



**Identificazione, sperimentazione e validazione di alcuni indicatori di qualità delle attività sanitarie e socio-sanitarie del territorio**

**Roma 14, 15 settembre 2006**

**ASSR**

AGENZIA per i  
SERVIZI SANITARI  
REGIONALI

UNIVERSITÀ DEGLI STUDI DI ROMA  
**U Tor Vergata**

**Il record linkage probabilistico – metodi di implementazione**

- La stima dei parametri
- Funzioni di confronto tra stringhe
- La necessità del blocking
- La determinazione dei valori soglia
- La stima degli errori, studi in letteratura
- Misure di performance
- Forzare 1-1 linkage
- Sistemi di record linkage

Fellegi e Sunter propongono due metodi per la stima dei parametri:

- Il primo assume che siano disponibili informazioni a priori sulla distribuzione delle variabili utilizzate per il confronto nelle due popolazioni (ad esempio cognome, nome, data di nascita), così come sulle probabilità dei differenti tipi di errori introdotti nei file dal processo di generazione dei record.



In questo modo è possibile fare delle stime dei parametri che tengano conto della frequenza con cui si distribuisce un variabile nella popolazione (frequency based)

Linkage probabilistico: metodi – *La stima dei parametri*

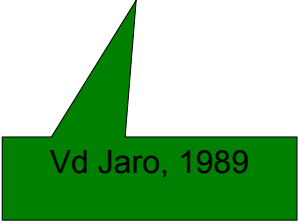
- Il secondo metodo utilizza le informazioni presenti nei due file per stimare le probabilità di interesse (nessuna informazione a priori).

Questo secondo metodo richiede confronti dicotomici e viene presentato per 3 variabili che soddisfino l'assunto di indipendenza delle probabilità condizionate.

Nel caso in cui le variabili siano più di tre sarà possibile applicare il metodo dei momenti per determinare le stime delle probabilità di interesse.

Linkage probabilistico: metodi – *La stima dei parametri*

Dal momento che il metodo dei momenti si è dimostrato numericamente instabile in alcune applicazioni e con alcune misture di distribuzioni, si è pensato di utilizzare metodi basati sulla massima verosimiglianza come l'algoritmo EM (Expectation-Maximization).



Vd Jaro, 1989

L'algoritmo EM, introdotto da Dempster e Laird, Si utilizza in presenza di dati mancanti, per modelli di misture finite o anche quando la funzione di verosimiglianza non sia trattabile analiticamente.

Linkage probabilistico: metodi – *La stima dei parametri*

Secondo Winkler (1999) in alcuni esperimenti preliminari il match basato sulle frequenze aveva dato risultati migliori rispetto al matching basato sulla semplice concordanza/discordanza.

Tuttavia grazie allo sviluppo di metodi più sofisticati per la stima dei parametri di concordanza/discordanza, quali gli algoritmi EM, il secondo metodo aveva dato risultati migliori

Linkage probabilistico: metodi – *La stima dei parametri*

Stima delle probabilità basate su confronti dicotomici (ipotesi di indipendenza degli esiti dei confronti sulle diverse variabili)

$$m(\gamma)$$

Una delle tecniche più usate consiste nell'applicare l'algoritmo di Expectation Maximization che fornisce stime di massima verosimiglianza dei parametri di interesse.

Linkage probabilistico: metodi – *La stima dei parametri*

$$u(\gamma)$$

Poiché generalmente  $\text{card}(U) \ll \text{card}(M)$  si opera un campionamento delle coppie di record e si stimano le frequenze dei diversi esiti di confronto  $\gamma$  ignorando il contributo di  $M$

Linkage probabilistico: metodi – *La stima dei parametri*

In molte situazioni non è sufficiente confrontare due stringhe carattere per carattere a causa degli errori tipografici.

Si necessita di avere funzioni che rappresentino concordanze parziali, con concordanze totali rappresentate da 1 e gradi di concordanza parziale compresi tra 0 e 1.

Linkage probabilistico: metodi – *Funzioni di confronto tra stringhe*

Nel 1989 Jaro ha introdotto un metodo per trattare con la presenza di errori tipografici.

L'idea di Jaro consiste di due passi:

- applicare un algoritmo di confronto tra stringhe che restituisce un valore basato su caratteri comuni, trasposizioni e lunghezza delle stringhe
- usare il valore ottenuto per aggiustare il peso tra quello totale di concordanza e di discordanza

Linkage probabilistico: metodi – *Funzioni di confronto tra stringhe*

SDO		ReNCaM		COG	NOM	w_cog	w_nom	w
ROSSI	CARLO	ROSSI	CARLO	1	1	5,80	4,32	10,12
ROSSI	CARLO	VERDE	MARIO	0	0	-4,60	-3,90	-8,50
ROSSI	CARLO	NERI	MARIO	0	0	-4,60	-3,90	-8,50
VERDI	MARIO	ROSSI	CARLO	0	0	-4,60	-3,90	-8,50
VERDI	MARIO	VERDE	MARIO	0	1	-4,60	4,32	-0,28
VERDI	MARIO	NERI	MARIO	0	1	-4,60	4,32	-0,28
GIALLI	ANNA	ROSSI	CARLO	0	0	-4,60	-3,90	-8,50
GIALLI	ANNA	VERDE	MARIO	0	0	-4,60	-3,90	-8,50
GIALLI	ANNA	NERI	MARIO	0	0	-4,60	-3,90	-8,50

Linkage probabilistico: metodi – Funzioni di confronto tra stringhe

SDO		ReNCaM		COG	NOM	w_cog	w_nom	w	score
ROSSI	CARLO	ROSSI	CARLO	1	1	5,80	4,32	10,12	10,12
ROSSI	CARLO	VERDE	MARIO	0	0	-4,60	-3,90	-8,50	-8,50
ROSSI	CARLO	NERI	MARIO	0	0	-4,60	-3,90	-8,50	-8,50
VERDI	MARIO	ROSSI	CARLO	0	0	-4,60	-3,90	-8,50	-8,50
VERDI	MARIO	VERDE	MARIO	0	1	-4,60	4,32	-0,28	6,12
VERDI	MARIO	NERI	MARIO	0	1	-4,60	4,32	-0,28	-0,28
GIALLI	ANNA	ROSSI	CARLO	0	0	-4,60	-3,90	-8,50	-8,50
GIALLI	ANNA	VERDE	MARIO	0	0	-4,60	-3,90	-8,50	-8,50
GIALLI	ANNA	NERI	MARIO	0	0	-4,60	-3,90	-8,50	-8,50

Linkage probabilistico: metodi – Funzioni di confronto tra stringhe

Esistono molti tipi di algoritmi per il confronto di stringhe tra gli altri possiamo ricordare:

- Jaro ha introdotto un algoritmo che tiene conto caratteri comuni e trasposizioni per passare da una stringa all'altra;
- Winkler ha modificato l'algoritmo introdotto da Jaro;
- N-gram distance considera l'insieme di tutte le sottostringhe di lunghezza n;
- Edit distance (Levenshtein distance) è il numero minimo di inserimenti, cancellazioni e sostituzioni di singoli caratteri richieste per trasformare una stringa nell'altra;

Linkage probabilistico: metodi – Funzioni di confronto tra stringhe

Algoritmo di Jaro per il confronto di stringhe:

I passi basilari dell'algoritmo includono il calcolo della lunghezza delle due stringhe e il numero di caratteri comuni e delle trasposizioni.

I caratteri comuni sono quelli che distano per meno della metà della stringa più corta, mentre le trasposizioni sono i casi in cui il carattere in una stringa è fuori posto rispetto al corrispondente carattere comune nell'altra stringa.

$$C(s_1, s_2) = 1/3 * \left( \frac{N_{common}}{L_{s1}} + \frac{N_{common}}{L_{s2}} + 0.5 * \frac{N_{transpositions}}{N_{common}} \right)$$

Linkage probabilistico: metodi – Funzioni di confronto tra stringhe

Il numero di tutte le possibili coppie è generalmente troppo elevata perché si possano confrontare tutte:

File 1	File 2	Numero coppie	Numero max possibili match
1000	1000	$10^6$	1000
10000	10000	$10^8$	10000
100000	100000	$10^{10}$	100000

**Per ridurre il numero di confronti** da fare si utilizza una tecnica detta **“blocking”** dei file, che consiste nel suddividere i file di partenza in blocchi mutuamente esclusivi ed esaustivi e di **confrontare solo i record contenuti nei blocchi equivalenti** dei due file.

Linkage probabilistico: metodi – *Blocking*

Se nel nostro esempio la variabile di blocking fosse l’iniziale del cognome non opereremmo molti confronti

SDO		ReNCaM	
ROSSI	CARLO	ROSSI	CARLO
<del>ROSSI</del>	<del>CARLO</del>	<del>VERDE</del>	<del>MARIO</del>
<del>ROSSI</del>	<del>CARLO</del>	<del>VERDE</del>	<del>MARIO</del>
<del>VERDI</del>	<del>MARIO</del>	<del>ROSSI</del>	<del>CARLO</del>
VERDI	MARIO	VERDE	MARIO
VERDI	MARIO	NERI	MARIO
<del>GIALLI</del>	<del>ANNA</del>	<del>ROSSI</del>	<del>CARLO</del>
<del>GIALLI</del>	<del>ANNA</del>	<del>VERDE</del>	<del>MARIO</del>
GIALLI	ANNA	NERI	MARIO

Linkage probabilistico: metodi – *Blocking*

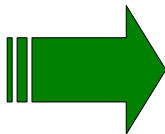
Scopo del blocking è diminuire il numero di coppie da confrontare **senza scartare** coppie che costituiscono un match. I record che appartengono a blocchi diversi nei due file **non vengono confrontati** e sono considerati a priori come non match.

Per esempio i blocchi potrebbero essere fatti per equivalenza di sesso e codice fonetico del cognome, oppure di data di nascita e codice fonetico del cognome.

Funzione  
soundex

Linkage probabilistico: metodi – *Blocking*

Il blocking dei file fa sì che siano confrontati solo i record che hanno lo stesso valore delle variabili di blocking.



Una conseguenza di questa strategia è che i record che non hanno lo stesso valore della variabile di matching sono automaticamente classificati come non match, con conseguente possibile introduzione di errori.

Linkage probabilistico: metodi – *Blocking*

Nasce il problema di **valutare l'impatto** del blocking nella generazione di errori di linkage:

- quali variabili utilizzare
- strategie di blocking: or blocking, and blocking, multipass blocking

Linkage probabilistico: metodi – *Blocking*

Il numero di record in ogni blocco deve essere:

- abbastanza piccolo da evitare molti confronti non produttivi e tuttavia
- abbastanza grande da prevenire che record riferiti alla stessa persona siano in blocchi diversi e quindi non siano confrontati.

Linkage probabilistico: metodi – *Blocking*

Le migliori variabili di blocking sono quelle con

- il più alto numero di valori,
- la più alta affidabilità e stabilità,
- le minori frequenze di errori.

Una buona variabile di blocking dovrebbe presentare un elevato numero di valori che siano abbastanza uniformemente distribuiti e abbiamo una bassa probabilità di presentare errori

Le variabili con i più alti pesi sono le migliori variabili di blocking.

Linkage probabilistico: metodi – *Blocking*

Gli errori nelle variabili di blocking potrebbero incrementare la frequenza di falsi non-match.

Idealmente le variabili di blocking non dovrebbero avere errori e dovrebbero essere immutabili.

Dal momento che questo non è possibile in pratica, i falsi non match possono essere ridotti considerando passi multipli di linkage.

Ad ogni passo vengono confrontati record che non sono stati appaiati al precedente. Ad ogni passo si utilizzano differenti variabili di blocking.

Linkage probabilistico: metodi – *Blocking*

Per variabili alfanumeriche sono spesso utilizzati come variabili di blocking codici fonetici delle variabili stesse,

Ad esempio si possono considerare i codici fonetici di nome e cognome al fine di controllare l'effetto di errori di spelling o di comprensione nella memorizzazione dei record.

Linkage probabilistico: metodi – *Blocking*

Soundex è un **algoritmo fonetico** per indicizzare nomi in base al loro suono

Il suo scopo principale è quello di consentire la **transcodifica in una medesima stringa** di nomi con pronuncia uguale ma diversa rappresentazione grafica

da **Paolo Borsa** (Politecnico di Milano)

Linkage probabilistico: metodi – *Blocking*

Un algoritmo di Soundex prende una parola – generalmente un nome di persona – come input e produce una stringa di caratteri alfanumerici che identifica un gruppo di parole la cui realizzazione fonetica è (più o meno) simile.

La Soundex è stata sviluppata per la **lingua inglese**, nella quale il rapporto di corrispondenza tra suoni e segni scritti è piuttosto libero, e per il contesto statunitense, in cui i nomi propri hanno origine etnica diversa.

Per questo un'eventuale applicazione richiede estrema cautela

da **Paolo Borsa** (Politecnico di Milano)

Linkage probabilistico: metodi – *Blocking*

Il codice Soundex è **una lettera seguita da tre numeri**:

la lettera è la prima lettera del nome,

i numeri transcodificano le rimanenti consonanti raggruppate secondo la seguente tabella

1	←	B, P, F, V
2	←	C, K, G, J, Q, S, Z, X
3	←	D, T
4	←	L
5	←	M, N
6	←	R

da **Paolo Borsa** (Politecnico di Milano)

Linkage probabilistico: metodi – *Blocking*

Tra i molti codici fonetici a disposizione, l'algoritmo Soundex risulta quello maggiormente adattabile a contesti non-anglofoni (in cui si utilizza l'alfabeto latino), proprio in virtù della sua **scarsa sofisticazione**.

Nella Soundex la transcodifica delle lettere in cifre avviene, infatti, in base a una suddivisione piuttosto grezza dei grafemi e dei fonemi corrispondenti:

da **Paolo Borsa** (Politecnico di Milano)

Linkage probabilistico: metodi – *Blocking*

La scelta dei valori di soglia è un passo del record linkage che diversi autori hanno affrontato in modo differente



Alcuni autori hanno affrontato la scelta basandosi sull'analisi dell'istogramma dei pesi generati (Gomatam et al. 2002)



Nel loro lavoro Fellegi e Sunter presentano un metodo di determinazione dei valori di soglia basato sulla creazione di un campione di configurazioni e sulla stima delle probabilità  $\lambda$  e  $\mu$  associate a diversi valori soglia

Linkage probabilistico: metodi – *valori soglia*

Nel lavoro di Jaro 1989 si propone questo metodo per confronti di tipo dicotomico:

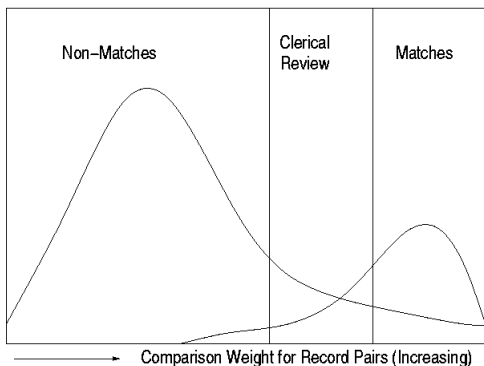
Ci sono  $2^n$  possibili configurazioni di concordanza/discordanza delle  $n$  variabili. Queste configurazioni possono essere ordinate per il valore del peso complessivo. Dopo aver ordinato le configurazioni è possibile calcolare  $P(.|M)$  e  $P(.|U)$

$$\Pr(\gamma^j | \mathbf{U}) = \prod_{i=1}^n u_i^{\gamma_i^j} (1 - u_i)^{1 - \gamma_i^j}$$

$$\Pr(\gamma^j | \mathbf{M}) = \prod_{i=1}^n m_i^{\gamma_i^j} (1 - m_i)^{1 - \gamma_i^j}$$

Linkage probabilistico: metodi – *valori soglia*

Il peso massimo per una decisione di non match (soglia inferiore) è il peso della configurazione dove la somma di  $P(.|M)$  non eccede la probabilità desiderata che un match sia classificato come non match



Il peso minimo per una decisione di match (soglia superiore) è il peso della configurazione per cui la somma di  $P(.|U)$  non eccede la probabilità desiderata che un non match possa essere classificato come un match.

Linkage probabilistico: metodi – *valori soglia*

La stima del numero di falsi match e dei falsi non match è affrontato in diversi modi in letteratura

Quando è possibile effettuare delle revisioni manuali si possono campionare dalle coppie di record per determinare il vero stato di match.

Una volta che le vere coppie sono state identificate si associa loro il peso e si costruiscono le distribuzioni cumulate per stimare gli errori associati ai valori soglia.

Linkage probabilistico: metodi – *stima errori*

Alternativamente si possono utilizzare modelli che consentano la stima automatica delle frequenze di errore.

Belin e Rubin (1991) hanno sviluppato un metodo per stimare le frequenze di falsi match associati a diversi valori soglia. La distribuzione dei pesi osservati è vista come una mistura di pesi dei veri match e dei falsi match.

Attraverso l'algoritmo EM sono stati stimati i parametri della mistura di distribuzioni.

Per modellizzare le curve di match e non match, gli autori richiedono di conoscere la verità sullo stato di matching di un insieme di coppie.

Linkage probabilistico: metodi – *stima errori*

## Vantaggi

- non è necessario considerare direttamente il metodo di creazione dei pesi, in questo modo è applicabile a molti modi di calcolare i pesi.
- una volta che il modello è stato messo a punto le frequenze di errore possono essere esaminate per infiniti valori soglia e bande di confidenza possono essere costruite per monitorare la precisione delle stime.

Linkage probabilistico: metodi – stima errori

## Limiti

- è necessario disporre di un training set (che potrebbe avere caratteristiche diverse dai file su cui si lavora)
- stima solo i falsi mach, non le frequenze complessive di falsi positivi e falsi negativi
- un assunto chiave del metodo di Belin-Rubin è che sia possibile trasformare le distribuzioni dei pesi negli insiemi di match e non match in normali, ma questo non è un assunto sempre plausibile.
- il metodo funziona meglio quando le curve dei peso verso le frequenze per match e non match sono piuttosto separate e non ci si discosta troppo dall'assunto di indipendenza

Linkage probabilistico: metodi – stima errori

Weengly *et al* (2005) propongono un metodo basato sulla simulazione di match e non match con distribuzioni multinomiali con parametri basati sulle stime delle probabilità  $m$  e  $u$ .

Metodi Monte Carlo vengono utilizzati per la simulazione di match e non match per i quali si calcolano poi i pesi di confronto.

Le distribuzioni cumulate dei pesi dei match e dei non match sono utilizzate per stimare gli errori associati ai valori di soglia

Linkage probabilistico: metodi – *stima errori*

**Sensibilità:** il numero di coppie di record appaiate correttamente diviso per il numero totale di coppie di record che costituiscono veri match

**Specificità:** il numero di coppie di record correttamente non appaiate diviso per il numero di coppie di record che costituiscono veri non match

**Frequenza di match:** il numero totale di coppie di record appaiate diviso per il numero totale di coppie di record che costituiscono veri match

**Valore predittivo positivo:** il numero totale di coppie di record correttamente appaiate diviso per il numero totale di coppie di record appaiate

Linkage probabilistico: metodi – *misura di performance*

In alcuni casi il linkage deve avvenire in modo che a un record nel primo file corrisponda al più un record nel secondo file e viceversa (one-to-one linkage)

Ci sono algoritmi grezzi che associano un record con il corrispondente record disponibile che ha il maggior peso, mentre i record successivi sono confrontati solo con quelli che non sono stati assegnati.

Linkage probabilistico: metodi – *forzare 1-1 linkage*

Nel suo lavoro Jaro (1989) propone un metodo più articolato:

dopo aver calcolato la matrice dei pesi complessivi per tutte le coppie di un blocco è possibile identificare uno schema di assegnazione 1-1 in modo che sia massimizzata la somma dei pesi composti assegnati alle coppie di record.

L'utilizzo di un tale modello di programmazione lineare rappresenta un avanzamento rispetto ai precedenti metodi di assegnazione.

Linkage probabilistico: metodi – *forzare 1-1 linkage*

## **Governativi e universitari**

GDRIVER: US Bureau of the Census Software

Febri (Freely Extensible Biomedical Record Linkage)

GRLS: Statistics Canada (UNIX)

EPILINK

## **Freeware**

Link Plus

Link King

## **Commerciali**

Charles Day

A checklist for evaluating record linkage software

Record Linkage Techniques - 1997

March 20-21, 1997 - Arlington, VA

Linkage probabilistico: metodi – sistemi di linkage

## **Bibliografia essenziale**

- Newcombe, H.B., *“Handbook of record linkage: Methods for health and statistical studies, administration, and business”*, Oxford University Press, New York, 1988;
- Fellegi, I.P., e Sunter, A.B. *“A theory for record linkage”*, Journal of the American Statistical Association, 1969, 64, 1183-1210;
- Jaro, M.A., *“Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida”*, Journal of the American Statistical Association, 1989, 84, 414-420;
- Belin, T.R. e Rubin, D.B., *“A method for calibrating falsematch rates in record linkage”*, Journal of the American Statistical Association, 1995, 90, 694-707;
- Winkler, W.E., *“Matching and record linkage”*, In Business Survey Methods, (Eds. B.G. Cox *et al.*), John Wiley & Sons, 1995, 355-384;
- Gomatam, S. *et al.*, *“An empirical companion of record linkage procedures”*, Statistics in Medicine, 2002, 21, 1485-1496;

Linkage probabilistico: metodi – bibliografia

## **Bibliografia essenziale**

- Gill, L., *“Methods for automatic record matching and linkage and their use in national statistics”*, National Statistics Methodological Series No 25, 2001;
- Gu L. *et al.*, *“Record Linkage: Current Practice and Future Directions”*, CMIS Technical Report No. 03/83, CSIRO Mathematical and Information Sciences, 2003, Canberra, Australia  
<http://datamining.csiro.au>