Office of Tax Analysis
U.S. Treasury Department
Washington, D.C. 20220
Issued: July, 1976

Reducing and Merging Microdata Files

J. Scott Turner
Gary B. Gilliam
U.S. Treasury Department

OTA Paper 7 October, 1975

TABLE OF CONTENTS

page

I. A Network Model to Reduce the Size of Microdata Files . . . 1

A. Introduction 1

B. The Distance Function 2

1. Definition 2

2. Complications 3

C. The File Reduction Problem as a Network Model 4

D. Conclusion 7

II. Network Models Applied to Microdata Set Merging 8

A. Introduction 8

B. The Matching Problem as the Assignment Model 9

1. The Distance Parameter C_{ij} 11

a. Sum of Squared Deviations 11

b. Adjustment for Relative Magnitudes 11

c. Weight Adjustments 12

C. The Merge Problem as the Transportation Model 13

D. Conclusion 14

I. A NETWORK MODEL TO REDUCE THE SIZE OF MICRODATA FILES

A. Introduction. Microdata files are typically very large.[1/] Many researchers using these computer-based files are faced with either clock time or central processing unit time constraints, and consequently smaller representative files are required. In such situations it is desirable to have a method of file reduction which "minimizes" the amount of information lost.

The problem of file reduction involves both the selection of records to be in the reduced file, and the selection of weights[2/] for these records. The most typical and also the most elementary way to select records for a reduced file is to use random selection (for example, if the file is to be reduced to one-third, each record in the original file has a selection probability of one-third). The most elementary way to reweight the records selected is to scale all the original weights of the selected records by the inverse of the selection probability (for example, if the file is reduced to one-third, all weights are multiplied by three). Random selection and reciprocal reweighting is an unbiased method, yet it exerts no direct control over the information which is lost as a file is reduced. Similar records might be selected and records dissimilar to all other records might be removed.

It is the intent of this section to demonstrate that file reduction can be viewed as a network model. An objective function can be generated using the parameters of a microdistance function and then minimized to reduce a file from n records to n* records. The solution to the network model indicates uniquely which records should be selected for the reduced file and the new weight of each selected record.

B. The Distance Function.

1. Definition. File reduction techniques are designed to select one set of records from the original file for preservation in the reduced file, and to remove the complementary sets of records. The extent to which records removed from the original file can be represented by reweighted records in the reduced file depends on the ability of the reduced file to represent both the original variance-covariance matrix, and the vector of data item means of the original file. For example, if a given record i is removed and is identical in every detail to a record j which remains in the reduced file, and record j's weight is increased by the weight of record i, then the means of data items and the variance-covariance matrix of the reduced file are identical to those of the original file.[3/] If however, a removed record is very dissimilar to any record included in the reduced file, then the variance-covariance matrix and the vector of means are affected.

There is not a unique quantitative measure of the degree of closeness between records removed and records retained by a file reduction technique. However, the problem of distance between records in a microdata file is similar to the problem of distance between coordinates in a multidimensional space in multivariate regressions. In effect, a record in a microdata file with m data items is a point in an m dimensional space. The squared distance between two microdata records i and j could be given by the microdistance function (1.1).

$$C_{ij} = \sum_{p=1}^{m} (a_{ip} - a_{jp})^2 \qquad (1.1)$$

a_{ip} = amount of item p in the i th record

a_{jp} = amount of item p in the j th record

A microdistance function such as (1.1) indicates the extent that the profile of attributes of one record matches the profile of the attributes of another record without any reference to the magnitude of the attributes. In this sense, the use of a microdistance function compares the profiles of the records in the file.

2. Complications. Problems of index variables and scale arise when specifying a distance function between records in a microdata file. Frequently the data items are indices such as race, sex, type of household, type of employment, and so forth, rather than magnitude items. It is common to indicate an infinite distance between two records if, for example, race or sex is different. In effect, this specification means that one of these records cannot represent the other in the reduced file. Also the data items which are magnitudes such as wages and salaries, business income, dividends, social security benefits, family size, and so on, have very different relative magnitudes. For example, average dividend income in a file might be $200 whereas average wages and salaries might be $10,000; consequently, small percentage deviations in wages would dominate large percentage deviations in dividend income. This problem can be addressed by normalizing differences by standard deviations.

C. The File Reduction Problem as a Network Model. Given a microdistance function such as (1.1), and given that the reduced file is a reweighted subset of the original file, a mathematical programming problem can be specified to minimize a macrodistance function if an original file with n records is reduced to n* records. In figure 1, each node under the heading i represents one of the n original records in a microdata file, while the nodes under the heading j denote all the records which could possibly be in the reduced file. Only n* of the records under the heading j will in fact be retained in the reduced file. If record j* is in the reduced file, then j* is record j of the original file with possibly a larger weight.

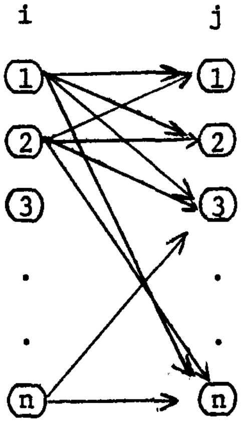

Figure 1

The arc linking record i with record j indicates that record i can be replaced with record j. Let C_{ij} denote the distance of the arc between i and j. Note that all arcs (i,i) are in the network and that the distance elements C_{ii} necessarily equal zero. The problem of minimizing arc distance can be specified in the following manner:

Minimize $\sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij}X_{ij}$ (1.2)

subject to $\sum_{i=1}^{n} X_{ii} = n*$ (1.3)

$\sum_{j=1}^{n} X_{ij} = 1$ for i=1,2,...n (1.4)

$\sum_{i=1}^{n} a_{ij}X_{ij} \leq 0$ for j=1,2,...n (1.5)

$$a_{ij} = \begin{cases} n* - n & \text{for } j=i \\ 1 & \text{for } j \neq i \end{cases} \quad (1.6)$$

X_{ij} is either zero or one (1.7)

This model will minimize the macrodistance function (1.2) if the file is reduced from n records to n* records. The constraint (1.3) specifies the size of the reduced file. If $X_{ii} = 1$, then record i is in the reduced file. If $X_{ii} = 0$, then the i th record is not in the reduced file, and constraint (1.4) requires that there will be one and only one positive X_{ij} (with i≠j) which identifies the record j which represents record i in the reduced file. Constraint (1.5) insures that record i will not replace any record k if $X_{ii} = 0$. However, if $X_{ii} = 1$, the coefficient of X_{ii} in (1.5) is sufficiently large and negative to allow any or all of the original records to be replaced by the i th record.

The basis for the above model is a matrix with n* diagonal elements X_{ii} and (n - n*) off-diagonal elements X_{ij}. The basis for the above model contains the linking of records which are not in the reduced file to the record representing such records in the reduced file, and therefore can be used to reweight the records in the reduced file. Every record in the original file is the origin for one and only one positive arc X_{ij} which terminates with record j. If a record is in the reduced file, it is the origin of a positive arc which terminates with itself, that is, $X_{ii} = 1$. If X_{ii} is in the basis, there will be no arc X_{ij}, $(i \neq j)$ in the basis. If arc X_{ii} is not in the basis, one and only one arc X_{ij} will be in the basis.

The result achieved by objective function (1.2) is dependent upon the analyst's choice of a microdistance function C_{ij}. This choice in turn depends on the data items which are relevant to the particular studies requiring the reduced file. For example, the function (1.1) when used with (1.2) minimizes the sum of squared distances between each original record and its corresponding destination in the reduced file. The microdistance function

$$C_{ij} = \sum_{k=1}^{m} |a_{ik} - a_{jk}| \tag{1.8}$$

when used with (1.2) minimizes the sum of absolute distances between each record in the original file and the record which is its destination in the reduced file. Other objectives, such as minimizing the sum of squared differences between the vectors of data means (or variances) in the original file and the reduced file, would require the specification of a quadratic objective function, as well as segmented distances C_{ijk}. Such topics will not be discussed in the study.

The model given by (1.2) - (1.7) demonstrates that the file reduction problem for a file with n records can be specified as a zero-one network problem with 2n nodes, n^2 possible arcs, and (2n + 1) restrictions. This model has the form of a generalized transportation problem with one additional restriction.

D. Conclusion. There are many occasions for reducing files with a large number of records. Specifying file reduction as a network problem improves upon older reducing techniques, such as random selection and reciprocal weighting. An objective function may be generated using the parameters of a microdistance function, and then minimized to reduce a file from n to n* records. The solution to this network model indicates which records should be selected for the reduced file, as well as their new weights.

II. NETWORK MODELS APPLIED TO MICRODATA SET MERGING

A. Introduction. Microdata set merging is increasingly being used to broaden the information of one data set by transferring information from records of another microdata set. It is the purpose of this section to demonstrate that the analytic form of merging can also be viewed in the more general framework of a network model. Analytic objective functions can be specified and algorithms can be used to produce the most efficient matching of data records.

A principal goal in merging is to select subsets of records in the two data sets which represent the same, or approximately the same, persons or families.[4/] For example, the subset of family records in microdata set A with family income between $11,000 and $12,000, no income other than wages and salaries, married husband and wife present, with two dependent children, is considered to represent families with the same attributes in microdata set B. After the subsets have been determined by the selection of record attributes, usually either a random selection technique is applied to data set B to select a record to be matched with a given record in data set A, or a sort merge procedure is employed.[5/] Many variants to the matching technique are possible, depending upon the relative sizes of the corresponding subsets of A and B. If there are more records in the subset of A than in the corresponding subset of B, selection with replacement can be used, and if the situation is reversed, selection without replacement is typically used.

It is our intent to show that, once the corresponding subsets in data sets A and B have been selected, it is possible to replace the random matching solution or the sort merge technique with a matching solution which is uniquely determined by an algorithm and which is the "best" matching solution (as defined by the analyst's criteria). At the outset, the analyst can quantitatively specify the criteria for the "closeness" of the match; that is, specify an objective function, and an algorithm that can be applied to minimize this objective function subject to constraints. The merging problem, after the selection of the corresponding subsets A and B, is a network problem and can be viewed as either the assignment model or the transportation model.

B. The Matching Problem as the Assignment Model. The most elementary form of a microdata merging problem is the special case where the two data sets A and B to be merged have the same number of physical records, and each record has the same weight. A given record in data set A can be matched with any given record in data set B or vice versa, and the question is to find the optimal match of records between the two data sets.

Let C_{ij} denote the value of the distance function of matching the ith record of data set A with the jth record of data set B. This parameter is a function of the common variables in the two data sets. If the common variables have exactly the same value in the ith record of A and the jth record of data set B, we can state that the value of the distance function is zero. When differences exist between the common variables, a positive distance between records exists.

Given parameters C_{ij}, the merging problem can be specified as a linear mathematical programming problem. If the number of records in a given subset of A is equal to the number of records in the corresponding subset of B, and each record has a weight of one, the linear mathematical programming model is identical to the assignment model. The matching problem as the assignment model is given below in equations (2.1), (2.2), and (2.3).

$$\text{minimize} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij} C_{ij} \tag{2.1}$$

$$\text{subject to} \quad \sum_{j=1}^{n} X_{ij} = 1 \text{ for } i = 1,2,\ldots.n \tag{2.2}$$

$$\sum_{i=1}^{n} X_{ij} = 1 \text{ for } j = 1,2,\ldots.n \tag{2.3}$$

X_{ij} = 1 if the ith record in A is matched with jth record in B

0 if the ith record in A is not matched with the jth record in B

C_{ij} = distance parameter to be defined by analyst.

The constraints denoted by equations (2.2) and (2.3) require that each record in data set A be matched with one record in data set B, and vice versa. The mathematical form of the merging problem is the same as the mathematical form of the assignment problem.

Algorithms exist to solve the model given by (2.1), (2.2), and (2.3). The restriction that the number of records in the subset of A is equal to the number of records in the corresponding subset of B can be relaxed, and in this situation the linear programming model is coincidental with the transportation model which is discussed later.

1. The Distance Parameter C_{ij}. The distance parameter C_{ij} can take any form, linear or nonlinear, deterministic or stochastic. The analyst decides which form is most appropriate given the overall objective of the study. For example, poverty studies are sensitive to transfer receipts, whereas personal tax studies are sensitive to items affecting the tax base.

a. Sum of Squared Deviations. A distance function which is the sum of squared deviations is specified in equation (3.1).

$$C_{ij} = \sum_{k=1}^{P} (a_{ik} - b_{jk})^2 \qquad (3.1)$$

a_{ik} = value of variable k in the ith record of A

b_{jk} = value of variable k in the jth record of B

Another possibility for C_{ij} is to use absolute rather than squared deviations. The important point is that C_{ij} is calculated outside the network model and enters the network algorithm as a parameter.

b. Adjustment for Relative Magnitudes. In many cases the distance function will not be adjusted for differences in absolute magnitudes of the common variables. However, it could be perceived that percentage deviations are the most important criterion for determining the closeness of fit of a match. For example, a difference of $100 in the level of interest received between a record in A and a record in B might be more important than a difference of $500 in wages and salaries. The $100 deviation in interest might represent a 50 percent deviation from average interest received whereas the $500 deviation in wages and salaries might represent only a 5 percent deviation from the average amount of wages and salaries. One way to confront the problem is to calculate the means for each of the common variables and use the ratio of the deviation to the mean in the distance function. Equation (3.2) represents a distance function using this concept.

$$C_{ij} = \sum_{k=1}^{P} \frac{(a_{ik} - b_{ik})^2}{Z_k} \tag{3.2}$$

Z_k = mean value of the kth common variable using observations from both data sets A and B.

c. Weight Adjustments. The issue of the importance of one common variable relative to another can be made quantitatively explicit through the inclusion of subjective weights h_k. For example, if interest is considered to be only 25 percent as important as wages and salaries the subjective weight for interest would be .25 and the corresponding subjective weight for wages and salaries would be 1.0. One approach is to first make adjustments for different magnitudes, and then make adjustments for the relative importance of each variable. Subjective weights can be incorporated into distance functions (3.1) and (3.2) as multiplicative factors.

$$C_{ij} = \sum_{k=1}^{P} \frac{(a_{ik} - b_{ik})^2}{V_k} \tag{3.3}$$

V_k = $1/h_k$ or Z_k/h_k

h_k = subjective importance weight for the kth common variable

C. The Merge Problem as the Transportation Model. The most typical form of a microdata merge problem is the situation where the records in data sets A and B have different weights with the provision that the sum of the weights in each set is equal. Data set A has n records and the weight of the ith record is w_i. Data set B has m records with the jth record having weight y_j.

$$\sum_{i=1}^{n} w_i = \sum_{j=1}^{m} y_j \qquad (4.1)$$

The parameter C_{ij} is the value of the micro distance function of matching the jth record in data set B with the ith record in data set A. If the w_i and y_j parameters are integers, the merging problem can be specified as the transportation model.

$$\text{Minimize} \quad \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij}X_{ij} \qquad (4.2)$$

$$\text{Subject to} \quad \sum_{j=1}^{m} X_{ij} = w_i \text{ for } i = 1,2,\ldots\ldots\ldots\ldots n \qquad (4.3)$$

$$\sum_{i=1}^{n} X_{ij} = y_j \text{ for } j = 1,2,\ldots\ldots\ldots\ldots m \qquad (4.4)$$

The solution to this problem identifies the records in data set B which are to be merged with each record in data set A. Record i in data set A might be associated with more than one positive X_{ij} which indicates that the merged file can contain split records. For example, record i in set A might have a weight of 15,000, record j in set B has weight 10,000, and record k in set B has weight 5,000. If $X_{ij} = 10,000$, and $X_{ik} = 5,000$, then the ith record in A has been split into two identical records (one with weight 10,000 merged with the jth record in B, and the other with weight 5,000 merged with the kth record in B). The maximum number of records in the merged file is n+m-1 which denotes the maximum amount of record splitting.

The merged file generated by solving the transportation problem preserves all the marginal and joint distributions of the variables in the original data sets A and B. This conclusion is based on the following observation: Any weighted a_{ik} or b_{jk} in files A or B will appear in the merged file in one or more records. If the data item appears in only one record, it will have the original weight, and if it appears in more than one record, its combined weight equals the original weight.

D. Conclusion. At the present time, efficient algorithms for the actual matching of data sets are not being used. Much of the effort is spent in the creation of meaningful subsets for merging the data sets A and B, with the actual matching being done using sort merge techniques. We clearly see a need for the creation of a distance function along the lines of those specified in this paper, and the use of mathematical programming techniques to improve the closeness of the matches.

Identifying data merging as a network problem allows the comparison of the mathematical structure of this model with the inventory of standard models in network analysis. It is our conclusion that the merging of data sets can be classified as the classical transportation model, and in certain situations, it can be classified as the classical assignment model. The benefit of this classification is that the merged file is the one which minimizes the macro distance function while preserving the marginal and joint distributions inherent in the original data sets.

FOOTNOTES

[1] For example, the *Current Population Survey* has approximately 200,000 weighted records, the *Statistics of Income* sample has approximately 250,000 records, and the *Public Use Decennial Population Census* has over 200,000 records.

[2] The files are weighted so as to maintain the control totals. The weight of the record indicates the number of units in the population which this observation represents. For example, in a probability sample where each household has one chance in 10,000 of being selected, the weight of each selected household is 10,000, the inverse of the selection probability. Reduced files are subsamples with records reweighted in accordance with the control totals of the original file.

[3] Suppose $a_{ik} = a_{jk}$ for all data items k. (See equation 1.1.) The weight of the i th record is w_i and the weight of the j th record is w_j. In the calculation of the mean of the k th data item in the original file the contribution of the i th and the j th records is $w_i\ a_{ik} + w_j\ a_{jk} = (w_i + w_j)\ a_{jk}$. In the reduced file, the contribution of the reweighted j th record is $(w_i + w_j)\ a_{jk}$. In the variance-covariance matrix of the original file, the contribution of the i th and j th records is $w_i(a_{ik} - u_k)(a_{ip} - u_p) + w_j\ (a_{jk} - u_k)(a_{jp} - u_p) = (w_i + w_j)(a_{jk} - u_k)(a_{jp} - u_p)$. In the variance-covariance matrix of the reduced file, the contribution of the reweighted j th record is $(w_i + w_j)(a_{jk} - u_k)(a_{jp} - u_p)$.

[4] "Constructing a New Data Base From Existing Microdata Sets: The 1966 Merge File", Benjamin Okner, *Annals of Economic and Social Measurement*, July 1972.

[5] "The Creation of a Microdata File for Estimating the Size Distribution of Income", Edward C. Budd, *The Review of Income and Wealth*, December 1971.