AN APPLICATION OF A THEORY FOR RECORD LINKAGE

Richard W. Coulter, Department of Agriculture

I. INTRODUCTION

As part of the effort by the Statistical Reporting Service to build a master list sampling frame of farms in each State, a record linkage system is being developed for use in detecting duplication in a list. To build this master, lists from several sources are combined and duplication, both between and within the lists, is removed. In selecting a linkage technique, an important consideration was the paucity of identifying data on most records. The table below illustrates the information available for one fairly typical State.

As the table indicates, only given name, surname, and place name are guaranteed to be present. Address information for the rural population is scarce and most often is only a rural route number. The presence of identifier numbers is rare. It is estimated that in making comparisons, nearly 60 percent of the comparison pairs will have no information in addition to given name, surname, place name, and possibly route number. In an attempt to best use this limited information in linkage, a probability model is used which incorporates some of the concepts developed by Ivan Fellegi and Alan Sunter [1]. A number of modifications and extensions have been made to portions of the original theory. (See [3].) Some of these will be examined in the following. Prior to this some background information on the model is necessary.

Let L_A be the set of records, $\alpha(a)$, pertaining to the population A, with elements $a_i \varepsilon A$, under consideration.

Define $M = \{(a_i, a_j);\ a_i = a_j,\ i < j\}$

$U = \{(a_i, a_j);\ a_i \neq a_j,\ i < j\}$

as the matched and unmatched sets, respectively. Denote by $\gamma = (\gamma^k)$ the coded result of the comparison of the variables in the comparison pair $\left[\alpha(a_i\ ),\ \alpha(a_j)\right]$ where the result of the comparison on the k^{th} component is denoted by γ^k.

The comparison space can be defined as the set of all realizations of γ generated as a result of the comparison of records associated with members of M or U. Two probabilities are estimated for each γ^k.

$$1.\quad m(\gamma^k) = P\{\gamma^k\left[\alpha(a_i), \alpha(a_j)\right];\ (a_i,\ a_j)\ \varepsilon\ M\}$$

$$2.\quad u(\gamma^k) = P\{\gamma^k\left[\alpha(a_i), \alpha(a_j)\right];\ (a_i,\ a_j)\ \varepsilon\ U\}$$

A component weight for each γ^k is defined by:

$$w(\gamma^k) = \log_{10}\left[m(\gamma^k)\ /\ u(\gamma^k)\right].$$

The component weights for those variables compared are then summed to yield a total weight, $w(\gamma)$, for each comparison pair.

Two threshold values are calculated to which the total weight is compared. If the total weight is less than the lower threshold, then the pair is classified as a nonlink. If the total weight is larger than the upper threshold, then the pair is classified as a link. Pairs with total weight between the two thresholds are classified as possible links.

As an illustration of this general technique, the specific calculations for surname - surname code will be examined. In addition, the manner in which several other variables are used will be briefly described. Since the same general technique is used for these, the specific

Table A.--Availability of Identifying Data

Variable	% Presence in File
Prefix	3 (82% of these are 'MR')
Given Name	100 (24% of these are an initial only)
Middle Name	52 (90% of these are an initial only)
Surname	100
Rural Route	76 (43% of these are 'RT 1')
Box Number	43
House Number	5
Street Name	8
Place Name	100
Social Security Number	0
Employer Identification Number	2
Telephone	4

computations (some of which are rather lengthy) will not be given at this time.

II. USE OF SURNAME - SURNAME CODE AS A MATCHING VARIABLE

Surname and surname code are used as a joint variable in the linkage model. (See [7].) When surnames agree, the appropriate weight is assigned and surname code is not considered. However, when surnames disagree, then surname codes are compared. Depending upon this outcome, the appropriate weight is assigned. Under the present blocking scheme, surname codes must agree and, thus, the weight assigned when surnames disagree will always be the weight for agreement on the particular surname code. The manner in which weights are calculated for this variable is described below.

A. Notation

Let, $X = \{x_j,\ j = 1,2,\dots,n\}$ represent the set of all possible realizations of surnames in the file;

$Y = \{y_k,\ k = 1,2,\dots,n'\}$ represent the set of all possible realizations of surname codes on the file;

$Y' = \{y_d,\ d = 1.2,\dots,n''\}$ represent the subset of Y that consists of surname codes associated with more than one surname;

$f_{x_1}, f_{x_2}, \dots, f_{x_n}$ denote the frequencies of the surname realizations;

$$\sum^{n} f_{x_j} = N$$

$f_{y_1}, f_{y_2}, \dots, f_{y_n}$ denote the frequencies of the surname realizations;

$$\sum^{n'} f_{y_k} = N,\ \sum^{n''} f_{y_d} = N'$$

e = P (surname in error in the file of records associated with the matched set);

e_T = P (error-free forms of the surnames in a pair associated with the matched set are different);

g_1 = P (a surname in error in a pair associated with the matched set receives the same code as the correct surname);

g_2 = P (a valid change in surname occurs in matched records and both receive the same surname code);

$m(\gamma_h) = P(\gamma_h$ | the pair represents records from M), h = 1,2,3; and

$u(\gamma_h) = P(\gamma_h$ | the pair represents records from U), h = 1,2,3;

where, γ_1 denotes agreement on surname,

γ_2 denotes agreement on surname code and disagreement on surname, and

γ_3 denotes disagreement on both surname and surname code.

B. Assumptions

1. The distribution of matching surnames (surname codes) in the matched set is the same as the distribution in the file.

2. The distribution of surnames (surname codes) in the unmatched set is the same as the distribution in the file.

3. The g_1 and g_2 probabilities are independent of surname code.

C. Calculations (for surname x_j and surname code y_d)

$$m\left[\gamma_1\ (x_j)\right] = (f_{x_j}/N)\ (1 - e)^2\ (1 - e_T)$$

$$u\left[\gamma_1\ (x_j)\right] = (f_{x_j}/N)^2$$

$$m\left[\gamma_2\ (y_d)\right] = (f_{y_d}/N')\ \Big[2g_1\ e(1 - e)(1 - e_T) + g_1^2\ e^2(1 - e_T) + g_2\ (1 - e)^2 \cdot e_T + 2g_1\ g_2\ e(1 - e)\ e_T + g_1^2\ g_2\ e^2\ e_T\Big]$$

$$u\left[\gamma_2\ (y_d)\right] = u(\text{agree on sn code}) \cdot u(\text{disagree on sn} \mid \text{agree on sn code}) = u(\text{agree on sn code}) \cdot \Big[1 - u(\text{agree on sn} \mid \text{agree on sn code})\Big] = (1/N^2)\left[f_{y_d}^{\ 2} - \sum_{j=1}^{n''_d} f_{x_j}^{\ 2}\right],$$

where n''_d = the number of surnames with surname code y_d

$$m(\gamma_3) = 2(1 - g_1)\ e(1 - e)(1 - e_T) + (1 - g_1^{\ 2})\ e^2(1 - e_T) + (1 - g_2)(1 - e)^2\ e_T + 2(i - g_1g_2)e\ (1 - e)e_T + (1 - g_1^{\ 2}g_2)\ e^2e_T$$

$$u(\gamma_3) = 1 - \sum_{k=1}^{n'} (f_{y_k}/N)^2$$

$$\text{weight} = w(\gamma_h) = \log_{10}\left[m(\gamma_h)/u(\gamma_h)\right], \quad h = 1,2,3$$

Under the present blocking scheme, surname code is used as the first blocking factor and, thus, γ_3 does not occur; i.e., $m(\gamma_3)$ and $u(\gamma_3)$ are both zero. To fit the supplied probabilities to the actual situation, the probabilities for both m and u should be redistributed over γ_1 and γ_2.

For h = 1,2 the revised probability functions would be:

$$m(\gamma_h)' = m(\gamma_h \mid \gamma_3 \text{ does not occur})$$
$$= m(\gamma_h) / \left[1 - m(\gamma_3)\right]$$
$$u(\gamma_h)' = u(\gamma_h \mid \gamma_3 \text{ does not occur})$$
$$= u(\gamma_h) / \left[1 - u(\gamma_3)\right].$$

Since most of the probability for the unmatched set will be concentrated in γ_3, the net effect of this redistribution would be a significant reduction in the derived weights for exact matches on surname and surname code. For this reason, we have chosen to ignore this effect of blocking for weight calculation purposes. For example, in a test file of 150,000 records, a surname which occurs 1,000 times receives a weight for agreement of 2.16. The revised weight using the redistributed probabilities would be -.51.

The weight for γ_1 depends primarily on the frequency of the particular surname, with the more rare surnames receiving the larger weights. The weight for γ_2 depends on the frequency of the surname code, on the size of the error rates e and e_T and on the number of distinct surnames within that codes. Infrequent surname codes, large error rates and few different surnames all tend to make the weight for this condition large.

III. OTHER VARIABLES

Modifications have been made to other variables in an attempt to improve the linkage results. These will be outlined below.

A. Given Name - First Name

As part of the processing prior to linkage, each given name on the file is assigned a formal or first name. (See [8].) A dictionary of the most common given name is utilized for this purpose. For given names not in the dictionary, the given name will also serve as the first name. Common examples of given - first names are: Bill=William, Dick=Richard, Jack=John.

First name is used in the model in a manner similar to surname code. If given names agree, then first names are not compared. However, if given names disagree, then first names may either agree or disagree. Weight calculation routines have been developed for the three possible conditions using the same general technique as discussed for surname - surname code. An additional factor which has to be considered for this variable is that one name may be an initial, while the other may be a complete name. In this case, the initial is compared against the first letter of both the given and first names of the complete name. The probability of this occurring is estimated using frequencies of initials on the file and weights for the various outcomes are also calculated.

B. Place Name

A place name dictionary for each State is utilized to standardize all spellings and abbreviations of place names and to assign a latitude - longitude location to each. (See [11].) The standardization eliminates disagreement due to different spellings of place names. The location of each is, then, used to compute the distance between two places, in a comparison when the place names are different. This distance is classified into one of seven intervals, and a different weight is calculated for each interval. The intervals are:

1. 0 to 1 miles
2. 1 to 10 miles
3. 10 to 25 miles
4. 25 to 50 miles
5. 50 to 100 miles
6. 100 to 200 miles
7. over 200 miles.

The m and u probabilities and subsequent weights for the agreement condition on place names are calculated in the same manner as is done for surname. The weight computation for place name disagreement is outlined below.

1. The m values are based on counts for each interval of matched pairs with place name disagreement taken from a sample. These are then fitted, using least squares estimates to a monotonically decreasing function of the form $y = ae^{bd}$. The fitted values form the distribution for m.

2. The u values are estimated from the file. Every pair of distinct place names is compared, their distance apart calculated, and the product of their relative frequencies summed in the appropriate interval. This yields the probability of getting place name disagreement in a particular interval by chance; i.e.,

   u(disagreement in Ith interval) = $2\,\Sigma(f_x/N)\,(f_y/N)$, where f_x, f_y are frequencies of place names whose distance apart is in interval I; and N = total number of records on file.

In practice, the further away two place names are, the larger their disagreement weight becomes.

C. Box Number and House Number

Disagreement weights for these variables are based on the amount of disagreement present. This is measured by comparing these on a character-by-character basis. (See [13].) Box and house number are up to five characters long and, thus, there are 15 different combinations of number of agreements - number of disagreements when the variable is present in both records and not identical. Different m and u probabilities and weights are calculated for each of these conditions. The key to the calculations is to estimate the appropriate probabilities for one character, given that data are present, and, then, to make the assumption that the probability of misreported data is independent of the particular character and is equal for each of them. In general, the more disagreement present, the larger the disagreement weight will be.

D. Social Security Number and Other Identifiers

Weights for identifier numbers, such as SSN, are also partitioned. Only one agreement weight is calculated for these. SSN, for example, is broken into four partitions which are assumed to be independent. (See [16].) The m and u values are calculated for one partition and independence assumptions allow these to be extrapolated to the entire number. For SSN, sixteen different weights are calculated for conditions ranging from complete agreement to complete disagreement.

See the following papers for additional information on identifier comparisons: [9] for derivation of the middle name comparison; [10] for a derivation of the negative weight to be used when one record has "Jr." and the other has no suffix; and [12] for a discussion of the additional negative weight when more than one address variable disagrees.

IV. ERROR RATES AND THRESHOLDS

Implicit in the use of the model is the assumption that the two error rates -- probability of a recording error and probability of a valid change for records associated with the matched set -- are known or can be estimated for each variable prior to processing the file through the linkage system. In the absence of prior knowledge, the current system is designed to process a sample of blocks through linkage in order to estimate these errors. (See [4] and [17].) Initial estimates are provided and the linkage decisions for the sample are manually reviewed and questionable decisions are resolved. Once this is completed, counts of error conditions are kept by variable for those pairs which are links. These are then used to estimate the necessary error rates.

To aid in this process, counts are maintained within the software for those pairs originally classified as definite links. As decisions are changed, based upon the review, these counts are updated. The importance of these estimates is demonstrated by the graph in Figure 1, which gives the frequency distribution of total comparison weights for three sets of error rates, where the rates were varied for four of the variables. As the graph indicates, the major effect of an increase in error rates (decrease in quality) is to shift the frequency curve to the right, particularly at the lower end of the scale, resulting in an increase in the number of pairs classified as possible links (weight between 5.0 and 7.5). That is, the model is unable to classify as many pairs as definite nonlinks. Pairs with small total weights are most affected, since it is in these pairs that there is the most disagreement in components, and the error rates affect most the weights assigned to the disagreement condition.

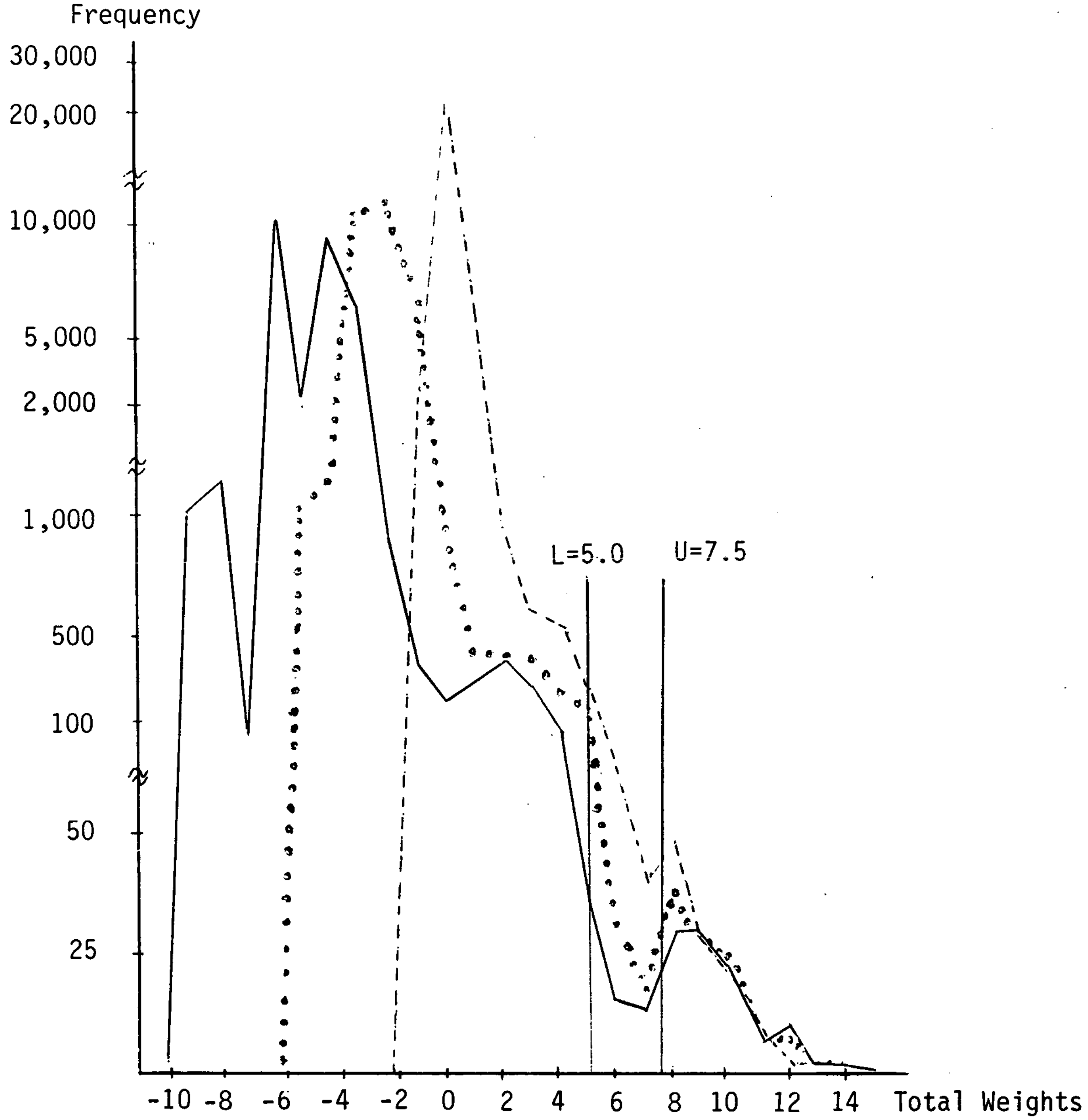
The final parameters to be supplied are the threshold values. It is these two values which ultimately determine the classification of each pair. Fellegi and Sunter suggest a technique of estimating these by sampling from the tails of the m and u probability distributions for the comparison pairs. In practice, a technique of initially estimating these -- based on a combination of weights for selected components-- and revising, as necessary -- as a result of the review of the sample used to estimate error rates -- has proven to be more satisfactory. The initial estimate of the lower threshold is made by summing the agreement weights for the most common given name, surname, and place name. This has proven to be an excellent "first guess." Another tool which can be useful in setting thresholds is the distribution of total weights. This distribution for one sample of 2,200 records is given in Figure 2. The thresholds could expect to be most efficiently set at points on either side of the lowest point on the u-shape portion of the curve (about a total weight of six in the example). The percentage of pairs classified as links after the manual resolution is also indicated for each interval in this example. Specifying the allowable rates of misclassification would, then, also determine where the thresholds will be set.

V. REMARKS

Research and analysis of results is continuing in order to further improve the procedure. For example, the possibility of using a coding procedure for given name is now being investigated. Also, questions concerning the stability of the error rates across States and, more generally the amount of preprocessing of a sample that is necessary are being investigated. The amount of manual review that is necessary after the automated procedure is also a concern. The limited amount of identifying data that is present on the lists necessitates using each item to the fullest extent possible, but it also implies that a manual review of, at least, some decisions will always be necessary.

Figure 1.--Total Weights by Frequency for Three Sets of Error Rates

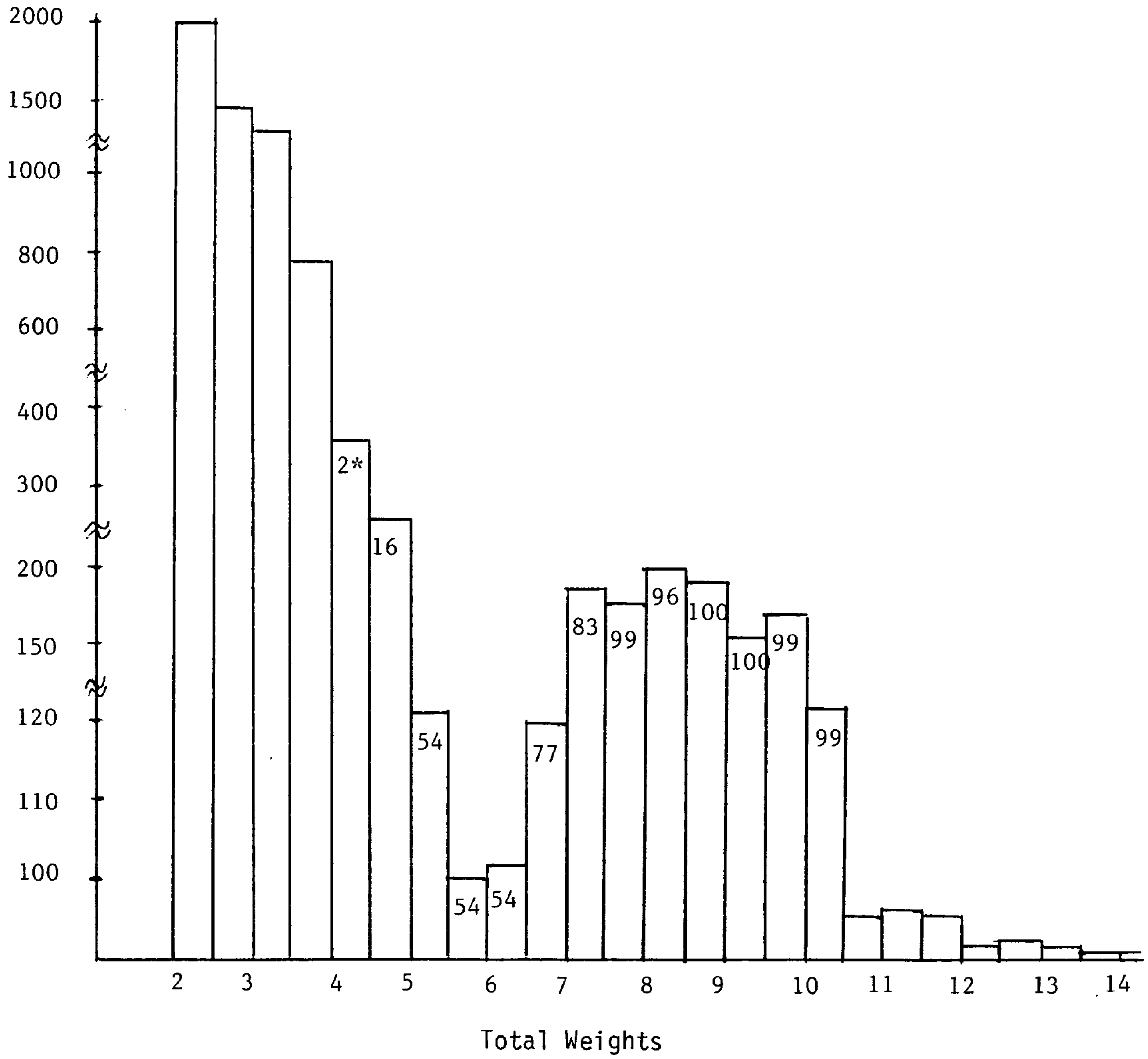
(Approximately 39,000 comparisons)



Key for Figure 1

Variable	Recording Error			Change Error		
	——	····	----	——	····	----
Given Name	.001	.01	.1	.001	.01	.1
Middle Name	.001	.01	.1	.001	.01	.1
Surname	.001	.01	.1	.001	.01	.1
Place Name	0	0	0	.001	.01	.1

Figure 2.--South Carolina Sample - Weight Distribution



*Numbers in each bar indicate the percentage of resolved pairs in that interval that were links.

The computed thresholds used prior to any resolution were 4.5 and 8.3.

NOTES AND REFERENCES

[1] Fellegi, Ivan P. and Sunter, Alan B. (1969) "A Theory for Record Linkage," Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183-1210. (Also reprinted in this volume.)

Editors' Note:

This report is part of a series of Working Papers documenting the development of a record linkage system by the Statistical Reporting Service (SRS) of the U.S. Department of Agriculture (USDA). The collection represents various stages in the research and modification of matching theory to construct a master list sampling frame of farm operators by State. The work was begun under the direction of Max Arellano and later refined by Richard Coulter and others.

Thanks to the help of Nancy Kirkendall, we have added annotated references to this paper to tie it in with related reports prepared as part of the same series. With the exception of [6], none of the papers have been previously published, and they are only available in draft form from:

Henry Power
Statistical Reporting Service
U.S. Department of Agriculture
S. Agriculture Bldg., Room 5864
Washington, DC 20250.

It is the hope of the editors that interest generated by this Workshop will lead to the eventual publication of this valuable set of papers.

[2] Arellano, Max G. (1976) "Application of the Fellegi-Sunter Record Linkage Model to Agricultural List Files," SRS, USDA.

[3] Arellano, Max G. (1976) "The Development of a Linkage Rule for Unduplicating Agricultural List Files," SRS, USDA. This paper describes the differences between the USDA assumptions and the Fellegi-Sunter assumptions as applied to probabilistic matching. Major differences are in the definition of the error rates and the assumptions concerning errors in the files used to derive agreement weights. (6 pages)

[4] Arellano, Max G. (1976) "The Estimation of P(M)," SRS, USDA.

[5] Coulter, Richard W. and Mergerson, James W. (1977) "An Application of a Record Linkage Theory in Constructing a List Sampling Frame," SRS, USDA. From the Coulter paper reprinted here, one might think that the SRS record linkage system is strictly an application of the probabilistic matching procedures. In [5], Coulter and Mergerson describe the SRS system in more detail than is found in any of the other papers. This latter paper describes preprocessing and variable identification procedures; it, then, discusses the method used to classify records as being partnership, corporate or individual records. The partnership and corporate record linkages are handled manually. Only the individual records are processed through the probabilistic linkage. The overall system adjusts for some of the matches missed because of blocking on surname by identifying for manual review all of the record pairs which agree exactly on address. This paper gives a nice overview of the entire system. (29 pages)

[6] Lynch, Billy T. and Arends, William L. (1977) "Selection of a Surname Coding Procedure for the SRS Record Linkage System," SRS, USDA. This is the only paper in the series which was published by SRS. In it, Lynch and Arends describe the analysis of surname coding systems performed by USDA. These efforts led to the selection of a revised NYSIIS (New York State Identification and Intelligence System) coding system as the most appropriate system for SRS purposes. (31 pages)

[7] Arellano, Max G. and Coulter, Richard W. (1976) "Weight Calculation for the Surname Comparison," SRS, USDA. This paper provides the mathematical derivation for the weights used for the comparison of surname, including surname code. It details the assumptions and the error terms needed in the implementation. (6 pages)

[8] Arellano, Max G. and Coulter, Richard W. (1976) "Weight Calculation for the Given Name Comparison," SRS, USDA. This paper provides the mathematical derivation for the weights used for the comparison of given names. It recognizes nicknames and initials. As in [7], it details the assumptions. (9 pages)

[9] Arellano, Max G. and Coulter, Richard W. (1976) "Weight Calculation for the Middle Name Comparison," SRS, USDA. This paper provides the mathematical derivation for the weights used for the comparison of middle names. It also accounts for agreement on middle initial. As in [7], it details assumptions. (5 pages)

[10] Coulter, Richard W. (1976) "A Weight for 'Junior' vs. Missing," SRS, USDA. This paper derives the disagreement weight for the case when one record includes "Jr." and the other record does not. (4 pages)

[11] Arellano, Max G. (1976) "Weight Calculation for the Place Name Comparison," SRS,

USDA. This paper provides the mathematical detail for the comparison of place names. Disagreement weights for the place name comparison are based on how far apart the two different places are (as calculated by using the latitude and longitude for each place). This paper also details assumptions. (5 pages)

[12] Coulter, Richard W. (1976) "Processing of Comparison Pairs in Which Place Names Disagree," SRS, USDA. This paper compares addresses and their components -- street name, street number, etc. Since these variables are probably not independent, the paper derives an additional negative weight for use when there is a disagreement on more than one address variable. (4 pages)

[13] Arellano, Max G. (1976) "Calculation of Weights for Partitioned Variable Comparisons," SRS, USDA. This paper describes the calculation of agreement weights when variables are to be compared by splitting them into different partitions and comparing the pieces -- for example, if two 3-digit numbers were compared by examining one digit at a time. (This is how house number and box number are compared.) (10 pages)

[14] "Partitioned Variable Comparison/Algorithm for Identifying Configurations," SRS, USDA. This paper translates three outcome comparison configurations on n variables to integers in the interval from 0 to 2**(n+1)-2 for purposes of indexing. (1 page)

[15] Nelson, D.O. (1976) "On the Solution of a Polynomial Arising During the Computation of Weights for Record Linkage Purposes," SRS, USDA. The procedure described in [13] for determing weights for partitioned variables needs a root of a polynomial. This paper shows that a root in the appropriate range exists and that it can be evaluated numerically. (2 pages)

[16] Arellano, Max G. (1976) "Optimum Utilization of the Social Security Number for Matching Purposes," SRS, USDA. This paper presents the derivation of weights to be used in the comparison of social security numbers. The social security number is partitioned into four pieces (of length 2,2,2, and 3) for purposes of comparison. For more on this technique, see also [13]. (10 pages)

[17] Arellano, Max G. and Arends, William L. (1976) "The Estimation of Component Error Probabilities for Record Linkage Purposes," SRS, USDA. This paper describes the estimation of error rates used in calculating most of the agreement and disagreement weights for individual variable comparisons. There are three types of errors recognized in the USDA system: errors resulting from the erroneous reporting or recording of a value, errors which are a result of a valid change in the value of a variable, and missing values. (14 pages)

[18] Coulter, Richard W. (1975) "Sampling Size in Estimating Component Error Probabilities," SRS, USDA. This paper describes the determination of the sample size required to estimate the error rates described in [17]. It also refers to [4]. (12 pages)