k$, costituito da molti elementi uguali a 1 costituisce un evento frequente quando $(a,b) \in \mathcal{M}$ e raro quando $(a,b) \in \mathcal{U}$.

Indichiamo con $|\mathcal{E}|$ la cardinalità di un generico insieme $\mathcal{E}$. È importante osservare che $|\mathcal{M}|$ è tipicamente di gran lunga minore[3] di $|\mathcal{U}|$.

([3]) Ad esempio, supponendo che il numero di casi comuni ai due archivi $\mathcal{A}$ e $\mathcal{B}$ sia 90 e $n_{\mathcal{A}} = n_{\mathcal{B}} = 100$, allora $|\mathcal{M}| = 90$ e $|\mathcal{U}| = 9.910$.

La procedura di abbinamento esatto implementata in *RECLINK* combina opportunamente l'informazione fornita dai dati - i vettori di confronto $\mathbf{y}_{ab} = \{y_{ab}^j\}$, $j = 1, 2, ..., k$, osservati su ogni coppia (a,b) - con altre informazioni disponibili a priori al fine di individuare l'insieme di coppie che meglio rappresenti il vero ma incognito insieme M.

L'insieme L prodotto in *output* da *RECLINK* è costituito in prevalenza da coppie (a,b) i cui elementi a e b si riferiscono con probabilità relativamente elevata alla medesima unità statistica; per contro, l'insieme complementare di L rispetto ad $A \times B$, $A \times B - L$, è costituito in prevalenza da coppie (a,b) i cui elementi a e b si riferiscono con probabilità relativamente elevata ad unità statistiche differenti.

Nel seguito sarà utilizzata la denominazione *link*/*non-link* per riferirsi all'etichetta assegnata definitivamente ad una coppia alla fine di un processo di abbinamento (eventualmente costituito da una combinazione di più modalità di abbinamento, sia di tipo automatico che manuale) e la denominazione *match*/*non-match* per indicare la vera ma incognita condizione di appartenenza di una coppia all'insieme M o U.

Le coppie candidate ad essere etichettate come *links* sono perlopiù contenute in L; tuttavia è quasi sempre opportuno far seguire l'esecuzione di *RECLINK* da un passo di ispezione manuale su una parte degli insiemi L e $A \times B - L$ al fine di cautelarsi il più possibile dal rischio di commettere i seguenti due tipi di errore:

1) coppie appartenenti all'insieme M possono essere erroneamente etichettate come *non-links*;
2) coppie appartenenti all'insieme U possono essere erroneamente etichettate come *links*.

Il risultato dell'applicazione *RECLINK* consiste essenzialmente nella scrittura su tabelle apposite di un *database Oracle* di informazioni relative all'esito della decisione di abbinamento automatico per ogni coppia di *records*.

Le modalità di utilizzazione di queste tabelle di *output* non saranno prese specificamente in esame in questo documento dal momento che possono differire notevolmente a seconda dell'obiettivo statistico da raggiungere (vedi paragrafo 2), delle variabili di abbinamento utilizzate, ma anche da altre informazioni eventualmente disponibili. In fase di progettazione di un processo di abbinamento di ampia portata è comunque utile prevedere la

predisposizione di un'interfaccia opportuna che agevoli il lavoro di ispezione manuale dell'*output* generato da *RECLINK*. Un esempio di tale tipo di interfaccia viene riportato nel paragrafo 7.

4. Elaborazioni preliminari dei dati da sottoporre all'applicazione *RECLINK*

4.1 Standardizzazione di stringhe

La definizione [1] utilizzata per operare il confronto tra le realizzazioni delle k variabili di abbinamento sulla coppia (a,b) presuppone la standardizzazione preliminare di informazioni eventualmente presenti in formato libero (variabili di tipo stringa).

La standardizzazione di stringhe quali nomi, cognomi, indirizzi o ragioni sociali, permette di ridurre notevolmente gli errori di abbinamento. Infatti, spesso tali campi, pur essendo relativi ad una stessa unità statistica, possono non coincidere a causa della presenza di parole non significative, sigle e abbreviazioni (ad esempio, "`sig.`", "`dott.`", "`prof.`" per gli individui, "`v.`", "`p.zza`", "`v.le`" per gli indirizzi, "`spa`", "`scarl`", "`sas`" per le imprese), banali errori di trascrizione e/o registrazione, o modi alternativi utilizzati per denominare una stessa entità; conseguentemente molti *matches* potrebbero essere erroneamente etichettati come *non-links*. Per tale motivo, prima di procedere con l'esecuzione di *RECLINK*, è opportuno standardizzare variabili di abbinamento in formato libero.

Le possibili modalità di standardizzazione non verranno approfondite nel presente documento; si accenna soltanto al fatto che possono differire notevolmente a seconda delle variabili prese in considerazione (nomi, indirizzi, ragioni sociali) e in fase di progettazione di un processo di abbinamento dovrebbero essere predisposte in modo tale da ridurre, per quanto possibile, l'errore di tipo 1), tenendo sotto controllo l'errore di tipo 2).

Esempio 1. *Per la standardizzazione di una stringa contenente il nome ed il cognome di un individuo potrebbe essere utile adottare un algoritmo che trasformi tutti i caratteri in formato maiuscolo, elimini i caratteri speciali ("*`(`*", "*`)`*", "*`.`*", "*`;`*", "*`/`*", "*`%`*", ecc.) e le parole di supporto* ("`sig.`", "`dott.`", "`prof.`", *ecc.), separi la parte relativa al nome da quella*

relativa al cognome e restituisca in output due stringhe di 3 caratteri, una per il cognome e una per il nome, secondo le regole adottate per la costruzione del codice fiscale[4] *di un individuo:*

```
INPUT
variabile originale
dott. ROSSI Bruno

OUTPUT
cognome standardizzato          nome standardizzato
RSS                             BRN
```

Le stringhe generate in output potranno poi essere utilizzate come variabili di abbinamento.

La definizione di algoritmi per la standardizzazione di indirizzi risulta generalmente più complicata rispetto al caso illustrato nell'esempio 1, data la natura stessa della stringa.
Attualmente presso l'Istat è disponibile il *software SISTER* che si basa su un dizionario relativo a tutte le strade italiane (Caccia, 2001). Il *software* prende in *input* la stringa dell'indirizzo da standardizzare, la confronta con quelle presenti nel dizionario e, se la riconosce[5], la restituisce in forma standardizzata. Nel caso in cui la stringa non venga riconosciuta automaticamente occorre effettuare un recupero manuale.

4.2 Formazione di blocchi

Al fine di rendere più efficiente l'individuazione dei *matches*, solitamente la ricerca viene effettuata solo su sottoinsiemi disgiunti (o blocchi) di $\mathcal{A}\times\mathcal{B}$, invece che su $\mathcal{A}\times\mathcal{B}$. Un blocco è costituito da tutte le coppie che presentano una stessa combinazione di modalità di alcune delle variabili di abbinamento (preferibilmente non affette da errori)[6]. In tale contesto le

---

([4]) L'adozione delle regole utilizzate per la costruzione del codice fiscale può essere giustificata dalla considerazione che, per i nomi e cognomi italiani, gli errori di trascrizione e/o di registrazione risultano meno frequenti per le consonanti e per i caratteri iniziali.

([5]) La stringa dell'indirizzo da standardizzare può non essere trovata o riconosciuta: questo può accadere perché registrata con evidenti errori, o perché troppo distante dalla stringa a cui il *software* la ritiene più vicina, o anche perché associata a due o più alternative.

([6]) E' evidente che, anche per archivi di dimensione non elevata, il numero di confronti da effettuare risulta enorme. Per fare un esempio, l'abbinamento di due archivi, ognuno costituito da 10.000 *records*, richiederebbe

variabili di abbinamento vengono chiamate criteri di formazione dei blocchi o, più sinteticamente, variabili di bloccaggio. Una volta partizionati gli archivi in blocchi, è ovvio che le coppie di unità relative a blocchi individuati da differenti combinazioni di modalità delle variabili di bloccaggio saranno implicitamente etichettate come *non-links*. L'introduzione dei blocchi può comportare un aumento dell'errore di tipo 1) e/o una riduzione dell'errore di tipo 2). Se le variabili di bloccaggio non sono affette da errori, le coppie implicitamente etichettate come *non-links* sono effettivamente *non-matches*.
È evidente, quindi, che la scelta delle variabili di bloccaggio può incidere notevolmente sui risultati di un processo di abbinamento.

5. La metodologia implementata in *RECLINK*

5.1 Fondamenti teorici della metodologia implementata nei principali software generalizzati per il record linkage

I principali *software* di tipo generalizzato attualmente disponibili per effettuare il *record linkage* del tipo *uno a uno* - tra i quali, ad esempio, *CANLINK* sviluppato da *Statistics Canada*, *Record Linkage Software* sviluppato dal *Bureau of the Census*, *OXLINK* sviluppato dall'Università di Oxford - si basano essenzialmente sulla regola di decisione proposta da Fellegi e Sunter (1969).
La formalizzazione dei due statistici canadesi segue le linee della teoria classica della verifica delle ipotesi statistiche. Dato il vettore $\mathbf{y}_{ab}$, il problema dell'abbinamento viene affrontato mediante l'introduzione di una statistica in base alla quale discriminare tra le ipotesi H_0 - i *records* relativi alla coppia (a,b) si riferiscono alla stessa unità - e H_1 - i *records* relativi alla coppia (a,b) si riferiscono ad unità differenti. Fellegi e Sunter propongono, quale statistica test per discriminare tra H_0 e H_1, il rapporto di verosimiglianza:

---

di esaminare 100 milioni di coppie di *records*. Suddividendo (ipoteticamente) ognuno dei due archivi in 50 blocchi da 200 *records*, per ogni blocco sarebbe necessario esaminare 200×200 = 40.000 coppie di *records*; quindi, con il ricorso ai blocchi, il numero totale di coppie di *records* da esaminare ammonterebbe ad appena 2 milioni (40.000×50).

$$r = \frac{P(\mathbf{y}|H_0)}{P(\mathbf{y}|H_1)} = \frac{P(\mathbf{y}|(a,b) \in \mathcal{M})}{P(\mathbf{y}|(a,b) \in \mathcal{U})} = \frac{m_{\mathbf{y}}}{u_{\mathbf{y}}}, \qquad [2]$$

rapporto che cresce all'aumentare della concordanza tra le variabili di abbinamento.

La decisione di classificare la coppia (a,b) come appartenente a $\mathcal{M}$ o a $\mathcal{U}$ viene presa in base a due valori t_μ e t_λ $(t_\lambda \le t_\mu)$ che individuano gli estremi, rispettivamente inferiore e superiore, degli intervalli di accettazione e di rifiuto di H_0. La regola di abbinamento di Fellegi e Sunter prende allora la forma seguente:

se $r_{ab} \ge t_\mu$ la coppia viene etichettata come *link*;

se $t_\lambda < r_{ab} < t_\mu$ per decidere se la coppia è un *link* o un *non-link*, è necessario procedere ad un'ispezione manuale dei *records*;

se $r_{ab} \le t_\lambda$ la coppia viene etichettata come *non-link*.

Seppure possa essere derivata teoricamente, la definizione della regola di abbinamento proposta da Fellegi e Sunter si scontra con problemi operativi non indifferenti. Vale la pena di richiamare alcuni tra gli ostacoli più rilevanti che si incontrano nelle applicazioni pratiche e il cui tentativo di superarli ha finalizzato gran parte degli studi sul *record linkage* condotti a partire dal contributo classico dei due studiosi:

a) le funzioni di probabilità $m_{\mathbf{y}}$ e $u_{\mathbf{y}}$ sono ignote; è pertanto necessario ricorrere ad una loro stima;
b) occorre scegliere i valori soglia[7] t_λ e t_μ;
c) la regola di abbinamento si applica ad ogni coppia considerata singolarmente; per tenere conto dei vincoli di compatibilità esistenti tra le coppie - un *record* dell'archivio $\mathcal{A}$ può essere abbinato ad un solo *record* dell'archivio $\mathcal{B}$ - si ricorre generalmente a tecniche di ricerca operativa.

([7]) Per una soluzione ottimale al problema della scelta dei valori soglia, sarebbe necessario conoscere la distribuzione del vettore di confronto, condizionatamente alle due ipotesi H_0 e H_1. Nella maggior parte delle applicazioni, è usuale determinare i valori soglia in modo empirico, ricorrendo all'ispezione della distribuzione osservata del rapporto [2].

5.2 Fondamenti della metodologia implementata in RECLINK

La metodologia implementata in *RECLINK* si basa su una logica bayesiana[8] secondo cui l'informazione fornita dai dati (i vettori di confronto) viene combinata con quella disponibile a priori al fine di ottenere una distribuzione a posteriori sull'insieme $\mathcal{A} \times \mathcal{B}$ opportunamente vincolato. Lo spazio di tutti i possibili abbinamenti tra unità viene rappresentato mediante una matrice $\mathbf{C} = \{C_{ab}\}$, di dimensione $v_A \times v_B$, che può assumere valori nell'insieme $\mathcal{C}$ di tutte le matrici che soddisfano le seguenti condizioni:

$$C_{ab} = \begin{cases} 1 & \text{se} \quad (a,b) \in \mathcal{M} \\ 0 & \text{se} \quad (a,b) \in \mathcal{U} \end{cases}$$

$$\sum_{b=1}^{v_B} C_{ab} \leq 1 \qquad \forall\, a \in \mathcal{A},$$

$$\sum_{a=1}^{v_A} C_{ab} \leq 1 \qquad \forall\, b \in \mathcal{B};$$

le ultime due relazioni costituiscono i vincoli di compatibilità tra coppie appartenenti all'insieme $\mathcal{M}$.

Questo tipo di approccio permette di calcolare:

i. la probabilità che una coppia costituisca un *match*, condizionatamente ai dati osservati (i vettori di confronto);
ii. la probabilità congiunta che più coppie siano *matches*, condizionatamente ai dati osservati (i vettori di confronto).

A tale proposito è importante osservare che probabilità condizionate ad eventi osservabili (quelli relativi ai vettori di confronto) sono molto più direttamente interpretabili di probabilità condizionate ad eventi non osservabili (quelli relativi alle ipotesi $(a,b) \in \mathcal{M}$ e $(a,b) \in \mathcal{U}$) e, di conseguenza, di maggiore interesse ai fini pratici. Inoltre, la possibilità di calcolare la probabilità condizionata congiunta che più coppie siano *matches* costituisce un avanzamento rispetto alle soluzioni ottenute in base alla metodologia classica che generalmente fornisce

---

([8]) Il nucleo di tale metodologia è stato proposto in Fortini, Liseo, Nuccitelli, Scanu (2001).

regole di decisione per ogni coppia considerata singolarmente, senza tener conto dei vincoli di compatibilità esistenti.

Una volta nota la distribuzione a posteriori per $\mathbf{C}$, è necessario scegliere il valore della matrice rappresentante gli abbinamenti più plausibile tra tutte le matrici appartenenti all'insieme $\mathcal{C}$. Allo scopo possono essere definiti diversi stimatori puntuali, tra cui la moda a posteriori e il punto di minimo della funzione di perdita quadratica attesa a posteriori.

5.2.1 La funzione di verosimiglianza

Denotiamo con $\mathcal{D}$ l'insieme di tutti i possibili vettori **i** costituiti da *k* elementi uguali a 0 o a 1 (quindi $|\mathcal{D}| = 2^k$).

Idealmente, se tutte le *k* variabili di abbinamento fossero osservate senza errore in entrambi gli archivi $\mathcal{A}$ e $\mathcal{B}$, il vettore di confronto $\mathbf{y}_{ab}$ sarebbe costituito da elementi tutti uguali a 1, per le coppie (a,b) appartenenti all'insieme $\mathcal{M}$, e da elementi tutti uguali a 0, per le coppie (a,b) appartenenti all'insieme $\mathcal{U}$. Dal momento che il verificarsi di errori e di inconsistenze tra i due archivi risulta inevitabile, il vettore di confronto può assumere tutti i 2^k valori **i** con opportune probabilità.

Condizionatamente alle coppie appartenenti a $\mathcal{M}$, si assume per il vettore di confronto $\mathbf{Y}$ una distribuzione multinomiale con parametri:

$$\mathbf{m} = \{m_{\mathbf{i}}\}, \qquad m_{\mathbf{i}} \geq 0 \qquad \sum_{\mathbf{i}\in\mathcal{D}} m_{\mathbf{i}} = 1.$$

Condizionatamente alle coppie appartenenti a $\mathcal{U}$, si assume per il vettore di confronto $\mathbf{Y}$ una distribuzione ancora multinomiale, ma caratterizzata da un diverso insieme di parametri:

$$\mathbf{u} = \{u_{\mathbf{i}}\}, \qquad u_{\mathbf{i}} \geq 0 \qquad \sum_{\mathbf{i}\in\mathcal{D}} u_{\mathbf{i}} = 1.$$

La funzione di verosimiglianza[9] associata alle osservazioni dei confronti $\mathbf{y}_{ab}$ sulle $\nu_{\mathcal{A}} \times \nu_{\mathcal{B}}$ coppie è dunque la seguente:

$$L(\mathbf{c}, \mathbf{m}, \mathbf{u} | \mathbf{y}) = \prod_{a=1}^{\nu_{\mathcal{A}}} \prod_{b=1}^{\nu_{\mathcal{B}}} \left(\prod_{\mathbf{i} \in \mathcal{D}} m_{\mathbf{i}}^{\, d(\mathbf{y}_{ab}, \mathbf{i})} \right)^{c_{ab}} \left(\prod_{\mathbf{i} \in \mathcal{D}} u_{\mathbf{i}}^{\, d(\mathbf{y}_{ab}, \mathbf{i})} \right)^{1 - c_{ab}} = \quad [3]$$

$$= \prod_{\mathbf{i} \in \mathcal{D}} m_{\mathbf{i}}^{\sum_{a,b} d(\mathbf{y}_{ab}, \mathbf{i}) c_{ab}} u_{\mathbf{i}}^{\sum_{a,b} d(\mathbf{y}_{ab}, \mathbf{i})(1 - c_{ab})}$$

in cui

$$d(\mathbf{y}_{ab}, \mathbf{i}) = \begin{cases} 1 & \text{se } \mathbf{y}_{ab} = \mathbf{i} \\ 0 & \text{se } \mathbf{y}_{ab} \neq \mathbf{i} \end{cases} \qquad \mathbf{i} \in \mathcal{D}.$$

5.2.2 La distribuzione a priori per la matrice **C**

La distribuzione a priori per $\mathbf{C}$ è definita dalla seguente relazione[10]:

$$P(\mathbf{C} = \mathbf{c}) = \pi_H(h) P(\mathbf{C} = \mathbf{c} | H = h) \qquad \mathbf{c} \in C \qquad [4]$$

in cui $\pi_H(h)$ rappresenta la distribuzione a priori per il numero di *matches* h ($h = 0,\ 1,\ ...,\ \nu_{\mathcal{A}} \wedge \nu_{\mathcal{B}}$) e l'altro fattore è la distribuzione a priori per $\mathbf{C}$, condizionatamente al numero di *matches*.

Per H si ipotizza una distribuzione binomiale di parametro ξ

$$\pi_H(h) = P(H = h) = \binom{\nu_{\mathcal{A}} \wedge \nu_{\mathcal{B}}}{h} \xi^h (1 - \xi)^{\nu_{\mathcal{A}} \wedge \nu_{\mathcal{B}} - h}$$

([9]) Tale verosimiglianza si basa sull'ipotesi, non valida, di indipendenza tra le variabili $\mathbf{Y}$. Anche se non è chiaro l'effetto di tale assunzione sui risultati di una procedura di *record linkage*, si ritiene di poter utilizzare la verosimiglianza [3]. Per un approfondimento del problema del mancato rispetto dell'ipotesi di indipendenza, si veda Fortini, Nuccitelli, Liseo, Scanu (2002).

([10]) L'uguaglianza è valida poiché $P(\mathbf{C} = \mathbf{c}) = P(\mathbf{C} = \mathbf{c}, H = h)$, dal momento che, data una matrice $\mathbf{c}$, risulta determinato il numero di *matches* h.

con $\xi \in (0,\ 1)$ e $h = 0,\ 1,\ ...,\ \nu_A \wedge \nu_B$. ξ rappresenta PCOPPIA, uno dei parametri di *input* di *RECLINK* e dovrebbe essere scelto in base all'informazione disponibile sul numero di *matches* per l'applicazione specifica, in modo che la distribuzione di *H* renda più plausibili i valori di *h* che le informazioni a priori favoriscono. Ad esempio, ξ potrebbe essere calibrato utilizzando la frequenza relativa media dei *matches* osservati in applicazioni simili.

Invece, dal momento che generalmente non si dispone di informazioni sufficienti per 'privilegiare' certe matrici rispetto ad altre caratterizzate da uno stesso numero di *matches h*, per $(\mathbf{C}|H = h)$ viene adottata la distribuzione uniforme

$$P(\mathbf{C} = \mathbf{c}|H = h) = \frac{1}{h!\binom{\nu_A}{h}\binom{\nu_B}{h}} \qquad \mathbf{c} \in C.$$

5.2.3 La distribuzione a priori per i parametri delle distribuzioni multinomiali

La verosimiglianza [3] dipende oltre che da $\mathbf{c}$, che costituisce l'oggetto di inferenza, anche da $\mathbf{m}$ e $\mathbf{u}$, che assumono il ruolo di parametri di disturbo.

La funzione di verosimiglianza dipendente solo da $\mathbf{c}$ si ottiene analiticamente integrando la verosimiglianza [3] rispetto ai parametri di disturbo, dopo avere specificato per questi un'appropriata distribuzione iniziale. Si ipotizza che le distribuzioni a priori per i vettori aleatori $\mathbf{M}$ e $\mathbf{U}$ siano indipendenti da $\mathbf{C}$. Inoltre, si assume per essi una distribuzione di Dirichlet[11], caratterizzata rispettivamente dai seguenti vettori di iperparametri:

$$\boldsymbol{\alpha} = \{\alpha_{\mathbf{i}}\} \qquad \alpha_{\mathbf{i}} > 0 \qquad \forall\, \mathbf{i} \in \mathcal{D},$$

$$\boldsymbol{\beta} = \{\beta_{\mathbf{i}}\} \qquad \beta_{\mathbf{i}} > 0 \qquad \forall\, \mathbf{i} \in \mathcal{D}.$$

([11]) La classe delle densità di tipo Dirichlet è coniugata al modello multinomiale, il che comporta una semplificazione notevole dal punto di vista computazionale.

La calibrazione dei vettori di iperparametri $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ richiede alcune precisazioni. In *RECLINK* è stata adottata la seguente riparametrizzazione:

$$\alpha_{\mathbf{i}} = \delta^{\sum_{j=1}^{k} i_j - \phi} \qquad \beta_{\mathbf{i}} = \delta^{\phi - \sum_{j=1}^{k} i_j} \qquad\qquad \mathbf{i} \in \mathcal{D},$$

in cui δ e ϕ sono costanti da specificare opportunamente. Un valore di δ, scelto in modo che $\delta > 1$ (con $\phi \in \Re$), ordina i possibili vettori $\mathbf{i} \in \mathcal{D}$ in modo tale che la distribuzione a priori per $\mathbf{M}$ attribuisca maggiore probabilità a valori elevati di $m_{\mathbf{i}}$, quando $\mathbf{i}$ presenta molti 1, e a valori bassi di $m_{\mathbf{i}}$, quando $\mathbf{i}$ presenta molti 0. Per la distribuzione a priori per $\mathbf{U}$ valgono considerazioni opposte.

In *RECLINK* ϕ è posto pari a $k/2$ mentre δ rappresenta un parametro di *input* (`FATSC`).

5.2.4 La distribuzione a posteriori

La verosimiglianza [3], dopo l'operazione di integrazione rispetto ai parametri di disturbo, assume la forma seguente:

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

Applicando il teorema di Bayes, la distribuzione a posteriori per $\mathbf{C}$, a meno della costante di normalizzazione, risulta

$$P(\mathbf{C} = \mathbf{c}|\mathbf{y}) \propto \pi_H(h) P(\mathbf{C} = \mathbf{c}|H = h) L(\mathbf{c}|\mathbf{y}) \qquad \mathbf{c} \in C. \qquad [5]$$

5.2.5 Approssimazione della distribuzione a posteriori

Il numero di matrici in corrispondenza delle quali è definita la [5] è dato da

$$\sum_{h=0}^{\nu_{\mathcal{A}} \wedge \nu_{\mathcal{B}}} h! \binom{\nu_{\mathcal{A}}}{h} \binom{\nu_{\mathcal{B}}}{h} = \sum_{h=0}^{\nu_{\mathcal{A}} \wedge \nu_{\mathcal{B}}} \frac{\nu_{\mathcal{A}}! \nu_{\mathcal{B}}!}{(\nu_{\mathcal{A}} - h)! (\nu_{\mathcal{B}} - h)! h!},$$

cosicché la loro esplorazione esaustiva comporta un onere computazionale non indifferente. Inoltre l'esplorazione del supporto della distribuzione a posteriori è resa ancora più complicata dal fatto che non esiste un ordinamento naturale dei possibili valori della matrice $\mathbf{C}$.

Al fine di ottenere un'approssimazione accurata della distribuzione di interesse $P(\mathbf{C} = \mathbf{c}|\mathbf{y})$, non trattabile analiticamente, si ricorre ad un metodo del tipo *MCMC* (*Markov Chain Monte Carlo*). L'idea è quella di generare un campione dalla distribuzione d'interesse affrontando il problema in modo indiretto, vale a dire costruendo un'opportuna catena di Markov $\{\mathbf{C}^{(0)}, \mathbf{C}^{(1)}, \ldots, \mathbf{C}^{(t)}, \ldots\}$, con spazio degli stati C e distribuzione stazionaria propria $P(\mathbf{C} = \mathbf{c}|\mathbf{y})$.

Per costruire la catena di Markov in questione è stato utilizzato un algoritmo Metropolis-Hastings (Hastings, 1970). A partire da una matrice iniziale $\mathbf{c}^{(0)} \in C$, viene generata una realizzazione della catena nel modo seguente.

Se al passo (t) la catena si trova nello stato $\mathbf{c}^{(t)} \in C$, al passo $(t+1)$ viene estratta una matrice candidata $\mathbf{c}^{(*)} \in C$ da un'opportuna distribuzione di probabilità condizionata di proposta di spostamento $Q(\mathbf{C}^{(*)} = \mathbf{c}^{(*)} | \mathbf{C}^{(t)} = \mathbf{c}^{(t)})$, che possiamo definire 'ausiliaria'.

Allora lo stato della catena al passo $(t+1)$ viene scelto tra $\mathbf{c}^{(*)}$ e $\mathbf{c}^{(t)}$ in base a una regola di decisione che dipende sia dalla distribuzione di interesse che dalla distribuzione ausiliaria.

Più precisamente, al passo $(t+1)$, la catena si muove nello stato $\mathbf{c}^{(*)}$ con probabilità

$$\alpha_{MH}(\mathbf{c}^{(t)}, \mathbf{c}^{(*)}) = min\left\{1, \frac{P(\mathbf{C} = \mathbf{c}^{(*)}|\mathbf{y}) Q(\mathbf{C}^{(t)} = \mathbf{c}^{(t)} | \mathbf{C}^{(*)} = \mathbf{c}^{(*)})}{P(\mathbf{C} = \mathbf{c}^{(t)}|\mathbf{y}) Q(\mathbf{C}^{(*)} = \mathbf{c}^{(*)} | \mathbf{C}^{(t)} = \mathbf{c}^{(t)})}\right\},$$

altrimenti rimane nello stato $\mathbf{c}^{(t)}$.

Per maggiori dettagli su come viene definita la distribuzione di proposta di spostamento si rimanda a Nuccitelli (2001).

5.2.6 Stimatori puntuali

Una volta estratto un campione dalla complessa distribuzione a posteriori per $\mathbf{C}$ mediante applicazione del metodo *MCMC*, è necessario scegliere il valore della matrice rappresentante gli abbinamenti più plausibile tra tutte le matrici appartenenti all'insieme C. Allo scopo possono essere definiti diversi stimatori puntuali; in base alle simulazioni effettuate, gli stimatori più affidabili (tra quelli finora proposti) sono risultati (Nuccitelli, 2002):

a. il punto di massimo assoluto a posteriori (o uno dei punti di massimo assoluto a posteriori, se ne esiste più d'uno);
b. il punto di minimo della funzione di perdita quadratica attesa a posteriori.

Per quanto riguarda il punto di massimo assoluto, supponiamo che sia stata generata la seguente realizzazione della catena in T passi (T corrisponde in *RECLINK* al parametro di *input* `NITER`):

$$\left\{\mathbf{c}^{(0)},\ \mathbf{c}^{(1)},\ \ldots,\ \mathbf{c}^{(t)},\ \ldots,\ \mathbf{c}^{(T)}\right\};$$

RECLINK fornisce allora come stima del punto di massimo assoluto a posteriori la matrice

$$\hat{\mathbf{c}}_M = \arg \max_{\mathbf{c}\in\left\{\mathbf{c}^{(0)},\ \mathbf{c}^{(1)},\ \ldots,\ \mathbf{c}^{(t)},\ \ldots,\ \mathbf{c}^{(T)}\right\}} P\left(\mathbf{C}=\mathbf{c}\middle|\mathbf{y}\right).$$

Per quanto riguarda il punto di minimo della funzione di perdita quadratica attesa a posteriori, sia $\mathbf{c}^{v}$ la matrice vera e indichiamo con $I_{\mathcal{E}}(\cdot)$ la funzione indicatrice dell'insieme $\mathcal{E}$.

La funzione di perdita quadratica associata alla decisione $\mathbf{c}\in C$ è

$$L\left(\mathbf{c},\mathbf{c}^{v}\right)=\sum_{a,b}\left(c_{ab}^{v}-c_{ab}\right)^2.$$

La perdita attesa a posteriori associata alla decisione $\mathbf{c} \in C$ risulta

$$W_L(\mathbf{c}) = \sum_{a,b} c_{ab} P(C_{ab} = 0|\mathbf{y}) + \sum_{a,b} (1 - c_{ab}) P(C_{ab} = 1|\mathbf{y}) =$$

$$= \sum_{a,b} c_{ab} + \sum_{a,b} P(C_{ab} = 1|\mathbf{y}) - 2\sum_{a,b} c_{ab} P(C_{ab} = 1|\mathbf{y}) \quad . \qquad [6]$$

La ricerca della matrice $\hat{\mathbf{c}}_Q$ che rende minima $W_L(\mathbf{c})$ si svolge risolvendo, indipendentemente l'uno dall'altro, h ($h = 0,\ 1,\ ...,\ \nu_{\mathcal{A}} \wedge \nu_{\mathcal{B}}$) problemi di minimizzazione: per ogni h, viene individuata la matrice $\hat{\mathbf{c}}_Q^h \in C^h$ che rende minima la perdita attesa a posteriori $W_L(\mathbf{c})$ nel sottoinsieme C^h delle matrici caratterizzate dallo stesso numero di *matches h*

$$C^h = \left\{ \mathbf{c} \in C : \sum_{a,b} c_{ab} = h \right\}.$$

Per $h = 0$, la soluzione è banale, dal momento che C^0 è costituito dalla sola matrice $\mathbf{c}_Q^0$ i cui elementi c_{ab} sono nulli per ogni $a \in \mathcal{A}$ e per ogni $b \in \mathcal{B}$.
Fissato un valore di h ($h \geq 1$), la matrice $\hat{\mathbf{c}}_Q^h$ che minimizza la [6] è quella che rende massima la quantità[12]

$$\sum_{a,b} c_{ab} P(C_{ab} = 1|\mathbf{y})$$

con i vincoli

$$\sum_{b=1}^{\nu_{\mathcal{B}}} c_{ab} \leq 1 \qquad \forall\, a \in \mathcal{A},$$

([12]) Un aspetto interessante di questo problema di massimizzazione (per h fissato) è che la soluzione ottima ottenuta mediante un metodo di programmazione lineare continua, come il metodo del simplesso, è intera. Infatti, si può dimostrare che la matrice dei coefficienti dei vincoli lineari del problema è totalmente unimodulare (cioè ogni sua sottomatrice quadrata ha determinante pari a 0 o a ± 1) e quindi ogni soluzione basica ammissibile risulta intera.

$$\sum_{a=1}^{\nu_A} c_{ab} \le 1 \qquad \forall\, b \in \mathcal{B},$$

$$\sum_{a,b} c_{ab} = h,$$

$$c_{ab} \ge 0 \qquad \forall\, a \in \mathcal{A} \qquad \forall\, b \in \mathcal{B}.$$

La matrice $\hat{\mathbf{c}}_Q$ che rende minima la [6] in C è quindi data da

$$\hat{\mathbf{c}}_Q = \arg \min_{\mathbf{c} \in \left\{ \hat{\mathbf{c}}_Q^0,\ \hat{\mathbf{c}}_Q^1,\ \ldots,\ \hat{\mathbf{c}}_Q^h,\ \ldots,\ \hat{\mathbf{c}}_Q^{\nu_A \wedge \nu_B} \right\}} \{W_L(\mathbf{c})\}.$$

In *RECLINK* il parametro di *input* `FLAG_FLUSSO_ELABORAZIONE` permette di selezionare il tipo di stimatore (con "`B`" si sceglie il punto di massimo assoluto a posteriori, con "`S`" il punto di minimo della funzione di perdita quadratica attesa a posteriori).

6. Guida all'utilizzo di *RECLINK*

L'applicazione *RECLINK* esegue l'abbinamento probabilistico di *records* contenuti in due diverse tabelle di un *database Oracle* residente su *server UNIX*. L'applicazione *RECLINK* è disponibile nelle due versioni per ambienti *MS DOS*/*WINDOWS* e *UNIX* ed è stata sviluppata in linguaggio *C*; per accedere al *database Oracle* è stato utilizzato *Pro-C* (pre-compilatore fornito da *Oracle* per la scrittura di codice *Embedded SQL*).

Per far funzionare l'applicazione è necessario installare *Oracle Client*, nella versione completa per sviluppatori, sulla macchina che si intende utilizzare.

L'applicazione si compone di due *files*:

- ✓ `rl21win.exe` *file* eseguibile;
- ✓ `rl20.ini` *file* contenente la stringa per la connessione al *database Oracle*.

Il *file* `rl21win.exe` può essere collocato in una *directory* a piacere; il *file* `rl20.ini` deve risiedere nella stessa *directory* del *file* `rl21win.exe`.

Per lanciare il programma occorre:

1. aprire una sessione *DOS*;

2. posizionarsi sulla *directory* nella quale si trovano i *files* `rl21win.exe` e `rl20.ini`;
3. digitare "`rl21win`";
4. premere il tasto INVIO.

Il tempo d'esecuzione dell'applicazione è legato, ovviamente, alla dimensione delle tabelle da abbinare e alle risorse del *computer*[13] che si utilizza (frequenza del microprocessore, RAM).

Il *file* `rl20.ini` deve avere il seguente formato:

`nomeutente/password@DataSourceName` (*esempio*: `pd_sites/pdsites@ARG`)

Il *file* eseguibile `rl21win.exe` è stato creato a partire dai seguenti *files*:

- `rl_21_Main.c`
- `rl_21_Oracle.c`
- `rl_20_Bosio.c`
- `rl_21_Simplex.c`
- `simplx.c` — *file* proveniente da *Numerical Recipes Software* v. 2.08
- `simp1.c` — *file* proveniente da *Numerical Recipes Software* v. 2.08
- `simp2.c` — *file* proveniente da *Numerical Recipes Software* v. 2.08
- `simp3.c` — *file* proveniente da *Numerical Recipes Software* v. 2.08
- `nrutil.c` — *file* proveniente da *Numerical Recipes Software* v. 2.08
- `nr.h` — *file* proveniente da *Numerical Recipes Software* v. 2.08
- `nrutil.h` — *file* proveniente da *Numerical Recipes Software* v. 2.08
- `rl_20_Bosio.h` — prototipi di `rl_20_Bosio.c`
- `rl_21_Simplex.h` — prototipi di `rl_21_Simplex.c`
- `rl_20_Global.h` — *extern* delle variabili globali
- `oraSQL8.LIB` — libreria *Oracle*.

([13]) Si consiglia di utilizzare almeno un Pentium II 300 Mhz con 128 Mb di RAM.

6.1 La tabella T_CONFIGURAZIONE

Le variabili di abbinamento da utilizzare, le unità statistiche cui va applicata la regola di decisione implementata in *RECLINK* e i valori dei parametri di *input* necessari per l'esecuzione dell'algoritmo vanno opportunamente specificati dall'utente nella tabella T_CONFIGURAZIONE[14]:

Tabella T_CONFIGURAZIONE:

campo	*tipo e lunghezza*	*modalità di accesso da parte di RECLINK*
ID_CONFIG	number	lettura
PROVINCIA	varchar2(3)	lettura
COMUNE	varchar2(3)	lettura
SEZIONE	varchar2(7)	lettura
GRADO_ESITO	number	lettura
EPSILON	number	lettura
CONC	number	lettura
FATSC	number	lettura
PCOPPIA	number	lettura
NITER	number	lettura
NUM_CAMPI	number	lettura
SELECT_IDC	varchar2(1024)	lettura
SELECT_CEN	varchar2(1024)	lettura
VALORI_WHERE_IDC	varchar2(200)	lettura
VALORI_WHERE_CEN	varchar2(200)	lettura
NOME_FILE_SOMME_UNI	varchar2(200)	lettura
NOME_FILE_COPPIE_SOL	varchar2(200)	lettura
N_RECORD_LETTI_IDC	number	scrittura
N_RECORD_LETTI_CEN	number	scrittura
N_COPPIE_SCRITTE	number	scrittura
N_CHIAVI_SCRITTE	number	scrittura
TIMESTAMP	varchar2(100)	scrittura
FLAG_FLUSSO_ELABORAZIONE	varchar2(1)	lettura

I campi descritti qui di seguito contengono la configurazione dei valori di *input* necessari per l'esecuzione di *RECLINK*:

ID_CONFIG — numero che identifica univocamente ogni riga della tabella T_CONFIGURAZIONE; questo

([14]) La tabella T_CONFIGURAZIONE non deve avere campi con valore null.

	campo rappresenta la chiave della tabella T_CONFIGURAZIONE;
PROVINCIA	campo di comodo che può essere utilizzato per la formazione di blocchi (vedi sottoparagrafo 4.2);
COMUNE	come per PROVINCIA;
SEZIONE	come per PROVINCIA;
GRADO_ESITO	come per PROVINCIA;
EPSILON	numero arbitrariamente piccolo (solitamente viene posto pari a "1E-5") che rappresenta la probabilità di estrazione di una coppia in corrispondenza della quale il vettore di confronto osservato è nullo;
CONC	numero (solitamente viene posto pari a "1") che permette di enfatizzare la probabilità di estrazione di coppie con vettore di confronto osservato costituito da molti elementi uguali a 1;
FATSC	per la descrizione di questo campo si rimanda al sottoparagrafo 5.2.3;
PCOPPIA	per la descrizione di questo campo si rimanda al sottoparagrafo 5.2.2;
NITER	per la descrizione di questo campo si rimanda al sottoparagrafo 5.2.6;
NUM_CAMPI	numero dei campi da leggere dalla tabella NOMETAB_A (o NOMETAB_B) indicati nella clausola SELECT di SELECT_IDC (o SELECT_CEN);
SELECT_IDC	stringa *SQL* per l'interrogazione sulla tabella NOMETAB_A da sottoporre all'abbinamento; il primo campo della clausola SELECT deve essere rappresentato dalla chiave (contatore dei *records*)

della tabella NOMETAB_A - il programma lo tratterà diversamente rispetto agli altri campi; gli altri campi sono costituiti dalle variabili di abbinamento;

esempio:
```
SELECT  PRG, CAMPO2, CAMPO3, …, CAMPOn
FROM    NOMETAB_A
WHERE   CAMPOa >:pippo and CAMPOb =:pippo
or CAMPOc <:pippo
```

a tale proposito è importante sapere che:

- ":pippo" è un segnaposto; quindi al suo posto si può scrivere una stringa qualsiasi purché sia preceduta da ":";
- per l'inserimento di "is null" nella clausola WHERE non è previsto il segnaposto ma si deve scrivere direttamente "is null";

esempio:
```
SELECT  PRG, CAMPO2, CAMPO3, …, CAMPOn
FROM    NOMETAB_A
WHERE   CAMPOa >:pippo and CAMPOb =:pippo
or CAMPOc <:pippo and CAMPOd is
null
```

SELECT_CEN — stringa *SQL* per l'interrogazione sulla tabella NOMETAB_B da sottoporre all'abbinamento; sono valide le stesse considerazioni fatte per la tabella SELECT_IDC;

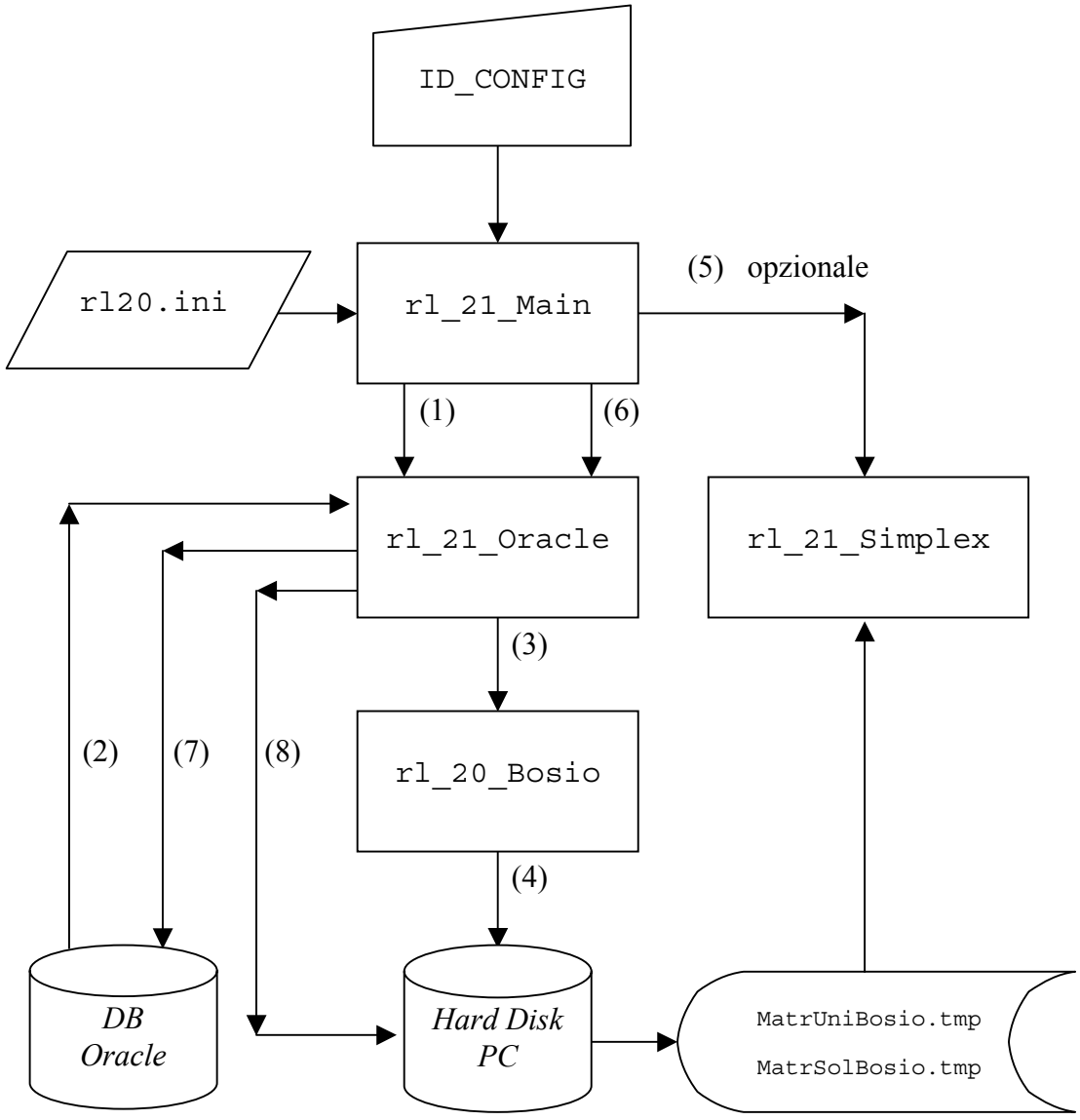
esempio:
```
SELECT  PRG, CAMPO2, CAMPO3, …, CAMPOn
FROM    NOMETAB_B
WHERE   CAMPOa >:pippo and CAMPOb =:pippo
or CAMPOc <:pippo
```

`VALORI_WHERE_IDC`	valori effettivi che vanno al posto di "`:pippo`"; devono ovviamente essere nello stesso ordine nel quale compaiono nella clausola `WHERE`; tali valori vanno separati con "`,`";
`VALORI_WHERE_CEN`	come per `VALORI_WHERE_IDC`;
`NOME_FILE_SOMME_UNI`	nome del *file* di *output* contenente, per ogni coppia di *records*, i valori delle rispettive chiavi nelle tabelle `NOMETAB_A` e `NOMETAB_B` ed il valore della probabilità a posteriori che la coppia costituisca un *match*; il nome del *file* può avere i seguenti formati: ▪ `fileout.txt` (in questo caso viene scritto nella *directory* in cui si trova l'eseguibile); ▪ `c:\\nomeDir\\fileout.txt` (in questo caso viene scritto nella *directory* indicata; attenzione, in *WINDOWS* va raddoppiato il carattere "\" – analogamente per *UNIX*);
`NOME_FILE_COPPIE_SOL`	nome del *file* di *output* contenente, per ogni coppia di *records* abbinati, i valori delle rispettive chiavi nelle tabelle `NOMETAB_A` e `NOMETAB_B`; il formato del nome del *file* è analogo a quello di `NOME_FILE_SOMME_UNI`;
`FLAG_FLUSSO_ELABORAZIONE`	per la descrizione di questo campo si rimanda al sottoparagrafo 5.2.6.

Per la descrizione dei campi contenenti informazioni di *output* - `N_RECORD_LETTI_IDC`, `N_RECORD_LETTI_CEN`, `N_COPPIE_SCRITTE`, `N_CHIAVI_SCRITTE`, `TIMESTAMP` - si rimanda al sottoparagrafo 6.3.

6.2 Schema di funzionamento di RECLINK

Lo schema della struttura interna di funzionamento dell'applicazione *RECLINK* è il seguente:



`rl_21_Main` Questo modulo prende in *input* il valore di `ID_CONFIG` e la stringa di connessione al *database Oracle* (contenuta nel *file* `rl20.ini`). Poi chiama la funzione `mainOracle` contenuta in `rl_21_Oracle` per leggere i dati contenuti nel *database Oracle* ed eseguire l'algoritmo implementato in `rl_20_Bosio`; nel caso in cui venga richiesta come soluzione il punto di minimo della funzione di perdita quadratica attesa a posteriori viene chiamato anche l'algoritmo implementato in

rl_21_Simplex. Alla fine richiama la funzione mainOracle per la scrittura dell'*output*.

rl_21_Oracle — Questo modulo accede al *database Oracle*, legge le tabelle NOMETAB_A e NOMETAB_B da abbinare (selezionate mediante le istruzioni *SQL* specificate nella configurazione contrassegnata da ID_CONFIG); chiama la funzione mainBosio contenuta in rl_20_Bosio per individuare il punto di massimo a posteriori; scrive i risultati sulle tabelle T_COPPIE, T_CHIAVI_NON_ACCOPPIATE, T_CONFIGURAZIONE e nei files di *output* il cui nome è stato specificato nei campi NOME_FILE_SOMME_UNI e NOME_FILE_COPPIE_SOL della tabella T_CONFIGURAZIONE.

rl_20_Bosio — Questo modulo individua il punto di massimo a posteriori; produce in *output* i due *files* MatrUniBosio.tmp e MatrSolBosio.tmp.

Il primo *file* contiene le coppie visitate dalla catena di Markov ad ogni iterazione (vedi sottoparagrafi 5.2.5 e 5.2.6) secondo il seguente formato: il primo campo rappresenta il numero di iterazione, il secondo campo contiene il numero d'ordine di lettura del *record* proveniente dalla tabella NOMETAB_A ed il terzo contiene il numero d'ordine di lettura del *record* proveniente dalla tabella NOMETAB_B.

Nel secondo *file* sono riportate le coppie soluzione che rappresentano il punto di massimo a posteriori; ognuna di esse è rappresentata su una riga mediante il numero d'ordine di lettura del *record* proveniente dalla tabella NOMETAB_A ed il numero d'ordine di lettura del *record* proveniente dalla tabella NOMETAB_B.

rl_21_Simplex — Questo modulo viene eseguito soltanto se nella configurazione contrassegnata da ID_CONFIG il campo FLAG_FLUSSO_ELABORAZIONE assume valore "S"; in tal caso, oltre al punto di massimo a posteriori già individuato da rl_20_Bosio, viene determinato il punto di minimo della funzione di perdita quadratica attesa a posteriori. I risultati scritti sulle tabelle T_

COPPIE, T_CHIAVI_NON_ACCOPPIATE, T_CONFIGURAZIONE e sui *files* di *output* (il cui nome è stato specificato nei campi NOME_FILE_SOMME_UNI e NOME_FILE_COPPIE_SOL) si riferiscono soltanto a quest'ultimo stimatore. I *files* MatrUniBosio.tmp e MatrSolBosio.tmp costituiscono l'*input* necessario per l'esecuzione di questo modulo.

6.3 Descrizione dell'output prodotto da RECLINK

Il risultato dell'applicazione *RECLINK* consiste nella scrittura sulle tabelle T_COPPIE, T_CHIAVI_NON_ACCOPPIATE, T_CONFIGURAZIONE e sui *files* NOME_FILE_SOMME_UNI e NOME_FILE_COPPIE_SOL di informazioni relative all'esito della decisione di abbinamento per ogni coppia di *records*.

Come già anticipato nel sottoparagrafo 6.1, il *file* NOME_FILE_SOMME_UNI contiene, per ogni coppia di *records*, i valori delle rispettive chiavi nelle tabelle NOMETAB_A e NOMETAB_B ed il valore della probabilità a posteriori che la coppia costituisca un *match*; il *file* NOME_FILE_COPPIE_SOL contiene per ogni coppia di *records* abbinati, i valori delle rispettive chiavi nelle tabelle NOMETAB_A e NOMETAB_B.

I campi di scrittura in *output* della tabella T_CONFIGURAZIONE sono N_RECORD_LETTI_IDC, N_RECORD_LETTI_CEN, N_COPPIE_SCRITTE, N_CHIAVI_SCRITTE, TIMESTAMP, secondo la descrizione che segue:

N_RECORD_LETTI_IDC	numero di *records* letti dalla tabella NOMETAB_A;
N_RECORD_LETTI_CEN	numero di *records* letti dalla tabella NOMETAB_B;
N_COPPIE_SCRITTE	numero di coppie di *records* abbinati;
N_CHIAVI_SCRITTE	numero totale di *records* non abbinati;
TIMESTAMP	data e ora di fine elaborazione.

Ogni riga della tabella T_COPPIE contiene informazioni relative ad ognuna delle coppie di *records* abbinati:

Tabella T_COPPIE:

campo	*tipo e lunghezza*	*modalità di accesso da parte di RECLINK*
ID_CONFIG	number	scrittura
CHIAVE_TAB_IDC	varchar2(20)	scrittura
CHIAVE_TAB_CEN	varchar2(20)	scrittura
PROB_COPPIA	number	scrittura
SOMMA_PROB_IDC	number	scrittura
SOMMA_PROB_CEN	number	scrittura
PROVINCIA	varchar2(3)	scrittura
COMUNE	varchar2(3)	scrittura
SEZIONE	varchar2(7)	scrittura
GRADO_ESITO	number	scrittura
TIMESTAMP	varchar2(100)	scrittura

ID_CONFIG	numero identificativo della configurazione dei valori di *input* sottoposta al programma;
CHIAVE_TAB_IDC	valore della chiave relativo al *record* abbinato proveniente dalla tabella NOMETAB_A;
CHIAVE_TAB_CEN	valore della chiave relativo al *record* abbinato proveniente dalla tabella NOMETAB_B;
PROB_COPPIA	valore della probabilità a posteriori che la coppia costituisca un *match*;
SOMMA_PROB_IDC	valore della probabilità a posteriori che il *record* proveniente dalla tabella NOMETAB_A faccia parte di un *match*;
SOMMA_PROB_CEN	valore della probabilità a posteriori che il *record* proveniente dalla tabella NOMETAB_B faccia parte di un *match*;
PROVINCIA	campo di comodo eventualmente utilizzato per la formazione di blocchi;
COMUNE	come per PROVINCIA;

SEZIONE	come per PROVINCIA;
GRADO_ESITO	come per PROVINCIA;
TIMESTAMP	data e ora di fine elaborazione.

Ogni riga della tabella T_CHIAVI_NON_ACCOPPIATE contiene informazioni relative ad ogni *record* non abbinato:

Tabella T_CHIAVI_NON_ACCOPPIATE:

campo	*tipo e lunghezza*	*modalità di accesso da parte di RECLINK*
ID_CONFIG	number	scrittura
TABELLA_PROVENIENZA	varchar2(50)	scrittura
CHIAVE	varchar2(20)	scrittura
PROVINCIA	varchar2(3)	scrittura
COMUNE	varchar2(3)	scrittura
SEZIONE	varchar2(7)	scrittura
GRADO_ESITO	number	scrittura
TIMESTAMP	varchar2(100)	scrittura

ID_CONFIG	numero identificativo della configurazione dei valori di *input* sottoposta al programma;
TABELLA_PROVENIENZA	tabella di provenienza del *record* non abbinato;
CHIAVE	valore della chiave relativo al *record* non abbinato;
PROVINCIA	campo di comodo eventualmente utilizzato per la formazione di blocchi;
COMUNE	come per PROVINCIA;
SEZIONE	come per PROVINCIA;
GRADO_ESITO	come per PROVINCIA;
TIMESTAMP	data e ora di fine elaborazione.

7. Esempio di interfaccia per l'immissione dei valori di *input* e per la revisione manuale dell'*output* di *RECLINK*

Come esempio di interfaccia predisposta per agevolare l'immissione dei valori di *input* e la revisione manuale dell'*output* prodotto da *RECLINK*, si riporta quella realizzata in *Oracle Forms Developer* (Release 6i)[15] per l'abbinamento dei *records* del 14° Censimento della Popolazione con quelli della relativa Indagine sul grado di Copertura.

L'Indagine sul grado di Copertura del 14° Censimento della Popolazione (d'ora in poi IDC) si è svolta a circa un mese di distanza dalle operazione censuarie, enumerando di nuovo la popolazione residente in un campione areale di sezioni di censimento. Le stime del grado di copertura del Censimento sono prodotte tramite il confronto e l'abbinamento dei dati elementari relativi alle due rilevazioni, IDC e Censimento (d'ora in poi CEN).

L'abbinamento tra le registrazioni censuarie e quelle di un'indagine sul grado di copertura costituisce generalmente un processo complesso il cui scopo è determinare il numero comune di famiglie e di individui rilevati in entrambe le occasioni. Dal momento che errori di abbinamento, anche molto contenuti, possono dar luogo a distorsioni di entità non trascurabile nelle stime del grado di copertura, l'intero processo deve essere progettato in modo tale da produrre risultati il più possibile accurati.

In occasione del Censimento della Popolazione del 2001, il processo di abbinamento è stato concepito come la composizione di tre fasi sequenziali, distinte a seconda della tipologia delle unità da abbinare (Nuccitelli, 2004): schematicamente si procede prima abbinando le famiglie (Fase 1), poi gli individui all'interno delle famiglie abbinate (Fase 2), infine gli individui residui non abbinati, a prescindere dalla loro famiglia di appartenenza (Fase 3). Ognuna delle tre fasi è costituita da una combinazione, più o meno articolata, delle varie modalità possibili di abbinamento, per cui si abbinano automaticamente, in modo deterministico e/o probabilistico, la maggior parte delle unità ed i restanti casi, più ambigui, mediante operazioni di tipo manuale.

Dal momento che il livello territoriale più fine rispetto al quale sono stati rilevati i dati in entrambe le occasioni è costituito dalle sezioni di censimento, il processo di abbinamento è

([15]) L'interfaccia è stata predisposta da Claudia Cianfarani e Fabrizio Delli Priscoli.

stato progettato in modo tale da elaborare i dati di una *sezione di censimento* campione alla volta. Quindi, la sezione di censimento costituisce, insieme alla *provincia* e al *comune*, una variabile di bloccaggio[16].

Le schermate che vengono mostrate in seguito a titolo esemplificativo si riferiscono al passo di *record linkage* eseguito mediante *RECLINK* e relativo alla Fase 1 di abbinamento delle famiglie residenti nella sezione "2275" del comune di Bologna.

Tale passo viene preceduto da un'operazione di abbinamento deterministico la cui chiave è costituita dalle seguenti variabili:

- *nome dell'intestatario del questionario*;
- *cognome dell'intestatario del questionario*;
- *indirizzo standardizzato*[17];
- *numero civico*.

Vengono sottoposte al passo di *record linkage* tutte le famiglie non abbinate al passo di abbinamento deterministico. La chiave di abbinamento utilizzata nel passo di *record linkage* è costituita dalle seguenti variabili:

- *nome standardizzato dell'intestatario del questionario*[18] (NOME_INT);
- *cognome standardizzato dell'intestatario del questionario* (COGNOME_INT);
- *indirizzo standardizzato* (INDIRI_INT);
- *numero civico* (NUMCIV_INT);
- *sesso dell'intestatario del questionario* (SESSO_INT);
- *giorno di nascita dell'intestatario del questionario* (GGNAS_INT);
- *mese di nascita dell'intestatario del questionario* (MMNAS_INT);
- *anno di nascita dell'intestatario del questionario* (AANAS_INT);
- *numero totale di maschi nella famiglia* (NUM_MASCHI_FAM);
- *numero totale di femmine nella famiglia* (NUM_FEMMINE_FAM).

([16]) Tale scelta ha reso necessaria la risoluzione di eventuali errori di geocodifica presenti nei dati, prima di procedere con l'abbinamento, in modo da evitare di introdurre l'errore di tipo 1).

([17]) In questo contesto, la standardizzazione della variabile *indirizzo* è stata effettuata mediante l'utilizzo di un dizionario, appositamente costruito, comprendente tutte le strade facenti parte degli itinerari percorsi da ciascun rilevatore nelle sezioni campione.

([18]) Per la standardizzazione delle variabili *nome* e *cognome* è stato adottato l'algoritmo cui si è accennato nell'Esempio 1.

La schermata di Figura 1 consente all'utente di inserire tutte le informazioni necessarie per configurare in modo adeguato il programma *RECLINK*.

Dal momento che - viste le caratteristiche delle rilevazioni coinvolte - ci si aspetta un numero molto elevato di abbinamenti, PCOPPIA è stato posto pari a "0.98".

La stringa SELECT_IDC utilizzata per l'interrogazione sulla tabella T_FAMIGLIA_IDC (contenente i *records* relativi alle famiglie rilevate nell'occasione IDC) è la seguente:

```
SELECT    PRGFAM_IDC, COGNOME_INT_NORM, NOME_INT_NORM,
          INDIRI_INT, NUMCIV_INT, SESSO_INT, GGNAS_INT,
          MMNAS_INT, AANAS_INT, NUM_MASCHI_FAM,
          NUM_FEMMINE_FAM
FROM      T_FAMIGLIA_IDC
WHERE     CODPRB=:PROVINCIA AND CODCOB=:COMUNE AND
          NSEZB=:SEZIONE AND LINKATA=:LINKATA
```

La stringa SELECT_CEN utilizzata per l'interrogazione sulla tabella T_FAMIGLIA_CEN (contenente i *records* relativi alle famiglie rilevate nell'occasione CEN) è la seguente:

```
SELECT    PRGFAM_CEN, COGNOME_INT_NORM, NOME_INT_NORM,
          INDIRI_INT, NUMCIV_INT, SESSO_INT, GGNAS_INT,
          MMNAS_INT, AANAS_INT, NUM_MASCHI_FAM,
          NUM_FEMMINE_FAM
FROM      T_FAMIGLIA_CEN
WHERE     CODPRB=:PROVINCIA AND CODCOB=:COMUNE AND
          NSEZB=:SEZIONE AND LINKATA=:LINKATA
```

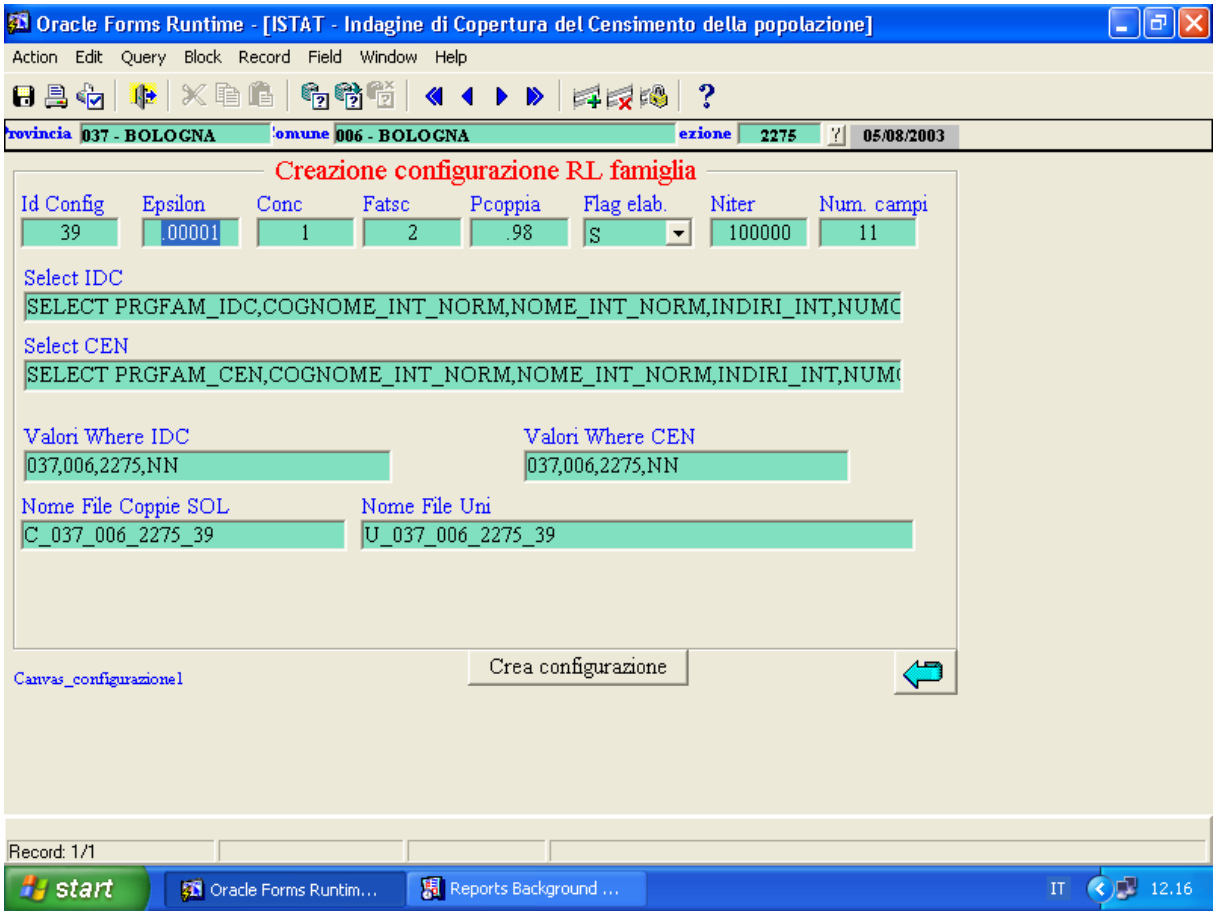
I campi PRGFAM_IDC e PRGFAM_CEN rappresentano la chiave (contatore dei *records*) rispettivamente delle tabelle T_FAMIGLIA_IDC e T_FAMIGLIA_CEN.

Il campo LINKATA assume il valore "SI" oppure "NN" a seconda che il *record* corrispondente sia stato abbinato o meno nel passo precedente di abbinamento deterministico.

La configurazione dei valori di *input* viene contrassegnata automaticamente mediante il numero identificativo "39".

Il campo di comodo `GRADO_ESITO` non è visualizzato in quanto non utilizzato; per comodità in questo contesto ad esso viene assegnato automaticamente il valore di `ID_CONFIG`.

Figura 1 – Esempio di schermata per l'immissione dei valori dei parametri di input



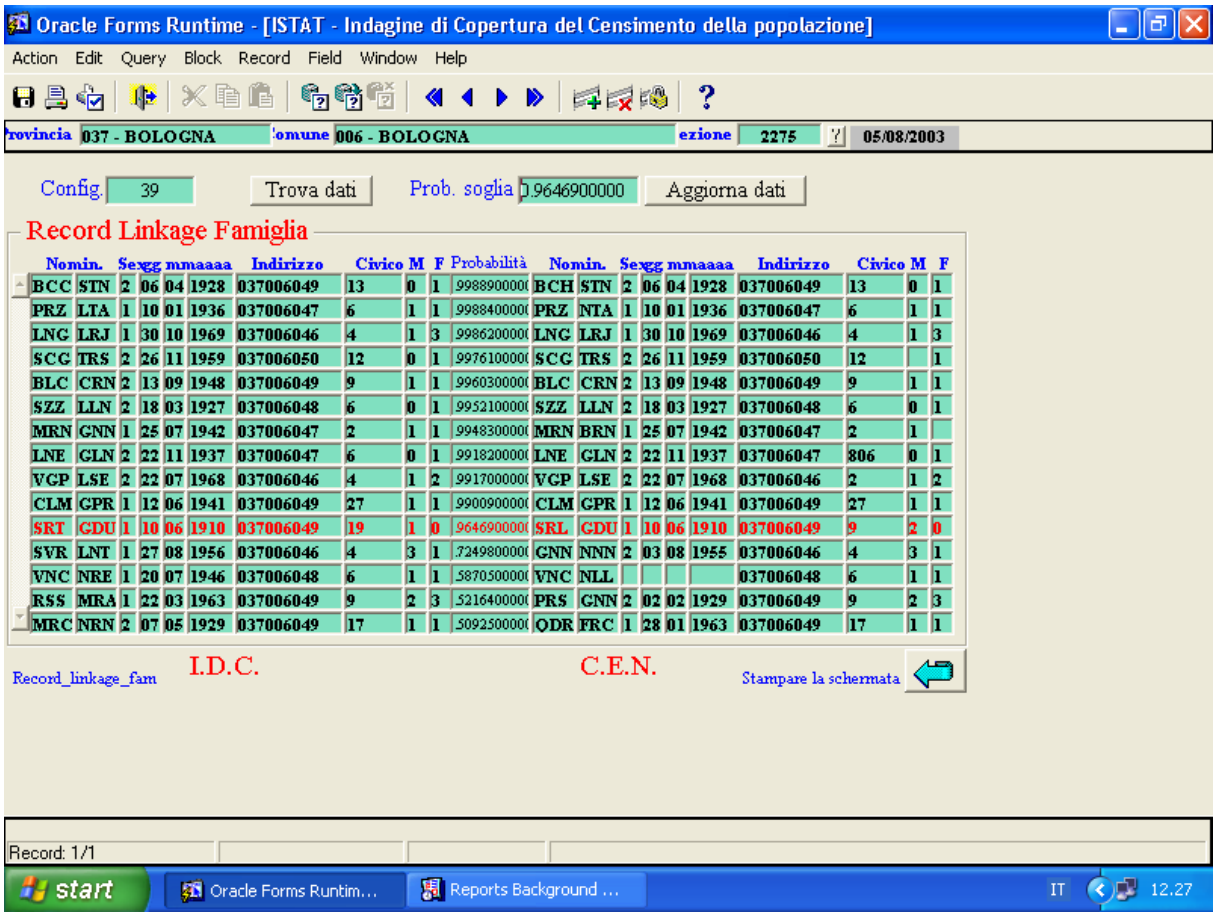
Con riferimento alla configurazione creata, *RECLINK* fornisce in *output* l'insieme di coppie soluzione (punto di minimo della funzione di perdita attesa a posteriori) visualizzato nella schermata di Figura 2. Ogni riga presenta i valori delle variabili di abbinamento, rispettivamente nelle occasioni IDC e CEN, per una certa coppia soluzione; il valore centrale rappresenta la probabilità a posteriori che la coppia proposta costituisca un *match*. Come si può vedere, l'insieme di abbinamenti fornito in *output* è costituito in prevalenza da coppie di *records* che con probabilità molto elevata si riferiscono ad una stessa famiglia. Tuttavia in

questo contesto, al fine di cautelarsi il più possibile dal rischio di etichettare come *links* coppie appartenenti all'insieme dei *non-matches*, un sottoinsieme delle coppie soluzione viene sottoposto a revisione manuale[19].

Più precisamente, le coppie soluzione vengono presentate secondo il valore decrescente della probabilità a posteriori; sulla base delle discordanze osservate per una stessa coppia tra le modalità delle variabili di abbinamento corrispondenti nelle occasioni IDC e CEN, l'utente può scegliere un opportuno valore soglia della probabilità. Tutte le coppie caratterizzate da una probabilità inferiore a tale valore soglia vengono sottoposte ad un esame più approfondito per essere confermate o meno. Nell'esempio preso in esame, il valore soglia della probabilità che meglio distingue gli accoppiamenti che quasi sicuramente si riferiscono ad una stessa famiglia da quelli più incerti è costituito da "`0.9646900000`". Dopo tale valore si possono osservare discordanze importanti tra le modalità assunte dalle variabili di abbinamento.

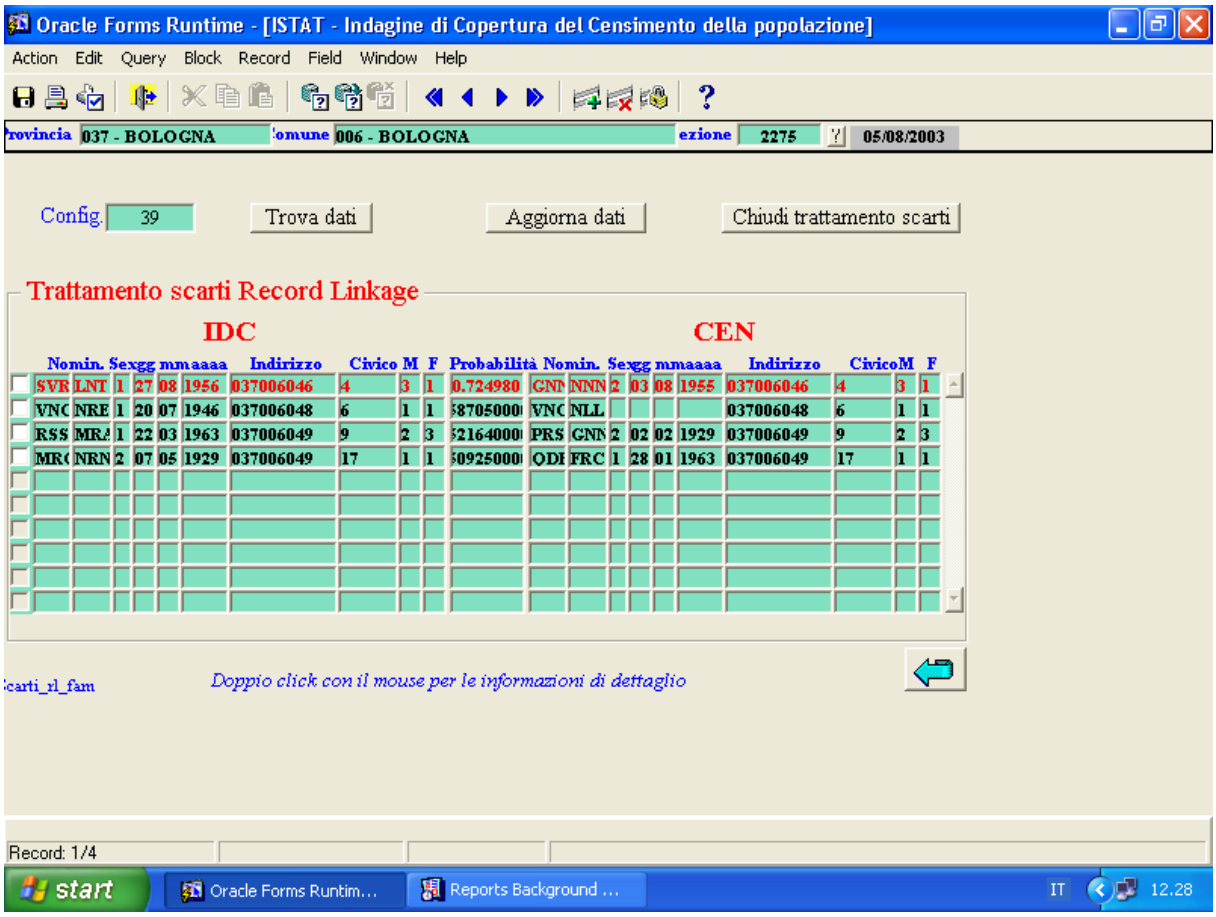
([19]) Al fine di cautelarsi il più possibile dal rischio di etichettare come *non-links* coppie appartenenti all'insieme dei *matches*, è generalmente opportuno sottoporre a revisione anche un sottoinsieme dei *records* non abbinati. Nel contesto specifico tale revisione non viene effettuata nella Fase 1 di abbinamento delle famiglie, ma in qualche modo rinviata dopo il passo di *record linkage* della Fase 3 ed eseguita con riferimento agli individui (per ulteriori dettagli si rimanda a Nuccitelli, 2004).

Figura 2 – Esempio di schermata per la revisione dell'output generato da RECLINK



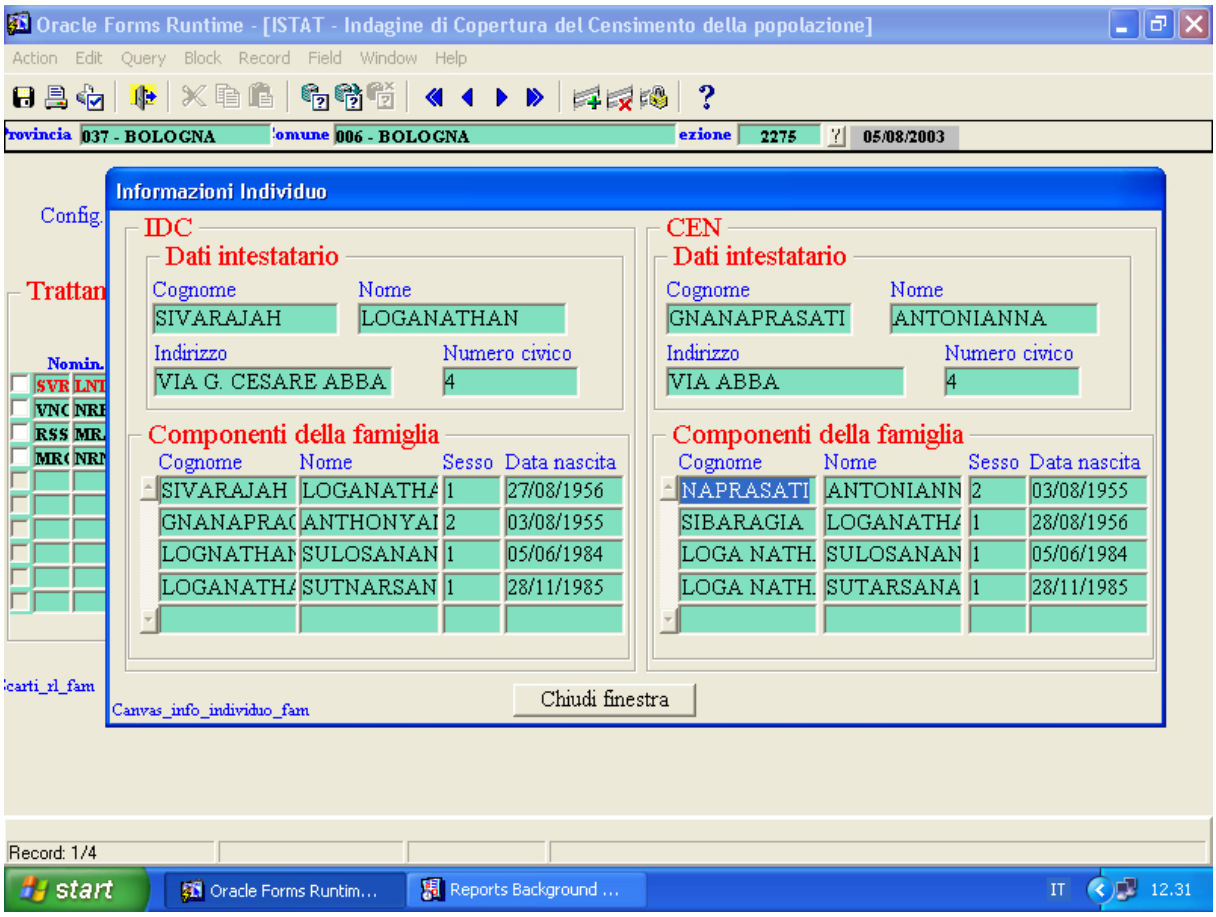
La schermata successiva (Figura 3) permette di confermare o meno, mediante biffatura della casella bianca corrispondente, ogni coppia sottoposta a revisione manuale. Al fine di decidere sull'eventuale scioglimento di un legame, all'utente viene data la possibilità di accedere ad informazioni supplementari.

Figura 3 – Esempio di schermata per l'eventuale scioglimento di coppie



Ad esempio, effettuando un doppio-clic sulla linea corrispondente alla prima coppia da controllare di Figura 3, appare la schermata di Figura 4 contenente, oltre ad alcune notizie - in versione non standardizzata - relative all'intestatario del questionario, anche informazioni di dettaglio relative agli altri componenti della famiglia (*nome*, *cognome*, *sesso* e *data di nascita*). Come si può constatare, nel caso preso in esame i *records* proposti per l'accoppiamento si riferiscono alla stessa famiglia, dal momento che le discordanze osservate tra le modalità delle variabili di abbinamento risultano imputabili al fatto di aver assunto come intestatario del questionario una persona diversa nelle occasioni IDC e CEN.

Figura 4 – Esempio di schermata riportante le informazioni di dettaglio



Bibliografia

Caccia M. (2001). *SISTER. Manuale tecnico*, DWI s.r.l., Verona.

Ceccarelli C., Discenza A. R., Rosati S., Paggiaro A., Torelli N. (2002). *Le matrici di transizione della Rilevazione trimestrale sulle forze di lavoro*. Nota metodologica presentata il 12 dicembre 2002 al seminario "Flussi del mercato del lavoro dal 1998 al 2002" presso l'Istat (http://www.istat.it/Comunicati/Fuori-cale/allegati/Mobilit--m/nota-metodologica.pdf).

Cella P., Garofalo G., Paggiaro A., Torelli N., Viviano C. (2003). Demografia d'impresa: l'utilizzo di tecniche di abbinamento per l'analisi della continuità, *Contributi ISTAT*, 5.

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

Fortini M., Liseo B., Nuccitelli A., Scanu, M. (2001). On Bayesian record linkage, *Research in Official Statistics*, 4, 185-198.

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*, New York, August 11-15, 2002.

Hasting W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, 57, 97-109.

Nuccitelli A. (2001). *Integrazione di dati mediante tecniche di abbinamento esatto: una rassegna critica ed una proposta in ambito bayesiano*, Tesi di dottorato in *Metodi Statistici per l'Economia e per l'Impresa*, XIII ciclo, Università degli Studi di Roma Tre, Roma.

Nuccitelli A. (2002). Bayesian point estimators for record linkage, *Atti della XLI Riunione Scientifica della Società Italiana di Statistica*, Sessioni Spontanee, Università Milano-Bicocca, 5-7 giugno 2002, 559-562.

Nuccitelli A. (2004). Integrating the 2001 Population Census and Post-Enumeration Survey data: the Italian experience, *Proceedings of the European Conference on Quality and Methodology in Official Statistics*, Mainz, Germany, May 24-26, 2004.

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P. (1993). *Numerical recipes in C: the art of scientific computing*, Cambridge University Press.

Yancey W. E., Winkler, W. E. (2002). *Record Linkage Software. User documentation*, U. S. Bureau of the Census, Washington.