# A PROPOSED IMPROVEMENT IN COMPUTER MATCHING TECHNIQUES

Thomas R. Belin, Harvard University
Department of Statistics, 1 Oxford Street, Cambridge, MA 02138

Key words: Census undercount, false match rate, Fellegi-Sunter algorithm, mixture model

Exact matching of multivariate records between two large databases relies on a statistical model for the likelihood that two records represent the same individual. A key assumption underlying the often-used algorithm proposed by Fellegi and Sunter (1969) and other matching algorithms is that agreement on different fields of information is independent. The determination of a cutoff threshold above which records will be considered matched relies heavily on the independence assumption, which is known not to be valid in many applications. This paper proposes a new method for automating the setting of cutoffs that incorporates past data to avoid relying on the independence assumption. The new method is illustrated in the context of matching records between the census and a large-scale post-enumeration survey taken after the census, which constitutes part of the process for estimating the census undercount rate.

## I. INTRODUCTION

Exact matching methods for linking together large databases of records are used in a variety of settings. In the context of evaluating the coverage of the decennial census, automated matching techniques are used as part of an extensive matching effort to compare the census to a large-scale post-enumeration survey (PES) conducted after the census. Records that remain unmatched after the overall matching process provide information about census coverage: individuals captured in the PES but not in the census constitute evidence of undercounting, and individuals enumerated in the census but not in the PES represent overcounting. [For a review of issues involved in census undercount estimation, see Citro and Cohen (1985).] The goals in the computer matching process are to declare as many records matched as possible and to avoid false declarations of match.

There is a body of statistical theory concerning properties of computer matching procedures (Newcombe, et. al., 1959; Tepping, 1968; Fellegi and Sunter, 1969; Chernoff, 1977; Chernoff, 1980; Newcombe, 1988). Exact matching of multivariate records relies on a statistical model for the likelihood that two records represent the same individual. In the literature, attention is focussed on matching procedures that ascribe weights to agreement or disagreement on each of several variables. Weights for agreement on individual variables are then aggregated into a composite weight, which is a univariate summary of the closeness of a pair of records. These weighting procedures are derived from models which assume that agreement on different variables is independent.

A prominent example of this type of matching procedure, which has been used in many record linkage applications, is the algorithm proposed by Fellegi and Sunter (1969). In the Fellegi-Sunter algorithm, the individual weights for different variables are obtained by taking the logarithm of the ratio of the likelihood of agreement given the pair of records is a match to the likelihood of agreement given the pair of records is not a match. Probabilities of agreement are estimated based on the assumption that agreement on different variables is independent.

In order to automate the matching process, a cutoff weight has to be specified above which records will be declared matched. In the Fellegi-Sunter framework, all of the possible patterns of agreement and disagreement among the matching variables are sorted by descending composite weight, and a cutoff is set when the cumulative probability of false match (calculated by multiplying together the same estimated probabilities that are obtained for the weight calculation) exceeds a supplied tolerable rate.

The Census Bureau has implemented a computer matching program based on the Fellegi-Sunter algorithm (Jaro, 1989; Laplant, 1988; Winkler, 1988; Winkler, 1989). In recent test censuses, over 70% of the records from the census and PES were declared matched by computer (see, e.g., Record Linkage Staff, 1986).

However, earlier research has shown (as one's intuition might suggest) that the independence assumption underlying the Fellegi-Sunter algorithm does not hold for the data used in computer matching of census individuals to individuals in the PES (Kelley, 1986). The procedure for setting a cutoff weight in the Fellegi-Sunter framework relies heavily on this independence assumption in that the probability of false match associated with different patterns of agreement is obtained by multiplying together probabilities of agreement on individual variables. Violations of the independence assumption raise questions about whether the procedure for setting a cutoff weight is correctly calibrated to the tolerable false match rate specified by the operator of the computer matching program.

## II. EVALUATION OF THE FELLEGI-SUNTER METHOD OF SETTING CUTOFF WEIGHTS

The following results (Table 1) show the observed false match rate (number of false matches / number of declared matches) associated with various levels of the tolerable false match rates supplied by the operator of the computer matching program. The data for this evaluation come from the 1986 test census of Los Angeles. These findings illustrate the failure of the Fellegi-Sunter technique for establishing cutoffs in the census matching operation.

Table 1. Tolerable false match rate supplied by user of matching program, and observed false match rates, based on data from 1986 test census of Los Angeles.

| Supplied false match rate | Observed false match rate |
|---|---|
| 0.05 | 0.0627 |
| 0.04 | 0.0620 |
| 0.03 | 0.0620 |
| 0.02 | 0.0619 |
| 0.01 | 0.0602 |
| 0.001 | 0.0497 |
| 0.0001 | 0.0365 |
| 0.00001 | 0.0224 |
| 0.000001 | 0.0067 |
| 0.0000001 | 0.0067 |
| 0.00000001 | 0.0067 |

The observed error rates in Table 1 are always higher than the supposedly "fixed" false match rate supplied by the user of the computer matching program. Also, the Fellegi-Sunter technique for setting cutoffs has the intuitively unsatisfying property that different prescribed error rates can lead to the same cutoff weight (e.g. for the second and third entries in the table, prescribed false match rates of 0.04 and 0.03 both led to 15816 records being declared matched with 980 false matches; prescribed false match rates of 0.000001, 0.0000001, and 0.00000001 all led to 14175 declared matches with 95 false matches).

For some time, computer matching experts at the Census Bureau have been aware of the poor calibration of the Fellegi-Sunter technique for setting cutoff levels. The method for setting cutoffs in computer matching that is currently being used at the Census Bureau is a fully manual approach in which a hardcopy printout of pairs of candidate matches is examined by "eyeballing" the data and setting the cutoff when the candidate pairs begin to look dissimilar. To obtain estimates of the false match rate associated with this manual procedure, it is necessary to have clerks review the declared matches to see whether they are accurate.

The research described in this paper focusses on a new idea for automating the setting of cutoffs so that accurate internal estimates of false match rates in computer matching would be routinely available without needing to resort to extensive clerical review of matched pairs. These methods for calibrating the false match rate could be applied to any matching technique that is based on a univariate composite weight (such as the Fellegi-Sunter weighting algorithm and other similar weighting approaches).

The main idea behind the proposed calibration technique is to use past data, with match status (i.e. whether a pair is a true or a false match) known from clerical matching, as a "training sample" to get information about the distribution of weights in a current computer matching problem. One advantage of using past data is that the approach does not need to rely on the assumption that agreement on different variables is independent.

## III. PROPOSED CALIBRATION TECHNIQUE

Two distributions are relevant to this problem, both conditional on declared matches: the distribution of weight given that a pair of records is a true match, p(w|true), and the distribution of weight given that a pair of records is a false match, p(w|false). The observed weights for declared matched pairs come from a mixture of these two distributions.

To fix ideas, suppose that after the computer matching program is run we obtain a list of potential matches ranked from highest weight to lowest weight. We can adjust the cutoff level to obtain more or fewer declared matches, thereby presumably obtaining more or fewer false matches. Underlying this framework is an assumption that the probability of true match is a monotone function of the weight; that is, the higher the weight, the higher the probability of true match and the lower the probability of false match.

The training sample is thus devoted to getting information about the two components of the mixture distribution. Although the distribution of weights in previous data sets may not be exactly like the distribution of weights in the current data set, it seems plausible that the shape of the weight distribution might be fairly similar. Specifically, the training sample is used to provide information about the distributional form of each of the components of the mixture distribution and to provide information about the ratio of the variances of the two components.

Data on weights from computer matching in the 1986 test census of Los Angeles are shown in Figure 1 and data from the 1988 test census of St. Louis are shown in Figure 2. The pictured distributions were obtained by declaring 80% of the records in the PES file from both sites to be matched. (Eighty percent of the file matched is greater than

Figure 1. Distributions of observed weights for true matches (top) and false matches (bottom) in 1986 Los Angeles data.

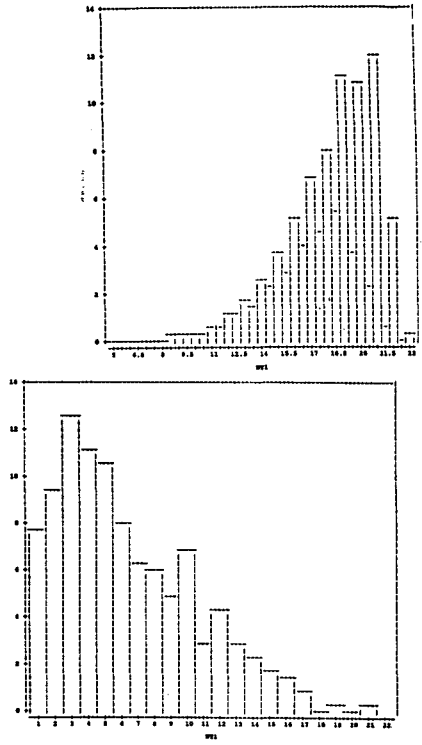
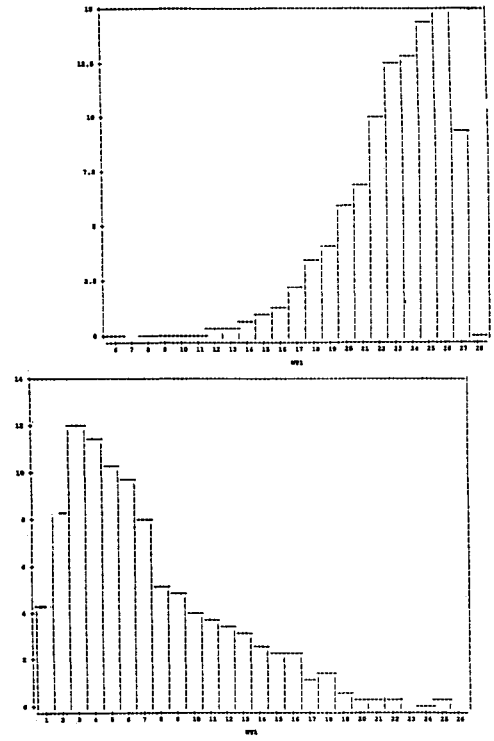


Figure 2. Distributions of observed weights for true matches (top) and false matches (bottom) in 1988 St. Louis data.



the proportion of records that were actually declared matched by computer in these test censuses and leads to a considerable number of false matches, but this is done intentionally here so as to provide a full view of the false match component.)

As is apparent from these figures, the shapes of the two components in the different data sets are very similar. The primary difference is that the proportion of declared matches in the Los Angeles data that are false matches is 5.9% (924 observations in the false match component, 14822 in the true match component), and for the St. Louis data the proportion of declared matches that are false matches is 9.7% (1049 in the false match component, 9820 in the true match component).

As can also be seen from the data, the distributions of weights are truncated both above and below. The cutoff weight truncates the weight distribution on the lower end (that is, no declared matches have weight lower than the cutoff weight), and the weight distribution is effectively truncated above by the maximum possible weight (which corresponds to perfect agreement between two records). The maximum possible weight is a known quantity, and if the goal of a calibration method is to estimate the false match rate for a given cutoff weight, then the cutoff weight can be considered known as well.

To proceed, a model is needed to describe the way that false matches enter the data. After fitting a model, it would then be possible to estimate the false match rate for different specified cutoff levels by using predicted values from the fitted model.

The approach used here was to find a suitable transformation of the components of the mixture distribution so that the resulting distribution would be normalized. The procedure for selecting a transformation would be based on the maximum likelihood method of Box and Cox (1964) that is often used in a regression framework to satisfy normal distribution assumptions in that contee|. The ensuing discussion will focus on two model-fitting efforts for this problem and assumes that a suitable transformation has already been applied to the data.

Let $w_i$ be the weight ascribed to the $i^{th}$ pair of records declared matched (i=1,2,...,n) after a transformation has been applied to normalize the weight distribution, let $\mu_T$ and $\mu_F$ be the means of the two components, and $\sigma_T^2$ and $\sigma_F^2$ be the variances of the two components, with subscript T referring to the distribution of weight for true matches and F referring to the distribution of weight for false matches. Let $\lambda$ be the proportion of weights coming from the false match component, and let $Z_i$ be an (unobserved) indicator variable for each pair of records which takes the value 1 if the pair is a false match and 0 if the pair is a true match.

Model 1. The observed weights are a transformed mixture of truncated normal distributions. The model can be written:

$$w_i|\mu_T,\mu_F,\sigma_T^2,\sigma_F^2,Z_i \sim$$

$$\text{Trunc-N}\ (\mu_T(1-Z_i)+\mu_F Z_i\ ,\ \sigma_T^2(1-Z_i)+\sigma_F^2 Z_i,\ A,\ B)$$

$$Z_i \sim \text{Bernoulli}(\lambda)$$

where A is the lower truncation point (i.e. the cutoff weight) and B is the upper truncation point (i.e. the maximum possible weight).

The motivation for the use of the truncated normal distribution is that the weight distributions are truncated and that some transformation of the components might appear to be fairly normal; the motivation for the mixture model framework was mentioned before.

This formulation of a mixture model in terms of unobserved indicator variables specifying which component of the mixture distribution an observation comes from has been suggested by several authors [see e.g. Dempster, Laird, and Rubin (1977); Aitkin and Rubin (1985); Titterington, Smith, and Makov (1985); Little and Rubin (1987)]. This framework motivates the use of the EM algorithm for fitting mixture models by maximum likelihood. The unobserved $Z_i$'s are treated as missing data, so the iterative steps of the EM algorithm involve obtaining the expected values of the $Z_i$'s given observed data and current parameter estimates, and maximizing the complete data likelihood conditional on current values of the sufficient statistics.

For the sake of fitting mixture models, it is necessary to constrain the variances of the two components. Otherwise, the fitted model would suggest that one component consists of a single observation (with zero variance) and that the other component consists of the rest of the observations, since with the $\sigma$ in the denominator of the normal density the likelihood is unbounded near $\sigma$=0, at the boundary of the parameter space [see Aitkin and Rubin (1985)].

For this reason, it is important to obtain from the training sample not only information about an appropriate transformation but also information about the ratio of the variances of the two components. The procedures used here constrained the ratio of the variances in the fitted model to equal the ratio of the variances observed in the training sample, thinking of the variance ratio as a piece of information that might reasonably be similar in different computer matching settings.

Maximum likelihood estimation for the parameters of the truncated normal distribution is discussed in Dempster, Laird, and Rubin (1977). Maximum likelihood estimates of the mean and variance based on a sample from a truncated normal distribution are obtained by an EM algorithm that considers what the contribution to the likelihood would have been had we seen observations in the tails of the normal distribution.

Instabilities arose in fitting Model 1. The problem was that without constraining the tail area of the truncated normal distributions by borrowing more information from past data, the fitted mixture could suggest that the portion of the normal distribution being truncated was very large, as if the observed data constituted a tiny portion of the right tail of the false match component and of the left tail of the true match component. In other words, the fitted distributions were way too variable to correspond to the reality of the situation and were way too variable to derive reliable estimates of the false match rate in the observed data. As a result, Model 1 was set aside in favor of the following simplified model.

Model 2. The observed weights are a transformed mixture of normal distributions. The model can be written (again assuming a suitable transformation has already been applied to the data):

$$w_i|\mu_T,\mu_F,\sigma_T^2,\sigma_F^2,Z_i \sim$$

$$N\ (\mu_T(1-Z_i)+\mu_F Z_i\ ,\ \sigma_T^2(1-Z_i)+\sigma_F^2 Z_i\ )$$

$$Z_i \sim \text{Bernoulli}(\lambda)$$

The assumption that the components are normally distributed ignores the presence of truncation in the data. However, since we have a fairly full view of the two components, it was thought that the presence of truncation would not affect the estimates of the parameters very much. Further, the presence of truncation could still be accounted for as an ad hoc adjustment to the calculation of an estimated false match rate under the mixture-of-normals model by ignoring the contribution of the tail areas above the upper truncation point.

Using Model 2, the fitting of the mixture model no longer suffers from the instabilities encountered in fitting Model 1. (The constraint that the variance ratio in the fitted model be equal to the variance ratio observed in the training sample is again invoked in this case.)

The false match rate can be expressed as a function of the parameters in the model in the following way:

$$\text{FMR} = \frac{\lambda\ \{1 - \Phi[(A-\mu_F)/\sigma_F]\}}{\lambda\{1-\Phi[(A-\mu_F)/\sigma_F]\}+(1-\lambda)\{1-\Phi[(A-\mu_T)/\sigma_T]\}}$$

where FMR denotes the false match rate. The contribution of the tail above the upper truncation point can be discarded by

substituting $\Phi[(B-\mu_j)/\sigma_j]$ for the 1's inside the bracketed expressions (j=F,T as appropriate).

## IV. RESULTS

The results of using Los Angeles data as a training sample to predict error rates in the St. Louis data are shown in Table 2. The results of using St. Louis data as a training sample to predict error rates in Los Angeles are given in Table 3. Table 4 shows the results of the modified procedure in which Model 2 is fitted to Los Angeles data (using St. Louis data as the training sample) but the truncation of the upper tail is taken into account in calculating false match rates.

The primary conclusions at this stage of the research are the following:

1. The calibration of the predicted error rates to the observed error rates in the St. Louis data (Table 2) is quite good across a range of cutoff levels. If this degree of success could be consistently replicated, then this calibration method would be of considerable practical value.

2. The predicted error rates in the Los Angeles data were not as good, with the predicted false match rates being several times the observed false match rates, suggesting that refinements in the model are needed.

3. The predicted error rates are always conservative in these examples. There is no guarantee that predicted error rates from such methods would always be conservative, but in this regard the procedure probably represents an improvement over the Fellegi-Sunter approach, which was always "anti-conservative" in the sense that the true error rates were always higher than the predicted error rates.

4. Discarding the contribution of the upper tail brings the predicted error rates for the Los Angeles data closer to the observed error rates.

5. The predicted false match rates are not monotone increasing as the cutoff weight decreases, which our intuition tells us should be the case. This phenomenon was observed in all of the examples, even when the upper tail of the components was discarded in calculating false match rates. The apparent explanation for this effect is that the false match component was estimated to have a variance between 20 and 25 times that of the true match component, so that the upper part of the false match density would be disproportionately large. This phenomenon only existed, however, in a region where we would probably not be interested in setting a cutoff level anyway, with less than 60% of the file being declared matched.

Table 2. Predicted false match rates from Model 2 in 1988 St. Louis data using 1986 Los Angeles data as past information. (Proportion matched refers to the proportion of the St. Louis PES file declared matched for the particular cutoff weight.)

| Expected false match rate | Observed false match rate | Proportion matched |
|---|---|---|
| 0.00422 | 0.00053 | 0.280 |
| 0.00351 | 0.00056 | 0.524 |
| 0.00429 | 0.00128 | 0.633 |
| 0.00630 | 0.00353 | 0.688 |
| 0.00951 | 0.00764 | 0.714 |
| 0.01392 | 0.01238 | 0.726 |
| 0.01963 | 0.01914 | 0.735 |
| 0.02665 | 0.02262 | 0.739 |
| 0.03488 | 0.02725 | 0.743 |
| 0.04409 | 0.03204 | 0.747 |
| 0.05390 | 0.03745 | 0.752 |

Table 3. Predicted false match rates from Model 2 in 1986 Los Angeles data using 1988 St. Louis data as past information.

| Expected false match rate | Observed false match rate | Proportion matched |
|---|---|---|
| 0.04583 | 0.00104 | 0.146 |
| 0.02707 | 0.00072 | 0.353 |
| 0.02127 | 0.00093 | 0.493 |
| 0.02085 | 0.00216 | 0.613 |
| 0.02338 | 0.00348 | 0.672 |
| 0.02770 | 0.00515 | 0.710 |
| 0.03304 | 0.00767 | 0.735 |
| 0.03897 | 0.00999 | 0.747 |
| 0.04530 | 0.01343 | 0.756 |
| 0.05188 | 0.01550 | 0.760 |

Table 4. Predicted false match rates from Model 2, with expected error rates calculated without including tail areas above the upper truncation point, in 1986 Los Angeles data using 1988 St. Louis data as past information.

| Expected false match rate | Observed false match rate | Proportion matched |
|---|---|---|
| 0.01762 | 0.00104 | 0.146 |
| 0.01377 | 0.00072 | 0.353 |
| 0.01300 | 0.00093 | 0.493 |
| 0.01442 | 0.00216 | 0.613 |
| 0.01765 | 0.00348 | 0.672 |
| 0.02229 | 0.00515 | 0.710 |
| 0.02787 | 0.00767 | 0.735 |
| 0.03406 | 0.00999 | 0.747 |
| 0.04067 | 0.01343 | 0.756 |
| 0.04757 | 0.01550 | 0.760 |

6. The shapes of the true and false match components are similar for the different sites, suggesting that approaches based on using past information in current computer matching problems may be successful. These results are very preliminary, and other models may do a better job of capitalizing on the similarity between the weight distributions from different sites.

## V. CURRENT RESEARCH

Experience up to this point has sparked ideas for other modelling approaches. One method being pursued would allow for different transformations of the true and false match component. This would be done by fitting the mixture-of-transformed-normals model and considering the Box-Cox power transformation parameters of each component as parameters whose likelihood should be maximized at each iteration of the EM algorithm, as opposed to viewing the transformation of the data as a "pre-processing" step and fitting the mixture model on the already transformed data. This approach would help account for the dissimilarity in the shapes of the true and false match components, which are skewed in different ways. This dissimilarity seems to be a source of some of the departure of the predicted false match rates from the truth.

Another issue that will have to be addressed is the fact that although the shapes of the weight distributions are similar across sites, the proportion of records ($\lambda$) in the false match component will vary from site to site. Since the estimated false match rate is sensitive to the estimated proportion of records in the lower component, it will be important to incorporate uncertainty about this proportion in the estimation process.

With data available from a few different sites, it should be possible to gather information about the site to site variability of the proportion of records in the lower component. Uncertainty in the estimated false match rate could then be characterized on the basis of the model and prior uncertainty in the parameters.

## References


Aitkin, M., and Rubin, D. (1985). "Estimation and Hypothesis Testing in Finite Mixture Models," JRSSB, 47, pp. 67-75.

Box, G.E.P., and Cox, D.R. (1964). "An Analysis of Transformations (with discussion)," JRSSB, 26, pp. 211-246.

Chernoff, H. (1977). "Some Applications of a Method of Identifying an Element of a Large Multidimensional Population," in Multivariate Analysis IV, P.R. Krishnaiah, ed., pp. 445-456.

Chernoff, H. (1980). "The Identification of an Element of a Large Population in the Presence of Noise," Annals of Statistics, 8, pp. 1179-1197.

Citro, C., and Cohen, M., eds., (1985). The Bicentennial Census: New Directions for Methodology in 1990, Washington, DC: National Academy Press.

Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum Likelihood Estimation from Incomplete data via the EM algorithm," JRSSB, 39, pp. 1-38.

Fellegi, I., and Sunter, A. (1969). "A Theory for Record Linkage," JASA, 64, pp.1183-1210.

Jaro, M. (1989). "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Test Census of Tampa, Florida," JASA, 84, pp. 414-420.

Kelley, R.P. (1986). "Robustness of the Census Bureau's Record Linkage System," ASA Proceedings, Section on Survey Research Methods, pp. 620-624.

Laplant, W. (1988). "User's Guide for the Generalized Record Linakge Program Generator (GENLINK)," Technical Report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Little, R., and Rubin, D. (1987). Statistical Analysis with Missing Data, New York: John Wiley.

Newcombe, H., Kennedy, J., Axford, S., and James, A. (1959). "Automatic Linkage of Vital Records," Science, 130, pp. 954-959.

Newcombe, H. (1988). Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business, Oxford: Oxford University Press.

Record Linkage Staff (1986). "Summary of Matching for 1986 Los Angeles PES," Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Tepping, B. (1968). "A Model for Optimum Linkage of Records," JASA, 63, pp.1321-1332.

Titterington, D., Smith, A., and Makov, U. (1985). Statistical Analysis of Finite Mixture Distributions, New York: John Wiley.

Winkler, W. (1988). "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," ASA Proceedings, Section on Survey Research Methods.

Winkler, W. (1989). "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, pp. 145-155.