Chapter
4

Modeling Issues and the Use of Experience in Record Linkage

Michael D. Larsen, Harvard University

Abstract

The goal of record linkage is to link quickly and accurately records corresponding to the same person or entity. Fellegi and Sunter (1969) proposed a statistical model for record linkage that assumes pairs of entries, one from each of two files, either are matches corresponding to a single person or nonmatches arising from two different people. Certain patterns of agreements and disagreements on variables in the two files are more likely among matches than among nonmatches. The observed patterns can be viewed as arising from a mixture distribution.

Mixture models, which for discrete data are generalizations of latent-class models, can be fit to comparison patterns in order to find matching and nonmatching pairs of records. Mixture models, when used with data from the U.S. Decennial Census — Post Enumeration Survey, quickly give accurate results.

A critical issue in new record-linkage problems is determining when the mixture models consistently identify matches and nonmatches, rather than some other division of the pairs of records. A method that uses information based on experience, identifies records to review, and incorporates clerically-reviewed data is proposed.

Introduction

Record linkage entails comparing records in one or more data files and can be implemented for unduplication or to enable analyses of relationships between variables in two or more files. Candidate records being compared really arise from a single person or from two different individuals. Administrative data bases are large and clerical review to find matching and nonmatching pairs is expensive in terms of human resources, money, and time. Automated linkage involves using computers to perform matching operations quickly and accurately.

Mixture models can be used when the population is composed of underlying and possibly unidentified subpopulations. The clerks manually identify matches and nonmatches, while mixture models can be fit to unreviewed data in the hopes of finding the same groups. However, mixture models applied to some variables can produce groups that fit the data but do not correspond to the desired divisions. A critical issue in this application is determining when the model actually is identifying matches and nonmatches.

A procedure is proposed in this paper that when applied to Census data seems to work well. The more that is known about a particular record linkage application, the better the procedure should work. The size of the two files being matched, the quality of the information recorded in the two files, and any clerical review that has already been completed are incorporated into the procedure. Additionally, the procedure should help clerks be more efficient because it can direct their efforts and increase the value of reviewed data through use in the model.

The paper defines mixture models and discusses estimation of parameters, clustering, and error rates. Then previous theoretical work on record linkage is described. Next, the paper explains the proposed procedure. A summary of the application of the procedure to five Census data sets is given. The paper concludes with a brief summary of results and reiteration of goals.

Mixture Models

An observation $\mathbf{y}_i$ (possibly multivariate) arising from a finite mixture distribution with G classes has probability density

$$p(\mathbf{y}_i \,|\, \Pi, \Theta\,) = \Sigma_{g=1,G}\, \pi_g\, p_g(\,\mathbf{y}_i \;|\, \theta_g), \tag{1}$$

where π_g $(\Sigma_{g=1,G}\, \pi_g\; =1)$, p_g, and θ_g are the proportion, the density of observations, and the distributional parameters, respectively, in class g, and Π and Θ are abbreviated notation for the collections of proportions and parameters, respectively. The likelihood for π and θ based on a set of n observations is a product with index i=1,...,n of formula (1).

The variables considered in this paper are dichotomous and define a table of counts, which can have its cells indexed by i. In the application, each observation is ten dimensional, so n=1024. The mixture classes are in effect subtables, which when combined yield the observed table. The density $p_g(\bullet\,|\,\bullet)$ in mixture class g can be defined by a log-linear model on the expected counts in the cells of the subtable. The relationship among variables described by the log linear model can be the same or different in the various classes. When the variables defining the table in all classes are independent conditional on the class, the model is the traditional latent-class model. Sources for latent-class models include Goodman (1974) and Haberman (1974, 1979).

Maximum likelihood estimates of π and θ can be obtained using the EM (Dempster, Laird, Rubin 1977) and ECM (Meng and Rubin, 1993) algorithms. The ECM algorithm is needed when the log linear model in one or more of the classes can not be fit in closed form, but has to be estimated using iterative proportional fitting.

The algorithms treat estimation as a missing data problem. The unobserved data are the counts in each pattern in each class and can be represented by a matrix $\mathbf{z}$ with n rows and G columns, where entry z_{ig} is the number of observations with pattern i in class g. If the latent counts were known, the density would be

$$p(\,\mathbf{y},\, \mathbf{z}\,|\,\Pi,\Theta) = \;\Pi_{i=1,n}\; \Pi_{g=1,G}\; (\pi_g\; p_g(\mathbf{y}_i \;|\, \theta_g))^{z_{ig}}. \tag{2}$$

Classified data can be used along with unclassified data in algorithms for estimating parameters. The density then is a combination of formulas (2) and a product over i of (1). Known matches and nonmatches, either from a previous similar matching problem or from clerk-reviewed data in a new problem, can be very valuable since subtables tend to be similar to the classified data.

Probabilities of group membership for unclassified data can be computed using Bayes' Theorem. For the kth observation in the ith cell, the probability of being in class g $(z_{igk}=1)$ is

$$p(z_{igk} =1 \;|\; \mathbf{y}_i,\, \pi,\theta) = \;\pi_g\; p_g(\mathbf{y}_i \;|\, \theta_g) \,/\, \Pi_{h=1,G}\; \pi_h\;\; p_h(\mathbf{y}_i\,|\,\theta_h)\;. \tag{3}$$

Probability (3) is the same for all observations in cell i. Probabilities of class membership relate to probabilities of being a match and nonmatch only to the degree that mixture classes are similar to matches and nonmatches.

The probabilities of group membership can be used to cluster the cells of the table by sorting the cells of the table according to descending probability of membership in a selected class. An estimated error rate at a given probability cut-off is obtained by dividing the expected number of observations not in a class by the total number of observations assigned to a class. As an error rate is reduced by assigning fewer cells to a class, the number of observations in a nebulous group not assigned to a class increases.

Before the match and nonmatch status is determined by clerks tentative declarations as probable match and probable nonmatch can be made using mixture models. It is necessary to choose a class or classes to be used as probable matches and probable nonmatches, which usually can be done by looking a probabilities of agreement on fields in the mixture classes. The estimated error rates from the mixture model correspond to the actual rate of misclassification of matches and nonmatches only to the degree that the mixture classes correspond to match and nonmatch groups.

Record Linkage Theory

Fellegi and Sunter (1969) proposed a statistical model for record linkage that assumes pairs of entries, one from each of two files, either are matches corresponding to a single person or nonmatches arising from two different people. Patterns of agreements and disagreements on variables have probabilities of occurring among matches and among nonmatches. If the pairs of records are ordered according to the likelihood ratio for being a match versus being a nonmatch and two cut-off points are chosen, one above which pairs are declared matches and one below which pairs are declared nonmatches, the procedure is optimal in the sense of minimizing the size of the undeclared set at given error levels.

Fellegi and Sunter (1969) suggested methods to estimate the unknown probabilities involved in the likelihood ratio. Some of their simplifying assumptions, such as the independence of agreement on fields of information within matches and nonmatches, have continued to be used extensively in applications.

In the methods proposed in this paper, the likelihood ratio is estimated using the mixture model. If the first class is the class of probable matches, the likelihood ratio is for pattern i is

$$p(g=1 \mid y_i, \Pi, \Theta)/\, p(g\neq 1 \mid \mathbf{y}_i, \Pi, \Theta) = \pi_1\, p_1(\mathbf{y}_i \mid \theta_1) \,/\, \Sigma_{g=2,G}\, \pi_g\, p_g(\mathbf{y}_i \mid \theta_g). \qquad (4)$$

The success depends on the relationship between the implied latent groups and the match and nonmatch categories.

The choice of cutoff values for declaring observations matches and nonmatches is critical, as demonstrated by Belin (1993). Belin and Rubin (1995) have shown that previous applications of the Fellegi-Sunter procedure do not always have their specified error levels. In applications, the cutoff values often are determined by manual review of observations in the “gray area” or likely to be sent to clerical review.

In the current paper, a cutoff can be chosen using mixture model results to achieve a specified error level, but the actual error level might or might not be close to the estimated level.

Winkler (1988, 1989a, 1989b, 1989c, 1990, 1992, 1993, 1994) and Thibaudeau (1989, 1993) have used mixture models of the type used in this article in record-linkage applications at the Census Bureau. The new procedure in this article addresses the critical question of when a mixture-model approach is appropriate for a new record-linkage situation.

Belin and Rubin (1995) developed a procedure for estimating error rates in some situations using the Fellegi-Sunter algorithm and applied it to Census data. Their method needs clerically-reviewed training data from similar record linkages and works well when the distributions of likelihood values (4) for matches and nonmatches are well separated. The new approach of this paper does not require training data, but could use it as classified observations, and provides its own estimates of error rates as described for mixture models.

Other applications of record linkage have used more information, such as frequency of names and string comparator metrics, than simple binary agree/disagree comparisons of fields. While there obviously is value in more detailed comparisons, this paper uses only multivariate dichotomous data and leaves development of model-based alternatives to current methods for more complicated data to later.

Procedure

The procedure for applying mixture models to record-linkage vector comparisons is specified below. It has many informal aspects some of which correspond to decisions often made in practical applications. Later work will investigate formalizing the procedure.

- Fit a collection of mixture models that have been successful in previous similar record-linkage problems to the data.
- Select a model with a class having (a) high probabilities of agreement on important fields, (b) probability of class membership near the expected percent of pairs that are matches, and (c) probabilities of class membership for individual comparison patterns near 0 or 1.
- Identify a set of records for clerks to review using the mixture model results.
- Refit the mixture model using both the classified and unclassified data.
- Cycle through the two previous steps as money and time allow, or until satisfied with results.

Models that can be used in step (1) are illustrated below. Some searching through other model possibilities might have to be done. In step (2), from the observed, unclassified data, it is possible to compute the probability of agreement on fields and combinations of fields. The probabilities should be higher in the probable match class than overall. The percent of pairs that are matches is limited by the size of the smaller of the two lists contributing to candidate pairs. If a class is much larger than the size of the smaller list, it must contain several nonmatches. Of course no single model may be clearly preferable given the informal statement of criteria.

In step (3), records to review can be identified by first accumulating pairs into the probable match class according to probability of membership until a certain point and then reviewing pairs at the chosen boundary. The boundary used in this paper is the minimum of the estimated proportion in the probable mixture class and the size of the small list divided by the total number of pairs.

In the next section, the procedure is applied to five Census data sets and produces good results. Many aspects of the procedure parallel successful applications of the Fellegi-Sunter approach to record linkage. Different mixture models give slightly different estimates of the likelihood ratio just as different estimation methods currently used in practice lead to different orderings of pairs.

Application

In 1988, a trial census and post-enumeration survey (PES) were conducted. Data from two urban sites are referred to as D88a and D88b in this section. In 1990, following the census, a PES was conducted in three locations. D90a and D90b are the data sets from urban sites, and D90c are data from a rural site. Table 1 contains summaries of the five data sets. Not all possible pairs of records from the census and PES were compared. Candidate match pairs had to agree on a minimal set of characteristics. Sites vary in size, proportion that are matches, and probabilities of agreeing on fields. D90c is rural, and its address information is not very precise. Thus, relatively more pairs are compared, yielding lower probabilities of agreement and a lower proportion of matches.

Table 1. -- Summary of Five Census/Post-Enumeration Survey Data Sets, Including Probabilities of Agreements on Fields Overall (and for Matches)

Data set	**D88a**	**D88b**	**D90a**	**D90b**	**D90c**
Census size	12072	9794	5022	4539	2414
PES size	15048	7649	5212	4859	4187
Total pairs	116305	56773	37327	38795	39214
Matches	11092	6878	3596	3488	1261
Nonmatch	105213	49895	33731	35307	37953
Last name	.32(.98)	.41(.99)	.31(.98)	.29(.98)	.26(.98)
First name	.11(.95)	.14(.98)	.12(.96)	.11(.95)	.06(.95
House #	.28(.97)	.18(.50)	.30(.95)	.27(.94)	.06(.42)
Street	.60(.96)	.28(.49)	.37(.67)	.59(.95)	.11(.44)
Phone #	.19(.71)	.31(.83)	.19(.69)	.18(.66)	.06(.45)
Age	.16(.85)	.23(.94)	.19(.89)	.17(.88)	.11(.89)
Relation to head of household	.19(.48)	.20(.54)	.16(.46)	.19(.48)	.25(.56)
Martial status	.41(.84)	.44(.89)	.36(.78)	.42(.85)	.42(.88)
Sex	.53(.96)	.53(.98)	.52(.97)	.52(.96)	.50(.96)
Race	.91(.97)	.93(.98)	.80(.93)	.83(.91)	.80(.86)

Mixture models considered in this application have either two or three classes. Models for the variables within each class include either main effects only, all two-way interactions, all three-way interactions, a five-way interaction between typically household variables (last name, house number, street name, phone number, and race) and a five-way interaction between typically personal variables (the other five), and a set of interactions described by Armstrong and Mayda (1993). The actual models are described in Table 2.

Table 2. -- Mixture Models Considered for Each Data Set

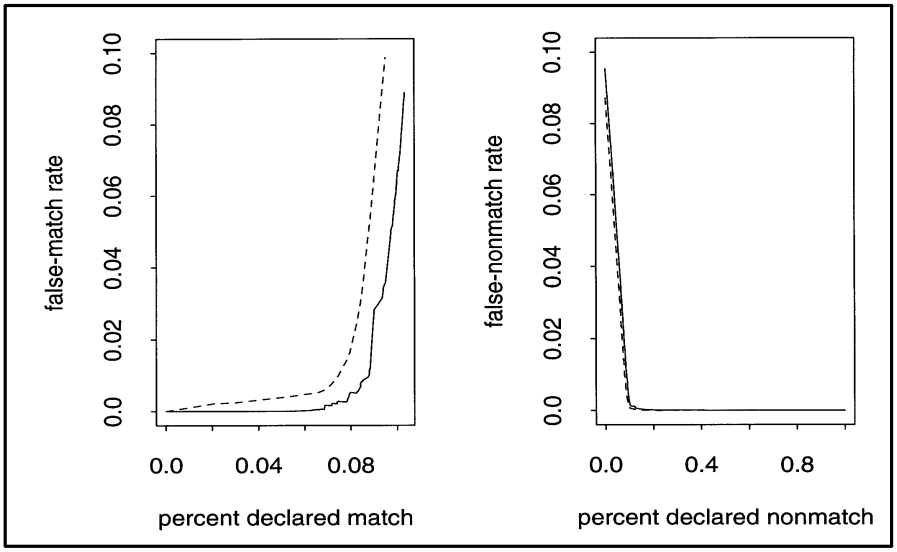
Abbreviation	**Model Class 1**	**Model Class 2**	**Model Class 3**
2C CI	Independent	Independent	
2C CI-2way	Independent	2way interactions	
2C CI-3way	Independent	3way interactions	
3C CI	Independent	Independent	Independent
3C CI-2way	Independent	2way interactions	2way interactions

2C CI-HP	Independent	5way interactions (Household, Personal)	
2C HP	5way interactions	5way interactions	
3C CI-HP	Independent	5way interactions	5way interactions
2C AM	Independent	Armstrong and Mayda, 1993	
3C AM	Independent	Armstrong and Mayda, 1993	Armstrong and Mayda, 1993

To illustrate the results from fitting a mixture model, the three-class conditional independence model (model 4) was fit to D88a. Figure 1 contains plots of the estimated and actual false-match and false-nonmatch rate. At an error rate of .005, using the estimated false-match curve, 7462 matches and 3 nonmatches are declared matches, giving an actual error rate of .0004. At an estimated error rate of .01, 8596 matches and 23 nonmatches are declared matches, giving an actual error rate of .0027.

Figure 1. -- False-Match and False-Nonmatch Rates From Fitting a Three-Class Conditional Independence Mixture to D88a

(The solid lines are actual and the dashed lines are estimated error rates)



The three-class conditional independence model works for the D88a data, because one of the classes tends to consist of pairs that agree on most comparisons. The medium-sized mixture class tends to agree on variables defining households, but to disagree on personal variables. This class can be called the same household-different person class. The third class tends to disagree on all comparisons. The three-class conditional-independence model also produces good results for the other data sets, except for D90c data. The difference could be caused by the fact that D90c is from a rural area, while the others have a lot of the population in urban settings with better household identifiers.

The search procedure was applied to each of the five data sets. The models selected to start with are given in the second row of Table 3. The number of matches and nonmatches declared matches with estimated false-match rates .005 and .01 are given in lines three and four of Table 3. The number of matches and nonmatches declared nonmatches with estimated false-nonmatch rates .001 and .005 are given in the next two lines. Most models are successfully separating matches and nonmatches. However, in some cases, the rapid rise of estimated false-match rates means few observations can safely be declared matches.

Table 3. -- Initial Model Selected for Each Data Set, Along with Matches and Nonmatches Declared Matches and Nonmatches for Two False-Match (FMR) and False-Nonmatch (FNMR) Rates

Parentheses enclose (match, nonmatch) counts

Data set	**D88a**	**D88b**	**D90a**	**D90b**	**D90c**
Model	3C CI	2C CI-2way	3C CI	3C CI	3C CI-HP
.005 FMR	(7442,2)	(0,0)	(2802,27)	(2421,12)	(766,99)
.01 FMR	(8596,23)	(24,0)	(3083,50)	(2812,25)	(997,112)
.001 FNMR	(260, 104587)	(3455, 49855)	(124, 33244)	(69, 34507)	(32, 36900)
.005 FNMR	(1021, 105117)	(3858, 49882)	(248, 33469)	(234, 35126)	(61, 37571)
Total Counts	(11092, 105213)	(6878, 49895)	(3596, 33731)	(3488, 35307)	(1261, 37953)

The models used for D88a, D88b, and D90c were clearly the best candidates among the proposed models for trying to identify matches. In the cases of D90a and D90b, the model with two classes, one with conditional independence between the variables and the other with all two-way interactions, were close competitors to the three-class conditional-independence model. The models chosen for D90a and D90b had estimated error rates that grew slowly until approximately the proportion in the smallest class. The models not chosen had rapidly rising estimated error rates right away.

Pairs were identified to be reviewed by clerks. For the data set D88a, 5000 pairs were reviewed and error rates reestimated. 1000 pairs were reviewed and then the model was refit 5 times. Then 5000 more pairs were reviewed, 1000 at a time. Table 4 contains results for all 5 data sets. For the smaller data sets, fewer observations were reviewed. Note that in Table 4, the reported estimated false-match rates have been reduced. After about ten percent of the pairs are reviewed, most of the matches and nonmatches can be identified with few errors.

Table 4. -- Matches and Nonmatches Declared Matches and Nonmatches for Two False-Match (FMR) and False-Nonmatch Rates (FNMR) After Reviewing Some Records and Refitting Models

Parentheses enclose (match, nonmatch) counts

Data set	**D88a**	**D88b**	**D90a**	**D90b**	**D90c**
~~Model~~	~~3C CI~~	~~2C CI-2way~~	~~3C CI~~	~~3C CI~~	~~3C CI-HP~~
Reviewed	5000	2500	2000	2000	2000
.001 FMR	(10764, 0)	(2703, 1)	(2620, 10)	(2562, 8)	(48, 2)
.005 FMR	(10917, 27)	(3105, 8)	(3447, 26)	(3347, 17)	(393,5)

.001 FNMR	(58, 102728)	(3339, 49694)	(104, 33657)	(76, 35227)	(40, 37633)
.005 FNMR	(255 , 105212)	(3448, 49866)	(316, 33718)	(206, 35298)	(121, 37863)
Reviewed	10000	5000	4000	4000	4000
.001 FMR	(10916, 13)	(5057, 1)	(3439, 1)	(3341, 3)	(1019, 5)
.005 FMR	(10917, 27)	(6479, 17)	(3456, 18)	(3352, 9)	(1217, 5)
.001 FNMR	(58, 102728)	(246, 49857)	(106, 33688)	(76, 35236)	(32, 37994)
.005 FNMR	(255, 105212)	(433, 49881)	(194, 33731)	(206, 35307)	(186, 37948)
Total counts	(11092, 105213)	(6878, 49895)	(3596, 33731)	(3488, 35307)	(1261, 37953)

Figure 2 (on the next page) illustrates the impact of the addition of clerk-reviewed data on false-match rate estimates for data set D90c. The method performs better on the other data sets with their models than on D90c.

Conclusion

The development of theory related to applications can be useful for several reasons. The mixture-modeling approach of this paper hopefully can provide some insight into adjustments that are made in applications to make current theory work. Aspects of the new procedure with models parallel actual practice without models. The modeling approach also could improve efficiency by helping clerks identify valuable records to review and then using the additional information through the model to learn more about unclassified observations. More formal model selection procedures and models that allow more complex comparison data will increase the usefulness of the theory.

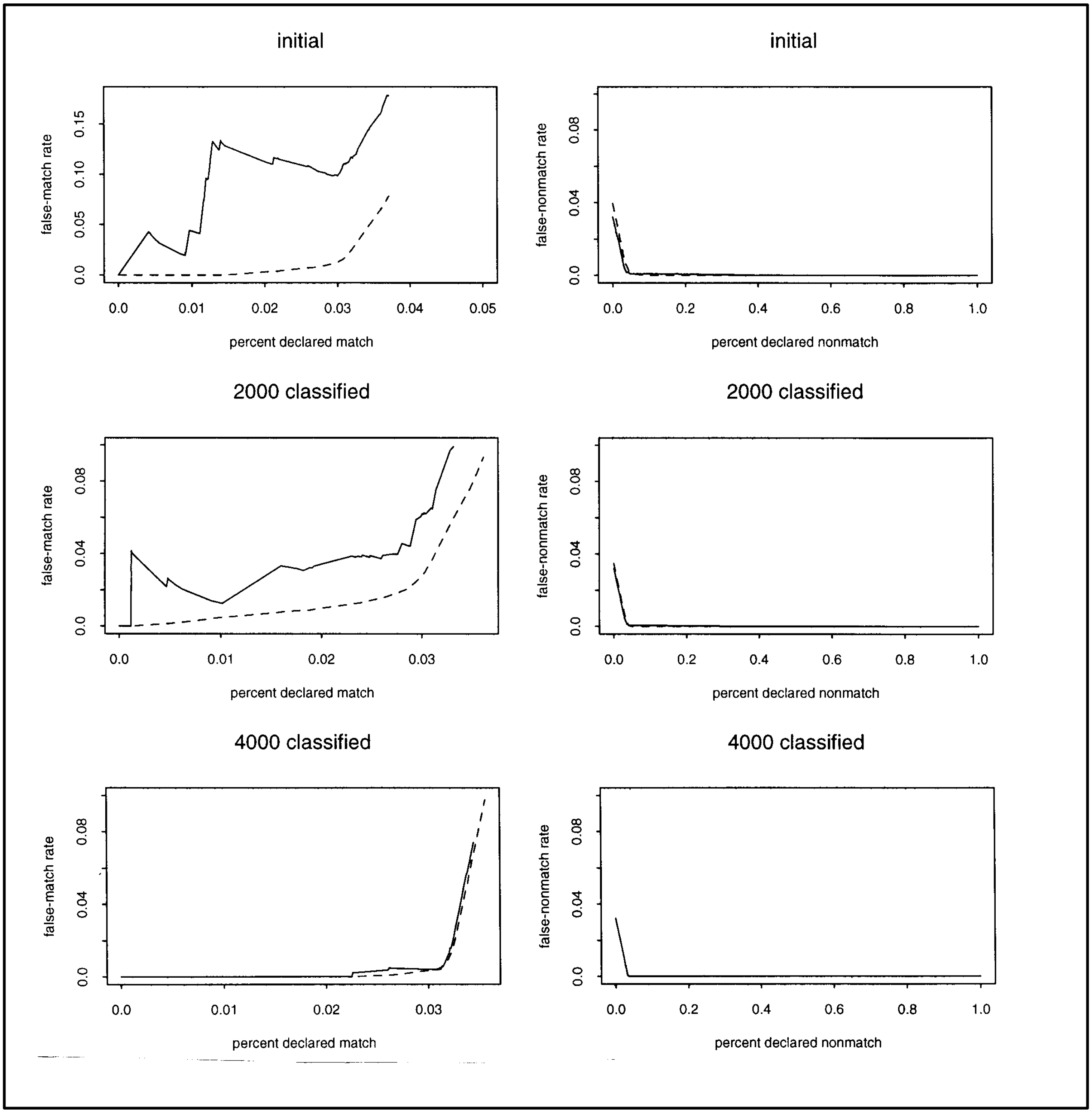
The goal of this paper has been to demonstrate methods that could be used in new record-linkage situations with big lists where accuracy, automation, and efficiency are needed. The procedure identifies matches and nonmatches, directs clerks in their work, and provides cut-offs and estimates of error rates on five Census data sets.

Acknowledgments

The author wishes to thank William E. Winkler and Donald B. Rubin for their support and guidance in this project. Thanks also to William E. Winkler and Fritz Scheuren for organizing the Record Linkage Workshop.

Figure 2. -- False-Match (FMR) and False-Nonmatch (FNMR) Rates for D90c: Initial Estimates, Estimates After Reviewing 2000, and Estimates After Reviewing 4000 Pairs

(Note that the initial FMR plot has different axes than the others)



initial

initial

2000 classified

2000 classified

4000 classified

4000 classified

References

Armstrong, J. B. and Mayda, J. E. (1993). Model-Based Estimation of Record Linkage Error Rates, *Survey Methodology*, 19, 137-147.

Belin, Thomas R. (1993). Evaluation of Sources of Variation in Record Linkage Through a Factorial Experiment, *Survey Methodology*, 19, 13-29.

Belin, Thomas R. and Rubin, Donald B. (1995). A Method for Calibrating False-Match Rates in Record Linkage, *Journal of the American Statistical Association*, 90, 694-707.

Dempster, A. P.; Laird, N. M.; and Rubin, Donald B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm, *Journal of the Royal Statistical Society*, Series B, 39, 1-22, (C/R: 22-37).

Fellegi, Ivan P. and Sumter, Alan B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Goodman, Leo A. (1974). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models, *Biometrika*, 61, 215-231.

Haberman, Shelby J. (1974). Log-Linear Models for Frequency Tables Derived by Indirect Observation: Maximum Likelihood Equations, *The Annals of Statistics*, 2, 911-924.

Haberman, Shelby J. (1979). *Analysis of Qualitative Data*, Vol. 2, New York: Academic Press.

Meng, Xiao-Li and Rubin, Donald B. (1993). Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework, *Biometrika*, 80, 267-278.

Thibaudeau, Yves. (1989). Fitting Log-Linear Models in Computer Matching, *Proceedings of the Section on Statistical Computing, American Statistical Association*, 283-288.

Thibaudeau, Yves. (1993). The Discrimination Power of Dependency Structures in Record Linkage, *Survey Methodology*, 19, 31-38.

Winkler, William E. (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 667-671.

Winkler, William E. (1989a). Frequency-Based Matching in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 778- 783.

Winkler, William E. (1989b). Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Bureau of the Census Annual Research Conference*, 5, 145- 155.

Winkler, William E. (1989c). Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage, *Survey Methodology*, 15, 101-117.

Winkler, William E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 354-359.

Winkler, William E. (1992), Comparative Analysis of Record Linkage Decision Rules, *Proceedings of the Survey Research Methods Section*, *American Statistical Association,* 829- 834.

Winkler, William E. (1993). Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, 274- 279.

Winkler, William E. (1994). Advanced Methods for Record Linkage, *Proceedings of the Survey Research Methods Section*, *American Statistical Association,* 467-472.