Survey Methodology, December 1998
Vol. 24, No. 2, pp. 171-183
Statistics Canada

Use of Statistical Matching Techniques in Calibration Estimation

ROBBERT H. RENSSEN[1]

ABSTRACT

This article deals with an attempt to cross-tabulate two categorical variables, which were separately collected from two large independent samples, and jointly collected from one small sample. It was assumed that the large samples have a large set of common variables. The proposed estimation technique can be considered a mix between calibration techniques and statistical matching. Through calibration techniques, it is possible to incorporate the complex designs of the samples in the estimation procedure, to fulfill some consistency requirements between estimates from various sources, and to obtain fairly unbiased estimates for the two-way table. Through the statistical matching techniques, it is possible to incorporate a relatively large set of common variables in the calibration estimation, by means of which the precision of the estimated two-way table can be improved. The estimation technique enables us to gain insight into the bias generally obtained, in estimating the two-way table, by sole use of the large samples. It is shown how the estimation technique can be useful to impute values of the one large sample (donor source) into the other large sample (host source). Although the technique is principally developed for categorical variables Y and Z, with a minor modification, it is also applicable for continuous variables Y and Z.

KEY WORDS: Consistency between estimates; General regression estimator; Imputation; Multivariate auxiliary information; Two-way table.

1. INTRODUCTION

Most statistical surveys are conducted to obtain estimates of simple descriptive finite population parameters. The estimates are often presented in tabular form, with cells containing estimates of population totals or subgroup totals. Often, data are collected on an extensive set of variables, producing numerous results for these variables and their relationships. In order to save resources and decrease response burden, statistical bureaus wish to reduce sample sizes and shorten questionnaires. They resort to administrative data sources and existing large-scale sample surveys, or applying splitting questionnaire survey designs (see Raghunathan and Grizzle 1995). As a consequence, methods for combining distinct data sources have become a popular tool in the production of statistics. Combining data sources can be done in many different ways; two well-known techniques in survey sampling are statistical matching and calibration estimation.

Singh, Mantel, Kinack and Rowe (1993) describe statistical matching as a special case of imputation in which there are two distinct micro-data sources containing different information on different units. One data source serves as a host or recipient file to which new information is imputed for each record, using data from the other source, which is the donor file. More specifically, they consider a host file A, containing information on variables (X, Y) and a donor file B containing information on variables (X, Z). The common variable X can be used to identify similar units in the two files. In general, statistical matching deals with the problem of completing the records in file A, by imputing values for Z using the information on the (X, Z) relationship in file B. These imputed Z-values suffer from a serious limitation in that, the real relationship between Y and Z may be completely lost in the enriched host file. This limitation amounts to the so-called assumption of conditional independence between Y and Z given X. In order to get rid of this conditional independence assumption, Singh *et al.* (1993) consider a third data set (file C) representing auxiliary information about the full set (X, Y, Z). For example, this data set could come from a small-scale specially conducted survey. They discuss several imputation methods to complete file A, by adding Z from file B using information from A, B, and C, on the joint relationships of X, Y, and Z. Singh *et al.* (1993) give many relevant references on statistical matching techniques. We only mention Rodgers (1984), Rubin (1986) and Paass (1986).

In Deville and Särndal (1992), calibration estimation is derived as a general technique to weight sample surveys, taking into account the complex design of the sample and auxiliary information obtained from external sources (see also Deville, Särndal, and Sautory 1993). The use of auxiliary information, *i.e.*, control variables, primarily aim at three goals: namely, reducing sampling variance, reducing bias due to non-response, and ensuring consistency between estimates from various sources with respect to the used control variables. There is an extensive body of literature on weighting methods in sample surveys. We refer to Bethlehem and Keller (1987), Alexander (1987), Lemaître and Dufour (1987), and Zieschang (1990).

[1] Robbert H. Renssen, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, Netherlands.

This article deals with the specific problem of how to estimate the cross-product between Y and Z (*e.g.*, the two-way table between Y and Z in case these variables are categorical or the covariance between Y and Z in case these variables are continuous), using statistical matching techniques as well as calibration estimation. We assume that two data files A and B represent two large-scale sample surveys, possibly both obtained by a complex design. In order to weight the specially conducted small sample (file C), auxiliary information is derived from these large samples. It might be difficult to judge whether the large samples should be considered as suppliers of auxiliary information for the small sample, or vice versa. Through the statistical matching, it is possible to incorporate a large set of X-variables in the estimation procedure, despite the sample size of the small sample. The use of calibration estimation makes it possible to take account of the complex design of all samples in the estimation procedure, and to fulfill some consistency requirements. Most of the article is devoted to categorical Y and Z, because of the specific properties of these variables. For example, it is shown that the marginal counts of the estimated YZ-table, always coincide with estimates for the population totals of Y and Z, when the ordinary calibration estimator is applied with the X-variables as control variables, on the first and second large sample respectively. Nevertheless, the proposed method is also applicable for continuous Y and Z. Throughout this article it will be assumed that X may consist of several variables, which may be categorical and/or continuous. It is argued that when the X-variables are highly correlated with either Y or Z, then our estimation method gives relatively precise estimates for the cross-product between Y and Z, *e.g.*, for the complete YZ-table when Y and Z are categorical.

The proposed estimation procedure closely resembles a method presented in Singh *et al.* (1993, Section 2) to estimate a correlation coefficient between Y and Z. These variables are assumed to be univariate in this article. Our method, however, differs from theirs in that it incorporates the complex designs of all data sources in the estimation procedure and that it uses the large data sources more efficiently in estimating population parameters from the small data source. When Y and Z are categorical, and there is no linear correlation between X and Y as well as between X and Z, then our method corresponds to incomplete post-stratification (Deville and Särndal 1992, Bethlehem and Keller 1987). On the other hand, if Y is perfectly correlated with X, then our method gives an estimated two-way table between Y and Z which corresponds to an estimated two-way table that would have been obtained from file B if first the Y-values were imputed. A similar result holds if Z and X are perfectly correlated.

Although combining distinct data sources across common variables may be fruitful from a theoretical point of view, in practice, complications may arise because common variables in the strict sense are not easily found, mainly due to discrepancies between definitions, methods of observation, and reference period. These complications may be reduced if the survey processes involved, are harmonized at an early stage. A promising application of the use of common variables, lies in integrated survey designs, such as the Dutch Household Survey on Living Conditions, see van Tuinen (1995), Bakker and Winkels (1998), Winkels and Everaers (1998), and Hofmans (1998). The questionnaire design of this survey has a three-shell structure. The first shell contains questions on demographic and socioeconomic issues, and level of education. The second shell contains a few easy to answer core questions, on every relevant aspect of living conditions. The questions in the third shell also concern living conditions, but they are more exhaustive than the questions in the second shell. In order to shorten the time it takes to answer, the third shell questionnaire is split. Each respondent has to fill in the complete questionnaire of the first and second shell and one sub-questionnaire of the third shell. On account of the third shell, the sample is split into sub- samples associated with each sub-questionnaire. The sampling design of each sub-sample can be described as two-phase sampling for the general regression estimator.

The organization of this article is as follows. The theoretical framework is developed in Section 2. For this purpose it is convenient to discuss a calibration estimator for the small sample, obtaining auxiliary information from two distinct registrations instead of two distinct large samples. One registration contains values on X and Y and the other registration on X and Z. Sections 2.1 to 2.4 deal with categorical Y- and Z-variables. In Section 2.1, the registrations are used to obtain a first synthetic estimate of the YZ-table by regression methods of imputation. It is shown that this synthetic two-way table has some interesting properties. In Section 2.2 we propose a set of calibration equations to weight the small sample, based on these properties. We briefly discuss its relationship to complete and incomplete post-stratification. A numerical illustration is given in Section 2.3. The linkage to statistical matching techniques as discussed in Singh *et al.* (1993) is given in Section 2.4. The treatment of categorical Y and Z is unnecessary and restrictive. In Section 2.5, it is shown that the proposed weighting technique is also applicable for continuous Y and Z or for continuous Y and categorical Z. In Section 3, the technique is modified, using auxiliary information from two distinct large samples instead of two registrations. By means of a simulation study, the modified weighting method is compared to the traditional incomplete two-way stratification. Finally, Section 4 contains some concluding remarks.

2. COMBINING REGISTRATIONS ACROSS COMMON VARIABLES

Consider a finite population $\Omega = \{1, ..., N\}$ of N persons and suppose there are two registrations available of these

persons. The first registration contains of each person k, a record with scores y_k and x_k of the variables Y and X respectively, and the second registration of each person k, a record with scores z_k and x_k of the variables Z and X respectively, $k = 1, ..., N$. Obviously, the variable X is present in both registrations. We note that the records from both registrations correspond to the same finite population. The process of merging these registrations, would be like exact matching if X is used to compare the records in the one registration with those in the other registration, in an effort to determine which pairs of records relate to the same population unit (see Fellegi and Sunter 1969). In this article we will proceed differently.

2.1 Formulating the Synthetic Population Totals

Let Y denote education with p categories and Z denote employment with q categories. Then y_k is a vector of order p, representing p dummy variables. Each dummy variable corresponds to a specific category; it equals 1 if person k belongs to that category, otherwise it equals 0. Analogously defined, z_k is a vector of order q. Further, X may be the result of a complete or incomplete crossing (stratification) of a number of characteristics (*e.g.*, sex, age, region, marital status, *etc.*). The scores x_k are vector valued, of order r. In case X consists of a complete stratification, x_k represents r dummy variables. In the remaining of this article, r should be considered large in comparison with $p \times q$. The population totals for Y and Z are the marginal frequency distributions with respect to education and employment. Using the common variable X, predictions for Y and Z can be defined with a multiple linear regression model:

$$\hat{y}_k = B^t x_k, \quad k = 1, ..., N,$$

and

$$\hat{z}_k = A^t x_k, \quad k = 1, ..., N,$$

where B and A are the ordinary least squares regression coefficients satisfying the normal equations

$$\left(\sum_{k=1}^{N} x_k x_k^t \right) B = \sum_{k=1}^{N} x_k y_k^t \qquad (1)$$

and

$$\left(\sum_{k=1}^{N} x_k x_k^t \right) A = \sum_{k=1}^{N} x_k z_k^t. \qquad (2)$$

The superscript 't' denotes transposition. This model is called a linear probability model, (see Maddala 1983, chap. 2). There are more elegant models, such as probit and logit models, to predict binary variables. However, we are not interested in the predictions themselves, but in the synthetic population totals of these predictions. These totals appear to have nice properties if the linear prediction model is used, and for this reason the model can be justified. Note that B is calculated from the first registration and A from the second one. By means of the common variable X and the regression coefficients B and A, we construct a synthetic registration, which contains a record of each person k with scores x_k, $B^t x_k$, and $A^t x_k$. In fact, either y_k or z_k may be added to this registration, but for our purposes this addition appears to be superfluous (see next paragraph). If there exists a vector a of order r of fixed numbers such that $a^t x_k = 1$ for all k, then the population totals of the new variables $B^t x_k$ and $A^t x_k$ equal the population totals of the corresponding original variables (see *e.g.*, Bethlehem and Keller 1987). This can be shown easily by first pre-multiplying the normal equations (1) and (2) by a^t and subsequently substituting $a^t x_k = 1$ into the resulting equations.

From the synthetic registration, a synthetic two-way table can be defined by $\sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t$. This synthetic two-way table can be considered as an approximation of the (simultaneous) frequency distribution $\sum_{k=1}^{N} y_k z_k^t$. Using the normal equations (1) and (2), the following identities can be derived:

$$\sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t = \sum_{k=1}^{N} y_k (A^t x_k)^t$$

$$= \sum_{k=1}^{N} (B^t x_k) z_k^t.$$

Clearly, the crossings between $B^t x_k$ and $A^t x_k$, y_k and $A^t x_k$, or $B^t x_k$ and z_k, all result in identical synthetic two-way tables. Therefore, it suffices to consider only $\sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t$, and delete either y_k or z_k in the synthetic registration. The difference between the real frequency distribution between Y and Z and its synthetic "approximation", can be obtained from the following decomposition

$$\sum_{k=1}^{N} y_k z_k^t = \sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t + \sum_{k=1}^{N} (y_k - B^t x_k)(z_k - A^t x_k)^t. \qquad (3)$$

Note the strong resemblance with the ordinary variance decomposition in regression analysis (see *e.g.*, Searle 1971). If either $B^t x_k = y_k$ or $A^t x_k = z_k$ for all k, then the two-way table derived from the synthetic registration, equals the real simultaneous frequency distribution between Y and Z.

Let l be a vector of appropriate order consisting of ones, and note that $l^t y_k = 1$ and $l^t z_k = 1$ for all k. If there exists a constant a such that $a^t x_k = 1$ for all k, then we also have

$$l'\hat{y}_k = l'B'x_k = l'\left(\sum_{k=1}^{N} y_k x_k'\right)\left(\sum_{k=1}^{N} x_k x_k'\right)^{-1} x_k =$$

$$a'\left(\sum_{k=1}^{N} x_k x_k'\right)\left(\sum_{k=1}^{N} x_k x_k'\right)^{-1} x_k = a'x_k = 1$$

for all k, and similarly $l'\hat{z}_k = l'A'x_k = 1$ for all k. It follows that

$$l'\sum_{k=1}^{N}(B'x_k)(A'x_k)' = \sum_{k=1}^{N}(A'x_k)' = \sum_{k=1}^{N} z_k' \qquad (4)$$

and

$$\sum_{k=1}^{N}(B'x_k)(A'x_k)'l = \sum_{k=1}^{N}(B'x_k) = \sum_{k=1}^{N} y_k. \qquad (5)$$

So, the row and column totals of the synthetic two-way table, equal the corresponding marginal population counts with respect to Y and Z.

What remains to consider, is the condition $a'x_k = 1$ for all k, for some constant a. This condition is satisfied if X represents a categorical variable. More generally, the condition is always satisfied if the vector X can be partitioned into two sub-vectors, one of which represents a categorical variable.

2.2 Formulating the Constraints in Calibration Estimation

Suppose a probability sample s of size n is drawn from the finite population $\Omega = \{1, ..., N\}$ according to a sampling design $p(s)$ such that the first and second order inclusion probabilities $\Pr(k \in s) = \pi_k$ and $\Pr(k, l \in s) = \pi_{kl}$ are strictly positive. For each $k \in s$ the vector of scores (x_k, y_k, z_k) is observed. Two distinct registrations are available to provide auxiliary information. The first registration contains for each $k \in \Omega$, records with scores on x_k and y_k, the second registration contains for each $k \in \Omega$, scores on x_k and z_k. The objective is to estimate the YZ- table from the sample s, using auxiliary information from both registrations. There exists a wide range of weighting type estimators in the presence of multivariate auxiliary information. In Särndal, Swensson and Wretman (1992), the general regression estimator is extensively discussed. It implicitly defines sample weights, which reproduce the known population totals of the auxiliary variables, used as control variables in the estimator. Such a consistency property is attractive if the auxiliary information is used both for publication and for weighting. As a generalization of the general regression estimator, the calibration estimator is developed (Deville and Särndal 1992 and Deville *et al.* 1993).

To be specific, let G be a real valued function as defined in Deville *et al.* (1993) and consider the following weighting type estimator for our YZ-table:

$$\hat{T} = \sum_{k=1}^{n} w_k(y_k z_k'), \qquad (6)$$

where w_k is a scalar, representing a weight assigned to person $k \in s$. Denote $d_k = \pi_k^{-1}$. A calibration estimator for the YZ-table uses weights which are obtained by minimizing $\sum_{k=1}^{n} d_k G(w_k/d_k)$ with respect to w_k subject to a set of constraints on w_k for any particular sample s. We first consider the following set of constraints:

$$\sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{N} y_k \text{ and } \sum_{k=1}^{n} w_k z_k = \sum_{k=1}^{N} z_k. \qquad (I)$$

This (first) set of constraints only uses the (marginal) counts with respect to Y and Z. No use is made of the common variable X. One of the $p + q$ equations is redundant, so to solve the minimization problem, one equation can be deleted. For $G(w_k/d_k) = (w_k/d_k - 1)^2$, the resulting calibration estimator corresponds to incomplete two-way stratification as defined in Bethlehem and Keller (1987). By taking $G(w_k/d_k) = 1 + w_k/d_k(\log(w_k/d_k) - 1)$, the classical raking ratio estimator is obtained (see *e.g.*, Oh and Scheuren 1987). Copeland, Peitzmeier and Hoy (1987) have compared these methods, based on data of the Current Population Survey. They conclude that the estimates produced by the two methods are very similar. In Deville *et al.* (1993), two other distance functions are discussed, which are especially interesting in view of the problem of extreme weights. Estimating two-way tables with constraints on the marginal counts, is frequently performed in sample surveys. Often, the constraints on the marginal counts are required for two reasons. The first reason is to reduce sampling error and sampling bias, and the second reason is to meet consistency requirements with published population counts.

Suppose that x_k is categorical with r categories. Since population information about the crossings between Y and X, and the crossings between Z and X are available, we may also consider the following set of constraints:

$$\sum_{k=1}^{n} w_k(y_k x_k') = \sum_{k=1}^{n} y_k x_k' \text{ and}$$

$$\sum_{k=1}^{n} w_k(z_k x_k') = \sum_{k=1}^{N} z_k x_k'.$$

The number of non-redundant constraints in this set equals $r(p + q - 1)$. For large r, this set may be not feasible because it contains too many constraints in comparison with

the sample size. Only if r is small, the set may be of practical interest. In the remaining of this article, this set of constraints will be disregarded.

In view of incorporating a large set of common variables in the weighting procedure, we consider a set of constraints, which exploits the bivariate population information that we have in the synthetic table:

$$\sum_{k=1}^{n} w_k (B'x_k)(A'x_k)' = \sum_{k=1}^{N} (B'x_k)(A'x_k)'. \qquad \text{(II)}$$

This (second) set of constraints is a straightforward application of the theory of calibration estimators. Population totals of the crossing between $B'x_k$ and $A'x_k$ are known, so these crossings are taken as auxiliary variables to formulate the set of constraints. Evidently, for large r, the number of non-redundant constraints remains bounded by $p \times q$. A major disadvantage of the resulting calibration weights is that, they do not necessarily reproduce the (marginal) population counts with respect to Y and Z, when applying these weights to y_k and z_k respectively. In other words, the resulting calibration weights do not necessarily satisfy the first set of constraints. Especially, if this set of constraints is formulated in view of consistency requirements, this is a serious drawback.

Therefore, as an alternative, we consider a third set of constraints:

$$\sum_{k=1}^{n} w_k \left(y_k z_k' - (y_k - B'x_k)(z_k - A'x_k)' \right) = \sum_{k=1}^{N} (B'x_k)(A'x_k)' \qquad \text{(III)}$$

Assuming that there exists a constant a, such that $a'x_k = 1$ for all k, this set of constraints meets the consistency objective. Let l denote a vector of ones of appropriate order and recall that $l'y_k = l'B'x_k = l'z_k = l'A'x_k = 1$ for all k, $B'\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} y_k$, and $A'\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} z_k$. By pre-multiplying the third set of equations on both sides with l', we obtain the first set of constraints with respect to Z, and post-multiplying the third set on both sides with l gives the first set of constraints with respect to Y. The resulting calibration estimator can be expressed as

$$\hat{T} = \sum_{k=1}^{n} w_k (y_k z_k') = \sum_{k=1}^{N} (B'x_k)(A'x_k)' + \sum_{k=1}^{N} w_k (y_k - B'x_k)(z_k - A'x_k)'.$$

Clearly, this estimator obeys the decomposition given by (3). It equals the synthetically defined two-way table plus an adjustment term. This adjustment term is a calibration estimate for the difference between the real frequency distribution between Y and Z and the synthetically defined two-way table. Similarly to the second set of constraints, the number of non-redundant constraints in the third set is bounded by $p \times q$.

An important special case is $G(w_k/d_k) = (w_k/d_k - 1)^2$. Then each estimated cell is a general regression estimate with $(y_k z_k)$, $\mathrm{vec}(B'x_k x_k' A)$, and $\mathrm{vec}(y_k z_k' - (y_k - B'x_k)(z_k - A'x_k)')$ as control variables in case of the first, second, and third set of constraints respectively. Analytical formulas for the design variance of the general regression estimator, are given in *e.g.*, Särndal *et al.* (1992, chap. 6). In fact, these formulas are approximations for large sample sizes. In Deville and Särndal (1992), sufficient conditions are given under which these approximations are valid for calibration estimators in general.

In Deville *et al.* (1993), complete post-stratification is described as a calibration method for which all population counts with respect to the cross-classifications, are used in the set of constraints. An elaboration of complete post-stratification, results in the ordinary post-stratification estimator, regardless of the distance function G. As an alternative, incomplete post-stratification is described as a calibration method, in which less detailed than a complete knowledge of all cell counts, is used in the constraint set. The calibration estimator defined under the first set of constraints, is a commonly used example of incomplete post-stratification. Several cases are discussed, in which incomplete post-stratification is preferable to complete post-stratification. Two of them are, lack of population information and, some zero or extremely small cell counts (see also Oh and Scheuren 1987). The calibration estimator defined under the second and third set of constraints, corresponds to complete post-stratification in the sense that, all crossings are used as auxiliary information. Except when a perfect linear relationship exists either between Y and X, or between Z and X, the method differs from complete post-stratification in using synthetic population totals instead of real population counts. Complete post-stratification gives unstable results, if some sample cells have only few observations. In such situations, incomplete post-stratification is of practical interest. Similarly, the calibration estimator under the second and third set of constraints may be unstable. Analogously to incomplete post-stratification, one might consider using an incomplete crossing in the constraints instead.

2.3 A Numerical Illustration

We illustrate the calibration estimator under the three different sets of constraints by means of a hypothetical example. The example is based on real data from a sample on behalf of the Dutch National Travel Survey (1994). The sampling design is roughly a self-weighted cluster sample of addresses. All persons living in a selected address, are included in the sample. The net sample size is approximately 80,000 persons within 34,000 addresses. From this sample, two hypothetical registrations of approximately

N = 80,000 persons are constructed. In the one registration, age is registered (in six categories), and in the other registration, car ownership (in two categories). The common variable between the registrations is a key number for addresses, resulting in r = 34,000 categories for the X-variable. For this particular example the synthetic two-way table simplifies to

$$\sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t = \sum_{j=1}^{r} N_j \bar{y}_j \bar{z}_j^t,$$

where N_j denotes the size of the j-th address, $\bar{y}_j$ the mean of the six age categories of the j-th address, and $\bar{z}_j$ the mean of the two car ownership categories of the j-th address.

In order to calculate the synthetic two-way table, both registrations are combined as follows. Firstly, they are sorted according to the key number for addresses. Secondly, the address counts of the six age categories and the two car ownership categories are calculated. Thirdly, each address count of age, is linked with its corresponding address count of car ownership. By means of this synthetic registration of r = 34,000 addresses, the synthetic two-way table can be calculated. The result is shown in Table 1. This table can be considered as a first approximation of the real frequency distribution between age and car ownership. A sufficient condition for a close approximation, is homogeneity with respect to either age or car ownership within all addresses, *i.e.*, all persons at the same address should either be in the same age category or in the same car ownership category. For most (multiple) person addresses, this seems to be an unlikely proposition. It follows from equations (4) and (5) that the row and column totals in table 1 coincide with the real (marginal) population counts of age and car ownership respectively.

By means of a simple random sample of n = 1000 persons, the population cell counts are estimated using a general regression estimator. Three sets of auxiliary variables are used, in accordance with the three sets of constraints mentioned in the previous section. The estimated tables are given below (for convenience we have taken the quadratic distance measure: $G(w_k/d_k) = (w_k/d_k - 1)^2$). The corresponding estimated standard deviations are within parenthesis. These estimates are based on the usual variance formulas of the general regression estimator, see Särndal *et al.* (1992, chap. 6).

Table 1
Synthetic Population Totals for Crossings Between Age and Car Ownership

	1	2	3	4	5	6	total
yes	3461	1659	5739	10770	6536	3334	31499
no	9827	4692	7902	17102	6424	5389	51336
total	13288	6351	13641	27872	12960	8723	82835

Table 2
Estimated Population Totals for Crossings Between Age and Car Ownership, Satisfying the First Set of Constraints

	1	2	3	4	5	6	total
yes	0 (0)	0 (0)	4968 (423)	15414 (543)	7518 (458)	3599 (375)	31499
no	13288 (0)	6351 (0)	8673 (423)	12458 (543)	5422 (458)	5124 (375)	51336
total	13288	6351	13641	27872	12960	8723	82835

Table 3
Estimated Population Totals for Crossings Between Age and Car Ownership, Satisfying the Second Set of Constraints

	1	2	3	4	5	6	total
yes	0 (0)	0 (0)	4791 (435)	13826 (811)	6887 (494)	3421 (321)	28923 (1005)
no	14385 (782)	7012 (595)	8118 (563)	12893 (796)	5853 (464)	5654 (306)	53912 (1005)
total	14385 (782)	7012 (595)	12908 (603)	26718 (958)	12739 (419)	9074 (177)	82835

Table 4
Estimated Population Totals for Crossings Between Age and Car Ownership, Satisfying the Third Set of Constraints

	1	2	3	4	5	6	total
yes	0 (0)	0 (0)	5501 (226)	15647 (227)	6898 (177)	3453 (78)	31499
no	13288 (0)	6351 (0)	8139 (226)	12224 (227)	6062 (177)	5270 (78)	51336
total	13288	6351	13641	27872	12960	8723	82835

In Table 2 the population counts are estimated according to the ordinary incomplete two-way stratification (Bethlehem and Keller 1987). There are no young people (age category 1 and 2) owning a car, observed in the sample, which is likely to be representative for the population, so these cells are estimated by zero. Due to the consistency requirements, *i.e.*, the first set of constraints, it follows that the estimated cell counts of young people without a car equal the corresponding marginal cell counts. An attempt to improve Table 2, is to use the common variable address in the weighting procedure. In Table 3, the cell estimates are given according to the second set of constraints. As already mentioned in the previous section, the estimated row and column totals may differ from the real population counts. A comparison between Table 2 and Table 3 shows that these differences can be considerable. In addition, almost all estimated cell counts in Table 2 have smaller estimated standard deviations than the corresponding estimated cell counts in Table 3. So, the second set of constraints gives quite unsatisfactory results. The third set of constraints covers the first set of constraints. This implies 1) consistency of the estimated marginal cell counts with respect to the corresponding known population cell counts, and 2) smaller asymptotic variances of all estimated cell counts. The results are shown in Table 4. Indeed, the estimated marginal cell counts are consistent, and the estimated standard deviations are at most half of the corresponding standard estimates given in Table 2.

2.4 Imputing Values of the one Registration into the Other Registration

Until now, we have developed a weighting method to estimate a two-way table between two variables, which are registered in two distinct registrations. Often, one is interested not only in estimated two-way tables, or more generally, estimated linear relations, but in complete registrations in which both variables are simultaneously registered. Users of statistics find such complete data-bases easy to analyze. The creation of such enriched registrations can be seen as a special case of imputation. One registration serves as a host or recipient source, and the other as a donor source. Assuming the second registration to be the donor source, the problem is imputing Z-values from the second registration, into the first registration using the estimated two-way table discussed in Section 2.2, as auxiliary information. Statistical matching problems using data from a third data source, have already been considered by Rubin (1986) and Paass (1986). Singh *et al.* (1993) gives a review of their methods. In addition, they propose some modifications to Rubin's (1986) and Paass's (1986) methods. Our imputation method is based on the regression method suggested by Rubin (1986) and Singh *et al.* (1993).

After having defined predictors for the Z-variables by means of the regression model

$$\hat{z}_k = A^t x_k, \; k = 1, ..., N,$$

where A is given by (2), we define new predictions for these variables by means of the enlarged regression model

$$\tilde{z}_k = \alpha_1^t x_k + \alpha_2^t y_k, \; k = 1, ..., N,$$

with

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \left[\sum_{k=1}^{N} \begin{pmatrix} x_k x_k^t & x_k y_k^t \\ y_k x_k^t & y_k y_k^t \end{pmatrix}^{-1} \right] \left[\sum_{k=1}^{N} \begin{pmatrix} x_k z_k^t \\ y_k z_x^t \end{pmatrix} \right].$$

Using well-known results about partial regression coefficients in the general linear model (see *e.g.*, Seber 1977), α_1 and α_2 can be expressed as

$$\alpha_1 = A - B\alpha_2$$

and

$$\alpha_2 = \left[\sum_{k=1}^{N} (y_k - B^t x_k)(y_k - B^t x_k)^t \right]^{-1} \times \left[\sum_{k=1}^{N} (y_k - B^t x_k)(z_k - A^t x_k)^t \right],$$

where B and A are given by (1) and (2) respectively. They can be calculated from the first and second registration. The partial regression coefficients should be estimated from the third source. We suggest

$$\hat{\alpha}_1 = A - B\hat{\alpha}_2$$

and

$$\hat{\alpha}_2 = \left[\sum_{k=1}^{N} (y_k - B^t x_k)(y_k - B^t x_k)^t \right]^{-1} \times \left[\sum_{k=1}^{n} w_k (y_k - B^t x_k)(z_k - A^t x_k)^t \right],$$

where w_k are calibration weights which are discussed in Section 2.2. Based on these estimates we define new predictions for the Z-values:

$$\hat{\hat{z}}_k = \hat{\alpha}_1^t x_k + \hat{\alpha}_2^t y_k = A^t x_k + \hat{\alpha}_2^t (y_k - B^t x_k), \; k = 1, ..., N. \quad (7)$$

These new predictions equal the old predictions (see Section 2.1) plus an adjustment term. This adjustment term depends on the difference between the Y-value and its (old) prediction. It can be viewed as an attempt to improve the prediction for Z, however, and more important, it is a means to reconstruct the weighting type estimator under the third set of constraints (Section 2.2). Indeed, the following equality holds:

$$\sum_{k=1}^{N} y_k \hat{\hat{z}}_k^t = \sum_{k=1}^{N} (B^t x_k)(A^t x_k)^t + \sum_{k=1}^{n} w_k (y_k - B^t x_k)(z_k - A^t x_k)^t.$$

This is just the weighting type estimator under the third set of constraints, if the corresponding calibration weights are used to estimate α_2. It is easy to show that

$$\sum_{k=1}^{N} x_k \hat{\hat{z}}_k^t = \sum_{k=1}^{N} x_k \hat{z}_k^t = \sum_{k=1}^{N} x_k z_k^t.$$

So, also the XZ-table can be reconstructed. At the beginning of this section, we assumed the second registration to be the donor source. This choice was arbitrary. If the Y-values were imputed instead of the Z-values, we would have obtained an identical estimate for the YZ-table. In addition, the XY-table could have been reconstructed.

The new predictions for the Z-values can be used for imputation. Singh *et al.* (1993) give algorithms for imputation using regression models. These Z-values can be imputed in the first registration in two steps. In the first step, the predictions given by (7) are calculated for each

(x_k, y_k) in the first registration. We have shown that the crossings between the Y-values and these predicted Z-values, can be considered as weighting type estimators. However, the calculated predictions have in general no realistic values, and therefore the first step is followed by a second step. In the second step, each predicted Z-value in the first registration is replaced by a live Z-value from the second registration, which is nearest under some Euclidean distance in (X, Z).

2.5 Estimating Cross-Products for Continuous *Y*- and *Z*-Variables

The consistency property of the third set of constraints (Section 2.2) also hold with respect to continuous Y- and Z-variables, provided that there exist constants a_y and a_z of proper order, such that $a_y^t y_k = 1$ and $a_z^t z_k = 1$ for all k. To see this, we slightly extend the results of Section 2.1. First note that

$$a_y^t B^t x_k = a_y^t \sum_{k=1}^{N} y_k x_k^t \left(\sum_{k=1}^{N} x_k x_k^t \right)^{-1} x_k =$$

$$a^t \sum_{k=1}^{N} x_k x_k^t \left(\sum_{k=1}^{N} x_k x_k^t \right)^{-1} x_k = a^t x_k = 1$$

(it is still assumed that there exists a constant a such that $a^t x_k = 1$ for all k). Similarly, it holds that $a_z^t A^t x_k = 1$. The equivalent equations of (4) and (5) for the continuous case are readily obtained. Consequently, pre-multiplying both sides of (III) with a_y^t gives $\sum_{k=1}^{n} w_k z_k^t = \sum_{k=1}^{N} z_k^t$ and post-multiplying both sides of (III) with a_z yields $\sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{N} y_k$. So, the third set of constraints meets the consistency objective, *i.e.*, the calibration equation of the first set of constraints, for quite general Y- and Z-variables. We will give two examples.

In the first example we take $y_k = (1, y_{2k})^t$ and $z_k = (1, z_{2k})^t$, where both y_{2k} and z_{2k} are assumed to be continuous. By taking $a_y = a_z = (1, 0)^t$ we see that $a_y^t y = a_z^t z = 1$ for all k. The cross-product between Y and Z equals

$$\sum_{k=1}^{N} y_k z_k^t = \begin{pmatrix} N & \sum_{k=1}^{N} z_{2k} \\ \sum_{k=1}^{N} y_{2k} & \sum_{k=1}^{N} y_{2k} z_{2k} \end{pmatrix},$$

from which the covariance between y_{2k} and z_{2k} is easily derived. This cross-product can be estimated using the third set of constraints. An elaboration of this set gives the following four constraints for this particular example:

$$\sum_{k=1}^{n} w_k = N, \sum_{k=1}^{n} w_k y_{2k} = \sum_{k=1}^{N} y_{2k}, \sum_{k=1}^{n} w_k z_{2k} = \sum_{k=1}^{N} z_{2k},$$

and

$$\sum_{k=1}^{n} w_k (y_{2k} z_{2k} - (y_{2k} - B_2^t x_k)(z_{2k} - A_2^t x_k)) = \sum_{k=1}^{N} (B_2^t x_k)(A_2^t x_k),$$

where the regression coefficients are given by

$$B_2 = \left(\sum_{k=1}^{N} x_k x_k^t \right)^{-1} \sum_{k=1}^{N} x_k y_{2k}$$

and

$$A_2 = \left(\sum_{k=1}^{N} x_k x_k^t \right)^{-1} \sum_{k=1}^{N} x_k z_{2k}.$$

If one is specially interested in the correlation coefficient between y_{2k} and z_{2k}, then following constraints may be considered in addition:

$$\sum_{k=1}^{n} w_k y_{2k}^2 = \sum_{k=1}^{N} y_{2k}^2 \text{ and } \sum_{k=1}^{n} w_k z_{2k}^2 = \sum_{k=1}^{N} z_{2k}^2.$$

In the second example, we suppose that $y_k = (1, y_{2k})^t$, where y_{2k} may be continuous, and z_k is categorical with q categories. By taking $a_y = (1, 0)^t$ and $a_z = l$, where l is a vector of ones of proper order, we see that $a_y^t y_k = a_z^t z_k = 1$ for all k. The cross-product between Y and Z is

$$\sum_{k=1}^{N} y_k z_k^t = \begin{pmatrix} N_1 & N_2 & \cdots & N_q \\ \sum_{k \in C_1} y_{2k} & \sum_{k \in C_2} y_{2k} & \cdots & \sum_{k \in C_q} y_{2k} \end{pmatrix},$$

where C_h denotes the set of population elements belonging to the h-th category of Z, and N_h the size of C_h. It is ensured that the calibration weights according to the third set of constraints, satisfy the 'marginal' calibration equations $\sum_{k=1}^{n} w_k z_k = \sum_{k=1}^{N} z_k = (N_1 \ \ldots \ N_q)^t$ and $\sum_{k=1}^{n} w_k y_{2k} = \sum_{k=1}^{N} y_{2k}$, which both may be of interest in view of consistency requirements.

3. COMBINING INDEPENDENT SAMPLES ACROSS COMMON VARIABLES

In the previous section, we have presented a method for combining two registrations across common variables, using auxiliary information from a small sample. In this section, the method is adjusted by combining two independent samples. We consider a complete registration of persons, two large-scale sample surveys, and a small-scale sample survey. The registration contains a limited set of variables such as sex, age, region, and marital status. These

variables are denoted by X. In the one large sample, the variables Y, U, and X are observed, and in the other large sample, the variables Z, U, and X. In the small sample all variables, *i.e.*, Y, Z, U, and X, are observed. The small sample could come from a specially conducted small-scale survey, or from sample overlap of the large-scale surveys. In Figure 1, the data sources are schematically given. For convenience, it is assumed that all samples correspond to different units, *i.e.*, it is assumed that there is no sample overlap.

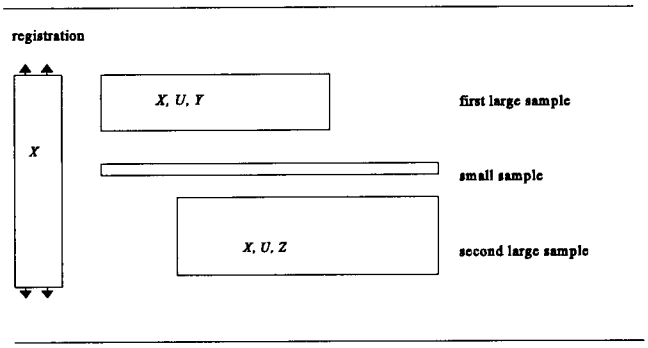

Figure 1. Overview of the Several Data Sources

The common variables X and U are partitioned into $C = (X\ U)$, where X denotes the set of common variables with known population totals, and U denotes the set of common variables with unknown population totals. All samples may be drawn by some complex sampling design. Both Y and Z are assumed to be categorical, however, as in Section 2.5, the suggested weighting methods are also applicable for continuous Y and Z. The purpose is to estimate the two-way table between Y and Z. We consider two estimators. One estimator is based on incomplete two-way stratification (analogous to the first set of constraints of Section 2.2), and the other estimator is based on a mix between statistical matching and calibration (analogous to the third set of constraints of Section 2.2).

3.1 Incomplete Two-Way Stratification

First the population totals of Y and Z are estimated by means of the first and second (large) sample respectively. These population totals are estimated in two phases. In the first phase, both (large) samples are weighted using X as a set of control variables. This implies that both (large) samples are weighted such that they reproduce the known population totals of X, which are denoted by t_x. Based on these weights, a pooled estimate for the population totals of U is

$$\hat{t}_u = \lambda \sum_{k \in n_1} w_{1k} u_k + (1 - \lambda) \sum_{k \in n_2} w_{2k} u_k,$$

where w_{1k} and w_{2k} denote the (first phase) calibration weights of the first and second sample, and $\lambda \in [0, 1]$. In the second phase, both samples are reweighted using simultaneously X and U as control variables. Let v_{1k} and v_{2k} denote these second phase calibration weights. The resulting estimators for the population totals of Y and Z can be considered as calibration estimators in two phases (see Renssen and Nieuwenbroek 1997, Section 6). These estimators are denoted by $\hat{t}_y$ and $\hat{t}_z$ respectively:

$$\hat{t}_y = \sum_{k \in n_1} v_{1k} y_k \text{ and } \hat{t}_z = \sum_{k \in n_2} v_{2k} z_k.$$

We note that both estimators are based on a similar set of control variables. If the common set of variables is large, one may consider using a smaller subset to weight both samples. In general, the subset to weight the first sample may differ from the subset to weight the second sample. However, we shall assume in the sequel that both (large) samples are weighted according to the same set of control variables.

The two-way table between Y and Z can be estimated by weighting the (small) third sample, using simultaneously Y and Z as control variables, *i.e.*,

$$\hat{T} = \sum_{k \in n_3} w_{3k} (y_k z_k^t),$$

where the calibration weights w_{3k} satisfy the constraints

$$\sum_{k \in n_3} w_{3k} y_k = \hat{t}_y \text{ and } \sum_{k \in n_3} w_{3k} z_k = \hat{t}_z.$$

This is incomplete two-way stratification, where the unknown population totals of Y and Z are replaced by their estimates. These sets of constraints ensure precisely estimated marginal counts of the YZ-table if the common variables C are highly correlated with Y and Z.

3.2 Synthetic Two-Way Stratification

In this section, we consider an alternative estimator for the YZ-table, which also uses the (large) samples as a source of auxiliary information. However, instead of using estimated marginal counts as auxiliary information, estimated synthetic cell counts are used. Let B denote the population regression coefficient between Y and C, which is estimated by the first (large) sample:

$$\hat{B} = \left(\sum_{k \in n_1} v_{1k} c_k c_k^t \right)^{-1} \left(\sum_{k \in n_1} v_{1k} c_k y_k^t \right).$$

Similarly, let A denote the population regression coefficient between Z and C, which is estimated by the second (large) sample:

$$\hat{A} = \left(\sum_{k \in n_2} v_{2k} c_k c_k^t \right)^{-1} \left(\sum_{k \in n_2} v_{2k} c_k z_k^t \right).$$

Note that these estimated regression coefficients are based on the second phase calibration weights instead of the inclusion weights. If there exists a constant a, such that $a^t c_k = 1$ for all k, then we still have $l^t \hat{B}^t c_k = l^t \hat{A}^t c_k = 1$ for all k. Now, inspired by the decomposition given by (3), *i.e.*,

$$\sum_{k=1}^{N} y_k z_k^t = B^t \sum_{k=1}^{N} (c_k c_k^t) A + \sum_{k=1}^{N} (y_k - B^t c_k)(z_k - A^t c_k)^t,$$

we suggest estimating the two-way table in two steps. In the first step the first term on the right-hand side is estimated by substituting the population regression coefficients B and A by their estimates $\hat{B}$ and $\hat{A}$. Furthermore, we suggest to estimate $\sum_c = \sum_{k=1}^{N} c_k c_k^t$ by the pooled estimate:

$$\hat{\sum}_c = \gamma \sum_{k \in n_1} v_{1k} (c_k c_k^t) + (1 - \gamma) \sum_{k \in n_2} v_{2k} (c_k c_k^t),$$

where v_{1k} and v_{2k} denote the (second phase) weights of the first and second sample and $\gamma \in [0, 1]$. Eventually, the first term is estimated by $\hat{B}^t \hat{\sum}_c \hat{A}$. Until now, no use of the third (small) sample has been made. If desired, estimates for B, A, and $\sum_c$ can be improved slightly by also using the small sample.

In the second step, the complete two-way table between Y and Z is estimated by weighting the third (small) sample according to the calibration estimator subject to the third set of constraints (see Section 2.2), where B, A, and $\sum_c$ are replaced by their estimates $\hat{B}$, $\hat{A}$, and $\hat{\sum}_c$. The resulting estimator equals

$$\sum_{k=1}^{n_3} w_3 (y_k z_k^t) = \hat{B}^t \hat{\sum}_c \hat{A} + \sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}^t c_k)(z_k - \hat{A}^t c_k)^t. \quad (8)$$

The first term on the right-hand side is an estimate for the synthetic two-way table. This estimate is approximately unbiased for the YZ-table, if the conditional independence assumption holds. We note that, this type of estimator is essentially obtained by applying the constrained statistical matching method (see *e.g.*, Barr and Turner 1980, Rodgers 1984, or Rubin 1986). The second term is an adjustment term to obtain an approximately unbiased estimate for the YZ-table, without this assumption. If there exists a constant a such that $a^t c_k = 1$ for all sampled elements, then we obtain by pre-multiplying both sides of (8) with l^t, the following estimator for the population total of Z:

$$\sum_{k \in n_3} w_{3k} z_k^t = \left(\gamma \sum_{k \in n_1} v_{1k} c_k^t + (1 - \gamma) \sum_{k \in n_2} v_{2k} c_k^t \right) \hat{A} = \left(\sum_{k \in n_2} v_{2k} c_k^t \right) \hat{A} = a^t \left(\sum_{k \in n_2} v_{2k} c_k c_k^t \right) \hat{A} = \hat{t}_z^t.$$

Similarly, we have by post-multiplying both sides with l, an estimator for the population total of Y:

$$\sum_{k \in n_3} w_{3k} y_k = \hat{B}^t \left(\gamma \sum_{k \in n_1} v_{1k} c_k + (1 - \gamma) \sum_{k \in n_2} v_{2k} c_k \right) = \hat{B}^t \left(\sum_{k \in n_1} v_{1k} c_k \right) = \hat{B}^t \left(\sum_{k \in n_1} v_{1k} c_k c_k^t \right) a = \hat{t}_y.$$

It follows that the marginal cell counts of the estimated two-way table, are the two-phase calibration estimators for the population totals of Y and Z as defined Section 3.1.

3.3 A Simulation Study; Integration of Household Surveys

In this subsection, we wish to compare the weighting techniques incomplete two-way stratification as discussed in subsection 3.1, and synthetic two-way stratification as discussed in subsection 3.2, by means of a simulation study. To that purpose, we use a data set, which stems from a pilot study of the Dutch Household Survey on Living Conditions, (see van Tuinen 1995). The data set consists of 1,085 records of which the following variables are observed: age (six categories: 15-24, 25-34, 35-44, 45-54, 55-64, 65+), sex (two categories: male or female), ownership of house (two categories: yes or no), occupation (five categories: work, housekeeping, education, voluntary, other), and health (two categories: yes or no). On behalf of the simulation study, this data set is considered as a finite population. The population totals of age and sex are assumed to be known.

In order to simulate the weighting techniques, we have carried out a Monte Carlo algorithm. Namely, we have drawn 500 samples, independently of each other, according to a two-phase sampling design. In the first phase, a simple random sample of size 20,500 is drawn with replacement. In this sample, age, sex, and ownership of house, are observed. In the second phase, the (first phase) sample is randomly divided into two large sub-samples of sizes 10,000 and one small sub-sample of size 500; in the one large sub-sample, occupation is observed (denoted by Y), in the other large sub-sample, health (denoted by Z), and in the small sub-sample, both occupation and health are observed. At each run, we have estimated the two-way table between Y and Z, according to four weighting methods which are discussed next.

The first phase sample is weighted with a crossing between sex and age as control variables. This is just post-stratification with twelve post-strata. Based on these weights, population totals can be estimated for all observed variables in the first phase sample, and for crossings between them. In particular, we may reproduce the population totals for the crossing between age and sex, and obtain estimated population totals for the crossings between age, sex, and ownership of house. Now, we distinguish two sets of common variables to weight the large sub-samples, as well as to obtain an estimate for the synthetic two-way table between Y and Z. The first set is a crossing between age and sex (12 categories) and the second set is a crossing between age, sex, and ownership (24 categories). For each simulation, this gives two different estimates for the marginal counts, *i.e.* two different estimates for the population totals of Y and Z – note that both estimates are based on post-stratification – and two different estimates for the synthetic two-way table. In order to weight the small sub-sample, we distinguish between the weighting method based on incomplete two-way stratification, and the weighting method based on synthetic two-way stratification. Since two different sets of common variables are used to weight the large sub-samples, as well as for statistical matching, we obtain four sets of calibration weights for each simulation run with respect to the small sub-sample, which in turn gives for each simulation run, four different estimated two-way tables between Y and Z. For the ease of computation, we have used the quadratic distance measure in the calibration estimation, implying that each estimated cell corresponds to a general regression estimate. Finally, we have taken the averages and variances of these two-way tables over the 500 simulations. The results are shown in tables 5 to 8.

The averages over the 500 simulations are almost identical for the four types of estimators, as can be seen from these tables. Note that the given cell counts are rounded off. We have also calculated the real YZ-table from the finite population. The real counts equal exactly the averages, which are given in Table 5 (or 6). For this particular simulation study, we conclude that all estimators have a very small bias.

The variances over these 500 simulations are given within the brackets. The variances of the estimated marginal counts of Tables 5 and 7 coincide, because these estimates are based on the same estimator. For the same reason it holds that the variances of the estimated marginal counts in tables 6 and 8 coincide. Note that the variances of the estimated marginal counts in tables 6 and 8 are slightly smaller than the variances of the estimated marginal counts in Tables 5 and 7, due to the larger set of common variables. However, for most estimated marginal counts this variance reduction can be considered negligible.

Tables 5 and 6 give identical variances with respect to all estimated cell counts. The variances for most estimated cell counts in Table 7, are plainly smaller than those in tables 5 and 6. In Table 8, this variance reduction is even greater. For this particular example, we conclude that the use of the larger set of common variables, in combination with the first weighting method, slightly reduces the variances of the estimated marginal counts, but leaves the variances of the estimated cell counts unaffected. Naturally, using the larger set of common variables in combination with the second weighting method, also slightly reduces the variances of the marginal cell counts. Finally, given a set of common variables, the weighting method based on synthetic matching, results in smaller variances for the estimated cell counts, than the weighting method based on incomplete two-way stratification.

Table 5
Incomplete Two-way Stratification Combined with the First Set of Common Variables

	1	2	3	4	5	total
yes	$447_{(96)}$	$232_{(97)}$	$89_{(28)}$	$25_{(21)}$	$59_{(49)}$	$852_{(17)}$
no	$61_{(79)}$	$104_{(90)}$	$11_{(21)}$	$11_{(19)}$	$46_{(46)}$	$233_{(17)}$
total	$508_{(23)}$	$336_{(19)}$	$100_{(8)}$	$36_{(3)}$	$105_{(10)}$	1085

Table 6
Incomplete Two-way Stratification Combined with the Second Set of Common Variables

	1	2	3	4	5	total
yes	$447_{(96)}$	$232_{(97)}$	$89_{(28)}$	$25_{(21)}$	$59_{(49)}$	$852_{(17)}$
no	$61_{(79)}$	$104_{(90)}$	$11_{(21)}$	$11_{(19)}$	$46_{(46)}$	$233_{(17)}$
total	$508_{(23)}$	$336_{(19)}$	$100_{(8)}$	$36_{(3)}$	$105_{(9)}$	1085

Table 7
Synthetic Two-way Stratification Combined with the First Set of Common Variables

	1	2	3	4	5	total
yes	$447_{(75)}$	$231_{(74)}$	$89_{(17)}$	$25_{(20)}$	$59_{(42)}$	$851_{(17)}$
no	$61_{(58)}$	$105_{(65)}$	$11_{(12)}$	$11_{(19)}$	$46_{(38)}$	$234_{(17)}$
total	$508_{(23)}$	$336_{(19)}$	$100_{(8)}$	$36_{(3)}$	$105_{(10)}$	1085

Table 8
Synthetic Two-way Stratification Combined with the Second Set of Common Variables

	1	2	3	4	5	total
yes	$447_{(70)}$	$231_{(70)}$	$89_{(16)}$	$25_{(18)}$	$59_{(40)}$	$851_{(17)}$
no	$61_{(52)}$	$105_{(60)}$	$11_{(11)}$	$11_{(16)}$	$46_{(37)}$	$234_{(17)}$
total	$508_{(23)}$	$336_{(19)}$	$100_{(8)}$	$36_{(3)}$	$105_{(9)}$	1085

3.4 Imputing Values of the one Large Sample into the Other Large Sample

By means of the two large samples and the small sample, one may construct a synthetic sample in which the real Y-values and predicted Z-values, and/or the predicted Y-values and the real Z-values are simultaneously recorded.

We define predictions for the Y- and Z-values analogously to (7), namely

$$\hat{y}_k = \hat{B}^t c_k + \tilde{\beta}_2^t \ (z_k - \hat{A}^t c_k), k = 1, ..., n_2, \qquad (9)$$

and

$$\hat{z}_k = \hat{A}^t c_k + \tilde{\alpha}_2^t \ (y_k - \hat{B}^t c_k), k = 1, ..., n_1, \qquad (10)$$

with

$$\tilde{\beta}_2 = \left[\sum_{k=1}^{n_2} v_{2k}(z_k - \hat{A}^t c_k)(z_k - \hat{A}^t c_k)^t\right]^{-1} \times \left[\sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}^t c_k)(z_k - \hat{A}^t c_k)^t\right],$$

and

$$\tilde{\alpha}_2 = \left[\sum_{k=1}^{n_1} v_{1k}(y_k - \hat{B}^t c_k)(y_k - \hat{B}^t c_k)^t\right]^{-1} \times \left[\sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}^t c_k)(z_k - \hat{A}^t c_k)^t\right].$$

For each (c_k, y_k) the Z-values can be imputed in the first large sample by means of (10), $k = 1, ..., n_1$, and similarly for each (c_k, z_k) the Y-values can be imputed in the second large sample by means of (9), $k = 1, ..., n_2$. Based on these imputed values, we may define the following estimates for the two-way table between Y and Z:

$$\sum_{k=1}^{n_1} v_{1k} y_k \hat{z}_k^t = \hat{B}^t \sum_{k=1}^{n_1} v_{1k} c_k c_k^t \hat{A} + \sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}^t c_k)(z_k - \hat{A}^t c_k)^t \qquad (11)$$

and

$$\sum_{k=1}^{n_2} v_{2k} \hat{y}_k z_k^t = \hat{B}^t \sum_{k=1}^{n_2} v_{2k} c_k c_k^t \hat{A} + \sum_{k=1}^{n_3} w_{3k}(y_k - \hat{B}^t c_k)(z_k - \hat{A}^t c_k)^t. \qquad (12)$$

One estimate is based on the first synthetic sample, the other on the second synthetic sample. By pooling the synthetic samples, one obtains a pooled synthetic sample of size $n_1 + n_2$, from which a pooled estimated for the two-way table can be constructed. This pooled estimate shows a close resemblance to (8). Note that if C and Z are perfectly correlated, then the left-hand side of (11) reduces to $\sum_{k=1}^{n_1} v_{1k} y_k z_k^t$, *i.e.*, our estimated two-way table corresponds to a weighted estimated two-way table based on the first sample, as if the real values of Z were imputed in this sample. Similarly, if C and Y are perfectly correlated, then (12) reduces to $\sum_{k=1}^{n_2} v_{2k} y_k z_k^t$.

An important special case to consider, is when c is categorical. Then the following equalities hold true:

$$\sum_{k \in n_1} v_{1k}(c_k c_k^t) = \sum_{k \in n_2} v_{2k}(c_k c_k^t) = \text{diag}\begin{pmatrix} t_x \\ \hat{t}_u \end{pmatrix},$$

so (11) and (12) coincide. Furthermore, we have for categorical c:

$$\sum_{k \in n_1} v_{1k} c_k \hat{z}_k^t = \sum_{k \in n_2} v_{2k} c_k z_k^t$$

and

$$\sum_{k \in n_2} v_{2k} \hat{y}_k c_k^t = \sum_{k \in n_1} v_{1k} y_k c_k^t.$$

Obviously, if c is categorical, then it suffices to create a synthetic sample, which is based on either the first synthetic sample or the second synthetic sample. In either case, the weighting type estimates for the CZ-table, the CY-table, and the YZ-table, can be reconstructed. Finally, we note that the imputed values in all synthetic samples may be unrealistic. As described in Section 2.4, the calculated predictions may be replaced by live values according to some algorithm.

4. SUMMARY

In this article we presented a weighting procedure to combine information from distinct sample surveys. The linking pin between these surveys, is a set of common variables, (see Figure 1). It is argued that these samples should be weighted according to a sequential structure. First, both large samples were weighted using X as control variables. Based on these weighted samples, we could obtain a pooled estimate for the population total of U. Then both large samples were reweighted using simultaneously X and U as control variables. This gave an estimate for the population total of Y and Z.

Using statistical matching techniques with X and U as common variables, we may also obtain an estimate for a synthetic two-way table between Y and Z. Eventually, the small sample was weighted according to two different sets of control variables. The first set of control variables corresponded to the estimated population totals of Y and Z, and the second set of control variables to the estimated synthetic two-way table. Using the first set of control variables, is strongly related to incomplete two-way stratification. The theoretical framework needed to develop the second weighting method, was discussed all through this article. By means of both weighting methods, the

YZ-table can be estimated (it is tacitly assumed that *Y* and *Z* are categorical). The marginal counts of the *YZ*-table corresponding to the first weighting method, equal by definition of the calibration equations, the estimated population totals of *Y* (which is based on the first large sample) and *Z* (which is based on the second large sample). It was shown, that this consistency property also holds for the second weighting method. A numerical study was conducted to evaluate the performance of the weighting methods with respect to the cell counts. It was found that both weighting methods yielded nearly (design) unbiased estimated two-way tables. The simulated (design) variances of the second weighting method, appeared to be smaller than the corresponding (design) variances of the first weighting method, with respect to all estimated cell counts. In principle, the *Y*- and *Z*-variables were assumed to be categorical, however, it was argued that the ideas presented were also applicable for continuous *Y* and *Z* or for continuous *Y* and categorical *Z*.

ACKNOWLEDGEMENTS

The author wishes to thank Peter Kooiman, Nico Nieuwenbroek, and Ger Slootbeek for their careful reading and useful remarks. The author also thanks two anonymous referees and an associated editor for their valuable suggestions to improve the article. The views expressed in this article are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

BAKKER, B.F.M., and WINKELS, J.W. (1998). Why integration of household surveys? – Why POLS?. *Netherlands Official Statistics*, 13, 5-7.

BARR, R.S., and TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Survey. Technical report, U.S. Department of the Treasury, Office of Tax Analysis.

BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

COPELAND, K.R., PEITZMEIER, F.K., and HOY, C.E. (1987). An alternative method of controlling Current Population Survey estimates to population counts. *Survey Methodology*, 13, 173-181.

DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.C., SÄRNDAL, C.-E., and SAUTORY, O., (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HOFMANS, M.G. (1998). Innovative weighting in POLS. Making use of core questions. *Netherlands Official Statistics*, 13, 12-15.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-208.

MADDALA, G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

OH, H.L., and SCHEUREN, F. (1987). Modified raking ratio estimation. *Survey Methodology*, 13, 209-219.

PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: Elsevier Science.

RAGHUNATHAN, T.E., and GRIZZLE, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.

RENSSEN, R.H., and NIEUWENBROEK, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.

RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2, 91-102.

RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4, 87-94.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

SEARLE, S.R. (1971). *Linear Models*. New York: John Wiley & Sons.

SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons.

SINGH, A.C., MANTEL, H.J., KINACK, M.D, and ROWE, R. (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19, 59-79.

TUINEN VAN, H.K. (1995). Social indicators, social surveys and integration of social statistics. *Statistical Journal of the United Nations ECE*, 12, 379-394.

WINKELS, J.W., and EVERAERS, P.C.J. (1998). Design of an integrated survey in the Netherlands. The case POLS. *Netherlands Official Statistics*, 13, 6-11.

ZIESCHANG, K.D. (1990). A generalized least squares weighting system for the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.