EXACT MATCHING USING ELEMENTARY TECHNIQUES

William E. Winkler, Energy Information Administration

1. INTRODUCTION

The purpose of this paper is to illustrate problems in identifying duplicates in name and address files. The illustration takes the form of a report on work in progress on developing methods that are readily applicable to many address lists consisting primarily of companies or establishments. The specific examples are taken from lists in use at the Energy Information Administration (EIA).

In developing matching methods, we wish to minimize Type I and Type II error rates. A Type I error is a false match identified as a potential match by software, and a Type II error is a duplicate that is unmatched.

This paper presents a methodology for determining and comparing error rates when various matching strategies are applied to files in which duplicates are identified and contain the corresponding active records' control numbers. It also presents a methodology for determining error rates based on samples. The sample related methodology is not new, but can be non-trivially deduced from cluster sampling techniques (see e.g., Cochran (1977)) and from techniques for estimating rates of duplicates (Deming and Glasser (1959)). The techniques of Deming and Glasser are widely applicable, but do not appear to be widely known.

The matching methods presented can be easily implemented on files containing less than 100,000 records because they require minimal programming expertise, no sophisticated methods of file structuring or sort mechanisms, and no historical knowledge of formatting conventions or the probability of a match given that various subportions of fields take certain values and agree.

2. BACKGROUND

2.1. Why Procedures are Needed

EIA has a continuing need for match/merge software because survey frames must be periodically updated using lists from State and commercial sources. The software must be sufficiently flexible to assure that it can be readily applied to a variety of surveys. The modifications in the match/merge procedures should be in the strategy of application rather than in modifications to the basic programs.

As there appears to be no common terminology connected with some of the concepts of association, we introduce some here. A duplicate record is a redundant record having both a similar name and similar street address to the active record that is maintained for mailing purposes. An associate record is a redundant record having either a different name and/or street address from the cand/corresponding active record that is regarded as its parent. The parent (or headquarters) record is the active record that is maintained for mailing purposes.

Those retained redundant records that are connected to parent records are referred to as dispensable records. Although dispensable records are not necessary for sampling or estimation, they are necessary for the frame maintenance and updating. To avoid redoing work during updates, each identified duplicate and associate record should be maintained in the master frame file and contain its respective parent's control number.

For this paper, the set of associates includes nonreporting subsidiaries of parents, predecessors of active firms, and nonreporting affiliates of active firms.

2.3. What is Presented in the Paper

The remainder of this paper shows how a new match/merge strategy was developed. Section 3 contains the procedures and describes data bases used for the evaluation.

The main results of applying the newly developed match/merge strategy to the constructed empirical data base are compared with the results from applying a previously existing strategy in section 4. Section 5 contains results from applying the new strategy to two special frames. It was used in identifying potential duplicates within the EIA-23 Oil and Gas Well Operators frame and matches in the EIA-7A Coal Production frame with a comparable file from the Mine Safety and Health Administration (MSHA).

Results from applying a discriminant analysis procedure to the first empirical data base are presented in section 6. The final section contains some conclusions and some ideas for future work.

3. METHODS USED TO DEVELOP NEW STRATEGIES

In order to develop a more effective match/merge strategy it was necessary to:

1. construct a suitable empirical data base for refining procedures,
2. define evaluation criteria,
3. refine procedures, and
4. evaluate procedures on additional data bases.

3.1. Creation of a Suitable Empirical Data Base

The basic empirical data base containing 66,414 records was constructed by removing easily identified duplicates from a set of 176,000 records obtained from 11 EIA and 47 State and industry lists of sellers of petroleum products. Of the 66,414 records, 3,091 are duplicates, 8,456 are associates, and 54,867 are headquarters records. The final set of duplicates were identified through both computer-assisted and manual procedures while the set of associates were identified through callbacks or surveying.

3.2. Criteria for Evaluation

3.2.1. Type I and II errors

As unmatched duplicates are considerably more difficult to identify than false matches, the primary emphasis in developing a new strategy was minimizing Type II errors before minimizing Type I errors.

It is important to note that if a file has no unmatched duplicate records, then any match/merge strategy applied will either yield no potential pairs or a Type I error rate of 100 percent and a Type II error rate of 0 percent. Because the basic empirical data base is relatively free of duplicates - as a result of reducing it from 176,000 to 66,000 records - application of any match/merge strategy will produce relatively high Type I error rates.

3.2.2. Rate of unmatched dispensable records

The number of unmatched dispensables as a percentage of the total number of records in a file is also an important evaluation criteria. We define the rate of unmatched dispensable records as Q/(X+Q)*100 where Q is either the number U of unmatched duplicate records or the number A of unmatched associate records and X is the number of parent records.

This additional evaluation criteria is important because the Type II error rate criteria will not provide a measure of how free of duplicates and associates a file is. The Type II error rate does not work well because, as the number of duplicates D in a file decreases, the Type II error rate (U/D*100 where U is the number of unmatched duplicates) will necessarily increase.

In the analysis of the empirical data base, D is held constant so that the comparative advantages of various strategies can be assessed using Type II error rates. The rate of unmatched dispensable records will not work well for these comparative evaluations because it is too dependent on the number of parent records X which does not change. That is, if U1 and U2 are the numbers of unmatched duplicates under two matching strategies and $U1 < U2 << X$, then U1/(U1+X) and U2/(U2+X) are approximately equal.

3.3. Evaluation of Selected Methods Using Other Data Bases

The most successful of the matching strategies developed may be dependent on the basic empirical data base. To evaluate how widely applicable the strategies are, they must be applied to other data bases.

3.3.1. Data bases selected

The data bases selected for additional work are the EIA-23 Oil and Gas Well Operators frame and the EIA-7A Coal Production frame.

The EIA-23 frame, which contains 21,637 records identified as active unduplicated entities, is used for identifying duplicates within a list. As it contains a large number of records associated with partnerships, it presents a new difficulty from those encountered in the empirical data base which generally did not contain partnerships. As there is no complete identification of duplicates, Type I and Type II error rates corresponding to duplicates must be computed based on samples.

The EIA-7A frame is a particularly useful data base because many of its records can be connected to records in a list from the Mine Safety and Health Administration (MSHA) using the MSHA ID. There are 3,262 pairs of records from the two files that are separated into duplicates and associates and that can be used in determining the Type I and Type II error rates for various matching strategies. The EIA-7A/MSHA comparison yields a new difficulty because the EIA-7A is a county-level frame of operators of facilities associated with surface mining, underground mining, or preparation plants, while the MSHA list does not always distinguish between types of mining operations, may aggregate some mining operations across counties, and may list owners instead of operators.

3.3.2. Determination of Type II error rates using samples

Type II error rates are determined via manual review of all records in a sample of three-digit ZIP codes (or any other suitable identifier) in files in which duplicates have not been previously identified. The set of identified duplicates is divided into those that would be identified by the matching criteria and those that would not. Estimators of the rate and its variance can be determined by the usual formulas from cluster sampling (see e.g., Cochran (1977), sec. 11.12). For more details, see Winkler (1984).

3.3.3. Determination of Type I error rates using samples

Type I error rates are determined via manual review of a simple random sample of potential matches in those files in which duplicates have not been previously identified. Use of results based on samples necessarily yields confidence intervals for the true parameter (see Winkler (1984) for theoretical details).

3.4. Discriminant Analysis

To determine the relative value of individual fields such as name, street address, city, state, ZIP code, and telephone in determining likely true and false matches, a discriminant analysis procedure was tried. The input data consisted of the numeric values corresponding to the largest matching character strings starting from the first character for each field associated with pairs of records matched by software.

As the actual truth or falsehood of each match was known, random samples could be drawn and used for calibrating the discrimination procedures.

4. RESULTS USING THE BASIC EMPIRICAL DATA BASE

The results in this section are based on applying three different sets of matching criteria that were developed as the understanding of matching strategies improved. The current set of matching criteria (given below) were developed using the empirical data base. Prior to use in the matching program, each address file has most punctuation deleted and the spelling of words such as STREET, NORTH, P O BOX, etc. standardized.

Basic Set of Matching Criteria

(1) 3 characters ZIP, 4 characters name
(2) 5 characters ZIP, 6 characters street address
(3) 10 digits telephone
(4) Sort name field into words of decreasing length and then match using 1.

An early set of 12 criteria were developed on an ad hoc basis and used in creating the empirical data base (they are described fully in Winkler (1984), but are not necessary for

this discussion). An intermediate set of seven criteria, which were developed using the empirical data base, include the four criteria in the basic set and (5) 15 characters of name, (6) word length sort and then match using criteria (5), and (7) word length sort of address and then match using criteria (2) of the basic set.

Table 1 shows the improvement in matching efficiency with each of the successive sets of criteria. The Type II error rate of approximately 17 percent with the early set of criteria is reduced to approximately 1 percent with the intermediate and later sets of criteria. Manual processing (as indicated by the number of potential matches) is reduced from 39,000 to 34,000 to 12,000 with the successive sets of criteria, while Type I error rates remained relatively constant.

5. APPLICATION TO OTHER DATA BASES FOR VERIFICATION

5.1. EIA-23 Oil and Gas Well Operators Frame

A preliminary review of the EIA-23 list and review of early outputs indicated that the basic set of matching criteria could be modified to lower Type I error rates with little increase in Type II error rates. The matching criteria used consisted solely of (1) 3 digits ZIP and 4 characters name and (2) 5 digits ZIP and 8 characters street address. For consistency with earlier results criteria (3) 15 characters name and (4) 10 digits phone were are also used.

The word length sort was dropped because the name and address fields in the EIA-23 file are consistently formatted. Matching using 15 characters of the name or the phone number were dropped because review of random samples of records matched using these criteria would have identified duplicates that all would have been identified by criterias (1) and (2) above. The number of characters in the street address was increased to eight because use of only six characters yielded more false matches with no apparent decrease in the number of unmatched duplicates.

Results are preliminary because the identification of true matches within the set of potential matches and of true duplicates within the set of records in three-digit ZIP codes have not been verified independently. The independent verification will be part of the normal EIA-23 frame updating which takes place in the fall.

5.1.1. Type II error rates

Review of a cluster sample of 1885 records from 19 three-digit ZIP codes yielded 29 duplicates, 1 of which would not be identified by criterias (1) and (2) or criterias (1), (2), (3), and (4) above. Thus, the Type II error rate is 0.034 (1/29) with standard deviation of 0.024 and the 95 percent confidence interval is (0,0.082) (see Winkler (1984) for details).

5.1.2. Type I error rates

The overall rate of false matches using criteria (1) and (2) is estimated at 90.2 percent (6,902 of 7,650 potential duplicates) with coefficient of variation of 1.83 percent. The 95 percent confidence interval for the true percentage of Type I errors is (86.9,93.5).

The overall rate of false matches using criteria (1) and (2) plus criteria (3) 10 characters name and (4) 10 digits phone is estimated at 91.2 percent (8,604 of 9.438 potential duplicates) with coefficient of variation of 1.49 percent. The 95 percent confidence interval for the true percentage of Type I errors is (88.4,93.9).

5.2. EIA-7A and MSHA Coal Operators Frames

A preliminary review of the outputs from matching the EIA-7A frame with the MSHA frame indicated that the matching criteria could be modified to lower Type I error rates with little increase in Type II error rates. The basic matching criteria used consisted solely of (1) three-digit county code and 4 characters name and (2) three-digit county code and 6 characters street address.

The word length sort combined with criteria (1) and (2) was also used to identify additional duplicates. Matching using company name only was not used because the large number of mines associated with large companies yielded excessively high Type I error rates. Matching using the phone number was not used because the MSHA file does not generally contain phone numbers.

Two comparisons were performed by matching two independent random samples of 1,000 records with the corresponding 3,262 records in the MSHA frame. Two samples instead of all 3,262 corresponding MSHA records were used to determine how much variation might occur if samples instead of entire files are used. Table 2 shows that little variation occurs.

5.2.1. Overall results

The overall results (Table 2) show that, in the first sample, the first set of criteria identify 75.83 and 48.47 percent of duplicates and associates, respectively; the second set identify 81.57 and 51.29 percent of duplicates and associates, respectively. In the second sample, the first set of criteria identify 77.50 and 49.88 percent of duplicates and associates, respectively; the second set identify 82.25 and 52.44 percent of duplicates and associates, respectively.

5.2.2. Type II error rates

From Table 2, we obtain that the Type II error rate for matching duplicates in the first sample using Criteria (1) and (2) and Criteria (1) and (2) with the word length sort are 5.57 and 3.30 percent respectively; for associates, 29.88 and 15.06, respectively. The Type II error rate for matching duplicates in the second sample using Criteria (1) and (2) and Criteria (1) and (2) with the word length sort are 5.10 and 2.99 percent, respectively; for associates, 31.09 and 15.55, respectively.

5.2.3. Type I error rates

Independently applying Criteria (1) and (2) and Criteria (1) and (2) plus the word length sort to the first sample yields 841 and 1,581 potential duplicates and Type I error rates of 23.9 and 32.3 percent, respectively. Independently applying the two sets of criteria to the second sample yields 859 and 1,618 potential duplicates and Type I error rates of 26.2 and 34.5 percent, respectively. These Type I error rates are necessarily biased low because they are based on comparing MSHA records with the subset of EIA-7A records having MSHA ID's, not the entire set of EIA-7A records.

5.2.4. Rate of unmatched dispensable records

For the first sample, the first set of criteria yields (Table 2) that the rate of unmatched duplicate records is 5.27 percent and the rate of unmatched associate records is 23.01 percent; the second set of criteria, 3.20 and 13.09 percent for duplicates and associates, respectively. For the second sample, the first set of criteria yields (Table 2) that the rate of unmatched duplicate records is 4.85 percent and the rate of unmatched associate records is 23.01 percent; the second set of criteria, 2.90 and 13.19 percent for duplicates and associates, respectively.

6. RESULTS OF DISCRIMINANT ANALYSIS

The basic results were that discriminant analysis procedures could not accurately delineate likely true and false matches and that numeric results varied substantially according to the different random samples used for calibration and evaluation.

Many false matches characterized as true (with "probability" greater than .90) had the form:

Smith Oil	114 Main St	Houston
77001	713/456-9986	
Jones Fuel	114 Main St	Houston
77001	713/456-9986	

Many true matches characterized as false (with "probability" greater than .90) had the form:

Solas Oil	1114 Main St	Pittsburgh
15134	412/763-1186	
Salas Fuel	114 N Main St	Greensburg
15201	412/763-1186	

The first example represents a pair of retail establishments, one having gone out of business and the other occupying the location of the previous establishment. The second example represents the situation in which typographical differences occur in multiple fields.

Tables 3 and 4 show the results of a discriminant analysis procedure based on independent samples. Results vary substantially because samples of size 500 used in calibrating the discriminant analysis procedures were not sufficiently large to yield stable covariance matrices. As there are only 3900 potential matches of duplicates, taking substantially larger samples is not possible.

7. CONCLUSIONS AND FUTURE WORK

The results from applying the match/merge software to the basic empirical data base, the EIA-23 frame, and the EIA-7A/MSHA files show that the matching strategies are straightforward to apply and can be easily modified. Because the modifications are in the strategy of application rather than in the code, match/merge strategies can be refined by statisticians or survey managers with minimal training.

7.1. Problems Remaining

7.1.1. Weighting methodologies

Although the discriminant analysis procedure (section 6) did not work well with numeric data obtained from examining entire fields, it may perform better if subportions of fields (such as initials, surname, street name, or street number) are used.

The Department of Agriculture, the Bureau of the Census, and Statistics Canada have all developed and implemented methods for using the information available in subdivided fields. The weighting methodologies, however, have used approaches based on the theory of Fellegi and Sunter (1969) (e.g., Agriculture, Census (under investigation)) and ad hoc methods (Statistics Canada and Census) that are different from the discriminant analysis approach.

7.1.2. Search mechanisms

Current software compares every address record with every record in the spelling standardization tables. If a suitable distance measure could be developed, then methods for searching subportions of files (Friedman, Bentley, and Finkel (1977)) could be used for substantially improving search times. At present, improving search times is a low priority because the standardization program (which requires between 10 and 60 minutes CPU time) only needs to be run once.

7.1.3. Capture/Recapture

Various authors (e.g., Scheuren (1983)) have suggested using capture/recapture methods for evaluating rates of unmatched duplicates. This paper (see also Winkler (1984)) shows that estimators of the rates that are based on samples can have high coefficients of variation. As formulas for estimating the effect of population (or list) overlap on population totals requires exact counts of population overlap (see e.g., Bishop, Fienberg, and Holland (1975), Chapter 6), additional variance in estimates of population totals will generally be induced by the use of estimators of population overlap.

Even if statisticians know the error rates associated with match/merge software and with manual followup, providing accurate estimates of population totals (or undercounts) will be difficult.

7.2. Summary

Organizations such as the Department of Agriculture, the Bureau of the Census, and Statistics Canada have developed and refined frame maintenance techniques. Such refinements can take the form of developing probabilities (weights) for successful matches based on followup of tens of thousands of records over a period of years, sophisticated data base structuring that allows faster access, and improved search algorithms that take advantage of both the data base structure and new input/output methods available on some computers.

EIA has many fuel-specific frames of relatively small size (generally less than 10,000 records). The construction of a vocabulary based weighting scheme for identifying true duplicates in each survey frame does not seem to be cost-effective; nor does the development of sophisticated data base structures or search algorithms which may require more maintenance than sequential files and sequential searches.

Given that the match/merge strategies presented in this paper require little programming expertise or maintenance and can be relatively quickly applied to standard

sequential data files, we believe the most cost-effective approach to frame maintenance in relatively small files is to reduce the rate of false matches within the structure of our existing matching strategies.

BIBLIOGRAPHY

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), "Discrete Multivariate Analysis," MIT Press, Cambridge, MA.

Bourne, C. P., and Ford, D. J. (1961), "A Study of Methods for Systematically Abbreviating English Words and Names," J. ACM 8, 538-552.

Cochran, W. G. (1977), Sampling Techniques (3rd edition), Wiley: New York.

Deming, W. E., and Glasser, G. J. (1959), "On the Problem of Matching Lists by Samples," JASA 30, 403-415.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," JASA 40, 1183-1210.

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977), "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. Math. Software 3, 209-226

Hall, P. A. V. and Dowling, G. R. (1980), "Approximate String Matching," Computing Surveys 12, 381-402.

Kruskal, J. B. (1983), "An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules," SIAM Review 25, 201-237.

Scheuren, F. (1983), "Design and Estimation for Large Federal Surveys Using Administrative Records," ASA Proceedings of the Section on Survey Research Methods, 377-381.

Statistics Canada, "Record Linkage Software."

U. S. Department of Agriculture, "List Frame Development: Procedures and Software."

U. S. Department of Commerce, Bureau of the Census/Agriculture Division, "Record Linkage for Development of the 1978 Census of Agriculture Mailing List."

Winkler, W. E., (1984), "Issues in Developing Frame Matching Procedures," Presented to the ASA Committee on Energy Statistics in April 1984.

Table 1: Results of Matching 3,091 Duplicates and 8,456 Associates with 54,867 Parents (Rates are percentages)

	Early Set of Matching Criteria	Intermed Set of Matching Criteria	Current Set of Matching Criteria
Type II Error Rate			
Duplicates	16.5	0.8	1.4
Associates	27.2	11.1	27.3
Dispensables	24.5	8.3	20.4
Type I Errors For Dispensables			
Potential Matches	39,000	34,000	12,000
Error Rate	44	37	41
For Duplicates			
Potential Matches	NA	12,000	3,100
Error Rate	NA	17	30
Rate of Unmatched			
Duplicates	0.86	0.05	0.08
Associates	3.97	1.68	4.04
Dispensables	4.78	1.72	4.11

NA- Not available

Table 2: Comparison of Matching Criteria Using EIA-7A and Two Samples of Size 1000 from the MSHA List 1/

Sample	Criteria	Status Code 2/	Match with Correct Parent	Match with Wrong Parent	Not Matched 3/	True Number of Matches
First	(1)+(2)	DD	436 (75.83)	107 (18.61)	32 (5.57)	575 (100.0)
		non-DD	206 (48.47)	92 (21.65)	127 (29.88)	425 (100.0)
		All	642 (64.20)	199 (19.90)	159 (15.90)	1000 (100.0)
	(1)+(2) plus word-length sort	DD	469 (81.57)	87 (15.13)	19 (3.30)	575 (100.0)
		non-DD	218 (51.29)	143 (33.65)	64 (15.06)	425 (100.0)
		All	687 (68.70)	230 (23.00)	83 (8.30)	1000 (100.0)
Second	(1)+(2)	DD	441 (77.50)	99 (17.40)	29 (5.10)	569 (100.0)
		non-DD	215 (49.88)	82 (19.03)	134 (31.09)	441 (100.0)
		All	656 (65.60)	181 (18.10)	163 (16.30)	1000 (100.0)
	(1)+(2) plus word-length sort	DD	468 (82.25)	84 (14.76)	17 (2.99)	569 (100.0)
		non-DD	226 (52.44)	138 (32.02)	67 (15.55)	441 (100.0)
		All	694 (69.40)	222 (22.20)	84 (8.40)	1000 (100.0)

1/ The numbers in parentheses are percentages of row totals.
2/ 'DD' means a duplicate having a similar name and a similar address. 'Non-DD' means an associate.
3/ Type II error rates are in parentheses.

Table 3: Summary of Misclassified Records Seed 69581, Sample Containing 118 False and 382 True Matches

Threshold Level	Matches		
	Misslassified		Not Classified
	True	False	
0.5	28	30	0
0.65	18	21	58
0.7	15	15	80
0.75	14	14	101
0.8	12	11	125
0.85	10	10	150
0.90	9	8	185
0.95	9	4	244

Table 4: Summary of Misclassified Records Seed 84429, Sample Containing 135 False and 365 True Matches

Threshold Level	Matches		
	Misslassified		Not Classified
	True	False	
0.5	22	44	0
0.65	14	36	41
0.7	11	34	57
0.75	10	30	73
0.8	9	22	104
0.85	9	18	131
0.90	8	16	160
0.95	5	11	210