TUTORIAL ON THE FELLEGI-SUNTER MODEL FOR RECORD LINKAGE

Ivan P. Fellegi, Statistics Canada

EDITORS' NOTE

The following exhibits, numbered 1 to 22, were used at the Workshop on Exact Matching Methodologies (in the form of transparencies) as the basis for a presentation of the essential features and some of the consequences of the Fellegi-Sunter model and theory for record linkage. Many Workshop participants commented favorably on the exhibits and requested copies. The exhibits are presented here, without additional commentary, for the benefit of those who would like to have a convenient summary of the main points. The following chart shows the relationship between groups of exhibits and specific sections of the article, "A Theory for Record Linkage," which can be found on pages 51-78 of this volume.

Figure 1.--Exhibits for Fellegi-Sunter Article

Exhibit Numbers	Topic	Section of Article	Pages
1 to 6, 7a	Basic model and theory	2	52-57
7b, 8 to 10	Method of constructing an optimum linkage rule; consequences	2.1	54-57
11 to 14	Assumptions used in estimating weights	3.2	57-59
15 to 17	Calculation of weights, Method I	3.3.1	60-62
18	Calculation of weights, Method II	3.3.2	62-63
19, 20	Blocking	3.4	64-65
21	Choice of comparison space	3.6	66-67
22	Calculation of threshold values	3.7	67-68

Exhibit 1

Two sets of units: $A = \{a\}$, $B = \{b\}$

Vector of characteristics $\alpha(a)$, $\beta(b)$ associated with units.

$L_A = \{\alpha(a);\ a \in A\}$, $L_B = \{\beta(b);\ b \in A\}$ (lists)

$L_A \times L_B = M + U$

where $M = \{[\alpha(a), \beta(b)];\ a = b,\ a \in A,\ b \in B\}$

$U = \{[\alpha(a), \beta(b)];\ a \neq b,\ a \in A,\ b \in B\}$

$L_A \times L_B$ unmanageable.

Exhibit 2

Code results of comparing $\alpha(a)$, $\beta(b)$: $\gamma(a, b)$

$\gamma[\alpha(a), \beta(b)] = \gamma(a, b) = (\gamma^1, \gamma^2, \ldots, \gamma^k)(a, b)$

Examples: $\gamma_i = 0$ if sex is same

1 if sex is different

2 if sex is missing on either record

Exhibit 3

γ_j = 0 if name is same and is Brown

1 if name is same and is Smith

2 if name is same and is Jones

3 if name is same and not Brown, Smith, Jones

4 if name is different

5 if name is missing on either record

$\Gamma = \{\gamma(a, b)\}$: comparison space.

Exhibit 4

Linkage rule: decision regarding match status of (a, b) based on $\gamma(a, b)$

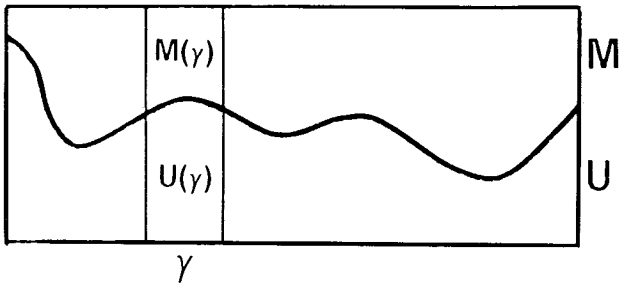
$d(\gamma) = A_1$: link (inference is "match")

$d(\gamma) = A_2$: possible link ("don't know")

$d(\gamma) = A_3$: non-link (inference is "unmatched")

Exhibit 5

$\gamma(a, b) = \gamma_0$ is a subset of $L_A \times L_B$

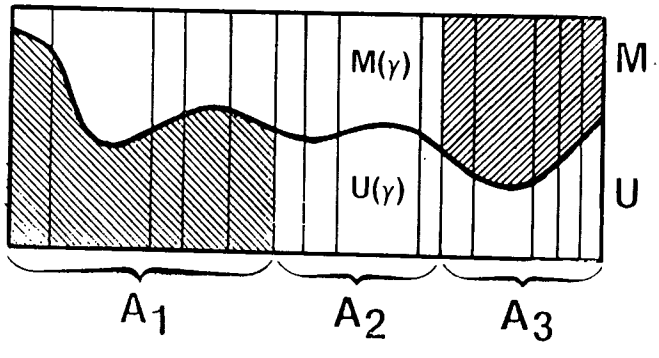


$$m(\gamma) = P\{\gamma(a, b) \mid (a, b) \in M\} = \frac{||\, M(\gamma)\,||}{||\, M\,||}$$

$$u(\gamma) = P\{\gamma(a, b) \mid (a, b) \in U\} = \frac{||\, U(\gamma)\,||}{||\, U\,||}$$

Exhibit 6

A linkage rule partitions $L_A \times L_B$:



For any $\gamma \in A_1$ all record pairs in $U(\gamma)$ are linked in error.

$$\mu = P(A_1 \mid U) = \sum_{\gamma \in A_1} u(\gamma)$$ proportion of linked record pairs in U

$$\lambda = P(A_3 \mid M) = \sum_{\gamma \in A_3} m(\gamma)$$ proportion of unlinked record pairs in M

Exhibit 7

a) Definition: Consider all linkage rules R on Γ with error levels μ_0, λ_0. Then R^1 is optimal if $P(A_2 \mid R^1) \leqq P(A_2 \mid R)$ for all R.

b) Heuristic: arrange $L_A \times L_B$ so that $m(\gamma)$ monotone decreases and $u(\gamma)$ increases. Choose A_1, A_3 to correspond to desired μ, λ. Then this linkage rule is optimal.

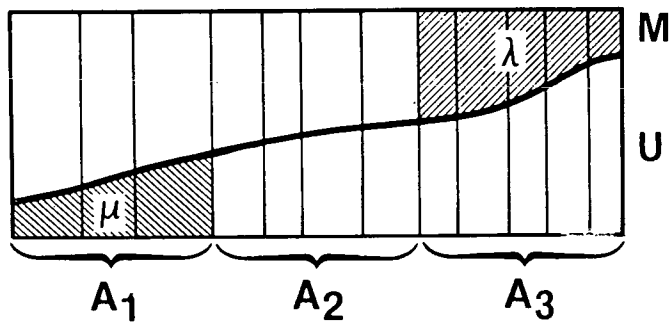


Exhibit 8

Optimal rule: order γ by decreasing values of $m(\gamma)/u(\gamma)$.

A_1 if $T_\mu \leqq m(\gamma)/u(\gamma)$

A_2 if $T_\lambda < m(\gamma)/u(\gamma) < T_\mu$

A_3 if $m(\gamma)/u(\gamma) \leqq T_\lambda$

T_μ chosen so that $\mu = \mu_0$, T_λ so that $\lambda = \lambda_0$

Likelihood ratio tests: A_1 at level μ, A_3 at level λ.

Uniformly most powerful.

Tepping's test (JASA, 1968) functionally equivalent.

Exhibit 9

HIGH → $m(\gamma)/u(\gamma)$ → LOW

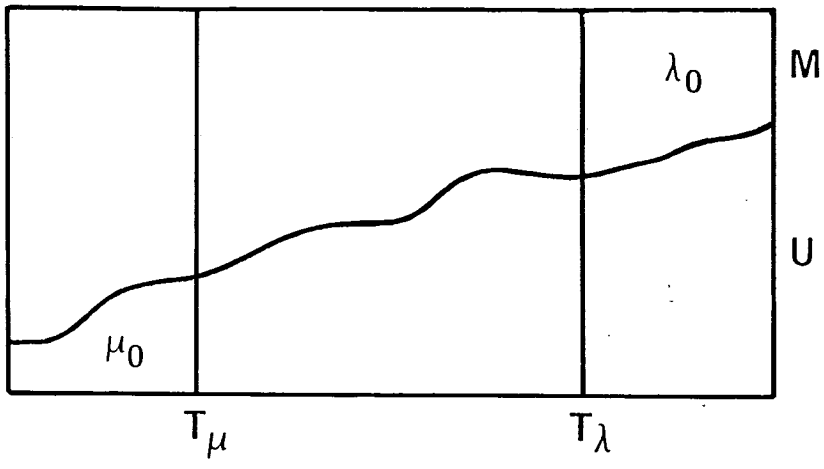


Exhibit 10

1. Trade-off between decreasing μ_0, λ_0 or A_2
2. A_2 can be eliminated if $T_\mu = T_\lambda$
3. Typically $\mu_0 < < \lambda_0$ should hold. If N is the number of matched record pairs, $(N_A N_B - N)$ the number of unmatched record pairs, then condition for number of linked record pairs to be N is

$$N(1 - \lambda_0) + (N_A N_B - N)\mu_0 = N.$$

$$\text{True if } \mu_0 = \frac{N}{N_A N_B - N}\lambda_0$$

4. Randomized decision may be needed to achieve $\mu = \mu_0$, $\lambda = \lambda_0$ exactly.

Exhibit 11

Estimating m/u

If $\gamma = (\gamma^1, \gamma^2, \ldots, \gamma^K)$

γ^k has n_k values

then γ has $n_1 . n_2 \ldots n_K$ values.

Simplifying assumption:

$$m(\gamma) = m(\gamma^1) . m(\gamma^2) \ldots m(\gamma^K)$$

$$u(\gamma) = u(\gamma^1) . u(\gamma^2) \ldots u(\gamma^K)$$

Components of γ are conditionally independent w.r. to m and u.

Exhibit 12

Matched records: Without errors, all γ^k should show "agreement". Hence independence $\longrightarrow$ errors in different ident. variables of a and b are independent.

Unmatched records: accidental agreement on one variable (e.g. name) is independent of accidental agreement on another (e.g. address).

Estimands: $m(\gamma^1), m(\gamma^2), \ldots m(\gamma^K)$ -- $n_1 + n_2 + \ldots + n_K$

(also for u).

Exhibit 13

Need care in defining γ :

$$\gamma^1 - \begin{cases} \text{agreement on female given name} \\ \text{agreement on male given name} \\ \text{disagreement on given name} \\ \text{given name missing on either record} \end{cases}$$

$$\gamma^2 - \begin{cases} \text{agreement on sex} \\ \text{disagreement on sex} \\ \text{sex missing on either record} \end{cases}$$

Accidental agreement on $\gamma^1 \rightarrow$ agreement on γ^2.
Independence might hold if first two codes of γ^1 combined.

Exhibit 14

Prefer to use log (m/u) - monotone incr. function of (m/u).

$\log (m/u) = w^1 + w^2 + \ldots + w^k$ where

$$w^k = \log [m(\gamma^k)/u(\gamma^k)]$$

We have

$$w^k \gtrless 0 \quad \text{if} \quad m(\gamma^k) \gtrless u(\gamma^k)$$

(intuitively appealing).

Similar to Newcombe-Kennedy (Communications of ACM, 1962).

Exhibit 15

METHOD 1 FOR WEIGHT CALCULATION (ILLUSTRATION)

Weights for "name" component.

Let proportions of different names in A, B and $A \cap B$ be

$p_A(i)$, $p_B(i)$, $p(i)$ $(\Sigma\, p=1)$. For simplicity:

$$p_A(i) = p_A(i) = p(i)$$

e_A, e_B: prob. of misreporting name in A, B respectively

p observable, e separately to be estimated.

Exhibit 16

$$w \text{ (agreement on jth name)} \approx \log (1/p_j)$$

- Positive
- The smaller p(j), the larger w
- I.e. large positive weight for agreement on rare characteristic

$$w(\text{agreement}) \approx \log (1/p) \quad \text{where} \quad p = \sum_j p_j^2$$

- Large for uniformly well discriminating variable
- p decreases fast if common outcomes are separated.

Exhibit 17

$$w \text{ (disagreement)} = \log \frac{e_A + e_B}{1-p}$$

- Typically negative
- The smaller the error, the larger the negative weight
- I.e. disagreement on well reported variable $\longrightarrow$ large negative weight
- E.g.: sex. Don't restrict linkage variables to high discrimination.

w (name missing on either file) = 0

- neutral contribution.

Exhibit 18. SECOND METHOD (ILLUSTRATION)

Assume only three components; each coded to two states: "agreement", "disagreement".

Conditional probabilities of "agreement" are m_h, u_h.

$$N_A N_B U_h = N m_h + (N_A N_B - N) u_h \qquad h = 1,2,3$$

where U_h: proportion of record pairs with "agreement" in h-th component.

U_h, N_A, N_B observable; N, m_h, u_h unknown.

Above 3 equations can be supplemented by other 4; all involve observable quantities + 7 unknown variables.

Solvable; generalizable; heavy dependence on independence.

Exhibit 19

Blocking

Objective: reduce number of comparisons.

Implicit assumption: comparisons not made are non-linked (A_3).

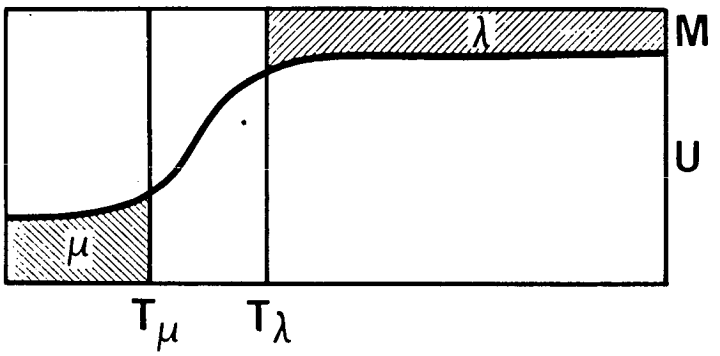


Exhibit 20. IDEAL BLOCKING VARIABLE

1. If a variable is such that disagreement results in very large negative weight -- corresponding e_A, e_B very small. Does not increase λ.

2. High discrimation results in maximum file blocking (comparisons restricted to records which agree on the blocking variable).

Frequent compromise: coded name where code is designed to reduce impact of misspellings.

Additional use of any well reported variable, even of low discrimination (e.g. sex), is net bonus.

Exhibit 21. CHOICE OF COMPARISON SPACE

1. How many separate values to recognize for agreement?

   Trade-off between complexity and reduction in Σp_j^2

2. How many of the variables common to both files should we use?

   Generally: the more the better.

3. w is positive for agreement, negative for disagreement almost certainly.

4. If $e_A + e_B < \frac{1}{2} < 1-p$, then each additional variable increases total weight for matched records, decreases total weight for unmatched records -- both with probability $> \frac{1}{2}$.

Exhibit 22. ESTIMATING THRESHOLDS

1. Select at random one value of <u>each</u> γ^k. Higher probabilities for high $|w|$;

2. Combine into γ; compute corresponding weight (w);

3. Repeat n times;

4. Arrange γ by decreasing w;

5. Set T_μ , T_λ as in Γ, but counting each γ with inverse of probability of selection.