DACSEIS
IST-2000-26057

Workpackage 7

A strategy to obtain consistency among tables of survey estimates

Deliverable 7.2

2003

List of contributors:

H.J.H. Boonstra, J.A. van den Brakel, P. Knottnerus, N.J. Nieuwenbroek, R.H. Renssen; Statistics Netherlands.

Main responsibility:

P. Knottnerus; Statistics Netherlands

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

Preface

This deliverable describes the new estimation strategy of Statistics Netherlands in view of the Dacseis project (workpackage 7). Part of this work is based on earlier studies at Statistics Netherlands, see Renssen *et al.* (2001). This deliverable focuses on the variance estimation of the estimators proposed in the earlier studies.

At Statistics Netherlands the design and organization of the statistical process is changing. This is motivated by the need to produce more consistent data and by political pressures to cut down the response burden. The idea underlying 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 deliverable gives the initial impetus of a more formal approach to an estimation strategy per micro database. The proposed strategy ensures that all estimated m-way tables are consistent with respect to common margins, even if these tables are estimated from different surveys. It is based on the regression estimator, but not necessarily on a fixed set of weights per survey. The applicability of the strategy is illustrated by means of a fictitious example. Although the estimation strategy proposed in this deliverable is closely related to the regression estimator, the corresponding variance formulas are somewhat more complicated.

H.J.H. Boonstra, J.A. van den Brakel, P. Knottnerus, N.J. Nieuwenbroek, R.H. Renssen; Statistics Netherlands.

Contents

List of figures **VII**

List of tables **IX**

1 Introduction **1**

2 The conceptual framework **3**

2.1 Terminology 3

2.2 Notation 4

3 A simple example of the estimation strategy **7**

4 The estimation strategy **11**

4.1 A prototypical micro database 11

4.2 The weighting strategy 12

4.2.1 Deriving starting weights based on sampling designs 12

4.2.2 Deriving regression weights based on overall weighting schemes . . . 13

4.2.3 Deriving final weights based on re-weighting schemes 13

4.3 The resulting estimators 16

5 Variance estimation **19**

5.1 A simple example; one registration and one sample 19

5.2 One registration and two independent samples 21

5.2.1 RW-estimators based on S 22

5.2.2 RW-estimators based on S_1 23

5.2.3 Recursions for calculating the e-variables 24

5.3 One registration and two dependent samples; two-phase sampling 26

6 Alternative approximations of the RW-estimator and its variance 29

7 Summary and further research 33

A Appendix 35

A.1 Estimates of five target distributions according to the weighting model "region + sex". 35

A.2 Estimates of five target tables according to re-weighting scheme, see table 3.3 36

References 37

List of Figures

4.1 A prototypical micro database 12

List of Tables

3.1 Region × sex 9
3.2 Sex × region × age 9
3.3 Re-weighting schemes 10

A.1 Region × employ 35
A.2 Employ × age 35
A.3 Sex × employ × region 35
A.4 Region × employ 36
A.5 Employ × age 36
A.6 Sex × employ × region 36

Chapter 1

Introduction

Traditionally, statistical offices organize their data collection, -processing, and -dissemi nation 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[1]. Secondly, in order to limit the response burden, providers of information should be questioned as little as possible, while the opposite may occur when using the stovepipe model. Thirdly, the accuracy of the estimates may be unnecessarily low if the estimation strategy makes no or less use of supplementary registrations and/or other surveys.

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

Currently, Statistics Netherlands distinguishes between a micro database for persons and a micro database for businesses. Globally speaking, the micro database for persons consists of the Municipal Base Administration as the backbone with sample surveys and registers about persons matched to it. Similarly, the micro database for businesses consists of the General Business Register as the backbone with sample surveys and registers about businesses matched to it. Typically, micro databases can be seen as rectangular arrays with individual (statistical) units in its rows and scores for variables in its columns. Obviously, only the observed scores are available. The unobserved scores correspond to empty cells, provided 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

[1] Willeboordse and Ypma (1996, 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.

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, however, is not suitable for a micro database, since there are several surveys (including registers) involved. Although consistency is achieved per survey, across surveys many estimates will be inconsistent. This is extensively illustrated in KROESE and RENSSEN (1999) and KROESE *et al.* (2000). The present paper provides an alternative estimation strategy. The strategy is still based on weighting, but not necessarily on one set of weights per survey.

Before describing this estimation strategy, we first introduce in Chapter 2 some basic concepts, resulting in both a terminology and a notation. The estimation strategy is introduced by means of an example in Chapter 3. The strategy involves 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 inconsistency between these estimates. The estimation strategy is discussed more generally in Chapter 4. Variance estimators for the estimators derived in Chapter 4 are given in Chapter 5. In Chapter 6 a simplification of the RW-estimator is proposed in order to give an alternative expression for its variance. Finally, Chapter 7 touches briefly on some subjects that need further research.

Chapter 2

The conceptual framework

In this chapter, we first explore the structure of *micro databases* as suppliers of the input for the estimation process. 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 a specific target population. Each unit of this population shapes a row in the database. The person-based micro database corresponds to the Municipal Base Administration and the business-based micro database to the General Business Register. Each unit is described according to one or more *variables*. With respect to the measurement of scales, a major distinction applies between *categorical* or *qualitative* variables and *quantitative* variables.

We introduce the adjective *first-order* for a variable defined at the level at which the variable was initially observed (e.g. income in euros) and *second-order* for a variable whose values are derived from a first-order variable (income in classes). In theory, it is possible to derive numerous second-order variables. However, in order to avoid a proliferation of finite population parameters, resulting in confusion or even inconsistent publication data, it makes sense to limit the categories allowed for second-order variables. This can be accomplished by imposing the rule that all variables measuring the same quantity at different levels of detail must conform to a nested structure. A variable A is *nested* within a variable B if each category of A fits into a single category of B. This restriction implies that first-order variables may only be accompanied by a hierarchical sequence of second-order variables.

A further distinction applies to the *role* of a variable in the aggregation/estimation process:

- *classification* variables, i.e. variables whose values represent categories, which partition the population into an exclusive and exhaustive set of subpopulations
- *quantification* variables, i.e. variables whose numerical values are used to compile totals, means, etc., for the classes defined by a classification variable.

We distinguish between *simple* and *multiple* classification variables. The former consist of a single categorical variable while the latter are obtained by crossings of two or more categorical variables. In this report, we only consider the estimation of frequency tables, which are defined exclusively in terms of classification variables. RENSSEN *et al.* (2001) also elaborate on quantification variables.

2.2 Notation

Let Ω denote a specific target population of size N and $\mathbf{x}^{(r)}$ a simple classification variable with $p = p(r)$ classes. The superscript r indicates the degree of classification; the larger r, the finer the classification. For mathematical convenience, we represent $\mathbf{x}^{(r)}$ as a p-vector of dummy variables, that is, we define

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

where $i = 1, \ldots, N$ and $j = 1, \ldots, p$. A *hierarchical sequence* of k simple classification variables is denoted by $\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(k)}$. Here, $\mathbf{x}_i^{(r)}$ is nested within $\mathbf{x}_i^{(s)}$ if $r > s$. According to this notation, $\mathbf{x}_i^{(k)}$ contains the largest number of classes. For convenience, we also define $\mathbf{x}_i^{(0)} \equiv 1$, which can be considered as a degenerate classification variable that refers to a single class, namely the complete population. If G first-order simple classification variables are distinguished, then G of such hierarchical sequences can be defined: $\mathbf{x}_{1i}^{(1)}, \ldots, \mathbf{x}_{1i}^{(k_1)}$, $\mathbf{x}_{2i}^{(1)}, \ldots, \mathbf{x}_{2i}^{(k_2)}, \ldots, \mathbf{x}_{Gi}^{(1)}, \ldots, \mathbf{x}_{Gi}^{(k_G)}$.

Let the vector $\mathbf{x}_{1i}^{(k_1)} \times \mathbf{x}_{2i}^{(k_2)} \times \ldots \times \mathbf{x}_{Gi}^{(k_G)}$ be the multiple classification variable that divides the population into the maximum number of disjoint groups or sub-populations. If the frequency distribution is to be viewed as a vector, then here and in the rest of this text a Kronecker product ("$\otimes$") is understood for the symbol "$\times$". The population frequency distribution of this variable is given by

$$\sum_{i \in \Omega} \mathbf{x}_{1i}^{(k_1)} \times \mathbf{x}_{2i}^{(k_2)} \times \ldots \times \mathbf{x}_{Gi}^{(k_G)}. \tag{2.1}$$

Only frequency distributions that can be derived as marginal distributions from (2.1), i.e.

$$\sum_{i \in \Omega} \mathbf{x}_{1i}^{(r_1)} \times \mathbf{x}_{2i}^{(r_2)} \times \ldots \times \mathbf{x}_{Gi}^{(r_G)},$$

where $0 \leqslant r_g \leqslant k_g$, $g = 1, \ldots, G$, are potential candidates for the aggregate database. Clearly all these distributions or, equivalently, contingency tables are related to distribution (2.1) through a G-vector $\mathbf{r} = (r_1, \ldots, r_G)^t$. The number of non-zero values of this vector equals the dimension of the distribution, while these non-zero values themselves refer to the levels of the simple classification variables involved. Recall that $\mathbf{r}$ corresponds to a partition of the finite population into an exclusive and exhaustive set of sub-populations. In particular, the population total is obtained by taking all $r_j = 0$.

Two frequency distributions $\mathbf{t}_1$ and $\mathbf{t}_2$ are necessarily related to each other, i.e. cell totals or combinations of cell totals of $\mathbf{t}_1$ equal cell totals or combinations of cell totals of $\mathbf{t}_2$,

since they always have at least the population total in common. Let $\mathbf{t}_1$ be characterized by $\mathbf{r}_1$ and $\mathbf{t}_2$ by $\mathbf{r}_2$, then these common totals define a common frequency distribution that is characterized by the vector $\mathbf{s} = \min(\mathbf{r}_1, \mathbf{r}_2)$, whose components are the minima of the corresponding components of $\mathbf{r}_1$ and $\mathbf{r}_2$. This rule rests on the nested structure of the classification variables. The common frequency distribution plays a central role in our estimation procedure in view of the consistency requirement. Notice that if $\mathbf{t}_1$ and $\mathbf{t}_2$ are defined by non-overlapping sets of simple classification variables, then the common part is the population total, which corresponds to a zero vector $\mathbf{s}$.

Chapter 3

A simple example of the estimation strategy

In the previous chapter we presented the conceptual framework for all frequency distributions considered for publication. Let

$$\mathbf{t}_\Gamma = \sum\nolimits_{i \in \Omega} \mathbf{\Gamma}_i \tag{3.1}$$

denote such a frequency distribution, where the vector $\mathbf{\Gamma}_i = \mathbf{x}_{1i}^{(r_1)} \times \mathbf{x}_{2i}^{(r_2)} \times \ldots \times \mathbf{x}_{Gi}^{(r_G)}$ refers to a multiple classification variable. Whether $\mathbf{t}$ is actually qualified for publication usually depends on three conditions: $\mathbf{t}$ should be worthwhile, $\mathbf{t}$ should be safe, i.e. it should pass the rules of disclosure control, and the estimate for $\mathbf{t}$, denoted by $\widehat{\mathbf{t}}$, should be sufficiently accurate, that is, all cell estimates of $\mathbf{t}$ should have a sufficiently small mean squared error. These conditions are all well known and commonly imposed by most statistical offices, and we will not elaborate on them.

In order to meet users' wishes with respect to comparability of statistical output, Statistics Netherlands also employs a fourth condition: $\widehat{\mathbf{t}}$ should be nummerically consistent with respect to all related distributions estimated from the micro database. The estimation strategy that takes into account this fourth condition is new and will be the subject of this chapter. It is based on the general regression estimator and 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 (overall) weighting scheme, and 3) consistently estimating a set of distributions. In this chapter we give a simple example to explain the main issues of this estimation strategy. In Chapter 4 the estimation strategy is discussed more extensively in a more general context.

Suppose the population size is $N = 1000$ persons. A complete register for this population provides scores of the variables region (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 employ (yes, no). We define one second-order variable for region by merging the municipalities into two provinces, namely Agria (Wheaton, Greenham and Newbay) and Induston (Oakdale, Smokeley, Crowdon and Mudwater).

Four hierarchical sequences of (simple) classification variables are distinguished: [region$^{(2)}$, region$^{(1)}$, region$^{(0)}$], [sex$^{(1)}$, sex$^{(0)}$], [age$^{(1)}$, age$^{(0)}$], and [employ$^{(1)}$, employ$^{(0)}$]. The most detailed classification variable consists of $7\times2\times3\times2 = 84$ classes corresponding with a partition that is denoted by $\mathbf{r} = (\ 2\ \ 1\ \ 1\ \ 1\ )^t$. Our purpose is to consistently estimate the following set of frequency distributions:

- $\mathbf{t}_1 = \text{region}^{(2)}\times\text{sex}^{(1)} \rightarrow \mathbf{r}_1 = (\ 2\ \ 1\ \ 0\ \ 0\ )^t$
- $\mathbf{t}_2 = \text{region}^{(1)}\times\text{sex}^{(1)}\times\text{age}^{(1)} \rightarrow \mathbf{r}_2 = (\ 1\ \ 1\ \ 1\ \ 0\ )^t$,
- $\mathbf{t}_3 = \text{region}^{(2)}\times\text{employ}^{(1)} \rightarrow \mathbf{r}_3 = (\ 2\ \ 0\ \ 0\ \ 1\ )^t$,
- $\mathbf{t}_4 = \text{employ}^{(1)}\times\text{age}^{(1)} \rightarrow \mathbf{r}_4 = (\ 0\ \ 0\ \ 1\ \ 1\ )^t$,
- $\mathbf{t}_5 = \text{sex}^{(1)}\times\text{employ}^{(1)}\times\text{region}^{(1)} \rightarrow \mathbf{r}_5 = (\ 1\ \ 1\ \ 0\ \ 1\ )^t$.

The first two frequency distributions can be obtained by straightforward register counts. The results are given in Tables 3.1 and 3.2. The counts for region$^{(2)}$ (Table 3.1) perfectly agree with the counts for region$^{(1)}$ (Table 3.2).

In order to estimate the remaining three frequency distributions we employ the following estimation strategy. As a start, we adopt the (overall) weighting scheme "region$^{(1)}$ + sex$^{(1)}$" to reduce sampling error and non-response bias. That is to say, the variables denoted by the vectors region$^{(1)}$ and sex$^{(1)}$ are used as auxiliary variables in a classical regression estimator. Subsequently, the resulting regression estimator of an arbitrary study variable can be written as a weighted mean of the sample observations of that study variable; see SÄRNDAL *et al.* (1992), pp. 232-233. Assuming that there is a unique matching key - so it is not difficult to identify the persons from the sample in the register (see e.g. Figure 4.1 in the next chapter) - the tables region$^{(2)}\times$employ$^{(1)}$, employ$^{(1)}\times$age$^{(1)}$, and sex$^{(1)}\times$employ$^{(1)}\times$region$^{(1)}$ can be estimated by means of the regression-based weights. The estimated tables are given in Appendix A.1. Although these tables are consistent, a comparison between these tables on the one hand and the register counts on the other hand (Tables 3.1 and 3.2) reveals several inconsistencies regarding e.g. age$^{(1)}$ and region$^{(2)}$.

There are several ways to fix these inconsistencies. An obvious way is to enlarge the (overall) weighting scheme. It is always a good thing to anticipate on the publication tables in an early stage. For example, the (overall) weighting scheme "region$^{(2)}$+ age$^{(1)}$ + region$^{(1)}\times$sex$^{(1)}$", would automatically avoid all inconsistencies. In this example with only a few variables available, we use the parsimonious weighting scheme "region$^{(1)}$ + sex$^{(1)}$" only to give a simple demonstration of how the general procedure works, apart from whether or not in this particular case a larger weighting scheme would be preferable. In practice, with many frequency tables to be estimated, it is often impossible to solve the consistency requirement by applying one (overall) weighting scheme, since the weighting scheme might become too large compared to the sample size. Below, we focus on $\mathbf{t}_3$, noting that $\mathbf{t}_4$ and $\mathbf{t}_5$ can be discussed in a similar way.

Minimal re-weighting schemes and the order problem

Observe for $\mathbf{t}_3$ that $\min(\mathbf{r}_3, \mathbf{r}_1) = (\ 2\ \ 0\ \ 0\ \ 0\ )^t$ and $\min(\mathbf{r}_3, \mathbf{r}_2) = (\ 1\ \ 0\ \ 0\ \ 0\ )^t$. So, in order to re-estimate $\mathbf{t}_3$ consistently with $\mathbf{t}_1$ and $\mathbf{t}_2$ it suffices to allow for the table that

Table 3.1: region$^{(2)}\times$ sex$^{(1)}$ = ($\mathbf{t}_1$)

	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 3.2: sex$^{(1)}\times$ region$^{(1)}\times$ age$^{(1)}$ = ($\mathbf{t}_2$)

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

is characterized by the partition (2 0 0 0)^t, which corresponds to region$^{(2)}$. This table is referred to as the 'minimal re-weighting scheme'. One could simply adjust $\mathbf{t}_3$ such that the re-estimate is consistent with region$^{(2)}$ by using the calibration property of the regression estimator, and proceed with $\mathbf{t}_4$ given the estimates for $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$, etc. However, the order in which the estimated tables are adjusted affects the final result.

The splitting-up procedure and re-estimation

To largely circumvent the order problem we apply the so-called splitting-up procedure. This means that we first estimate the most detailed margins of $\mathbf{t}_3$ and next $\mathbf{t}_3$ itself, given the estimates of these margins. A most detailed margin of some m-way table[1] is obtained by replacing one of the simple classification variables of the table by its next lower variable in the hierarchical sequence, keeping the other $m-1$ simple classification variables unchanged. This leads to a set of m different most detailed margins for an m-way table. So, before re-estimating $\mathbf{t}_3$ = region$^{(2)}\times$employ$^{(1)}$ we first estimate its most detailed margins:

1. region$^{(1)}\times$ employ$^{(1)}$ (= region$^{(2-1)}\times$ employ$^{(1)}$),
2. region$^{(2)}$ (= region$^{(2)}\times$ employ$^{(1-1)}$)

The proper data set to estimate region$^{(2)}$ is the register. Therefore region$^{(2)}$ can be obtained by register counts. The proper data set for estimating region$^{(1)}\times$employ$^{(1)}$is the sample. The margin "region$^{(1)}\times$employ$^{(1)}$" estimated by the traditional regression estimator with the (overall) weighting scheme "region$^{(1)}$ + sex$^{(1)}$", is consistent with Tables 3.1 and 3.2.

[1]An m-way table is a table of the form (3.1) where m of the G components $r_1 \ldots r_G$ are non-zero.

Table 3.3: Re-weighting schemes

Frequency distributions	Proper subset	Re-weighting schemes
region$^{(2)}$	R	-
employ$^{(1)}\times$ region$^{(1)}$	S	-
region$^{(2)}\times$ employ$^{(1)}$	S	employ$^{(1)}\times$region$^{(1)}+$ region$^{(2)}$
employ$^{(1)}$	S	-
age$^{(1)}$	R	-
employ$^{(1)}\times$ age$^{(1)}$	S	employ$^{(1)}+$ age$^{(1)}$
sex$^{(1)}\times$employ$^{(1)}$	S	-
sex$^{(1)}\times$region$^{(1)}$	R	-
employ$^{(1)}\times$region$^{(1)}$	S	-
sex$^{(1)}\times$employ$^{(1)}\times$region$^{(1)}$	S	sex$^{(1)}\times$ employ$^{(1)}+$ sex$^{(1)}\times$ region$^{(1)}$ + employ$^{(1)}\times$region$^{(1)}$

Next we re-estimate $\mathbf{t}_3$ taking into account the estimates of its most detailed margins region$^{(1)}\times$employ$^{(1)}$ and region$^{(2)}$. The proper data set for re-estimating this table is the sample. Re-estimation entails the application of the regression estimator, where "region$^{(2)}$ + region$^{(1)}\times$ employ$^{(1)}$" is the weighting scheme. The starting weights for this regression estimator, which usually correspond to the inclusion weights of the sampling scheme, are now taken to be the regression weights obtained with the overall weighting scheme. Since the re-weighting scheme "region$^{(2)}$ + region$^{(1)}\times$employ$^{(1)}$" encloses the minimal re-weighting scheme region$^{(2)}$, the resulting re-estimates are consistent with $\mathbf{t}_1$ and $\mathbf{t}_2$. Table 3.3 exhibits the re-weighting schemes for $\mathbf{t}_3$, $\mathbf{t}_4$ and $\mathbf{t}_5$. The re-estimated tables are displayed in Appendix A.2.

Chapter 4

The estimation strategy

In this chapter we extend the ideas of the previous chapter to two more general situations based on one registration. Besides, the estimation strategy is given in more general terms. In the first situation two independent samples with some common variables are added. In the second situation we have one sample that is partitioned into two subsamples. Note that each subsample of the second situation can be described by a two-phase sampling design.

In Section 4.1 we discuss a prototypical micro database in case of one registration and two (sub)samples. In Section 4.2 a weighting strategy for this micro database is discussed and in Section 4.3 the resulting estimators are given.

4.1 A prototypical micro database

Figure 4.1 depicts the micro database for both situations. Only the shaded surfaces are filled with observations. The R-columns correspond to a complete registration. The register variables are denoted by $R_1 \ldots R_q$. In the first situation, the two sample surveys are combined. In the one sample survey, V- and U-variables are observed and in the other sample survey Z- and U-variables. Note that the U-variables are observed in both sample surveys. Also note that it is tacitly assumed that the sample sizes are relatively small so that the intersection of the (independent) samples can be neglected. In the second situation, it is assumed that the U-variables are observed in a first-phase sample, and the $V-$ and $Z-$variables in two complementary subsamples of this first-phase sample.

We note that the register precisely corresponds to the (finite) population of interest. This population is denoted by Ω and consists of N units. In correspondence with Figure 4.1, we associate with each unit a vector with scores of potential target variables. Some of these target variables are observed from registrations, the remaining target variables are observed in sample surveys. The database can be divided into four micro subsets, denoted by R, S, S_1 and S_2; R corresponds to the administrative registration and contains the R-variables; S corresponds to the union of both (sub)samples and contains $R-$ and $U-$variables; S_1 corresponds to the first (sub)sample and contains R-, U-, and V-variables; S_2 corresponds to the second (sub)sample and contains R-, U-, and Z-variables.

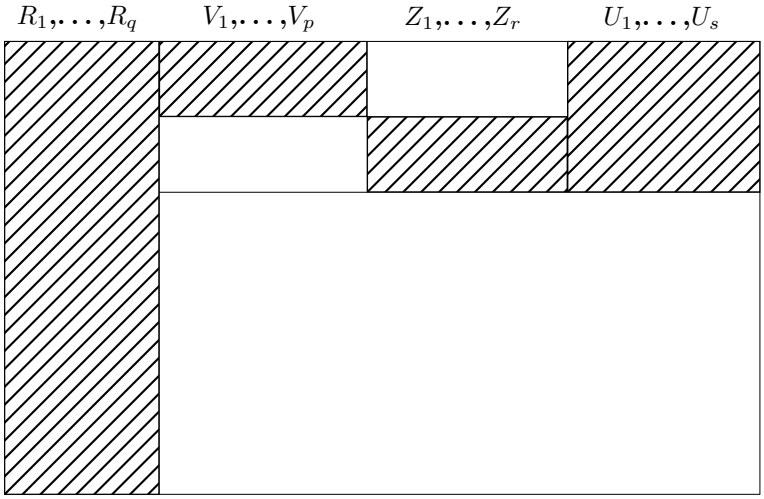

Figure 4.1: A prototypical micro database

Estimation of a specific m-way table starts by determining the proper micro subset. The proper micro subset is the largest subset of records for which scores on all variables of this m-way table are available. 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.

4.2 The weighting strategy

The weighting strategy involves assigning a fixed set of regression weights to each micro subset according to some overall weighting scheme to reduce sampling error and non-response and to meet some (not all) consistency requirements. We first have to derive the starting weights for each micro subset.

4.2.1 Deriving starting weights based on sampling designs

For R the starting weights are equal to 1 as this subset corresponds to a complete register. For S_1 and S_2 they equal the inverse of the (net) first-order inclusion probabilities of the first and second sample respectively. This also holds for the starting weights with respect to S in the case that S corresponds to a first-phase sample. In Section 5.3 these inclusion probabilities will be discussed for two-phase sampling. If S corresponds to the union of two independent samples, the starting weights for S can be derived in several ways. Here we use starting weights which ensure that all estimators can eventually be written as linear combinations of Horvitz-Thompson estimators based on the two samples. Thus, let π_{1i} and π_{2i} denote the first-order inclusion probabilities of the i-th unit, $i \in \Omega$, with respect to S_1 and S_2 respectively. Then the starting weights for S are defined as $d_i = \lambda\pi_{1i}^{-1}$ for $i \in S_1 \backslash S_2$, $d_i = (1-\lambda)\pi_{2i}^{-1}$ for $i \in S_2 \backslash S_1$, and $d_i = \lambda\pi_{1i}^{-1} + (1-\lambda)\pi_{2i}^{-1}$ for $i \in S_1 \cap S_2$, where $\lambda \in [0,1]$. Note that there may be a (small) overlap between the independent samples, which we have not tried to depict in Figure 4.1. Furthermore, it should be noted that due

to the starting weights d_i thus defined an estimator based on S can always be written as a weighted mean of two Horvitz-Thompson estimators from S_1 and S_2, that is, in obvious notation, $\widehat{\mathbf{t}}_\Gamma^{d(S)} = \lambda\widehat{\mathbf{t}}_\Gamma^{d(S_1)} + (1-\lambda)\widehat{\mathbf{t}}_\Gamma^{d(S_2)}$; see also (4.5). The choice of λ may reflect the confidence in the one sample compared to the other. For example, it may depend on indicators for several survey errors, such as sampling errors or non-response errors. As a simple choice one may take λ proportional to the relative sample size of S_1 with respect to S_2.

4.2.2 Deriving regression weights based on overall weighting schemes

After calculating the starting weights, a strategy has to be developed to assign a fixed set of regression weights to each of the micro subsets. One strategy is to weight S first, using some suitable $R-$variables in the weighting scheme and to weight S_1 and S_2 next. For both S_1 and S_2 we may use R- and U-variables as well as crossings between R- and U-variables in the weighting schemes. Such weighting schemes are called *overall weighting schemes*, in order to discriminate them from the *re-weighting schemes*. The resulting estimates based on S_1 and S_2 are by construction consistent with the counted or estimated totals of the variables from the overall weighting scheme. Such a weighting strategy corresponds to adjusted general regression estimates as discussed in Renssen and Nieuwenbroek (1997) in the case of two independent samples. It corresponds to two-phase regression estimates in case of the partitioned subsamples. For further details on the (two-phase) regression estimator we refer to e.g. Särndal *et al.* (1992).

Let B denote the proper micro subset for the estimation of a frequency table and $w_i^{(B)}$ the regression weights of micro subset B, then

$$w_i^{(B)} = d_i^{(B)}[1 + \mathbf{a}_i^t\widehat{\mathbf{G}}_B(\widehat{\mathbf{t}}_\mathbf{a} - \widehat{\mathbf{t}}_\mathbf{a}^{d(B)})], \tag{4.1}$$

where $d_i^{(B)}$ denote the starting weights of micro subset B, $\mathbf{a}_i$ denotes the vector of auxiliary variables determined by the overall weighting model of B, $\widehat{\mathbf{t}}_\mathbf{a}$ is a vector of counted or estimated population totals of $\mathbf{a}$ derived from a larger micro subset, while

$$\widehat{\mathbf{t}}_\mathbf{a}^{d(B)} = \sum\nolimits_{i\in B} d_i^{(B)}\mathbf{a}_i$$

is a vector of starting estimates based on micro subset B, and $\widehat{\mathbf{G}}_B$ is a generalized inverse of

$$\sum\nolimits_{i\in B} d_i^{(B)}\mathbf{a}_i\mathbf{a}_i^t.$$

4.2.3 Deriving final weights based on re-weighting schemes

By means of the overall weighting schemes one may produce general regression estimates for a set of target distributions. These estimates are asymptotically design unbiased (ADU) and there exist approximation formulas for their design variances. All general regression estimates for distributions with the same underlying micro subset are consistent

by construction since they use the same weights based on the overall weighting scheme. Since only a limited number of auxiliary variables can be incorporated into a weighting scheme, complete consistency between estimates based on different subsets cannot be guaranteed.

If one is willing to abandon the common practice of using one set of (regression) weights per micro subset, one can obtain consistency to a larger extent. Then, given a specific distribution to be estimated, it is necessary to determine all its 'margins' that have been estimated before, and to use these margins as auxiliary information in a re-weighting process.

Let $\{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_T\}$ be the series of target distributions to be estimated, and suppose $\mathbf{t}_t$ is the first target distribution in this series that has to be re-estimated in view of the consistency requirement. All its predecessors: $\mathbf{t}_1$, $\mathbf{t}_2$, ..., $\mathbf{t}_{t-1}$ must be taken into account. The common parts of $\mathbf{t}_t$ and its predecessors are reflected by $\mathbf{t}_{(1,t)}$, $\mathbf{t}_{(2,t)}$, ..., $\mathbf{t}_{(t-1,t)}$, where $\mathbf{t}_{(i,t)}$ refers to the common frequency distribution between $\mathbf{t}_i$ and $\mathbf{t}_t$, corresponding to the classification vector $\min(\mathbf{r}_i, \mathbf{r}_t)$. Now, one way to re-estimate $\mathbf{t}_t$ such that its re-estimate is consistent with its predecessors is to find the proper micro subset and to apply the general regression estimator to this subset using $\mathbf{t}_{(1,t)}$, $\mathbf{t}_{(2,t)}$, ..., $\mathbf{t}_{(t-1,t)}$ as auxiliary information. It is important to note that the regression estimates of these auxiliary tables are consistent by assumption since $\mathbf{t}_t$ is the first distribution that needs to be re-estimated. The re-weighting scheme that corresponds to the set of auxiliary information, given by

$$\mathbf{t}_{(1,t)} + \mathbf{t}_{(2,t)} + \ldots + \mathbf{t}_{(t-1,t)},$$

is called the *minimal re-weighting scheme*. The re-weighting procedure for, say $\mathbf{t}_s$, where $\mathbf{t}_s$ is the second target distribution to be re-estimated, is carried out similarly. Then, one has to account for $\mathbf{t}_1$, $\mathbf{t}_2$, ..., $\mathbf{t}_{s-1}$, which includes $\mathbf{t}_t$. The re-weighting scheme is generally over-specified in the sense that many variables in the re-weighting scheme can be expressed as linear combinations of other variables in the re-weighting scheme. Therefore the re-weighting scheme can often be written in a simpler form. For example in Chapter 3 the minimal re-weighting scheme of table $\mathbf{t}_3$ is written as

$$\mathbf{t}_{(1,3)} + \mathbf{t}_{(2,3)} = \text{region}^{(2)} + \text{region}^{(1)} = \text{region}^{(2)}.$$

Minimal re-weighting schemes should be derived 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. An important issue that should be considered therefore is the order in which the frequency distributions are estimated. The splitting-up procedure is introduced to circumvent this order problem.

In order to explain the estimation strategy based on the splitting-up procedure, consider an arbitrary m-way table given by

$$\mathbf{t}_\Gamma = \sum_{i \in \Omega} \mathbf{\Gamma}_i \qquad (\mathbf{\Gamma}_i \equiv \mathbf{x}_{1i}^{(r_1)} \times \mathbf{x}_{2i}^{(r_2)} \times \ldots \times \mathbf{x}_{mi}^{(r_m)}).$$

It is assumed that $r_k > 0$, $k = 1, 2, \ldots, m$. Let B denote the proper micro subset used to estimate $\mathbf{t}_\Gamma$. The splitting-up procedure is described in a recursive way. In the first

recursion step we consider the m most-detailed margins of $\mathbf{t}_\Gamma$, of which the corresponding classification variables are

$$\begin{aligned}
\Gamma^-_{1i} &= \mathbf{x}^{(r_1-1)}_{1i} \times \mathbf{x}^{(r_2)}_{2i} \times \vdots \times \mathbf{x}^{(r_m)}_{mi} \\
\Gamma^-_{2i} &= \mathbf{x}^{(r_1)}_{1i} \times \mathbf{x}^{(r_2-1)}_{2i} \times \vdots \times \mathbf{x}^{(r_m)}_{mi} \\
&\bullet \\
&\bullet \\
&\bullet \\
\Gamma^-_{mi} &= \mathbf{x}^{(r_1)}_{1i} \times \mathbf{x}^{(r_2)}_{2i} \times \vdots \times \mathbf{x}^{(r_m-1)}_{mi}
\end{aligned}$$

The recursive way of estimating means that, in a second recursion step, 1) each of these margins is estimated on its own proper micro subset and 2) each of these estimates is obtained according to the splitting-up procedure in the same way as described below for $\mathbf{t}_\Gamma$. This implies that each estimated margin in turn is an RW-estimate. Now, assuming that these margins are estimated consistently, $\mathbf{t}_\Gamma$ can also be estimated consistently by re-weighting $\mathbf{t}_\Gamma$ according to the re-weighting scheme

$$\Gamma^-_1 + \Gamma^-_2 + \ldots + \Gamma^-_m \quad . \tag{4.2}$$

Formally, the recursion ends by the zero dimensional margin $\mathbf{r} = (0, 0, \ldots, 0)^t$. In practice, however, there is no need to continue the recursion in a particular branch when a margin satisfies the decomposition criterion explained below.

Let $\mathbf{t}_\Gamma$ denote an arbitrary m-way table of which B is the proper micro subset, where $\mathbf{\Gamma}$ denotes a (multiple) classification variable. Then, $\mathbf{t}_\Gamma$ is said to satisfy the decomposition criterion if $\mathbf{\Gamma}$ can be decomposed as $\mathbf{\Gamma} = \mathbf{\Gamma}_\mathrm{a} \times \mathbf{\Gamma}_\mathrm{b}$, where $\mathbf{\Gamma}_\mathrm{a}$ corresponds to a (margin of a) term of the overall weighting scheme of B and $\mathbf{\Gamma}_\mathrm{b}$ is a crossing of simple classification variables of which each simple classification variable is observed exclusively within B. Let $\mathbf{x}^{(r_j)}_j$ denote such a simple classification variable. Then, by the condition for $\mathbf{\Gamma}_\mathrm{b}$ it is meant that there is no $\mathbf{x}^{(s_j)}_j$, $1 \leqslant s_j \leqslant r_j$, observed in a larger micro subset. This implies that B is the proper micro subset not only for $\mathbf{t}_{\mathbf{\Gamma}_\mathrm{b}}$ but also for all its margins. In this situation there is no need to reweight $\mathbf{t}_\Gamma$, i.e. to apply the splitting-up procedure, since consistency is already achieved by the regression weights. Namely,

1. the margins of $\mathbf{t}_\Gamma$, which are included in the overall weighting scheme are consistent due to the calibration property of the generalized regression estimator,

2. the variables among $\mathbf{x}^{(r_1)}_{1i}, \mathbf{x}^{(r_2)}_{2i}, \ldots, \mathbf{x}^{(r_m)}_{mi}$ observed only within micro subset B cannot give rise to inconsistencies with other tables.

Note that a special case of decomposition occurs when $\mathbf{t}_\Gamma$ is known from the register. It is self-evident that such a $\mathbf{t}_\Gamma$ need not be re-weighted and can directly be obtained from the register. In fact, this corresponds to the trivial decomposition $\Gamma = \Gamma_\mathrm{a} \times \Gamma_\mathrm{b} = 1 \times \Gamma_\mathrm{b}$ because, by definition, $\mathbf{t}_\Gamma$ cannot be observed outside the corresponding B, i.e., the register. For some examples, see Sections 5.1 and Chapter 6.

The splitting-up procedure implies that for the estimation of the lower-dimensional margins the proper micro subset is used. Since estimates for lower-dimensional margins of a distribution are generally based on larger micro subsets, they are more precise than estimates for higher-dimensional margins of the distribution. Using these margins as auxiliary information in the re-weighting scheme for a distribution $\mathbf{t}_\Gamma$ might improve the precision of the final estimates.

The final weights corresponding to the last re-weighting step are denoted by $r_i^{(B)}$. One finds similarly to the formula for the regression weights that

$$r_i^{(B)} = w_i^{(B)}[1 + \mathbf{m}_i^t \widehat{\mathbf{G}}_B(\widehat{\mathbf{t}}_\mathbf{m} - \widehat{\mathbf{t}}_\mathbf{m}^{w(B)})], \tag{4.3}$$

where $\mathbf{m}_i^t = (\Gamma_{1i}^{-t}, \ldots, \Gamma_{mi}^{-t})$ denotes the vector of most-detailed margins defined by (4.2), and $\widehat{\mathbf{t}}_\mathbf{m}$ is a vector of population totals of $\mathbf{m}_i$, counted or estimated from a larger micro subset. Furthermore,

$$\widehat{\mathbf{t}}_\mathbf{m}^{w(B)} = \sum\nolimits_{i \in B} w_i^{(B)} \mathbf{m}_i$$

is a vector of regression estimates based on micro subset B, and $\widehat{\mathbf{G}}_B$ a generalized inverse of

$$\sum\nolimits_{i \in B} w_i^{(B)} \mathbf{m}_i \mathbf{m}_i^t.$$

The same approach has to be followed for the tables of most-detailed margins as well as for tables with finer classifications, and so on.

In the end, it seems that the number of restrictions imposed by the re-weighting scheme may become very large. We notice, however, that the number of non-redundant consistency restrictions cannot exceed the number of cells of the target table under consideration. Re-estimation problems, which are due to too small sample sizes compared to the re-weighting scheme, indicate that the target table itself is probably too detailed to be estimated by means of weighting techniques.

4.3 The resulting estimators

Following the weighting strategy, three types of estimators for $\mathbf{t}_\Gamma$ are distinguished. The first type of estimator corresponds to the class of Horvitz-Thompson type estimators (HT-estimators) based on the proper micro subset B:

$$\widehat{\mathbf{t}}_\Gamma^{d(B)} = \sum_{i \in B} d_i^{(B)} \Gamma_i. \tag{4.4}$$

Strictly speaking, (4.4) only is a HT-estimator if the starting weights equal the reciprocal of the first order inclusion weights. Especially, when B corresponds to the union of two independent samples, (4.4) corresponds not to a HT-estimator but to a linear combination of two HT-estimators, e.g.

$$\widehat{\mathbf{t}}_\Gamma^{d(S)} = \lambda \widehat{\mathbf{t}}_\Gamma^{d(S_1)} + (1 - \lambda) \widehat{\mathbf{t}}_\Gamma^{d(S_2)} \quad . \tag{4.5}$$

The second type of estimators corresponds to the class of general regression estimators:

$$\widehat{\mathbf{t}}_{\Gamma}^{w(B)} = \sum_{i \in B} w_i^{(B)} \Gamma_i = \widehat{\mathbf{t}}_{\Gamma}^{d(B)} + (\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}}^{d(B)})^t (\widehat{\mathbf{t}}_{\mathbf{a}} - \widehat{\mathbf{t}}_{\mathbf{a}}^{d(B)}), \tag{4.6}$$

where $w_i^{(B)}$ are the general regression weights defined in (4.1). The matrix $\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}}^{d(B)}$ of estimated regression coefficients is defined by

$$\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}}^{d(B)} = \widehat{\mathbf{G}}_B \sum_{i \in B} d_i^{(B)} \mathbf{a}_i \Gamma_i^t \ , \tag{4.7}$$

with Γ as target variable and $\mathbf{a}$ as regressor (cf. (4.1)). Note that the regression coefficient is estimated from the micro dataset B using the starting weights. Strictly speaking (4.6) is not a general regression estimator as the vector of population totals $\widehat{\mathbf{t}}_{\mathbf{a}}$ might be estimated instead of known. However, it is assumed that $\widehat{\mathbf{t}}_{\mathbf{a}}$ is based on a micro dataset larger than B and hence that $\widehat{\mathbf{t}}_{\mathbf{a}}$ is more accurate than $\widehat{\mathbf{t}}_{\mathbf{a}}^{d(B)}$.

The last type of estimators are repeated weighting (RW) estimators resulting from a complete splitting-up procedure. As explained in section 4.2.3, such estimators can be written in a recursive way:

$$\widehat{\mathbf{t}}_{\Gamma}^{RW} = \sum_{i \in B} r_i^{(B)} \Gamma_i = \widehat{\mathbf{t}}_{\Gamma}^{w(B)} + (\widehat{\mathbf{B}}_{\Gamma;\mathbf{m}}^{w(B)})^t \begin{pmatrix} \widehat{\mathbf{t}}_{\Gamma_1^-}^{RW} - \widehat{\mathbf{t}}_{\Gamma_1^-}^{w(B)} \\ \vdots \\ \widehat{\mathbf{t}}_{\Gamma_m^-}^{RW} - \widehat{\mathbf{t}}_{\Gamma_m^-}^{w(B)} \end{pmatrix}, \tag{4.8}$$

where $r_i^{(B)}$ are the weights defined in (4.3) resulting from re-weighting with respect to $\Gamma_1^- = \mathbf{x}_1^{(r_1-1)} \times \mathbf{x}_2^{(r_2)} \times \ldots \times \mathbf{x}_m^{(r_m)}$, $\ldots$, $\Gamma_m^- = \mathbf{x}_1^{(r_1)} \times \mathbf{x}_2^{(r_2)} \times \ldots \times \mathbf{x}_m^{(r_m-1)}$. The matrix $\widehat{\mathbf{B}}_{\Gamma;\mathbf{m}}^{w(B)}$ of estimated regression coefficients is defined similarly to (4.7) but with regressors $\mathbf{m}^t = (\Gamma_1^{-t}, \ldots, \Gamma_m^{-t})$ and based on regression weights $w_i^{(B)}$ instead.

Chapter 5

Variance estimation

In the previous chapters the estimation strategy of repeated weighting is explained. In this chapter approximation formulas for the variance of the repeated weighting estimator (RW-estimator) are derived. First, variance formulas are derived for the simplest situation, i.e. one register and one sample from the example given in Chapter 3. Subsequently, variance estimators are derived for more general situations, i.e. one register and two independent samples, and one register and one sample that is partitioned into two dependent subsamples according to a two-phase sampling design.

5.1 A simple example; one registration and one sample

Consider the example described in Chapter 3. A complete registration R is available, which contains the variables region ($\mathbf{x}_1$), sex ($\mathbf{x}_2$), and age ($\mathbf{x}_3$). In addition, a sample S is drawn in which the variable employment ($\mathbf{x}_4$) is observed. Instead of a simple random sampling design (without replacement) a complex sampling design (without replacement) is assumed, where π_i and π_{ij} denote the first and second-order inclusion probabilities for $i, j \in \Omega$, respectively. The overall weighting scheme of S is $\mathbf{x}_1^{(1)} + \mathbf{x}_2^{(1)}$, so that $\mathbf{a}^t = (\ \mathbf{x}_1^{(1)^t}\ \ \mathbf{x}_2^{(1)^t}\ )$. In this subsection the variance for the RW-estimator for $\mathbf{t}_3 = \text{region}^{(2)} \times \text{employ}^{(1)}$,

$$\mathbf{t}_3 = \mathbf{t}_{\mathbf{x}_1^{(2)} \times \mathbf{x}_4^{(1)}} = \sum_{i \in \Omega} \mathbf{x}_{1i}^{(2)} \times \mathbf{x}_{4i}^{(1)},$$

is derived. The idea is to express the RW-estimator in terms of regression type estimators, which in turn can be expressed in terms of HT-estimators. According to Table 3.3, the reweighting scheme is $\mathbf{x}_1^{(1)} \times \mathbf{x}_4^{(1)} + \mathbf{x}_1^{(2)}$. Consequently, the first step of the reweighting procedure yields the expression

$$\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)} \times \mathbf{x}_4^{(1)}} = \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1^{(2)} \times \mathbf{x}_4^{(1)}} + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1^{(2)} \times \mathbf{x}_4^{(1)}; \mathbf{x}_1^{(1)} \times \mathbf{x}_4^{(1)} + \mathbf{x}_1^{(2)}})^t \begin{pmatrix} \widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(1)} \times \mathbf{x}_4^{(1)}} & - & \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1^{(1)} \times \mathbf{x}_4^{(1)}} \\ \widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}} & - & \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1^{(2)}} \end{pmatrix} . \quad (5.1)$$

The second step involves the estimation of $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}}$ and $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}}$. As both tables satisfy the decomposition criterion explained below (4.2), it is not necessary to proceed with the splitting-up procedure; $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}}$ can be obtained straightforwardly by register counting,

$$\widehat{\mathbf{t}}^{\mathrm{RW}}_{\mathbf{x}_{1(2)}} = \mathbf{t}_{\mathbf{x}_1^{(2)}} \quad , \tag{5.2}$$

and $\mathbf{t}^{RW}_{\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}}$ is estimated consistently from S using the overall weighting scheme $\mathbf{x}_1^{(1)}+\mathbf{x}_2^{(1)}$,

$$\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}} = \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}}. \tag{5.3}$$

Inserting (5.2) and (5.3) into (5.1) gives

$$\begin{aligned}\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} &= \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}+\mathbf{x}_1^{(2)}})^t \begin{pmatrix} 0 \\ \mathbf{t}_{\mathbf{x}_1^{(2)}} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1^{(2)}} \end{pmatrix} \\ &= \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{x}_1^{(2)};\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}+\mathbf{x}_1^{(2)}})^t \left(\mathbf{t}_{\mathbf{x}_1^{(2)}} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1^{(2)}}\right),\end{aligned} \tag{5.4}$$

where $\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{x}_1^{(2)};\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}+\mathbf{x}_1^{(2)}}$ is the estimated coefficient for $\mathbf{x}_1^{(2)}$ in a weighted regression of $\mathbf{x}_1^{(2)} \times \mathbf{x}_4^{(1)}$ on $\mathbf{x}_1^{(1)} \times \mathbf{x}_4^{(1)}$ and $\mathbf{x}_1^{(2)}$. Using equation (4.6) for the regression estimators, we can rewrite (5.4) in terms of HT-estimators as

$$\begin{aligned}\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} &= \widehat{\mathbf{t}}^{d(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} + [(\widehat{\mathbf{B}}^{d(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{a}})^t - (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{x}_1^{(2)};\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}+\mathbf{x}_1^{(2)}})^t (\widehat{\mathbf{B}}^{d(S)}_{\mathbf{x}_1^{(2)};\mathbf{a}})^t] \left(\mathbf{t}_{\mathbf{a}} - \widehat{\mathbf{t}}^{d(S)}_{\mathbf{a}}\right) \\ &\quad + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{x}_1^{(2)};\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}+\mathbf{x}_1^{(2)}})^t \left(\mathbf{t}_{\mathbf{x}_1^{(2)}} - \widehat{\mathbf{t}}^{d(S)}_{\mathbf{x}_1^{(2)}}\right) \\ &\equiv \widehat{\mathbf{t}}^{d(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} + \widehat{\mathbf{M}}_1 \left(\mathbf{t}_{\mathbf{a}} - \widehat{\mathbf{t}}^{d(S)}_{\mathbf{a}}\right) + \widehat{\mathbf{M}}_2 \left(\mathbf{t}_{\mathbf{x}_1^{(2)}} - \widehat{\mathbf{t}}^{d(S)}_{\mathbf{x}_1^{(2)}}\right),\end{aligned} \tag{5.5}$$

where we introduced coefficient matrices $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$ given by

$$\begin{aligned}\widehat{\mathbf{M}}_{\mathbf{1}} &= (\widehat{\mathbf{B}}^{d(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{a}})^t - (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{x}_1^{(2)};\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}+\mathbf{x}_1^{(2)}})^t (\widehat{\mathbf{B}}^{d(S)}_{\mathbf{x}_1^{(2)};\mathbf{a}})^t \quad , \\ \widehat{\mathbf{M}}_{\mathbf{2}} &= (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)};\mathbf{x}_1^{(2)};\mathbf{x}_1^{(1)}\times\mathbf{x}_4^{(1)}+\mathbf{x}_1^{(2)}})^t.\end{aligned}$$

To find an expression for the variance of the RW-estimator, (5.5) is approximated by means of a first-order Taylor series expansion:

$$\begin{aligned}\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} &\approx \widehat{\mathbf{t}}^{d(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} + \mathbf{M}_1 \left(\mathbf{t}_{\mathbf{a}} - \widehat{\mathbf{t}}^{d(S)}_{\mathbf{a}}\right) + \mathbf{M}_2 \left(\mathbf{t}_{\mathbf{x}_1^{(2)}} - \widehat{\mathbf{t}}^{d(S)}_{\mathbf{x}_1^{(2)}}\right) \\ &= \widehat{\mathbf{t}}^{d(S)}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}} - \mathbf{M}_1\widehat{\mathbf{t}}^{d(S)}_{\mathbf{a}} - \mathbf{M}_2\widehat{\mathbf{t}}^{d(S)}_{\mathbf{x}_1^{(2)}} + \mathbf{c},\end{aligned} \tag{5.6}$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are the finite population analogs of $\widehat{\mathbf{M}}_1$ and $\widehat{\mathbf{M}}_2$ and $\mathbf{c}$ is some constant. Denoting

$$\mathbf{e}_i = \mathbf{x}_{1i}^{(2)} \times \mathbf{x}_{4i}^{(1)} - \mathbf{M}_1\mathbf{a}_i - \mathbf{M}_2\mathbf{x}_{1i}^{(2)},$$

we can approximate the variance-covariance matrix (hereafter simply called variance) of the table of estimates (5.5) by

$$\mathrm{Var}(\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}}) \approx \sum_{i\in\Omega}\sum_{j\in\Omega}(\pi_{ij}-\pi_i\pi_j)\frac{\mathbf{e}_i\mathbf{e}_j^t}{\pi_i\pi_j}. \tag{5.7}$$

An estimator for this variance expression is given by

$$\widehat{\mathrm{Var}}(\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}}) \approx \sum_{i\in S}\sum_{j\in S}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{\widehat{\mathbf{e}}_i\widehat{\mathbf{e}}_j^t}{\pi_i\pi_j}, \tag{5.8}$$

with $\widehat{\mathbf{e}}_i = \mathbf{x}_{1i}^{(2)}\times\mathbf{x}_{4i}^{(1)} - \widehat{\mathbf{M}}_1\mathbf{a}_i - \widehat{\mathbf{M}}_2\mathbf{x}_{1i}^{(2)}$. The residuals in the variance estimator (5.8) might be multiplied by correction weights or g-weights as an alternative (see Särndal *et al*, 1992, Result 6.6.1). One possible option is to take $g_i^{(S)} \equiv r_i^{(S)}/d_i^{(S)}$, however, this should be further investigated. In addition, note that (5.7) and (5.8) can be further elaborated for various complex sampling designs (see Särndal *et al*, 1992, Chapter 3). If the sample size is small in comparison with the size of the whole population one could borrow the simpler variance formula

$$\widehat{\mathrm{Var}}(\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1^{(2)}\times\mathbf{x}_4^{(1)}}) \approx \frac{n}{n-1}\sum_{i\in S}\frac{\widehat{\mathbf{e}}_i\widehat{\mathbf{e}}_i^t}{\pi_i^2},$$

see Särndal *et al.* (1992), Section 11.2.

From the splitting-up procedure for $\mathbf{t}_3$ as given in (5.1) to (5.5) it follows that, in principle, all RW-estimators based on S can be approximated by linear combinations of HT-estimators:

$$\widehat{\mathbf{t}}_\Gamma^{RW} = \widehat{\mathbf{t}}_\Gamma^{d(S)} - \sum_{t=1}^{T}\widehat{\mathbf{M}}_t\widehat{\mathbf{t}}_{\phi_t}^{d(S)} + \sum_{t=1}^{T}\widehat{\mathbf{M}}_t\mathbf{t}_{\phi_t} \approx \widehat{\mathbf{t}}_\Gamma^{d(S)} - \sum_{t=1}^{T}\mathbf{M}_t\widehat{\mathbf{t}}_{\phi_t}^{d(S)} + \mathbf{c}. \tag{5.9}$$

Here ϕ_t runs (by its index t) over all tables used as auxiliary information in the estimation procedure to estimate $\mathbf{t}_\Gamma$, that is, ϕ_t denotes either a term of the overall weighting scheme or a margin of Γ encountered in the re-weighting process; cf. (5.5). In the case at hand ϕ_t runs over the terms $\mathbf{x}_1^{(1)}$ and $\mathbf{x}_2^{(1)}$ of the overall weighting scheme and the re-weighting variable $\mathbf{x}_1^{(2)}$, as can be seen from (5.6).

5.2 One registration and two independent samples

In this section the variance approximation formulas for the RW-estimator in case of one register and two independent samples are derived. Recall from Chapter 4 that four micro datasets are considered in this situation, namely the registration R, the union of the two samples S, the first sample S_1, and the second sample S_2. Both S_1 and S_2 are drawn according to some complex design without replacement. Let π_{1i} and π_{1ij} denote the first and second-order inclusion probabilities with respect to S_1 for $i, j \in \Omega$. A similar notation holds for the first and second-order inclusion probabilities with respect to S_2.

The starting weights $d_i^{(B)}$ for the different micro subsets have been discussed in Subsection 4.2.1 Given these starting weights, regression weights $w_i^{(B)}$ are assigned to each micro subset according to (4.1). According to the weighting strategy as suggested in Subsection 4.2.2, S is weighted first, using some suitable $R-$variables in the weighting scheme. Subsequently, S_1 and S_2 are weighted, using R- and U-variables as well as crossings between R- and U-variables in the weighting schemes. This implies that $\widehat{\mathbf{t}}_{\mathbf{a}}$ in (4.1) corresponds to register totals when weighting S, whereas $\widehat{\mathbf{t}}_{\mathbf{a}}$ may consist of both known population totals and estimated population totals when weighting S_1 or S_2. In fact, in the latter case, $\widehat{\mathbf{t}}_{\mathbf{a}}$ may consist of known population totals as well as RW-estimators that are based on S. In the next two subsections, the overall weighting scheme for S and S_1 will be denoted by $\mathbf{a}$ and $\mathbf{a_1}$ respectively.

5.2.1 RW-estimators based on S

The RW-estimator based on one registration R and one sample S, which happens to be the union of two independent samples is discussed first. Almost all of the reasoning of section 5.1 can be repeated here, that is, all RW-estimators based on S can be written as linear combinations of HT-estimators. Similarly to (5.9), using in addition (4.5), we get

$$\widehat{\mathbf{t}}_\Gamma^{RW} = \widehat{\mathbf{t}}_\Gamma^{d(S)} - \sum_{t=1}^{T} \widehat{\mathbf{M}}_t \widehat{\mathbf{t}}_{\phi_t}^{d(S)} + \sum_{t=1}^{T} \widehat{\mathbf{M}}_t \mathbf{t}_{\phi_t} \tag{5.10}$$

$$\approx \widehat{\mathbf{t}}_\Gamma^{d(S)} - \sum_{t=1}^{T} \mathbf{M}_t \widehat{\mathbf{t}}_{\phi_t}^{d(S)} + \mathbf{c} \tag{5.11}$$

$$= [\lambda \widehat{\mathbf{t}}_\Gamma^{d(S_1)} + (1-\lambda)\widehat{\mathbf{t}}_\Gamma^{d(S_2)}] - \sum_{t=1}^{T} \mathbf{M}_t [\lambda \widehat{\mathbf{t}}_{\phi_t}^{d(S_1)} + (1-\lambda)\widehat{\mathbf{t}}_{\phi_t}^{d(S_2)}] + \mathbf{c}$$

$$= \lambda [\widehat{\mathbf{t}}_\Gamma^{d(S_1)} - \sum_{t=1}^{T} \mathbf{M}_t \widehat{\mathbf{t}}_{\phi_t}^{d(S_1)}] + (1-\lambda)[\widehat{\mathbf{t}}_\Gamma^{d(S_2)} - \sum_{t=1}^{T} \mathbf{M}_t \widehat{\mathbf{t}}_{\phi_t}^{d(S_2)}] + \mathbf{c}$$

$$= \lambda \sum_{i \in S_1} \pi_{1i}^{-1} (\Gamma_i - \sum_{t=1}^{T} \mathbf{M}_t \phi_{ti}) + (1-\lambda) \sum_{i \in S_2} \pi_{2i}^{-1} (\Gamma_i - \sum_{t=1}^{T} \mathbf{M}_t \phi_{ti}) + \mathbf{c}$$

$$= \lambda \sum_{i \in S_1} \pi_{1i}^{-1} \mathbf{e}_i + (1-\lambda) \sum_{i \in S_2} \pi_{2i}^{-1} \mathbf{e}_i + \mathbf{c}. \tag{5.12}$$

Since S_1 and S_2 are independent, the variance of $\widehat{\mathbf{t}}_\Gamma^{RW}$ can be approximated by

$$\mathrm{Var}(\widehat{\mathbf{t}}_\Gamma^{RW}) \approx \lambda^2 \sum_{i \in \Omega} \sum_{j \in \Omega} (\pi_{1ij} - \pi_{1i}\pi_{1j}) \frac{\mathbf{e}_i \mathbf{e}_j^t}{\pi_{1i}\pi_{1j}} + (1-\lambda)^2 \sum_{i \in \Omega} \sum_{j \in \Omega} (\pi_{2ij} - \pi_{2i}\pi_{2j}) \frac{\mathbf{e}_i \mathbf{e}_j^t}{\pi_{2i}\pi_{2j}}. \tag{5.13}$$

As expected, (5.13) not only depends on the residuals $\mathbf{e}_i = \Gamma_i - \sum_t \mathbf{M}_t \phi_{ti}$ and on the sampling designs of S_1 and S_2 but also on the choice of λ. The variance estimation procedure of (5.13) is similar to that of (5.7) and we will not elaborate on this any further.

5.2.2 RW-estimators based on S_1

Consider the RW-estimator based on S_1, and note that the RW-estimator based on S_2 can be discussed similarly. Recall that the overall weighting scheme of S_1 consists of R-variables, U-variables, or combinations of these variables. Therefore, without loss of generality, it is assumed that $\widehat{\mathbf{t}}_{\mathbf{a}_1}$ partly consists of known population totals obtained from the register and partly of unknown population totals estimated from S. Equation (4.8) expresses a general RW-estimator in terms of RW-estimators of its margins and regression estimators based on the overall weighting scheme. After expanding the recursion, i.e. after repeatedly substituting (4.8) into itself, one always arrives at an expression in terms of regression estimators,

$$\widehat{\mathbf{t}}_{\Gamma}^{RW} = \widehat{\mathbf{t}}_{\Gamma}^{w(S_1)} + \sum_{t=1}^{T} (\widehat{\mathbf{C}}_{\Gamma;\phi_t}^{w(S_1)})^t (\widehat{\mathbf{t}}_{\phi_t} - \widehat{\mathbf{t}}_{\phi_t}^{w(S_1)}), \tag{5.14}$$

where t runs over margins of Γ encountered in the successive steps of the recursive re-weighting procedure, $\widehat{\mathbf{C}}_{\Gamma;\phi_t}$ denote matrices consisting of products of regression coefficients, and $\widehat{\mathbf{t}}_{\phi_t}$ is either a vector of known population totals obtained from R or an estimated vector of population totals based on S. Inserting

$$\widehat{\mathbf{t}}_{\Gamma}^{w(S_1)} = \widehat{\mathbf{t}}_{\Gamma}^{d(S_1)} + (\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}_1}^{d(S_1)})^t (\widehat{\mathbf{t}}_{\mathbf{a}_1} - \widehat{\mathbf{t}}_{\mathbf{a}_1}^{d(S_1)})$$

and

$$\widehat{\mathbf{t}}_{\phi_t}^{w(S_1)} = \widehat{\mathbf{t}}_{\phi_t}^{d(S_1)} + (\widehat{\mathbf{B}}_{\phi_t;\mathbf{a}_1}^{d(S_1)})^t (\widehat{\mathbf{t}}_{\mathbf{a}_1} - \widehat{\mathbf{t}}_{\mathbf{a}_1}^{d(S_1)})$$

into (5.14) results in

$$\begin{aligned}\widehat{\mathbf{t}}_{\Gamma}^{RW} = \widehat{\mathbf{t}}_{\Gamma}^{d(S_1)} &+ \left[(\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}_1}^{d(S_1)})^t - \sum_{t=1}^{T} (\widehat{\mathbf{C}}_{\Gamma;\phi_t}^{w(S_1)})^t (\widehat{\mathbf{B}}_{\phi_t;\mathbf{a}_1}^{d(S_1)})^t \right] (\widehat{\mathbf{t}}_{\mathbf{a}_1} - \widehat{\mathbf{t}}_{\mathbf{a}_1}^{d(S_1)}) \\ &+ \left[\sum_{t=1}^{T} (\widehat{\mathbf{C}}_{\Gamma;\phi_t}^{w(S_1)})^t \right] (\widehat{\mathbf{t}}_{\phi_t} - \widehat{\mathbf{t}}_{\phi_t}^{d(S_1)}),\end{aligned} \tag{5.15}$$

where $\widehat{\mathbf{t}}_{\mathbf{a}_1}$ is the vector of (counted or estimated) population totals of the overall weighting scheme $\mathbf{a}_1$ of S_1. Finally, noting that both $\widehat{\mathbf{t}}_{\mathbf{a}_1}$ as well as $\widehat{\mathbf{t}}_{\phi_t}$ are based on S, both can be approximated by a linear combination of HT-estimators because of (4.5), (4.6), and (5.12). Hence,

$$\widehat{\mathbf{t}}_{\Gamma}^{RW} \approx \widehat{\mathbf{t}}_{\Gamma}^{d(S_1)} + \sum_{k=1}^{2} \sum_{t=1}^{T_k} \mathbf{M}_t^{(k)} \widehat{\mathbf{t}}_{\phi_t}^{d(S_k)} + \mathbf{c}, \tag{5.16}$$

where the first summation runs over the two samples and the second summation over tables whose underlying variables are used in the re-weighting procedure and/or the overall weighting schemes. The $\mathbf{M}$-matrices consist of sums of products of population regression coefficients and depend also on λ, and $\mathbf{c}$ is a vector of sums of products of population regression coefficients multiplied by population totals. The exact form of an $\mathbf{M}$-matrix naturally depends on Γ, the involved tables ϕ_t, and the overall weighting schemes of

involved micro subsets. Given these $\mathbf{M}$-matrices, one may approximate the variance of $\widehat{\mathbf{t}}_\Gamma^{RW}$ by

$$\begin{aligned}
\mathrm{Var}(\widehat{\mathbf{t}}_\Gamma^{RW}) &\approx \mathrm{Var}\left(\widehat{\mathbf{t}}_\Gamma^{d(S_1)} + \sum_{k=1}^{2}\sum_{t=1}^{T_k} \mathbf{M}_t^{(k)}\widehat{\mathbf{t}}_{\phi_t}^{d(S_k)}\right) \\
&= \mathrm{Var}\left(\widehat{\mathbf{t}}_\Gamma^{d(S_1)} + \sum_{t=1}^{T_1} \mathbf{M}_t^{(1)}\widehat{\mathbf{t}}_{\varphi_t}^{d(S_1)} + \sum_{t=1}^{T_2} \mathbf{M}_t^{(2)}\widehat{\mathbf{t}}_{\varphi_t}^{d(S_2)}\right) \\
&= \mathrm{Var}\left(\sum_{i\in S_1}\pi_{1i}^{-1}\left(\Gamma_i + \sum_{t=1}^{T_1}\mathbf{M}_t^{(1)}\phi_{ti}\right) + \sum_{i\in S_2}\pi_{2i}^{-1}\left(\sum_{t=1}^{T_2}\mathbf{M}_t^{(2)}\phi_{ti}\right)\right) \\
&\equiv \mathrm{Var}\left(\sum_{i\in S_1}\pi_{1i}^{-1}\mathbf{e}_{1i} + \sum_{i\in S_2}\pi_{2i}^{-1}\mathbf{e}_{2i}\right) \\
&= \sum_{i\in\Omega}\sum_{j\in\Omega}(\pi_{1ij} - \pi_{1i}\pi_{1j})\frac{\mathbf{e}_{1i}\mathbf{e}_{1j}^t}{\pi_{1i}\pi_{1j}} + \sum_{i\in\Omega}\sum_{j\in\Omega}(\pi_{2ij} - \pi_{2i}\pi_{2j})\frac{\mathbf{e}_{2i}\mathbf{e}_{2j}^t}{\pi_{2i}\pi_{2j}},
\end{aligned} \tag{5.17}$$

where we introduced the new variables

$$\mathbf{e}_{1i} = \Gamma_i + \sum_{t=1}^{T_1}\mathbf{M}_t^{(1)}\phi_{ti} \quad , \quad \mathbf{e}_{2i} = \sum_{t=1}^{T_2}\mathbf{M}_t^{(2)}\phi_{ti}.$$

Corresponding variance estimators for with replacement designs are recently implemented by Statistics Netherlands in a software package called VRD; see SNIJDERS and HOUBIERS (2002) as well as the formula below (5.8) and its accompanying comment for a similar situation. Given a specific table to be re-weighted and the available micro datasets, this package automatically derives the $\mathbf{e}_k$-variables for $k = 1, \ldots, K$; see also the next subsection.

5.2.3 Recursions for calculating the *e*-variables

For computational convenience we present in this subsection a recursive relationship between the $\mathbf{e}_1$- and $\mathbf{e}_2$-variables corresponding to the estimator $\widehat{\mathbf{t}}_\Gamma^{RW}$ on the one hand, and the $\mathbf{e}_1$- and $\mathbf{e}_2$-variables corresponding to the estimators $\widehat{\mathbf{t}}_{\Gamma_q^-}^{RW}$ on the other hand ($q = 1, \ldots, m$); cf. (4.8) and (5.17). For the sake of simplicity, we confine ourselves to the same case as before, i.e., one registration and two independent samples while the RW-estimator is based on S_1. It is not difficult to generalize these results for more than two independent samples.

By construction the first RW-estimator can always be written as a linear combination of regression estimators from samples S_1 and S_2 provided the estimated $\widehat{\mathbf{B}}$-matrices are replaced by their population analogs. Writing these regression estimators as linear combinations of HT-estimators is straightforward; see (4.6) and the comments below that equality. Hence, the $\mathbf{e}_1$- and $\mathbf{e}_2$- variables corresponding to the first RW-estimator are easily found.

Now in order to derive the required recursion formula in an arbitrary step of the RW-estimation procedure, write according to the notation used in (5.17) $\widehat{\mathbf{t}}_{\Gamma}^{RW}$ as a sum of HT-estimators of two **e**-variables. That is,

$$\widehat{\mathbf{t}}_{\Gamma}^{RW} = \widehat{\mathbf{t}}_{\mathbf{e}_1(\Gamma)}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\Gamma)}^{d(S_2)} + \mathbf{c}_{\Gamma}.$$

Likewise, for the most detailed margins

$$\widehat{\mathbf{t}}_{\Gamma_q^-}^{RW} = \widehat{\mathbf{t}}_{\mathbf{e}_1(\Gamma_q^-)}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\Gamma_q^-)}^{d(S_2)} + \mathbf{c}_{\Gamma_q^-} \qquad (q = 1, \ldots, m). \tag{5.18}$$

Also write the vector $\widehat{\mathbf{t}}_{\mathbf{a}}$ of (counted or estimated) population totals of the auxiliary variables as a sum of two HT-estimators plus a constant. In obvious notation,

$$\widehat{\mathbf{t}}_{\mathbf{a}} = \widehat{\mathbf{t}}_{\mathbf{e}_1(\mathbf{a})}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\mathbf{a})}^{d(S_2)} + \mathbf{c}_{\mathbf{a}}. \tag{5.19}$$

Recall from (4.8) that the RW-estimator for the actual sample S_1 is equal to

$$\widehat{\mathbf{t}}_{\Gamma}^{RW} = \widehat{\mathbf{t}}_{\Gamma}^{w(S_1)} + \sum_{q=1}^{m} (\widehat{\mathbf{B}}_{\Gamma;\Gamma_q^-;\mathbf{m}}^{w(S_1)})^t (\widehat{\mathbf{t}}_{\Gamma_q^-}^{RW} - \widehat{\mathbf{t}}_{\Gamma_q^-}^{w(S_1)}). \tag{5.20}$$

Making use of (4.6) and (5.19), we can write $\widehat{\mathbf{t}}_{\Gamma}^{w(S_1)}$ as

$$\widehat{\mathbf{t}}_{\Gamma}^{w(S_1)} = \widehat{\mathbf{t}}_{\Gamma}^{d(S_1)} + (\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}}^{d(S_1)})^t (\widehat{\mathbf{t}}_{\mathbf{e}_1(\mathbf{a})}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\mathbf{a})}^{d(S_2)} + \mathbf{c}_{\mathbf{a}} - \widehat{\mathbf{t}}_{\mathbf{a}}^{d(S_1)}). \tag{5.21}$$

Also,

$$\widehat{\mathbf{t}}_{\Gamma_q^-}^{w(S_1)} = \widehat{\mathbf{t}}_{\Gamma_q^-}^{d(S_1)} + (\widehat{\mathbf{B}}_{\Gamma_q^-;\mathbf{a}}^{d(S_1)})^t (\widehat{\mathbf{t}}_{\mathbf{e}_1(\mathbf{a})}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\mathbf{a})}^{d(S_2)} + \mathbf{c}_{\mathbf{a}} - \widehat{\mathbf{t}}_{\mathbf{a}}^{d(S_1)}) \qquad (q = 1, \ldots, m). \tag{5.22}$$

Now combining (5.18)-(5.22), we can write $\widehat{\mathbf{t}}_{\Gamma}^{RW}$ as

$$\begin{aligned}
\widehat{\mathbf{t}}_{\Gamma}^{RW} &= \widehat{\mathbf{t}}_{\Gamma}^{d(S_1)} + (\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}}^{d(S_1)})^t (\widehat{\mathbf{t}}_{\mathbf{e}_1(\mathbf{a})}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\mathbf{a})}^{d(S_2)} - \widehat{\mathbf{t}}_{\mathbf{a}}^{d(S_1)}) \\
&\quad + \sum_{q=1}^{m} (\widehat{\mathbf{B}}_{\Gamma;\Gamma_q^-;\mathbf{m}}^{w(S_1)})^t \{\widehat{\mathbf{t}}_{\mathbf{e}_1(\Gamma_q^-)}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\Gamma_q^-)}^{d(S_2)} - \widehat{\mathbf{t}}_{\Gamma_q^-}^{d(S_1)} - (\widehat{\mathbf{B}}_{\Gamma_q^-;\mathbf{a}}^{d(S_1)})^t (\widehat{\mathbf{t}}_{\mathbf{e}_1(\mathbf{a})}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\mathbf{a})}^{d(S_2)} - \widehat{\mathbf{t}}_{\mathbf{a}}^{d(S_1)})\} + \mathbf{c}_{\Gamma} \\
&\equiv \widehat{\mathbf{t}}_{\mathbf{e}_1(\Gamma)}^{d(S_1)} + \widehat{\mathbf{t}}_{\mathbf{e}_2(\Gamma)}^{d(S_2)} + \mathbf{c}_{\Gamma},
\end{aligned}$$

where the variables $\mathbf{e}_{1i}(\Gamma)$ and $\mathbf{e}_{2i}(\Gamma)$ are defined recursively by

$$\begin{aligned}
\mathbf{e}_{1i}(\Gamma) &\equiv \Gamma_i + \sum_{q=1}^{m} (\widehat{\mathbf{B}}_{\Gamma;\Gamma_q^-;\mathbf{m}}^{w(S_1)})^t \{\mathbf{e}_{1i}(\Gamma_q^-) - \Gamma_{qi}^-\} \\
&\quad + \left((\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}}^{d(S_1)})^t - \sum_{q=1}^{m} (\widehat{\mathbf{B}}_{\Gamma;\Gamma_q^-;\mathbf{m}}^{w(S_1)})^t (\widehat{\mathbf{B}}_{\Gamma_q^-;\mathbf{a}}^{d(S_1)})^t \right) \{\mathbf{e}_{1i}(\mathbf{a}) - \mathbf{a}_{\mathbf{i}}\} \\
\mathbf{e}_{2i}(\Gamma) &\equiv \sum_{q=1}^{m} (\widehat{\mathbf{B}}_{\Gamma;\Gamma_q^-;\mathbf{m}}^{w(S_1)})^t \mathbf{e}_{2i}(\Gamma_q^-) + \left((\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}}^{d(S_1)})^t - \sum_{q=1}^{m} (\widehat{\mathbf{B}}_{\Gamma;\Gamma_q^-;\mathbf{m}}^{w(S_1)})^t (\widehat{\mathbf{B}}_{\Gamma_q^-;\mathbf{a}}^{d(S_1)})^t \right) \mathbf{e}_{2i}(\mathbf{a}).
\end{aligned}$$

It should be noted that the structure of these recursions resembles strongly the structure of the first equality in (5.5). In addition, these recursions suggest at least that for multi-dimensional tables it is no longer self-evident that the random character of the regression coefficients can be ignored because there are so many of them. In Chapter 6 as well as in

the case study we will pay more attention to the question of to what extent the random character of the regression coefficients can be ignored.

An interesting interpretation of the **e**-variables is as follows. In case of SRS sampling it is well known that after replacing the estimated regression coefficients in a regression estimator by their population analogs the variance of the resulting regression estimator is equal to the variance of the *sample* mean of the corresponding *population* residuals from the underlying regression, provided that the variables are rescaled with a factor N for the population total to be estimated. As we have seen above, the first RW-estimator appearing in the RW-estimation procedure can always be written as a linear combination of regression estimators. Consequently, repeatedly applying the RW-estimator always yields estimators which remain within the class of linear combinations of regression estimators. Apart from a constant the **e**-variable or the so-called superresidual per sample can be interpreted as the corresponding linear combination of the *population* residuals from the successive regression estimators involved in the RW-estimator. Discarding a constant term, for unequal probability sampling the RW-estimator can be interpreted similarly as a sum of HT-estimators of the population totals of the superresiduals ($\mathbf{e}_{Si}$) from the corresponding underlying samples S; see also KNOTTNERUS (2001).

5.3 One registration and two dependent samples; two-phase sampling

In this section, variance approximation formulas for the RW-estimator in case of two-phase sampling are derived. Consider one register R, a first-phase sample S, and one second-phase sample S_1. Both the first and the second-phase samples are drawn according to some complex design without replacement. Let π_i and π_{ij} denote the first and second order inclusion probabilities with respect to the first-phase sample S for $i, j \in \Omega$, respectively. Given the realization of the first-phase sample, the first and second-order inclusion probabilities with respect to the second-phase sample are denoted by $\pi_{i|S}$ and $\pi_{ij|S}$, $i, j \in S$. The starting weights for S are $d_i^{(S)} = \pi_i^{-1}$ and, given S, for S_1, $d_i^{(S_1)} = \pi_i^{-1}\pi_{i|S}^{-1}$. SÄRNDAL *et al.* (1992), Sections 9.2 and 9.3 call the resulting two-phase estimator a π^*-estimator, to emphasize the fact that it is not exactly a HT-estimator. The overall weighting scheme for S is represented by $\mathbf{a}$. This scheme consists of register variables only. The overall weighting scheme for S_1 is represented by $\mathbf{a}_1$. This scheme may consist of register variables and variables observed in the first-phase sample.

We distinguish between RW-estimators based on S and RW-estimators based on S_1. The former situation is completely covered by the case of one register and one sample, discussed in section 5.1 This subsection deals with RW-estimators based on S_1. Now, the only difference with the case discussed in subsection 5.2.2 where S consisted of two independent samples, is the sampling design. Moreover, the difference is essentially in the second-order inclusion probabilities. Therefore we can simply repeat here the RW-estimator (5.15)

$$\widehat{\mathbf{t}}_\Gamma^{RW} = \widehat{\mathbf{t}}_\Gamma^{d(S_1)} + \left[(\widehat{\mathbf{B}}_{\Gamma;\mathbf{a}_1}^{d(S_1)})^t - \sum_{t=1}^{T} (\widehat{\mathbf{C}}_{\Gamma;\phi_t}^{w(S_1)})^t (\widehat{\mathbf{B}}_{\phi_t;\mathbf{a}_1}^{d(S_1)})^t \right] (\widehat{\mathbf{t}}_{\mathbf{a}_1} - \widehat{\mathbf{t}}_{\mathbf{a}_1}^{d(S_1)})$$

$$+\left[\sum_{t=1}^{T}(\widehat{\mathbf{C}}_{\Gamma;\phi_t}^{w(S_1)})^t\right](\widehat{\mathbf{t}}_{\phi_t}-\widehat{\mathbf{t}}_{\phi_t}^{d(S_1)}).$$

However, in the context of two-phase sampling it is more appropriate to replace the estimated parts of $\widehat{\mathbf{t}}_{\mathbf{a}_1}$ and $\widehat{\mathbf{t}}_{\phi_t}$ in components based on S and S_1 rather than on the two subsamples S_1 and S_2. As we have seen before, it follows that an RW-estimator based on S_1 can be approximated by a linear combination of HT-estimators, which are partly based on S and partly on S_1:

$$\widehat{\mathbf{t}}_{\Gamma}^{RW}\approx\widehat{\mathbf{t}}_{\Gamma}^{d(S_1)}+\sum_{t=1}^{T}\mathbf{M}_t\widehat{\mathbf{t}}_{\phi_t}^{d(S)}+\sum_{t=1}^{T_1}\mathbf{M}_t^{(1)}\widehat{\mathbf{t}}_{\phi_t}^{d(S_1)}+\mathbf{c}. \tag{5.23}$$

The variance of (5.23) can be approximated using the standard approach for two-phase sampling. That is, conditioning on the realization of S in the first phase, we have:

$$\mathrm{Var}(\widehat{\mathbf{t}}_{\Gamma}^{RW})=\mathrm{EVar}(\widehat{\mathbf{t}}_{\Gamma}^{RW}\,|S)+\mathrm{VarE}\,(\widehat{\mathbf{t}}_{\Gamma}^{RW}\,|S),$$

where

$$\begin{aligned}\mathrm{EVar}(\widehat{\mathbf{t}}_{\Gamma}^{RW}\,|S)&\approx\mathrm{EVar}\left((\widehat{\mathbf{t}}_{\Gamma}^{d(S_1)}+\sum_{t=1}^{T_1}\mathbf{M}_t^{(1)}\widehat{\mathbf{t}}_{\phi_t}^{d(S_1)})\,|S\right)\\&=\mathrm{E}\left(\sum_{i\in S}\sum_{j\in S}(\pi_{ij|S}-\pi_{i|S}\pi_{j|S})\frac{\mathbf{e}_{1i}\mathbf{e}_{1j}^t}{\pi_i\pi_{i|S}\pi_j\pi_{j|S}}\right),\end{aligned}$$

with $\mathbf{e}_{1i}=\Gamma_i+\sum_{t=1}^{T_1}\mathbf{M}_t^{(1)}\phi_{ti}$, and

$$\mathrm{VarE}(\widehat{\mathbf{t}}_{\Gamma}^{RW}\,|S)\approx\mathrm{Var}\left((\widehat{\mathbf{t}}_{\Gamma}^{d(S)}+\sum_{t=1}^{T_1}\mathbf{M}_t^{(1)}\widehat{\mathbf{t}}_{\phi_t}^{d(S)}+\sum_{t=1}^{T}\mathbf{M}_t\widehat{\mathbf{t}}_{\phi_t}^{d(S)}\right)=\sum_{i\in\Omega}\sum_{j\in\Omega}(\pi_{ij}-\pi_i\pi_j)\frac{\mathbf{e}_i\mathbf{e}_j^t}{\pi_i\pi_j},$$

with $\mathbf{e}_i=\Gamma_i+\sum_{t=1}^{T}\mathbf{M}_t\phi_{ti}+\sum_{t=1}^{T_1}\mathbf{M}_t^{(1)}\phi_{ti}$. As a result the following approximate expression for the variance of (5.23) is obtained:

$$\mathrm{Var}(\widehat{\mathbf{t}}_{\Gamma}^{RW})=\mathrm{E}\left(\sum_{i\in S}\sum_{j\in S}(\pi_{ij|S}-\pi_{i|S}\pi_{j|S})\frac{\mathbf{e}_{1i}\mathbf{e}_{1j}^t}{\pi_i\pi_{i|S}\pi_j\pi_{j|S}}\right)+\sum_{i\in\Omega}\sum_{j\in\Omega}(\pi_{ij}-\pi_i\pi_j)\frac{\mathbf{e}_i\mathbf{e}_j^t}{\pi_i\pi_j}. \tag{5.24}$$

An estimator for the approximate variance of (5.23) is given by

$$\widehat{\mathrm{Var}}(\widehat{\mathbf{t}}_{\Gamma}^{RW})=\sum_{i\in S_1}\sum_{j\in S_1}\frac{(\pi_{ij|S}-\pi_{i|S}\pi_{j|S})}{\pi_{ij|S}}\frac{\widehat{\mathbf{e}}_{1i}\widehat{\mathbf{e}}_{1j}^t}{\pi_i\pi_{i|S}\pi_j\pi_{j|S}}+\sum_{i\in S_1}\sum_{j\in S_1}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}\pi_{ij|S}}\frac{\widehat{\mathbf{e}}_i\widehat{\mathbf{e}}_j^t}{\pi_i\pi_j}, \tag{5.25}$$

with $\widehat{\mathbf{e}}_{1i}=\Gamma_i+\sum_{t=1}^{T_1}\widehat{\mathbf{M}}_t^{(1)}\phi_{ti}$ and $\widehat{\mathbf{e}}_i=\Gamma_i+\sum_{t=1}^{T}\widehat{\mathbf{M}}_t\phi_{ti}+\sum_{t=1}^{T_1}\widehat{\mathbf{M}}_t^{(1)}\phi_{ti}$, see Särndal *et al.* (1992), Chapter 9.

Note that the first term on the right-hand side of (5.25) is an estimator for the first term on the right-hand side of (5.24). Therefore, the first variance component in (5.24) can be conveniently stated as an expected value over the first phase.

Chapter 6

Alternative approximations of the RW-estimator and its variance

In Chapter 5, approximation variance formulas for the RW-estimator that result from a full splitting-up procedure are derived. These formulas, however, are not very informative. The residual terms may have complicated forms depending on **M**-matrices, and it is hard to see if and how the splitting-up procedure influences the accuracy of the estimates. HOUBIERS *et al.* (2003) have performed some simulations to get insight into both the accuracy of the RW-estimator itself and the accuracy of the approximation of the variance estimator. More insight into these variance formulas can be obtained by simplifying the expression of the RW-estimator. In this section this is done by assuming some approximation conditions that are satisfied in most practical situations. For DACSEIS deliverable D7.3 an additional simulation study is carried out in order to get more insight into the performance of the RW-estimator and the corresponding variance formulas derived so far.

Continuing the situation of one register and one sample (see Section 5.2), we describe the splitting-up procedure of the table $\mathbf{t}_6 = \text{region}^{(1)} \times \text{age}^{(1)} \times \text{employ}^{(1)} = \mathbf{x}_1 \times \mathbf{x}_3 \times \mathbf{x}_4$, where, for short, the superscripts '(1)' are omitted. The RW-estimator is expressed in terms of regression type estimators, and it is argued how this estimator can be simplified. The estimation of $\mathbf{t}_6$ according to the splitting-up procedure starts with:

$$\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4} = \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4} + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_3\times\mathbf{x}_4+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ \widehat{\mathbf{t}}^{RW}_{\mathbf{x}_3\times\mathbf{x}_4} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4} \\ \widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_4} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_4} \end{pmatrix}.$$

Subsequently, $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_3}$, $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_3\times\mathbf{x}_4}$, and $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_4}$ must be estimated. The proper micro subset for the first table is the registration: $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_3} = \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3}$. The proper micro subset for the second and third table is the sample. The second table is estimated by

$$\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_3\times\mathbf{x}_4} = \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4} + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_3+\mathbf{x}_4})^t \begin{pmatrix} \widehat{\mathbf{t}}^{RW}_{\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_3} \\ \widehat{\mathbf{t}}^{RW}_{\mathbf{x}_4} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_4} \end{pmatrix}.$$

Since, the third table satisfies the decomposition criterion it is estimated by

$$\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_4} = \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_4},$$

see text below (4.2). Finally $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_3}$ and $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_4}$ must be estimated. The first table is estimated from the registration, i.e. $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_3} = \mathbf{t}_{\mathbf{x}_3}$ and the second table from the sample, i.e. $\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_4} = \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_4}$. Collecting results, the following RW-estimator is obtained:

$$\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4} = \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4} + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_3\times\mathbf{x}_4+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_3+\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_3} \\ 0 \end{pmatrix} \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{pmatrix}. \tag{6.1}$$

In the following it is suggested that the expression of the RW-estimator can be approximated by a simpler one. To illustrate this, consider the inner adjustment term:

$$(\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_3+\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_3} \\ 0 \end{pmatrix}.$$

This term can be interpreted as an adjustment term to improve the regression type estimate for $\mathbf{t}_{\mathbf{x}_3^{(1)}\mathbf{x}_4^{(1)}}$, where the most detailed margins of $\mathbf{t}_{\mathbf{x}_3^{(1)}\mathbf{x}_4^{(1)}}$, i.e. $\mathbf{x}_3^{(1)} + \mathbf{x}_4^{(1)}$, are used as explanatory variables. If these margins are the most important explanatory variables, then adding some more explanatory variables will not alter the adjustment term largely. In particular, the following approximation may be valid for many practical situations:

$$(\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_3+\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_3} \\ 0 \end{pmatrix} \approx (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix}, \tag{6.2}$$

i.e. $\mathbf{x}_1 \times \mathbf{x}_3 + \mathbf{x}_1 \times \mathbf{x}_4$ is used as re-weighting scheme for $\mathbf{t}_{\mathbf{x}_3\times\mathbf{x}_4}$, instead of $\mathbf{x}_3 + \mathbf{x}_4$. This particular choice is motivated by the decomposition criterion for $\mathbf{t}_6$. Namely, $\mathbf{x}_1 \times \mathbf{x}_3$ and $\mathbf{x}_1 \times \mathbf{x}_4$ are the only margins of $\mathbf{t}_6$ that satisfy the decomposition criterion. Accepting (6.2), (6.1) can be approximated as

$$\begin{aligned} \widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4} \approx \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4} + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_3\times\mathbf{x}_4+\mathbf{x}_1\times\mathbf{x}_4})^t \\ \times \begin{pmatrix} (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix} \\ (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix} \\ (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix} \end{pmatrix}. \end{aligned} \tag{6.3}$$

Due to the choice of the weighting scheme $\mathbf{x}_1^{(1)} \times \mathbf{x}_3^{(1)} + \mathbf{x}_1^{(1)} \times \mathbf{x}_4^{(1)}$, the first and third inner adjustment terms equal the first and third inner terms of (6.1), since effectively $\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4} = (\mathbf{I}\ \mathbf{0})^t$ and $\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4} = (\mathbf{0}\ \mathbf{I})^t$. So, the approximation of (6.3) is only due to the second inner term. Next, (6.3) can be rewritten as an expression with a single regression adjustment term

$$\widehat{\mathbf{t}}^{RW}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4} \approx \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4} + (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix}. \tag{6.4}$$

This expression uses all most detailed margins of $\mathbf{t}_6$ that satisfy the decomposition criterion as auxiliary information.

The equality between the right-hand side of (6.3) and the right-hand side of (6.4) can be seen from the following equations:

$$
\begin{aligned}
& (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_3\times\mathbf{x}_4+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix} \\ (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix} \\ (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix} \end{pmatrix} \\
= \; & (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_3\times\mathbf{x}_4+\mathbf{x}_1\times\mathbf{x}_4})^t (\widehat{\mathbf{B}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_3\times\mathbf{x}_4+\mathbf{x}_1\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix} \\
= \; & (\widehat{\mathbf{B}}^{(S)}_{\mathbf{x}_1\times\mathbf{x}_3\times\mathbf{x}_4;\mathbf{x}_1\times\mathbf{x}_3+\mathbf{x}_1\times\mathbf{x}_4})^t \begin{pmatrix} \mathbf{t}_{\mathbf{x}_1\times\mathbf{x}_3} - \widehat{\mathbf{t}}^{w(S)}_{\mathbf{x}_1\times\mathbf{x}_3} \\ 0 \end{pmatrix} .
\end{aligned}
$$

The first equality holds by definition of the regression coefficients. The second equality is an application of the following simple result. Let ξ_i, ζ_i, and η_i denote some (multiple) categorical variables, assume that $\eta_i = \mathbf{A}\zeta_i$ for all i for some constant matrix $\mathbf{A}$, and consider the following regression coefficients:

$$
\widehat{\mathbf{B}}_{\xi;\zeta} = \left(\sum_{i\in S} w_i\zeta_i\zeta_i^t\right)^{-1} \sum_{i\in S} w_i\zeta_i\xi_i^t, \quad \widehat{\mathbf{B}}_{\zeta;\eta} = \left(\sum_{i\in S} w_i\eta_i\eta_i^t\right)^{-1} \sum_{i\in S} w_i\eta_i\zeta_i^t, \text{ and}
$$
$$
\widehat{\mathbf{B}}_{\xi;\eta} = \left(\sum_{i\in S} w_i\eta_i\eta_i^t\right)^{-1} \sum_{i\in S} w_i\eta_i\xi_i^t .
$$

Then it follows that

$$
\begin{aligned}
\widehat{\mathbf{B}}^t_{\xi;\zeta}\widehat{\mathbf{B}}^t_{\zeta;\eta} &= \sum_{i\in S} w_i\xi_i\zeta_i^t \left(\sum_{i\in S} w_i\zeta_i\zeta_i^t\right)^{-1} \sum_{i\in S} w_i\zeta_i\eta_i^t \left(\sum_{i\in S} w_i\eta_i\eta_i^t\right)^{-1} \\
&= \sum_{i\in S} w_i\xi_i\zeta_i^t \left(\sum_{i\in S} w_i\zeta_i\zeta_i^t\right)^{-1} \sum_{i\in S} w_i\zeta_i\zeta_i^t\mathbf{A}^t \left(\sum_{i\in S} w_i\eta_i\eta_i^t\right)^{-1} = \widehat{\mathbf{B}}^t_{\xi;\eta}.
\end{aligned}
$$

Since re-weighting schemes always consist of margins of the target table, the relation $\eta_i = \mathbf{A}\zeta_i$ is always fulfilled for the RW-estimators.

To generalize, let $\Gamma_1^-, \ldots, \Gamma_L^-$ denote all margins of Γ encountered in the distinct steps of the RW-estimation procedure that satisfy the decomposition criterion. Without loss of generality, suppose that the registration is the proper micro subset for $\Gamma_1^-, \ldots, \Gamma_K^-$ and that the sample is the proper micro subset for $\Gamma_{K+1}^-, \ldots, \Gamma_L^-$, $K \leqslant L$. Let $\widehat{\mathbf{t}}_\Gamma^{RW}$ denote the

RW-estimator that results from the splitting-up procedure. Then, this estimator can be approximated by

$$\widehat{\mathbf{t}}_{\Gamma}^{RW} \approx \widehat{\mathbf{t}}_{\Gamma}^{w(S)} + (\widehat{\mathbf{B}}^{w(S)}_{\Gamma;\Gamma_1^-,\ldots,\Gamma_K^-;\Gamma_1^-,\ldots,\Gamma_L^-})^t \begin{pmatrix} \mathbf{t}_{\Gamma_1^-} - \widehat{\mathbf{t}}^{w(S)}_{\Gamma_1^-} \\ \vdots \\ \mathbf{t}_{\Gamma_K^-} - \widehat{\mathbf{t}}^{w(S)}_{\Gamma_K^-} \end{pmatrix} \tag{6.5}$$

$$= \widehat{\mathbf{t}}_{\Gamma}^{w(S)} + \sum_{j=1}^{K} (\widehat{\mathbf{B}}^{w(S)}_{\Gamma;\Gamma_j^-;\Gamma_1^-,\ldots,\Gamma_L^-})^t (\mathbf{t}_{\Gamma_j^-} - \widehat{\mathbf{t}}^{w(S)}_{\Gamma_j^-}). \tag{6.6}$$

Naturally, this approximation can be elaborated in terms of HT-estimators. Note that the RW-estimator coincides exactly with (6.5) if all margins of Γ satisfy the decomposition criterion. For example, the RW-estimator for $\mathbf{t}_3$ as given by (5.4) has the same form as (6.5). For the estimation of tables, the original RW-estimator from Section 4 is used. To facilitate the variance computation, the simplified RW-estimator suggested in this section can be used, provided the approximation conditions like (6.2) are satisified.

It should be noted that the approximation of the RW-estimator described in this section can also be used as a justification for the variance approximation proposed in the previous section. This variance approximation is based on a Taylor series expansion leading to neglecting a number of random (estimated) $\mathbf{B}$-matrices. For instance, assuming that all classifications have five different classes, it can be seen from (6.1) that each cell of the study table $\mathbf{t}_6$ is based on $175 (= 25 + 150)$ estimated regression coefficients. Note that 25 corresponds to the number of regression coefficients of $\mathbf{t}_{\mathbf{x}_1 \times \mathbf{x}_3}$ in the first adjustment term in (6.1), while 150 (=25+5×25) corresponds to the number of regression coefficients involved in the second adjustment term for $\mathbf{t}_{\mathbf{x}_3}$. In contrast, (6.4) is only based on 25 estimated regression coefficients. In a worst case scenario this means that in (6.1) the effect of the 175 errors of order $1/\sqrt{n}$ may exceed substantially the effect of the 25 errors of order $1/\sqrt{n}$ in (6.4). For a survey with, for instance, $n = 100,000$ this difference might be substantial, especially for categorical variables. However, when under the regularity conditions the approximation of the RW-estimator is accurate, this means that the Taylor series approximation of the original RW-estimator is accurate as well, provided that the error of the Taylor series expansion of the approximate RW-estimator is negligible indeed.

Chapter 7

Summary and further research

In this deliverable an estimation strategy for combined data sources is proposed. The reasons to combine data sources and to come up with a new estimation strategy are political pressure to reduce response burden and to accommodate user demands to produce outputs that are numerically consistent. The estimation strategy is based on regression techniques but not necessarily on one weighting scheme per survey. It involves three steps: 1) constructing rectangular micro subsets from the combined data sources, 2) assigning to each micro subset a (fixed) set of regression weights according to some weighting scheme, and 3) for each target table adjusting the estimates based on the original weighting scheme using a re-weighting scheme that is tailored to the consistency demand. The estimation strategy is 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 consistent estimates assumes the existence of 'perfect' micro databases, i.e. micro databases that consist of (not too large) a number of rectangular micro datasets that are consistent at the micro level. However, constructing such micro databases is a very complex task, in which many difficult choices have to be made. In addition, due to e.g. discrepancies in definitions, different questionnaires, data collection methods or different reference periods between surveys, it is difficult to obtain common variables in real terms. In practice, common variables between different surveys can be obtained through vigorous harmonization of the different phases of the survey designs.

This deliverable focuses on classification variables in order to avoid a number of complications with respect to quantitative variables. Two complications are the duality of quantitative variables and the phenomenon of 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 concerning sub-variables 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 have developed a first methodology to cope with these complications, but this research is still ongoing.

Appendix A

Appendix

A.1 Estimates of five target distributions according to the weighting model "region + sex".

Table A.1: region$^{(2)}$ $\times$ employ$^{(1)}$($= \mathbf{t}_3$)

	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 A.2: employ$^{(1)}$ $\times$ age$^{(1)}$($= \mathbf{t}_4$)

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

Table A.3: sex$^{(1)}$ $\times$ employ$^{(1)}$ $\times$ region$^{(1)}$($= \mathbf{t}_5$)

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

A.2 Estimates of five target tables according to reweighting scheme, see table 3.3

Table A.4: region$^{(2)}$ × employ$^{(1)}$(= $\mathbf{t}_3$)

	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 A.5: employ$^{(1)}$ × age$^{(1)}$(= $\mathbf{t}_4$)

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

Table A.6: sex$^{(1)}$ × employ$^{(1)}$ × region$^{(1)}$(= $\mathbf{t}_5$)

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

References

Deville, J. C. and Särndal, C. E. (1992): Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Houbiers, M., Knottnerus, P., Kroese, A. H., Renssen, R. H. and Snijders, V. (2003): *Estimating consistent table sets: Position paper on repeated weighting.* Voorburg: Statistics Netherlands.

Knottnerus, P. (2001): *Variances in Repeated Weighting (in Dutch).* Voorburg: Statistics Netherlands.

Kroese, A. H. and Renssen, R. H. (1999): *Weighting and imputation at Statistics Netherlands (draft version).* Presented at the International Association of Survey Statisticians Satellite Conference. Small Area Estimation, Riga, Latvia, 20-21 August.

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**, 23–31. Special Issue: Integrating administrative registers and household surveys.

Renssen, R. H., Kroese, A. H. and Willeboordse, A. (2001): *Aligning Estimates by Repeated Weighting.* Heerlen: 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**, 368–374.

Särndal, C. E., Swensson, B. and Wretman, J. (1992): *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Snijders, V. and Houbiers, M. (2002): *Variance estimation in repeated weighting for VRD 1.4 (in Dutch).* Voorburg: Statistics Netherlands.

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

Willeboordse, A. and Ypma, W. (1996): From rules to tools. New opportunities to establish coherence among statistics. In *Proceedings of the Conference on output Databases.* Voorburg: Statistics Netherlands.

Willeboordse, A. and Ypma, W. (1998): *Meta Tools in support of a Corporate Dissemination Strategy.* Voorburg: Statistics Netherlands.