AN IMPLEMENTATION OF A TWO-POPULATION FELLEGI-SUNTER PROBABILITY LINKAGE MODEL

Max G. Arellano, University of California, San Francisco

I. INTRODUCTION

The California Automated Mortality Linkage System (CAMLIS) has been in operation at the University of California, San Francisco, since the fall of 1981. It was organized under the sponsorship of the Department of Epidemiology and International Health to facilitate the clearance of study population files submitted by qualified investigators against mortality files for the State of California.

The linkage of two independently generated data files has long been thought to be the exclusive province of highly trained clerks because of the need to process the discrepancies which frequently occur between sets of identifying information for the same person on the two files.

A computerized approach to the record linkage problem can adopt either deterministic or probabilistic decision criteria. Deterministic linkage criteria require the formulation of a 'match key' to establish the relationship between records on the two files to be linked. This match key functions on an 'either or' basis, i.e., if an identical value of the match key is found on both files, the records with the identical values are said to be matched. Otherwise, the records are said to be unmatched. In order to perform its required function with minimal error, this match key must possess as many of the characteristics of a unique identifier as possible. Match keys can be constructed from any conceivable combination of last name, first name, sex, social security number, birth date (or portions thereof), or any other identifying items present on the file. Although it is not a true unique identifier, the ready availability of the social security number has led to its widespread use as the match key of choice in deterministic linkage applications.

Probabilistic linkage criteria are based on a linkage weight calculated for each pairwise comparison between records on the two files to be linked; these linkage weights are the sum of component weights calculated for each item of identification contained on the two files. The component weights are functions of occurrence probabilities and of the reliability of the data items. Probabilistic decision criteria provide an attractive alternative to deterministic linkage criteria as a means of computerizing the record linkage activity primarily because: 1) they assign weights in a manner that is consistent with our own human intuition and 2) they can accommodate partial agreements. On the debit side: 1) they require the estimation of many parameters, some of which are inestimable, 2) they are much more difficult to program and 3) they are more costly to use.

Our decision to adopt probabilistic decision criteria was based primarily on our conviction, based on a careful analysis of the available information, that the requirements of investigators in the health and medical care research fields could not be met solely by deterministic linkage criteria. Our experience over the last four years has served to confirm the validity of that decision.

II. THE FELLEGI-SUNTER WEIGHTING ALGORITHM

The Fellegi-Sunter [1] weighting algorithm requires the estimation of two probability distribution functions:

If we let,

p_{jA} = P(Occurrence of the jth configuration in population A)

p_{jB} = P(Occurrence of the jth configuration in population B)

$p_{jA\cap B}$ = P(Occurrence of the jth configuration in $A\cap B$)

$w(\gamma_j)$ = Probability linkage weight for the jth agreement configuration

$m(\gamma_j)$ = P(Occurrence of the jth agreement configuration | the record pairs are associated with members of the matched set)

$$= P(\gamma_j \mid (a,b)\in M)$$

$$= p_{jA\cap B}(1-e_A)(1-e_B)(1-e_T)$$

$u(\gamma_j)$ = P(Occurrence of the jth agreement configuration | the record pairs are associated with members of the unmatched set)

$$= P(\gamma_j \mid (a,b)\in U)$$

$$= p_{jA}p_{jB}$$

Then, $w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$

Among the obvious difficulties encountered in the implementation of this model are:

(A) It does not address the problem of estimating the e or e_T terms. We generally refer to these as the "component error probabilities."

(B) The $p_{A\cap B}$ term requires information which can only be obtained when the linkage has been completed in a satisfactory manner, if then.

If the populations represented by the files that are being linked can be regarded as samples drawn from the same population, i.e., the "one-population" model, some simplifications can be introduced into the above expressions:

$$m(\gamma_j) = p_j(1-e)^2(1-e_T)$$

$$u(\gamma_j) = p_j^2$$

$$w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$$

$$= \log[p_j^{-1}(1-e)^2(1-e_T)]$$

Moreover, if the data are being collected continuously, as is generally the case under the circumstances to which the one-population model is

applicable, procedures can readily be developed to iteratively obtain "good" estimates of the component error probabilities. This is, unfortunately, not the case for situations to which the two-population model would generally be applied. For one thing, if the populations being linked do not overlap, the $p_{A\cap B}$ term is meaningless. The model also requires estimates of component error probabilities specific to the files that are being linked.

Prior information on the record-pairs that correspond to the intersection of the two populations is obviously desirable, if not absolutely necessary, before probability linkage can be initiated. However, since this is precisely the information we are attempting to obtain by means of probability linkage, if it can be obtained by other means, one may legitimately question the need for probability linkage.

In this paper I will describe the approach that has been adopted by the CAMLIS project to the problem of implementing a two-population Fellegi-Sunter model.

III. THE CAMLIS IMPLEMENTATION OF THE TWO-POPULATION FELLEGI-SUNTER MODEL

Central Concepts

The CAMLIS approach is based on the following central concepts:

(A) A two-stage linkage process, consisting of a deterministic first stage (primarily based on the social security number) followed by a probabilistic second stage, is necessary to achieve the desired performance characteristics. This strategy has several benefits:
  (1) Each stage is capable of detecting valid linkages which will escape detection by the other stage.
  (2) Since deterministic linkage is carried out first, the correctly matched records which it produces can be used to derive estimates of the component error probabilities required by probability linkage.

(B) A phonetic name encoding algorithm with superior operating characteristics must be used to form the basic comparison groups for probability linkage to minimize the number of pairwise record comparisons that must be carried out. We chose to adopt a modified version of the New York State Identification and Intelligence System (NYSIIS) phonetic coding system for this purpose. It is doubtful if CAMLIS could be operated on a cost-effective basis without the use of a phonetic name coding system with the superior performance characteristics of NYSIIS.

(C) A modification of the weighting algorithm for the two-population Fellegi-Sunter model is necessary to compensate for the inestimable parameters.

(D) Component error probabilities can be estimated from the "matched set" produced by first stage or deterministic linkage.

In this presentation, I will focus primarily on points (C) and (D) above, i.e., on our approach to the estimation of the parameters required by the two-population Fellegi-Sunter weighting algorithm.

The Estimation of Relative Frequency Parameters

In CAMLIS applications, a user file, which we denote as file L_A, is linked to a California State mortality file, which we denote as file L_B. Since the characteristics of most user files are significantly different from those of the California mortality file, the two-population model is obviously called for. However, many of the parameters required by the two-population model, e.g., $p_{A\cap B}$ and e_A, are inestimable. We therefore carefully scrutinized the expressions for the two probability distribution functions to determine whether a simplification was possible. We first made the observation that the characteristics of the user file are always subsets of the characteristics of the mortality file; we also observed that, for those components that are independent of mortality, $p_A \sim P_{A\cap B}$. These observations resulted in the elimination of the p_A term from the weighting algorithm and served to justify the use of relative frequencies derived only from the mortality files. Since these relative frequencies can change over time, files have been developed which contain the necessary relative frequencies at five-year intervals; CAMLIS procedures retrieve them as necessary.

The component for which the assumption is not tenable is birth year; an entirely different approach to weight computation for the birth year component has, therefore, been developed.

The Estimation of Component Error Probabilities

Within the context of a mortality clearance system, it is not possible to derive separate estimates of component error probabilities for files L_A and L_B; there is just not enough information available. We therefore made the simplifying assumption that the corresponding component error probabilities in the two files were identical, i.e., we assume that:

$$e = e_A = e_B$$

Estimates of e and e_T are derived from the matched record-pairs produced by first stage deterministic linkage. To eliminate spurious matches, we require a high concordance among the identifying elements on the two files that are not incorporated into the match key.

The basic algorithm that we utilize to calculate agreement configuration weights is therefore:

$$m(\gamma_j) = p_{jA}(1-e)^2(1-e_T)$$

$$u(\gamma_j) = p_{jA}p_{jB}$$

$$w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$$

$$= \log[p_{jB}^{-1}(1-e)^2(1-e_T)]$$

IV. CONCLUSION

The Fellegi-Sunter model requires an assumption regarding the independence of the components of the comparison vector; this assumption is frequently a major concern in linkage applications. It is not my intention to minimize the importance of this assumption. The real concern, however, must be the extent to which violations of

this assumption affect the results produced by the model.

(A) The components of the comparison vector should be carefully chosen. Only one of several highly dependent components should be incorporated into the model.

(B) Although it is possible to correct for the effect of dependence, for moderately dependent components, these efforts are hardly ever worth the small gain in precision that can be realized.

(C) We have done a great deal of difference analysis. Our conclusion is that the estimated component error probabilities and relative frequencies must differ considerably from the appropriate values to significantly affect the computed weights.

(D) For matches that achieve a linkage weight significantly greater than the upper threshold value, a bias in the weight is obviously of no consequence. Similarly, for matches that achieve a linkage weight significantly below the lower threshold value, a bias in the weight is also of no consequence. The vast majority of record-pairs achieve either very low or very high linkage weights.

(E) Record-pairs which achieve a linkage weight between the lower and upper threshold values are subject to manual review. Since record-pairs fall into this category because they either contain ambiguous or sparse identifying information, it is extremely doubtful whether they would differ significantly if the weights were computed according to a more precise model. In any case, comparable results could be obtained by redefining the upper and lower threshold values.

The major advantage of probability linkage is that it permits a meaningful ranking of matched record-pairs. The ranking makes it possible to focus review efforts on the comparisons which have been assigned borderline weights. It can readily be shown that the gain achieved by verifying the probability linkage decisions above a certain threshold value and below a certain threshold value is negligible.

Our experience with the Fellegi-Sunter probability linkage criteria has been uniformly favorable. It is our considered opinion, however, that probabilistic linkage and deterministic linkage are best utilized as complimentary procedures and that both are necessary to achieve optimum results.

REFERENCES

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