Robustness of the Census Bureau's Record Linkage System

R. Patrick Kelley, Bureau of the Census

INTRODUCTION

Record Linkage is the name given to any process which identifies the common reporting units in two data files.

At the Census Bureau record linkage has long been used for coverage and content evaluation. Early Census record linkage projects were conducted clerically. However, as the census has become larger and the demands on it more complex, such clerical operations have become too slow, expensive and error prone. So, the Bureau has developed an automated record linkage system.

The robustness of this automated system to errors in the data or deviations from the basic assumptions that were used in its development is of great interest to its users. In this paper we present a study of the system's robustness with respect to violations of the independence assumption. There are two questions of interest:

1. How do we test for the presence of dependence?
2. How robust is the record linkage operation to slight perturbations from independence?

To better define these questions we must consider some background information.

The methodology which forms the foundation for the Census Bureau's automated record linkage system is an extension of the record linkage theory presented in Fellegi and Sunter (1969).

Given a randomly selected record pair (α,β), the basic aim of this theory is to classify (α,β) into one of the following categories:

A_1: The pair is a match,
A_2: no determination - clerical review,
A_3: The pair is not a match

The statistical model that is used to compare various ways to make this classification is the same one used in discrete discriminant analysis (see, for example, Goldstein and Dillon (1978)).

The first component of this model is a vector valued function, referred to as the <u>comparison vector</u> and denoted by Γ, which measures <u>the similarity</u> between α and β. Given that we have selected a particular Γ, the second component of the model concerns the behavioral difference of Γ between matched and unmatched pairs. The model accounts for this difference by assuming that Γ is a random vector generated by $P(\cdot|M)$ on matched pairs and $P(\cdot|U)$ on unmatched pairs.

Using this discriminant analysis model, the best decision function is defined to be the one which

minimizes $P(A_2)$
subject to $P(A_1|U) \leq \mu$ and $P(A_3|M) \leq \lambda$,

where μ and λ are prespecified error rates.

It is proved that the best decision function is given by

$$(1)\quad D(\gamma) = \begin{cases} A_3 \text{ if } L(\gamma) \leq t_1 \\ A_2 \text{ if } t_1 < L(\gamma) < t_2 \\ A_1 \text{ if } t_2 \leq L(\gamma)\,, \end{cases}$$

where γ is a realization of Γ, $L(\gamma)=\ln 2[P(\gamma|M)/P(\gamma|U)]$, t_1 is the largest value in the range of $L(\cdot)$ for which $P(A_3|M) \leq \lambda$ and t_2 is the smallest value in the range of $L(\cdot)$ for which $P(A_1|U) \leq \mu$.

There are several tasks that need to be performed before this theory can be transformed into an operational system. First we need to determine the form Γ will take. Next we need to specify a model for $P(\cdot|M)$ and $P(\cdot|U)$. Third we need to estimate $P(\cdot|M)$ and $P(\cdot|U)$.

In performing these tasks we must balance two competing forces. The first is our desire to get as much information from the data as possible. The second is our desire to keep the cost of the linkage operation as low as possible.

In developing the Census Bureau's automated record linkage system Γ was selected in favor of keeping down cost. As a result the Γ vector selected consists of the agreement-disagreement pattern between the record identifier fields. Thus, for example, the first component of the comparison vector might be defined as

$$\Gamma_1(\alpha,\beta) = \begin{cases} 1 & \text{Surname on } \alpha = \text{surname on } \beta \\ 0 & \text{Otherwise.} \end{cases}$$

Further, it was assumed that the components of Γ are independent for both $P(\cdot|M)$ and $P(\cdot|U)$.

Throughout the rest of this paper we will assume that it is appropriate to select a comparison vector whose components are dichotomous.

A MODEL FOR DEPENDENT DATA

Of the many potential models which could be used to account for dependence between the components of the comparison vector we will study the following model which was presented in Bahadur (1961):

Let X be a random vector of n dichotomous components, $\theta_i=P(X_i=1)$ and

$$Z_i(X) = \frac{X_i-\theta_i}{\sqrt{\theta_i(1-\theta_i)}}\,.$$

Then it is proved that

$$P(X=x)=\prod_{i=1} \theta_i^{x_i}(1-\theta_i)^{1-x_i}\Big(1+\sum_{j<k}\rho_{jk}Z_j(x)Z_k(x)+$$
$$\sum_{j<k<l}\rho_{jkl}Z_j(x)Z_k(x)Z_l(x)$$
$$+ \ldots + \rho_{1\ldots n}Z_1(x)\ldots Z_n(x)\Big)$$

for all x, where $\rho_{jk}=E(Z_jZ_k)$, $\rho_{jkl}=E(Z_jZ_kZ_l),\ldots,$ and $\rho_{1\ldots n}=E(Z_1\ldots Z_n)$.

We will restrict our attention to those models where interactions for more than two factors are assumed to be zero.

So our model for $P(\Gamma=\gamma|M)$ is

$$P(\Gamma=\gamma|M)=\prod_{i=1}^{n} m_i^{\gamma_i}(1-m_i)^{1-\gamma_i}\Big(1+\sum_{j<k}\rho_{jk}Z_j(\gamma)Z_k(\gamma)\Big)$$

while

$$P(\Gamma=\gamma|U)=\prod_{i=1} u_i^{\gamma_i}(1-u_i)^{1-\gamma_i}\ (1+\sum_{j<k}\Psi_{jk}Z_j(\gamma)Z_k(\gamma))$$

In discussing this model the first thing to note is that not all combinations of ρ_{ij}'s and Θ_i's yield a true probability distribution. Bahadur demonstrates that $\sum_x P\ (X=x) = 1$, but shows that there can be cases where $P(X=x)<0$ for some x. As an example let $\Theta_1=.9$, $\Theta_2=.85$, $\Theta_3=.95$, $\rho_{12} = .1$, $\rho_{13} = -.1$ and $\rho_{23} = .05$ then $P(X=(010))= -.002$.

We will refer to a parameter vector which yields a true distribution as feasible. It is clear that $(\Theta_1\ldots\Theta_n;\rho_{12}\ldots\rho_{(n-1)n})$ is feasible if

$$(1+\sum_{i<j}\rho_{ij}Z_i(x)Z_j(x)) > 0 \quad \text{for all } x.$$

Checking this condition requires considerable computation. Therefore, a more tractable formulation is desirable. The following sufficient condition is provided by Bahadur:

Let λ^* = minimum eigenvalue of the matrix of correlations
and $\beta_i = \max\ \{\Theta_i/(1-\Theta_i),\ (1-\Theta_i)/\Theta_i\}$

then the parameter values are feasible if

$$\lambda^* \geq 1- 2/(\sum_{i=1}^n \beta_i).$$

Now this condition offers considerable reduction in computation. Unfortunately, it appears to be quite restrictive. For example, if $\Theta_1=.9$ and $\Theta_2=.85$ then $\beta_1=9$ and $\beta_2=5.66$, so $\lambda^* \geq .86$. But $\lambda^* = \min\ \{1+\rho_{12},\ 1-\rho_{12}\}$; thus $0 < \rho_{12} \leq .14$ is the interval of feasible parameter values given by the Bahadur condition. By computing the probability mass function for different ρ_{12} values we see that for this example we actually have a true distribution for all ρ_{12} such that $0 \leq \rho_{12} \leq .7925$. Thus, either a sharper bound needs to be worked out or we will have to check $P(X=x)$ for all x.

TESTS FOR INDEPENDENCE

Now let's look at ways to test for independence using the Bahadur second order model. There are two separate ways to approach the testing of independence. The first is to assume that we have independent samples from both the matched and unmatched population. We then separately test for independence using each sample. The second is to assume that we have a random sample from the mixture.

Let's consider the first method.
For a random vector with a second order Bahadur distribution, independence is equivalent to the null hypothesis

H_0: $\rho_{ij}=0$ for all $i<j$.

We can test this null against its negation using a likelihood ratio test. The likelihood under H_0 takes its maximum at $\hat\theta_i = \sum_{k=1}^N x_i^k/N$ where x_i^k is the ith component of the kth sample vector. To compute the maximum over the entire parameter space appears to be a relatively intractable problem. Goldstein and Dillon (1978) offer

$$\widehat{P(X=x)}= \prod_{i=1}^n \hat\theta_i^{x_i}\ (1-\hat\theta_i)^{1-x_i}\ (1+\sum_{j<k}\hat\rho_{jk}Z_j(x)Z_k(x))$$

where $\hat\rho_{jk}=(\sum_{\lambda=1}^N \gamma_j^\lambda\ \gamma_k^\lambda/N - \theta_j\ \theta_k)/\ \theta_j(1-\theta_j)\ \theta_k(1-\theta_k)$ as an estimate of $P(X=x)$. If we use $\widehat{P(X=x)}$ in the denominator in place of max. likelihood then our approximate likelihood ratio is given by,

$$LR=\frac{\prod_{k=1}^N\prod_{i=1}^n \hat\theta_i^{x_i^k}\ (1-\hat\theta_i)^{1-x_i^k}}{\prod_{k=1}^N\prod_{i=1}^n \theta_i^{x_i^k}\ (1-\theta_i)^{1-x_i^k}\ (1+\sum_{1<j}\hat\rho_{ij}Z_i(x^k)Z_j(x^k))}$$

$$=\frac{1}{\prod_{k=1}^N(1+\sum_{i<j}\hat\rho_{ij}Z_i(x^k)Z_j(x^k))}\ .$$

Our test statistic T is given by

$$T = -2\ \ln\ (LR)$$

$$= 2\sum_{\text{all } x} F(x)\ \ln(1+\sum_{i<j}\hat\rho_{ij}Z_i(x)Z_j(x))$$

where F(x) is the frequency of pattern x.

Now, -2 ln (likelihood ratio) is asymptotically χ^2 with $\frac{n(n-1)}{2}$ degrees of freedom. It seems that this may also be a good approximation for the asymptotic distribution of T.

Now, it is clear by inspection that T cannot be computed from non-feasible $\hat\rho_{ij}$'s and $\hat\theta_i$'s. To get around this problem we propose replacing $\hat\rho_{ij}$ by ρ_{ij}^* where ρ_{ij}^* solves the following problem:

$$\text{Minimize} \sum_{i<j}\ (\rho_{ij}-\hat\rho_{ij})^2$$

$$\text{Subject to}\ \ 1+\sum_{i<j}\rho_{ij}Z_i(x)Z_j(x)>0$$

Since $\hat\rho_{ij}$ is consistent for ρ_{ij} we see that asymptotically $\hat\rho_{ij} = \rho_{ij}^*$.

So, replacing $\hat\rho_{ij}$ by ρ_{ij}^* should not affect the asymptotic distribution T.

Thus, in summary, to test the null hypothesis of independence we propose using the test statistic

$$T = 2\sum_{\text{all } x} F(x)\ \ln\ (1+\sum_{i<j}\rho_{ij}^*\ Z_i(x)Z_j(x)).$$

We will reject the null if $T>\chi_\alpha$ where $P(\chi^2(n(n-1)/2)\leq \chi_\alpha) = 1-\alpha$.

Now let us consider an example using data from the 1985 Tampa pretest.

Example 1:

As an example let's consider the matched record pairs from the 1985 Tampa pretest PES/Census match. For brevity we will study the following variables:

Variable name	Probability of agreement
Last name	.86
First name	.78
Relation to head of household	.83
Street name	.93
House number	.99

The ρ_{ij} matrix for these data is

$$\begin{array}{cccc} .0022 & .0268 & -.002 & -.0111 \\ & .0087 & -.0052 & -.0379 \\ & & -.0168 & -.0012 \\ & & & -.0126, \end{array}$$

which is infeasible. The $\overset{*}{\rho_{ij}}$ matrix is

$$\begin{array}{cccc} .0032 & .0267 & -.0001 & -.0046 \\ & .0087 & -.0036 & -.0324 \\ & & -.0168 & -.0012 \\ & & & -.0019. \end{array}$$

Using $\overset{*}{\rho_{ij}}$, T is computed to be 14.54 which yields a p value between .1 and .25 for a χ^2 with 10 degrees of freedom.

To test the independence hypothesis using a sample of size N from the mixture we first obtain estimates of the m_i's and u_i's using the method of moments (Fellegi-Sunter type II estimation). In other words, for all γ we solve

$$pP(\Gamma=\gamma|M) + (1-p)P(\Gamma=\gamma|U) = P^*(\Gamma=\gamma),$$

where $P^*(\Gamma=\gamma)$ is the sample proportion of the event $\{\Gamma=\gamma\}$.

We then compute a χ^2 statistic on the fitted model. The obvious problem with this test is that it doesn't specifically test the independence hypothesis.

ROBUSTNESS

Moving on to question 2, that of system robustness, we first need to examine the nature of the incoming sequence of comparison vector values we are trying to match. The decision procedure given in (1) was developed under the hypothesis that the comparison vectors between separate record pairs are independent. However, since the record pairs that are considered for possible matches are elements of the cross product of the two files we are attempting to match, the comparison vectors are in fact dependent. Further, this cross product is often reduced to nonoverlapping blocks of data and matching is carried out on each block separately. The overall effect of this "blocked data" structure on λ and μ is currently unknown. For further discussion of the blocking process see Kelley (1985). We will begin our study of the effects of violations of the independence assumption by assuming the data to be classified are independent. We will then consider blocked data.

Suppose that we form a decision procedure assuming that the ρ_{ij}'s and Ψ_{ij}'s are zero.

What is the effect of nonzero ρ_{ij}'s and Ψ_{ij}'s on γ and μ, respectively? The true value of λ is

$$\lambda_t = \sum_{\gamma\varepsilon A_3} \prod_{i=1} m_i^{\gamma i} (1-m_i)^{1-\gamma i} (1+\sum_{j<k} \rho_{jk}Z_j(\gamma)Z_k(\gamma))$$

$$= \sum_{\gamma\varepsilon A_3} \prod_{i=1}^{n} m_i^{\gamma_i} (1-m_i)^{1-\gamma_i} +$$

$$\sum_{\gamma\varepsilon A_3} \sum_{j<k} \prod_{i=1} m_i^{\gamma_j} (1-m_i)^{1-\gamma_i} \rho_{jk}Z_j(\gamma)Z_k(\gamma)$$

$$= \lambda_0 + \sum_{j<k} \rho_{jk} \sum_{\gamma\varepsilon A_3} \prod_{i=1} m_i^{\gamma i} (1-m_i)^{1-\gamma_i} Z_j(\gamma)Z_k(\gamma)$$

$$(2) \qquad = \lambda_0 + \sum_{j<k} \rho_{jk} \lambda_{jk}$$

where $\lambda_{jk} = \sum_{\gamma\varepsilon A_3} \prod_{i=1}^{n} m_i^{\gamma i} (1-m_i)^{1-\gamma i} \frac{\gamma_j-m_j}{\sqrt{m_j(1-m_j)}} \frac{\gamma_k-m_k}{\sqrt{m_k(1-m_k)}}$

Likewise,

$$\mu_t = \mu_0 + \sum_{j<k} \Psi_{jk} u_{jk}$$

where

$$u_{jk} = \sum_{\gamma\varepsilon A_1} \prod_{i=1} u_i^{\gamma i} (1-u_i)^{1-\gamma i} \frac{\gamma_j-u_j}{\sqrt{u_j(1-u_j)}} \frac{\gamma_k-u_k}{\sqrt{u_k(1-u_k)}}.$$

Let's now consider a numerical example.

Example 2:

Let $m_1=.9$ $u_1=.05$
$m_2=.85$ $u_2=.1$
$m_3=.95$ $u_3=.05$
$\lambda_0=.026$ $\mu_0=.02525$

then

$$\lambda_t=.026 + .096\,\rho_{12} + .046\,\rho_{13} + .062\,\rho_{23}$$

and

$$\mu_t=.02525 + .036\,\Psi_{12} + .098\,\Psi_{13} - .007\,\Psi_{23}.$$

For this example both λ_t and μ_t are relatively sensitive to the data's actual correlation structure.

When this method was applied to the Tampa pretest match, it was found that the effect of any one correlation coefficient was negligible. But, the combined effect of a slight increase in all the coefficients could have a considerable effect on the error under study. Thus, even though no individual correlation may be large, the overall effect might be serious.

Now to test the system's robustness in the presence of blocked data, we are required to perform a series of simulation experiments. All of these experiments generated data, in the form of comparison vector values, according to the mixture $pP(\Gamma|M)+(1-p)P(\Gamma|U)$. To generate an observation we used the following two steps:

1. Randomly select a number ,R, between 0 and 1.
2. If R < proportion of matched record pairs p, then generate a matched vector value; or else generate an unmatched vector value.

To complete step 2 we needed an algorithm to generate data from a second order Bahadur model. Our algorithm is based on the following decomposition:

$$P(\Gamma_1=\gamma_1,\ldots,\Gamma_n=\gamma_n)=P(\Gamma_1=\gamma_1)P(\Gamma_2=\gamma_2|\Gamma_1=\gamma_1)\cdots$$

$$P(\Gamma_n=\gamma_n|\Gamma_{n-1}\cdots\Gamma_1=\gamma_1).$$

The algorithm itself uses n randomly selected numbers $r_1 \ldots r_n$. For the ith component

$$g_1 = \begin{cases} 1 \text{ if } r_i \leq P(\Gamma_i=1|\Gamma_1=\overset{*}{\gamma_1}\cdots\Gamma_{i-1}=\overset{*}{\gamma_{i-1}}) \\ 0 \quad \text{Otherwise} \end{cases}$$

where $\overset{*}{\gamma_1}\ldots\overset{*}{\gamma_{i-1}}$ are the 1st through the i-1 selections.

It can be shown that for the Bahadur model

$$P(\Gamma_m=\gamma_m|\Gamma_1=\gamma_1\cdots\Gamma_{m-1}=\gamma_{m-1}) =$$

$$\Theta_1^{\gamma_m}\;(1-\Theta_1)^{1-\gamma_m}\left[1+\frac{\sum_{j=1}^{m-1}\rho_{jm}Z_jZ_m}{1+\sum_{j<k<m-1}\rho_{jk}Z_{jk}}\right].$$

From this equation we have developed an iterative algorithm to generate data from a second order Bahadur model.

The simulation studies for robustness from independence were carried out on an IBM/PC. The data were generated by the means of the generalized feedback shift register generator given in Lewis and Payne (1973). This generator was implemented on the PC in PASCAL.

The basic experiment consisted of the classification of 3000 data points. These data were generated as 100 5x6 blocks. Each block contained 5 randomly assigned matched pairs for which we generated a Γ value according to P(.|M) and 25 randomly assigned unmatched pairs for which we generated a Γ value according to P(.|U). We then classified the data using the Fellegi-Sunter decision procedure with a linear sum assignment to break ties.

The numbers of false matches and non-matches were then computed. Each trial consisted of 10 replications of this basic experiment. The following table gives the results of our experiments:

Example 3:

m_1=.9 m_2=.85 m_3=.95	TM = Total Match	
	TNM = Total Non-Match	
u_1=.05 u_2=.1 u_3=.45	TFNM = Total False Non-Match	
	TFM = Total False Match	

indpndnt	TM = 2994	TFNM = 182	error rate ≅ .061
	TNM = 27006	TFM = 240	error rate ≅ .009
matched	TM = 2994	TFNM = 204	error rate ≅ .068
ρ_{ij}=.05	TNM = 27006	TFM = 240	error rate ≅ .009
unmatched			
Ψ_{ij}=0.0			
matched	TM = 2994	TFNM = 217	error rate ≅ .07
ρ_{ij}=0.0	TNM = 27006	TFM = 317	error rate ≅ .012
unmatched			
Ψ_{ij}=.05			
matched	TM = 2994	TFNM = 236	error rate ≅ .08
ρ_{ij}=0.05	TNM = 27006	TFM = 317	error rate ≅ .012
unmatched			
Ψ_{ij}=0.05			
matched	TM = 2994	TFNM = 279	error rate ≅ .09
ρ_{ij}=0.0	TNM = 27006	TFM = 462	error rate ≅ .017
unmatched			
Ψ_{ij}=.15			
matched			
data	TM = 2994	TFNM = 232	error rate ≅ .08
ρ_{ij}=.15	TNM = 27006	TFM = 235	error rate ≅ .009
unmatched			
Ψ_{ij}=0.0			
matched			
data	TM = 2994	TFNM = 326	error rate ≅ .11
ρ_{ij}=.15	TNM = 27006	TFM = 458	error rate ≅ .02
unmatched			
Ψ_{ij}=.15			

From these results it appears that the error rates are an increasing function of the ρ_{ij}'s and Ψ_{ij}'s, and so, the Fellegi-Sunter decision procedure is fairly sensitive to violations of the independence assumption when classifying blocked data.

In reviewing the results for both the blocked and iid data it is clear that these effects of correlation are different in these two cases.

For example with iid data if

ρ_{ij} = .05 and Ψ_{ij} = 0 then
λ = .026 + .01 = .036

while $\lambda \cong .068$ for blocked data.

The cause of this difference however, appears to be, at least in part, the result of the overall effect of the blocked data. The base false non-match rate for blocked data is approximately .06. For p_{ij}=.05 and Ψ_{ij}=0 λ-.06 = .008 ≃ .01 while for p_{ij} = .15 and Ψ_{ij} = 0 λ-.06 = .02. So we see that by replacing λ_0 in equation 2 by the false non-match rate for blocked data with zero correlation we obtain an approximation for the effect of correlation on false non-match rate for blocked data. Thus, the coefficients provided by equation 2 can be used to compute the approximate effect of ρ_{ij} on the false non-match rate for blocked data. The same argument applies for the effect of Ψ_{ij} on the false match rate for blocked data.

CONCLUSION

This paper represents a preliminary study of the robustness of a Fellegi-Sunter type record linkage procedure to violations of the independence assumption. As such, the model given in equations (1) and (2) for the effect of correlation on matching error should be used only as a guideline.

REFERENCES

1. Bahadur, R.R. (1961). A Representation of the Joint Distribution of Response to n Dichotomous Items. 'Studies in Item Analysis and Prediction.'

2. Fellegi, Ivan and Sunter, Alan (1969). A Theory for Record Linkage. 'Journal of the American Statistical Association.' Vol. 64, pp. 1183-1210.

3. Goldstein, Matthew and Dillon, William (1978). 'Discrete Discriminant Analysis,' Wiley.

4. Kelley, Patrick (1985). Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy. 'Proceedings of the Workshop on Exact Matching Methodologies.' Department of the Treasury Publication #1299(2-86).

5. Lewis, T.G., and Payne, W.H. (1973), Generalized Feedback Shift Register Pseudorandom Number Algorithm, JACM 20, 456-468.