Edit rules and the strategy of Consistent Table Estimation

Discussion paper 04013 (version 1)

Rob van de Laar

The views expressed in this paper are those of the authors
and do not necessarily reflect the policies of Statistics Netherlands

Statistics Netherlands

Voorburg/Heerlen, December 2004

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2003–2004	= 2003 to 2004 inclusive
2003/2004	= average of 2003 up to and including 2004
2003/’04	= crop year, financial year, school year etc. beginning in 2003 and ending in 2004

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314
Key figure: X-10
Production code: 6008304013

Statistics Netherlands

Edit rules and the strategy of Consistent Table Estimation

Abstract: The method of repeated weighting aims at obtaining numerical consistency among tables estimated from different surveys. However, in its common form, it does not take into account the existing edit rules. Consequently, the repeated weighting estimates will generally be not in agreement with existing edit rules. This report describes how to deal with linear categorical and numerical edit rules within the framework of repeated weighting estimation. Some basic examples will be given of tables estimated in agreement with edit rules. It will be shown that the presence of edit rules affects the order in which tables should be estimated. Particularly the splitting-up order of table estimation is shown to be in easy conflict with existing edit rules. A step-by-step plan will be proposed of an estimation procedure yielding numerically consistent tables in agreement with edit rules.

Keywords: consistency, linear edit rules, repeated weighting, general regressions estimators, weighting schemes, estimation strategy.

1. Introduction

In section 2 the repeated weighting (RW) procedure will be roughly described, with an emphasis on its limitations when applied in a practical situation with many edit rules and samples of limited size. Examples will be given of the application of different weighting schemes, in order to show that resulting estimated tables can either agree or disagree with an edit rule. Section 3 deals with linear edit rules, dividing them in conditional and unconditional edit rules, with numerical or categorical variables. As a result it is recommended to estimate any variable or variables that are part of an edit rule simultaneously with at least all register variables of the same edit rule. All other variables that are part of an edit rule - but not of the register - may be estimated simultaneously as well. If some estimated variables of an edit rule are estimated before, that earlier estimate should appear as a marginal term in the weighting scheme - like the register variables - for numerical consistency. Eventually, it should also be possible to estimate all variables of an edit rule simultaneously from one sample - while calibrating on all available margins - with each cell in the corresponding frequency table greater than or equal to zero. In section 3 some consequences concerning the choice and order of the table estimates will be considered, and in the remainder of section 3 some examples will be given. First in section 3.5 an example is given of a conditional edit rule with categorical variables, in fact a continuation of the example in section 2. Then in 3.6 an unconditional edit rule with numerical variables. In this example several methods of table estimation will be compared, for instance estimating all edit variables simultaneously with or without calibration on register values or on margins that were estimated before. In section 3.6.3, as an alternative approach, data from two different samples is used simultaneously, without either first merging both blocks or estimating one sample before the other. Next, in section 4, a general plan is proposed for a sample design and estimation procedure aimed at numerically consistent estimates that are in agreement with edit rules. However, the proposed approach is not strictly necessary, because for a particular sample the RW estimation procedure may yield estimates in accordance with all existing edit rules by mere chance. But in order to be less dependent of the actual sample data, the proposed step-by-step plan may be used. It will be shown that in that case too, the samples can limit the number of tables that can be estimated consistently. Section 5 summarizes that both numerical consistency and consistency of the estimates with the existing edit rules, make additional demands on the sample sizes and the variables observed in each sample. Generally, the larger the sample sizes, the better. If possible, registries should be substituted for samples.

2. The repeated weighting procedure

The main aim of the repeated weighting (RW) estimation procedure is obtaining a set of tables that are mutually consistent in a numerical sense. Margins of a detailed table must be estimated first if they can be based on a larger amount of data than used for the detailed table. In particular when some variables are part of a register, these known population values must be used, apart from any sample data. Then the RW procedure results in estimated tables that are numerically consistent with the register data and with all previously estimated marginal tables. In the RW procedure this calibration on all margins of each estimated table is accomplished using the regression estimator. In this way reliable tables based on registries or large samples are used to adapt samples consisting of less records, but containing additional variables that are not known otherwise. This weighting or reweighting is a way to correct for the random character of small samples. By weighting or reweighting a relatively small sample is made to resemble the population for all the variables used in calibration. With a sample being only a subset of the population, this resemblance will concern only a limited number of variables.

2.1 Limitations of the RW procedure

In the RW procedure the calibration is accomplished by adjusting the sampling weights by means of the regression estimator. Unfortunately this procedure has its limits. If the information from the register is too rich, e.g. if the register includes categorical variables with many categories, it can be mathematically impossible to adapt the sampling weights so that they exactly reproduce all given population tables from the registers. (After reweighting the sampling weights are usually called regression weights) This will also occur when too many estimated marginal tables are part of the weighting scheme. This can easily be the case while using the splitting-up procedure, because then each margin is estimated first, even if a marginal table cannot be derived from a different larger sample. Due to a too large weighting scheme, the intended small corrections of the first-order inclusion weights of the sample, will degenerate. This can cause negative cell estimates, or - if no solution exists - the reweighting process will crash. Then numerical consistency cannot be accomplished by correction of the design weight using the regression estimator. Only in rather uncomplicated cases the RW procedure can be used to produce numerically consistent tables: if

- the registries do not contain very detailed categorical variables
- relatively few independent data sources are used
- different data sources do not contain many variables in common
- the samples are large enough
- only a limited set of tables is estimated

Because of the random character of the samples, the impact of these conditions on the use of the RW procedure can be different in each actual situation.

2.2 Edit rules

Still one other condition should be fulfilled in order to get acceptable results from the RW procedure. This is related to the edit rules that exist in the actual population. In general, edit rules are some kind of relations between two or more variables on the level of the individual elements in the population. An edit rule applies to all individual elements of the population, or to a subset of all elements that meet some condition. For instance an edit rule can be a relation between numerical variables that only applies if a categorical variable belongs to a particular category. Apart from the individual elements in a population, edit rules do also have an effect on the properties of the whole population, i.e. on the values of the actual population totals. Consequently, the estimated population totals should be consistent with the edit rules. An important question is whether the RW procedure will yield estimates in accordance with the existing edit rules. Otherwise the RW procedure should be adapted so that it will yield estimates consistent with all existing edit rules. This is an important point, because in many cases estimated population totals that violate the edit rules can easily be revealed as being erroneous by an informed user of the estimated tables. The following examples illustrate this.

2.3 Example

A population of $N = 400$ persons, has two municipalities: M_1 and M_2. The variable *age* is part of the register R and has two categories: 18$^-$ and 18$^+$:

Register	*number of inhabitants 18$^-$*	*number of inhabitants 18$^+$*	*total*
municipality M_1	35	75	110
municipality M_2	90	200	290
total	125	275	400

A simple random sample without replacement (SRS) S of size $n = 100$ contains three categorical variables: *age* (18- or 18+), *municipality* (M_1 or M_2), and *possession of a driving licence* (*true* or *false*):

S	*#(18- without driving licence)*	*# (18- with driving licence)*	*# (18+ without driving licence)*	*# (18+ with driving licence)*	*total*
municipality M_1	3	0	2	15	20
municipality M_2	17	0	8	55	80
total	20	0	10	70	100

In this population every person with a driving licence must be at least 18 years old:

$$\textit{if}\ \ (\textit{having a driving licence})\ \ \textit{then}\ \ (\textit{age} \geq 18)$$

In the sample all 100 elements satisfy this edit rule. We use this sample to estimate the frequency table

$$[\textit{municipality}\ \otimes \textit{driving licence}] \times 1.$$

In this case the register can be used in several different ways, and we will check whether or not the population estimates turn out to be consistent with the edit rule. In order to get an interesting case, we have chosen the rather unlikely situation of a sample featuring very many persons with a driving licence, i.e. relatively few persons younger than 18 years, and relatively few persons aged 18 years and older but without driving licence. Nevertheless it is a correct sample, as no one in the sample younger than 18 years owns a driving licence.

2.3.1 *Example 1*

Without using the register, the Horvitz-Thompson estimator with design weights $d = N / n = 4$ yields:

S	*# (without driving licence)*	*# (with driving licence)*	*total*
municipality M_1	20	60	80
municipality M_2	100	220	320
total	120	280	400

In this case municipality M_2 has 220 driving licences, but according to the register only 200 people aged 18 years and older live in municipality M_2. Also the total number of driving licences in the population is estimated 280, and according to the register the total number of people aged 18 years and older is 275. Also, the total number of inhabitants in both municipalities differs from the register values. So this HT-estimation is not consistent with the edit rule and is not numerically consistent with the register. Is the edit rule satisfied when using the register information, i.e. by weighting or by reweighting? We'll apply four different weighting schemes that include register information in the next four examples.

2.3.2 *Example 2*

Calibrating on the number of inhabitants in each municipality yields the following table. The bold numbers are marginal cell values that were used for calibration.

$S+R$	*# (without driving licence)*	*# (with driving licence)*	*total*
municipality M_1	27,500	82,500	**110**
municipality M_2	90,625	199,375	**290**
total	118,125	281,875	400

These estimates do not agree with the edit rule and the register, because in municipality M_1 the number of driving licences (82,5) exceeds the number of inhabitants aged 18 years and older (75). What's worse, if both this table and the register table [*municipality* ⊗ *age*] are used in the weighting scheme of the more detailed table [*municipality* ⊗ *age* ⊗ *driving licence*] × 1, a negative cell value occurs in category [*municipality* M_1 ⊗ *18*$^+$ ⊗ *without driving licence*]:

$S+R$	*without driving licence*		*with driving licence*		*total*
	18-	*18+*	*18-*	*18+*	
municipality M_1	35	-7,5	0	82,5	
total of mun. M_1	**27,500**		**82,500**		110
municipality M_2	90	0,625	0	199,375	
total of mun. M_2	**90,625**		**199,375**		290

So an estimated marginal table that violates an edit rule may cause negative cell values in the more detailed table. This is another reason to avoid estimating marginal tables that are incompatible with the edit rules.

In the next three examples we first estimate the more detailed frequency table [*municipality* ⊗ *age* ⊗ *driving licence*] × 1. After that *age* can simply be removed by aggregation, to arrive at [*municipality* ⊗ *driving licence*] × 1.

2.3.3 *Example 3*

If, as part of the register, the total number of driving licenses is known to be 240,625, it must be included in the weighting scheme. For instance, the weighting scheme [*municipality* + *age* + *driving licence*] yields the following table:

$S + R$	*# (18- without driving licence)*	*# (18- with driving licence)*	*# (18+ without driving licence)*	*# (18+ with driving licence)*	*total*
municipality M_1	24,001	0	10,169	75,830	**110**
municipality M_2	100,999	0	24,206	164,795	**290**
mun. $M_1 + M_2$	125	0	34,375	**240,625**	
total	**125**		**275**		400

In this case too the number of driving licences in municipality M_1 (75,830) exceeds the number of inhabitants aged 18 years and older (75), so the edit rule is violated. Also, the marginal table [*municipality* ⊗ *age*] differs from the register table [*municipality* ⊗ *age*]. Note that apart from the numerical inconsistencies with the register, this table itself is consistent with the edit rule, because $75,830 < 85,999$, $164,795 < 189,001$, $240,625 < 275$, and no persons younger than 18 years have a driving licence. Also no negative cell values occur.

2.3.4 *Example 4*

Poststratification with weighting scheme [*municipality* ⊗ *age*]:

$S + R$	*# (18- without driving licence)*	*# (18- with driving licence)*	*# (18+ without driving licence)*	*# (18+ with driving licence)*	*total*
municipalit M_1	35	0	8,824	66,176	
total of mun. M_1	**35**		**75**		110
municipality M_2	90	0	25,397	174,603	
total of mun. M_2	**90**		**200**		290
total of $M_1 + M_2$	125	0	34,221	240,779	400

This table is consistent with the edit rule, because $66,176 < 75$, $174,603 < 200$, and $240,779 < 275$. Also it is numerically consistent with the register table [*municipality* ⊗ *age*].

2.3.5 *Example 5*

Including also the total number of driving licences, the weighting scheme is [*municipality* ⊗ *age* + *driving licence*]. This yields the table:

$S+R$	*# (18- without driving licence)*	*# (18- with driving licence)*	*# (18+ without driving licence)*	*# (18+ with driving licence)*	*total*
municipality M_1	35	0	8,855	66,145	
total of mun. M_1	**35**		**75**		110
municipality M_2	90	0	25,520	174,480	
total of mun. M_2	**90**		**200**		290
total of $M_1 + M_2$	125	0	34,375	**240,625**	400

In this case too the estimated table is consistent the register and with the edit rule, because the number of inhabitants with a driving licence is smaller than the number of inhabitants aged 18 years and older, in both municipalities, as well as in the total population: $66{,}145 < 75$, $174{,}480 < 200$, and $240{,}625 < 275$.

In the last three examples, we first estimated the more detailed table including *age,* although the target table is [*municipality* ⊗ *driving licence*] × 1. The variable *age* is part of the edit rule, and it was included in the weighting schemes 3, 4, and 5. Then, according to the RW procedure at least the term [*municipality* ⊗ *age*] must be part of the weighting scheme, as in cases 4 and 5. According to the splitting-up procedure the term [*driving licence*] in the weighting scheme is correct, because it is a marginal term of [*municipality* ⊗ *driving licence*] × 1. But apart from the splitting-up procedure, the term [*driving licence*] is relevant only if it can be estimated from a data block larger than S, or if it is known from a register. Note that because the target table [*municipality* ⊗ *driving licence*] × 1 is a marginal table of [*municipality* ⊗ *age* ⊗ *driving licence*] × 1, according to the RW procedure - with or without splitting-up – the table including *age* must not be estimated before [*municipality* ⊗ *driving licence*] × 1. For that reason the estimation procedures 4 and 5 differ from the RW procedure: instead of estimating margins of [*municipality* ⊗ *driving licence*] × 1 first, a more detailed table was estimated first. Subsequently the target table can be found by aggregation. In examples 4 and 5, both variables being part of the edit rule - *age* and *driving licence* – are estimated simultaneously. In both cases this results in an estimated table in agreement with the edit rule, and with all cell values greater than or equal to zero. This contrasts with example 2, where subsequent simultaneous estimation of the same three variables in [*municipality* ⊗ *age* ⊗ *driving licence*] × 1 yields a negative cell value. Whether the weighting scheme of example 4 or 5 is used seems not to be crucial, as both yield estimates consistent with the edit rule and consistent with the register [*municipality* ⊗ *age*]. After these examples, the next section will elaborate on a more general case.

3. Linear edit rules and table estimates

For simplicity the examples in the previous section used only an edit rule with categorical variables, and only one simple frequency table was estimated. In the more general case categorical and numerical variables are part of the estimated tables, and both types of variables are part of the existing edit rules. Instead of according to the types of variables involved, an alternative and more profound way of classifying edit rules is the distinction between unconditional and conditional edit rules. In section 3.1 different types of linear edit rules are distinguished. In section 3.2 the implications of linear edit rules on the estimation of population totals are discussed. In section 3.3 several numerical examples are given that illustrate the methods described in section 3.2.

3.1 Types of linear edit rules

Linear edit rules can be divided into different types:

a. unconditional linear edit rules

e.g. $x_i = ay_i + b$, $x_i \geq ay_i + b$, $c_i \in [c_1, c_2, c_3]$,

with x_i and y_i numerical variables, and c_i a categorical variable of element i.

b. conditional edit rules of the *if then* type:

if (condition = true) *then* (result = true)

e.g. $\textit{if} \;\; a_i \in [a_1, a_2, a_3] \;\; \textit{then} \;\; c_i \in [c_1, c_2]$,

$\textit{if} \;\; a_i \in [a_1, a_2, a_3] \;\; \textit{then} \;\; z_i > 100$,

$\textit{if} \;\; z_i > 100 \;\; \textit{then} \;\; c_i \in [c_1, c_2]$

c. conditional edit rules of the *if and only if* type:

if and only if (condition = true) *then* (result = true)

e.g. $\textit{if and only if} \;\; x_i = 1 \;\; \textit{then} \;\; y_i = 2$,

$\textit{if and only if} \;\; c_i \in [c_1, c_2, c_3] \; \textit{then} \;\; a_i \in [a_1, a_2]$

Unconditional linear edit rules apply to all elements in a population, whereas conditional linear edit rules are confined to the subset of elements that satisfy a condition. Both numerical and categorical variables can be part of an edit rule. In case of a conditional edit rule, the variables can be part of the condition as well as the result, for instance

$$\textit{if} \;\; (a_i \in [a_1, a_2, a_3]) \vee (z_{1i} > z_{2i} + 10) \;\; \textit{then} \;\; (c_i \in [c_1, c_2]) \wedge (z_{3i} > 100)$$

with a_i and c_i categorical variables and z_{1i}, z_{2i} and z_{3i} numerical variables of element i. a_1, a_2, a_3 and c_1, c_2 are categories of a_i and c_i respectively.

3.2 Population totals estimated in accordance with linear edit rules

In this section the implications of linear edit rules on the estimation of population totals are discussed. An unconditional edit rule (type a) can be regarded as an *if then* edit rule with the condition equal to *true* for all elements. Also, an *if and only if* edit rule (type c) can be seen as a combination of two *if then* edit rules:

$$[\textit{if}\ \ (\text{condition} = \text{true})\quad \textit{then}\ \ (\text{result} = \text{true})]\ \textit{and}$$

$$[\textit{if}\ \ (\text{result} = \text{true})\qquad \textit{then}\ \ (\text{condition} = \text{true})].$$

Because both edit types a. and c. are special cases of the *if then* edit rule, we will first examine the *if then* edit rules (type b) in more detail in section 3.2.1. Then in section 3.2.2 the conditional *if and only if* edit rules (type c), and in section 3.2.3 the unconditional edit rules (type a) are discussed.

3.2.1 *If then* edit rules and table estimation

Let's define two dichotomous variables, x_i and y_i for each element i in the population as:

- $\textit{if}\ \ (\text{condition} = \text{true})\quad \textit{then}\ (x_i = 1)\ \textit{else}\ (x_i = 0)$
- $\textit{if}\ \ (\text{result} = \text{true})\qquad \textit{then}\ (y_i = 1)\ \textit{else}\ (y_i = 0)$

The following *if then* edit rule applies to all elements in the population:

$$\textit{if}\ \ (\text{condition} = \text{true})\ \textit{then}\ \ (\text{result} = \text{true})$$

It can be rewritten as

$$\textit{if}\ \ (x_i = 1)\ \textit{then}\ \ (y_i = 1)$$

Let the actual population totals of x_i and of y_i be X and Y respectively. Then

- X is the actual total number of elements in the population with $x_i = 1$
- Y is the actual total number of elements in the population with $y_i = 1$

With N the total number of elements in the population, the number of elements with $x_i = 0$ equals $N - X$, and the number of elements with $y_i = 0$ equals $N - Y$.

Because for each element in the population $x_i \leq y_i$, X en Y will satisfy the inequality $X \leq Y$. Therefore the estimates $\hat{X}$ and $\hat{Y}$ are consistent with the edit rule if and only if $\hat{X} \leq \hat{Y}$.

When $\hat{X}$ and $\hat{Y}$ are estimated independently from two different data sources, e.g. two different samples, or one variable counted from a register and the other estimated independently from a sample, the inequality is not necessarily true. Even if $\hat{X}$ and $\hat{Y}$ are estimated from the same sample but using two different weighting schemes, the inequality may not hold. This can happen easily when using the repeated weighting procedure and x_i and y_i belong to different tables. In this case the two tables will usually have different weighting schemes because of their different marginals. In order to satisfy $\hat{X} \leq \hat{Y}$ for random samples, $\hat{X}$ and $\hat{Y}$ should be estimated *simultaneously* from the *same* sample, because this implies the same weighting scheme for both variables. Let $\hat{X}$ and $\hat{Y}$ be estimated using the same weighting scheme. Then they are estimated with the same set of regression weights w_i:

$$\hat{X} = \sum_i (w_i x_i) \text{ and } \hat{Y} = \sum_i (w_i y_i)$$

Because of the condition $\hat{X} \leq \hat{Y}$, we have $\sum_i (w_i x_i) \leq \sum_i (w_i y_i)$. This implies that $\sum_i w_i \geq 0$ for all elements of the sample that belong to the category $(x_i = 0) \wedge (y_i = 1)$, because

$$\begin{aligned} \hat{Y} = \sum_i w_i y_i &= \sum_{\substack{i \in (x_i = 0,\\ y_i = 0)}} (w_i \times 0) + \sum_{\substack{i \in (x_i = 0,\\ y_i = 1)}} (w_i \times 1) + \sum_{\substack{i \in (x_i = 1,\\ y_i = 1)}} (w_i \times 1) \\ \geq \quad \hat{X} = \sum_i w_i x_i &= \sum_{\substack{i \in (x_i = 0,\\ y_i = 0)}} (w_i \times 0) + \sum_{\substack{i \in (x_i = 0,\\ y_i = 1)}} (w_i \times 0) + \sum_{\substack{i \in (x_i = 1,\\ y_i = 1)}} (w_i \times 1) \end{aligned}$$

implies

$$\sum_{\substack{i \in (x_i = 0,\\ y_i = 1)}} w_i + \sum_{\substack{i \in (x_i = 1,\\ y_i = 1)}} w_i \geq \sum_{\substack{i \in (x_i = 1,\\ y_i = 1)}} w_i \,,$$

so that

$$\sum_{\substack{i \in (x_i = 0,\\ y_i = 1)}} w_i \geq 0 \,. \; \blacksquare$$

This will certainly be the case for a random sample (x_i, y_i) if $w_i \geq 0$ for all i. In case of many *if then* edit rules, for each of them the elements of the sample belonging to a category with (condition = false) and (result = true) must have

$\sum_{\substack{i \in (x_i = 0, \\ y_i = 1)}} w_i \geq 0$. Also, all variables of each *if then* edit rule should be estimated simultaneously from the same data block. This general rule has an exception: if some variables that are part of edit rules are known from a register, they need not be estimated simultaneously with all other variables that are part of these edit rules. For each edit rule the register variables are calibrated on the register values. Even if variables from different edit rules are estimated apart from each other and using different weighting schemes, the common register values will be the same. The register variables of each edit rule are part of the weighting scheme, and therefore they do not depend on any sample, and will be consistent for all samples. Therefore the required relations between estimated population totals can be maintained.

We summarize some results:

1. If $\hat{X}$ and $\hat{Y}$ are estimated simultaneously, i.e. from one data block and using the same weighting scheme, and all regression weights w_i are greater than or equal to zero, then $\hat{X}$ and $\hat{Y}$ will be consistent with the *if then* edit rule.

2. If $\hat{X}$ and $\hat{Y}$ are estimated independently, i.e. from two different data blocks, or from one data block using two different weighting schemes, then $\hat{X}$ and $\hat{Y}$ need not be consistent with the *if then* edit rule.

3. If $\hat{X}$ en $\hat{Y}$ are estimated from one data block using the same weighting scheme, but with negative values of some resulting regression weights, the estimated $\hat{X}$ and $\hat{Y}$ may not be consistent with the *if then* edit rule.

4. Calibration on a poorly estimated marginal that does not obey an *if then* edit rule will cause some regression weights to be negative. That will result in negative cells in the estimated frequency table. For instance, if both X and Y are part of a marginal, and the edit rule is: *if* $(x_i = 1)$ *then* $(y_i = 1)$. If in the marginal $X > Y$, then in the estimated table the estimated total number of elements in the category $(x_i = 0) \wedge (y_i = 1)$ is negative. This implies that, depending on the size of the estimated table, one or more cells in the category $(x_i = 0) \wedge (y_i = 1)$ are negative, and some regression weights are negative too. Depending on the detail of the table, the category $(x_i = 0) \wedge (y_i = 1)$ may cover several marginal cells, but for this argument all marginal cells should be single valued for x_i and y_i.

5. In order to get estimates satisfying several linear (numerical and categorical) edit rules, all variables contained in the edit rules should be estimated simultaneously from the same data block. This implies that the data block should include the complete set of variables that are mutually linked in the edit rules. This does not apply to variables that are part of a register. It is a general

requirement, but neglecting it, a particular sample estimated with a particular weighting scheme could yield a table satisfying the edit rules by chance.

6. If regression weights are derived - using some appropriate weighting scheme - that have a positive total $\sum w_i$ of the sample elements in category $(x_i = 0) \wedge (y_i = 1)$, then the estimates are consistent with the *if then* edit rule and with the margins of the estimated table. This will certainly be the case if all regression weights are nonnegative. But a situation with some negative regression weights may also be correct.

7. Poststratification preserves the edit rules if the correction weights and the original sampling weights are positive. Then the resulting poststratification weights will also be positive.

8. In order to get estimates that apply to an *if then* edit rule, no extra terms need to be added to the weighting scheme. But if a table does not include all edit variables, it should first be extended with the missing edit variables. If marginals of the extended table are known from register or were estimated before, the extension of the estimated table will in turn necessitate the inclusion of additional terms in the weighting scheme, provided we want consistency with those marginals, and the resulting weighting scheme is small enough.

9. If a table does not include all edit variables, it should be extended with the missing edit variables before it is estimated. Categorical variables are usually part of a classification, so they exist at different levels of detail. For consistency with the edit rule it suffices to add a categorical variable at the level of the edit rule. A more detailed level is also allowed, but will yield a larger weighting scheme.

10. Estimates resulting from the design weights apply to *if then* edit rules if all design weights are positive, what is usually the case. However, numerical consistency with the register and with tables estimated from other data sources will be lacking.

11. Calibration on many marginals of the estimated table, e.g. the splitting up procedure, increases the risk of negative regression weights. This is also the case when all marginals used in the calibrations agree with all edit rules. Therefore a detailed table can best be estimated straightaway, calibrating on all its register margins. The disadvantage of estimating marginals first is that they will appear in the weighting scheme of the detailed table as compared to a weighting scheme consisting only of all register variables. Estimation of marginal tables can only be recommended if they use a much larger block of data. Then the smaller sample containing extra variables is corrected for sampling error.

12. Whether or not negative regression weights result in estimated population totals that contradict the edit rules, i.e. resulting in negative cells of a frequency table, depends on the actual sample and the estimated table. A detailed table estimated

from a small sample, that must be numerically consistent with many marginals, will probably contradict the edit rules. But a detailed table estimated using only one large data block with calibration on few marginals, will likely be in agreement with the edit rules.

When $\hat{X}$ and $\hat{Y}$ cannot be estimated from the same data block because no data block is available containing all variables contained in the edit rule, obviously no records exist that satisfy the edit rule. If $\hat{X}$ and $\hat{Y}$ are estimated from different data blocks, and they appear not to satisfy the edit, the design of the sample should be modified in a way that all variables of the edit rule are sampled simultaneously, i.e. for the same elements. Then $\hat{X}$ and $\hat{Y}$ can be estimated simultaneously from the same data block. An alternative approach can be carried out for *if and only if* edit rules and for numerical equalities, as will be shown in the next two sections.

3.2.2 *If and only if* edit rules and table estimation

The *if and only if* edit rules and unconditional edit rules can be examined like the previous section on *if then* edit rules, in order to get suitable regression weights. The *if and only if* edit rules can be viewed as a combination of two *if then* edit rules:

$$[\mathit{if}\ (x_i = 1)\ \mathit{then}\ (y_i = 1)]\ \wedge\ [\mathit{if}\ (y_i = 1)\ \mathit{then}\ (x_i = 1)]$$

The condition $\hat{X} = \hat{Y}$ yields $\sum_i (w_i x_i) = \sum_i (w_i y_i)$. Contrary to the *if then* edit rules there is no additional constraint to the regressions weights w_i now. Because the categories $(x_i = 0) \wedge (y_i = 1)$ and $(x_i = 1) \wedge (y_i = 0)$ have no elements in the sample, the estimated $\hat{X}$ and $\hat{Y}$ are

$$\hat{Y} = \sum_i w_i y_i = \sum_{\substack{i \in (x_i = 0, \\ y_i = 0)}} (w_i \times 0) + 0 + 0 + \sum_{\substack{i \in (x_i = 1, \\ y_i = 1)}} (w_i \times 1) = \sum_{\substack{i \in (x_i = 1, \\ y_i = 1)}} w_i$$

and $$\hat{X} = \sum_i w_i x_i = \sum_{\substack{i \in (x_i = 0, \\ y_i = 0)}} (w_i \times 0) + 0 + 0 + \sum_{\substack{i \in (x_i = 1, \\ y_i = 1)}} (w_i \times 1) = \sum_{\substack{i \in (x_i = 1, \\ y_i = 1)}} w_i$$

So, for all sets of regression weights w_i, the equality $\hat{X} = \sum_{\substack{i \in (x_i = 1, \\ y_i = 1)}} w_i = \hat{Y}$ holds, even if some weights w_i get negative because of a large weighting scheme. ■

The variables that are part of the *if and only if* edit rules should be estimated simultaneously, i.e. from the same data block using the same weighting scheme. Other data blocks that have one or more of the variables of the *if and only if* edit rule can be calibrated on the table that was estimated first containing all those variables. If no sample exists containing all variables of an *if and only if* edit rule, all but one of the variables can be estimated first. Then the population totals of the remaining variable can be calculated using the edit rule. Alternatively the supplying data blocks can be reweighted simultaneously with the edit rule as an additional constraint, as demonstrated in section 3.6.3.

3.2.3 Unconditional edit rules and table estimation

Regarding unconditional edit rules, two different types can be distinguished: equalities and inequalities, for instance $z_{1i} = az_{2i} + b$, or $z_{3i} > cz_{4i} + d$, with a, b, c and d constants, and z_{1i}, z_{2i}, z_{3i} and z_{4i} numerical variables.

Concerning the unconditional equalities: if the numerical variables are estimated simultaneously from the same data block, the equality of estimated population totals $\hat{Z}_1 = a\hat{Z}_2 + bN$ will hold for each set of regression weights. In this case a weighting scheme can be used for consistency with existent marginal tables if the regression weights result in an frequency table with no negative cells. With this condition marginal tables may be estimated first, also if they don't contain all variables that are part of the unconditional equality. This contrasts with the previous result for *if then* edit rules. If no sample exists that provides all variables of the equality, all but one of the population totals can be estimated. Then the population total of the remaining variable can be calculated from the edit rule. Alternatively two different data blocks can be reweighted simultaneously with the edit rule as an additional constraint.

Concerning the inequality, if $z_{3i} > cz_{4i} + d$ then the estimates of the population totals must satisfy $\hat{Z}_3 > c\hat{Z}_4 + Nd$. Therefore

$$\begin{aligned}\hat{Z}_3 - c\hat{Z}_4 - Nd &= \sum_i w_i\left(z_{3i} - cz_{4i} - d\right) \\ &= \sum_{i\in(w_i>0)} w_i\left(z_{3i} - cz_{4i} - d\right) + \sum_{i\in(w_i<0)} w_i\left(z_{3i} - cz_{4i} - d\right) > 0 .\end{aligned}$$

So the regression weights w_i must satisfy the following inequality:

$$\sum_{i\in(w_i>0)} w_i\left(z_{3i} - cz_{4i} - d\right) > \sum_{i\in(w_i<0)} - w_i\left(z_{3i} - cz_{4i} - d\right)$$

If due to the weighting scheme there are only few negative w_i, and the terms with negative w_i have relatively small values of $z_{3i} - cz_{4i}$, then the inequality will hold. This also depends on which elements are part of the sample. If all regression weights w_i are positive, and the variables are estimated simultaneously from the same data block i.e. with the same weighting scheme, the population estimates will satisfy the inequality. If no data source exists that contains all variables of the inequality, then whether or not the inequality $\hat{Z}_3 > c\hat{Z}_4 + Nd$ is satisfied depends on the elements in the data blocks and the weighting schemes used. If the inequality does not hold and data from several data blocks is used, then one variable that is part of the inequality can be calculated using the edit rule and using the other estimated variables. Preferably the variable with the worst quality should be calculated from the edit rule, after replacing the inequality by an equality. Alternatively the supplying data blocks can be reweighted simultaneously with the edit rule as an additional constraint. From a practical point of view this is the most cumbersome method. Examples of this, and of an alternative approach, will be given in the next two sections.

3.3 Numerical examples

In this section the methods described in the previous section are illustrated with numerical examples. The first example in section 3.3.1 is a conditional inequality with categorical variables. The examples in section 3.3.2 compare the results of three different methods of estimation of numerical variables with an unconditional equality.

3.3.1 Example of a conditional edit rule with categorical variables

As an example of a conditional edit rule, we continue the case presented in section 2.3, with two municipalities, and the edit rule

$$\textit{if} \;\; (\textit{having a driving licence}) \;\; \textit{then} \;\; (\textit{age} \geq 18)$$

The table to be estimated is [*municipality* ⊗ *driving licence*] × 1. The sample contains all variables that are part of the edit rule, i.e. *driving licence* and *age.* These two variables must be estimated simultaneously with the same weighting scheme, so the table is first extended with *age:*

$$[\textit{municipality} \otimes \textit{age} \otimes \textit{driving licence}] \times 1$$

In order to estimate consistently with the register, the weighting scheme must contain

$$[\textit{municipality} \otimes \textit{age}].$$

This is example 4 in section 2.3.4 mentioned before. The resulting table is numerically consistent with the register, and also in agreement with the edit rule.

In order to estimate also consistently with the total number of persons with a drivers licence - if that was estimated first or known from a register – the same table must be estimated using the weighting scheme

$$[\textit{municipality} \otimes \textit{age} + \textit{driving licence}].$$

This is example 5 in section 2.3.5 as mentioned before. It estimates the variables *age* and *driving licence* simultaneously, and the resulting table agrees to all conditions: i.e. consistency with the register and with the edit rule, and all cell values ≥ 0. For the sample used, both weighting schemes appear to be small enough to prevent negative cells.

All variables that are part of the same edit rule but not of the register (*driving licence*) must be estimated with the corresponding edit variables that are part of the register (*age*). Only edit variables known from a register (*age*) may be used separately for calibration while estimating tables. However, it can be favourable to estimate variables separately that are part of edit rules (*driving licence*) if a larger sample is available than is used for the detailed table (c.f. example 5 in section 2.3.5). However, such estimates should contain all edit variables that are part of the register (*age*) at a hierarchical level at least as detailed as in the edit rule. In this case *age* must have at least the categories <18 and ≥ 18, or ≤ 18 and >18, here the table to be estimated is [*age* ⊗ *driving licence*]. Afterwards such estimated table must be compatible with the edit rules. If the addition of such a marginal table to the weighing model of the more detailed table that contains all edit variables yields negative cells in the detailed table, the detailed table must be estimated without calibrating on the estimated marginal table (*driving licence*). Then the estimated margin must be ignored on behalf of an aggregation based on the estimated detailed table. Alternatively - if the detailed table still can't be estimated without negative cells - only the marginal should be estimated, omitting the detailed table. The inclusion of the term [*driving licence*] in the weighting scheme instead of [*age* ⊗ *driving licence*] is appropriate only if it is known from register. Inclusion of an estimated term [*driving licence*] should be avoided, even though it could be compatible when used. This applies even more to an estimated term [*municipality* ⊗ *driving licence*] instead of [*municipality* ⊗ *driving licence* ⊗ *age*], that is in fact the table itself without any estimated margin.

Tables involving variables that are part of edit rules but are also part of a register can be estimated without extending them with all other variables of the related edit rules. Because the actual values of the register variables are known, they may be part of an estimated table, as it is calibrated on these register values anyway. According to the RW procedure, if tables containing other variables - that are part of edit rules but not of the register - are estimated later on, they will also be calibrated on the variables that are part of the edit and of the register. This guarantees consistency with the edit rules concerning the register variables that are part of them.

3.3.2 Examples of an unconditional edit rule with numerical variables

A population consists of $N = 1000$ companies. In the register of each company its region – region 1 or region 2, and its size - small or large – are known. A sample S_1 consists of $n_1 = 200$ companies, of each company its *region*, *size* and *income*. S_2 is a sample of size $n_2 = 50$, featuring the variables *region*, *size, income, expenses* and *profit*. The sampling weights of S_1 are equal to 5, and of S_2 they are 20. S_1

and S_2 can either be two independent samples, or S_2 can be sampled as a subset of S_1 ($S_2 \subset S_1$). In the population the following numerical edit rule applies for each company:

$$income = expenses + profit.$$

From the register, the following register table is counted:

Register	*small*	*large*	*total*
region 1	88	227	315
region 2	183	502	685
total	271	729	1000

The 200 elements of sample S_1 aggregate to the following table:

S_1	*number of companies*		*income*	
	small	*large*	*small*	*large*
region 1	19	46	311,3574	3.081,4758
region 2	35	100	602,6832	6.500,8336
total	54	146	914,0406	9.582,3094

The 50 elements of S_2 aggregate to the following table:

S_2	*number of companies*		*income*		*profit*		*expenses*	
	small	*large*	*small*	*large*	*small*	*large*	*small*	*large*
region 1	8	11	123,8635	574,5667	63,8974	272,6188	59,9661	301,9479
region 2	7	24	270,1167	1.538,3375	123,4310	675,3744	146,6857	862,9631
total	15	35	393,9802	2.112,9042	187,3284	947,9932	206,6518	1.164,9110

In sample S_2 each element of the sample satisfies the edit rule, and consequently also the aggregated sample totals of *income*, *profit* and *expenses* of sample S_2 satisfy the edit rule for each category [*region* ⊗ *size*].

The tables to be estimated are [*region* ⊗ *size* ⊗ *income*], [*region* ⊗ *size* ⊗ *expenses*], and [*region* ⊗ *size* ⊗ *profit*], in short the table

$$region \otimes size \otimes (income + expenses + profit).$$

The estimated tables must satisfy the edit rule, and must be mutually consistent and consistent with the register [*region* ⊗ *size*]. Three different strategies are applied:

method 1: using only S_2 with poststratification

Estimating all three tables from S_2, with poststratification with respect to *region* ⊗ *size.* So the *income* data from S_1 is not used at all. All edit variables are estimated simultaneously, that is using the same data block S_2 and the same weighting scheme [*region* ⊗ *size*]. The total weights in each cell of the estimated table will be greater than or equal to zero because it is a simple poststratification instead of a regression. This will result in the required consistencies. However the small sample S_2 will yield estimates with a relatively large variance.

Using only S_2 and calibrating on [*region* ⊗ *size*] × 1 from the register yields:

Register	*number of companies*			*income*		
+ S_2	*small*	*large*	*total*	*small*	*large*	*total*
region 1	88	227	315	1.362,4979	11.856,9684	13.219,4663
region 2	183	502	685	7.061,6224	32.176,8921	39.238,5145
total	271	729	1000	8.424,1203	44.033,8605	52.457,9808

Register	*profit*			*expenses*		
+ S_2	*small*	*large*	*total*	*small*	*large*	*total*
region 1	702,8708	5.625,8609	6.328,7317	659,6271	6.231,1075	6.890,7346
region 2	3.226,8385	14.126,5816	17.353,4201	3.834,7839	18.050,3105	2.1885,0944
total	3.929,7093	19.752,4425	23.682,1518	4.494,4110	24.281,4180	28.775,8290

These estimates are consistent with the register values of the number of companies in each category [*region* ⊗ *size*]. They also meet the edit rule for each category [*region* ⊗ *size*], as can easily be verified. Consequently also the margins *region* and *size* meet the edit rule, as well as the estimated population totals of *income, profit,* and *expenses.*

method 2: calibration of S_2 using the register and an estimate from S_1

In this method the table [*region* ⊗ *size* ⊗ *income*] is estimated first from S_1 with poststratification with respect to [*region* ⊗ *size*]. Subsequently the other two tables are estimated simultaneously from S_2 with weighting scheme [*region* ⊗ *size*] × (1 + *income*). S_2 can either be a subset of S_1, or S_2 can be a sample independent of

S_1. Irrespective of the resulting regression weights in S_2, the estimated table will be in agreement with the edit rule because the edit is an unconditional equality. This method is preferable to method 1 if S_1 is larger or of better quality than S_2. This applies to both S_1 independent of S_2, and $S_2 \subset S_1$. This method may be used only if the resulting total weights in each cell are greater than or equal to zero.

Estimating [*region* ⊗ *size*] × *income* from S_1 with poststratification with regard to [*region* ⊗ *size*] from the register yields:

Register	*number of companies*			*income*		
+ S_1	*small*	*large*	*total*	*small*	*large*	*total*
region 1	88	227	315	1.442,0766	15.206,4135	16.648,4901
region 2	183	502	685	3.151,1725	32.634,1846	35.785,3571
total	271	729	1000	4.593,2491	47.840,5981	52.433,8472

Subsequent estimation of [*region* ⊗ *size*] × *profit* and [*region* ⊗ *size*] × *expenses* from S_2, with weighting scheme [*region* ⊗ *size*] × *income* + [*region* ⊗ *size*] × 1 using the regression estimator yields:

Register	*number of companies*			*income*		
$+ S_1 + S_2$	*small*	*large*	*total*	*small*	*large*	*total*
region 1	88	227	315	1.442,0766	15.206,4135	16.648,4901
region 2	183	502	685	3.151,1725	32.634,1846	35.785,3571
total	271	729	1000	4.593,2491	47.840,5981	52.433,8472

The total *number of companies* equals the register values for each category [*region* ⊗ *size*], and the *income* equals the table estimated from S_1. The same S_2 regression weights yield the following *profit* and *expense*:

Register	*profit*			*expenses*		
$+ S_1 + S_2$	*small*	*large*	*total*	*small*	*large*	*total*
region 1	739,9262	7.555,2815	8.295,2077	702,1504	7.651,1320	8.353,2824
region 2	1.423,7245	14.349,4267	15.773,1512	1.727,4480	18.284,7579	20.012,2059
total	2.163,6507	21.904,7082	24.068,3589	2.429,5984	25.935,8899	28.365,4883

These estimated values of *profit* and *expenses* are consistent with the values of *income* as estimated from sample S_1. Therefore the edit rule is observed for each estimated population total in the categories [*region* ⊗ *size*], and all its marginals. The advantage compared to method 1 is that the larger sample S_1 is used. The resulting weighting scheme was small enough to use it for consistency of the S_2 table with the *income*-marginal from S_1 and with the register. So the *income* margin could be estimated first, although it did not contain all edit variables. In the following method, the *income* is estimated using both samples S_1 and S_2, instead of fully adapting the *income* in sample S_2 to the *income* estimated from sample S_1. A combined approach for both samples is preferable if both samples are of comparable quality and size.

method 3: estimating S_1 and S_2 simultaneously

If S_1 and S_2 are independent and of comparable size and quality, they can both be weighted simultaneously in order to calibrate on the register [*region* ⊗ *size*], with the edit rule as an additional constraint. In this case a modified version of the regression estimator can be used to get the desired results. Compared to method 2 this method has the advantage of a smaller weighting scheme and of the variable *income* estimated from both samples instead of only sample S_1.

This method uses data from S_1 and S_2 simultaneously, with weighting scheme [*region* ⊗ *size*] and the numerical edit rule as an additional constraint. In this case the function f to be minimized is (see Van de Laar 2004):

$$f(w_{i,1}, w_{j,2}) = \sum_{i \in S_1} \left\{ (w_{i,1} - d_{i,1})^2 \sigma_{i,1}^2 / 2d_{i,1} \right\} + \sum_{j \in S_2} \left\{ (w_{j,2} - d_{j,2})^2 \sigma_{j,2}^2 / 2d_{j,2} \right\}$$

The $\sigma_{i,1}$ and $\sigma_{j,2}$ are the respective underlying model's variance parameters of S_1 and S_2. The n_1 final weights $w_{i,1}$ and the n_2 final weights $w_{j,2}$ must be so that the following 12 constraint equations $c_a = c_a\left(w_{i,1}, w_{j,2}\right) = 0$ are satisfied:

$$a = 1,..,4: \; c_a\left(w_{i,1}\right) \equiv X_a - \sum_{i \in S_1} w_{i,1} x_{ai,1} = 0,$$

$$a = 5,..,8: \; c_a\left(w_{j,2}\right) \equiv X_{[a-4]} - \sum_{j \in S_2} w_{j,2} x_{[a-4]j,2} = 0,$$

$$a = 9,..,12: c_a\left(w_{i,1}, w_{j,2}\right) \equiv \sum_{j \in S_2} \left(w_{j,2} x_{[a-8]j,2} \left[W_{j,2} + U_{j,2}\right]\right) - \sum_{i \in S_1} \left(w_{i,1} x_{[a-8]i,1} I_{i,1}\right) = 0.$$

The $X_1,..,X_4$ are the total numbers of companies known from the register in each of the 4 categories of [*region* ⊗ *size*]. The $x_{1i,1},..,x_{4i,1}$ and $x_{1j,2},..,x_{4j,2}$ are dichotomous variables indicating the category of [*region* ⊗ *size*] of sampling element $i \in S_1$ or $j \in S_2$ respectively. If we introduce the Lagrangian $F = F\left(\lambda_a, w_{i,1}, w_{j,2}\right)$ as

$$F = f(w_{i,1}, w_{j,2}) - \sum\nolimits_{a=1}^{12} \left[\lambda_a c_a \left(w_{i,1}, w_{j,2}\right)\right]$$

then at the constrained extremum of f the derivatives of F should be zero:

$$\partial F / \partial \lambda_a = \partial F / \partial w_{i,1} = \partial F / \partial w_{j,2} = 0$$

where the λ_a are called Lagrange multipliers.

For this constrained minimization problem, equating the partial derivatives of F to zero, and taking $\sigma_{i,1} = \sigma_1$ and $\sigma_{j,2} = \sigma_2$, yields (after multiplying by $x_{ni,1}$ or $x_{nj,2}$ and summarizing over the sample) the following 12 equations for the Lagrange multipliers λ_a with $a = 1,..,12$:

$$n = 1,..,4: \qquad \lambda_n \hat{X}_{nHT,1} + \lambda_{[n+8]} \hat{I}_{nHT,1} = \sigma_1^2 \left(\hat{X}_{nHT,1} - X_n\right)$$

$$n = 1,..,4: \qquad \lambda_{[n+4]} \hat{X}_{nHT,2} - \lambda_{[n+8]} \left(\hat{W}_{nHT,2} + \hat{U}_{nHT,2}\right) = \sigma_2^2 \left(\hat{X}_{nHT,2} - X_n\right)$$

$$n = 1,..,4: \qquad \hat{I}_{nHT,1} - \hat{W}_{nHT,2} - \hat{U}_{nHT,2} =$$

$$= \lambda_n \left(\frac{\hat{I}_{nHT,1}}{\sigma_1^2}\right) - \lambda_{[n+4]} \left(\frac{\hat{W}_{nHT,2} + \hat{U}_{nHT,2}}{\sigma_2^2}\right) +$$

$$+ \lambda_{[n+8]} \left\{ \sum_{i \in S_1} \left[\frac{d_{i,1} x_{ni,1} \left(I_{i,1}\right)^2}{\sigma_1^2} \right] + \sum_{j \in S_2} \left[\frac{d_{j,2} x_{nj,2} \left(W_{j,2} + U_{j,2}\right)^2}{\sigma_2^2} \right] \right\}$$

In order to calculate the twelve λ_a a 12-dimensional matrix is inverted[1]. This can easily be done using S-Plus. For this example, we take for both data blocks $\sigma_1 = \sigma_2 = 1$. In S-Plus, the following tables result:

The *income* from S_1 and S_2 for each category $n = 1,..,4$ equals:

$$\hat{I}_n = \left(1 - \frac{\lambda_n}{\sigma_1^2}\right) \hat{I}_{nHT,1} - \frac{\lambda_{[n+8]}}{\sigma_1^2} \sum_{i \in S_1} \left[d_{i,1} x_{ni,1} \left(I_{i,1}\right)^2 \right]$$

The resulting table of the *number of companies* and *income* is:

[1] If efficiency is essential these equations can also be solved by application of the Newton-Raphson method. This can be accomplished without explicit inversions of the matrices involved. The procedure is iterated until the equations are satisfied to a given precision.

Register	*number of companies*			*income*		
$+S_1+S_2$	*small*	*large*	*total*	*small*	*large*	*total*
region 1	88	227	315	1.406,9542	13.182,0511	14.589,0053
region 2	183	502	685	4.258,9358	32.416,7265	36.675,6623
total	271	729	1000	5.665,8900	45.598,7776	51.264,6676

The *profit* and *expenses* from S_1 and S_2 for each category $n = 1,..,4$ are:

$$\hat{W}_n = \left(1-\frac{\lambda_{[n+4]}}{\sigma_2^2}\right)\hat{W}_{nHT,2} + \frac{\lambda_{[n+8]}}{\sigma_2^2}\sum_{j\in S_2}\left[d_{j,2}x_{nj,2}W_{j,2}\left(W_{j,2}+U_{j,2}\right)\right]$$

and

$$\hat{U}_n = \left(1-\frac{\lambda_{[n+4]}}{\sigma_2^2}\right)\hat{U}_{nHT,2} + \frac{\lambda_{[n+8]}}{\sigma_2^2}\sum_{j\in S_2}\left[d_{j,2}x_{nj,2}U_{j,2}\left(W_{j,2}+U_{j,2}\right)\right]$$

with resulting table:

Register	*profit*			*expenses*		
$+S_1+S_2$	*small*	*large*	*total*	*small*	*large*	*total*
region 1	723,5716	6.389,1640	7.112,7356	683,3826	6.792,8871	7.476,2697
region 2	1.934,5157	14.243,4563	16.177,9720	2.324,4201	18.173,2702	20.497,6903
total	2.658,0873	20.632,6203	23.290,7076	3.007,8027	24.966,1573	27.973,9600

In each category [*region* ⊗ *size*] the numerical edit rule $\hat{I}_n = \hat{U}_n + \hat{W}_n$ is satisfied. When $\sigma_1^2 << \sigma_2^2$ the resulting *profit* and *expenses* are close to the S_2 estimates with poststratification with respect to [*region* ⊗ *size*]. With $\sigma_1^2 >> \sigma_2^2$ the resulting *income* is close to the S_1 income with poststratification with respect to [*region* ⊗ *size*].

Comparing the three resulting tables from method 1, 2, and 3, we observe considerable differences in the estimated values, indicating that the mean square errors of the estimates of the three methods are rather large in this example, with sample sizes 50 and 200 and the assumed population variance. For instance, the approximated values of *profit* of the three methods in the category [*small* in *region 2*] are 3227, 1424, and 1935 respectively, and in category [*large* in *region 1*] 5626, 7555, and 6389 respectively. In both categories method 1 yields a deviating value, using only the small sample S_2.

4. Edit rules, sampling, and the order of table estimation

The mere availability of register data and several sampling data blocks, and the existence of edit rules in the population, complicates the process of estimation because of the desired consistencies. Aiming at numerically consistent tables that agree with the edit rules, the following step-by-step plan is proposed:

1. list all variables needed for publication, both categorical and numerical.
2. examine which edit rules exist in the sampled population for these categorical and numerical variables.
3. determine which variables are known or can be derived from a register (RV's), and which variables must be estimated from sample data (SV's).
4. for all edit rules collectively, divide the SV's in mutually disjoint groups (EG's) so that no edit rules exist containing SV's of two different groups. This also divides the edit rules in corresponding groups (EG's).
5. for each EG a sample survey may be constituted that contains all SV's of the EG. Then for each EG all SV's are sampled simultaneously and for the same elements, see section 3.2.1. Usually they can be combined with all RV's of the EG by linking each sample element to the corresponding register data. If for some practical reason this linking is impossible, all RV's of the EG must also be sampled for all sample elements.
    a. If step 5 results in several EG's with a small number of variables, then EG's may be combined in order to reduce the number of samples. The SV's that are not part of an edit rule may be added to one or several of the EG's. No edit rules should exist that contain SV's of different EG's. Shared RV's in different EG's are allowed, see section 3.2.1.
    b. If there are additional sample data blocks available, e.g. from external sources, that contain a subset of the SV's and all RV's of one or more EG's, these sample data blocks may be used as additional data as indicated in step 6b below. Then the EG samples in step 5 must be large enough to calibrate them on the additional estimated marginal term derived from the 'SV-subset' sample data block. The 'SV-subset' sample blocks must be larger than the samples of step 5 in order to improve estimates. If corresponding to a 'SV-subset' sample block no sample containing all EG variables is available, it may still be useful in step 8 below.
6. The samples of step 5 must be large enough so that at least a weighting scheme equal to the RV's of the EG yields a nonnegative sum of regression weights for each cell in the estimated frequency table defined by the product of all categorical variables of the EG (RV's and SV's).
    a. If this is impossible, the sample is too small compared to the available RV's of the EG. Either the sample must be enlarged, or it should not be used at all. In the latter case no separate term with all categorical

variables of the EG will be part of the weighting scheme in step 8 below.

b. If all cells are nonnegative, an extra categorical RV or an additional marginal term from step 5b may be added to the weighting scheme. If all cells of the resulting table are nonnegative, the larger weighting scheme can be used for more detailed or better estimates. An advantage of adding the extra RV without estimating the original table first, is that the resulting more detailed table can be estimated without an estimated marginal term in it, i.e. with a smaller weighting scheme.

A special case of step 6a is the so called empty cell problem. Then the table to be estimated, as defined in step 6, has one or more nonzero marginal cells in its weighting scheme without any sample elements contributing to these nonzero marginal cells.

7. In this way for each EG or combination of EG's with sample data, one table is estimated in a numerically consistent way. Each cell of all frequency tables must be nonnegative, that is for each cell $\sum w_i \geq 0$. Margins of these tables that were not part of the weighting scheme, are calculated by aggregation.

8. Detailed frequency tables containing categorical RV's and categorical SV's from different EG's can be estimated if a sample with all these variables is available, and if the weighting scheme including a term with all categorical RV's and a separate term for each available estimated margin, is small enough. Then the sample is large enough to avoid the empty cell problem, and to get nonnegative cells. If a detailed table to be published also contains a categorical RV that is not part of any edit rule, such a RV must also be added to the estimated table and its weighting scheme, in order to obtain consistency with the register. This yields a larger weighting scheme, and in order to avoid the empty cell problem, the sample must be large enough. The resulting weights must also be used to estimate the corresponding tables of the numerical SV's in the EG's. If the weighting scheme is too large, some estimated margins may be omitted to reduce the weighting scheme. In that case there will be no numerical consistency with the removed estimated marginal terms. Either the detailed table, or the removed marginal table should not be published. If the mere register term makes a too large weighting scheme, the detailed table cannot be estimated from the sample in a numerically consistent way, i.e. the sample is too small or the table too large.

9. Tables containing RV's that are not part of the weighting scheme and EG's cannot be estimated consistently from the same set of regressions weights. If an additional RV cannot be added to the weighting scheme of the sample in step 8, such a more detailed table may be estimated using the RW method if it has only few categorical SV's. For an actual random sample this can be verified by calculation of the regression weights resulting from the RW weighting scheme of the table. Again, for each cell of the table the sum of the regression weights must be nonnegative. Note that in step 8 an extra RV works as a condition on

the sample size, whereas in step 9 (and also 6b) a certain sample is used to estimate a more detailed table.

The more information is available from the register, and the more edit rules exist that contain register variables, the larger the samples must be for reweighting them according to the product of all register variables of an EG. Numerical consistency with many register variables, and estimates in agreement with many edit rules, make higher demands on the samples: on their size, as well as on their number of variables.

5. Results and discussion

In a fictitious situation without edit rules, the RW procedure can be used to estimate numerically consistent tables from different data blocks. The number of estimated tables is limited only due to the increasing number of terms in the weighting schemes necessary for numerical consistency with all known or estimated marginals. This happens if no regressions weights can be calculated because of a too large weighting scheme. Another problem mainly of the estimation of detailed tables is the so-called empty cell problem. This occurs when a sample does not include some categories with a known or estimated marginal value unequal to zero. The samples must be large enough to prevent these problems, even if the estimated table does include register variables with many categories. If the sample is too small, or the table to be estimated is too detailed, it can not be estimated in consistence with its margins. If numerical consistency includes consistency with estimated marginals, and these yield a too large weighting scheme, one can try to abandon the estimated margin, and using the weighting scheme with only the product of all register variables that are part of the estimated table. Then any marginal table can be calculated by aggregation, resulting in numerically consistent tables.

In the more common situation with many edit rules existing in the population, an important additional requirement is that the estimated tables must be consistent with them. Table estimates that violate edit rules can often easily be revealed as being erroneous by an informed user of the tables, because edit rules are often common knowledge. In order to get estimates consistent with the edit rules, any variable that is part of an edit rule should be estimated simultaneously with at least all register variables that are part of the EG. Also, a data block must be available containing all variables of an EG. The estimated table containing all categorical variables of the EG, and also all desirable estimated marginal tables, must have in each cell a nonnegative total regression weight $\sum w_i$. If this is not the case, a larger sample must be used, or some of the estimated marginals must be left out from the weighting scheme. If the weighting scheme containing only the product of all register variables of the EG yields negative total weights in some cells, the table is too detailed. It can not be estimated consistently without enlarging the sample size.

The existence of edit rules, as well as the extensive use of register information, impose additional requirements on the samples. Also, the edit rules often require the inclusion of register variables in estimated tables. Instead of estimating many relatively small marginal tables first, the edit rules require first estimating the more detailed tables including all edit variables. In order to enable that, the number of estimated margins should be limited, particularly if the weighting scheme is already large because of the edit variables that are part of the register. The existence of edit rules and the availability of register information should be examined both during the design of the surveys, and while deciding on the number and order of tables to be estimated. Considering the edit rules is inevitable in order to get the most out of the available data.

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. Statistics Netherlands, Voorburg, The 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, 1999.

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., A.H. Kroese en A.J. Willeboordse (2001), Aligning estimates by repeated weighting, CBS, Heerlen.

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.

Renssen, R.H. Kroese, A.H., and Willeboordse, A. (2001). Aligning Estimated by Repeated Weighting. Statistics Netherlands, Heerlen, The Netherlands.

Särndal, C.-E., B. Swensson en J. Wretman (1992), Model Asisted Survey Sampling, Springer Verlag, New York.

Van de Laar, R.W.A. (2001), Weegproces SSB, volume 3: Inlezen bestanden, CBS, Voorburg.

Van de Laar, R.W.A. (2004), Simultaan schatten van editvariabelen bij herhaald wegen, CBS, Voorburg.

Weegproces SSB, volume 1: Handleiding voor het schatten van tabellen met herhaald wegen (2001), CBS, Voorburg.

Willeboordse, A. (2000). Towards a new Statistics Netherlands. Blueprint for a process oriented organisation structure. Statistics Netherlands, Voorburg, The 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, November 1996, Statistics Netherlands, Voorburg, The Netherlands.

Willeboordse, A. and Ypma, W. (1998). Meta Tools in support of A Corporate Dissemination Strategy. Research Paper 9839. Statistics Netherlands, Voorburg, The Netherlands.