ADVANCED METHODS FOR RECORD LINKAGE

William E. Winkler*, Bureau of the Census, Washington DC 20233-9100

KEY WORDS: string comparator, assignment algorithm, EM algorithm, latent class

Record linkage, or computer matching, is a means of creating, updating, and unduplicating lists that may be used in surveys. It serves as a means of linking individual records via name and address information from differing administrative files. If the files are linked using proper mathematical models, then the files can be analyzed using statistical methods such as regression and loglinear models (Scheuren and Winkler 1993).

Modern record linkage represents a collection of methods from three different disciplines: computer science, statistics, and operations research. Whereas the foundations are from statistics, beginning with the seminal work of Newcombe (Newcombe et al 1959, also Newcombe 1988) and Fellegi and Sunter (1969), the means of implementing the methods have primarily involved computer science. Methods from the three disciplines are needed for dealing with the three different types of problems arising in record linkage.

Because pairs of strings often exhibit typographical variation (e.g., Smith versus Smoth), the first need of record linkage is for effective string comparator functions that deal with typographical variations. While approximate string comparison has been a subject of research in computer science for many years, the most effective ideas in the record linkage context were introduced by Jaro (1989; see also Winkler 1990). Budzinsky (1991), in an extensive review of twenty string comparision methods, concluded that the original Jaro method and the extended method due to Winkler (1990) worked second best and best, respectively. Statistics Canada (Nuyens 1993) subsequently added string comparators based on Jaro and Winkler logic to CANLINK, Statistics Canada's matching system.

The second need of record linkage is for effective means of estimating matching parameters and error rates. In addition to proving the theoretical optimality of the decision rule of Newcombe, Fellegi and Sunter (1969) showed how matching parameters could be estimated directly from available data. Their estimation methods admit closed-form solutions only if there are three matching variables and a conditional independence assumption is made. With more variables, the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) can be used. If conditional independence is not assumed (i.e., interactions between agreements of variables such as house number, last name, and street name are allowed), then general computational algorithms (Winkler 1989) can be used. The general algorithm is an example of the MCECM algorithm of Meng and Rubin (1993). An enhancement to the basic algorithm (Winkler 1993) allows weak use of a priori information via convex constraints that restrict the solutions to subportions of the parameter space. The enhancement generalizes the MCECM algorithm.

The third need of record linkage is for a means of forcing 1-1 matching. Jaro (1989) introduced a linear sum assignment procedure (lsap) due to Burkard and Derigs (1980) as a highly effective means of eliminating many pairs that ordinarily might be clerically reviewed. With a household data source containing multiple individuals in a household, it effectively keeps the four pairs associated with father-father, mother-mother, son-son, and daughter-daughter pairs while eliminating the remaining twelve pairs associated with the household. An enhanced algorithm that uses less storage was used during the 1990 Decennial Census (Winkler and Thibaudeau 1991). This paper describes a new algorithm (Winkler 1994a) that can use 0.002 as much storage as the earlier algorithm and can eliminate some subtly erroneous matches that often occur in pairs of general administrative lists having only moderate overlap.

The next three sections describe the string comparator, the parameter-estimation algorithm, and the assignment algorithm, respectively. The results of section 5 provide empirical examples of how matching efficacy is improved for three, small pairs of high quality lists. Section 5 also presents a new method for estimating error rates and compares it to the method of Belin and Rubin (1995). The sixth section provides discussion. The final section consists of a summary and conclusion.

2. APPROXIMATE STRING COMPARISON

Dealing with typographical error can be vitally important in a record linkage context. If comparisons of pairs of strings are only done in an exact character-by-character manner, then many matches may be lost. An extreme example is the Post Enumeration Survey (PES) (Winkler and Thibaudeau 1991, also Jaro 1989) in which, among true matches, almost 20 percent of last names and 25 percent of first names disagreed character-by-character. If matching had been performed on a character-by-character basis, then more than 30 percent of matches would have been missed by computer algorithms that were intended to delineate matches automatically. In such a situation, required manual review and (possibly) matching error would have greatly increased.

In a large study of twenty from the computer science literature, Budzinsky (1991) concluded that the comparators due to Jaro (1989) and Winkler (1990) were the second best and best, respectively. The existing string comparator is augmented with a new algorithm (McLaughlin 1993) that deals with scanning errors (('1' versus 'I') and certain common keypunch errors ('V' versus 'B'). More details of the string comparators are given in Lynch and Winkler (1994) and in the longer technical report.

3. PARAMETER-ESTIMATION VIA THE EM ALGORITHM

The record linkage process attempts to classify pairs in a product space **A × B** from two files A and B into M, the set of true matches, and U, the set of true nonmatches. Fellegi and Sunter (1969), making rigorous concepts introduced by Newcombe (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma \mid M) \;/\; P(\gamma \in \Gamma \mid U) \tag{3.1}$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur.

The decision rule is given by:

If $R > UPPER$, then designate pair as a link.
If $LOWER \leq R \leq UPPER$, then designate pair as a possible link and hold for clerical review. (3.2)
If $R < LOWER$, then designate pair as a nonlink.

The cutoff thresholds *UPPER* and *LOWER* are determined by a priori error bounds on false matches and false nonmatches. The three components of Rule (3.2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (3.1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (3.1) would be small.

Fellegi and Sunter (1969, Theorem) showed that the decision rule is optimal in the sense that for any pair of fixed upper bounds on the rates of false matches and false nonmatches, the clerical review region is minimized over all decision rules on the same comparison space Γ. The theory holds on any subset such as pairs agreeing on a postal code, on street name, or on part of the name field. Ratio R or any monotonely increasing transformation of it (such as given by a logarithm) is defined as a matching weight or *total agreement weight*. In actual applications, the optimality of the decision rule (3.2) is heavily dependent on the accuracy of the estimates of the probabilities given in (3.1). The probabilities in (3.1) are called *matching parameters or matching weights.*

The matching parameters are estimated via the EM algorithm. The EM algorithm allows modelling when interactions between fields occur (i.e., conditional independence does not hold). A generalization of the ECM algorithm of Meng and Rubin (1994) allows use of convex constraints (Winkler 1993, 1994b) that restrict (predispose) solutions to subportions of the parameter space. For instance, a convex constraint might take the form:

$$P(\text{agree first, agree last} \mid \text{match}) \leq a, \tag{3.3}$$

for some $0 < a < 1$. Convex restrictions can be based on a priori knowledge of subspace regions in which modes of the likelihood yield good matching performance.

4. ASSIGNMENT

Jaro introduced a linear sum assignment procedure (lsap) to force 1-1 matching because he observed that greedy algorithms often made erroneous assignments. A greedy algorithm is one in which a record is always associated with the corresponding available record having the highest agreement weight. Subsequent records are only compared with available remaining records that have not been assigned. In the following, the two households are assumed to be the same, individuals have substantial identifying information, and the ordering is as shown.

```
HouseH1     HouseH2
husband
wife        wife
daughter    daughter
son         son
```

A new assignment algorithm (Winkler 1994a) reduces storage requirements by a much as 0.0002 (from 100 to 0.02 megabytes) with no loss in speed. Examples and additional details are given in the longer technical report.

5. RESULTS

Results are presented in two parts. The first section provides an overall comparison of matching methods that utilize various combinations of the new and old string comparators, the new and old assignment algorithms, and the generalized interaction weighting methods and independent weighting methods. The second provides results showing how accurately error rates can be estimated using the best matching methods from the first section. Error rates are compared with rates obtained via a method of Belin and Rubin (1995) that is known to work well in a narrow range of situations (Winkler and Thibaudeau 1991, Scheuren and Winkler 1993).

5.1. Overall Comparison of Matching Methods

For comparison purposes, results are produced using three pairs of files having known matching status. The baseline matching is done under 3-class, latent class models with interactions and under independence, respectively. The 3-class models are essentially the same ones used in Winkler (1992, 1993). The interactions are (1) 8-way between last name, first name, house number, street name, phone, age, relationship to head of household, and marital status, (2) 4-way between first name, house number, phone, and sex, and (3) 2-way between last name and race. The weights associated with interaction models are referred to as *generalized weights* and other weights obtained via independence models are referred to as *independent weights*. Results are reported for error rates of 0.002, 0.005, 0.01, and 0.02, respectively. *Link*, *Nonlink*, and *Clerical (or Possible Link)* are the computer designations, respectively. *Match* and *Nonmatch* are the true statuses, respectively. The baseline results (designated by *base)* are produced using the existing lsap algorithm and the previous string comparator but use the newer, 3-class EM procedures for parameter estimation (Winkler 1993). The results with the new string comparator (designated *s_c*) are produced with the existing string comparator replaced by the new one. The results with the new assignment algorithm (designated *as*) use both the new string comparator and the new assignment algorithm. For comparison, results produced using the previous string comparator but with the new assignment algorithm (designated by *os_l*) are also given.

Matching efficacy improves if more pairs can be designated as links and nonlinks at fixed error rate levels. In Tables 5.1-3, computer-designated links and clerical pairs are subdivided into (true) matches and nonmatches. Only the subset of pairs produced via 1-1 assignments are considered. In producing the tables, pairs are sorted by decreasing weights. The weights vary according to the different model assumptions and string comparators used. The number of pairs above different thresholds (i.e., *UPPER* of section 3) at different link error rates (0.002, 0.005, 0.01, and 0.02) are presented. False match error rates above 2 percent are not considered because the sets of pairs above the cutoff threshold *UPPER* contain virtually all of the true matches from the entire set of pairs when error rates rise to slightly less than 2 percent. In each line under the Interaction and Independent columns, the proportion of nonmatches (among the sum of all pairs in the Link and Clerical columns) is 2 percent.

The results generally show that the combination of generalized weighting with the new assignment algorithm performs slightly better than the baseline with independent weighting. In all of the best situations, error

Table 5.1 Match Results, Different Error Rates
1st Files, 4539 and 4859 records
38795 Pairs Agreeing on Block and
First Character of Last Name

	Interaction		Independent	
Link Error Rate	Link mat/nonm	Cler mat/non	Link mat/nonm	Cler mat/non
0.002				
base	3266/ 7	83/61	3172/ 6	242/64
s_c	2995/ 6	320/62	3176/ 6	236/64
as	3034/ 6	334/63	3176/ 6	234/64
os_l	3299/ 7	93/63	3174/ 6	242/64
0.005				
base	3312/17	37/51	3363/17	51/53
s_c	3239/17	76/51	3357/17	55/53
as	3282/17	86/52	3357/17	53/53
os_l	3354/17	38/52	3364/17	52/53
0.010				
base	3338/34	11/34	3401/34	13/36
s_c	3287/34	28/34	3396/34	16/36
as	3352/34	16/35	3396/34	14/36
os_l	3380/34	13/35	3402/34	14/36
0.020				
base	3349/68	0/ 0	3414/70	0/ 0
s_c	3315/68	0/ 0	3411/70	0/ 0
as	3368/69	0/ 0	3410/70	0/ 0
os_l	3393/69	0/ 0	3416/70	0/ 0

Table 5.2 Match Results, Different Error Rates
2nd Files, 5022 and 5212 records
37327 Pairs Agreeing on Block and
First Character of Last Name

	Interaction		Independent	
Link Error Rate	Link mat/nonm	Cler mat/non	Link mat/nonm	Cler mat/non
0.002				
base	3415/ 7	102/65	3475/ 7	63/65
s_c	3308/ 7	182/64	3414/ 7	127/65
as	3326/ 7	184/65	3414/ 7	127/65
os_l	3430/ 7	107/65	3477/ 7	63/65
0.005				
base	3493/18	24/54	3503/18	35/54
s_c	3349/17	41/54	3493/18	48/54
as	3484/18	26/54	3493/18	48/54
os_l	3511/18	26/54	3505/18	36/54
0.010				
base	3501/35	16/37	3525/36	13/36
s_c	3478/35	12/38	3526/36	15/36
as	3498/35	12/37	3526/36	15/36
os_l	3519/36	18/36	3527/36	14/36
0.020				
base	3517/72	0/ 0	3538/72	0/ 0
s_c	3490/71	0/ 0	3541/72	0/ 0
as	3510/72	0/ 0	3541/72	0/ 0
os_l	3537/72	0/ 0	3541/72	0/ 0

Table 5.3 Match Results, Different Error Rates
3rd Files, 15048 and 12072 Records
116305 Pairs Agreeing on Block and
First Character of Last Name

Link Error Rate	Interaction		Independent	
	Link mat/nonm	Cler mat/non	Link mat/nonm	Cler mat/non
0.002				
base	9519/19	287/181	9696/19	155/182
s_c	9462/19	338/181	9434/19	407/182
as	9418/19	410/182	9436/19	406/182
os_l	9695/19	151/182	9692/19	157/182
0.005				
base	9760/49	46/151	9792/49	59/152
s_c	9747/49	53/151	9781/49	60/152
as	9776/49	52/152	9783/49	57/152
os_l	9809/50	37/151	9791/49	58/152
0.010				
base	9784/99	22/101	9833/99	18/102
s_c	9774/99	16/101	9822/99	19/102
as	9803/99	25/102	9823/99	17/102
os_l	9828/99	18/102	9831/99	18/102
0.020				
base	9806/200	0/ 0	9851/201	0/ 0
s_c	9800/200	0/ 0	9841/201	0/ 0
as	9828/201	0/ 0	9842/201	0/ 0
os_l	9846/201	0/ 0	9849/201	0/ 0

levels are very low. The new string comparator produces worse results than the previous one (see e.g., Winkler 1990) and the new assignment algorithm (when combined with the new string comparator) performs slightly worse (between 0.1 and 0.01 percent) than the existing string comparator and lsap algorithm. In all situations (new or old string comparator, generalized or independent weighting), the new assignment algorithm slightly improves matching efficacy.

5.2. Estimation of Error Rates

Belin and Rubin (1995) introduced a method for estimating error rates that is known to work well in practice when the conditional independence assumption is reasonaably valid and matching is 1-1 (Winkler and Thibaudeau 1991, Scheuren and Winkler 1993). The method requires suitable calibration data and that the weighting curves corresponding to nonmatches and matches be well separated. The longer technical report introduces an alternate method that does not require calibration data and holds in a variety of situations for which the Belin-Rubin method does not converge. The basic idea is to begin with probabilities obtained for non-1-1 matching and adjust them to account (partially) for the effect of 1-1 assignment. Results are shown for generalized weights (Figures 1-6) and independent weights (Figures 7-12) for the same three pairs of files used in the previous section. In the comparisons, all matching methods use the previously existing string comparator and the new assignment algorithm.

Error rate estimates using the methods of this paper are compared with the method of Belin and Rubin (1994) via Figures 13-15 for independent weights and the distributions of nonmatches. With the independent weights of this paper, Belin-Rubin estimates are roughly as accurate as the independence estimates of this paper (Figures 10-12). To obtain the estimates in producing Figures 13-15, I modified Belin's software to yield estimates in a form consistent with the method of this paper. The current Belin-Rubin method is not intended to yield estimates for the distribution of matches and would not converge (even upon recalibration) with generalized weights.

6. DISCUSSION

This section provides discussion of the new string comparator and the methods of error rate estimation.

6.1. String Comparator

The new string comparator is primarily designed to assist on-line searches using last name, first name, or street name. In such situations, the new comparator is believed to be superior to the old (Lynch and Winkler 1994). The reason that the new comparator performs somewhat more poorly in matching situations is that error rates with the existing methods are very low and the redundancy of extra matching fields plays a more important role than single fields in isolation. Because the new string comparator often assigns slightly higher comparator values, a few isolated true nonmatches can receive slightly higher weighting scores and observed false match rates can increase above those obtained when the original string comparators were used.

Presently, since there are no suitable test decks for checking scanning errors (i.e., 'I' versus '1') and some types of keypunch errors (i.e., adjacent keys 'V' versus 'B'), there has been no empirical testing whether the associated adjustment for these types of errors helps.

6.2. Error Rate Estimation under the Belin-Rubin Model

The method of Belin and Rubin (1994) was designed for data situations similar to PES matching. In those situations, it performed very well (Winkler and Thibaudeau 1991). Because of the weighting adjustments that were used in PES matching, the shapes of curves of matches and nonmatches were somewhat different than the corresponding shapes of the curves under the independence model used in this paper. The Belin-Rubin method is not designed to work with non-1-1 matching, for situations in which the curves of matches and nonmatches are not very well separated, or for cases in which the shapes of curves are very different from those on which Belin and Rubin originally did their modelling.

The primary advantage of the Belin-Rubin method is in its conceptual simplicity and accuracy of the estimates in those situations for which it was designed. Belin and Rubin also obtain confidence intervals via the SEM algorithm. Because of the strong simplifying assumptions, the Belin-Rubin method can be subject to bias as Belin and Rubin showed in a large simulation experiment.

With data that is somewhat similar to the data of this paper and independence model weights, I have also observed bias similar to the bias that Belin and Rubin encountered in their simulation.

6.3. Error Rate Estimation under the Model of this Paper

Using non-1-1 matching, the general interaction model of this paper provided accurate decision rules and estimates of error rates with the three pairs of data files of the results sections plus two others. Estimates were relatively more accurate than the 1-1 adjusted estimates of this paper. An example is covered in Winkler (1993).

The reason that the generalized weighting model of this paper is useful is that it can be used in a variety of non-1-1 matching situations and, with adjustments like the one of this paper, can be used in 1-1 matching situations. Because the error-rate-estimation procedure of this paper uses more information, it also may be subject to less bias than the Belin-Rubin procedure. The bias of the error-rate-estimation procedures with a variety of different types of data is a topic of future research.

7. SUMMARY AND CONCLUSION

This paper describes enhancements to a record linkage methodology that employ string comparators for dealing with strings that do not agree character-by-character, an enhanced methodology for addressing differing, simultaneous agreements and disagreements between matching variables associated with pairs of records, and a new assignment algorithm for forcing 1-1 matching. Because of the interactions between the differing techniques, improving one method without accounting for how the method interacts with the others can actually reduce matching efficacy.

The results of this paper show that a sufficiently experienced practitioner can produce effective matching results and reasonably accurate estimates of error rates. I conclude that considerably more research is needed before the techniques can be used by naive practitioners on a large variety of administrative lists. The difficulties have the flavor of early regression analysis for which techniques for dealing with outliers, colinearity, and other problems had not been developed. The techniques, however, can be used with a narrow range of high-quality lists such as those for evaluating Census undercount that have known matching characteristics.

*The views expressed are attributable to the author and do not necessarily reflect those of the Bureau the Census. A longer version of this paper is available by request.

REFERENCES

Belin, T. R., and Rubin, D. B. (1995), "A Method of Calibrating False-Match Rates in Record Linkage," *Journal of the Amer. Statistical Assn.*, to appear.

Budzinsky, C. D. (1991), "Automated Spelling Correction," Statistics Canada.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. of the Royal Stat. Soc.*, **B**, **39**, 1-38.

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

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **89**, 414-420.

Lynch, M. P., and Winkler, W. E. (1994), "Improved String Comparator," technical report, Statistical Res. Div., Washington, DC: U.S. Bureau of the Census.

McLaughlin, G. (1993), Private communication of C-string-comparison routine.

Meng, X., and Rubin, D. B. (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, **80**, 267-278.

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

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

Nuyens, C. (1993), "Generalized Record Linkage at Statistics Canada," *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, 926-930.

Scheuren, F., and Winkler, W. E. (1993), "Regression Analysis of Data Files that are Computer Matched," *Survey Methodology*, **19**, 39-58.

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

Winkler, W. E. (1994a), "Improved Matching via a New Assignment Algorithm," technical report, Statistical Res. Div., Washington, DC: U.S. Bureau of the Census.

Winkler, W. E. (1994b), "Improved Parameter Estimation in Record Linkage," tech. report, Statistical Res. Div., Washington, DC: U.S. Bureau of the Census.

Winkler, W. E. (1994c), "Matching and Record Linkage," in B. G. Cox (ed.) *Survey Methods for Businesses, Farms, and Institutions*, New York: J. Wiley, to appear.

Winkler, W. E., and Thibaudeau, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census," Statistical Res. Div. Report 91/09, Washington, DC: U.S. Bureau of the Census.

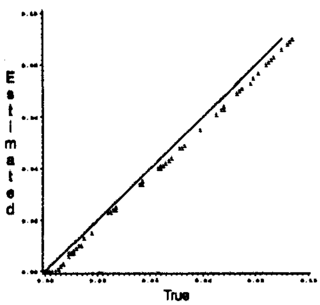

Figure 1. Estimates vs Truth
Cumulative Distribution of Matches
1st Files, Interaction EM, 1–1

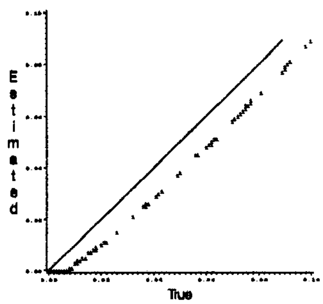

Figure 2. Estimates vs Truth
Cumulative Distribution of Matches
2nd Files, Interaction EM, 1–1

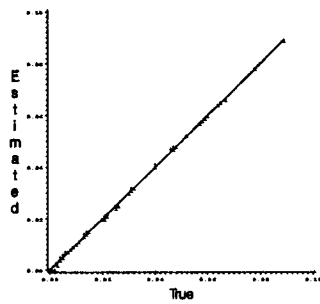

Figure 3. Estimates vs Truth
Cumulative Distribution of Matches
3rd Files, Interaction EM, 1–1

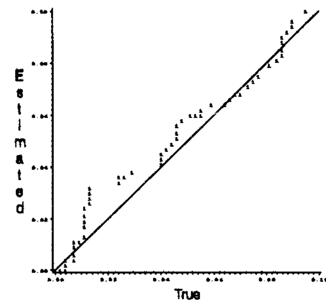

Figure 4. Estimates vs Truth
Cumulative Distribution of Nonmatches
1st Files, Interaction EM, 1–1

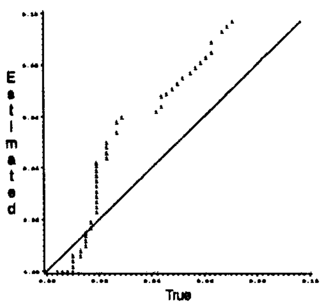

Figure 5. Estimates vs Truth
Cumulative Distribution of Nonmatches
2nd Files, Interaction EM, 1–1

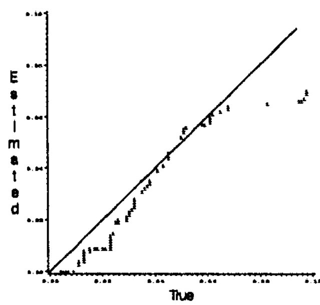

Figure 6. Estimates vs Truth
Cumulative Distribution of Nonmatches
3rd Files, Interaction EM, 1–1

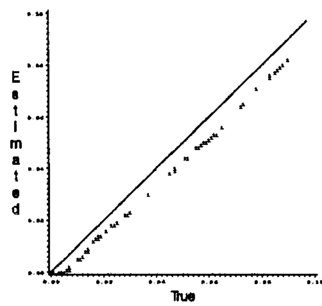

Figure 7. Estimates vs Truth
Cumulative Distribution of Matches
1st Files, Independent EM, 1–1

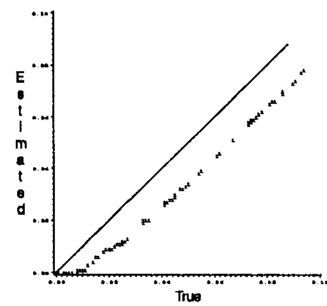

Figure 8. Estimates vs Truth
Cumulative Distribution of Matches
2nd Files, Independent EM, 1–1

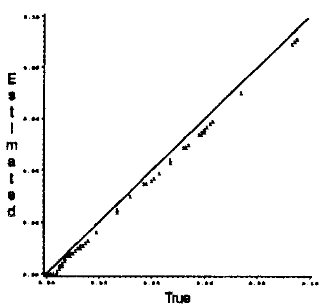

Figure 9. Estimates vs Truth
Cumulative Distribution of Matches
3rd Files, Independent EM, 1–1

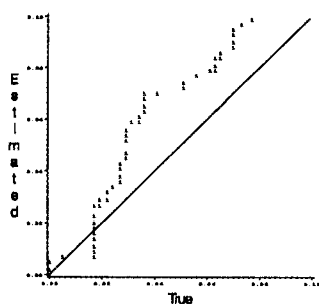

Figure 10. Estimates vs Truth
Cumulative Distribution of Nonmatches
1st Files, Independent EM, 1–1

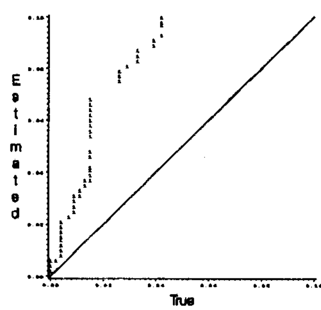

Figure 11. Estimates vs Truth
Cumulative Distribution of Nonmatches
2nd Files, Independent EM, 1–1

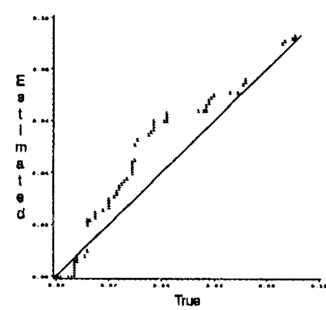

Figure 12. Estimates vs Truth
Cumulative Distribution of Nonmatches
3rd Files, Independent EM, 1–1

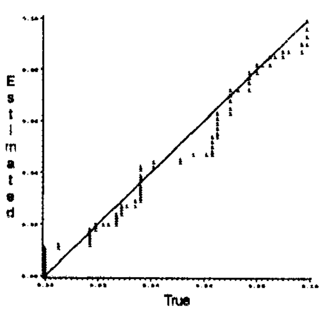

Figure 13. Estimates vs Truth
Cumulative Distribution of Nonmatches
1st Files, Independent EM, 1–1, TB

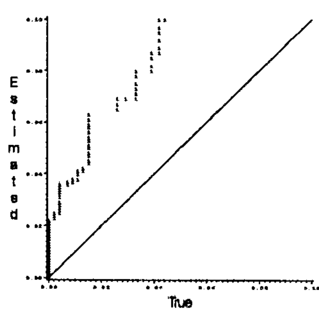

Figure 14. Estimates vs Truth
Cumulative Distribution of Nonmatches
2nd Files, Independent EM, 1–1, TB

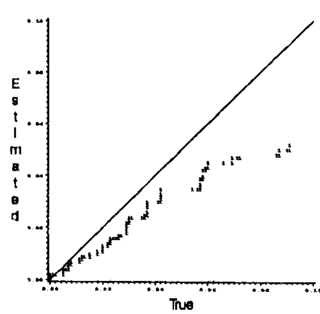

Figure 15. Estimates vs Truth
Cumulative Distribution of Nonmatches
3rd Files, Independent EM, 1–1, TB