Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

ALIGNING ESTIMATES BY REPEATED WEIGHTING

(shortened version)

R.H. Renssen, A.H. Kroese, and A. J. Willeboordse[1]

ABSTRACT

At Statistics Netherlands the design and organization of the statistical process is changing rapidly. This change is motivated by the need to produce more consistent data and by political pressures to cut down the response burden. The idea behind the new production process is to integrate all survey and administrative data into a limited number of micro databases and to develop an estimation strategy for these databases. This paper gives the initial impetus of an estimation strategy per micro database. The proposed strategy ensures that all estimated *m*-way tables are numerically consistent with respect to common margins, even if these tables are estimated from different surveys. It is still based on the calibration principle, however, not necessarily on a fixed set of weights per survey. The practicability of the strategy is tested by means of a fictitious example.

Key Words: Calibration estimators, Consistency, Frequency distributions, General regression estimators.

1. INTRODUCTION

Traditionally, statistical bureaus organize their data collection, -processing, and -dissemination according to a stovepipe model, i.e. many different surveys are carried out more or less independently of each other, while each survey has its own way of processing. There are several reasons why such an approach is unsatisfactory. Firstly, statistical data may be incomparable due to lack of coherence between the various surveys[2]. Secondly, in order to limit the response burden, providers of information should be questioned as few as possible, while the opposite may occur when using the stovepipe model. Thirdly, the accuracy of the estimates may be unnecessary low if in the estimation strategy no or less use is made of supplementary registrations and/or other surveys.

In order to cope with these disadvantages, Statistics Netherlands decided to reorganize its production processes drastically, see e.g. Willeboordse (2000), Al and Bakker (2000), and Laan van der (2000). The idea is to integrate all primary and secondary micro data sources into a limited number of micro databases and to develop an efficient estimation strategy, such that all estimates that are presented as *m*-way tables are numerically consistent, by which we mean that no contradictions may occur when comparing two or more tables, even not on account of sampling error.

Currently, Statistics Netherlands distinguishes between several micro databases, the micro databases for persons and the micro database for businesses being the most important ones. Somewhat simplified, the micro database for persons consists of the Municipal Base Administration as the backbone with sample surveys about persons and registers about persons matched to this. Similarly, the micro database for businesses consists of the General Business Register as the backbone with business surveys and registers about businesses matched to this. Typically, micro databases can be seen as rectangular arrays with individual (statistical) units in its rows and scores for variables in its col-

[1] R.H. Renssen, A.H. Kroese, and A.J. Willeboordse, Methods and Informatics Department, P.O. Box 4481, 6401 CZ, Heerlen, The Netherlands. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors wish to thank Peter Kooiman, Nico Nieuwenbroek, and René Achenbach for their careful reading and helpful comments. The authors also thank Jeroen Pannekoek for his useful contributions.

[2] Willeboordse and Ypma (1996 and 1998) distinguish two levels of coherence, namely “coordination of concepts”, requiring the coordination and standardization of variables and classifications and “internal consistency of data”, requiring the harmonization of data collection and -processing.

umns. Obviously, only the observed scores are registered. The unobserved scores correspond to empty cells, in the understanding that imperfections like measurement errors and item non-response have already been dealt with by some editing and imputation strategy.

The traditional way of constructing estimates is to use one set of weights per survey. Given the first order inclusion probabilities, such a set can be obtained by calibration techniques as discussed in Deville and Särndal (1992). When using one set of weights per survey, all variables are inflated in the same way. The main advantage of such an approach is that once the set of weights has been calculated, it can be applied directly to any set of study variables giving numerically consistent estimates. This approach is, however, not suitable for a micro database, since there are several surveys (including registers) involved. Although numerical consistency is achieved per survey, across surveys many estimates will be numerically inconsistent. This is extensively illustrated in e.g. Kroese, Renssen and Trijssenaar (2000). This paper provides for an alternative estimation strategy. This estimation strategy is still based on weighting, however, not necessarily on one set of weights per survey.

Before describing our estimation strategy, we first introduce in Section 2 some basic conceptual notions, resulting in both a terminology and a notation applying for the issue. Among others, we introduce the concept of *umbrella frequency distribution*, which forms a formal framework within which all potential target frequency distributions should fit. Such a framework is important, not only to formally define the notion of ‘numerical consistency’, but also to bring structure in the enormous number of target distributions statistical bureaus are usually faced with. The estimation strategy is given in Section 3. It consists of three steps. The first step divides the micro database into a number of rectangular micro subsets. The second step assigns preliminary regression weights to each of the micro subsets, by means of which a large number of estimates can be made. The third step fixes any inconsistencies between these estimates. In Section 4 we give a numerical example. Finally, Section 5 touches briefly on some subjects that need further research.

2. THE CONCEPTUAL FRAMEWORK

Before discussing the estimation strategy itself, it is important to have a clear understanding of the nature and structure of the statistical data that are subject of estimation. Therefore, in this section we first explore the data structures as they apply for *micro databases* as suppliers of the input for the estimation operation. Next, the data structure of *aggregate databases*, as resulting from the estimation process is explained. The conceptual framework will be laid down by providing both a terminology of relevant concepts and a mathematical notation.

2.1 Terminology

It is assumed that each micro database covers and defines a specific target population, namely the complete set of *statistical objects* that instantiate a certain object type. Each object shapes a row in the database listing. For the persons micro database this set corresponds to the Municipal Base Administration and for the businesses micro database to the General Business Register. Each object is described according to one or more *attributes*, expressed in terms of *values* that the object scores on a variety of *variables*. These variables are operationally defined by reference to a specific measurement of scale, which provides for a valid method to measure them. With respect to measurement of scales, a major distinction applies between

- *Categorical* (or *qualitative*) variables, the measurement of which is carried out in a *nominal* or *ordinal* scale. Its values refer to categories that often derived from a classification.
- *Quantitative* variables, which are measured in a *ratio* or an *interval* scale. Its values are denoted as numbers.

Furthermore, it is important to be aware that variables of both types can be defined on different levels of generalization. We introduce the adjective *first order* for a variable that – given the information available from observation – is defined at its most elementary level (e.g. income in euro’s) and *second order* for a variable whose values are obtained by generalization from a first order variable (income in classes). Second order variables can be derived from both categorical and quantitative variables. The former case comes down to collapsing existing categories up to a higher level in a classification hierarchy. The latter implies the *creation* of categories from numbers. As a matter of terminology, second order variables, derived from first order quantitative variables, keep the status of 'quantitative'. Indeed, by using the information from the corresponding first order variable, it remains possible to apply them in arithmetic operations.

In theory, it is possible to derive numerous second order variables from first order variables. However, in order to prevent the emergence of a proliferation of finite population parameters, resulting in confusion or even inconsistent publication data, it makes sense to deliberately control and limit the number of categories allowed for second order variables, and to explicitly document these variables in the micro database, next to the first order variables. A certain degree of restriction and discipline can be obtained by imposing the rule that all second order variables that are derived from a certain first order variable should have a nested structure. A variable *A* is said to be *nested* within a variable *B*, or more briefly, *A* is nested within *B*, if every category of *A* fits into a single category of *B*. The practical implication of this restriction is that first order variables may only be accompanied by a hierarchical sequence of second order variables.

When considering the data structure at the level of aggregates, a further (not an alternative) distinction applies, notably with respect to the *role* of a variable in the aggregation/estimation process, by which micro data are transformed to aggregate data:

- *classification* variables, i.e. variables whose values (= categories) are used to subdivide the object-type of the micro database into subtypes, and accordingly to subdivide the population into sub-populations or classes;
- *quantification* variables, i.e. variables whose values (= numbers) are used to compile totals, means, etc., applying for the classes as defined above.

Notice that classification variables in the aggregate database are not just counterparts of the categorical variables in the micro database. Indeed, it is not only categorical variables, but also second order quantitative variables that can act as classification variables. Notice further that a classification variable can be a *crossing* of several (categorical or second order quantitative) variables, e.g., age and sex, or rather age *by* sex. Therefore, it makes sense to distinguish between simple (classification) variables and multiple (classification) variables, the former referring to one categorical or second order quantitative variable, the latter referring to more than one. The categories of multiple classification variables can be said to refer to cross-classifications. In the remaining of this report, we only consider classification variables. Renssen et al. (2001) also elaborate on quantification variables.

2.2 Notation

Let *U* denote a specific target population and $X^{(r)}$ a simple classification variable with *p* classes. For mathematical convenience, we represent $X^{(r)}$ as a *p*-vector of dummy variables, that is, we define

$$\mathbf{x}_i^{(r)} = (x_{i1},\dots,x_{ip})^t \text{ with } x_{ij} = \begin{cases} 1 \text{ if the } i\text{-th object belongs to the } j\text{-th class} \\ 0 \qquad\qquad \text{otherwise} \end{cases}.$$

An *hierarchical sequence* of *k* simple classification variables is denoted by $\mathbf{x}_i^{(1)},\dots,\mathbf{x}_i^{(k)}$. Here, $\mathbf{x}_i^{(r)}$ is nested within $\mathbf{x}_i^{(s)}$ if $s < r$. According to this notation, $\mathbf{x}_i^{(k)}$ contains the most classes. For the sake of convenience we also define $\mathbf{x}_i^{(0)} \equiv 1$. This always constant can be considered as a degenerated classification variable that refers to just one class, namely the complete population. If we distinguish

G first order (classification) variables, then, obviously, *G* of such hierarchical sequences can be defined: $\mathbf{x}_{1i}^{(1)},...,\mathbf{x}_{1i}^{(k_1)}$, $\mathbf{x}_{2i}^{(1)},...,\mathbf{x}_{2i}^{(k_2)}$, $\ldots$, $\mathbf{x}_{Ji}^{(1)},...,\mathbf{x}_{Gi}^{(k_G)}$.

Let $\mathbf{x}_{1i}^{(r_1)},\mathbf{x}_{2i}^{(r_2)},...,\mathbf{x}_{mi}^{(r_m)}$ denote *m* simple classification variables with $p_1, p_2,..., p_m$ classes respectively, $0 \le r_j \le k_j$, $j = 1,...,m$. The complete crossing between these variables, which is denoted by $\mathbf{x}_{1i}^{(r_1)} \otimes \mathbf{x}_{2i}^{(r_2)} \otimes ... \otimes \mathbf{x}_{mi}^{(r_m)}$, results in a multiple classification variable with $p_1 \times p_2 \times ... \times p_m$ classes. The population frequency distribution of this variable can be written as

$$\sum_{i \in U} \mathbf{x}_{1i}^{(r_1)} \otimes \mathbf{x}_{2i}^{(r_2)} \otimes ... \otimes \mathbf{x}_{mi}^{(r_m)} .$$

Notice that the population total is obtained by taking all $r_j = 0$; one-dimensional frequency distributions are obtained by taking all $r_j = 0$ but one; two-dimensional frequency distributions by taking all $r_j = 0$ but two, etc..

Let $\mathbf{x}_{1i}^{(k_1)} \otimes \mathbf{x}_{2i}^{(k_2)} \otimes ... \otimes \mathbf{x}_{Gi}^{(k_G)}$ be the multiple classification variable that divides the population into the maximum number of disjoint groups or sub-populations. Then

$$\sum_{i \in U} \mathbf{x}_{1i}^{(k_1)} \otimes \mathbf{x}_{2i}^{(k_2)} \otimes ... \otimes \mathbf{x}_{Gi}^{(k_G)} . \qquad (1)$$

can be considered as an *umbrella frequency distribution*. It forms the framework within which the enormous number of simple descriptive finite population parameters statistical offices are typically faced with, should be defined. Only frequency distributions that can be derived as marginal distributions from (1), i.e.

$$\sum_{i \in U} \mathbf{x}_{1i}^{(r_1)} \otimes \mathbf{x}_{2i}^{(r_2)} \otimes ... \otimes \mathbf{x}_{gi}^{(r_g)} , \qquad (2)$$

where $0 \le r_g \le k_g$, $g = 1,...,G$, are potential candidates for the aggregate database. Clearly, all these distributions are related with the umbrella frequency distributions through a *G*-vector $\mathbf{r} = (r_1,...,r_G)^t$. The number of non-zero values of this vector represents the dimension of the distribution, while these non-zero values themselves refer to the levels of the simple classification variables that are involved. From a different point of view, $\mathbf{r}$ specifies a partition of the finite population into an exclusive and exhaustive set of sub-populations.

By construction, two frequency distributions $\mathbf{D}_1$ and $\mathbf{D}_2$ are necessarily related to each other, i.e. cell totals or combinations of cell totals of the one quantity distribution equal cell totals or combinations of cell totals of the other quantity distribution. Let $\mathbf{D}_1$ be characterized by $\mathbf{r}_1$ and $\mathbf{D}_2$ by $\mathbf{r}_2$, then these common totals define a common quantity distribution that is characterized by $\mathbf{s} = \min(\mathbf{r}_1, \mathbf{r}_2)$, where the minimum is component wise determined. This common quantity distribution plays a central role in our estimation procedure in view of the consistency requirement. Notice that if $\mathbf{D}_1$ and $\mathbf{D}_2$ are defined by non-overlapping sets of simple classification variables, they still have a common part, namely the population total, which corresponds to a zero $\mathbf{s}$-vector.

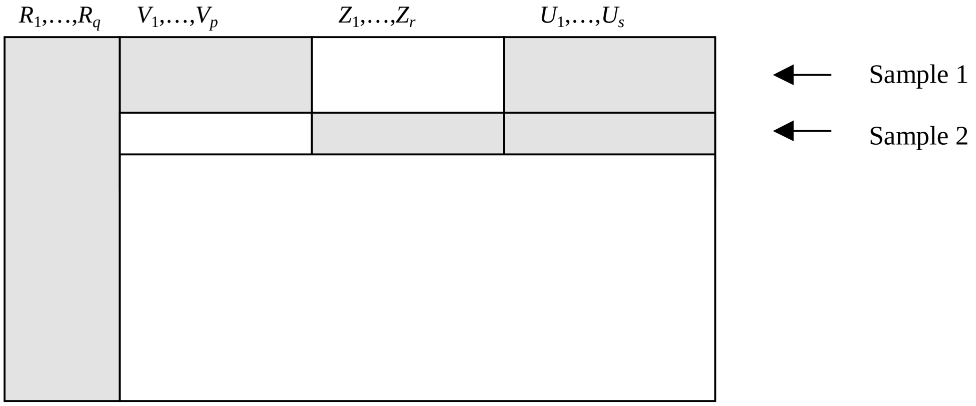
3. THE ESTIMATION STRATEGY

In the previous section we presented the conceptual framework within which each frequency distribution that is considered for publication should fit. Let $\mathbf{D}$ denote such a frequency. Whether $\mathbf{D}$ is actually qualified for publication usually depends on three conditions: $\mathbf{D}$ should be worthwhile, $\mathbf{D}$ should be safe, i.e. it should pass the rules of disclosure control, and $\hat{\mathbf{D}}$ should be sufficiently accurate, that is, all cell estimates of $\mathbf{D}$ should have a sufficiently small mean squared error. These conditions are all well known and commonly imposed by most statistical agencies, and we will not elaborate on them.

In order to meet the user wishes with respect to comparability of statistical output, Statistics Netherlands also employs a fourth condition: $\hat{\mathbf{D}}$ should be consistent with respect to all related distributions that already have been published (and stored) before, even if these distributions were estimated from other surveys. The estimation strategy that takes into account this fourth condition is new and will be the subject of this section. It is based on the general regression estimator, noting that it could more generally have been based on the calibration estimator. It involves three steps: 1) constructing rectangular micro subsets from a given micro database, 2) assigning to each micro subset a set of regression weights according to some weighting scheme, and 3) consistently estimating a set of quantity distributions.

By way of illustration we consider a simplified version as depicted in figure 1. Only the shaded surfaces are filled with observations. The *R*-columns correspond to a complete registration, to which two sample surveys are matched. In the one sample survey, denoted by S_1, *V*- and *U*-variables are observed; in the other sample survey, which is denoted by S_2, *Z*- and *U*-variables are observed. Obviously, the *U*-variables are observed in both sample surveys. We note that the register precisely corresponds to the (finite) population of interest. This population is denoted by *U* and consists of *N* units. In correspondence with figure 1, we associate with each unit a vector with scores of potential target variables. A part of these target variables is observed by administrative registrations, the remaining target variables are observed by sample surveys. Now, if $\mathbf{D}$ solely consists of register variables, then it can be estimated by straightforward counting. Otherwise, if $\mathbf{D}$ also consists of sampled variables, we use a general regression estimator.

Figure 1. A prototypical micro database



Step 1: constructing rectangular micro subsets
On behalf of the estimation strategy we divide the database as depicted in figure 1 into four subsets. In a slight abuse of our previous notation, they are denoted by R, $S_1 \cup S_2$, S_1 and S_2; R corresponds to the administrative registration and contains the *R*-variables; $S_1 \cup S_2$ corresponds to the union of the first and second samples and contains *R*- and *U*-variables; S_1 corresponds to the first sample and contains *R*-, *U*-, and *V*-variables; S_2 corresponds to the second sample and contains *R*-, *U*-, and *Z*-

variables. Now, estimating a specific *m*-way table involves among others the determination of the proper micro subset. For example, if the estimation concerns only *R*-variables the proper micro subset is R, while a crossing between *V*- and *R*-variables should be estimated from S_1.

Step 2: assigning regression weights
The next step is assigning a (fixed) set of weights to each micro subset according to some weighting scheme to (globally) adjust for sampling error and non-response and to meet some (not all) consistency requirements. To that purpose we first have to derive the starting weights for each micro subset.

Starting weights are easily derived for R, S built up by the weighting terms $\mathbf{D}_{(1,t)}, \mathbf{D}_{(2,t)}, \ldots, \mathbf{D}_{(t-1,t)}$, where $\mathbf{D}_{(i,t)}$ stands for the common frequency distribution between $\mathbf{D}_i$ and $\mathbf{D}_t$, i.e. $\mathbf{D}_{(i,t)}$ refers to the partition $\min(\mathbf{r}_i, \mathbf{r}_t)$.

Principally, one has to derive the minimal re-weighting scheme for each distribution separately. Given a specific order of the set of target distributions, these re-weighting schemes are uniquely defined. However, a different order generally implies a different set of re-weighting schemes. In the next section we will illustrate by means of an example how this order problem can be (largely) circumvented.

4. A numerical example; Samplone

The complete estimation strategy is easiest illustrated by means of a numerical example. Suppose the population size is N = 1000 persons. An integral registration provides scores of the variables region (there are seven municipalities; Wheaton, Greenham, Newbay, Oakdale, Smokeley, Crowdon, and Mudwater), age (three age classes; young, middle, old), and sex (male, female). Furthermore, a simple random sample of size n = 100 provides scores of the variable employment (yes, no). We define one second order variable for region by combining municipalities up to two provinces, namely Agria (consisting of Wheaton, Greenham, and Newbay) and Induston (consisting of Oakdale, Smokeley, Crowdon, and Mudwater).

Table 4.1. region$^{(2)}$ × sex$^{(1)}$

	Wheaton	Greenham	Neybay	Oakdale	Crowdon	Smokeley	Mudwater	Total
Male	70	44	31	36	128	80	122	511
Female	74	50	24	25	116	67	133	489
Total	144	94	55	61	244	147	255	1000

Table 4.2. sex$^{(1)}$ × region$^{(1)}$ × age$^{(1)}$

Male				Female			
	Agria	Induston	Total		Agria	Induston	Total
Young	80	146	226	Young	61	148	209
Middle	47	156	203	Middle	57	135	192
Old	18	64	82	Old	30	58	88
Total	145	366	511	Total	148	341	489

Analogous to the conceptual model, we distinguish four hierarchical sequences of (simple) classification variables: [region$^{(2)}$, region$^{(1)}$, region$^{(0)}$], [sex$^{(1)}$, sex$^{(0)}$], [age$^{(1)}$, age$^{(0)}$], and [employ$^{(1)}$, employ$^{(0)}$]. By means of these four hierarchical sequences one may construct maximally 3×2×2×2 = 24 (multiple) classification variables. We notice that the most detailed classification variable consists of 7×3×2×2 = 84 classes and corresponds with a partition that is denoted by $\mathbf{r} = (2 \quad 1 \quad 1 \quad 1)^t$. Our purpose is to consistently estimate the following set of distributions:

- region$^{(2)}$×sex$^{(1)}$,
- region$^{(1)}$×sex$^{(1)}$×age$^{(1)}$,
- region$^{(2)}$×employ$^{(1)}$,
- employ$^{(1)}$×age$^{(1)}$, and
- sex$^{(1)}$×employ$^{(1)}$×region$^{(1)}$

Now, the first two target distributions can be obtained by straightforward register counts. The results are given in tables 4.1 and 4.2. The counts for region$^{(2)}$ (table 4.1) agree nicely with the count for region$^{(1)}$ (table 4.2), which is requisite in a view of the consistency requirement.

In order to estimate the remaining target distributions we employ the following estimation strategy. As a start, we adopt the (overall) weighting scheme "region$^{(1)}$ + sex$^{(1)}$ " to adjust for sampling error (and any non-response bias). Ignoring the requirement of consistency, the resulting preliminary regression weights can be used to obtain the desired (preliminary) estimates. The results are given in appendix I. Although these tables are mutually consistent, a comparison between these tables on the one hand and the register counts on the other hand (table 4.1 and 4.2) reveals several inconsistencies with respect to e.g. the age$^{(1)}$-variable and the region$^{(2)}$-variable.

To deal with these inconsistencies we apply the theory of minimal re-weighting as developed in the previous section. The results are summarized in table 4.3. We only discuss $\mathbf{D}_3$ = region$^{(2)}$×employ$^{(1)}$ in more detail. The other target distributions can be discussed in a similar way. Now, in order to (largely) circumvent the order problem, we first estimate its one-dimensional margins: region$^{(2)}$ and employ$^{(1)}$. As these margins can be estimated by means of the overall weighting model without violating any consistency requirement, no re-weighting is needed. The minimal re-weighting scheme for $\mathbf{D}_3$ itself therefore is employ$^{(1)}$ + region$^{(2)}$.

Table 4.3. Re-weighting schemes[3]

Frequency distributions	Proper subset	Re-weighting needed	Re-weighting schemes
region$^{(2)}$×sex$^{(1)}$	R	No	-
region$^{(1)}$×sex$^{(1)}$×age$^{(1)}$	R	No	-
region$^{(2)}$	R	No	-
employ$^{(1)}$	S	No	-
region$^{(2)}$×employ$^{(1)}$	S	Yes	employ$^{(1)}$ + region$^{(2)}$
employ$^{(1)}$	S	No	-
age$^{(1)}$	R	No	-
employ$^{(1)}$×age$^{(1)}$	S	Yes	employ$^{(1)}$ + age$^{(1)}$
sex$^{(1)}$	R	No	-
employ$^{(1)}$	S	No	-
region$^{(1)}$	R	No	-
sex$^{(1)}$×employ$^{(1)}$	S	No	-
sex$^{(1)}$×region$^{(1)}$	R	No	-
employ$^{(1)}$×region$^{(1)}$	S	No	-
sex$^{(1)}$×employ$^{(1)}$×region$^{(1)}$	S	Yes	sex$^{(1)}$×employ$^{(1)}$ + sex$^{(1)}$×region$^{(1)}$ + employ$^{(1)}$×region$^{(1)}$

The resulting estimates are given in appendix II. Naturally, these estimates are consistent with respect to all register counts as well as mutually consistent.

5. SUMMARY AND FURTHER RESEARCH

In this paper we proposed an estimation strategy for combined data sources. The reasons to combine data sources and to come up with a new estimation strategy were political pressure to reduce response burden and to accommodate user demands to produce outputs that are consistent and mutually related. The estimation strategy was still based on regression techniques - or more generally on calibration techniques - but not necessarily on one weighting scheme per survey. It involved three steps: 1) constructing rectangular micro-datasets from the combined data sources, 2) assigning to each micro-dataset a (fixed) set of regression (or calibration) weights according to some weighting scheme, and 3) for each target table minimally adjusting the original weighting scheme to obtain a so-called minimal re-weighing scheme that is tailored to the consistency demand. The practicability of the estimation strategy was illustrated by means of a fictitious example.

The estimation strategy presented is preliminary, and a more extensive study is needed. Below, we briefly mention some difficulties. The idea of repeated weighting to obtain numerical consistent estimates assumes the existence of 'perfect' micro databases, i.e. micro databases that consists of a (not too large) number of rectangular micro datasets that are mutually consistent at the micro level. However, constructing such micro databases is a very complex task, in which many difficult choices have to be made. A (prototype) software tool has been developed to support the involved three steps of the repeated weighting process. Although it is possible to derive approximation formulas for the

[3] To circumvent some technical details, these re-weighting schemes may slightly differ from those given in Renssen et al. (2001).

variance, the obtained expressions are rather complicated and therefore not implemented yet. Finally, we mention two theoretical complications that are related to quantitative variables, namely the dual role of quantitative variables and the phenomenon sub-variable. Referring to the former, a quantitative variable, such as age, can be used both as classification and as quantification variable. Especially when a quantification variable assumes a finite number of values, where each value corresponds to a class of a corresponding classification variable, the consistency problem becomes manifest. Examples of the latter are often formulated in terms of edit rules, such as 'material costs' + 'personal' costs = 'total costs'. Then, the variables 'material costs' and 'personal costs' can be considered as sub-variables of 'total costs'. We already have developed methodology to cope with these complications, but this research is still going on.

REFERENCES

Al, P.G. and Bakker, F.M. (2000), "Re-engineering Social Statistics by Micro-integration of Different Sources: an Introduction", *Netherlands Official Statistics*, 15, pp. 4-6, Special Issue, *Integrating Administrative Registers and Household Surveys*.

Bethlehem, J. and Keller, W. (1987), "Linear Weighting of Sample Survey Data", *Journal of Official Statistics*, 3, pp. 141-153.

Deville, J.C. and Särndal, C.E. (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87, pp. 376-382.

Kroese, A.H., Renssen, R.H., and Trijssenaar, M. (2000), " Weighting or Imputation: constructing a consistent set of estimates based on data from different sources", *Netherlands Official Statistics*, 15, pp. 23-31, Special Issue, *Integrating Administrative Registers and Household Surveys*.

Laan van der, P. (2000), "Integrating Administrative Registers and Household Surveys", *Netherlands Official Statistics*, 15, pp. 7-15, Special Issue, *Integrating Administrative Registers and Household Surveys*.

Renssen, R.H., A.H. Kroese, and Willeboordse, A.J. (2001), "Aligning Estimates by Repeated Weighting", unpublished report (BPA H 491-01-TMO), Heerlen, The Netherlands: Statistics Netherlands.

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*, 90, pp. 368-374.

Willeboordse, A. (2000), "Towards a new Statistics Netherlands. Blueprint for a process oriented organisation structure", Unpublished report, Voorburg, The Netherlands: Statistics Netherlands.

Willeboordse, A. and Ypma, W. (1996), "”From Rules to Tools. New Opportunities to Establish Coherence among Statistics", *Proceedings of the Conference on output Databases*, Voorburg, November 1996, Statistics Netherlands, Voorburg, The Netherlands.

Willeboordse, A. and Ypma, W. (1998), "Meta Tools in Support of a Corporate Dissemination Strategy", unpublished report (research paper 9839), Voorburg, The Netherlands: Statistics Netherlands.

Appendix I:

Estimates of three target distributions according to the weighting model “region$^{(1)}$ + sex$^{(1)}$”.

Table 1. region$^{(2)}$ × employ$^{(1)}$

	Wheaton	Greenham	Neybay	Oakdale	Crowdon	Smokeley	Mudwater	Total
Job	66	25	32	33	65	77	66	363
No job	66	73	32	22	123	111	211	637
Total	131	97	65	55	188	187	276	1000

Table 2. employ$^{(1)}$ × age$^{(1)}$

	Young	Middle	Old	Total
Job	86	255	22	363
No job	357	106	174	637
Total	443	361	196	1000

Table 3. sex$^{(1)}$ × employ$^{(1)}$ × region$^{(1)}$

	Male				Female		
	Agria	Induston	Total		Agria	Induston	Total
Job	48	207	255	Job	75	34	108
No Job	104	152	256	No Job	67	314	381
Total	152	359	511	Total	141	348	489

Appendix II:

Estimates of three target tables according to re-weighting scheme, see table 4.3.

Table 1. region$^{(2)}$ × employ$^{(1)}$

	Wheaton	Greenham	Neybay	Oakdale	Crowdon	Smokeley	Mudwater	Total
Job	72	24	27	36	84	60	60	363
No job	72	70	28	25	160	87	195	637
Total	144	94	55	61	244	147	255	1000

Table 2. employ$^{(1)}$ × age$^{(1)}$

	Young	Middle	Old	Total
Job	76	270	17	363
No job	359	125	153	637
Total	435	395	170	1000

Table 3. sex$^{(1)}$ × employ$^{(1)}$ × region$^{(1)}$

	Male				Female		
	Agria	Induston	Total		Agria	Induston	Total
Job	46	209	255	Job	77	31	108
No Job	99	157	256	No Job	71	310	381
Total	145	366	511	Total	148	341	489