Report on WP2
Methodological developments

ISTAT, CBS, GUS, INE, SSB, SFSO, EUROSTAT

ESSnet on Data Integration

Contents

Preface

Jeroen Pannekoek and Arnout van Delden[1]

Data integration is an opportunity for NSIs. The availability of many more data sources than the traditional survey can be put to use by data integration, opening up possibilities for reducing costs and response burden while maintaining or increasing the statistical output, its quality and timeliness. From the description of the current state of the art in data integration methodology in the Report on WP1, it is apparent that the methodology for different aspects of data integration has been developing rapidly in the recent years. However from this Report on WP1 it also becomes clear that there remain some methodological issues to be tackled in order to better profit from the availability of multiple sources. In WP2 of the ESSnet on Data Integration a number of these methodological issues are further investigated. The progress that has been made and the solutions that have been obtained are reported in this Report on WP2.

The methodological developments described in the chapters of this volume can be classified according to three areas of integration activities connected with different phases in the statistical processing of data from multiple sources. In particular, we consider the following areas of integration tasks. *Record linkage*, which is an obvious first stage when combining data from different sources on the same units. When no unique error-free identifiers are available, the problem of linkage errors has to be dealt with at this stage. *Inference with multiple sources*, where estimation problems are studied when multiple sources are available but the sets of units in the different sources are not or only partially overlapping. *Micro and macro consistency*, where the problem

[1]Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: jpnk@cbs.nl.

is tackled that information from different sources can be inconsistent both at the micro-level, that is within linked records, as well as at the macro-level, that is between estimates of aggregates. Below, an overview is given of the seven chapters in this volume, organized by the three perspectives described above.

Record linkage

Chapter 1. Methodological developments on quality measures: estimation of probabilities of correct linkage.

Unless unique error free unit-identifiers are available, record linkage is a more or less error prone process and this can be expressed by the concept of the probability of a correct linkage. This probability can be seen as a measure of the quality of the linkage process. This chapter investigates methods to estimate this probability of a successful linkage.

Chapter 2. Editing errors in the relations between units when linking economic data sets to a population frame.

Linkage problems in economic statistics can often be traced back to the problem that there is no single, univocal, unit type. The base unit of companies are the legal units according to the chambers of commerce. Both administrative units, as well as the statistical units in the general business register, may be combinations of legal units. This is especially the case for larger companies. However, for a specific company, the two composite types may consist of different combinations of legal units. To combine information from different sources, the relations between these unit types must be established which can be a source of linkage errors. This chapter contributes methodology to detect and correct errors in this linkage process.

Inference with multiple and incomplete sources

Chapter 3. Methodological developments on the use of samples drawn according to complex survey designs.

The subject of this chapter is statistical matching. Statistical matching techniques combine information available in data sources with distinct units (the sets of units do not overlap) referring to the same target population. The data sources have some variables in common but other variables are measured in only one of the sources and the sources are in this sense „incomplete". The challenge in statistical matching is to make inference on the relation between the variables that are never observed together and, in particular, to measure the uncertainty about such inference. This chapter further elaborates on the

approaches to this problem discussed in chapter 2 of the Report on WP1, thereby paying especially attention to the application of these methods in the context of complex survey designs. An empirical evaluation of the efficiency of the methods using the results of a simulation study is also provided.

Chapter 4. Handling incompleteness after linkage to a population frame: incoherence in unit types, variables and periods.

The composite units as discussed in chapter 2 can lead to difficulties in the linkage process but they also lead to special forms of incompleteness in linked data sets. In particular, if for some but not all administrative units belonging to the same statistical unit the target variable is not (yet) available and the information on the statistical unit is partially missing. Also, variable definitions and measurement periods in administrative sources may differ from the target ones, which may lead to additional forms of incompleteness. In this chapter an overview is given of different patterns of incompleteness when tax-units are linked to statistical units and different imputation methods are proposed to solve these problems.

Chapter 5. Bootstrapping combined estimators based on register and survey data.

In this chapter we consider the problem of combining, in the estimation process, register and survey data with the following properties, often encountered in practice: the register only covers a selective part of the target population and the definition of the target variable as measured in the register differs from the definition aimed in the design of the survey. The problem considered in this chapter is how to combine, in the estimation process, the information from both sources. Three estimators are proposed for this and similar situations and ways to asses the variance of these estimators, using analytic formulae as well as via resampling (Bootstrap method) are provided. A simulation study using real data is performed to evaluate the performance of the different methods.

Micro and macro consistency

Chapter 6. Models and algorithms for micro-integration.

Especially in business statistics, there are many logical relations between the variables (like profit = turnover - costs) also known as edit-rules). When information on units comes from linking different sources, these logical relations may not hold and a micro-integration step is necessary to integrate the different pieces of information, the data sources and the edit-rules, to arrive

at consistent integrated micro-data. In this chapter a variety of methods are proposed that can solve this integration problem based on a minimum adjustment principle: the data will be modified (adjusted) as little as possible but such that all edit-rules are satisfied.

Chapter 7. Applications of macro-integration.

In order to obtain consistency between estimates at a macro-level, methods based on calibration can be used (e.g. repeated weighting, see Report on WP1, ch. 5). An alternative is, however, to apply methods based on a minimal adjustment principle directly to the estimates to solve inconsistency problems. This can have advantages because adjustment methods are more flexible in incorporating constraints and no access to the original micro-records is necessary. In this chapter the relation between calibration by adjusting weights and by adjusting of estimates directly is investigated and the applicability of macro-adjustment as an alternative to (repeated) weighting is investigated.

The methods in this volume can also be related to the three topics of data integration distinguished throughout this ESSnet: record linkage, statistical matching and micro integration. Chapters 1 and 2 are on record linkage and chapter 3 is on statistical matching. Chapters 4 and 5 have in common with the problem of statistical matching that the units in the sources cannot be linked to form a single enriched data file, because the sources do not contain the same units. In contrary to the statistical matching problem, in chapters 4 and 5 matching of some of the units is possible because the units in the different sources do overlap to some extend. Chapter 6 clearly treats a topic of micro-integration, in chapter 5 of the Report on WP1 this particular topic is discussed as a problem of correction for measurement errors for which the solutions shown here are much more general than those in WP1. The adjustment of estimates at a macro-level treated in chapter 7 is mathematically similar to the problem of adjusting micro-data discussed in chapter 6 but it is not a micro-integration method since it uses estimates at an aggregated level only. However the macro-adjustment methods of chapter 7 solve the same problem (i.e. consistency between estimates at an aggregate level) that was treated in the Report on WP1 (chapter 5) by micro-integrations methods (i.e. consistent repeated weighting).

Special thanks are due to Marcin Szymkowiak for the efforts in transforming all the files in this document in LaTeX.

Chapter 1

Application of Bayesian record linkage in a real life context: a comparison with traditional record linkage approaches and comments on the analysis of linked data

Brunero Liseo[a], **Mauro Scanu**[b], **Andrea Tancredi**[a], **Tiziana Tuoto**[b], **Luca Valentino**[b]

[a] *La Sapienza Universita di Roma, Italy*
[b] *Istituto nazionale di statistica – Istat, Italy*

1.1 Problem description: Survey on live births

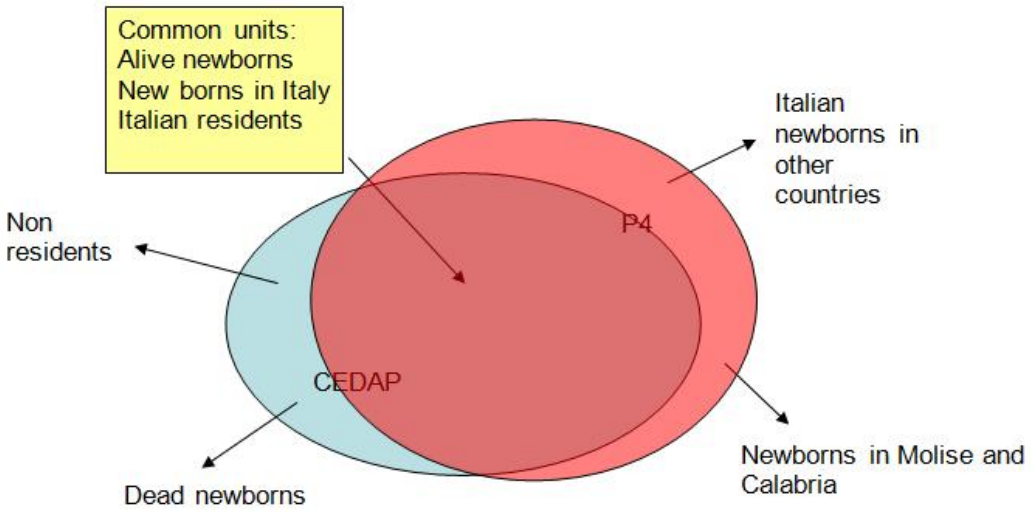
A relevant problem in demographic statistics is to link demographic and health data sets relative to births. The purpose of this linkage activity is to obtain an integrated data set with more information (in one data set there are the characteristics of the newborns as weight, type of birth, how many brothers he/she has, week of birth, while the other data set contains data on the characteristics of the household as marital status, nationality, education of the parents). The two data sets are named:

1. Resident administrative register (survey on new born inclusion, henceforth P4)

2. Health register (survey on the assistance certificates in the childbirth moment, henceforth CEDAP)

They are not managed directly by Istat: P4 is managed by municipal register offices while CEDAP by the Health Minister. Both data sets must be sent monthly to Istat.

The universes of interest for the two sources are not exactly the same. Indeed there is a large intersection consisting of alive newborns born and resident in Italy, but CEDAP considers also dead newborns and births of non residents. Moreover the P4 source considers also Italian newborns in other countries. Finally, the regions Molise and Calabria have not yet provided their CEDAP. The following figure shows the intersection between the two sources and the characteristics of the common units.



The two data sets consist of approximately 50 thousand records per month. The record linkage experiment was performed only for data relative to the month of March 2005. Exclusion of the non eligible newborns (i.e. those not belonging to the intersection of CEDAP and P4) leads to the following file sizes:

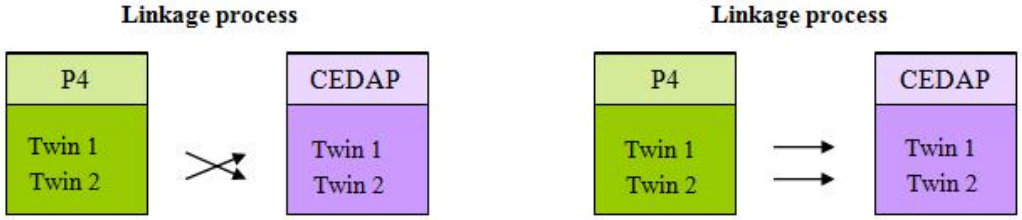
P4	**March 2005**	**43 336 units**
CEDAP	**March 2005**	**41 381 units**

These data sets are subject to confidentiality constraints that prevent a linkage with the help of personal identifiers (as the fiscal code) or identifying

Entities involved in births	Variables available in P4 and CEDAP
Mother	Mother's birth date (day, month and year)
Mother	Mother's marital status
Mother	Mother's citizenship
Mother	Place of residence of the mother (hence)
Newborn	Birthplace
Newborn	Newborn's birth date (day, month and year)
Newborn	Newborn gender
Father	Father's birth date (day, month and year)
Father	Father's citizenship

variables (as names and surnames). The variables that can be used for linkage purposes are in the table on next page.

These variables are usually more than enough for linking the records from the two data sets, with the exception of same sex twins. Indeed in such cases the available variables coincide and we have no evidence to distinguish one twin from the other.



These cases are classified as a common link.

1.2 Analysis of available data and selection of the matching variables

The analysis of the metadata suggests not using the variables citizenship (of mother and father) and marital status as matching variables. The problem is that these variables have a very poor identification power. Missing data affect the variables connected to the newborn's father (approximately 10% missing data rate). Finally, the variables residence of the mother and birthplace are highly correlated, and only birthplace is retained.

1.3 Objective

The idea is to apply the Liseo and Tancredi model, and to study the result comparing them with the results that can be obtained from traditional approaches. This work is organized in the following paragraphs:

1. apply the approach by Liseo and Tancredi
2. develop a practical guide for its application in NSIs,
3. check possible alternatives: e.g. those based on clerical review,
4. draw final considerations comparing the Liseo and Tancredi approach with other approaches.

1.4 Application of the Tancredi and Liseo approach and comparison with the probabilistic record linkage method based on conditional independence

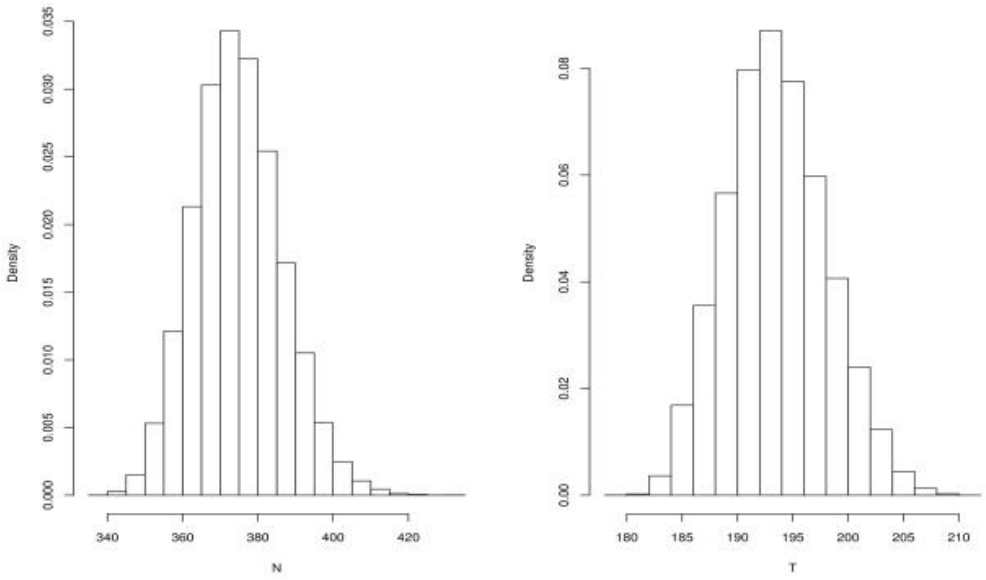
We illustrate the results of the Tancredi and Liseo model on a single block consisting of newborns with a mother of Chinese nationality. The two files consist of n_A=287 records from the P4 register and n_B=253 records from the CEDAP data set.

Selection of the matching variables - As matching variables we considered day, month and year of the mother's birth date. The total number of entries in Cartesian product of all the categories of the matching variables (V in the instructions of the application of this approach, see the Appendix A) is k=12x31x25=9300.

Setting the hyperparameters - The hyperparameter g appearing in the prior distribution $p(N)$ has been set equal to 2 in order to have a proper prior. Moreover the probability vectors θ of frequencies distributions of the matching variables are assumed independent Dirichlet random variables. This is equivalent to assume, at the super-population level, that day, month and year of the mother's birth date are mutually independent. We also assume that all the Dirichlet distributions are uniform on their support.

Simulation of the posterior distribution - To simulate the posterior distribution of the model parameters we have used the algorithm described in Tancredi and Liseo (2011). For this application a single Markov chain of length

100000 has been used. The following figure shows the posterior distributions for the population size N and for the number of matches T.



Results - The posterior means for the population N and T are respectively 375 and 194 while the 95% posterior credibility intervals are [354; 400] for N and [186; 203] for T. Regarding the probabilities of a measurement error $1-\beta$ we have the following posterior means: 0.007 for the year, 0.014 for the month and 0.014 for the day. Hence, the mother's birth year seems to be the matching variable with the smallest measurement error. However note that the hypothesis of uniform distribution for the matching variables, which is assumed by the model in case of measurement error, could not be justified for the month and day of the mother's birth date. In fact for these variables, if the true values are not reported, there would be more chance to observe values like 1 or 15 for the day and 1 for the month respect to other values.

Comparison with the approach defined in Jaro (1989) - Finally note that, applying a classical probabilistic record linkage approach and using the Jaro model for the comparison vectors with the same matching variables used for the Bayesian model, we have obtained 215 matches. Employing this number of matches to estimate the population size N would lead to a posterior estimate equal to 339 and a 95% posterior interval equal to [331, 348]. Hence, the Bayesian and the classical approaches based on conditional independence in this case seem to produce quite different results.

Data on computational aspects of the algorithm - Note that, from a computational point of view, Bayesian inference for the Tancredi and Liseo approach may require very intensive simulation methods also when, as in this case, we have to consider only a block of moderate size. In fact at each iteration of the simulation algorithm we need to update the true values μ for each sample unit, the population frequencies F and the super-population parameter θ, and the simulation of all these random variables may present formidable problems when the dimension of the multidimensional contingency table V becomes large. In multiple blocks applications, computing problems are also harder but parallel computations for separated blocks may reduce the computing time in a significant way.

1.5 Application of the probabilistic record linkage approach based on conditional independence and comparison with the deterministic approach used in Istat

The classical approach based on conditional independence has been also performed with the probabilistic model implemented in the software Relais.

Selection of the blocking variables – In order to reduce the search space reduction, Newborn gender and Newborn's birth date are used as blocking variables. CEDAP and P4 contain data from 31 birth dates then the number of expected blocks is 62, with a homogeneous distribution of the number of pairs in the blocks. The frequency of maximum expected link pairs in each block is approximately 1/750.

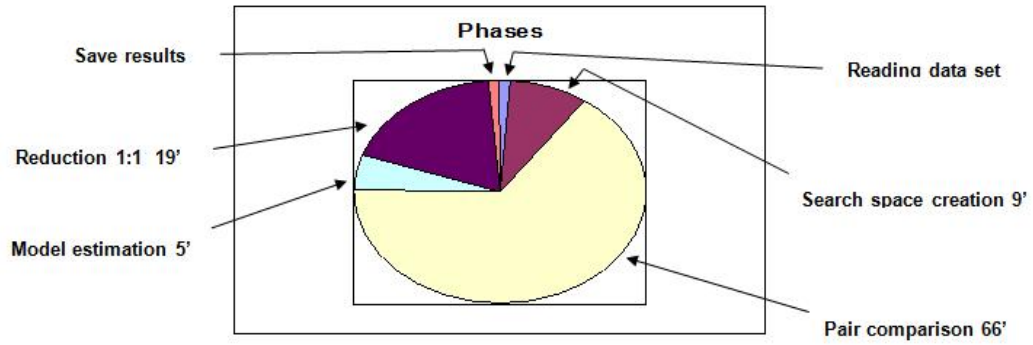
Selection of the matching variables – All the available variables but the blocking ones and those excluded for problems have been used as matching variables:

1. day of mother's birth date
2. month of mother's birth date
3. year of mother's birth date
4. birthplace
5. father's birth date (day, month and year)

Comparison metric – The simplest metric (that detects equality or inequality of the states of a matching variable in each pair of units in a block) is used for all matching variables with the exception of father's birth date. In this last case the Jaro distance with 0.9 as threshold is used (in other words two dates are considered equivalent even if they differ by one character)

Identification of the matched pairs – In order to declare a pair as a match, a single threshold is used (posterior probability equal to 0.9). In order to obtain consistent results, the one-to-one constraint algorithm that forces each unit to be linked with at most one unit has been implemented. The final set consists of 36'562 matches.

The whole process takes 1 hour 40 minutes (AMD Sempron 3400+ 2GHz, 1G RAM). The next figure represents how much time was devoted to each record linkage phase.



Application of a deterministic record linkage approach - The deterministic approach declares a pair as a match when some rules are fulfilled. A rule is defined on a combination of the key variables able to identify a unit: a pair is declared as a match when all the key variables in a rule agree. Different rules can be adopted. In this case, it is useful to associate a weight to each rule, the weight should reflect the importance of the key variables used in the rule for the unit identification.

The set of pairs is reduced by means of these two steps.

1. The first step checks if a pair is declared as a match for more than one rules. In this case, only the rule with the highest weight is retained.

2. The second step is a 1:1 constraint, that each unit can be linked to no more than one other unit. Also in this case, only the pair with the highest weight is retained.

Example:

Set:

P4	CEDAP
A1	A2
B1	B2
C1	

Comparisions:

Rule:	Pairs
Rule1 (weight3)	A1-A2
Rule2 (weight 2)	A1-A2
	B1-B2
Rule3 (weight 1)	C1-B2

Solution:

Pair	Weight
A1-A2	3
A1-A2	2
B1-B2	2
C1- B2	1

Step 1:

Pair	Weight
A1-A2	3
B1-B2	2
C1-B2	1

Step 2:

Pair	Weight
A1-A2	3
B1-B2	2

The rules used in this example are:

1. Agreement on all the key variables, with weight 2.
2. Agreement on 14 of the 15 common variables, with weight 1 (birthdates are split in three distinct variables: day, month and year)

Following these rules, the number of declared matches is 32595. *Comparison between probabilistic and deterministic record linkage* – The probabilistic and deterministic approaches determine two data sets of declared matches with the following characteristics:

- 31 675 pairs are classified as matches by both methods
- 256 pairs are not identical but can be considered equivalent because they consist of same-sex twins.

Naming „expert's rule" this deterministic procedure, the result of comparison between these approaches gives:

– 87% of matches according to the probabilistic record linkage approach are matches also for the expert's rule (31'931 in 36'562)

– - 98% of matches according to the expert's rule are matches also for the probabilistic record linkage approach (31'931 in 32'595)

An assessment of the quality of the linkage procedures can be performed through an evaluation of samples of pairs to be carefully evaluated (by clerical review). The clerical review consists in the analysis of the careful analysis of all the common variables observed in the two records. If for all the variables the differences are minimal the pair is classified as a true link.
Among the declared matches for the record linkage procedure, we distinguish:

- the common matches (A),
- matches consisting of twins (B),
- pairs declared as matches only by the expert's rule (C),
- pairs declared as matches only for the probabilistic record linkage approach (this last set is split in the one consisting only of pairs similar on at least half of the variables not used in the linkage procedure - D -, and the other pairs - E).

Class of pairs	**Total number of pairs**	**Sample size**	**True link**	**False link**
A	31675	45	45	0
B	256	39	39	0
C	664	41	41	0
D	4338	134	132	2
E	293	58	53	5

In search of possible undetected links, attention was given to two pair sets:

– Pairs not matched according to the expert's rule, available in the constrained 1:1 solution obtained through Relais but with a posterior probability value below the threshold (F)

– Pairs not matched according to the expert's rule and not , available in the constrained 1:1 solution obtained through Relais that coincide in at least one of the most significant variables (G)

The results obtained on the checked samples give the following false match and false non match rates:

Class of pairs	Total number of pairs	Sample size	True link	False link
F	847	56	41	15
G	65778	440	2	438

Expert's rule:

False match rate	0%
False non match rate	14,35%

Probabilistic record linkage:

False match rate	0,25%
False non match rate	4,16%

1.6 Bayesian uses of linked data

In general, from a statistical methodology perspective, the merge of two (or more) data files can be important for two different reasons:

- *per sé*, i.e. to obtain a larger and integrated reference data set.
- to perform a subsequent statistical analysis based on the additional information which cannot be extracted from either of the two single data files.

The first situation needs not any further comment: a new data set is created and appropriate statistical analyses will be performed based on it. On the other hand, the second situation is more interesting both from a practical and a theoretical perspectives. Let us consider a toy example to fix the ideas.

Suppose we have two computer files, say A and B, whose records respectively relate to units (e.g. individuals, firms, etc.) of partially overlapping populations PA and PB. The two files consist of several fields, or variables, either quantitative or qualitative. For example, in a file of individuals, fields can be "surname", "age", "sex", etc. The goal of a record linkage procedure is to detect all the pairs of units $(a; b)$, $a \in A$ and $b \in B$, such that a and b refer actually to the same unit.

Suppose that the observed variables in A are denoted by $(Z, X_1, \ldots, X_k)$ and the observed variables in B are $(W, X_1, \ldots, X_k)$. Then one might be interested in performing a linear regression analysis (or any other more complex association model) between Z and W, restricted to those pairs of records

which we declare as matches. The intrinsic difficulties which are present in such a simple problem are well documented and discussed in Scheuren and Winkler (1993) and Lahiri and Larsen (2005).

In the statistical practice it is quite common that the linker (the researcher who matches the two files) and the analyst (the statistician doing the subsequent analysis) are two different persons working separately. However, we agree with Scheuren and Winkler (1993), which say " *... it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly.*"

In a more general framework, suppose one has $(Z_1,\ldots,\ Z_h,\ X_1,\ldots,\ X_k)$ observed on n_A units in file A and $(W_1,\ldots,\ W_p,\ X_1,\ldots,\ X_k)$ observed on n_B units in file B. Our general goal can be stated as follows:

- use the key variables $(X_1,\ldots,\ X_k)$ to infer about the true matches between A and B.
- perform a statistical analysis based on variables Z's and W's restricted to those records which have been declared matches.

To perform this double task, we present here a fully Bayesian analysis. The main point to stress is that in our approach all the uncertainty about the matching process is automatically accounted in the subsequent inferential steps. This approach uses, generalizes and improves the Bayesian model for record linkage described in Fortini et al. (2001). We present the general theory and illustrate its performance via simple examples. In Section 1.6.1 we briefly recall the Bayesian approach to record linkage proposed by Fortini et al. (2001) to which to refer for details. Section 1.6.2 generalizes the method to include the inferential part. Section 1.6.3 concentrates on the special case of regression analysis, the only situation which has been already considered in literature: see Scheuren and Winkler (1993) and Lahiri and Larsen (2005).

1.6.1 Bayesian Record Linkage

In Fortini et al. (2001) a general technique to perform a record linkage analysis is performed. Starting from a set of key variables $(X_1,\ldots,\ X_k)$, observed in two different sets of units, the method defines, as the main parameter of interest a matching matrix C, of size n_A times n_B, whose generic element is

$$c_{a,b} = \begin{cases} 1 & \text{units a and b refer to the same units} \\ 0 & \text{otherwise} \end{cases} \tag{1.1}$$

The statistical model is based on a multinomial likelihood function where all the comparisons between key variables among units are measured on a 0/1 scale. As in the mixture model proposed by Jaro (1995) a central role is played by the parameter vectors m and u, both of length 2^k, with

$$m_i = P\left(Y_{a,b} = y_i \,|c_{a,b} = 1\right), u_i = P\left(Y_{a,b} = y_i \,|c_{a,b} = 0\right), \tag{1.2}$$

for $i=1, \ldots, 2^k$, and $Y_{a,b}$ represents the 2^k-dimensional vector of comparisons between units $a \in A$ and $b \in B$. For a different approach, based on the actual observed values see Tancredi and Liseo (2011).
Then a MCMC approach is taken which produce a sample from the posterior distribution of the matrix valued parameter C. See Fortini et al. (2001) for a discussion about the appropriate choices for the prior distribution on C and on the other parameters of the model, mainly m and u.

1.6.2 A general method for inference with linked data

In this section we illustrate how to construct and calibrate a statistical model based on a data set which is the output of a record linkage procedure. As we already stressed, the final output provided by the procedure described in the previous section will be a simulation from the joint posterior distribution of the parameters, say $(C, m, u; \xi)$, where ξ includes all the other parameters in the model.

This can be used according to two different strategies. In fact we can either

I. Compute a "point" estimate of the matrix C and then use this estimate to establish which pairs are passed to the second stage of the statistical analysis. In this case, the second step is performed with a fixed number of units (the declared matches). It must be noticed that, given the particular structure of the parameter matrix C, no obvious point estimates are available. The posterior mean of C is in fact useless since we need to estimate each single entry $c_{a,b}$ either with 0 or 1. The posterior median is difficult to define as well, and the most natural candidate, the maximum a posteriori (MAP) estimate typically suffers from sensitivity (to the prior and to Monte Carlo variability) problems: this last issue is particularly crucial in official statistics. For a deep discussion on these issues see Tancredi et al. (2005) and, for related problems, in a different scenario, Green and Mardia (2006).

II. Alternatively, one can transfer the "global" uncertainty relative to C (and to the other parameters), expressed by their joint posterior distribution, to the second step statistical analysis.

We believe that this latter approach is more sensible in the way it deals with uncertainty. Among other things, it avoids to over-estimate the precision measures attached to the output of the second step analysis.

The most obvious way to implement the approach II simply consists in performing the second step analysis at the same time as the record linkage analysis, that is, including the second step analysis into the MCMC procedure. This will cause a feed-back propagation of the information between the record linkage parameters and the more specific quantities object of interest. Here we illustrate these ideas in a very general setting; in the next section we will consider the regression example in details.

Let $D = (y, z, w) = (y_{11}, \ldots, y_{n_A n_B},\ z_1, ..., z_{n_A}, w_1, ..., w_{n_B})$ be the entire set of available data where, as in the Introduction, y_{ab} represents the vector of comparisons among variables which are present in both files, z_a is the value of covariate Z observed on individual $a \in A$ and w_b is the value of covariate W observed on individual $b \in B$. The statistical model can then be written as

$$p\,(y, z, w\,|C, m, u, \theta\,) \tag{1.3}$$

where *(C; m, u)* are the record linkage parameters and θ is the parameter vector related to the joint distribution of $(W;Z)$. The above formula can always be re-expressed as

$$p\,(y\,|C, m, u, \theta\,)\,p\,(z, w\,|C, y, m, u, \theta\,) \tag{1.4}$$

It sounds reasonable to assume that, given C, the comparison vector Y does not depend on θ; also, given C, the distribution of *(W;Z)* should not depend both on the comparison vector data Y and the parameters related to those comparisons. It follows that model can be simplified into the following general expression:

$$p\,(y\,|C, m, u\,)\,p\,(z, w\,|C, \theta\,) \tag{1.5}$$

The first term in the last expression is related to the record linkage step; the last term refers to the second step analysis and must be specified according to the particular statistical analysis. The presence of C in both the terms allows the feed-back phenomenon we mentioned before. Approaches I and II can be re-phrased using the last formula. In the case I the first factor of the model is used to get an estimate $\hat{C}$ of C. Then $\hat{C}$ is plugged into the second factor and a standard statistical analysis is performed to estimate θ.

In approach II the two factors are considered together within the MCMC algorithm thus providing a sample from the joint posterior distribution of

all the parameters. There is actually a third possible approach to consider and we call it approach III. In fact, one can perform a MCMC algorithm with the first factor only and, at each step t=1, ..., I, of the algorithm III. perform the statistical analysis expressed by the second factor of the model using the information contained in $C^{(t)}$, the actual value of the Markov chain for the parameter C at time t. This way, one can obtain an estimate $\hat{\theta}_t$ of θ at each step of the MCMC algorithm and then somehow summarize the set of estimates. In the next section we will illustrate the three approaches in the familiar setting of the simple linear regression.

We anticipate that approach I seems to miss to account for the uncertainty in the first step of the process and it tends to produce a false impression of accuracy in the second step inferences.

We consider approach II as the most appropriate in terms of the use of statistical information provided by the data. However, approach III can be particularly useful especially if the set of linked data must be used more than one time, for different purposes. In fact, while in approach II information flows back and forth from C to θ, in the case of III the information goes on one sense only, from C to θ.

1.6.3 An example: simple linear regression

Consider again the toy example in the Introduction and assume that our object of interest is the linear relation between W and Z, say

$$W_j = \theta_0 + \theta_1 Z_j + \varepsilon_j; \quad j \in M \tag{1.6}$$

Here we describe how to implement the three different approaches discussed in previous section. We assume that our statistical model can be simplified according to 1.6.2.

Method I

1. Use any Record Linkage procedure to decide which pairs of records are true matches,
2. Use the subset of matched pairs to perform a linear regression analysis and provide an estimate of $\theta = (\theta_0, \theta_1)$ via ordinary least squares, maximum likelihood or Bayesian method.

Method II

1. Set a MCMC algorithm relative to model 1.5, that is, at each iteration t=1, ..., T,
2. draw $C^{(t)}$from the full conditional distribution,
3. draw $m^{(t)}, u^{(t)}, \xi^{(t)}$from the full conditional distribution,
4. draw $\theta^{(t)}$ from the full conditional distribution.

From steps II-b and II-c one can notice that the marginal posterior distribution of C will be potentially influenced by the information on θ.

Method III

1. Set up a MCMC algorithm restricted to the first factor of (1.5) to provide a posterior sample form the joint posterior distribution of (C,m,u) To do that one can follows, for example, the methods illustrated in Fortini et al. (2001) and Tancredi and Liseo (2011).
2. At each iteration *t=1, ..., T* of the MCMC algorithm, use $C^{(t)}$ to perform a linear regression analysis restricted to those pairs of records (a,b) such that $c_{a,b}^{(t)} = 1$, and produce a point estimate of (θ_0, θ_1) say $\left(\hat{\theta}_0, \hat{\theta}_1\right)$)
3. Use the list of estimates $\hat{\theta}^{(t)}$ as an "approximation" of the posterior distribution of θ

In this third approach the estimation of C is not influenced by the regression part of the model and it might be safer to use it for a future and different statistical analysis on the same merged data set.

Chapter 2

Editing errors in the relations between units when linking economic data sets to a population frame

Arnout van Delden, Jeffrey Hoogland[1]

Centraal Bureau voor de Statistiek

Summary: In this chapter we concentrate on methodological developments to improve the accuracy of a data set after linking economic survey and register data to a population frame. Because in economic data different unit types are used, errors may occur in relations between data unit types and statistical unit types. A population frame contains all units and their relations for a specific period. There may also be errors in the linkage of data sets to the population frame. When variables are added to a statistical unit by linking it to a data source, the effect of an incorrect linkage or relation is that the additional variables are combined with the wrong statistical unit. In the present paper we formulate a strategy for detecting and correcting errors in the linkage and relations between units of integrated data. For a Dutch case study the detection and correction of potential errors is illustrated.

Keywords: Linkage error, relations between unit types, detection and correction of errors.

[1]Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail:adln@cbs.nl. Remark: The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

2.1 Background

2.1.1 Introduction

There is a variety of economic data available that is either collected by statistical or by public agencies. Combining those data at micro level is attractive, as it offers the possibility to look at relations / correlations between variables and to publish outcomes of variables classified according to small strata. National statistical institutes (NSI's) are interested to increase the use of administrative data and to reduce the use of survey data because population parameters can be estimated from nearly integral data and because primary data collection is expensive.

The economic data sources collected by different agencies are usually based on different unit types. These different unit types complicate the combination of sources to produce economic statistics. Two papers, the current paper and Van Delden and Van Bemmel (2011) deal with methodology that is related to those different unit types. Both papers deal with a Dutch case study in which we estimate quarterly and yearly turnover, where we use VAT data for the less complicated companies[2] and survey data for the more complicated ones.

Handling different unit types starts with the construction of a general business register (GBR) that contains an enumeration of the different unit types and their relations. From this GBR the population of statistical units that is active during a certain period is derived, the population frame. This population frame also contains the relations of the statistical units with other unit types, such as legal units. In the current paper we formulate a strategy for detecting and correcting errors in the linkage and relations between units of integrated data.

In the Dutch case study, after linkage, we handle differences in definitions of variables and completion of the data. After both steps, population parameters are computed. Both steps are treated by Van Delden and Van Bemmel (2011) and resemble micro integration steps as described by Bakker (2011). After the computation of population parameters, an additional step of detecting and correcting errors is done as treated in the current paper.

In a next step, the yearly turnover data are combined at micro level (enterprise) with numerous survey variables collected for Structural Business

[2]In the current paper 'company' is used as a general term rather than as a specific unit type.

Statistics. The paper by Pannekoek (2011) describes algorithms to achieve numerical consistency at micro level between some core variables collected by register data and variables collected by survey data. Examples of such core variables in economic statistics are turnover, and wages. There are also other European countries that estimate such a core variable, e.g. turnover, from a combination of survey and administrative data. Total turnover and wage sums are central to estimation of the gross domestic product, from the production and the income side respectively.

Because the current paper and Van Delden and Van Bemmel (2011) share the same background, the current section 2.1.1 and the sections 2.1.2 and 2.2 are nearly the same in both papers.

2.1.2 Problem of unit types in economic statistics

The different unit types in different economic data sources complicate their linkage and subsequent micro integration. When a company starts, it registers at the chamber of commerce (COC). This results in a so called 'legal unit'. The government raises different types of taxes (value added tax, corporate tax, income tax) from these "companies". Depending on the tax legislation of the country, the corresponding tax units may be composed of one or more legal units of the COC, and they may also differ for each type of tax. Finally, Eurostat (EC, 1993) has defined different statistical unit types (local kind of activity unit, enterprise, enterprise group) which are composed of one or more legal units.

In the end, for each country, the set of unit types of companies may be somewhat different. But generally speaking, for each country, the legal units are the *base* units whereas tax and statistical units are *composite* units (see Figure 2.1). In some countries, like France, there is one-to-one relationship between legal units and tax units and tax units are one-to-one related to statistical units. In other countries, like the Netherlands, units that declare tax may be groupings of legal units that belong to different enterprises (Vaasen and Beuken, 2009). Likewise, in Germany, tax units may declare turnover for a set of enterprises (Wagner, 2004). As a consequence, at least in the Netherlands and Germany, for the more complex companies tax units may be related to more than one enterprise. In other words, the tax and statistical units are both composed of legal units, but their composition may be different.

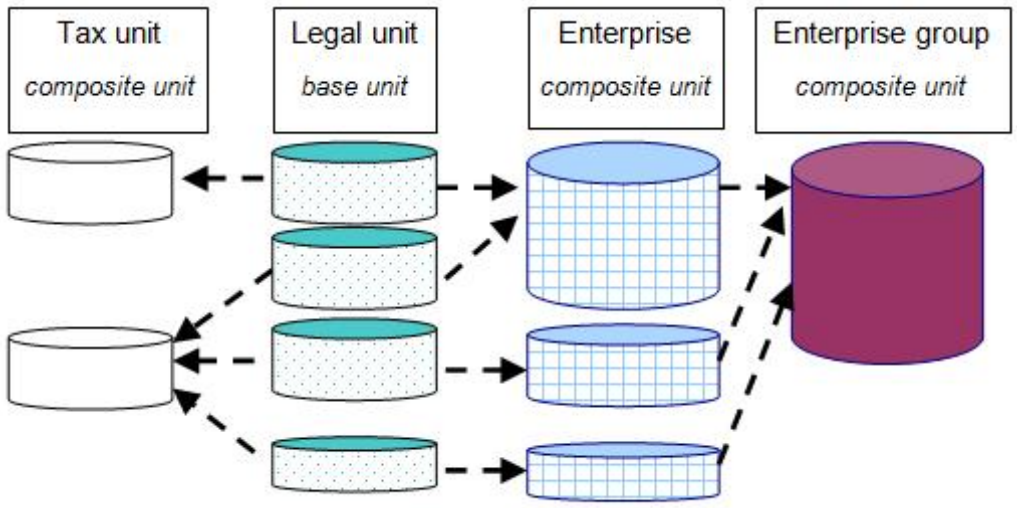


Figure 2.1. Different unit types in economic statistics. Each cylinder represents a single unit; arrows indicate the groupings of units.

2.1.3 General Business Register

NSI's have a GBR that contains an enumeration of statistical units and the underlying legal units. The GBR contains the starting and ending dates of the statistical units, their size class (SC code) and their economic activity (NACE code). In 2008, Eurostat has renewed its regulation on a business register (EC, 2008) in order to harmonise outcomes over different European countries. NSI's also use a GBR to harmonise outcomes over different economic statistics within an NSI. In addition, the Netherlands - and other NSI's, also added the relations between legal units and tax units to the GBR, to be able to use tax office data for statistical purposes.

2.1.4 Problem description

Within the GBR errors may occur in the relations between the unit types, which is explained in more detail in section 2.3. An example of a relation error is a tax unit in the GBR which is related to the wrong statistical unit. This statistical unit belongs to a certain NACE stratum. The consequence of this wrong relation may be that the tax unit 'belongs' to the wrong NACE stratum.

Also, linkage errors may occur between units in the data sets and the corresponding unit in the population frame, for example due to time delays. An example of a linkage error is a tax unit in the VAT declaration file that is

wrongly not linked to the corresponding tax unit in the population frame because a new identification number is used in the VAT declaration file where the population frame still contains the old identification number.

The focus of the current paper is to describe a strategy for detecting and correcting errors in the linkage and relations between units of integrated data. For a Dutch case study the detection of potential errors is illustrated.

2.1.5 Outline of the paper

The remainder of the paper is organised as follows. Section 2.2 describes a Dutch case study. Section 2.3 gives a classification of the errors that are considered in the current paper. In section 2.4 we describe the strategy of detecting and correcting the errors. Section 2.5 gives an example of a preliminary test on the effectiveness of a score function that we use. Finally, section 2.6 concludes and gives topics for future research.

2.2 Description of the case study

2.2.1 Background: statistical output

In the current paper we deal with the estimation of Dutch quarterly and yearly turnover levels and growth rates, based on VAT declarations and survey data. The work is part of a project called "Direct estimation of Totals". Turnover is estimated for the target population which consists of the statistical unit type the *enterprise*. Turnover output is stratified by NACE code size class. An overview of all processing steps from input to output data can be found in Van Delden (2010).

The estimated quarterly figures are directly used for short term statistics (STS). Also, the quarterly and yearly turnover levels and growth rates are input to the supply and use tables of the National Accounts, where macro integration is used to obtain consistent estimates with other parameters. Also, results are used as input for other statistics like the production index (micro data) and the consumption index (the estimates). Finally, yearly turnover is integrated at micro level with survey data of the Structural Business Statistics (SBS). Next, the combined data is used to detect and correct errors in both the turnover data as well as in the other SBS variables. Yearly turnover results per stratum are used as a weighting variable for SBS data.

In fact we deal with four coherent turnover estimates:

- net total turnover: total invoice concerning market sales of goods and services supplied to third parties excluding VAT
- gross total turnover: total invoice concerning market sales of goods and services supplied to third parties including VAT
- net domestic turnover: net turnover for the domestic market, according to the first destination of the product
- net non-domestic turnover: net turnover for the non-domestic market, according to the first destination of the product

More information on the turnover definition can be found in EC (2006). In the remainder of the paper we limit ourselves to net total turnover further referred to as turnover.

Table 2.1. Overview of the releases of the case study

Release	Period of estimation	Moment	Explanation
Flash estimate	Quarter	30-35 days after end of target period	Provisional estimate delivered for Quarterly Accounts, STS branches with early estimates
Regular estimate	Quarter	60-70 days after end of target period	Revised provisional estimate for Quarterly Accounts and for STS
Final STS estimate	Year and corresponding 4 quarters	April $y+1$, one year after target year	The estimates of the four quarters are consistent with the yearly figure
Final SBS estimate	Year and corresponding 4 quarters	April $y+2$, two years after target year	The estimates of the four quarters are consistent with the yearly figure. The yearly figure is based on STS and SBS turnover data

The quarterly and yearly figures are published in different releases, as shown in Table 2.1. The quarterly releases vary from a very early estimate delivered at 30-35 days after the end of the corresponding quarter to a final estimate for SBS publication delivered April year y+2 where y stands for the year in which the target period falls.

2.2.2 Target population and population frame

The statistical target population of a period consists of all enterprises that are active during a *period*. This true population is unknown. We represent

this population by a frame which is derived from the GBR. Errors in this representation are referred to as frame errors. Each enterprise has an actual and a coordinated value for the SC and NACE code. The coordinated value is updated only once a year, at the first of January and is used to obtain consistent figures across economic statistics. In the remainder of the paper we always refer to the coordinated values of SC and NACE code unless stated otherwise.

The population frame is derived as follows. First, each month, we make a view of the GBR that represents the population of enterprises that are active at the first day of the month; in short: the population state. This population state also contains the legal units, tax units and the 'enterprise groups'-units that are related to the enterprise population at the first day of the month. Next, the population frame for a period is given by the union of the relevant population states. For example, the frame for the first quarter of a year consists of the union of the population states on 1 January, 1 February, 1 March and 1 April.

For the case study, the frame contains four unit types: the legal unit (base unit), the enterprise (composite unit) and two tax units namely the base tax unit and the VAT unit. In the Netherlands each legal unit (that has to pay tax) corresponds one-to-one to a base tax unit. For the VAT, base tax units may be grouped into a VAT unit (composite unit). So this is an extension of the more general situation of Figure 2.1.

The units and their relations are shown in Figure 2.2. We consider:

1. the relation between the legal unit and the enterprise
2. the relation between the base tax unit and the legal unit
3. the relations between the VAT unit and the base tax unit

During the production of the GBR relation 1 is automatically derived from ownership relations in COC and tax office data, using business rules. Relation 2 is based on matching of name, postal code and house number, which Statistics Netherlands (SN) obtains from a National Basic Business Register (BBR). Relation 3 is automatically derived from tax office data using business rules.

The linkage between data sets and population frame is split into:

4. the linkage of a VAT unit in the VAT declaration to the (identical) VAT unit in the population frame

5. the linkage of an enterprise of the survey to an enterprise in the population frame

VAT declared by VAT units are linked to the corresponding VAT units in the frame, using exact matching of identification numbers (relation 4). Likewise, survey data as obtained for enterprises are linked to enterprises in the frame using exact matching of identification numbers (relation 5).

As explained in Vaasen and Beuken (2009), in the case of smaller companies each VAT unit is related to one enterprise and each enterprise may consist of one or more VAT units. For the more complicated companies, referred to as topX units, a VAT unit may be related to more than one enterprise.

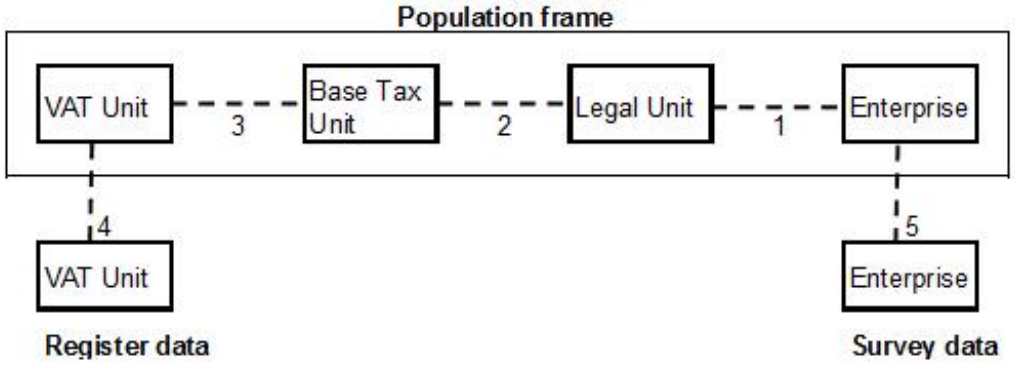


Figure 2.2. Relations between units types in the population frame, and the unit types of data sets that are linked to this frame.

2.2.3 Data

In the case study we use two types of source data. We use VAT data for non-topX enterprises. For topX enterprises we use survey data because VAT units may be related to more than one enterprise. This approach is quite common, also at other NSI's in Europe (e.g. Fisher and Oertel, 2009; Koskinen, 2007; Norberg, 2005; Orjala, 2008; Seljak, 2007). For non-topX units, we only use observations of VAT units that are related to the target population of enterprises.

Concerning VAT, a unit declares the value of sales of goods and services, divided into different sales types. The different sales types are added up to the total sales value, which we refer to as turnover according to the VAT declaration.

In the current paper we use VAT data and survey data for Q4 2009 - Q4 2010 and focus on quarterly estimates. Data are stratified according to NACE 2008 classification.

2.3 Classification of errors in relations and in linkage between units

2.3.1 Overview

In the case study, errors in the relations/linkages as shown in Figure 2.2 seems most likely in relation 1 and 3. Errors in relation 1 and 3 are due to errors in the data sources used in the production of the GBR and due to time delays. For 'relation' 4 and 5 errors may occur in the exceptional case of a mistake in the identification number. For more complex companies with more than one legal unit at the same address and postal code, errors may occur in relation 2.

In the next two sections we give a classification of errors that are considered in the present paper. We distinguish between (a) errors in the relations between different unit types within the population frame (section 2.3.2) and (b) errors in linkages between observations and the population frame (section 2.3.3). Note that the errors in the relations between unit types have been called *frame errors* by Bakker (2011) and *alignment* and *unit errors* by (Zhang, 2011).

2.3.2 Errors in the relations between unit types

Below we classify the different error types in order to understand the causes of errors and to assist the correction process. At SN relations between units are established in the GBR, from this GBR the population frame for a specific period is derived, see section 2.2.2. Any corrections in these relations are also made in the GBR and thereafter, new release of the population frame will be made.

Therefore, for the classification, we look into errors in the GBR and the consequences that they have for the population frame. We divide errors in the relations between unit types (see Figure 2.2) into four subtypes: namely errors in relations that are present versus relations that are wrongly absent and errors in relations that result in coverage errors and those that do not.

Erroneous positive relations between unit types

(a) **Error leading to over coverage**. The presence of a relation between unit types in the GBR where at least one unit is non domestic, resulting in over coverage in the population frame.

For example, two VAT units are linked to an enterprise of the target population. One of the two VAT units is domestic, the other is non-domestic, see Table 2.2. According to the Dutch tax rules the non-domestic unit has to declare tax in the Netherlands and is found within the Dutch tax data.

Note that for error type (a) the relation itself may be correct, but because we wish to make statistics by country the non-domestic units should not be included.

(b) **Erroneous relation.** An incorrect relation between unit types in the GBR where all units are domestic.

For example a domestic VAT unit is related to the wrong enterprise. This wrong enterprise may belong to another economic activity than the correct enterprise.

Errors concerning missing relations between unit types

(c) **Error leading to under coverage.** A relation between unit types that is wrongly missing in the GBR, resulting in a domestic unit that is not covered by the population frame.

For example, an enterprise consists of two legal units, but only one of them is found in the GBR. Another example is a domestic legal unit that is present in the GBR, but is incorrectly not yet related to an enterprise.

(d) **Erroneous missing relation.** An incorrect missing relation between unit types, where all the corresponding units are present in the GBR but just the relation itself is wrongly missing.

For example within the GBR, VAT unit A is related to enterprise A of the population, but should have been related to both enterprise A and B.

Table 2.2. Example of a domestic enterprise in the GBR related to a domestic and non-domestic VAT unit.

Enterprise	VAT unit	Domestic
47823457	0015720983001	Yes
47823457	0015720983001	No

2.3.3 Linkages errors between observations and the population frame

Likewise to errors in the relations between unit types in the GBR, errors in the linkage of data sets to the population frame can be divided into incorrect positive links (mismatches) and incorrect missing links (missed matches). In the case study we use exact matching, so we do not expect many linkage errors between observations and the population frame. Therefore we do not divide linkage errors into subtypes in the present paper.

2.4 Strategy of detection and correction of errors

2.4.1 Introduction

We distinguish between three phases in the production process where we can detect the above-mentioned errors, analyse them, and correct them if possible. The first phase is during the formation of the GBR. The second phase is just after linkage of VAT and survey data to the population frame. Those two phases focus on incorrectly missed links.

The third phase is done after the first estimation of population totals. Population totals are checked for implausible outcomes at aggregate level and next we zoom into the micro data using selective editing. In this phase we check for all possible sources of error. If a record value is suspicious we need aids to find out what type of error occurred. In section 2.4.2-2.4.4 we describe the three phases.

2.4.2 Phase 1: analysing VAT and legal units in the GBR that are not related to enterprises

Within the GBR, some relations between unit types may be absent. E.g., VAT units may, even though they might be related to legal units, not be related to enterprises. This happens e.g., due to the time delay the Dutch GBR applies when effectively introducing new enterprises. Phase one focuses on detecting *errors leading to under coverage* and correcting them. Wrongly missing relations lead to over coverage of the VAT unit population compared to the enterprise population.

To reduce the effects of this phenomenon we are thinking about two actions:

- Analyse VAT units in the GBR that are related to legal units, but are not (yet) related to enterprises. Sort these VAT units according to historical (quarterly) turnover and select the units with the largest turnover to be analysed first. Profilers should analyse these units in depth and decide upon possible relations with enterprises.

- Reduce the time delay between forming an enterprise and effectively introducing an enterprise in the GBR, by making use of information from other sources.

The first action tries to trace errors leading to under coverage and yields the introduction of new relations in the GBR. This can be effectuated in a new release of the GBR. At Statistics Netherlands it is not possible to "introduce" newly emerged enterprises in an already released GBR. The second action reduces coverage errors due to time delays.

2.4.3 Phase 2: analysing VAT units that cannot be linked to the population frame

Linking tax declarations to the population frame via VAT units, it turns out that not all VAT units in the tax-declarations-file can be linked to the population frame. Phase 2 tries to detect VAT units that are wrongly not linked to the population frame, this concerns *errors leading to under coverage.*

We should distinguish between two situations:

- Not possible to link a VAT-declaration to a VAT-unit in the population frame

- Not possible to link a VAT-declaration to an enterprise in the population frame

The first situation may occur e.g., when the file with the tax-declarations contains non-domestic units that are not present in the GBR. Another reason could be time delay: a new VAT-unit is already present in the tax-declaration-file but not yet present in the GBR. Again, sorting these VAT-units with respect to their turnover and profiling the units with the largest turnover, might help to reduce the effect of these linkage errors. First results for Q3 2010 show that, after sorting, 1,5 per cent of the units with a VAT-declaration that cannot be linked to the GBR corresponds to 90 per cent of the declared turnover.

The second situation is similar to the situation as mentioned in section 2.4.2. However, now we have additional information from the tax declaration that profilers could use in analysing the "missing" relations.

2.4.4 Phase 3: strategy of editing after making population estimates

2.4.4.1 Introduction

The third phase detects all kinds of errors. In terms of errors in the relations between unit types, we expect that in this phase we can find *erroneous positive relations* making use of the estimated stratum turnover levels and changes. Strata with extremely large stratum turnover values, may have 'errors leading to over coverage' or 'erroneous relations'.

2.4.4.2 Indicators for detection of potentially wrong population estimates

For each publication cell (combination of NACE codes) we obtain a set of population estimates concerning turnover level and yearly/quarterly growth rate. We use several indicators to assess whether population estimates are plausible (Hoogland, 2011; Van Delden et al., 2010). The main indicators are

- Difference between estimated growth rate and expected growth rate;
- Turnover share of unedited potential influential errors.

The difference between the estimated growth rate and the expected growth rate is too large if

$$|G_{h,r}^{k,k-s} - E(G_h^{k,k-s})| > d_h^E$$

with:

$G_{h,r}^{k,k-s}$: the growth rate for quarter k with respect to quarter k-s for publication cell h and release r. In practice, we mainly consider s=4.

$E(G_{h,r}^{k,k-4}) = G_{h,r'}^{k-1,k-5}$, that is, the expected year-to-year growth rate for a specific quarter is the year-to-year growth rate for the most recent release (r') for the previous quarter.

d_h^E: user-specified marginal value for the difference between the estimated growth rate and the expected growth rate.

The following indicator can be used to assess micro data that is used to estimate the yearly growth rate and turnover level for publication cell h in quarter k.

$$R_h^{k,k-4} = \frac{\sum_{j \in h} V_j^{k,k-4} \max\left\{O_j^k, O_j^{k-4}\right\}}{\sum_{j \in h} \max\left\{O_j^k, O_j^{k-4}\right\}},$$

where

$V_j^{k,k-4} = 1$, if turnover value O_j^k for enterprise j in quarter k is a potential influential error (PIE) and it is not checked or an editor indicated that the record was checked, but there was insufficient information for editing. To determine whether O_j^k is a PIE O_j^{k-4} is used as a reference value.

$V_j^{k,k-4} = 0$, otherwise.

In the next section we explain how potential influential errors can be detected using reference values.

2.4.4.3 Indicator(s) for detecting potential errors at microlevel

To detect potential influential errors, score functions are used to assess the influence and risk associated with the net total turnover for an enterprise in a publication cell and quarter. A detailed description is available in Van Delden *et al.* (2010). The basic idea is described in Hoogland (2009). The turnover values O_j^k and O_j^{k-4} are used to determine the influence and suspiciousness of O_j^k. These turnover values can be either observed or imputed. We try to construct homogeneous strata in order to detect deviant turnover values and growth rates within strata. The score for influence (I) and suspiciousness (S) for a specific enterprise j and quarter k are multiplied:

$$R_j^{k,k-4} = I_j^{k,k-4} \times S_j^{k,k-4}$$

The total score $R_j^{k,k-4}$ is between 0 and ∞, and a higher score means that the net total turnover for an enterprise is considered more influential and/or suspicious. All enterprises j with a value

$$R_j^{k,k-4} \geq R_{\min}$$

are listed on the 'PIE list' that is shown to analysts (see below).

We give a rough description of the score functions used to assess influence and suspiciousness. To assess the influence for enterprise j we use the turnover

values O_j^k and O_j^{k-4}, and the robust estimates of O_j^k and O_j^{k-4}. The idea is that we do not want to underestimate the influence of an enterprise if a turnover value is too small. The influence is large if the maximum value of these four turnover values is large relative to an estimate of the total turnover in the publication cell.

To assess the suspiciousness for enterprise j we compute partial suspicion scores. These partial scores represent suspiciousness regarding one of the features below:

$S_{1,j}^{k-4}$: turnover value in quarter *k-4;*

$S_{2,j}^{k}$: turnover value in quarter *k;*

$S_{3,j}^{k}$: yearly growth rate in quarter k;

$S_{4,j}^{k}$: inverse of yearly growth rate in quarter k;

A feature is suspicious if the corresponding partial suspicion score is larger than 1. This is the case if a feature is extremely high within a stratum with enterprises, otherwise the partial suspicion score is equal to 1. For each enterprise we determine whether a feature is suspicious. A 4-digit code (I_1, I_2, I_3, I_4) is used to summarize the suspiciousness of features, see figure 2.3.

$I_1 = 1$, if $S_{1,j}^{k-4}$ is suspicious, otherwise $I_1 = 0$

$I_2 = 1$, if $S_{2,j}^{k}$ is suspicious, otherwise $I_2 = 0$

$I_3 = 1$, if $S_{3,j}^{k}$ is suspicious, otherwise $I_3 = 0$

$I_4 = 1$, if $S_{4,j}^{k}$ is suspicious, otherwise $I_4 = 0$

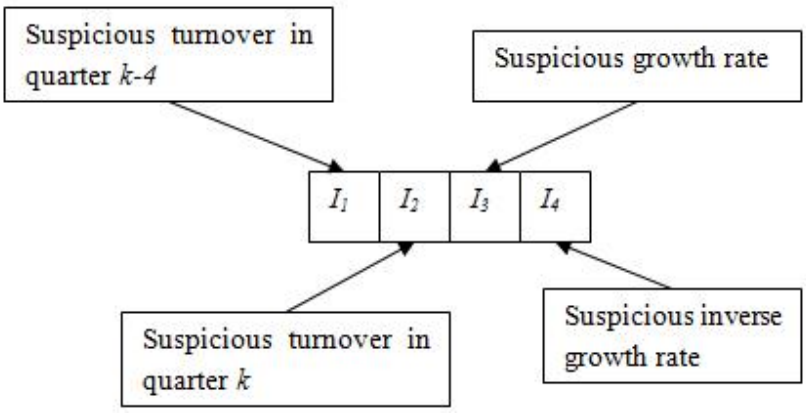


Figure 2.3. 4-digit code for suspiciousness

The partial suspicion scores are combined into an overall suspicion score.

$$S_j^k = \max(S_{1,j}^{k-4}, S_{2,j}^k) S_{3,j}^k S_{4,j}^k - 1$$

The overall suspicion score is 0 for enterprises where all partial suspicion scores are equal to 1, otherwise it is larger than 0.

An analyst decides whether enterprises on the PIE list are edited. He/she may consider additional information such as sorted tables with micro data and scatterplots showing the relation between a turnover variable in quarter k and the same turnover variable in quarter k-s. This may result in enterprises on the PIE list that are not selected for editing or additional enterprises with potential influential errors that have been selected for editing by an analyst.

2.4.4.4 Correction at microlevel: determining the error type

An editor has to determine the type of possible errors and a correction if there is an actual error. In the present paper, we focus on the detection of errors in the relation between unit types and in linkage errors. There are several aids for detection of relation and linkage errors

1) partial suspicion scores
2) seasonal patterns of turnover for an enterprise
3) seasonal patterns of turnover for VAT units related to an enterprise
4) features of an enterprise, linked legal units, and linked VAT units in combination with profiler knowledge

ad 1) Partial suspicion scores

The suspicion code gives an indication of the possible error type of a potential influential error that is detected by a score function. For example, suppose that an enterprise has a suspicious turnover value in both quarter k and quarter $k-4$ (code 1100). This indicates a potential error in the NACE code or size class, or a possible relation/linkage error that has been present for more than a year. Suppose that an enterprise has a suspicious turnover value in quarter k, a 'normal' turnover in quarter $k-4$ and a suspicious growth rate (code 0110). This is an indication of a possible erroneous turnover value or a possible linkage error in quarter k.

ad 2) Seasonal patterns of turnover for an enterprise

Another indication of a possible relation/linkage error is a shift in turnover for an enterprise from quarter $k-1$ to quarter k, in combination with a change in the VAT units that are related to the enterprise in quarter k. An enterprise is *stable* in quarter k if it has the same relation with VAT units as in quarter $k-1$, otherwise it is *unstable*.

ad 3) Seasonal patterns of turnover for VAT units related to an enterprise

In Table 2.3 an example is shown of a change in the VAT units that are related to an enterprise, which seems to cause a zero turnover in quarter 3. It is plausible that VAT unit 3333301 already declared turnover for the third quarter, but was (wrongly) not yet related to the enterprise 2 within the GBR.

Table 2.3. Turnover (× 1000 euros) patterns of VAT units related to enterprise 2.

VAT unit	Period	turnover
2222201	Quarter 1	2000
2222201	Quarter 2	2500
2222201	Quarter 3	0
3333301	Quarter 4	2200

In Table 2.4 an example is shown of an additional VAT unit that is related to enterprise 1 with a large effect on the quarterly turnover of enterprise 1. It seems that the large increase in turnover is not due to an erroneous turnover, because a large turnover is reported in each month in Q4. Features of VAT unit 3333301 and the corresponding legal unit and enterprise help to determine the type of possible error.

Table 2.4. Turnover (×1000 euros) patterns of VAT units related to enterprise 1.

VAT unit	Period	turnover
2222201	Quarter 1	20
2222201	Quarter 2	25
2222201	Quarter 3	24
2222201	Quarter 4	26
3333301	Q4, month 1	9000
3333301	Q4, month 2	10000
3333301	Q4, month 3	12000

ad 4) features of an enterprise, linked legal units, and linked VAT units in combination with profiler knowledge

An editor could check the available name and address of VAT units, and name address, legal form and number of working persons for legal units to check whether all VAT units are correctly related to the enterprise. Within the population frame information is available about changes in the composition of the enterprise (such as mergers). An event such as a merger may explain a structural shift in turnover for an enterprise and changes in the relation between the enterprise and VAT units.

2.5 Preliminary test on the effectiveness of score functions

We consider all enterprises that were edited within the DRT project for car trade and wholesale trade for Q4 2010. We use VAT data that were not edited before and survey data that were already edited for production of STS with the current statistical proces. We investigate the relationship between the type of error and whether an enterprise is on the PIE list.

The ratio 'number of records on PIE list that are selected for editing / total number of records selected for editing' says something about the effectiveness of the score function used to detect potential influential errors. Assuming that an analyst makes an effort to detect potential influential errors that are not on the PIE list.

Table 2.5 shows results for edited enterprises for car trade and wholesale trade for Q4 2010. Only 76 of the 92.225 active enterprises are edited. For 90.457 active enterprises we used VAT data instead of survey data. The number of edited records is small, because of time constraints and because part of the survey data was already edited for STS Q4 2010.

There were 38 enterprises on the PIE list, of which 23 were edited (Table 2.5) and 15 were not. From the 23 enterprises on the PIE list that were checked, the values of 7 enterprises were left unchanged (Table 2.5, category "no error"). Furthermore, about two third of the edited enterprises are not on the PIE list. This can mean that automatic detection of potential influential errors needs to be improved. It can also mean that some analysts select records for editing that are not influential or suspicious, or ignore enterprises on the PIE list that are influential and suspicious. The 4-digit code for suspiciousness shows that of the 38 enterprises on the PIE list, 28 enterprises

Table 2.5. Results for edited enterprises in 'car trade and wholesale trade', for the fourth quarter of 2010.

Error type	On PIE list	Not on PIE list	Total
Erroneous size class	2	2	4
Erroneous NACE code	0	0	0
Over coverage	0	1	1
Erroneous turnover value	13	23	36(1)
Linkage error	1	5	6
Unknown error type	0	0	0
No error	7	22	29(2)
Total	23	53	76

(1) 18 of them in survey data and 18 in tax declaration data

(2) 9 of them in survey data and 20 in tax declaration data

have a suspicious turnover level, 7 enterprises have a suspicious turnover growth and 3 enterprises have both.

Table 2.5 shows that six of the edited enterprises contain linkage errors. Several editors indicated that it was difficult to detect linkage errors by means of the available information. That is, there might be more linkage errors in records that were selected for editing. Most of the enterprises where linkage errors are detected are not on the PIE list. We have to adjust parameters and strata for the score functions in paragraph 2.4.4.3 in order to detect more linkage errors.

Based on the comments that editors made we conclude that they were not always sure that data were correct, but they did not correct data in these cases. So the actual number of errors in the edited records might be larger. Furthermore, we discovered an implementation error in the computation of our suspicion scores in the production system: imputed values were wrongly not included in the PIE list. In practice, imputed values were often edited. The effectiveness of the score function as given in Table 2.5 is therefore probably underestimated.

2.6 Summing up and topics for further research

Statistical and public agencies face the challenge to obtain plausible population estimates from combined data sources with different units. In this

paper we focused on economic data in which different unit types are used. The objective was to describe a strategy for detecting and correcting errors in the linkage and relations between units of integrated data. We used a case study concerning the estimation of quarterly and yearly turnover levels and growth rates, based on VAT declarations and survey data.

Potential influential errors have to detected, which can be due to errors in the relations between unit types in the GBR, from which the population frame is derived, and due to errors in linkage of data sets to the population frame. As a first result, a classification of relation and linkage errors is proposed based on our case study. We believe that the classification will also be useful for other National Statistical Institutes. We focused on errors in the relations between units, because we expect them to occur most in our case study. For the situation where errors in the linkage of a data set to a frame are likely to occur, the proposed classification can easily be extended using the categories: wrong positive linkage leading to over coverage, erroneous link, a missing link leading to under coverage and a missed link.

A second result from our study is that we split the detection of potential errors and their correction into different phases of the production process. We do so in order to find errors as early in the production process as possible although mainly errors leading to under coverage may be found in the first two phases. The use of different editing phases may also be interesting for any situation in which data are re-used. In the first phase we try to correct errors in the GBR as soon as possible, which is important as different statistics may use this GBR to derive a survey or population frame. The second phase focuses on VAT and survey data just after linkage. At SN, these data are stored in a central place, where it is made available for re-use by other statistics e.g. by international trade statistics. So also statistics that re-use data may profit from the data correction that has already been done.

Within each phase, we have some method that supports selective editing: we try to sort the errors according to their potential impact on the outcomes. In phase 1 and 2 we sort the units by turnover. In phase 3 we use a score function to detect potential influential errors.

In the first two phases we only deal with errors in the relations between units. In the third phase however, all kinds of errors may have occurred so we would like to be able to distinguish between them. We have developed several aids for the detection of a linkage error, such as a visualisation of seasonal patterns of turnover for VAT units related to an enterprise with a suspicious turnover. However, we believe that there is still a need to improve on this

aspect. The strategy to distinguish linkage from other errors needs to be specified explicitly and should be automatized if possible. For instance, we can automatize the determination of the stability of the relation between the VAT units and the enterprise. Additional indicators should be developed to determine the type of linkage error.

In terms of correcting the errors: in the case of economic statistics, so far, we need profilers to do this. Larger companies can have all kinds of constructions with underlying legal units and VAT units. Especially international companies may be very complicated and they may also be rather dynamic in terms of their underlying (base) units. In case of the smaller companies it may be useful to automatically collect internet information on name & activities. That may help to correct errors in the NACE code, for example.

In addition to the issues mentioned above, we came across several topics for future research. Additional VAT units may be linked using "loosened" matching rules as to increase the number of linked units. So far, tax and survey data are linked to the population frame using exact matching. Still, not all VAT-observations can be linked to the corresponding population frame which is partly due to time delays in the formation of the GBR. We found that we can link additional units by using just the fiscal base number, which is the core of the VAT identification number. Errors may occur more often when loosened linkage rules are used.

The efficiency of score functions might be improved. This can be done by optimizing the strata and parameters used to detect enterprises with a suspicious turnover. Another possible way to improve the efficiency of score functions is incorporate the effect of the number of working persons per enterprise on the turnover.

Chapter 3

Methodological developments on statistical matching

Marcello D'Orazio, Marco Di Zio and Mauro Scanu

Istituto nazionale di statistica - Istat, Italy

3.1 Problem description, illustrated with practical examples

Statistical matching techniques (D'Orazio *et al.*, 2006b) combine information available in distinct data sources referred to the same target population. The two datasets, A and B, are assumed to contain data collected in two independent sample surveys and such that:

1. the two samples contain distinct units (the sets of units observed in A and B do not overlap);

2. the two samples contain information on some variables X (common variables), while other variables are observed distinctly in one of the two samples, say, Y in A and Z in B.

In statistical matching, the key problem is the relationship among the variables Y and Z, given that they are never jointly observed in the data sets at hand. Analysts have always questioned if it was possible to either create synthetic records containing information on (X, Y, Z), as in Okner (1972), or through inference on model parameters describing the joint behaviour of

the variables, as correlation coefficients (Kadane, 1978, Rubin, 1986, Moriarity and Scheuren, 2001, 2003, Raessler, 2002), conditional probabilities (Renssen, 1998, D'Orazio *et al.*, 2006a), and so on.

The model is not identifiable given the available data, unless some untestable models are assumed, as conditional independence between Y and Z given X. Most of the literature on this topic is based on the conditional independence assumption. A more truthful study would limit inference on "how unidentifiable" is the model: this problem leads to the analysis of "uncertainty", as suggested by Raessler (2002} and D'Orazio *et al.* (2006b).

In this context, an inferential problem on (X,Y,Z) does not end up with a punctual estimate for the target quantities (function of the variable of interests), but, broadly speaking, consists of an interval that encloses all the possible values coherent with the observed information (*e.g.* an interval for a correlation coefficient of the variables "household expenditures" observed in a survey and "household consumption" observed in a second survey).

We remark that this interval is different from the inference based on confidence intervals, in the latter the uncertainty taken into account is due to sampling variability, in this case the uncertainty is due to the lack of information that implies model unindentifiability (Manski, 1995).

The computation of an interval instead of a single punctual value is inherently related to the absence of joint information on the variables observed distinctly in the two sample surveys. Papers in this context provide algorithms to build those intervals both in case of Gaussian distributions (Moriarity and Scheuren, 2001, Raessler, 2002, Kiesl and Raessler, 2008) and of multinomial distributions (D'Orazio, 2006a), however they refer to independent and identically distributed (iid) samples. In National Statistical Institutes (NSIs), survey data to be integrated generally refer to samples selected from the same finite population through a complex sampling design. There are not many studies concerning this issue (Rubin, 1986, Renssen, 1998), and they are not directly related to the analysis of "uncertainty".

Our aim is

1. to enclose the statistical matching procedures dealing with survey data drawn from finite populations in a unique framework that facilitates their comparison;

2. to study "uncertainty" in statistical matching in the framework of complex sampling designs;

3. to compare the different procedures in terms of their efficiency.

The continuous case, when the variables of interest are assumed to be normally distributed, has already been addressed in many papers and summarized in the ESSnet-ISAD WP1 document. In the sequel we will consider only categorical variables.

Practical examples are those related to the statistical matching applications considered in Istat so far:

- Joint use of FADN and FSS (Torelli *et al.*. 2008, Ballin *et al.*. 2009): in this case, the objective was to have joint tables on the economic variables (available on FADN) and structural variables (available on FSS) of the farms given some common characteristics (as the farm size, cattle characteristics, and so on).
- Joint use of Labour Force and Time use surveys (Gazzelloni *et al.*. 2007), where the time dedicated to daily work and to study its level of "fragmentation" (number of intervals/interruptions), flexibility (exact start and end of working hours) and intra-relations with the other life times (available on TUS) are studied together with the vastness of the information gathered in LFS on the aspects of the Italian participation in the labour market: professional condition, economic activity sector, type of working hours, job duration, profession carried out, etc.
- Joint use of the Survey of household income and wealth together with the household budget survey (Coli *et al.*, 2006), in order to have joint information on household expenditures and income given the socio-economic characteristics of the households.

All these samples are drawn according to complex survey designs.

3.2 Description of the available methods

When samples are drawn according to complex survey designs Rubin (1986) and Renssen (1998) are the two main references for statistical matching. A third approach given in Wu (2004), although not explicitly designed for statistical matching, can be used also in this context. In order to assess the uncertainty of statistical matching, it is enough to illustrate these approaches under the assumption of conditional independence of Y and Z given X.

Before showing the various approaches it is worth introducing some notations. Let U be the finite population of size N. Let us consider the random sample

A selected from U with the sampling design $p(A)$ consisting of n_A sample units and the random sample B selected from U with the sampling design $p(B)$ consisting of n_B sample units. Let $d_{A,a}=1/\pi_{A,a}$ be the direct weight associated to each sample unit in A, and $d_{B,b}=1/\pi_{B,b}$ be the corresponding direct weight for the units in B. The variables (X, Y) are observed in A while (X,Z) are observed in B.

3.3 Statistical matching with data from complex survey sampling

In literature it is possible to identify three statistical matching approaches when the data sources derive from complex sample surveys carried out on the same population:

a) Renssen's *calibrations based approach* (Renssen, 1998)

b) Rubin's *file concatenation* (Rubin, 1986)

c) Wu's approach based on *empirical likelihood* methods (2004)

The first two approaches can be considered as traditional ones while the latter is relatively new. Details concerning these methods are reported in Section 2.2 ("State of the art on statistical methodologies for data integration") of the Report of WP1.

The suggestion to use the pseudo empirical likelihood (PEL) to combine data from two sample surveys appeared in Wu (2004). This paper does not refer explicitly to SM but it nevertheless can be applied to such purpose. Two different approaches can be identified in the Wu (2004), the *separate* and the *combined* PEL.

3.3.1 Introduction to pseudo empirical likelihood

The finite population U $(y_1, y_2, \ldots, y_N)$ is viewed as an i.i.d. sample from a superpopulation and the *population log empirical likelihood* is (Chen and Sitter, 1999)

$$l_U(p) = \sum_{k \in U} \log(p_k)$$

where k is an indicator of population units, each one drawn with probability $p_k = P(y = y_k)$.

Given a generic probabilistic sample s selected from U, the Horvitz-Thompson estimator of $l_U(p)$ is

$$l_{HT}(p) = \sum_{k \in s} \frac{1}{\pi_k} \log(p_k)$$

with π_k the usual inclusion probability of unit k: $\pi_k = P(k \in s)$. Wu and Rao (2006) suggest a modified version called *pseudo empirical log-likelihood* (PELL):

$$l_{ns}(p) = n \sum_{k \in s} \tilde{d}_k(s) \log(p_k) \quad \tilde{d}_k(s) = d_k \Big/ \sum_{k \in s} d_k \quad d_k = 1/\pi_k$$

which takes into account the design effect when dealing with general unequal probability sampling without replacement.

In absence of auxiliary information, maximizing $l_{ns}(p)$ subject to the following constraints

$$1) p_k > 0, \quad k \in s \quad 2) \sum_{k \in s} p_k = 1$$

gives $\hat{p}_k = \tilde{d}_k(s)$ and the *maximum pseudo empirical likelihood* (MPEL) estimator of $\bar{Y}$ is the so called *Hajek estimator* $\hat{\bar{Y}}_H = \sum_{k \in s} \tilde{d}_k(s) y_k$. This latter estimator is less efficient than the Horvitz-Thompson estimator.

In presence of auxiliary information $\mathbf{X}$, a more efficient MPEL estimator of $\bar{Y}$ is obtained by maximizing $l_{ns}(p)$ subject to:

$$1) p_k > 0, \quad k \in s \quad 2) \sum_{k \in s} p_k = 1 \quad 3) \sum_{k \in s} p_k x_k = \bar{X}$$

Now p_k can be estimated using the Lagrange multiplier method, obtaining

$$\hat{p}_k = \frac{\tilde{d}_k(s)}{1 - \lambda' \left(x_k - \bar{X}\right)};$$

where λ is the solution of

$$g(\lambda) = \sum_{k \in s} \frac{d_k(s) \left(x_k - \bar{X}\right)}{1 + \lambda' \left(x_k - \bar{X}\right)} = 0.$$

It is worth noting that the new optimization problem is very similar to a calibration one. The only difference among the two methods consists in the

way of deriving the final weights. Wu suggests various methods to estimate the new weights (Wu, 2005). The advantage of PEL when compared to calibration, is that it produces only positive weights (constraint (1) in the maximization problem) while calibration may produce negative weights (e.g. in the case of linear calibration). Wu (2004) states: "GREG and EL are asymptotically equivalent but the latter has clear maximum likelihood interpretation and the resulting weights are always positive".

From the practical viewpoint the PEL have some further advantages. It is developed to deal with continuous $\mathbf{X}$ variables but it can handle also categorical variables (the dummy variables have to be considered). On the other hand, the calibration approach may fail when dealing with continuous variables or in presence of mixed type variables, in particular when the joint distribution of the $\mathbf{X}$ variables has to be considered: e.g. when dealing with a continuous and a categorical variable the joint distribution of the two variables can be maintained by calibrating with respect to the totals of the continuous variable computed in each category of the other variable, when dealing with two continuous variables their association is maintained if it is possible to include in the calibration their crossproduct.

Unfortunately, the theory underlying the PEL has to be modified to deal with stratified sampling when the allocation is not proportional (for more details see Wu, 2004; Wu and Rao, 2006). This is necessary in the general case of arbitrary sampling design within each stratum.

3.4 The EL to combine information from multiple surveys

Let us consider the simplest case of two sets of data sets A and B:

$$\{(y_a, x_a, d_a), a \in A\} \quad \{(z_b, x_b, d_b), b \in B\}$$

Resulting from a complex probabilistic not stratified sampling from the same target population U. The *separate approach* consists in keeping the two samples separate. More precisely, two separate estimation problems are considered:

Problem 1)

Maximize $l(p) = \sum_{a \in A} \frac{1}{\pi_{A,a}} \log(p_a), \quad p_a = P(y = y_a)$

Subject to: 1) $p_a > 0, a \in A$; 2) $\sum_{a \in A} p_a = 1$; 3) $\sum_{a \in A} p_a x_a = \bar{x}$.

Problem 2)

Maximize $l(q) = \sum_{b \in B} \frac{1}{\pi_{B,b}} \log(q_b), \quad q_b = P(z = z_b)$

Subject to: 1) $q_b > 0, \quad b \in B; \quad 2) \sum_{b \in B} q_b = 1; \quad 3) \sum_{b \in B} q_b x_b = \bar{x}.$

$\bar{x}$ can be known from external sources or can be estimated combining (linearly) $\bar{x}_A$ and $\bar{x}_B$ derived separately from the two surveys:

$$\bar{x}_{pool} = \delta \bar{x}_A + (1 - \delta) \bar{x}_B, \quad 0 \leq \delta \leq 1$$

The linear combination requires some choices concerning the importance, δ, of the estimate derived from A. Importance can be defined according to the quality of the estimates obtained from the two samples, for instance as a function of the mean square errors of $\bar{x}_A$ and $\bar{x}_B$. A rule of thumb can be given by $\delta = n_A/(n_A+n_B)$.

The EL separate approach is very similar to the Renssen (1998) harmonisation phase based on calibration of the weights of the two surveys in order to reproduce $\bar{x}$; in PEL approach calibration is substituted by alternative iterative algorithms.

As far as estimation is concerned, the Renssen's approach after the harmonization phase allows to estimate the parameter involving Y and Z or Y, Z and $\mathbf{X}$ under the Conditional Independence Assumption[1]. In particular, the parameters $\theta_{Y|X}$ can be estimated on A by using final calibrated weights; $\theta_{Z|X}$ is estimated on B by considering its final calibrated weights and θ_X can estimated either on A or on B (the same result comes out when using the final calibrated weights). It is straightforward to apply the same approach after harmonization based on the separate PEL.

The *combined* EL approach introduces a unique estimation problem:

Maximize

$$l(p, q) = \sum_{a \in A} \frac{1}{\pi_{A,a}} \log(p_a) + \sum_{b \in B} \frac{1}{\pi_{B,b}} \log(q_b)$$

Subject to the following constraints:

[1]Note that Renssen suggests also two ways to overcome uncertainty by introducing additional information, in terms of an additional complete third file with joint (X,Y,Z) observations. These approaches do not use the conditional independence assumption, but they are out of scope in the context of uncertainty where external information is assumed to be not available.

$$p_a > 0, \quad a \in A; \quad q_b > 0, \quad b \in B$$

$$\sum_{a \in A} p_a = 1, \quad \sum_{b \in B} q_b = 1$$

$$\sum_{a \in A} p_a x_a = \sum_{b \in B} q_b x_b$$

With respect to the separate case, a new EL is considered, and a new constraint (the last one) is added to the problem. This new constraint allows harmonizing the final sample to reproduce the same value of $\bar{x}$ without the need of knowing or estimating it.

3.5 Comparison among the approaches

The PEL approach to SM has been compared with the Renssen's one and with Rubin's file concatenation in a simulation study by D'Orazio *et al.* (2010). The simulation study is limited to the case of categorical variables: the objective is that of comparing the methods when estimating the relative frequencies involved in the Conditional Independence Assumption:

$$P(x, y, z) = P(y|x) \times P(z|x) \times P(x)$$

being $P(x = i) = (1/N) \sum_{k \in U} I(x_k = i) = N_i/N$.

The finite population U used in this study consists of an artificial population of $N = 5000$ individuals with age greater than 15 years and being occupied in a dependent position. The following variables are considered:

- Geographical Area (grouped respectively in 3 or 5 categories);
- Gender (X_1) (1='M', 2='F');
- Classes of Age (X_2) (3 classes: '16–22', '23–44', '45 and more');
- Education Level (Y) (4 categories: 1='No title or elementary school', 2='compulsory school', 3='Secondary school', 4='university degree or higher');
- Professional Status (Z) (3 categories: 1='worker', 2='employee', 3='manager').

In each simulation run two random samples, A and B, are selected from U. The samples are selected using a stratified random sampling with proportional allocation; the stratification is based on the Geographical Area: three strata ('North', 'Center' and 'South and Islands') are considered when selecting the sample A, while the strata are five ('NorthWest', 'North East', 'Center', 'South' and 'Islands') when selecting B. As far as the sampling fractions are concerned, three combinations are considered:

(a) $f_A = 0.10$ $(n_A = 500)$ and $f_B = 0.06$ $(n_B = 300)$;

(b) $f_A = f_B = 0.10$ $(n_A = n_B = 500)$;

(c) $f_A = 0.06$ $(n_A = 300)$ and $f_B = 0.10$ $(n_B = 500)$.

In sample A the variable Z is removed, in sample B the Y variable is removed. The whole process is repeated 10,000 times for each combination of the sampling fractions.

Table 3.1. Average of absolute distance (total variation distance x 100) among estimated and true (population) relative frequencies in case (a)

	$P(x)$	$P(x,y)$	$P(y\,\vert x)$	$P(x,z)$	$P(z\,\vert x)$
Rubin's file concatenation	2.617663	6.130204	5.494019	6.667832	5.494019
Renssen's calibration	2.617715	6.130005	5.493756	6.667069	5.493756
Wu separate	2.617715	6.130005	5.493756	6.667069	5.493756
Wu combined	2.795445	6.219056	5.493756	6.852094	5.493756

Table 3.2. Average of absolute distance (total variation distance x 100) among estimated and true (population) relative frequencies in case (b)

	$P(x)$	$P(x,y)$	$P(y\,\vert x)$	$P(x,z)$	$P(z\,\vert x)$
Rubin's file concatenation	2.318032	5.993903	5.482598	5.218500	5.482598
Renssen's calibration	2.318045	5.994167	5.482812	5.218904	5.482812
Wu separate	2.318045	5.994167	5.482812	5.218904	5.482812
Wu combined	2.499612	6.069208	5.482812	5.307666	5.482812

Tables 3.1, 3.2 and 3.3 report the averages, over the whole set of simulations, of the absolute distance (total variation distance) among the estimated and the true population relative frequencies (N_c/N) computed using:

$$\bar{d}\left(\hat{P}_c, P_c\right) = \frac{1}{10000} \sum_{t=1}^{10000} \left[\frac{1}{2} \sum_{c=1}^{C} \left| \hat{P}_{c,t} - P_c \right| \right],$$

Table 3.3. Average of absolute distance (total variation distance x 100) among estimated and true (population) relative frequencies in case (c)

	$P(x)$	$P(x,y)$	$P(y\,\vert x)$	$P(x,z)$	$P(z\,\vert x)$
Rubin's file concatenation	2.621003	7.746804	7.264135	5.382213	7.264135
Renssen's calibration	2.620999	7.747073	7.264329	5.382442	7.264329
Wu separate	2.620999	7.747073	7.264329	5.382442	7.264329
Wu combined	3.037953	7.894812	7.264329	5.487439	7.264329

being $\hat{P}_{c,t}$ the estimate of $P_c = N_c/N$ at the tth simulation run.

By looking at the simulation results, it appears that the methods here considered provide quite close results. Renssen's approach and Wu separate provide the same results, an expected result given that GREG and EL are asymptotically equivalent. When estimating the marginal distribution of the $\mathbf{X}$ variables, file concatenation is slightly better than the other ones, but the differences with Renssen approach and Wu separate are really negligible. Note that the Wu combined approach gives always the worst results.

As far as the joint or the conditional distributions are concerned, there are not manifest patterns in the results.

Given that all the methods provide estimates for $P(x)$,$P(y\,\vert x)$ and $P(z\,\vert x)$, the various approaches have been compared in terms of width uncertainty bounds for the cell probabilities of the table $P(y,z)$. In particular, using the (Fréchet class) it is possible to estimate the lower and the upper bounds for the cells in the table $P(y,z)$ through the following expressions:

$$\hat{P}_{kl}^{(low)} = \sum_{i,j} \left\{ \hat{P}(x_1 = i, x_2 = j) \;\times\; \max\left[0;\; \hat{P}(y = k\,\vert x_1 = i, x_2 = j) + \hat{P}(z = l\,\vert x_1 = i, x_2 = j) - 1\right]\right\}$$

$$\hat{P}_{kl}^{(up)} = \sum_{i,j} \left\{ \hat{P}(x_1 = i, x_2 = j) \;\times\; \min\left[\hat{P}(y = k\,\vert x_1 = i, x_2 = j)\,; \hat{P}(z = l\,\vert x_1 = i, x_2 = j)\right]\right\}$$

The following tables show the average estimate lower and upper bounds for each cell in the various cases.
Tables 3.10, 3.11 and 3.12 show summary results related to the average width of the intervals built using the estimates of $P(x)$,$P(y\,\vert x)$ and $P(z\,\vert x)$ obtained under the various approaches:

$$\bar{R}_c = \frac{1}{10000} \sum_{t=1}^{10000} \left(\hat{P}_{c,t}^{(up)} - \hat{P}_{c,t}^{(low)} \right).$$

Table 3.4. Average lower bounds of the uncertainty intervals for the cells in table of Y vs. Z in case (a)

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Est. bound with pop. Rel freq	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0	0.000058	0.000058	0.000058	0.000057
2	1	0.2806	0.010800	0.017578	0.017570	0.017570	0.017562
3	1	0.1536	0.040600	0.043473	0.043533	0.043533	0.043527
4	1	0.0074	0	0	0	0	0
1	2	0.0024	0	0	0	0	0
2	2	0.0536	0	0.000788	0.000789	0.000789	0.000782
3	2	0.2810	0.0148	0.021480	0.021509	0.021509	0.021507
4	2	0.0872	0	0.000002	0.000002	0.000002	0.000002
1	3	0	0	0	0	0	0
2	3	0.0060	0	0.000015	0.000015	0.000015	0.000014
3	3	0.0412	0	0.000054	0.000054	0.000054	0.000053
4	3	0.0386	0	0	0	0	0
	Average	1.0000					

Table 3.5. Average upper bounds of the uncertainty intervals for the cells in table of Y vs. Z in case (a)

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Est. bound with pop. Rel freq	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0.050800	0.050900	0.050867	0.050867	0.050869
2	1	0.2806	0.340200	0.338042	0.338009	0.338009	0.338007
3	1	0.1536	0.434400	0.427320	0.427347	0.427347	0.427355
4	1	0.0074	0.133200	0.133076	0.133091	0.133091	0.133090
1	2	0.0024	0.050800	0.050839	0.050806	0.050806	0.050809
2	2	0.0536	0.312600	0.305489	0.305486	0.305486	0.305489
3	2	0.2810	0.405000	0.393846	0.393867	0.393867	0.393881
4	2	0.0872	0.133200	0.133098	0.133112	0.133112	0.133112
1	3	0	0.048200	0.043179	0.043140	0.043140	0.043140
2	3	0.0060	0.085800	0.085791	0.085719	0.085719	0.085721
3	3	0.0412	0.085800	0.085835	0.085763	0.085763	0.085764
4	3	0.0386	0.076200	0.073677	0.073653	0.073653	0.073652
	Average	1.0000					

Table 3.6. Average lower bounds of the uncertainty intervals for the cells in table of Y vs. Z in case (b)

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Est. bound with pop. Rel freq	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0	0.000030	0.000030	0.000030	0.000030
2	1	0.2806	0.010800	0.015877	0.015894	0.015894	0.015892
3	1	0.1536	0.040600	0.042597	0.042546	0.042546	0.042539
4	1	0.0074	0	0	0	0	0
1	2	0.0024	0	0	0	0	0
2	2	0.0536	0	0.000397	0.000396	0.000396	0.000396
3	2	0.2810	0.014800	0.020028	0.020005	0.020005	0.020006
4	2	0.0872	0	0	0	0	0
1	3	0	0	0	0	0	0
2	3	0.0060	0	0.000005	0.000005	0.000005	0.000005
3	3	0.0412	0	0.000031	0.000031	0.000031	0.000031
4	3	0.0386	0	0	0	0	0
	Average	1.0000					

Tables 3.10-3.12 show that the average width of the cell proportions computed by considering the estimates of the Fréchet bounds remains essentially the same under the various approaches. In general, there is not an approach that outperforms the other ones. File concatenation seems to provide slightly better results than Renssen and Wu separate approaches. Wu combined approach tends to perform slightly worst than the other ones.

Table 3.7. Average upper bounds of the uncertainty intervals for the cells in table of Y vs. Z in case (b)

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Est. bound with pop. Rel freq	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0.050800	0.050828	0.050863	0.050863	0.050870
2	1	0.2806	0.340200	0.338501	0.338531	0.338531	0.338530
3	1	0.1536	0.434400	0.429847	0.429819	0.429819	0.429811
4	1	0.0074	0.133200	0.133083	0.133074	0.133074	0.133076
1	2	0.0024	0.050800	0.050795	0.050829	0.050829	0.050837
2	2	0.0536	0.312600	0.307095	0.307086	0.307086	0.307088
3	2	0.2810	0.405000	0.396211	0.396185	0.396185	0.396192
4	2	0.0872	0.133200	0.133087	0.133078	0.133078	0.133080
1	3	0	0.048200	0.044193	0.044227	0.044227	0.044230
2	3	0.0060	0.085800	0.085671	0.085712	0.085712	0.085707
3	3	0.0412	0.085800	0.085697	0.085738	0.085738	0.085733
4	3	0.0386	0.076200	0.074583	0.074595	0.074595	0.074591
	Average	1.0000					

Table 3.8. Average lower bounds of the uncertainty intervals for the cells in table of Y vs. Z in case (c)

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Est. bound with pop. Rel freq	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0	0.000049	0.000049	0.000049	0.000048
2	1	0.2806	0.010800	0.017456	0.017473	0.017473	0.017455
3	1	0.1536	0.040600	0.043622	0.043576	0.043576	0.043577
4	1	0.0074	0	0	0	0	0
1	2	0.0024	0	0	0	0	0
2	2	0.0536	0	0.000787	0.000787	0.000787	0.000772
3	2	0.2810	0.014800	0.021971	0.021964	0.021964	0.021928
4	2	0.0872	0	0.000001	0.000001	0.000001	0.000001
1	3	0	0	0	0	0	0
2	3	0.0060	0	0.000017	0.000017	0.000017	0.000016
3	3	0.0412	0	0.000059	0.000059	0.000059	0.000058
4	3	0.0386	0	0	0	0	0
	Average	1.0000					

Table 3.9. Average upper bounds of the uncertainty intervals for the cells in table of Y vs. Z in case (c)

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Est. bound with pop. Rel freq	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0.050800	0.050656	0.050686	0.050686	0.050674
2	1	0.2806	0.340200	0.338074	0.338106	0.338106	0.338138
3	1	0.1536	0.434400	0.427483	0.427446	0.427446	0.427495
4	1	0.0074	0.133200	0.132868	0.132848	0.132848	0.132845
1	2	0.0024	0.050800	0.050603	0.050633	0.050633	0.050622
2	2	0.0536	0.312600	0.305526	0.305528	0.305528	0.305556
3	2	0.2810	0.405000	0.393997	0.393992	0.393992	0.393997
4	2	0.0872	0.133200	0.132882	0.132862	0.132862	0.132858
1	3	0	0.048200	0.042914	0.042940	0.042940	0.042939
2	3	0.0060	0.085800	0.085609	0.085632	0.085632	0.085627
3	3	0.0412	0.085800	0.085657	0.085680	0.085680	0.085675
4	3	0.0386	0.076200	0.073646	0.073650	0.073650	0.073648
	Average	1.0000					

3.6 Comments

Though from the point of view of results there are not great differences, it is worth spending a few words on the practical aspects concerning the application of the various methods.

The file concatenation approach appears very simple: the samples are concatenated and estimation is carried out on a unique file. Unfortunately in

Table 3.10. Average width of the uncertainty intervals for the cells in table of Y vs. Z in case (a).

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0.050842	0.050809	0.050809	0.050812
2	1	0.2806	0.320464	0.320439	0.320439	0.320445
3	1	0.1536	0.383846	0.383815	0.383815	0.383827
4	1	0.0074	0.133076	0.133091	0.133091	0.133090
1	2	0.0024	0.050839	0.050806	0.050806	0.050809
2	2	0.0536	0.304700	0.304697	0.304697	0.304707
3	2	0.2810	0.372366	0.372359	0.372359	0.372375
4	2	0.0872	0.133096	0.133110	0.133110	0.133110
1	3	0	0.043179	0.043140	0.043140	0.043140
2	3	0.0060	0.085776	0.085704	0.085704	0.085707
3	3	0.0412	0.085780	0.085709	0.085709	0.085711
4	3	0.0386	0.073677	0.073653	0.073653	0.073652
	Average	1.0000	0.169804	0.169778	0.169778	0.169782

Table 3.11. Average width of the uncertainty intervals for the cells in table of Y vs. Z in case (b).

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0.050798	0.050833	0.050833	0.050840
2	1	0.2806	0.322623	0.322637	0.322637	0.322638
3	1	0.1536	0.387250	0.387273	0.387273	0.387272
1	1	0.0074	0.133083	0.133074	0.133074	0.133076
2	2	0.0024	0.050795	0.050829	0.050829	0.050837
3	2	0.0536	0.306699	0.306690	0.306690	0.306692
1	2	0.2810	0.376183	0.376180	0.376180	0.376186
2	2	0.0872	0.133087	0.133078	0.133078	0.133080
3	3	0	0.044193	0.044227	0.044227	0.044230
1	3	0.0060	0.085666	0.085707	0.085707	0.085702
2	3	0.0412	0.085666	0.085707	0.085707	0.085702
3	3	0.0386	0.074583	0.074595	0.074595	0.074591
	Average	1.0000	0.170886	0.170902	0.170902	0.170904

Table 3.12. Average width of the uncertainty intervals for the cells in table of Y vs. Z in case (c).

Education Level (Y)	Professional Status (Z)	Rel freq. in population	Rubin's File conc.	Renssen calibration	Wu separate	Wu combined
1	1	0.0484	0.050607	0.050637	0.050637	0.050626
2	1	0.2806	0.320618	0.320633	0.320633	0.320683
3	1	0.1536	0.383861	0.383870	0.383870	0.383917
1	1	0.0074	0.132868	0.132848	0.132848	0.132845
2	2	0.0024	0.050603	0.050633	0.050633	0.050622
3	2	0.0536	0.304739	0.304741	0.304741	0.304784
1	2	0.2810	0.372025	0.372028	0.372028	0.372069
2	2	0.0872	0.132882	0.132861	0.132861	0.132858
3	3	0	0.042914	0.042940	0.042940	0.042939
1	3	0.0060	0.085592	0.085615	0.085615	0.085611
2	3	0.0412	0.085598	0.085621	0.085621	0.085617
3	3	0.0386	0.073646	0.073650	0.073650	0.073648
	Average	1.0000	0.169663	0.169673	0.169673	0.169685

order to estimate the target quantities it is necessary to derive the new inclusion probability for each unit in the concatenated file. This task can be quite difficult to be achieved. In particular, the estimation of the $\pi_{k,A\bigcup B}$ requires knowledge of many kinds of information: (i) the sampling design $p(A)$ used to select A; (ii) the sampling design $p(B)$ used to select B; (iii) the A design variables in bith samples A and B; (iv) the B design variables in both samples A and B.

An approximate method to compute the $\pi_{k,A\bigcup B}$ is presented in Ballin *et al.* (2008a) but it requires the further availability of the sampling frame (assuming that both the samples have been selected from the same sampling frame).

Estimation of the inclusion probabilities for the units in the concatenated file permits a more accurate estimation of θ_X if compared to the other approaches (Renssen and PEL). On the other hand, the concatenated file does not allow for the direct estimation of $\theta_{Y|X}$ or $\theta_{Z|X}$. Methods to deal with missing values have to be chosen for estimating these parameters.

Finally it has to be considered that computation of the concatenated weights starts from the theoretic sample and unit nonresponse is not considered. In common practice, unit nonresponse has to be expected and file concatenation has to be performed by joining the respondents at two surveys carried out in the same population. This aspect introduces an element of further complexity into the problem and it is not clear how to handle unit nonresponse in this framework.

The Renssen's method allows to harmonize the marginal (joint) distributions of the $\mathbf{X}$ variables in both the samples but it requires knowledge or estimation of the totals of the $\mathbf{X}$ variables (when this information is not available from external sources). After the calibration step, it allows the direct estimation of $\theta_{Y|X}$ on A and $\theta_{Z|X}$ on B. Unfortunately, the calibration of the weights may not be successful. When dealing with too many variables it may be difficult to find a system of calibrated weights that satisfies all the constraints. This is likely to happen when dealing with continuous or mixed type X variables. The simplest way to solve the problem consists in the categorization of the continuous variables. This transformation introduces a further non-objective factor (how many classes? How to decide the break points? Etc.). Moreover, when linear calibration is considered there is the possibility of obtaining negative weights. This problem can be solved by introducing a further constraint, the positiveness of calibrated weights, into the optimization problem, but this may affect the successful solution of the problem.

A very important feature of the approach suggested by Renssen is that it can be applied also in the presence on unit nonresponse. In practice, it is possible to apply the methods by considering just the respondents to survey A and the respondents in survey B. This is important in practice, given that a certain amount of unit nonresponse is always encountered in sample surveys.

The approaches suggested by Wu (2004), have some interesting features.

As already mentioned, the separate approach is very similar to the calibration one. As for the calibration, the PEL approach consists in deriving new weights for the units in A and for the units in B that satisfy some constraints concerning the totals of $\mathbf{X}$ variables; in other words the two samples are harmonized in terms of the $\mathbf{X}$ variables. The PEL approach appears more flexible if compared to calibration because it does not provide negative weights and can be used in the presence of a mixed set of $\mathbf{X}$ variables. Moreover, when the totals of the $\mathbf{X}$ variables are unknown, the combined approach avoids their estimation; this advantage unfortunately produces slightly worse results when estimating the population parameters (as shown by the simulation results). From the theoretical view point the PEL introduces a major complexity, and the complexity increases in the presence of a stratified sampling with allocation that is not proportional (a very common situation in sample surveys). Moreover, the PEL approaches consider the theoretical samples and unit nonresponse is not considered. This latter feature represents an important limitation to the application of these approaches to the set of responding units at complex sample surveys (involving non-proportional stratification and clustering) carried out in NSIs.

Chapter 4

Handling incompleteness after linkage to a population frame: incoherence in unit types, variables and periods

Arnout van Delden, Koert van Bemmel[1]

Summary: In this chapter we concentrate on methods for handling incompleteness caused by differences in units, variables and periods of the observed data set compared to the target one. Especially in economic statistics different unit types are used in different data sets. For example, an enterprise (statistical unit type) may be related to one or more units in the Value Added Tax register. In addition those VAT units may declare turnover on a monthly, quarterly or yearly basis. Also the definition of turnover may differ from the targeted one. When tax data are linked to a population frame we may have response for only a part of the VAT units underlying the enterprise. We give an overview of different missingness patterns when VAT data are linked to enterprises and formulate methods to harmonize and complete the data at micro level.

Keywords: unit types, completion at micro level, harmonization

[1]Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail:adln@cbs.nl. Remark: The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

4.1 Background

4.1.1 Introduction

There is a variety of economic data available that is either collected by statistical or by public agencies. Combining those data at micro level is attractive, as it offers the possibility to look at relations / correlations between variables and to publish outcomes of variables classified according to small strata. National statistical institutes (NSI's) are interested to increase the use of administrative data and to reduce the use of primary data because population parameters can be estimated from nearly integral data and because primary data collection is expensive.

The economic data sources collected by different agencies are usually based on different unit types. These different unit types complicate the combination of sources to produce economic statistics. Two papers, the current paper and Van Delden and Hoogland (2011) deal with methodology that is related to those different unit types. Both papers deal with a Dutch case study in which we estimate quarterly and yearly turnover, where we use VAT data for the less complicated companies[2] and survey data for the more complicated ones.

Handling different unit types starts with the construction of a general business register (GBR) that contains an enumeration of the different unit types and their relations. From this GBR the population of statistical units that is active during a certain period is derived, the population frame. This population frame also contains the relations of the statistical units with other unit types, such as legal units. In the current paper we formulate a strategy for detecting and correcting errors in the linkage and relations between units of integrated data.

In the Dutch case study, after linkage, we handle differences in definitions of variables and completion of the data. After both steps, population parameters are computed. Both steps are treated in the current paper and resemble micro integration steps as described by Bakker (2011). After the computation of population parameters, an additional step of detecting and correcting errors is done as treated in the current paper.

In a next step, the yearly turnover data are combined at micro level (enterprise) with numerous survey variables collected for Structural Business Statistics. The paper by Pannekoek (2011) describes algorithms to achieve

[2]In the current paper 'company' is used as a general term rather than as a specific unit type.

numerical consistency at micro level between some core variables collected by register data and variables collected by survey data. Examples of such core variables in economic statistics are turnover, and wages. There are also other European countries that estimate such a core variable, e.g. turnover, from a combination of primary and secondary data. Total turnover and wage sums are central to estimation of the gross domestic product, from the production and the income side respectively.

Because the current paper and Van Delden and Hoogland (2011) share the same background, the current section 4.1.1 and the sections 4.1.2 and 4.2 are nearly the same in both papers.

4.1.2 Problem of unit types in economic statistics

The different unit types in different economic data sources complicate their linkage and subsequent micro integration. When a company starts, it registers at the chamber of commerce (COC). This results in a so called 'legal unit'. The government raises different types of taxes (value added tax, corporate tax, income tax) from these "companies". Depending on the tax legislation of the country, the corresponding tax units may be composed of one or more legal units of the COC, and they may also differ for each type of tax. Finally, Eurostat (EC, 1993) has defined different statistical unit types (local kind of activity unit, enterprise, enterprise group) which are composed of one or more legal units.

In the end, for each country, the set of unit types of companies may be somewhat different. But generally speaking, for each country, the legal units are the *base* units whereas tax and statistical units are *composite* units (see Figure 4.1). In some countries, like France, there is one-to-one relationship between legal units and tax units and tax units are one-to-one related to statistical units. In other countries, like the Netherlands, units that declare tax may be groupings of legal units that belong to different enterprises (Vaasen and Beuken, 2009). Likewise, in Germany, tax units may declare turnover for a set of enterprises (Wagner, 2004). As a consequence, at least in the Netherlands and Germany, for the more complex companies tax units may be related to more than one enterprise. In other words, the tax and statistical units are both composed of legal units, but their composition may be different.

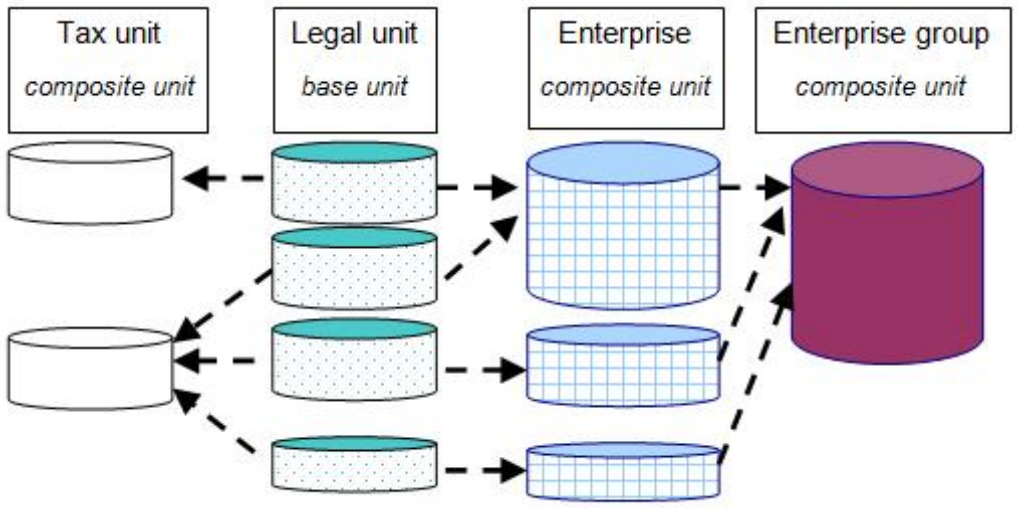


Figure 4.1. Different unit types in economic statistics. Each cylinder represents a single unit; arrows indicate the groupings of units.

4.1.3 General Business Register

NSI's have a GBR that contains an enumeration of statistical units and the underlying legal units. The GBR contains the starting and ending dates of the statistical units, their size class (SC code) and their economic activity (NACE code). In 2008, Eurostat has renewed its regulation on a business register (Eurostat, 2008) in order to harmonise outcomes over different European countries. NSI's also use a GBR to harmonise outcomes over different economic statistics within an NSI. In addition, the Netherlands - and other NSI's, also added the relations between legal units and tax units to the GBR, to be able to use tax office data for statistical purposes.

4.1.4 Problem description

The focus of the current paper is on incompleteness of the data at the level of the statistical unit after linkage of register and survey data to a population frame. This incompleteness can be due to the absence of source data (observations) or because the source data first need to be harmonised in order to get estimated values for the statistical unit and for the intended reporting period. We can also have partially missing information if for some but not all administrative units belonging to the same statistical unit the target variable is not (yet) available.

In the current paper we distinguish three main reasons for incompleteness:

1. observations are lacking;
2. the unit types of the source data differs from the statistical unit type;
3. the definition of the source variable differs from the target variable.

Each of these three reasons is explained in detail in section 4.3.

The objective of the current paper is to describe methods for handling incompleteness of a variable at the level of statistical units due to incoherencies in unit types and in variable definitions of source compared to target data.

In this study we used the following starting points. Firstly, we do not aim for perfect estimates for each single statistical unit but to have accurate estimates for publication cells. Secondly, we produce outcomes for different customers where each customer wants different strata. The basic publication cells from which all those strata can be constructed are rather small. We therefore wanted a method that uses the observed data as much as possible. Thirdly, the different customers have different moments at which they want their output, varying form very early (25 days after the end of the publication period) to very late (two years after the end of the publication year). We wanted to have a single estimation method that could be used for each of the releases. Finally, the method should be able to deal with all kinds of missingness patterns. We wanted a general approach that is also useful for NSI's with somewhat different missingness patterns. Given the third and fourth starting point we chose to use imputation, rather than weighting. We think that imputation provides flexibility to adjust the model to the corresponding missingness pattern and to deal with different publication moments.

4.1.5 Outline of paper

The remainder of the paper is organised as follows. Section 4.2 describes the Dutch case study. In section 4.3 we give a classification of missingness patterns, followed in section 4.4 by methodology to handle each of the missingness patterns. In section 4.5 we give an example of an accuracy test for some of the methods presented. Section 4.6 deals with special situations for which we wanted to make the outcomes more robust. Finally, in section 4.7 we sum up and suggest issues for further research.

4.2 Description of case study

4.2.1 Background: statistical output

In the current paper we deal with the estimation of Dutch quarterly and yearly turnover levels and growth rates, based on VAT declarations and survey data. The work is part of a project called "Direct estimation of Totals". Turnover is estimated for the target population which consists of the statistical unit type the *enterprise*. Turnover output is stratified by NACE code $\times$ size class. An overview of all processing steps from input to output data can be found in Van Delden (2010).

The estimated quarterly figures are directly used for the short term statistics (STS). Also, the quarterly and yearly turnover levels and growth rates are input to the supply and use tables of the National Accounts, where macro integration is used to obtain consistent estimates with other parameters. Also, results are used as input for other statistics like the production index (micro data) and the consumption index (the estimates). Finally, yearly turnover is integrated at micro level with survey data of the Structural Business Statistics (SBS). Next, the combined data is used to detect and correct errors in both the turnover data as well as in the other SBS variables. Yearly turnover results per stratum are used as a weighting variable for SBS data.

In fact we deal with four coherent turnover estimates:

- net total turnover: total invoice concerning market sales of goods and services supplied to third parties excluding VAT;
- gross total turnover: total invoice concerning market sales of goods and services supplied to third parties including VAT;
- net domestic turnover: net turnover for the domestic market, according to the first destination of the product;
- net non-domestic turnover: net turnover for the non-domestic market, according to the first destination of the product.

More information on the turnover definition can be found in EC (2006). In the remainder of the paper we limit ourselves to net total turnover further referred to as turnover.

The quarterly and yearly figures are published in different releases, as shown in Table 4.1. The quarterly releases vary from a very early estimate delivered at 30-35 days after the end of the corresponding quarter to a final estimate

for SBS publication delivered April year $y+2$ where y stands for the year in which the target period falls.

Table 4.1. Overview of the releases of the case study

Release	Period of estimation	Moment	Explanation
Flash estimate	Quarter	30–35 days after end of target period	Provisional estimate delivered for Quarterly Accounts, STS branches with early estimates
Regular estimate	Quarter	60–70 days after end of target period	Revised provisional estimate for Quarterly Accounts and for STS
Final STS estimate	Year and corresponding 4 quarters	April y+1, one year after target year	The estimates of the four quarters are consistent with the yearly figure
Final SBS estimate	Year and corresponding 4 quarters	April y+2, two years after target year	The estimates of the four quarters are consistent with the yearly figure. The yearly figure is based on STS and SBS turnover data

4.2.2 Target population and population frame

The statistical target population of a *period* consists of all enterprises that are active during a *period*. This true population is unknown. We represent this population by a frame which is derived from the GBR. Errors in this representation are referred to as frame errors. Each enterprise has an actual and a coordinated value for the SC and NACE code. The coordinated value is updated only once a year, at the first of January and is used to obtain consistent figures across economic statistics. In the remainder of the paper we always refer to the coordinated values of SC and NACE code unless stated otherwise.

The population frame is derived as follows. First, each month, we make a view of the GBR that represents the population of enterprises that are active at the first day of the month; in short: the population state. This population state also contains the legal units, tax units and the 'enterprise groups'-units that are related to the enterprise population at the first day of the month. Next, the population frame for a period is given by the union of the relevant population states. For example, the frame for the first quarter of a year consists of the union of the population states on 1 January, 1 February, 1 March and 1 April.

For the case study, the frame contains four unit types: the legal unit (base unit), the enterprise (composite unit) and two tax units namely the base tax unit and the VAT unit. In the Netherlands each legal unit (that has to pay tax) corresponds one-to-one to a base tax unit. For the VAT, base tax units may be grouped into a VAT unit (composite unit). So this is an extension of the more general situation of Figure 4.1. A more extensive description can be found in Van Delden and Hoogland (2011).

As explained in Vaasen and Beuken (2009), in the case of smaller companies each VAT unit is related to one enterprise and each enterprise may consist of one or more VAT units. For the more complicated companies, referred to as topX units, a VAT unit may be related to more than one enterprise.

4.2.3 Data

In the case study we use two types of source data. We use VAT data for the non-topX enterprises. For the topX enterprises we use primary data because VAT units may be related to more than one enterprise. This approach is quite common, also at other NSI's in Europe (e.g. Fisher and Oertel, 2009; Koskinen, 2007; Norberg, 2005; Orjala, 2008; Seljak, 2007). For the non topX units, we only use observations of VAT units that are related to the target population of enterprises.

Concerning the VAT, a unit declares the value of sales of goods and services, divided into different sales types. The different sales types are added up to the total sales value, which we refer to as turnover according to the VAT declaration.

In the current paper we use VAT and survey data for 2008 and 2009 and the first two quarters of 2010. Data are stratified according to NACE 2008 classification.

4.3 Classification of missingness patterns

4.3.1 Introduction

We classify the missingness patterns along three main reasons for missingness:

- observations are lacking;
- the unit types of the source data differs from the statistical unit type;
- the definition of the source variable differs from the target variable.

These three main reasons are further subdivided in section 4.3.2-4.3.4. Although this classification is derived from the Dutch case study, we believe that the three main reasons of missingness also apply to other situations (variables and NSI's).

Note that we structure *patterns* not units: we do not make a classification where each unit with a pattern only falls into one class. In practice, units can have two missingness patterns simultaneously. For example: a different unit structure can coincide with a different variable meaning. The imputation method handles the missingness patterns in a certain order, as explained in section 4.4.6.

4.3.2 Missingness due to lack of observations

4.3.2.1 Classification

We distinguish four kinds of missingness due to lack of observations:

(1a) Units in the population frame that did not respond (yet) but that have responded in the past.

For the quarterly data, a flash estimate is made and used in National Accounts. This flash estimate is delivered at about 30 days after the end of the period, see Table 4.1 of section 4.2.1. The processing time is currently around five days, so we can use data that are delivered to the tax office up to 25 days after the end of the period. At that moment 40-50% of the expected VAT units have not (yet) responded. Many of them have historical observations. Some will respond later, others are ended.

(1b) Units in the population frame with a structural lack of observations

Some structural non responding VAT units have dispensation from the tax office, because they are very small or because their activities require no tax obligations. Others may evade tax or they may be wrongly present in the frame. Also sample survey units may be structurally non respondent.

(1c) Units in the population frame that did not respond (yet) and that are new in the population frame

(1d) Units that do belong to the conceptual population but are wrongly not present in the population frame

Under coverage in the population frame is not solved by imputation but by the correction of linkages or relations between units, as explained in Van

Delden and Hoogland (2011).

4.3.2.2 Quantification

To quantify the occurrence of the different patterns of missingness, we counted the number of VAT-units with a tax declaration and their corresponding turnover after linkage to the population frame over Q4 2009 and Q1 2010, for NACE-code I: "Accommodation and food service activities". For both quarters we made two estimates: an early one and a late one (see Table 4.2) in which we "simulated" the releases as shown in Table 4.1. For units with historical turnover (missingness class 1a) we used units that had at least one declaration since January 2008. The results are shown in Table 4.3-Table 4.6).

Table 4.3 shows that 24494 VAT units have responded at the flash estimate for Q4 2009, compared to 46059 VAT units at the final estimate, corresponding to 4.26 and 7.42 milliard Euros respectively. When we only count the turnover of those enterprises where all related VAT units have responded for the full three months at the flash estimate of Q4 2009 (see complete declarations in Table 4.3) we get only 2.95 milliard Euros.

Figures of Q1 2010 are similar, see Table 4.4. Note that units that declare on a yearly basis were included in the Q4 2009 but not in the Q1 2010 counting's. We could not include the latter because our data file was up to Q3 2010 and therefore their yearly declarations over 2010 were not present in the file.

Table 4.5 shows that 46059 VAT units have responded over Q4 2009 at the final estimate. A subset of 26628 VAT units has historical responses but did not yet respond at the flash estimate (missingness pattern 1A). At the final estimate, 26023 of the 26628 non respondents were shown to be late respondents; the other 605 VAT units never responded and (probably) were ended.

In Table 4.6 we can see that patterns 1B and 1C occur far less frequently than pattern 1A. At the flash estimate over Q1 2010 24699 units were non respondent with a historical turnover (pattern 1A), 1541 units were non respondent with a structural lack of turnover (pattern 1B) and 1606 units were non respondents that were new in the population frame (pattern 1C). Also in terms of turnover, pattern 1A is far more important than pattern 1B and 1C. Note that some of the units with a structural lack of turnover as well as some of the new units have a tax office code $\neq 0$ which means they have dispensation from the tax office.

Table 4.2. Approximation of releases for Q4 2009 and Q1 2010 to estimate the frequency of missingness patterns among VAT units.

Period	Release (approximated)	Latest arrival date of declarations at tax office	Remark
Q4 2009	Flash estimate	25–1–2010	Partial response for monthly, quarterly and yearly respondents.
	Final estimate	26–8–2010*	Response (nearly) complete for monthly, quarterly and yearly respondents.
Q1 2010	Flash estimate	25–4–2010	Partial response for monthly, and quarterly respondents.
	Regular estimate	26–8–2010*	Response (nearly) complete for monthly and quarterly respondents. No yearly respondents yet.

* We took the latest available date in the data file

Table 4.3. Number of VAT units and corresponding turnover for Q4 2009 and NACE "Accommodation and food service activities".

Type of declaration	Flash estimate			Final estimate		
	Total	TopX	non-TopX	Total	TopX	non-TopX
	Number of units					
Total	24494	206	24288	46059	284	45775
Monthly	8854	139	8715	9080	139	8941
Quarterly	15159	66	15093	32536	140	32396
Yearly[1]	480	1	479	4389	5	4384
Other[2]	1	0	1	55	0	55
	Declared Turnover (× 10^9 Euros)					
Total	4.26	2.28	1.98	7.42	3.94	3.48
Monthly	2.80	1.67	1.13	3.86	2.41	1.46
Quarterly	1.45	0.61	0.84	3.48	1.53	1.95
Yearly	0.01	0.00	0.01	0.08	0.00	0.07
Other	0.00	0.00	0.00	0.00	0.00	0.00
	Turnover of complete declarations (× 10^9 Euros)					
Total	2.95	1.64	1.31			
Monthly	1.62	1.12	0.50			
Quarterly	1.32	0.52	0.80			
Yearly	0.01	0.00	0.01			
Other	0.00	0.00	0.00			

[1]Quartely turnover of units that declare on yearly basis is computed as yearly turnover divided by 4.
[2] Shifted calendar quarter (stagger)

We compared the number and turnover of VAT declarations that could be linked to the population frame of Q4 2009 versus those that could not be linked, over the *full range* of NACE codes at the final estimate, see Table 4.7. Results show that about 3 per cent of the turnover in the declaration

Table 4.4. Number of VAT-declarations and corresponding turnover for Q1 2010 and NACE "Accommodation and food service activities".

Type of declaration[1]	Flash estimate			Regular estimate		
	Total	TopX	non-TopX	Total	TopX	non-TopX
	Number of units					
Total	26241	202	26039	42417	278	42139
Monthly	9033	140	8893	9297	140	9157
Quarterly	17209	62	17147	33121	138	32983
Yearly	0	0	0	0	0	0
Other	0	0	0	0	0	0
	Declared Turnover ($\times$ 10^9 Euros)					
Total	4.00	2.11	1.89	6.20	3.21	2.99
Monthly	2.55	1.54	1.01	3.13	1.87	1.26
Quarterly	1.45	0.57	0.88	3.06	1.34	1.73
Yearly	0.00	0.00	0.00	0.00	0.00	0.00
Other	0.00	0.00	0.00	0.00	0.00	0.00
	Turnover of complete declarations ($\times$ 10^9 Euros)					
Total	3.29	1.87	1.42			
Monthly	2.00	1.43	0.57			
Quarterly	1.29	0.44	0.85			
Yearly	0.00	0.00	0.00			
Other	0.00	0.00	0.00			

[1] see footnotes of Table 4.3

Table 4.5. Number of VAT-units and corresponding turnover for Q4 2009, missingness pattern 1A[1] and NACE "Accommodation and food service activities".

		Number of units			Turnover ($\times$ 10^9 Euros)		
		Total	TopX	non-TopX	Total	TopX	non-TopX
Response at final estimate	Total	46059	284	45775	7.42	3.94	3.48
Pattern 1A	Total missing at flash estimate[1]	26628	157	26471			
	Response after flash estimate						
	Total	26023	156	25867	3.14	1.66	1.48
	Monthly	5377	80	5297	1.06	0.73	0.33
	Quarterly	17105	74	17031	2.01	0.92	1.09
	Yearly[2]	3487	2	3485	0.06	0.00	0.06
	Other[2]	54	0	54	0.00	0.00	0.00
	No later quarterly response	605	1	604			

[1] VAT units with at least one historical VAT declaration since January 2008.
[2] see footnotes of Table 4.3.

file could not be linked to the population frame. Since 2010, the turnover that cannot be linked to the population frame has gradually been reduced

Table 4.6. Number of VAT-units and corresponding turnover for Q1 2010, missingness patterns 1A–1C and NACE "Accommodation and food service activities".

		Number of units			Turnover ($\times$ 10^9 euros)		
		Total	TopX	non-TopX	Total	TopX	non-TopX
Response at regular estimate		42417	278	42139	6.20	3.21	2.99
Pattern 1A	Total missing at flash estimate[1]	24699	146	24553			
	Response after flash estimate	19967	141	19826	2.15	1.10	1.06
	Monthly [1]	5245	66	5179	0.60	0.33	0.26
	Quarterly	14722	75	14647	1.56	0.77	0.79
	Yearly	0	0	0	0.00	0.00	0.00
	Other	0	0	0	0.00	0.00	0.00
	No later quarterly response	4732	5	4727			
Pattern 1B	Total missing at at flash estimate	1541	8	1533			
	Code = 01 [2]	239	0	239			
	Code $\neq$ 01	1302	8	1294			
	Response after flash estimate	56	0	56	0.001	0.000	0.001
	Code = 01	56	0	56	0.001	0.000	0.001
	Code $\neq$ 01	0	0	0	0.000	0.000	0.000
Pattern 1C Pattern 1C	Total missing at flash estimate	1609	4	1602			
	Code = 01	1110	1	1109			
	Code $\neq$ 01	148	0	148			
	Code missing	348	3	345			
	Response after flash estimate	951	1	950	0.04	0.00	0.04
	Code = 01	948	1	947	0.04	0.00	0.04
	Code $\neq$ 01	1	0	1	0.00	0.00	0.00
	Code missing	2	0	2	0.00	0.00	0.00

[1] see footnotes of Table 4.3 [2] Code = 01: have to declare tax, Code $\neq$ 01: tax dispensation

to about 1 per cent of the total declared turnover due to improvement in the population frame.

4.3.3 Missingness due to different unit structure

4.3.3.1 Classification

We distinguish between two kinds of missingness due to a different unit structure:

(2a) Observations (e.g tax declarations) are related to one enterprise group but to more than one underlying enterprise.

Table 4.7. Number of VAT-units and corresponding turnover in the tax declaration file split into 'linked' and 'not linked' to the population frame of Q4 2009.

Linked to pop.frame Q42009	Type of declaration[1]	STS-domain			non STS-domain	
		Total	TopX	non-TopX	TopX	non-TopX
		Number of VAT units				
Linked	Total	1132741	7306	813312	3380	308743
	Monthly	179631	3413	144700	1133	30385
	Quarterly	826622	3636	592131	2014	228841
	Yearly	119895	256	76039	232	43368
	Other	6600	1	443	1	6155
Not linked	Total	289652				
Linked later	Total	43447				
		Declared Turnover ($\times$ 10^9 Euros)				
Linked	Total	327.2	147.3	144.1	11.5	24.3
	Monthly	164.8	68.8	79.1	6.6	10.2
	Quarterly	160.7	78.3	64.2	4.9	13.3
	Yearly	1.2	0.1	0.8	0.0	0.3
	Other	0.5	0.0	0.0	0.0	0.4
Not linked	Total	12.4				
Linked later	Total	4.2				

[1] see footnotes of Table 4.3

As explained in section 4.2.2 VAT declarations are mostly related to one enterprise group, but can be related to more than one enterprise underlying the enterprise group. The latter can be a problem because we wish to make estimates for strata defined by NACE codes and size classes which are properties of enterprises.

(2b) Observations (e.g tax declarations) are related to more than one Enterprise Group and also to more than one underlying enterprise.

Note that sometimes a VAT declaration is related to more than one enterprise group. This may for example occur with a unit that declares on a yearly basis and within that year the VAT unit has been taken over by a new enterprise group.

4.3.3.2 Quantification

We counted the occurrence of missingness patterns 2A and 2B for the Accommodation and food service activities over 2010 (Table 4.8. In Q1 2010 a total of 60775 tax declarations were linked to the corresponding population frame at the regular estimate. A total of 556 VAT declarations were related to topX enterprises. For the majority of them, 507, a VAT declara-

tion is related to one Enterprise Group and to one enterprise, 40 declarations were related to one Enterprise Group but to more than one enterprise, and 9 were related to more than one Enterprise Group as well as to more than one enterprise. Although only 49 declarations were related to more than one enterprise this corresponded to a quarterly turnover of 2.44 milliard euros compared to the total of 3.21 milliard euros for topX enterprises. From this we can conclude that mainly tax declarations of topX entities are related to more than one enterprise. From Table 4.8 we can see that also 56 declarations were related to more than one non-topX enterprise, corresponding to a much smaller quarterly turnover of 0.02 milliard euros.

Table 4.8. Number and turnover of VAT-declarations for Q1 2010 at the regular estimate by type or relation, for Accommodation and food service activities.

Type of declaration	Number of declarations			Turnover ($\times$ 10^9 Euros)		
	Total	TopX	non-TopX	Total	TopX	non-TopX
All types of relations of VAT unit to Enterprise group and enterprise						
Total	60755	556	60199	6.20	3.21	2.99
Monthly	27634	418	27216	3.13	1.87	1.26
Quarterly	33121	138	32983	3.06	1.34	1.73
Yearly	0	0	0	0.00	0.00	0.00
Other	0	0	0	0.00	0.00	0.00
A VAT-unit related to one Enterprise Group and to one Enterprise						
Total	60650	507	60143	3.74	0.77	2.97
Monthly	27553	379	27174	1.61	0.36	1.25
Quarterly	33097	128	32969	2.13	0.41	1.72
Yearly	0	0	0	0.00	0.00	0.00
Other	0	0	0	0.00	0.00	0.00
A VAT-unit related to one Enterprise Group but to more than one Enterprise						
Total	40	40	0	1.83	1.83	0.00
Monthly	30	30	0	0.90	0.90	0.00
Quarterly	10	10	0	0.93	0.93	0.00
Yearly	0	0	0	0.00	0.00	0.00
Other	0	0	0	0.00	0.00	0.00
VAT-unit related to more than one Enterprise Group and to more than one Enterprise						
Total	65	9	56	0.63	0.61	0.02
Monthly	51	9	42	0.62	0.61	0.01
Quarterly	14	0	14	0.01	0.00	0.01
Yearly	0	0	0	0.00	0.00	0.00
Other	0	0	0	0.00	0.00	0.00

4.3.4 Missingness due to different meaning of variable

4.3.4.1 Classification

The third cause of missingness is because the meaning of the variable in the data set differs from the target variable. We subdivide this into three types:

(3a) Observations are available for a period that is longer than the target period.

In the Netherlands and in many European countries, a VAT unit may report to the tax office on a monthly, quarterly or yearly basis. Generally speaking, the larger the unit, the more frequently it has to declare its turnover to the tax office[3]. The exact rules differ from country to country (Statistics Finland, 2009).

At the end of the year we make a release where we estimate yearly and quarterly turnover, which are numerically consistent with each other. For this final estimate, we use yearly turnover observations and divided those over the four underlying quarters.

(3b) Observations are available for a period that is shifted compared to target period.

Some VAT units declare tax on a quarterly basis, but the period is shifted compared to a calendar quarter. For example, units declare tax for February–April, or for March–May. Those units are referred to as "staggers". In the Netherlands staggers are rare but they occur frequently in the United Kingdom (Orchard *et al.*, 2010).

(3c) Observations are available but need to be transformed due to definition differences.

In the tax declaration form, the VAT-units declare the value of products and services that have been sold, the turnover. From this declaration, the amount of tax to be paid is calculated. However, the turnover found on the tax declaration form may differ from the one defined by Eurostat (EC, 2006):

- Some units have dispensation for part of their turnover. This is the case for some branches;
- Other tax rules; for example for some activities units don't have to declare the total turnover but only the profit margin[4].
- Intra-enterprise turnover. The statistical variable turnover only consists of market-oriented sales. When an enterprise consists of two or

[3]In the Netherlands only very small units are allowed to report tax on a yearly basis. However, since July 2009, many units are allowed to report on a quarterly basis instead of on a monthly basis.

[4]This applies to trade in second-hand goods that are sold to enterprises without a tax number or to private persons.

more VAT units, the declared turnover may partly consist of deliveries of goods or services within the enterprise which is not market-oriented.

4.3.4.2 Quantification

In Table 4.3 we can see that of the total of 46059 VAT units that declared tax over Q4 2009 in the Accommodation and food service activities at the final estimate, 4389 units declared tax on a yearly basis (pattern 3A) and 55 VAT units were "staggers' (pattern 3B). In terms of turnover this corresponded to 7.42 milliard euros for the total, 0.08 milliard euros for the yearly tax reporters and less than 0.01 milliard euros for the staggers.

As far as we know, the differences in the Netherlands between tax and target turnover are limited, within the domain of the STS regulation. Based on an analysis of tax data and SBS survey data over 2009, we found that for about 10 per cent of the 4 digit NACE codes within the STS domain, VAT turnover cannot be used because differences in definition are too large. For a further small number (less than 10) of 4 digit NACE codes we derive the target turnover from the statistical turnover. For the remaining nearly 90 per cent of the 4 digit NACE codes in the STS domain the VAT turnover corresponds closely to the target turnover. All those figures concern the total net turnover. For some STS domains the net turnover has to be split into sales to customers within the Netherlands (domestic) versus sales to customers outside the Netherlands (non domestic). This subdivision may sometimes be more difficult to estimate from VAT declarations.

4.4 Solutions for each type of missingness

4.4.1 Introduction

In section 4.4.2 explains at which unit level missing values are imputed. Sections 4.4.3 and 4.4.4 deal with completion, section 4.4.5 deals with harmonization. Finally section 4.4.6 explains some practical implementation issues. Some methodology to make the imputations more robust is treated in section 4.6. The methodology as described in the paper has slightly been simplified. Imputation at the level of the legal unit has been omitted because we use it only in some exceptional situations.

4.4.2 Level of imputation: statistical unit versus VAT unit

We analysed whether we wanted to impute at enterprise or at VAT unit level. To explain this, Table 4.9 shows an example of an enterprise is related to two VAT units. The quarterly turnover of the enterprise is given by the sum of the turnover of the two VAT units. Say we are in Q1 2010 and wish to publish data for Q4 2009 and for the whole year of 2009 In 2009 Q4 turnover of VAT unit 2 is missing. VAT unit 2 has reported for the previous quarters. This is a situation which often occurs with early estimates, see section 4.3.2. To complete the turnover of enterprise 1, we could impute the quarterly turnover of VAT unit 2 or we could discard the observed turnover and impute directly the total turnover for enterprise 1.

In order to make a choice we counted the number of VAT-units, classified according to the type of relation between VAT units and the enterprise within the STS-domain in the population frame of December 2009. 86 per cent of the enterprises were related to just one VAT unit and the remaining 14 per cent were related to two or more VAT-units.

Table 4.9. Part of the turnover of the statistical unit is completely missing for some quarter, but turnover is complete for historical quarters.

Enterprise Id	VAT id	Quarterly turnover					Yerly turnover
		2008	2009				2009
		Q4	Q1	Q2	Q3	Q4	
1	1	102	100	105	95	103	403
1	2	27	25	30	30	?	?
Total		129	125	135	125	?	?

We compared the turnover response from VAT declarations for early and late estimates in the Accommodation and Food service activities in Q4 2009 (Table 4.3) and in Q1 2010 (Table 4.4). In Q4 2009, the VAT units related to non-topX enterprises declared 3.48 milliard euros turnover at the final estimate. For the flash estimate we can only use declarations that were sent up to 25 days after the end of the quarter. For the flash, just 1.98 milliard euros were declared by non-topX VAT units. If we further limit the declarations to those that have a complete quarterly turnover for the enterprise, it drops down to 1.31 milliard euros. For Q1 2010 we found similar results. The main reason why quarterly turnover is incomplete at enterprise level for the flash estimate is that VAT units that declare monthly have declared just two of the three monthly periods. Another reason is that

some of the enterprises consist of more than one VAT unit, of which one did not yet respond.

The simplest imputation method would directly impute quarterly turnover at the level of the enterprise. In that case, the turnover of all units that did not have a complete quarterly turnover at enterprise level would be ignored and instead a value would be imputed. In the case of the Accommodation and Food service activities for Q4 2009 this would mean that we would discard for the non-topX units 1.98 - 1.31 = 0.67 milliard euros. Because a good quality of our first quarterly estimates is crucial to Statistics Netherlands we decided to use (nearly) all available turnover and impute turnover for the missing VAT declarations Thus in the case of a monthly reporter that has declared already two of the three months, we impute only the turnover of the third month.

We are aware that we could have used simpler methods, e.g. impute at the level of the enterprise and check for incompleteness on a monthly basis. In practice, this may be nearly as complicated because decision rules are needed to derive whether the reported turnover is complete or not.

4.4.3 Missingness due to lack of observations

4.4.3.1 Determine whether turnover is to be expected

When a VAT-unit that is found in the population frame has not responded yet we first have to decide whether we can expect turnover for this unit. To do so, we designed a decision tree (Figure 4.2). The first step is that we verify whether the unit has tax dispensation, using a variable from a tax office client data base. Our experience is that those units have hardly any turnover. Therefore, we impute no turnover in those cases.

The second step is that we check how many days there are after the end of the reporting period. If we make a late estimate at which we are beyond threshold *Tr1* (see Figure 4.2) we do not expect a declaration anymore, so probably the unit has ended but is falsely in the frame. If we make an early estimate, in the third step we look into the available historical declarations. If the last *Tr2* (see Figure 4.2) historical periods we did not have a tax declaration, we assume that also in the current period there will be no turnover. In case there is historical turnover we go to the fourth step. If the unit has been declared inactive from the second half of the reporting period (value *Tr3* in Figure 4.2) we assume that we do not expect turnover anymore. Otherwise, turnover is imputed.

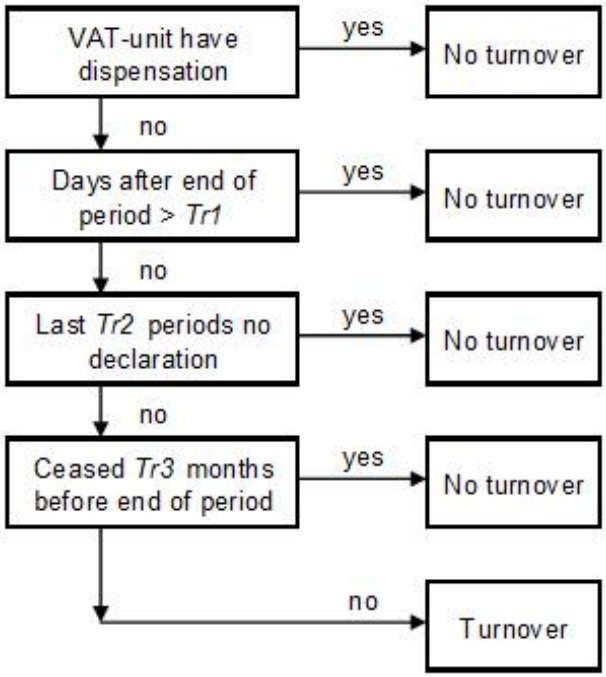


Figure 4.2. Decision tree to determine whether turnover is to be expected or not.

Table 4.10. Threshold values depending on periodicity of tax declaration.

Threshold value	Periodicity of tax declaration		
	monthly	quarterly	Yearly
Tr1	130 days	130 days	130 days
Tr2	5 months	2 quarters	2 years
Tr3	0,5 month*	1,5 month	6 months

* In fact we check whether the unit is active on the first day of the month but not on the last day of the month.

The first settings that we use for parameters *Tr1*, *Tr2* and *Tr3* are given in Table 4.10. The parameters can be determined by computing the difference between the imputed turnover at flash estimate versus the observed turnover at final estimate and likewise for the regular versus the final estimate, and minimising that difference over the publications cells.

4.4.3.2 Introduction to the formula

For those VAT-units where we expect turnover, we will compute an imputation value. Also for non-responding enterprises that received a questionnaire we compute an imputation value. For both unit types we use the same for-

mulas. In the explanation of the notation, below, we use "unit" where unit can be a VAT unit or an enterprise. The unit types will be specified when needed. As explained before, the VAT units can report on a monthly, quarterly or yearly basis. Therefore, the "period", as given below, can refer to a month, a quarter or a year. This will be specified later when needed.

We use the following notation:

O_i^t observed turnover of unit i in period t

$G^{t,s}$ ratio of turnover of period t compared to period s

$G_B^{t,s}$ ratio of turnover in a panel of units in reference group B of period t compared to period s

$R_{B(t,s)}$ set of units in reference group B that responded both in period s and t

In the following sections 4.4.3.3–4.4.3.5 we describe four imputation methods in case of lack observations. The order in which the imputation methods are used depends on the availability of auxiliary variables and on the number of units available to estimate the imputed value. The order of the methods is described in section 4.4.3.6. All the formula given in section 4.4 are computed per stratum h. The size of these strata is discussed in section 4.4.3.6. For simplicity of notation subscript h is omitted.

4.4.3.3 Pattern 1a: Units with historical turnover values

For units that have historical turnover values, the imputed value of unit i, $\hat{O}_i^t$, is computed as:

$$\hat{O}_i^t = \hat{G}_B^{t,s} O_i^s \tag{4.1}$$

where a hat stands for an estimation, s stands for a historical period, and $\hat{G}_B^{t,s}$ stands for the turnover ratio for period t and s in a reference group, with

$$\hat{G}_B^{t,s} = \frac{\sum_{j \in R_{B(t,s)}} O_j^t}{\sum_{j \in R_{B(t,s)}} O_j^s}. \tag{4.2}$$

We distinguish between two methods. We use either the turnover ratio of the current period compared to the corresponding period of a year ago (method A), or the ratio of two subsequent periods (method B). In Table 4.11 we specify the formula for the periodicity of response (monthly, quarterly and yearly).

Responding units in the reference group that declare tax on a monthly basis are used to impute non respondents that declare on a monthly basis. Likewise, responding units in the reference group that declare tax on a quarterly basis are used to impute non respondents that declare on a quarterly basis. For units that report monthly, we impute only turnover for the months that are missing. Quarterly turnover of units that declare on a monthly basis is obtained as the sum of the three corresponding monthly turnovers. For units that report quarterly we impute a quarterly turnover. For units that report yearly, we also impute a quarterly turnover for period *k(y)* which stands for quarter k in current year y, and use a growth rate based on the turnover of *k(y)* compared to the yearly turnover of last year *(y-1)*.

Table 4.11. Specification of method A and B for periodicity of response

Periodicity of response	Duration of historical period s	Duration of actual period t	Specific notation	Method A	Method B
Monthly	month	month	t=m	s=m–12	s=m–1
Quarterly	quarter	quarter	t=k	s=k–4	s=k–1
Yearly	year	quarter	t=$k(y)$		s=y–1

4.4.3.4 Pattern 1b–c. Units without historical turnover

When a unit has no historical turnover we impute an average value. Again we have two methods: C and D. Method C makes use of an auxiliary variable, namely the number of working persons. Turnover of unit i is imputed according to:

$$\hat{O}_i^t = \hat{\bar{Z}}^t \times m_i^t \times WP_i^t \tag{4.3}$$

where WP_i^t stands for the number of working persons of unit i in period t, m_i^t stands for the number of months in period t that unit i is active and $\hat{\bar{Z}}^t$ stands for the estimated average monthly turnover per working person among a reference group of respondents. $\hat{\bar{Z}}^t$ is computed as

$$\hat{\bar{Z}}^t = \frac{\sum_{j \in R^t} O_j^t}{\sum_{j \in R^t} m_j^t \times WP_j^t} \tag{4.4}$$

where R^t stands for the respondents in a reference group for which turnover and number of working persons are available.

Method D can be used when neither historical turnover nor number of working persons is available for unit i with a lacking observation. Turnover is then imputed according to the average of a reference group:

$$\hat{O}_i^t = \hat{\bar{O}}^t = \frac{1}{r^t} \sum_{j \in R^t} O_j^t, \tag{4.5}$$

where r^t stands for the number of respondents in reference group R^t. In Table 4.12 we specify the formula of methods C and D for periodicity of response (monthly, quarterly and yearly).

Table 4.12. Specification of method C and D for periodicity of response

Periodicity of response	Duration of period t	Specific notation
Monthly	month	t=m
Quarterly	quarter	t=k
Yearly	quarter	t=k

For units that declare tax on a yearly basis, method C and D is only used for the flash and regular estimate. Their imputation values are renewed at the final STS estimate when their declared yearly turnover is available. Then the units are imputed according to the method described in section 4.4.5.1.

4.4.3.5 Some exceptions to methods A–D

The above described methods cannot always be applied. We use the following exceptions:

- Units with negative turnover values in either historical period s or actual period t or both are excluded from the reference group when method A or B is applied.
- Units with negative turnover values in the actual period t are excluded from the reference group when method C or D is applied.
- If $\sum_{j \in R_{B(t,s)}} O_j^s = 0$ method A and B cannot be applied.

- When the composition of enterprises change between period s and t, the corresponding VAT units are excluded from the reference group, where t stands for the actual period and s for the corresponding period the previous year (method A) or for previous period (method B–D).
- The units to be imputed are excluded from method A and B when their turnover for historical period s is negative.

4.4.3.6 Order of methods A–D

The reference groups are determined by strata that are defined by the crossing response periodicity × size class (SC) × NACE code. Furthermore we use a minimum size of 20 units within a reference group in order to obtain a reasonable accurate estimate. We are aware that this is only rule a thumb; it might be better to estimate the accuracy of the imputed value and to use a minimum accuracy level.

Preliminary results have shown that growth rates (and means) differ per size class, economic activity and response periodicity. In order to impute as accurately as possible, the starting point is to use a reference stratum that is rather detailed: response periodicity × 1 digit SC code × 5 digit NACE code. If there are not enough units available we use a less detailed reference stratum. Table 4.13 shows the methods crossed by detail of reference stratum. The cell with number 1 represents the most accurate method and the cell with number 64 the least accurate one. Depending on the available data and the number of units per reference stratum, the most accurate method is selected. The order has been filled in based on experience and expert knowledge. Section 4.5 describes a first test to determine whether the order is correct. After testing, the scheme of Table 4.13 has been simplified.

4.4.4 Missingness due to different unit structure

4.4.4.1 Introduction

In the topX entities, tax declarations have a many-to-one relationship with the enterprise group and within the enterprise group the declarations may be related to more than one enterprise. The question is then how to divide the turnover among the enterprises of the enterprise group. For a comparable situation, the German Statistical Office (Gnoss, 2010) uses a linear regression model where log-transformed turnover per enterprise is estimated from NACE code, from number of employees and number of local units. The resulting estimated turnover is summed up to the estimated total of a group of

Table 4.13. Order of methods A–D in relation to the level of detail of the reference stratum

Method/ Reference stratum	NACE 5-digit × SC 1-digit × Periodicity type	NACE 5-digit × SC 1 digit	NACE 5-digit × SC group	NACE 5-digit	NACE 4-digit × SC 1-digit × Periodicity type	NACE 4-digit × SC 1 digit	NACE 4-digit × SC group	NACE 4-digit	NACE 3-digit × SC 1digit × Periodicity type	NACE 3-digit × SC 1digit	NACE 3-digit × SC group	NACE 3-digit	NACE 2-digit × SC 1 digit × Periodicity type	NACE 2-digit × SC 1 digit	NACE 2-digit × SC group	NACE 2-digit
A	1	3	5	13	7	9	11	21	15	17	19	23	43	45	47	49
B	2	4	6	14	8	10	12	22	16	18	20	24	44	46	48	50
C	25	26	27	31	28	29	30	35	32	33	34	36	51	52	53	54
D	37	38	57	61	39	40	58	62	41	42	59	63	55	56	60	64

enterprises. The result is adjusted to the observed VAT turnover at enterprise group level.

Because in the Netherlands mainly for largest topX entities the tax declarations are related to more than one enterprise, SN chose send all topX enterprises a survey to ask for their turnover.

4.4.4.2 Non response in the survey of topX enterprises

Within the topX enterprises, a considerable number of VAT units is related to just one enterprise, see e.g. Table 4.8. We can use those VAT declarations in the case of non response in the survey of topX enterprises. If historical turnover is available for the non responding topX enterprise i, if VAT units have a many-to-one relation to this enterprise i and if the turnover of all those VAT-units is available then the imputed turnover $\hat{O}_i^t$ for the actual period t is given by:

$$\hat{O}_i^t = \hat{G}_i^{t,s} O_i^s \tag{4.6}$$

where $\hat{G}_i^{t,s}$ stands for the estimated turnover ratio of enterprise i. $\hat{G}_i^{t,s}$ is computed from responding VAT-units j as

$$\hat{G}_i^{t,s} = \frac{\sum_{j \in R_{i(t,s)}} O_j^t}{\sum_{j \in R_{i(t,s)}} O_j^s}. \tag{4.7}$$

where $R_{i(t,s)}$ stands for the set of units j that is uniquely related to enterprise i in period t and s. First we try impute with a yearly growth rate and if that is not possible we use a period-to-period growth rate. Note that the restrictions as described in section 4.4.3.5 should also be taken into account, e.g. $\sum_{j \in R_{i(t,s)}} O_j^s > 0$.

When the conditions for formulas (4.6) and (4.7) are not fulfilled, we apply method A–D as described in section 4.4.3 but then directly at the level of the enterprise rather than at the level of the VAT units.

4.4.4.3 VAT declarations related to more than one non-topX enterprise

When VAT declaration i is related to more than one enterprise during period t, the observed turnover O_i^t will be divided among L related enterprises. Below, we describe methods for two situations:

(I) each enterprise ℓ ($\ell = 1, \ldots, L$) is only related to VAT declaration i, and

(II) at least one enterprise ℓ ($\ell = 1, \ldots, L$) is not only related to VAT declaration i but also to one or more other VAT declarations.

Note that for the method of estimation it makes no difference whether the VAT declaration is related to more than one enterprise at one single time point or whether it relates first to enterprise A and then to enterprise B.

Situation (I) Each enterprise ℓ is related to one VAT unit

Denote R_ℓ^t as a reference population of enterprises that contains ℓ and $\hat{\bar{Z}}^t(\ell)$ as the average monthly turnover per working person for that reference population. Likewise to formula (4.4), $\hat{\bar{Z}}^t(\ell)$ is computed as:

$$\hat{\bar{Z}}^t(\ell) = \frac{\sum_{j \in R_\ell^t} O_j^t}{\sum_{j \in R_\ell^t} WP_j^t \times m_j^t} \tag{4.8}$$

where R_ℓ^t stands for the population of j enterprises in period t in a reference stratum that also contains enterprise ℓ, m_j^t stands for the number of months within a quarter that enterprise j is part of the quarterly population frame and WP_j^t is the number of working persons of enterprise j in period t.

Next, we make a preliminary turnover estimate of enterprise ℓ, denoted by $\tilde{O}_\ell^t$, as

$$\tilde{O}_\ell^t = \hat{\bar{Z}}^t(\ell) \times WP_\ell^t \times m_\ell^t \tag{4.9}$$

The final turnover estimate of $\hat{O}_\ell^t$ is obtained by calibrating the preliminary turnover estimates to the observed turnover of VAT declaration i,O_i^t:

$$\hat{O}_\ell^t = O_i^t \times \frac{\tilde{O}_\ell^t}{\sum_\ell \tilde{O}_\ell^t} \tag{4.10}$$

Enterprises with a negative turnover for period t are excluded from R_ℓ^t, but O_i^t is allowed to be negative. The determination of the reference stratum is likewise to section 4.4.3.6. It should hold that $\sum_{j \in R_\ell^t} WP_j^t \times m_j > 0$ and $\sum_\ell \tilde{O}_\ell^t > 0$, otherwise a less detailed stratum must be taken. If the latter is not possible we do not use the VAT declaration but we impute a turnover for the non responding enterprise using the methods of "missingness due to lack of observations".

Situation (II) Each enterprise ℓ is related to more than one VAT unit

If one or more of the enterprises ℓ is not only related to observation VAT unit i, but also to other VAT-units, then we try to use the whole procedure of situation I but in stead of enterprises we estimate the turnover of the related *legal units*. In the end we sum up the turnover of the legal units to the total of the enterprise. When that is also not possible – likewise to situation II – we impute at the enterprise level using the methods of "missingness due to lack of observations".

4.4.5 Missingness due to different meaning of variable

4.4.5.1 Pattern 3a. Observations available for a period longer than the target period

The yearly VAT declarations are divided over the four quarters of the year by making use of the turnover distribution for a reference population, adjusted for the months that units are active during the year.

Denote R^y as the reference population with units i' for which we know the turnover of the whole year y, i.e. four quarters in case of a VAT unit that reports on a quarterly basis and 12 months for a VAT unit that reports on a monthly basis. The quarterly turnover of period k for the reference population, now denoted by $KO^k(R^Y)$, is computed as:

$$KO^k(R^Y) = \sum_{i' \in R^y} KO^k_{i'} \tag{4.11}$$

The fraction of quarterly turnover, $F^k_{KO}(R^Y)$, in quarter k of year y is given by:

$$F^k_{KO}(R^Y) = KO^k(R^Y) \Big/ \sum_{k \in y} KO^k(R^Y) \tag{4.12}$$

The quarterly turnover of period k for yearly VAT declaration of unit i is now estimated as

$$\hat{KO}^k_i = \frac{m^k_i F^k_{KO}(R^Y)}{\sum_{k=1}^{4} m^k_i F^k_{KO}(R^Y)} \times JO^y_i \tag{4.13}$$

where JO^y_i stands for the observed yearly turnover of unit i in year y and m^k_i stands for the number of months in quarter k that unit i is active.

Some special rules apply to the imputation method for pattern 3a. The stratum level containing the reference population is determined according to the scheme for method A as described section 4.4.3.6. Units are excluded from R^y when their quarterly turnover is negative. VAT units that are related to enterprises with a changing VAT unit composition during the year are also excluded from the reference group. The observed JO^y_i of the unit to be imputed is allowed to be smaller than 0. Also, units to be imputed are excluded from this method when they belong to an enterprise with a changing composition of VAT units during the year. When $\sum_{k \in y} KO^k(R^Y) = 0$ or when $\sum_{k=1}^{4} m^k_i F^k_{KO}(R^Y) = 0$ the method cannot be used. In those cases that the method cannot be applied, method A–D of section 4.4.3 is used.

4.4.5.2 Pattern 3b. Observations available for a period that is shifted compared to the target period

Some VAT-units declare tax for a three months period that is shifted compared to a calendar quarter. So far, this concerns less than 1 per cent of

all VAT units that declare on a quarterly basis. Therefore, we use a simple correction. We attribute a shifted three months value to the calendar quarter that overlaps most, in terms of calendar days, with the shifted period.

4.4.5.3 Pattern 3c. Observations with differences in definition

VAT turnover may differ from target turnover. Below we describe solutions for two issues.

Issue (I) VAT declaration patterns

Some VAT-units have remarkable temporal VAT turnover patterns. For example, a unit declares exactly the same turnover during three subsequent quarters followed by a different turnover value in the fourth quarter. For those remarkable patterns we first sum up the turnover of the four quarters to a yearly turnover. Next, we estimate the quarterly turnover as described in section 4.4.5.1. These pattern corrections can only be done after the declarations for the whole year have been received, which corresponds to the final STS estimate.

Issue (II) Definition differences

For some VAT-units VAT turnover deviates from the target definition. We estimate the target turnover of period t of VAT unit i from the VAT data using a linear transformation:

$$\hat{O}_i^t = \hat{a}^y O_i^t(*) + \hat{b}^y/c \tag{4.14}$$

where t can stand for a month, a quarter or a year depending on the response periodicity and c=1 for yearly turnover, c=4 for quarterly turnover and c=12 for a monthly turnover.

We estimate the parameters $\hat{a}^y$ and $\hat{b}^y$ using SBS survey data and VAT data at the level of the *enterprises* for historical year y^* (= y–2). We use only enterprises that are active during the whole year and for which we have response for both the SBS and the VAT; the latter needs to be complete for all underlying VAT units. The parameters in formula (4.14) are estimated by applying a linear regression to enterprises j within a stratum:

$$O_j^{y*}(SBS) = \hat{a}^{y*} O_j^{y*}(VAT) + \hat{b}^{y*} + e_j^{y*}, \tag{4.15}$$

where e_j^{y*} stands for the residual of enterprise j in year $y*$. Parameters are estimated per stratum, where a stratum corresponds approximately with 4

digit NACE level. In fact base strata are chosen from which all output can be made.

The residuals in formula (4.15) are minimized using weighted least squares, where the weights are defined by $w_j^{y*} = 1/\pi_j^{y*} O_j^{y*}(VAT)$. π_j^{y*} stands for the inclusion probability of unit j in year y^* of the SBS survey and is included so the regression represents the population. The component $1/O_j^{y*}(VAT)$ is included because we assume that the variance is proportional to the size of the enterprise. We compute the standard error of the estimated parameters, accounting for the sampling design. When $\hat{a}^{y*}$ is not significantly different (t-distribution) from 1, we use $\hat{a}^{y}$=1 otherwise $\hat{a}^{y} = \hat{a}^{y*}$. Likewise, when $\hat{b}^{y*}$ is not significantly different from 0, we use $\hat{b}^{y}$=0 otherwise $\hat{b}^{y} = \hat{b}^{y*}$.

For NACE codes with a poor correlation between SBS and VAT data, i.e. a correlation coefficient smaller than 0.7 on log transformed data, target turnover cannot be estimated from VAT turnover. For those NACE codes we use sample survey data instead.

4.4.6 Some practical implementation issues

In the current section we mention some practical implementation issues. The first issue is that some of the missingness patterns described in section 4.4.3–4.4.5 can occur simultaneously. For example a unit that declares tax on a quarterly basis can be a non-respondent for the flash estimate. Simultaneously, this VAT unit can be related to two enterprises.

The different patterns of missingness are handled in the following order:

1. The VAT declaration patterns are corrected.

2. The target turnover is estimated from the VAT turnover. After step 1 and 2, the VAT turnover is comparable to the turnover of the survey data: both comply with the target turnover. In any imputation after step 2 that is done at the level of the enterprise, its source (VAT or sample data) is no longer relevant. Harmonisation is done before completion in order to have more enterprises available for the reference populations.

3. Turnover of VAT-units that declare on a yearly basis is divided over the four quarters of the year. Step 3 is only needed for the final STS and SBS release. For the flash and regular quarterly release this step is skipped.

4. Missing observations are imputed.

5. Turnover of step 3 and 4 is divided over two or more enterprises in case the VAT unit is related to more than one enterprise.

Furthermore, there are many implementation issues that concern the treatment of auxiliary information, such as the reporting periodicity of the VAT units, and the actual number of working persons and NACE code of the enterprise. We also deal with cases where auxiliary information is missing and cases with conflicting auxiliary information because it varies from period to period or by source.

4.5 Example of a test of accuracy of imputation methods

4.5.1 Data and methodology of the test

4.5.1.1 General setting and data set

In this section we give an example of an accuracy test of the imputation methods. At Statistics Netherlands it is crucial to have a small difference between early and final quarterly estimates. In line with that, we test imputation methods in the case of "lack of observations" as given in section 4.4.3. We use VAT data and compared imputed values at an early date with observed values at final response. We limited ourselves to non-topX units.

We took VAT data from Q1 2008–Q2 2010, where each quarter was linked to the population frame of the corresponding quarter. We removed non-domestic VAT units, VAT units linked to topX enterprises, VAT units that did not link to enterprises and VAT-units that linked to more than one enterprise. We also removed extreme values: about 500–1000 per quarter. The resulting data file contained 13.5 million records.

We show test results for the imputation of Q1 2010. The tests are done at 2- and 3-digit NACE level within the domain of the STS statistics. At 2-digit NACE level there are five very small strata, namely 06, 12, 36, 97, 99. Imputation results of those small strata were poor compared to the others. Those five strata were excluded from the results shown below.

Note that in the evaluation we have used two restrictions

- we included only those units for which the quarterly turnover is complete at the final estimate.

- we included only those units that could be imputed by methode A–D

4.5.1.2 Indicators

To explain the indicators, we first introduce some notation:

$O^k_{h,early}$ total observed turnover of VAT declarations in quarter k and stratum h at the early estimate;

$\hat{O}^k_{h,early}$ total imputed turnover in quarter k and stratum h at the early estimate;

$O^k_{h,final}$ total observed turnover of VAT declarations in quarter k and stratum h at the final estimate.

Within each stratum, only those units that fulfil the restrictions mentioned in section 4.5.1.1 are included.

We evaluate the imputation results using three (base) indicators. The first base indicator is the relative difference for quarter k between the total turnover per stratum h based on observations and imputations at the early estimate and the total turnover at the final estimate:

$$D^k_h = 100 \left(\frac{O^k_{h,early} + \hat{O}^k_{h,early}}{O^k_{h,final}} - 1 \right) \tag{4.16}$$

The second indicator is its absolute value denoted by $\left|D^k_h\right|$.

The third base indicator is the overall relative difference over all strata, given by

$$D^k = 100 \left(\frac{\sum_{h=1}^{H} (O^k_{h,early} + \hat{O}^k_{h,early})}{\sum_{h=1}^{H} O^k_{h,final}} - 1 \right) \tag{4.17}$$

4.5.1.3 Description of the tests

Test 1

In test 1 we concentrate on method A and B of section 4.4.3. Table 4.13 shows the crossing 'method × reference stratum,' a cell in this crossing will

be referred to as a sub method. Each sub method has been given a number which shows the order that is used in production, as explained before in section 4.4.3.6. In test 1 we test the accuracy of the sub methods 1–24 at a registration date corresponding to roughly 50 percent response. The registration date is the date that the VAT-declaration of a company is registered at the tax office. As indicators we computed the average and the median of $\left|D_h^k\right|$ over the strata h at two and three digit NACE level.

To have a fair comparison between methods, we included only those records that could be imputed by all methods. The minimal size of the reference population was set to 20 units. We included only those strata h that fulfilled the above condition.

Test 2

In test 2 we analysed the accuracy of the imputation in production given the order of the sub methods in Table 4.13 at two registration dates corresponding to roughly 50 and 75 per cent response (see Table 4.14). As indicators we computed (1) D^k, (2) the average, median, 10 and 90 percentile of D_h^k, and (3) the average and median of $\left|D_h^k\right|$. As strata we took the two and three digit NACE level.

Table 4.14. Response rate at two registration dates for Q1 2010

Registration date	Response (%)	
	Quarterly reporter	Monthly reporter, 3^e month
28 April 2010	58	54
30 April 2010	77	75

4.5.2 Test results and first conclusions

Test 1

Table 4.15 shows that indicator values for imputation accuracy of units that report on a monthly basis are much smaller than of those that report quarterly. This is simply because the indicators are computed for quarterly turnover. At the chosen registration date (28-4-2011) most monthly reporters have already declared turnover for the first two months of the quarter and about 50% of the units have reported for the third month.

Table 4.15 clearly shows that the average and median values of $\left|D_h^k\right|$ are larger at 3 digit than at 2 digit NACE level. Table 4.15 also shows that the results for method A, using a yearly growth rate, is more accurate than method B,

Table 4.15. Average and median of $|D_h^k|$ for different imputation methods, Q1 2010 at 50% response.

Method / sub method	NACE 2 digit				NACE 3 digit			
	Quarterly reporter		Monthly reporter		Quarterly reporter		Monthly reporter	
	Avg	P_{50}	*Avg*	P_{50}	*Avg*	P_{50}	*Avg*	P_{50}
A								
1	1.98	0.85	0.42	0.14	2.46	1.04	0.59	0.21
3	1.79	0.93	0.42	0.14	2.35	1.15	0.59	0.21
5	1.62	0.92	0.42	0.12	2.41	1.13	0.55	0.24
7	1.95	0.87	0.40	0.10	2.43	1.01	0.57	0.18
9	1.77	0.76	0.40	0.10	2.33	1.13	0.57	0.18
11	1.57	0.92	0.41	0.13	2.37	1.14	0.54	0.24
13	1.68	0.72	0.40	0.13	2.51	1.11	0.53	0.20
15	1.87	0.90	0.40	0.13	2.45	1.04	0.58	0.20
17	1.79	0.78	0.40	0.13	2.39	1.16	0.58	0.20
19	1.62	0.78	0.40	0.15	2.42	1.13	0.54	0.24
21	1.71	0.78	0.38	0.13	2.50	1.04	0.51	0.18
23	1.79	0.86	0.38	0.12	2.59	1.14	0.51	0.19
B								
2	3.09	0.76	0.36	0.20	2.81	0.89	0.51	0.22
4	3.12	0.86	0.36	0.20	2.93	1.05	0.51	0.22
6	2.80	0.94	0.33	0.19	2.81	1.12	0.44	0.18
8	2.80	0.78	0.32	0.20	2.60	0.87	0.46	0.20
10	2.65	0.79	0.32	0.20	2.59	1.02	0.46	0.20
12	2.56	0.94	0.30	0.17	2.64	1.13	0.42	0.20
14	2.79	1.09	0.27	0.11	2.61	0.97	0.44	0.21
16	2.79	1.02	0.32	0.18	2.65	0.96	0.44	0.19
18	2.65	0.81	0.32	0.18	2.63	1.09	0.44	0.19
20	2.63	1.07	0.28	0.17	2.70	1.16	0.40	0.19
22	2.82	1.12	0.25	0.10	2.61	1.04	0.42	0.18
24	2.89	1.21	0.25	0.09	2.67	1.04	0.42	0.20

using a period-to-period growth rate. Moreover, differences between method A and B within the same sub method are larger than differences among the sub methods within method A and within method B. Finally, we found some patterns in the performance among the sub methods for quarterly reporters:

- sub methods that differentiate among periodicity type perform slightly less than those that do not; A (1 vs. 3; 7 vs. 9; 15 vs. 17); B(8 vs. 10 and 16 vs. 18).
- at 2 digit level, sub methods that use a size class group perform slightly better than those that use a 1-digit size class; A (5 vs. 3; 11 vs. 9; 19 vs.17); B (6 vs. 4; 12 vs.10; 20 vs. 18). At 3 digit level differences between those size class sub methods are usually smaller than at 2 digit level and sometimes opposite.

For monthly reporters differences among sub methods were much smaller and such patterns were not found.

Test 2

At 2- and 3-digit NACE level (Table 4.16) most indicators for imputation accuracy clearly improved from 50 to 75 per cent response. The difference between $P_{10}(D_h^k)$ and $P_{90}(D_h^k)$ is smaller at 75 than at 50 per cent response and also average and median values for $\left|D_h^k\right|$ are smaller. Note that the average value for $\left|D_h^k\right|$ is larger than its median, that means that the distribution of $\left|D_h^k\right|$ across strata h is skewed to the right. The average for D^k at 2-digit NACE level and 50 per cent response is 0.15 per cent, whereas D^k at the total STS domain was found to be *minus* 0.53 per cent. This sign-difference is because when computing the average, all strata have the same weight.

The example presented here illustrates that test results can be used to improve the initially designed approach. The results in Table 4.16 are based on the sub methods order of Table 4.13. We expect that the results in Table 4.16 can be improved by changing the sub methods order. First of all Table 4.15 suggests that, for method A and B, reference strata can start at 3 digit NACE level rather than at 5 digit level. The advantage of starting at 3 digit NACE level is that results are more robust against outliers because the reference strata contain more units. Furthermore, Table 4.13 assumes that, starting at method A for given reference stratum, the second best choice is method B within that same reference stratum. Table 4.15 however indicates that the second best choice is to stay within method A and move to the next (less detailed) reference stratum.

Note that the test results might depend on the method that has been used to remove extreme values. Preferably, imputation methods are tested with data that have been edited in production.

Table 4.16. Imputation accuracy for Q1 2010 at two NACE levels and starting dates.

NACE level	Response	D_h^k				$\left\|D_h^k\right\|$	
		Avg	P_{10}	P_{50}	P_{90}	*Avg*	P_{50}
2 digit	50%	0.15	-4.44	0.49	5.37	3.27	1.71
	75%	0.34	-1.84	0.70	3.40	2.10	1.25
3 digit	50%	1.23	-6.66	0.57	10.77	5.48	2.59
	75%	1.37	-2.22	0.55	5.76	3.26	1.12

4.6 Improving robustness in special situations

4.6.1 Negative turnover values

The use of negative turnover values in the reference group can cause implausible imputation values. Therefore, we excluded them from the reference group as explained in more detail in section 4.4. Furthermore, negative historical values of a unit to be imputed can cause implausible imputation values when this unit is imputed using a turnover ratio of the actual to the historical period (method A or B). Therefore, those units are imputed using an average value based on the actual period (method C or D).

4.6.2 Stable versus unstable unit structure

Implausible imputation values can occur when in the reference group the VAT units that are related to enterprises change during the periods that are used to compute a reference turnover ratio or mean. This is best explained by giving an example.

Table 4.17 shows three VAT-units that belong to the same enterprise. The columns stand for the subsequent months of 2010. VAT unit 2 ended in month 5 and VAT unit 3 is new from month 4 onwards in the population frame. We can see that the enterprise has shifted turnover from VAT unit 1 and 2 to VAT unit 3.

Imagine that we make an early estimate for Q2 2010 and some units that declare on a monthly basis did not yet declare their tax over June. To impute the missing values according to method B, we compute a turnover ratio of June to May for a reference group. Say that unit 3 has responded at the early estimate but units 1 and 2 have not. If we would include unit 3 in the reference group of method B, we would overestimate the turnover ratio because we did not account for the fact that turnover is shifted from VAT unit 1 and 2 to VAT unit 3.

Table 4.17. Monthly turnover of three VAT units related to the same enterprise.

VAT Id	**Monthly turnover (× 1000 euros)**									
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
1	10.0	13.0	12.0	14.0	11.0	2.0	3.0	2.0	3.0	2.0
2	0.8	0.8	1.1	0.4	0.3					
3				0.0	0.0	12.0	13.0	15.0	14.0	12.0

To avoid implausible imputation values, VAT-units that belong to enterprises with a changing VAT unit composition during the period to be imputed, are excluded from the reference groups, see also section 4.4.

Note that implausible imputation values can also occur when we impute at the level of the VAT unit and the VAT unit *to be imputed* belongs to an enterprise with a changing composition of VAT-units. From Table 4.17 follows that if we would use the historical monthly values (month 5) of unit 1 and 2, and unit 3 has already responded that we would overestimate the total turnover. We wish to make these imputations more robust by using the following rules:

- Enterprises where the composition of the underlying VAT units in the historical turnover differs from that in period t are imputed at the level of the enterprise.
- Enterprises where the composition of the underlying VAT units in the historical turnover differs from that in the actual period t are excluded from the reference group.

4.6.3 Late versus ended respondent

Figure 4.2 of section 4.4.3.1 presents rules to decide whether turnover is to be expected in case of lacking observations. For estimates that are made before threshold *Tr1*, we have a deterministic decision scheme that assumes turnover is ether expected or not depending on historical observations (*Tr2*) and the period that the unit is active according to the population frame (*Tr3*). This deterministic approach might lead to a bias in the estimation of turnover. Alternatively to using *Tr2* and *Tr3*, we could impute all units and multiply them by a 'proportion late respondents'. This proportion represents the fraction of VAT units that did not respond at the time of the current release but does respond before *Tr1*. These proportions should be determined for a group of homogeneous units. One could think of NACE code, periodicity of response, size, number of months without historical observations etcetera.

Results in Table 4.5 about the Accommodation and food service activities showed that from the 26628 units that did not respond 25 days after the end of quarter Q4 2009, 26023 units (98 per cent) did respond at the final estimate. The remaining 605 units (2 per cent) were probably ended.

4.6.4 Dealing with frame errors in smallest size classes

4.6.4.1 Description of methodology

In preliminary tests we sometimes found extreme imputation values in the smallest enterprises (SC 0–2, up to two working persons), because there were units in the reference group with an unexpectedly large turnover. There are different reasons for those extreme turnover values. Firstly, some units truly have a very large turnover combined with only 0–2 working persons. This is especially the case with units dealing with "royalties" of artists and in the case of holdings. Both dominate in certain NACE codes. Secondly, the SC of the unit may be wrong due to a missing relation between the enterprise and a legal unit that does have a certain number of working persons. Also, the SC may be wrong because when the number of working persons is unknown a value of 0 working persons is taken during derivation of the GBR.

To give an idea about the frequency of this problem, we counted the number of VAT units with a quarterly turnover larger than 10 million euros in SC 0. Averaged over Q1 2008 – Q2 2010, there were 77 of those extreme VAT units per quarter on a total of 47 thousand VAT-units per quarter for SC 0.

To avoid implausible imputation values in the smaller size classes, we compute an imputation SC that can deviate from the coordinated SC in the frame. The imputation SC is derived as follows.

Firstly, for enterprises in small size classes (SC 0 and 1) we check whether the actual number of working persons of the enterprise corresponds with the coordinated SC. If the actual number working persons corresponds to a much larger size class (SC 6 and larger), the imputation SC is based on the actual number of working persons. The values for the lower size classes (SC0 and 1) and the upper ones (SC 6 and larger) have been taken from the next step.

Secondly, for the remaining enterprises, we compare their historical quarterly turnover with the median value per SC. If the turnover per enterprise is considered to be too large, the imputation SC will be larger than the coordinated SC. This is done as follows. Denote L_ℓ as the set of size classes of the small enterprises for which we compute an imputation SC. Some of those will be assigned a new, larger, imputation SC. The set of these larger imputation size classes are denoted by L_u. Note that subscript ℓ stands for lower and u for upper, with $\ell < u$. Let sc denote an individual size class and $O^{k-1}_{sc,med}$ the median quarterly turnover per SC of period k–1. Now we compute the smallest value of $O^{k-1}_{sc,med}$ withinL_u, and the largest value of $O^{k-1}_{sc,med}$ within L_ℓ. For enterprise j we now check the conditions:

$$O_j^{k-1} > \min_{sc \in L_u} \{O_{sc,med}^{k-1}) \text{ and } O_j^{k-1} > \max_{sc \in L_\ell} \{O_{sc,med}^{k-1}) \tag{4.18}$$

If the conditions in formula 4.18 are not fulfilled, the imputation SC of enterprise j for period k equals the coordinated SC. Otherwise, the imputation SC is the SC for with distance d_j^{k-1} is minimal, with

$$d_j^{k-1} = |\ln(O_j^{k-1}) - \ln(O_{sc,med}^{k-1})| \tag{4.19}$$

and use as imputation SC of enterprise j for period k that SC for which d_j^{k-1} is minimal.

To get an idea about the threshold value that we needed to take for L_ℓ and L_u, we conducted a small test with real data (see appendix B). This resulted in L_ℓ=1 and L_u=6.

4.7 Summing up and topics for further research

We have described a methodology to handle incompleteness due to differences in unit types, variable definitions and periods of observed data compared to target ones. We use a regression method to harmonise differences in definition and use imputation for completion. Our method of mass imputation can only be used to estimate a limited number of (related) core variables. Although we have described the methodology only for net turnover, we have implemented it for four related turnover variables; therefore we have some additional steps. When many variables need to be completed, mass imputation is not suitable because it is hard to impute values at unit level that are plausible for all possible combinations of variables.

We use an imputation method that tries to use all the available observations and impute only the "missing parts". For each missing value we try to make use of 'the best' available auxiliary information. We took that approach to produce early estimates of good quality and to have good results for relatively small domains. The approach is flexible: models can be adapted if accuracy tests show that the quality of the imputations for certain missingness patterns is not good enough. Other NSI's may have different unit types and likewise their missingness patterns may be different. Still, the general approach can also be used by other NSI's.

The method presented in the current paper can be cumbersome to implement, mainly because we use various types of auxiliary information depending on

the pattern of missingness. If an NSI is not interested in using all observed register data, the method can be simplified considerably by always imputing quarterly turnover at the level of the enterprise.

We see some possible refinements of the current approach. Firstly, we could use a weighted combination of a yearly (method A) and a period-to-period (method B) turnover ratio. Tests so far showed little difference in the quality of method A and B, but maybe the results of a composite estimator are more robust for small strata.

A second refinement would be to correct somehow for the difference in the turnover ratio of respondents to non respondents, for example by using time series techniques. Such a correction will only be an improvement when this difference is more or less stable over time.

A third optional refinement is to include the effect of the number of VAT units that is related to the enterprise into the imputation model of the reference group and of the recipient unit. So far, our experience is that the variation in turnover among VAT units that are related to the same enterprise is large. Therefore we do not expect that this will improve the results. Fourthly, rather than using a fixed order of imputation methods $\times$ strata, we could re-compute the order of the methods based on historical data – as part of the production process. Re-computation of the preferable imputation method during the production process is done by Finland (Koskinen, 2007).

In the case study presented, we first link the observed turnover of VAT-units to statistical units (enterprises), then complete turnover at the level of the enterprises and finally add up to obtain total turnover of the stratum population. Alternatively, we might have estimated the stratum totals directly from completing the turnover of VAT units, thus ignoring the relation with the enterprises. Problem with the direct approach is that we need to classify the units according to economic activity. At SN, the VAT units are classified by economic activity at the tax office, referred to as a 'branch code'. Preliminary research showed that this branch code deviates from the coordinated NACE code and is not good enough to determine small NACE strata. However, research may be done to find out whether the branch code is good enough to determine the total for the whole STS domain. For late estimates, when all VAT-units have responded, we could add up turnover to the STS domain. This total could then be used as a restriction to the imputations at micro level. Such a restriction may improve the quality of late estimates.

Chapter 5

Bootstrapping Combined Estimators based on Register and Survey Data

Léander Kuijvenhoven and Sander Scholtus[1]

This paper describes how the bootstrap resampling method may be used to assess the accuracy of estimates based on a combination of data from registers and sample surveys. We consider three different estimators that may be applied in this context. The validity of the proposed bootstrap method is tested in a simulation study with realistic data from the Dutch Educational Attainment File.

5.1 Introduction

In this paper, we consider the situation where estimates are based on data from different sources. In particular, producers of official statistics are increasingly making use of existing registers. There are several reasons for doing this, like reducing costs and reducing the burden on respondents. Also, businesses and individuals are becoming less tolerant of surveys, as is reflected by lower response rates. National Statistical Institutes (NSIs) are therefore seeking ways to put different sources of information together, to increase the ability to produce information with good quality attributes in an efficient way.

A problem, however, is that registers are often primarily used for non-

[1]Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: sshs@cbs.nl. Remark: The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

statistical purposes, and therefore not always ideal from a statistical perspective. In some cases, an additional sample survey is needed to obtain reliable statistical results. The problem of assessing the accuracy of estimates based on a combination of administrative sources and sample surveys has, therefore, become very relevant to NSIs. In this paper we examine a relatively simple way to evaluate the accuracy of an estimate based on combined data, namely by using a form of bootstrap resampling.

The primary objective of this paper is to develop a methodology for assessing the accuracy of particular estimates from combined data. We do not discuss the problem of micro integration itself, e.g. how to construct a statistical database or how to handle inconsistencies between data from different sources. Instead, we assume that a statistical database has already been constructed.

The outline of this paper is as follows. In Section 5.2, the setting at hand is further clarified, and three types of estimators that may be used in this context are introduced. One of these estimators is a classical regression estimator which does not use register data and serves as a benchmark for comparing the other estimators. Section 5.3 introduces the proposed bootstrap method for combined data. Section 5.4 describes a simulation study in which the proposed bootstrap method is applied to realistic data from the Dutch Educational Attainment File. Finally, Section 5.5 closes the paper with a short discussion and some ideas for further research.

5.2 Combining Register and Survey Data

5.2.1 Description of the Situation

For convenience, we describe the case that a target variable is observed in one register and one sample. We denote the register by R and the sample by s. Let $\mathcal{U}$ denote the target population. Figure 5.1 shows the relationship between $\mathcal{U}$, R and s graphically.

Let y_k denote the value of a target variable y for an element $k \in \mathcal{U}$. The objective of the survey is to estimate the total value of y for the target population:

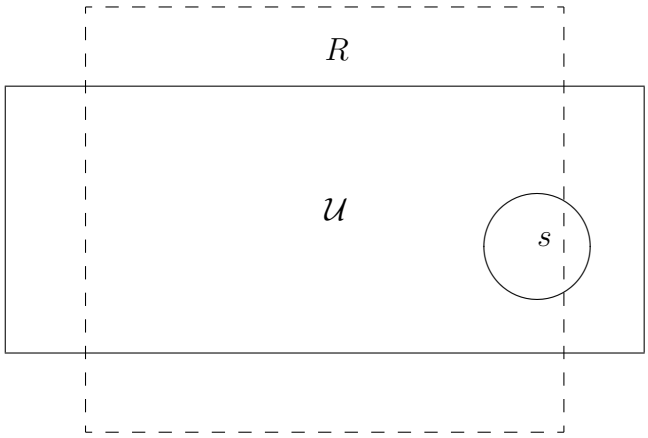
$$\theta_y = \sum_{k \in \mathcal{U}} y_k. \tag{5.1}$$

In this paper, we are interested in the case where the register only covers a selective part of the target population, so that a simple register total would not

be a valid estimate for (5.1). This happens for instance with the Dutch registers on education, which have come into existence only recently, and hence mainly contain information on younger persons. In order to obtain valid estimates, the information from the register has to be combined with information from a sample survey. There are many ways to form such a combined estimator. We consider three types of estimators in the next subsection.

In general, the sample s may be drawn originally from the target population $\mathcal{U}$. Therefore, it can be partitioned into a subsample that overlaps with the register (s_R) and another subsample that does not (s_{NR}). The information from s_{NR} is our only source for inference about $\mathcal{U}_{NR}$, the part of the population that is not covered by the register. Of course, it seems more effective to draw the sample only from $\mathcal{U}_{NR}$. However, this means that the sample cannot be drawn before the register data is obtained, which in turn might conflict with timeliness constraints for the survey. Another reason why the sample may partly overlap with the register, is that an NSI may decide to use existing sample data in the data integration process, instead of introducing a new sample survey. This is, for instance, the case for the Dutch Educational Attainment File, which re-uses data from several cycles of the Dutch Labour Force Survey.

Figure 5.1. The target population (rectangle), the register (dashed rectangle), and the sample (circle)



Throughout this paper, we make the following simplifying assumptions:

- The target population can be partitioned into two disjoint strata: $\mathcal{U}_R = \mathcal{U} \cap R$ and $\mathcal{U}_{NR} = \mathcal{U} \backslash \mathcal{U}_R$, and this stratification is fixed in the sense that it does not depend on an actual realisation of the register R. Note that there is supposed to be no overcoverage in the register, i.e. we

assume that any register records pertaining to elements outside $\mathcal{U}$ can be identified and removed from R.

- The register contains values of a random variable $z_k = y_k + \xi_k e_k$, where ξ_k is a dichotomous variable with $\mathrm{P}(\xi_k = 1) = \lambda_k$ and $\mathrm{P}(\xi_k = 0) = 1 - \lambda_k$, indicating whether an error occurs in the recorded value for element k, and the error e_k is drawn from a distribution with mean μ_k and variance σ_k^2. Moreover, the z_k are drawn independently. Note that λ_k represents the probability that the register value for element k contains an error.

- By contrast, we assume that the target variable is recorded without error in the sample survey. In practice, of course, some measurement errors are bound to occur, but we assume that the effect of these errors is negligable compared to the sampling variance and the effect of errors in the register. This assumption reflects the fact that a statistical institute has direct control over the quality of the sample data, whereas the register data is usually collected by an external party for non-statistical purposes.

A straightforward calculation shows that, under the assumed error model,

$$\mathrm{E}(z_k) = \mathrm{E}_{\xi_k}[\mathrm{E}(z_k \mid \xi_k)] = \mathrm{E}_{\xi_k}(y_k + \xi_k \mu_k) = y_k + \lambda_k \mu_k, \tag{5.2}$$

and

$$\begin{aligned} \mathrm{V}(z_k) &= \mathrm{E}_{\xi_k}[\mathrm{V}(z_k \mid \xi_k)] + \mathrm{V}_{\xi_k}[\mathrm{E}(z_k \mid \xi_k)] \\ &= (1 - \lambda_k)\mathrm{V}(z_k \mid \xi_k = 0) + \lambda_k \mathrm{V}(z_k \mid \xi_k = 1) + \mathrm{V}_{\xi_k}(y_k + \xi_k \mu_k) \\ &= 0 + \lambda_k \sigma_k^2 + \mu_k^2 \lambda_k (1 - \lambda_k) \\ &= \lambda_k [\sigma_k^2 + \mu_k^2 (1 - \lambda_k)]. \end{aligned} \tag{5.3}$$

5.2.2 Three Types of Estimators

5.2.2.1 The Ordinary Regression Estimator

The first estimator that we consider does not use any information from the register, but is solely based on the sample survey. In principle, one could use the direct (or Horvitz-Thompson) estimator $\sum_{k \in s} y_k / \pi_k$, where π_k denotes the inclusion probability of element k in the sample[2]. This is in fact an unbiased estimator of θ_y. It is common practice, however, to apply a linear

[2]Note that π_k denotes the probability of inclusion in the sample, not the register, so it does not automatically follow that $\pi_k = 1$ for all $k \in \mathcal{U}_R$.

regression model to increase the precision of the estimator, and to correct for nonresponse in the original sample. This leads to the well-known regression estimator:

$$\hat{\theta}_{1y} = \sum_{k \in s} w_{1k} y_k, \tag{5.4}$$

with

$$w_{1k} = \frac{1}{\pi_k}\Big[1 + \mathbf{x}'_{1k}\big(\sum_{l \in s} \frac{\mathbf{x}_{1l}\mathbf{x}'_{1l}}{\pi_l}\big)^{-1}\big(\sum_{l \in \mathcal{U}} \mathbf{x}_{1l} - \sum_{l \in s} \frac{\mathbf{x}_{1l}}{\pi_l}\big)\Big]$$

the final weight of element $k \in s$. In this expression, $\mathbf{x}_{1k}$ denotes a vector of auxiliary variables that are observed for all elements of $\mathcal{U}$, corresponding to the chosen linear model. By construction, the final weights satisfy the so-called calibration equations:

$$\sum_{k \in s} w_{1k}\mathbf{x}_{1k} = \sum_{k \in \mathcal{U}} \mathbf{x}_{1k}.$$

The properties of the regression estimator are well-established (Särndal et al., 1992; Knottnerus, 2003). In particular, it is an asymptotically unbiased estimator. For future reference, we note the trivial fact that the variance of $\hat{\theta}_{1y}$ does not depend on the register, i.e. $\mathrm{V}(\hat{\theta}_{1y}) = \mathrm{V}_s(\hat{\theta}_{1y} \mid R)$.

5.2.2.2 An Additive Combined Estimator

Next, we consider the following estimator for θ_y:

$$\hat{\theta}_{2y} = \sum_{k \in \mathcal{U}_R} z_k + \sum_{k \in s_R} (y_k - z_k) + \sum_{k \in s_{NR}} w_{2k} y_k, \tag{5.5}$$

with

$$w_{2k} = \frac{1}{\pi_k}\Big[1 + \mathbf{x}'_{2k}\big(\sum_{l \in s_{NR}} \frac{\mathbf{x}_{2l}\mathbf{x}'_{2l}}{\pi_l}\big)^{-1}\big(\sum_{l \in \mathcal{U}_{NR}} \mathbf{x}_{2l} - \sum_{l \in s_{NR}} \frac{\mathbf{x}_{2l}}{\pi_l}\big)\Big]$$

the final weight of element $k \in s_{NR}$. In this case the regression estimator is used to calibrate s_{NR} on known or previously estimated marginal counts of $\mathcal{U}_{NR}$. This leads to an estimate for the total of y in $\mathcal{U}_{NR}$, which is added to the estimate for $\mathcal{U}_R$. The latter estimate is obtained as the observed total of z_k in $\mathcal{U}_R$, except for the elements of s_R, for which we use y_k because this value is taken to be more accurate.

In appendix C the following expressions for the bias and variance of $\hat{\theta}_{2y}$ are derived, under the assumptions made in Section 2.1:

$$\text{bias}(\hat{\theta}_{2y}) = \text{E}(\hat{\theta}_{2y}) - \theta_y \doteq \sum_{k \in \mathcal{U}_R} (1 - \pi_k)\lambda_k \mu_k \tag{5.6}$$

and

$$\text{V}(\hat{\theta}_{2y}) = \sum_{k \in \mathcal{U}_R} (1 - \pi_k)\lambda_k[\sigma_k^2 + \mu_k^2(1 - \lambda_k)] + \text{V}_s\Big(- \sum_{k \in s_R} \lambda_k \mu_k + \sum_{k \in s_{NR}} w_{2k} y_k \Big). \tag{5.7}$$

It is interesting to examine the properties of $\hat{\theta}_{2y}$ for some special cases of the error model from Section 2.1.

1. If it is assumed that $\mu_k = 0$ for all $k \in \mathcal{U}_R$, then it follows from (5.6) that $\hat{\theta}_{2y}$ is an asymptotically unbiased estimator for θ_y. In this case, it is expected that the errors in the register cancel out in aggregates. The variance of $\hat{\theta}_{2y}$ reduces to

   $$\text{V}(\hat{\theta}_{2y}) = \sum_{k \in \mathcal{U}_R} (1 - \pi_k)\lambda_k \sigma_k^2 + \text{V}_s\Big(\sum_{k \in s_{NR}} w_{2k} y_k \Big).$$

   Note that the first term is the expected error variance in the register (after correction with information from s_R) and the second term is the sampling variance in s_{NR}. However, the assumption that $\mu_k = 0$ may be too optimistic in practice.

2. An important special case occurs when y and z are binary variables, such that $y_k = 1$ if element k belongs to a domain of $\mathcal{U}$, and $y_k = 0$ otherwise. The population total θ_y then measures the size of the domain. Errors in the register correspond to misclassifications of elements of $\mathcal{U}_R$ with respect to the domain. In this case it is natural to assume the following error model for z_k:

   $$z_k = (1 - \xi_k)y_k + \xi_k(1 - y_k),$$

   making $\text{P}(z_k = y_k) = 1 - \lambda_k$ and $\text{P}(z_k = 1 - y_k) = \lambda_k$. In the model of Section 2.1, this leads to $\mu_k = 1 - 2y_k$ and $\sigma_k^2 = 0$. From this and (5.6) and (5.7), it follows that

   $$\text{bias}(\hat{\theta}_{2y}) \doteq \sum_{k \in \mathcal{U}_R} (1 - \pi_k)\lambda_k(1 - 2y_k)$$

and

$$\begin{aligned} \mathrm{V}(\hat{\theta}_{2y}) &= \sum_{k\in\mathcal{U}_R}(1-\pi_k)\lambda_k(1-\lambda_k)(1-2y_k)^2 \\ &\quad +\mathrm{V}_s\Big[-\sum_{k\in s_R}\lambda_k(1-2y_k)+\sum_{k\in s_{NR}}w_{2k}y_k\Big]. \end{aligned}$$

3. Kuijvenhoven and Scholtus (2010) consider the errors in the register to be deterministic: $\sigma_k^2 = 0$ and either $\lambda_k = 0$ or $\lambda_k = 1$, for all $k \in \mathcal{U}_R$. Under this model, the observed register value $z_k = y_k + \lambda_k\mu_k$ with probability one, and λ_k reduces to a simple indicator of error occurrence in z_k. In this case the bias of $\hat{\theta}_{2y}$ can be written as

$$\mathrm{bias}(\hat{\theta}_{2y}) \doteq \sum_{k\in\mathcal{U}_R}(1-\pi_k)(z_k-y_k), \tag{5.8}$$

and the variance of $\hat{\theta}_{2y}$ simplifies to

$$\begin{aligned} \mathrm{V}(\hat{\theta}_{2y}) &= \mathrm{V}_s\Big[\sum_{k\in s_R}(y_k-z_k)+\sum_{k\in s_{NR}}w_{2k}y_k\Big] \\ &= \mathrm{V}_s\Big[\sum_{k\in\mathcal{U}_R}z_k+\sum_{k\in s_R}(y_k-z_k)+\sum_{k\in s_{NR}}w_{2k}y_k\Big] \\ &= \mathrm{V}_s\big(\hat{\theta}_{2y}\,\big|\,R\big), \end{aligned}$$

Hence, under this assumption the variance of $\hat{\theta}_{2y}$ can be evaluated by focusing purely on the sampling variance. In the remainder of this paper, we will in fact assume that the errors in the register satisfy this deterministic model.

5.2.2.3 A Regression-Based Combined Estimator

The last estimator of θ_y that we consider is based on two separate regression models for $\mathcal{U}_R$ and $\mathcal{U}_{NR}$. Specifically:

$$\hat{\theta}_{3y} = \sum_{k\in s_R}w_{3Rk}y_k + \sum_{k\in s_{NR}}w_{3NRk}y_k, \tag{5.9}$$

with

$$w_{3Rk} = \frac{1}{\pi_k}\Big[1 + \big(\sum_{l\in\mathcal{U}_R} z_l - \sum_{l\in s_R}\frac{z_l}{\pi_l}\big)\big(\sum_{l\in s_R}\frac{z_l^2}{\pi_l}\big)^{-1} z_k\Big]$$

the final weight of element $k \in s_R$ and

$$w_{3NRk} = \frac{1}{\pi_k}\Big[1 + \mathbf{x}'_{3NRk}\big(\sum_{l\in s_{NR}}\frac{\mathbf{x}_{3NRl}\mathbf{x}'_{3NRl}}{\pi_l}\big)^{-1}\big(\sum_{l\in\mathcal{U}_{NR}}\mathbf{x}_{3NRl} - \sum_{l\in s_{NR}}\frac{\mathbf{x}_{3NRl}}{\pi_l}\big)\Big]$$

the final weight of element $k \in s_{NR}$.

For the non-registered part of the population, this estimator uses a similar approach to $\hat{\theta}_{2y}$. For $\mathcal{U}_R$, the estimator uses a regression model with the register variable z as predictor variable, since z is likely to be highly correlated with the target variable y. Since the regression estimator is asymptotically unbiased, it holds asymptotically that $\mathrm{E}(\hat{\theta}_{3y}) = \theta_y$. Hence, the advantage of this approach is that it incorporates the information from the register into the estimation process without the risk of introducing a substantial bias. However, this approach is mainly suited for surveys with only one target variable, since otherwise it leads to a different set of regression weights for each target variable. In particular, this type of estimator is not useful if the objective of the survey is to create a general purpose data file for researchers.

It is not difficult to see that, in the extreme case that the register contains no measurement errors at all, i.e. $z_k = y_k$ for all $k \in \mathcal{U}_R$, the two estimators $\hat{\theta}_{2y}$ and $\hat{\theta}_{3y}$ become identical if they use the same weighting model for s_{NR}.

We observe that $\hat{\theta}_{3y}$ can in fact be written as an ordinary regression estimator, which – unlike $\hat{\theta}_{1y}$ – uses auxiliary information from the register. To see this, define a new auxiliary vector $\mathbf{x}_{3k}$ for each $k \in \mathcal{U}$ by:

$$\mathbf{x}_{3k} = \begin{cases} (z_k, \mathbf{0}')' & \text{if } k \in \mathcal{U}_R \\ (0, \mathbf{x}'_{3NRk})' & \text{if } k \in \mathcal{U}_{NR} \end{cases}$$

and define new regression weights

$$w_{3k} = \frac{1}{\pi_k}\Big[1 + \mathbf{x}'_{3k}\big(\sum_{l\in s}\frac{\mathbf{x}_{3l}\mathbf{x}'_{3l}}{\pi_l}\big)^{-1}\big(\sum_{l\in\mathcal{U}}\mathbf{x}_{3l} - \sum_{l\in s}\frac{\mathbf{x}_{3l}}{\pi_l}\big)\Big].$$

Then it is easily derived that $w_{3k} = w_{3Rk}$ for all $k \in s_R$ and $w_{3k} = w_{3NRk}$ for all $k \in s_{NR}$. Therefore it holds that $\hat{\theta}_{3y} = \sum_{k\in s} w_{3k} y_k$.

Finally, we remark that under the deterministic error model from Section 5.2.2.2, it clearly holds that $\mathrm{V}(\hat{\theta}_{3y}) = \mathrm{V}_s(\hat{\theta}_{3y} \mid R)$.

5.3 A Bootstrap Method for Combined Data

5.3.1 Introduction to the Bootstrap

Loosely speaking, the bootstrap idea is to mimic the process that generated the originally observed data, by estimating the underlying distribution from the sample and then resampling from this estimated distribution. In some special cases the bootstrap can be performed analytically, but usually one resorts to Monte Carlo approximation, by generating a large number of bootstrap replicates of the target estimate. These replicates are obtained by taking the algorithm that produces the original estimate when applied to the original sample, and applying it to resamples taken from the estimated distribution. We refer to Efron and Tibshirani (1993) for an introduction to the classical bootstrap.

An important problem with the classical bootstrap arises when it is applied to finite population sampling, namely how to mimic the effect of sampling without replacement. In order to obtain a valid measure of the variance of an estimate, it is crucial to capture the effect of the sampling design. In particular, sampling without replacement leads to a smaller variance than sampling with replacement.

There are various methods suggested in the literature to adapt the classical bootstrap to finite population sampling, including the with-replacement bootstrap (McCarthy and Snowden, 1985), the rescaled bootstrap (Rao and Wu, 1988), the mirror-match bootstrap (Sitter, 1992b) and the without-replacement bootstrap (Gross, 1980; Bickel and Freedman, 1984; Chao and Lo, 1985; Sitter, 1992a). A summary of these methods can be found in Shao and Tu (1995). However, these methods tend to be difficult to apply in practice. Antal and Tillé (2011) describe yet another bootstrap method for finite population sampling.

A newer form of the without-replacement bootstrap has been suggested by Booth et al. (1994), Canty and Davison (1999) and Chauvet (2007). In the next section we describe a variant of the latter method and apply it to the case of combined register and survey data. In line with the deterministic error model from Section 5.2.2.2, we treat the register data as fixed in this bootstrap method.

5.3.2 The Proposed Bootstrap Method

The approach taken by Booth et al. (1994) entails generating pseudo-populations. A pseudo-population is an estimated version of the target population, obtained by taking d_k copies of each element from the original sample, where $d_k = 1/\pi_k$ is the inclusion weight. Bootstrap resamples are drawn by applying the original sampling design to the pseudo-population, and a replicate of the original estimator is calculated from each bootstrap resample. Finally, estimates of the accuracy of the original estimator, such as its variance or confidence intervals, are obtained from the distribution of these replicates, analogous to the classical bootstrap method.

In general d_k need not be an integer, which makes it necessary to round the inclusion weights. Writing $d_k = \lfloor d_k \rfloor + \varphi_k$ (with $\varphi_k \in [0, 1)$), a stochastic form of rounding is used that rounds d_k down to $\lfloor d_k \rfloor$ with probability $1 - \varphi_k$, and up to $\lfloor d_k \rfloor + 1$ with probability φ_k[3]. In order to eliminate the effect of the stochastic rounding on the outcome of the bootstrap method, multiple pseudo-populations can be formed, each based on a different rounding of the inclusion weights.

The diagram in Figure 5.2 summarises the bootstrap method. In this description, B denotes the number of constructed pseudo-populations and C the number of replicates computed from each pseudo-population. The total number of bootstrap replicates equals $B \times C$. Suitable values of B and C are discussed in Section 4.

Following results of Chauvet (2007), a single pseudo-population could also be used as an approximation of the above-mentioned approach. Using a single pseudo-population is, as one would expect, less computer-intensive and faster than using multiple pseudo-populations. The bootstrap method with a single pseudo-population is obtained as a special case of the algorithm in Figure 5.2 with $B = 1$, so that Steps 1 to 3 are only run once. Note that compared to the multiple pseudo-population approach, a higher value of C is now needed to achieve convergence of the Monte Carlo approximation. In the simulation study in Section 4, both the multiple and single pseudo-population approach are investigated.

In the above algorithm we have not defined which estimator is used specif-

[3]This stochastic rounding can be executed in different ways. Kuijvenhoven and Scholtus (2010) apply Fellegi's method for consistent rounding directly to the inclusion weights. Booth et al. (1994) and Chauvet (2007) round the weights implicitly, by taking $\lfloor d_k \rfloor$ copies of each element from the original sample and then drawing an additional subsample from the original sample using the drawing probabilities φ_k.

Figure 5.2. A bootstrap algorithm for finite population sampling.

Step 1 Writing $d_k = \lfloor d_k \rfloor + \varphi_k$, define a random inflation weight $\delta_k = \lfloor d_k \rfloor$ with probability $1 - \varphi_k$ and $\delta_k = \lfloor d_k \rfloor + 1$ with probability φ_k. Generate a pseudo-population $\hat{\mathcal{U}}$ by taking δ_k copies of each element k from the original sample s.

Step 2 Draw a sample s^* from $\hat{\mathcal{U}}$ with the original sample design. That is, for $j \in \hat{\mathcal{U}}$ there is an inclusion probability $\pi_j^* \propto \pi_k$, if j is a copy of $k \in s$, where the π_j^* are scaled so that $\sum_{j \in \hat{\mathcal{U}}} \pi_j^*$ equals the original sample size. For each bootstrap resample, compute the replicate $\hat{\theta}^* = t(s^*, R)$, where $t(.)$ denotes the algorithm such that $\hat{\theta} = t(s, R)$.

Step 3 Step 2 is repeated C times to obtain replicates $\hat{\theta}_1^*, \ldots, \hat{\theta}_C^*$. From these replicates, compute:

$$v_{boot} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\hat{\theta}_c^* - \overline{\hat{\theta}^*}\right)^2$$

$$\overline{\hat{\theta}^*} = \frac{1}{C} \sum_{c=1}^{C} \hat{\theta}_c^*$$

Step 4 Steps 1 to 3 are repeated B times to obtain $v_{boot}^1, \cdots, v_{boot}^B$. The estimated variance of the original estimator is

$$v_{boot}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} v_{boot}^b.$$

ically. In fact, a different choice of $t(.)$ is used in Step 2 of the algorithm, depending on the estimator. For the estimators from Section 5.2.2, we define the following expressions for the bootstrap replicate $\hat{\theta}^* = t(s^*, R)$:

$$
\begin{aligned}
t_1(s^*, R) &= \sum_{j \in s^*} w^*_{1j} y^*_j \\
t_2(s^*, R) &= \sum_{k \in \mathcal{U}_R} z_k + \sum_{j \in s^*_R} (y^*_j - z^*_j) + \sum_{j \in s^*_{NR}} w^*_{2j} y^*_j \\
t_3(s^*, R) &= \sum_{j \in s^*_R} w^*_{3Rj} y^*_j + \sum_{j \in s^*_{NR}} w^*_{3NRj} y^*_j
\end{aligned}
$$

In these expressions the following notation is used: s^*_R and s^*_{NR} denote the parts of the resample that consist of copies of elements from s_R and s_{NR}, respectively; y^*_j and z^*_j are by definition equal to y_k and z_k if j is a copy of k; $w^*_{.j}$ denotes a regression weight obtained from the bootstrap resample by applying the same regression model that led to $w_{.k}$ in the original sample. Thus for each bootstrap resample a new set of regression weights is obtained. In this manner the effect on the variance of the estimator due to weighting is taken into account.

Due to nonresponse only a part of the original sample is usually observed in practice. Note that with nonresponse present also nonrespondents are duplicated in the pseudo-population. Therefore, nonresponse will also occur in the bootstrap resamples, namely when copies of original nonrespondents are drawn. Canty and Davison (1999) argue that the bootstrap is valid, provided that the same weighting model that was used to correct for nonresponse in the original sample is also applied to each bootstrap resample, under the assumption that the weighting model indeed explains nonresponse behaviour. Through this approach, each weighted bootstrap resample will correctly represent the original population. Shao and Sitter (1996) use a similar approach, but they impute data for nonrespondents instead of weighting.

5.4 Simulation Study

In this section, we assess the correctness of the bootstrap method from Section 5.3.2 in a simulation study. For this simulation, we used a small subset of the Dutch Educational Attainment File (EAF) as our target population. The EAF contains information on the highest attained education level of persons living in the Netherlands. Moreover, the EAF can be linked to other files containing background variables for these persons. The information on educational attainment is obtained from the Dutch educational registrations and from the Labour Force Survey. For persons that are present in more

than one source, the scores for education levels in different sources are compared and one of the values is chosen (usually the highest one). This process is called harmonisation. We refer to Linder and Van Roon (2011) for more details on the EAF.

As our target population we selected a subset of 49647 persons aged over 14 years[4] from the EAF. For the purpose of this simulation study, the file containing the records of these 49647 persons was considered to be a complete enumeration of the target population. In this target population, register information was available for 8904 persons, so the size of $\mathcal{U}_R$ was 8904. The remaining 40743 persons, for which no register information was available, constituted the subpopulation $\mathcal{U}_{NR}$. It is noted that for persons in $\mathcal{U}_R$, the education level from the register may differ from the final, harmonised education level. Using the notation from Section 5.2, the true value y_k corresponds to the harmonised education level and the register value z_k corresponds to the unharmonised education level from the register. For the purpose of the simulation study, differences between these two values were considered to be caused by errors in the register.

Next, we drew samples from our target population and used these samples, together with $\mathcal{U}_R$, to estimate certain parameters of the target population. Since the true values of these target parameters were also known in this study, the theoretical accuracy of the survey estimators could be measured directly. We also computed estimates of accuracy using the bootstrap method, and compared these with the theoretical accuracy. In order to comply with the assumption from Section 5.2.1 that measurement errors only occur in the register, for the purpose of this simulation study, we always took the harmonised education levels as observed data in our samples.

A stratified simple random sampling design was used to draw the samples, where the stratification was by *Sex* (values: Male and Female) and *Age* (values: Young, Middle, and Old). The total sample size equaled 3615. The sampling fractions were: 30% for the two strata with *Age* = Young, 6% for the two strata with *Age* = Middle, and 3% for the two strata with *Age* = Old. These sampling fractions were chosen such that the corresponding d_k had large non-integer parts: the inclusion weights were 3 1/3, 16 2/3, and 33 1/3 for young persons, middle-aged persons, and old persons, respectively. Thus, we expected to see a relatively large effect of the stochastic rounding on the outcome of the bootstrap method.

[4]The education levels are sometimes deductively imputed for persons younger than 15 years, so these cannot be considered as typical register or sample data.

The education levels come from a hierarchic classification. The highest level of the classification consists of five codes, ranging from primary education (code 1) to education leading to a master's degree or higher (code 5). In this simulation study, we estimated the number of persons with educational attainment code 3 (which corresponds to the second stage of secondary education) in each stratum. These parameters can be seen as population totals of suitably defined binary variables. Table 5.1 displays the actual values in the target population.

We used the three estimators (5.4), (5.5), and (5.9) from Section 5.2 to estimate the target parameters. The regression weights w_{1k}, w_{2k}, and w_{3NRk} were obtained using the following linear model:

$$Region(5) \times Age(3) + Region(5) \times Sex(2) \times Marital\ Status(3),$$

where the number in brackets denotes the number of classes for each auxiliary variable.

The true statistical properties of the three estimators were approximated by drawing 20000 samples from the target population. Table 5.1 displays the approximate standard errors of the three estimators based on these 20000 realisations. Since estimator 2 is known to be potentially biased, Table 5.1 also shows the approximate relative bias of this estimator based on 20000 realisations. As the target population was completely known, the theoretical bias of estimator 2 could also be calculated directly from expression (5.8), which led to similar values.

It can be seen that the register count overestimates the number of persons with educational attainment code 3. This bias is caused by the fact that for some persons with educational attainment code 4, not all forms of education attained by these persons have been properly registered, so that the reported educational attainment code 3 in the register is too low by mistake. Of course, this effect could be neutralised by the presence of persons with an actual educational attainment code 3 who are registered with a lower educational attainment code, but apparently the latter type of error occurs less often, so that the net result is a positive bias. The bias is much larger for the strata of young persons than for the other strata, because, as mentioned in Section 5.2.1, older persons are underrepresented in the register. In fact, hardly any register information is available from the strata of old persons, so that the three estimators are actually almost identical for these strata (which explains why the standard errors are more or less equal).

Table 5.1. Target parameters, standard errors, and relative bias based on 20000 simulations. Abreviations: YM = Young Males, MM = Middle-Aged Males, OM = Old Males, YF = Young Females, MF = Middle-Aged Females, OF = Old Females.

	YM	MM	OM	YF	MF	OF
target parameters						
	1178	4459	4423	1164	5386	3880
standard errors (with full response)						
estimator 1	49	203	298	48	208	293
estimator 2	36	186	297	36	190	291
estimator 3	40	190	297	38	193	291
relative bias (with full response)						
estimator 2	+12%	+3%	+0%	+7%	+2%	+0%
standard errors (with nonresponse)						
estimator 1	58	241	349	57	248	341
estimator 2	42	223	347	43	229	339
estimator 3	47	227	347	45	231	339
relative bias (with nonresponse)						
estimator 2	+13%	+3%	−0%	+8%	+1%	−1%

For the above-mentioned results it was assumed that all persons responded when sampled. It is more realistic to assume that some nonresponse occurs in the sample. To keep matters simple, we adopted the so-called fixed response model (Bethlehem et al., 2011), whereby each person in the target population either always responds or never responds when sampled. Response indicators were randomly assigned to the persons in the target population, in such a way that the weighting model in *Region*, *Age*, *Sex*, and *Marital Status* explained most of the nonresponse behaviour. The last two sections of Table 5.1 show the approximate standard errors of the three estimators and the relative bias of estimator 2 with nonresponse, again based on 20000 realisations.

The approximate standard errors in Table 5.1 serve as a benchmark for the bootstrap results to be discussed below.

In order to apply the bootstrap method proposed in Section 5.3.2, suitable values for B and C had to be chosen. Chauvet (2007) reported results based

Table 5.2. Standard errors from the bootstrap method with multiple pseudo-populations (average of 20 realisations). In brackets the relative standard deviation of the 20 realised values.

	YM	MM	OM	YF	MF	OF
standard errors (with full response)						
estimator 1	50	203	299	49	212	289
	(4%)	(2%)	(2%)	(3%)	(2%)	(3%)
estimator 2	37	188	297	38	195	289
	(7%)	(2%)	(2%)	(6%)	(2%)	(3%)
estimator 3	40	192	297	40	198	289
	(6%)	(2%)	(2%)	(5%)	(2%)	(3%)
standard errors (with nonresponse)						
estimator 1	60	244	350	59	250	339
	(9%)	(2%)	(3%)	(4%)	(2%)	(4%)
estimator 2	45	230	347	47	233	337
	(9%)	(3%)	(3%)	(9%)	(3%)	(3%)
estimator 3	49	233	347	49	237	337
	(8%)	(2%)	(3%)	(9%)	(3%)	(3%)

on $B = 100$ and $C = 30$ for the multiple pseudo-population approach, and $C = 1000$ for the single pseudo-population approach. In contrast to Chauvet (2007), we considered only variance estimates and not the estimation of bootstrap confidence intervals in this study. It is acknowledged in the bootstrap literature that, compared to the estimation of confidence intervals, a smaller number of replicates suffices for variance estimation. Therefore, to limit the amount of computational work, we chose $B = 50$ and $C = 30$ for the multiple pseudo-population approach in this simulation study. For the single pseudo-population approach, we chose $C = 1000$.

Table 5.2 reports the estimated standard errors for the three estimators obtained from the bootstrap method with multiple pseudo-populations, both without and with nonresponse. To trace the sampling variability of the bootstrap estimates, these results were based on 20 realisations of the bootstrap method, and Table 5.2 shows both the mean and the relative standard deviation of 20 bootstrap estimates. Similar results for the bootstrap method with a single pseudo-population are reported in Table 5.3.

These results do not exhibit large differences between the multiple and single

pseudo-population approaches. The estimated standard errors are in both cases close to the approximate true values from Table 5.1, with a tendency to slightly overestimate the standard errors in most strata. The relative standard deviations of the bootstrap estimates are small and the two approaches perform about equally well in this respect also. The similar performance of the multiple and single pseudo-population approaches seen here is in line with results reported by Chauvet (2007) in a simulation study involving an artificial population of normally distributed data.

Table 5.3. Standard errors from the bootstrap method with a single pseudo-population (average of 20 realisations). In brackets the relative standard deviation of the 20 realised values.

	YM	MM	OM	YF	MF	OF
standard errors (with full response)						
estimator 1	50	205	297	50	209	290
	(5%)	(2%)	(3%)	(4%)	(2%)	(3%)
estimator 2	37	189	295	38	192	289
	(6%)	(3%)	(3%)	(6%)	(2%)	(3%)
estimator 3	41	193	295	41	195	290
	(6%)	(3%)	(3%)	(5%)	(2%)	(3%)
standard errors (with nonresponse)						
estimator 1	59	246	349	59	248	337
	(7%)	(3%)	(3%)	(4%)	(3%)	(4%)
estimator 2	45	230	346	47	232	335
	(8%)	(3%)	(3%)	(8%)	(3%)	(4%)
estimator 3	49	235	346	49	235	335
	(7%)	(3%)	(3%)	(7%)	(3%)	(4%)

For the stratified simple random sampling design used in this simulation study, a practical alternative method for estimating the variance of a regression estimator is to apply Taylor linearisation (Särndal et al., 1992; Knottnerus, 2003). For estimator 1, the following variance estimator is readily found in the literature:

$$\hat{\mathrm{V}}(\hat{\theta}_{1y}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s^2_{\hat{\varepsilon}_{1h}}}{n_h},$$

where H denotes the number of strata (in this case: $H = 6$), N_h and n_h denote the population and sample size in stratum h, and $s^2_{\hat{\varepsilon}_{1h}}$ is the sample

variance of the residuals of the fitted regression model. A similar expression is found for estimator 3, since we already noted that this estimator can be written in the same form as estimator 1. For estimator 2, we used the following variance estimator:

$$\hat{\mathrm{V}}(\hat{\theta}_{2y}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{u_{1h}}^2 + s_{u_{2h}}^2 - 2s_{u_{1h}u_{2h}}}{n_h},$$

with

$$u_{1hk} = \begin{cases} (y_{hk} - z_{hk})n_h/N_h & \text{if } k \in s_R \\ 0 & \text{if } k \in s_{NR} \end{cases}$$

and

$$u_{2hk} = \begin{cases} 0 & \text{if } k \in s_R \\ \hat{\varepsilon}_{2hk} & \text{if } k \in s_{NR} \end{cases}$$

Here, $s_{u_{1h}u_{2h}}$ denotes the sample covariance of u_1 and u_2 in stratum h. This formula is obtained from the derivation given in appendix C below expression (C.3), for the particular case of stratified simple random sampling and the deterministic error model. In addition, the population (co)variances have been estimated by their sample equivalents.

Table 5.4 reports the results of 20 realisations of the linearisation method. In contrast to the bootstrap method, it is seen that the linearisation method tends to underestimate the standard errors, and this effect becomes more pronounced in the situation with nonresponse. On the other hand, the linearisation method appears to be less sensitive to sampling variability, since the relative standard deviations of the 20 realisations are mostly smaller than for the bootstrap method.

To conclude, we give some information on the practical execution of the above simulation study. Most of the computational work for the bootstrap method was done in Blaise, a survey processing system developed at Statistics Netherlands. The bootstrap method was implemented as a series of so-called Manipula setups in Blaise, and the Blaise weighting tool Bascula was used to compute the regression weights. Finally, the statistical software R was used to compile and analyse the results of the simulation study. The estimated standard errors for the linearisation method were also calculated in R.

Table 5.4. Standard errors from the linearisation method (average of 20 realisations). In brackets the relative standard deviation of the 20 realised values.

	YM	MM	OM	YF	MF	OF
standard errors (with full response)						
estimator 1	48	200	294	47	206	284
	(2%)	(1%)	(1%)	(2%)	(< 1%)	(2%)
estimator 2	33	184	292	34	187	282
	(2%)	(1%)	(1%)	(2%)	(1%)	(2%)
estimator 3	37	187	292	37	190	282
	(2%)	(1%)	(1%)	(2%)	(1%)	(2%)
standard errors (with nonresponse)						
estimator 1	56	229	343	56	235	325
	(2%)	(1%)	(2%)	(2%)	(1%)	(2%)
estimator 2	39	211	340	40	213	323
	(2%)	(1%)	(2%)	(2%)	(1%)	(3%)
estimator 3	43	214	339	43	216	322
	(2%)	(1%)	(2%)	(2%)	(1%)	(2%)

5.5 Discussion

In this paper, we have described different estimators based on a combination of register data and sample data, and we have introduced a bootstrap method for assessing the variance of these estimators. Moreover, the performance of the bootstrap method was examined in a simulation study, using realistic data from the Dutch Educational Attainment File. The results of this simulation study show that the bootstrap provides valid variance estimates for estimators based on combined data, and that the quality of the bootstrap estimates compares favourably to an alternative method based on linearisation. It also appears from the study that the bootstrap method with a single pseudo-population is not outperformed by the multiple pseudo-population approach, although the latter has a more sound theoretical basis (Chauvet, 2007). The single pseudo-population approach is less complex than the multiple pseudo-population approach and, in principle, requires less computational work. However, in practice both approaches require the computation of a similar number of replicates; in our simulation study, the total number of replicates BC equals 1500 for the multiple pseudo-population approach and 1000 for the single pseudo-population approach.

Given a combination of register data and sample survey data, there are of course many different estimators that one could consider. In this paper we have only treated three such estimators. Another interesting estimator, suggested by De Heij (2011), is the following:

$$\hat{\theta}_{4y} = \sum_{k \in \mathcal{U}_R} z_k + \sum_{k \in s_R} \frac{y_k - z_k}{\pi_k} + \sum_{k \in s_{NR}} w_{4k} y_k,$$

with w_{4k} a regression weight, defined analogously to w_{2k}. Like our additive estimator $\hat{\theta}_{2y}$, this estimator uses a regression estimator for $\mathcal{U}_{NR}$ and an adjusted register total for $\mathcal{U}_R$, where the adjustment is based on information from s_R. In $\hat{\theta}_{2y}$ the adjustment term only corrected the observed individual errors in the overlap. Consequently, as was confirmed in the simulation study, this estimator is stable but it may have a large bias. In $\hat{\theta}_{4y}$, the adjustment term is based on a Horvitz-Thompson estimate of the total error in the register. This approach has the advantage that it leads to an asymptotically unbiased estimator, unlike $\hat{\theta}_{2y}$. On the other hand, the variance of the adjustment term – and hence of the estimator as a whole – might be large. It would be interesting for a future study to compare the performance of $\hat{\theta}_{4y}$ and the other estimators in practical situations.

The bootstrap method described here only considers the variance due to sampling and treats the observed register data as fixed. In Section 5.2.1 we considered a general measurement error model for register values, which includes the possibility of stochastic errors in the register. From a theoretical point of view, it might be an interesting topic for future research to extend the bootstrap method so that it can also be used when the errors in the register are of a stochastic nature. However, a practical application of this theory would require accurate estimates of the model parameters λ_k, μ_k, and σ_k^2, and these might be difficult to obtain if s_R is our only source of information.

Another assumption made in this paper is that the target variable is observed without error in the sample survey, or, if errors do occur, that the effect of these errors on the estimates is negligable compared to the sampling variance. It may be of interest to relax this assumption and to also assume a model for measurement errors in the sample survey. Note that this implies that more complex estimators are needed, because we can no longer simply use the sample data to correct errors in the register data.

Chapter 6

Models and algorithms for micro-integration

Jeroen Pannekoek[1]

6.1 Introduction

Many statistical institutes aim to increase the use of administrative sources for producing there statistical outputs. The advantages of the use of administrative sources have often been spelled out: the costs involved with using administrative data is much less than the costs of conducting a survey, the amount of data of administrative sources is usually much larger than what could reasonably be obtained by a survey and the re-use of already existing data decreases the response burden. However, in many cases the administrative sources alone do not contain all the information that is needed to produce the detailed statistics that national statistical offices are required to produce. For example, administrative sources can be sufficient to produce the short term business statistics with only a few variables (mainly turnover) but for the yearly structural business statistics with close to a hundred variables, much more information is needed then can be obtained from administrative sources. For such more detailed statistics additional surveys are conducted.

In this paper we consider business statistics based on different sources. A few main variables can be obtained from reliable administrative sources and ad-

[1]Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: jpnk@cbs.nl. Remark: The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

ditional more detailed variables are obtained by surveys that can be linked to the administrative data. For business statistics, there are many logical relations between the variables (also known as edit-rules or edit-constraints). The linked records do not necessarily respect the edit rules and a micro-integration step is necessary to integrate the different pieces of information, the data sources and the edit-constraints, to arrive at consistent integrated micro-data. We will investigate ways to reconcile the possibly conflicting information and emphasize the assumptions underlying these methods.

In section 6.2 of this paper the problem of adjusting micro-data such that a consistent integrated record results is illustrated by an example. In section 6.3 the different adjustment models are described and these methods are illustrated by a numerical example in section 6.4 In section 6.6 the choice of variables and the addition of edit-rules to better preserve the structure of the data is briefly discussed. Section 6.7 extends the methods to adjustment to multiple sources and inclusion of soft edit constraints. An algorithm that can be used to implement the adjustment procedures of this paper is described in Appendix D.

6.2 The adjustment problem for micro-data

To illustrate the adjustment problem at micro-level, we consider the following situation that arises in business statistics. There is information on some key variables available from reliable administrative data. Let these variables be the total turnover *(Turnover)*, the number of employees *(Employees)* and the *Total costs.* These variables are used to compile the short term economic statistics (STS) and are published quarterly as well as yearly. The yearly structural business statistics (SBS), requires much more detail and this more detailed information is not available from registers. Therefore, a sample survey is conducted to obtain the additional details. The situation then arises that for the key variables, two sources are available for each respondent: the register value and the survey value. As an example, consider the following business record with eight variables. For three variables, *Employees, Turnover* and *Total Costs,* both register and survey values are available and for the other variables only survey values are obtained.

Besides the data from the survey and the register(s), other knowledge on the values in the record is available in the form of logical relations between variables and the fact that all values, except *Profit*, should be non-negative. The logical relations, also called edit-rules, for this record can be formulated

Table 6.1. Example Business record with data from two sources

Variable	*Name*	*Sample Value*	*Register Value*
x_1	Profit	330	
x_2	Employees (Number of employees)	20	25
x_3	Turnover main (Turnover main activity)	1000	
x_4	Turnover other (Turnover other activities)	30	
x_5	Turnover (Total turnover)	1030	950
x_6	Wages (Costs of wages and salaries)	200	
x_7	Other costs	500	
x_8	Total costs	700	800

as:

a1: $x_1 - x_5 + x_8 = 0$ (*Profit = Turnover – Total Costs*)

a2: $x_5 - x_3 - x_4 = 0$ (*Turnover = Turnover main + Turnover other*)

a3: $x_8 - x_6 - x_7 = 0$ (*Total Costs = Wages + Other costs*)

These restrictions can be formulated in matrix notation as

$$\mathbf{Ax} = \mathbf{0}$$

with the rows of $\mathbf{A}$ corresponding to the restrictions a1 – a3, so $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 \end{pmatrix}$$

Now, we need to combine the different pieces of information, survey values, register values and edit-rules, to obtain a record such that

1. the record contains the register values for variables for which these are available and
2. the record satisfies the edit constraints.

To obtain such a record, we start by replacing the survey values with the register values when these are available. Let the values of the resulting record be denoted by the vector $\mathbf{x}_0$. This record will, in general, not satisfy the second requirement.

One possible strategy to solve the consistency problem is to adjust the survey values, as little as possible, such that the edit-rules are satisfied. If the resulting adjusted record is denoted by $\tilde{\mathbf{x}}$, this adjustment problem can be formulated as:

$$\begin{array}{l} \tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}_0) \\ s.t. \quad \mathbf{A}\tilde{\mathbf{x}} = 0 \end{array},$$

with $D(\mathbf{x}, \mathbf{x}_0)$ a function measuring the distance or deviance between $\mathbf{x}$ and $\mathbf{x}_0$. In the next section we will consider different functions D for the adjustment problem. In addition to the equality constraints, we also often have inequality constraints, the simplest of which is the non-negativity of most economic variables. Other inequality constraints arise, for instance, when it is known that *Wages* should not be less than a certain factor $f_{\min}$ (the minimum wage) times *Employees*. To also include linear inequality constraints the adjustment problem can be extended as

$$\begin{array}{l} \tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}_0). \\ s.t. \quad \mathbf{A}_1\tilde{x} = 0 \ \text{ and } \ \mathbf{A}_2\tilde{\mathbf{x}} \leq 0 \end{array} \tag{6.1}$$

For ease of exposition we will write the constraints in (6.1) more compactly as $\mathbf{A}\tilde{\mathbf{x}} \leq \mathbf{0}$.

In the minimization problems above, it is tacitly assumed that the register values are not to be changed and so the minimization is over the survey variables in $\mathbf{x}$ only. Similarly, the adjustment models in the next paragraph apply only to the the values that are adjusted and not to the fixed register values. In Appendix D this distinction is made more explicit. In section 6.6 we consider situations where the distinction between fixed and adjustable values does not coincide with the distinction between register and survey values but values of both sources can be adjusted.

6.3 Loss-functions and adjustment models

The conditions for a solution to the problem formulated in (6.1) can be found by inspection of the Lagrangian for this problem, which can be written as

$$L(\mathbf{x}, \boldsymbol{\alpha}) = D(\mathbf{x}, \mathbf{x}_0) + \sum_k \alpha_k \left(\sum_i a_{ki} x_i \right), \tag{6.2}$$

with $\boldsymbol{\alpha}$ a vector of Langrange multipliers, one for each of the constraints k, a_{ki} the element in the k-th row (corresponding to constraint k) and i-th column (corresponding to variable x_i) of the restriction matrix $\mathbf{A}$ and $D(\mathbf{x}, \mathbf{x}_0)$ a loss-function measuring the distance or discrepancy between $\mathbf{x}$ and $\mathbf{x}_0$. From optimisation theory it is well known that for a convex function $D(\mathbf{x}, \mathbf{x}_0)$ and linear (in)equality constraints, the solution vector $\tilde{\mathbf{x}}$ must satisfy

the so-called Karush-Kuhn-Tucker (KKT) conditions (see, e.g. Luenberger, 1984). One of these conditions is that the gradient of the Lagrangian w.r.t. $\mathbf{x}$ is zero, i.e.

$$L'_{x_i}(\tilde{x}_i, \boldsymbol{\alpha}) = D'_{x_i}(\tilde{x}_i, \mathbf{x}_0) + \sum_k \alpha_k a_{ki} = 0, \tag{6.3}$$

with L'_{x_i} the gradient of L w.r.t. $\mathbf{x}$ and D'_{x_i} the gradient of D w.r.t. $\mathbf{x}$. From this condition alone, we can already see how different choices for D lead to different solutions to the adjustment problem. Below we shall consider three familiar choices for D, Least Squares, Weighted Least Squares and Kullback-Leibler divergence, and show how these different choices result in different structures of the adjustments, which we will refer to as the adjustment models.

6.3.1 Least Squares (LS)

First, we consider the least squares criterion to find an adjusted $\mathbf{x}$-vector that is closest to the original unadjusted data, that is:

$$D(\mathbf{x}, \mathbf{x}_0) = \tfrac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0),$$

is the Least Squares (LS) criterion, $D'_{x_i}(\tilde{x}_i, \mathbf{x}_0) = \tilde{x}_i - x_{0,i}$, and we obtain from (6.3)

$$\tilde{x}_i = x_{0,i} + \sum_k a_{ki}\alpha_k \tag{6.4}$$

This shows that the least squares criterion results in an additive structure for the adjustments: the total adjustment to variable $x_{o,i}$ decomposes as a sum of adjustments to each of the constraints k. Each of these adjustments consists of an adjustment parameter α_k that describes the amount of adjustment due to constraint k and the entry a_{ki} of the constraint matrix $\mathbf{A}$ pertaining to variable i and constraint k (with values 1,-1 or 0) that describes whether variable $x_{o,i}$ is adjusted by α_k, $-\alpha_k$ or not at all.

For variables that are part of the same constraints and have the same value a_{ki}, the adjustments are equal and the differences between adjusted variables are the same as in the unadjusted data. In particular, this is the case for variables that add up to a fixed total, given by a register value, and are not part of other constraints.

6.3.2 Weighted Least Squares (WLS)

For the weighed least squares criterion,

$$D(\mathbf{x}, \mathbf{x}_0) = \tfrac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T Diag(\mathbf{w})(\mathbf{x} - \mathbf{x}_0),$$

with $Diag(\mathbf{w})$ a diagonal matrix with a vector with weights along the diagonal, we obtain from (6.3)

$$\tilde{x}_i = x_{0,i} + \frac{1}{w_i} \sum_k a_{ki}\alpha_k. \tag{6.5}$$

Contrary to the least squares case where the amount of adjustment to a constraint is equal in absolute value (if it is not zero) for all variables in that constraint, the amount of adjustment now varies between variables according to the weights: variables with large weights are adjusted less than variables with small weights.

For variables that are part of the same constraints and have the same value a_{ki}, the adjustments are equal up to a factor $1/w_i$ and the differences of the weighted adjusted variables are the same as in the unadjusted data $w_i\tilde{x}_i - w_j\tilde{x}_j = w_i x_{0,i} - w_j x_{0,j}$.

The weighted least squares approach to the adjustment problem has been applied by Thomson et al. (2005) in the context of adjusting records with inconsistencies caused by imputation. Some of the variables were missing and the missings were filled in by imputed values without taking care of edit constraints. This caused inconsistencies that were resolved by minimal adjustments, in principle to all variables, observed or imputed, according to the WLS-criterion. They used weights of 10,000 for observed values and weights of 1 for imputed values. Effectively, this means that if a consistent solution can be obtained by changing only imputed variables, this solution will be found. Otherwise (some of the) observed variables will also be adjusted.

One specific form of weights that is worth mentioning is obtained by setting the weight w_i equal to $1/x_{0,i}$ resulting, after dividing by $x_{0,i}$ in the adjustment model

$$\frac{\tilde{x}_i}{x_{o,i}} = 1 + \sum_k a_{ki}\alpha_k, \tag{6.6}$$

which is an additive model for the *ratio* between the adjusted an unadjusted values. It may be noted that the expression on the right-hand side of (6.6) is the first-order Taylor expansion (i.e. around 0 for all the α_k's) to the multiplicative adjustment given by

$$\frac{\tilde{x}_i}{x_{o,i}} = \prod_k (1 + a_{ki}\alpha_k) \tag{6.7}$$

From (6.6) we see that the α_k's determine the difference from 1 of the *ratio* between the adjusted and unadjusted values, which is usually much smaller than unity in absolute value (e.g. an effect of 0.2 implies a 20% increase due to adjustment which is large in practice). The products of the α_k's are therefore often much smaller than the α_k's themselves, in which cases (6.6) becomes a good approximation to (6.7), i.e. the corresponding WLS adjustment is roughly given as the product of the constraint-specific multiplicative adjustments.

6.3.3 Kullback-Leibler divergence (KL)

The Kullback-Leibler divergence measures the difference between $\mathbf{x}$ and $\mathbf{x}_0$ by the function $D_{KL} = \sum_i x_i(\ln x_i - \ln x_{0,i} - 1)$. It can be shown that for this discrepancy measure, the adjustment model takes on the following form

$$\tilde{x}_i = x_i \times \prod_k \exp\left(-a_{ik}\alpha_k\right). \tag{6.8}$$

In this case the adjustments have a multiplicative form and the adjustment for each variable is the product of adjustments to each of the constraints. The adjustment factor $\gamma_k = \exp(-a_{ik}\alpha_k)$ in this product represents the adjustment to constraint k and equals 1 for a_{ik} is 0 (no adjustment), $1/\gamma_k$ for a_{ik} is 1 and the inverse of this factor, γ_k, for a_{ik} is -1.

For variables that are part of the same constraints and have the same value a_{ki}, the adjustments factors are equal and the ratios between adjusted variables are the same as between the unadjusted variables, $\tilde{x}_i/\tilde{x}_j = x_{0,i}/x_{0,j}$.

6.4 Numerical illustrations

The different methods (LS, WLS and KL) have been applied to the example record. For the WLS method we used as weights the inverse of the $\mathbf{x}_0$-values so that the relative differences between $\mathbf{x}$ and $\mathbf{x}_0$ are minimized and the adjustments are proportional to the size of the $\mathbf{x}_0$-values. The results for the different methods are in table 6.2. The register values that are treated as fixed are shown in bold, the other values may be changed by the adjustment procedure.

The LS adjustment procedure leads to one negative value, which is not allowed for this variable. Therefore the LS-procedure was run again with a non-negativity constraint added for the variable *Turnover other.* This results simply in a zero for that variable and a change in *Turnover main* to ensure that *Turnover = Turnover main + Turnover other.* Without the non-negativity constraint, the LS-results clearly show that for variables that are part of the same constraints (in this case the pairs of variables x_3, x_4 and x_6, x_7 that are both appearing in one constraint only), the adjustments are equal: -40 for x_3, x_4 and +50 for x_6, x_7.

The WLS procedure with weights equal to the inverse of the original values and the KL procedures lead to the same solution. It can shown analytically that this should be the case for this simple example. In general, with this weighting scheme, the solutions should be similar but not identical. It is clear that for the WLS/KL solution, the adjustments are larger for large values of the original record than for smaller values. In particular, the adjustment to *Turnover other* is only -2.3, so that no negative adjusted value results in this case, whereas the adjustment to *Turnover main* is 77.7. The multiplicative nature of these adjustments (as KL-type adjustments) also clearly shows since the adjustment *factor* for both these variables is 0.92. The adjustment factor for *Wages* and *Other costs* is also equal (to 1.14) because these variables are in the same single constraint and so the ratio between these variables is unaffected by this adjustment.

Table 6.2. Example business record and adjusted values

Variable	*Name*	*Sample/Register Value*	*Adjusted value*		
			LS	LS non-negative	WLS/KL
x_1	Profit	330	150	150	150
x_2	Employees	**25**	**25**	**25**	**25**
x_3	Turnover main	1000	960	950	922
x_4	Turnover other	30	-10	0	28
x_5	Turnover	**950**	**950**	**950**	**950**
x_6	Wages	200	250	250	228
x_7	Other costs s	500	550	550	571
x_8	Total costs	**800**	**800**	**800**	**800**

6.5 Other solutions for adjusting the sample data

The current solutions all use no more information then the data from the survey and register and the edit rules. Other knowledge about the relations

between variables is not taken into account and, as a consequence, these relations may be distorted by the adjustment process. Below we give two ways of using substantive knowledge to improve the preservation of the relations between variables; by the choice of variables to be adjusted and by adding edit rules. In these approaches, the minimization problem is not changed. In section 6.7 other ways to use additional are proposed, by slightly modifying the objective function.

6.5.1 Adjusting fewer variables

Thus far, in the solution to the minimal adjustment problem, all survey variables appearing in any of the constraints could be adjusted. However, it is not necessary for the solution of the inconsistency problem to adjust all these variables. Variables that are not related to the variables from the register, should preferably not be adjusted.

For example, imputation models for the structural business statistics at Statistics Netherlands use, for most variables, *Turnover* as a predictor. However, it turned out that for some variables related to *Employment*, such as *Wages,* Turnover is not a good predictor. When available *Employment* would obviously be a good predictor but even a stratum mean or a historical value is better fort these variables.

Because of the weak relation between *Wages* and *Turnover* one can choose, in the adjustment procedure, not to adjust *Wages* and let the adjustment to *Costs other* alone take care of satisfying the edit rule a3.

6.5.2 Adding edit rules

Variables that do not appear in any edit-rules will not be adjusted. This can distort relations between the variables that are in the edit rules and the variables that are not. One way to remedy this problem is to add extra edit rules. An example is the variable *Employment* that is in none of the edit rules. However, if *Wages* changes it seems reasonable to let *Wages* change also. This can be accomplished by adding the edit rule

$$Employment = 10 \times Wages,$$

which reflects the ratio of these two variables observed in the sample. However, since this ratio is an estimate from a sample one may allow some deviation from this edit-rule by restricting the value of *Employment* to an interval

$$12 \times Wages \geq Employment \geq 8 \times Wages.$$

Probably better ways to deal with the problem of variables that do not appear in edit rules will be discussed in section 6.7.

6.6 Adjusting to multiple sources and using soft constraints

In this section we consider the possibilities for further modelling of the adjustment problem by using, simultaneously, information from multiple sources. First, we consider the situation that both register and survey values are considered to provide information for the final adjusted consistent estimates rather than discarding survey values for which register values are available. Then we show that the approach used to combine information from multiple sources can be viewed as using, in addition to the "hard" constraints that are to be satisfied exactly, also "soft" constraints that only need to be fulfilled approximately.

6.6.1 Adjusting to both survey and register values

So far we considered the case where one of the sources (the administrative one) provides the reference values that are considered to be the correct ones and these values replace the values of the corresponding survey variables. Another situation arises when both data sources are considered to be fallible. In this situation we do not want to discard the data from one of the sources but we consider both sources to provide useful information on the variables of interest. This means that in the final consistent estimated vector we should not simply copy the values from the register values but obtain adjusted values that depend on both the survey values and the available register values. The data from the survey will be denoted by $\mathbf{x}_{0,S}$ and the data from the register by $\mathbf{x}_{0,R}$. In particular, for the example in table 6.1, the data that are used are the following:

$\mathbf{x}_{0,S}^T$=(*Profit, Employees, Turnover main, Turnover other, Turnover, Wages, Other costs, Total costs*),

$\mathbf{x}_{0,R}^T$=(*Employees_reg, Turnover_reg, Total costs_reg).*

where the suffix *_reg* is used to distinguish the register variables from their survey counterparts.

The use of both sources can be accomplished by setting up the loss function as follows:

A consistent minimal adjustment procedure based on the information from the survey values, the register values and the edit rules can be set up by considering the following constrained optimization problem

$$\begin{aligned} &\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}}\{D(\mathbf{x}, \mathbf{x}_{0,S}) + D(\mathbf{x}_R, \mathbf{x}_{0,R})\} \\ &\text{s.t. } \mathbf{Ax} \leq \mathbf{0} \end{aligned} \tag{6.9}$$

where the vector $\mathbf{x}_R$ denotes the subvector of $\mathbf{x}$ that contains the variables that are observed in the register. The vectors $\mathbf{x}$ and $\mathbf{x}_{0,S}$ both contain all variables and can be partitioned as $\mathbf{x} = (\mathbf{x}_{\bar{R}}^T, \mathbf{x}_R^T)^T$ and $\mathbf{x}_{0,S} = (\mathbf{x}_{0,S\bar{R}}^T, \mathbf{x}_{0,SR}^T)^T$, with $\bar{R}$ denoting the set of variables not in the register. Using this partitioning and the property that the distance functions considered in this paper are all decomposable in the sense that they can be written as a sum over variables, (6.9) can be re-expressed as

$$\begin{aligned} &\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}}\{D(\mathbf{x}_{\bar{R}}, \mathbf{x}_{0,S\bar{R}}) + D(\mathbf{x}_R, \mathbf{x}_{0,SR}) + D(\mathbf{x}_R, \mathbf{x}_{0,R})\} \\ &\text{s.t. } \mathbf{Ax} \leq \mathbf{0} \end{aligned} . \tag{6.10}$$

This clearly shows that the values of the variables R that are in both the register and the survey are adjusted to satisfy the edit constraints and remain as close as possible to both the register value and the survey value. Note that variables that are in both the register and the survey will be adjusted, if the two values are not equal, even if they do not appear in any edit rules, which is different from the situation considered before.

6.6.2 Soft constraints

The adjustment towards the register values due to a separate component in the objective function can also be interpreted as adding "soft" constraints to the optimization problem. These soft constraints express that $\tilde{x}_R$ should be approximately equal to the register values $x_{0,R}$ but need not "fit" these data exactly as was required before.

The notion of soft constraints opens up a number of possibilities for further modelling the adjustment problem. In particular, the hard ratio constraint on *Employment* and *Wage*s, used as an example in section 6.6.1 can be made into a soft constraint by adding to the loss function the component $D\left(x_{wages}, 10 \times x_{employment}\right)$. This soft constraint is often more reasonable then using hard upper and lower bounds on the adjusted value for *Employment* and *Wages*. In fact we can do both, for instance to bound *Wages* within certain hard limits and use soft constraints to draw the value of *Wages* within these bound towards the expected value of ten times the number of employees.

6.7 Conclusions

We have considered procedures to make a linked record, containing variables obtained from different sources, consistent with edit rules. This can be seen as integrating information from five different sources:

1. Data from an administrative source;
2. Data from a survey;
3. An assumed model for the differences between the inconsistent linked record and the consistent adjusted (or integrated) record.
4. Knowledge in the form of logical "hard" edit constraints;
5. Knowledge in the form of "soft" edit constraints.

Each of these five pieces of information has its influence on the values in the resulting integrated record.

When the differences between the survey variables and the register variables are small measurement errors, the effect of the choice of model will also be small and therefore unimportant in practice. However, the adjustment procedure can also be used in other settings where the differences between the adjusted and unadjusted records are larger and the choice of model becomes more important.

One such a setting occurs when the data from the survey are missing for some unit (unit non-response). For large firms with a sampling fraction of one (the continuing units that are always in the sample), it is customary to impute for the missing values, for instance with the values for that same unit on a previous occasion (t-1). If current register values are available for some variables, these can be used as imputations in stead of the (t-1) values. This will, however, result in an inconsistent record since it contains values from two different sources. An adjustment procedure can then be applied to this record consisting of current register values and (t-1) survey values. The adjustment model may matter in this case and the model can be seen as part of the imputation model.

Another setting in which the adjustment model may matter is when one of the sources is from a different point in time then the other. If we adjust towards the most recent data, the adjustment model can be seen as part of an extrapolation procedure and should be chosen with this purpose in mind.

Chapter 7

Macro-integration techniques with applications to census tables and labor market statistics

Nino Mushkudiani, Jacco Daalmans and Jeroen Pannekoek[1]

Macro-integration is widely used for the reconciliation of macro figures, usually in the form of large multi-dimensional tabulations, obtained from different sources. Traditionally these techniques have been extensively applied in the area of macro-economics, especially in the compilation of the National Accounts. Methods for macro-integration have developed over the years and have become very versatile techniques for solving integration of data from different sources at a macro level. In this paper we propose applications of macro-integration techniques in other domains than the traditional macro-economic applications. In particular, we present two possible applications for macro-integration methods: reconciliation of tables of a virtual census and combining estimates of labor market variables.

7.1 Introduction

Macro-integration is widely used for the reconciliation of macro figures, usually in the form of large multi-dimensional tabulations, obtained from different sources. Traditionally these techniques have been extensively applied in the area of macro-economics, especially in the compilation of the National

[1]Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: nmsi@cbs.nl. Remark: The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Accounts, for example to adjust input-output tables to new margins (see, e.g. Stone et al., 1942). Combining different data at the macro level, while taking all possible relations between variables into account, is the main objective of reconciliation or macro-integration. Combining different data sources also makes possible to detect and correct flaws in data and to improve the accuracy of estimates. The methods for macro-integration have developed over the years and have become very versatile techniques for solving integration of data from different sources at a macro level. In this paper we propose several new applications of macro-integration techniques in other domains than the traditional macro-economic applications.

Currently, at Statistics Netherlands, methods for macro-integration are applied to match quarterly and yearly values of the National Accounts. The multivariate Denton method (see Bikker and Buijtenhek, 2006) was extended for this application with ratio restrictions, soft restrictions, inequalities and variances. Besides a large number of variables, this model can also handle a very large number of restrictions (see Bikker et al., 2010). Another application of macro-integration at SN uses a Bayesian approach to deal with inclusion of inequality constraints, for integration of international trade statistics and transport statistics (see Boonstra et al., 2010).

In this paper we investigate the application of macro-integration techniques in the following areas:

- Reconciliation of tables for the Census 2011;
- Early estimates for the labor market variables.

The paper is organized as follows: in Section 7.2 we will give a short outline of macro-integration methods used in this paper, including the extended Denton method. In Section 7.3, we describe the Census 2011 data problem for SN and the use of macro-integration for it. In Section 7.4, we will do the same for the early estimates of labor market variables. The conclusions can be found in Section 7.5.

7.2 Methods

7.2.1 The macro-integration approach

We consider a set of estimates in tabular form. These can be quantitative tables such as average income by region, age and gender or contingency tables arising from the cross-classification of categorical variables only, such

as age, gender, occupation and employment. If some of these tables have certain margins in common and if these tables are estimated using different sources, these margins will often be inconsistent. If consistency is required, a macro-integration approach can be applied to ensure this consistency.

The macro-integration approach to such reconciliation problems is to view them as constrained optimization problems. The totals from the different sources that need to be reconciled because of inconsistencies are collected in a vector $\mathbf{x}$ $(x_i : i = 1, \ldots, N)$. Then a vector $\widehat{\mathbf{x}}$, say, is calculated that is close to $\mathbf{x}$, in some sense, *and* satisfies the constraints that ensure consistency between the totals. For linear constraints, the constraint equations can be formulated as

$$\mathbf{C}\widehat{\mathbf{x}} = \mathbf{b}. \tag{7.1}$$

where $\mathbf{C}$ is a $c \times N$ matrix, with c the number of constraints and $\mathbf{b}$ a c-vector. These linear constraints include equality constraints that set the corresponding margins of tables estimated from different sources equal to each other as well as benchmarking constraints that set the estimates of certain margins from all sources equal to some fixed numbers. The equality constraints are likely to apply to common margins that can be estimated from different sample surveys but cannot be obtained from a population register, while the benchmarking constraints are likely to apply when the common margins can be obtained from register data in which case the fixed numbers are the values for this margin obtained from the register.

Consider a class of penalty functions represented by $(\mathbf{x} - \widehat{\mathbf{x}})'\mathbf{A}(\mathbf{x} - \widehat{\mathbf{x}})$, a quadratic form of differences between the original and the adjusted vectors, here $\mathbf{A}$ is a symmetric, $N \times N$ nonsingular matrix. The optimization problem can now be formulated as:

$$\min_{\widehat{\mathbf{x}}} \ (\mathbf{x} - \widehat{\mathbf{x}})'\mathbf{A}(\mathbf{x} - \widehat{\mathbf{x}}), \quad \text{with} \ \ \mathbf{C}\widehat{\mathbf{x}} = \mathbf{b}.$$

In the case that $\mathbf{A}$ is the identity matrix, we will be minimizing the sum of squares of the differences between the original and new values:

$$(\mathbf{x} - \widehat{\mathbf{x}})'(\mathbf{x} - \widehat{\mathbf{x}}) = \sum_{i=1}^{N} (x_i - \widehat{x}_i)^2.$$

To solve this optimization problem, the Lagrange method can readily be applied. The Lagrangian is

$$L = (\mathbf{x} - \widehat{\mathbf{x}})'\mathbf{A}(\mathbf{x} - \widehat{\mathbf{x}}) - \boldsymbol{\lambda}'(\mathbf{C}\widehat{\mathbf{x}} - \mathbf{b}) \tag{7.2}$$

with $\boldsymbol{\lambda}$ a vector with Lagrange multipliers. For an optimum, we must have that the gradient of $L(\boldsymbol{\lambda}, \widehat{\mathbf{x}})$ with respect to $\widehat{\mathbf{x}}$ is zero. This gradient is:

$$\frac{\partial L}{\partial \widehat{\mathbf{x}}} = -2(\mathbf{x} - \widehat{\mathbf{x}})'\mathbf{A} - \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0}$$

and hence,

$$2(\mathbf{x} - \widehat{\mathbf{x}}) = -\mathbf{A}^{-1}\mathbf{C}'\boldsymbol{\lambda}. \tag{7.3}$$

By multiplying both sides of this equation with $\mathbf{C}$ and using equation (7.1) we obtain for $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = -2(\mathbf{C}\mathbf{A}^{-1}\mathbf{C}')^{-1}(\mathbf{C}\mathbf{x} - \mathbf{b}),$$

where $\mathbf{C}\mathbf{A}^{-1}\mathbf{C}'$ is a square matrix that is nonsingular as long as there are no redundant constraints. Substituting this result in (7.3) leads to the following expression for $\widehat{\mathbf{x}}$:

$$\widehat{\mathbf{x}} = \mathbf{x} - \mathbf{A}^{-1}\mathbf{C}'(\mathbf{C}\mathbf{A}^{-1}\mathbf{C}')^{-1}(\mathbf{C}\mathbf{x} - \mathbf{b}). \tag{7.4}$$

7.2.2 Comparison with the GREG-estimator

In survey methodology it is common to make use of known marginal totals of variables that are also measured in the survey by the use of calibration or generalized regression (GREG) estimation, (see Särndal C., Swensson B. and Wretman J., 1992). Following Boonstra H.J., (2004), we will compare in this subsection the GREG-estimator with the adjusted estimator given by (7.4) for the estimation of contingency tables with known margins.

The situation in which calibration or GREG-estimation procedures can be applied is as follows. There is a target variable y, measured on a sample of n units, for which the population total, x_y say, is to be estimated. Furthermore, there are measurements on a vector of q auxiliary variables on these same units for which the population totals are known. For the application of the GREG-estimator for the total of y, first the regression coefficients for the regression of y on the auxiliary variables are calculated. Let the measurements on y be collected in the n-vector $\mathbf{y}$ with elements $y_i, (i = 1, \ldots, n)$, and the measurements on the auxiliary variables in vectors $\mathbf{z}_i$ and let $\mathbf{Z}$ be the $n \times q$ matrix with the vectors $\mathbf{z}_i$ as rows. The design-based estimator of the regression coefficient vector $\boldsymbol{\beta}$ can then be obtained as the weighted least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\boldsymbol{\Pi}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\boldsymbol{\Pi}^{-1}\mathbf{y}, \tag{7.5}$$

with $\mathbf{\Pi}$ a diagonal matrix with the sample inclusion probabilities π_i along the diagonal.

Using these regression coefficients the regression estimator for the population total of y is estimated by

$$\hat{x}_{y.greg} = \hat{x}_{y.ht} + (\mathbf{x}_{z.pop} - \hat{\mathbf{x}}_{z.ht})' \hat{\boldsymbol{\beta}}, \tag{7.6}$$

with $\hat{x}_{y.ht}$ and $\hat{\mathbf{x}}_{z.ht}$ the 'direct' Horwitz-Thompson estimators, $\sum_i y_i/\pi_i$ and $\sum \mathbf{z}_i/\pi_i$, for the population totals of y and $\mathbf{z}$, respectively and $\mathbf{x}_{z.pop}$ the known population totals of the auxiliar variables. The regression estimator $\hat{x}_{y.greg}$ can be interpreted as a 'weighting' estimator of the form $\sum_i w_i y_i$ with the weights w_i given by

$$w_i = \frac{1}{\pi_i}\left[1 + (\mathbf{x}_{z.pop} - \hat{\mathbf{x}}_{z.ht})' \left(\mathbf{Z}'\mathbf{\Pi}^{-1}\mathbf{Z}\right)^{-1} \mathbf{z}_i\right]. \tag{7.7}$$

From (7.7) two important properties of the GREG-estimator are directly apparent. Firstly, the weights depend only on the auxiliary variables and not on the target variable. This means that the GREG-estimators for different target variables can be obtained by the same weights as long as the auxiliary variables remain the same. Secondly, the GREG-estimates of the totals of the auxiliary variables, $\hat{x}_{z.greg} = \sum_i w_i \mathbf{z}_i$, are equal to their known population totals.

For multiple target variables, $\mathbf{y}_i = (y_{i1} \ldots y_{ip})$ the GREG-estimators can be collected in a $p-$vector $\hat{\mathbf{x}}_{y.greg}$ and (7.6) generalizes to

$$\hat{\mathbf{x}}_{y.greg} = \hat{\mathbf{x}}_{y.ht} + \mathbf{B}\left(\mathbf{x}_{z.pop} - \hat{\mathbf{x}}_{z.ht}\right), \tag{7.8}$$

with $\hat{\mathbf{x}}_{y.ht}$ the $p-$vector with Horvitz-Thompson estimators for the target variables and $\mathbf{B}$ the $p \times q$-matrix with the regression coefficients for each target variable on the rows. Generalizing (7.5), we have for the coefficient matrix $\mathbf{B} = \mathbf{Y}'\mathbf{\Pi}^{-1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{\Pi}^{-1}\mathbf{Z}\right)^{-1}$, where $\mathbf{Y}$ is the $n \times p$-matrix with the vectors of target variables, $\mathbf{y}_i$, on the rows.

Now, consider the case when the totals to be estimated are the cell-totals of a contingency table obtained by the cross-classification of a number of categorical variables. For instance, the target totals could be the numbers of individuals in the categories 1.Unemployed and 2.Employed of the variable Employment by age category and sex in some (sub)population. If we assume, for ease of exposition, that Age has only two categories, 1.Young and 2.Old and Sex has the categories 1.Male and 2.Female, then there are eight totals to be estimated, one for each cell of a $2 \times 2 \times 2$ contingency table.

Corresponding to each of these eight cells we can define, for each individual, a zero-one target variable indicating whether the individual belongs to this cell or not. For instance $y_1 = 1$ if Employment = 1, Age = 1 and Sex = 1, and zero in all other cases and $y_2 = 1$ if Employment = 2, Age = 1 and Sex = 1, and zero in all other cases, etc. Each individual scores a 1 in one and only one of the eight target variables.

For such tables, some of the marginal totals are often known for the population and GREG-estimators that take this information into account are commonly applied. In the example above, the population totals of the combinations of Sex and Age could be known for the population and the auxiliary variables then correspond to each of the combinations of Sex and Age. The values for the individuals on these auxiliary variables are sums of values of the target variables. For instance, the auxiliary variable for Age = 1 and Sex = 1 is the sum of y_1 and y_2 and will have the value 1 for individuals that are young and male and either employed or unemployed and the value 0 for individuals that are not both young and male. Similarly, we obtain for each of the four Age × Sex combinations zero-one auxiliary variables as the sum of the corresponding target variables for Unemployed and Employed. In general, if there are p target variables and q auxiliary variables corresponding to sums of target variables, we can write the values of the auxiliary variables as

$$\mathbf{z}_i = \mathbf{C}\mathbf{y}_i, \tag{7.9}$$

with $\mathbf{C}$ the $q \times p$ constraint matrix (consisting of zeroes and ones) that generates the sums of the y_i values corresponding to the auxiliary variables. Since (7.9) applies to each row of $\mathbf{Z}$ and $\mathbf{Y}$, we can write $\mathbf{Z} = \mathbf{Y}\mathbf{C}'$ and so

$$\mathbf{B} = \mathbf{Y}'\mathbf{\Pi}^{-1}\mathbf{Y}\mathbf{C}'\left(\mathbf{C}\mathbf{Y}'\mathbf{\Pi}^{-1}\mathbf{Y}\mathbf{C}'\right)^{-1}. \tag{7.10}$$

In the case considered here, where the target variables correspond to cells in a cross-classification of categorical variables, this expression can be simplified as follows. The rows of $\mathbf{Y}$ contain a 1 in the column corresponding the cell to which the unit belongs and zeroes elsewhere. Afther rearranging the rows such that the units that belong to the same cell (score a one on the same target variable) are beneath each other, $\mathbf{Y}$ can be written as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_q} \end{pmatrix},$$

where n_j is the number of units scoring a one on target variable j and $\mathbf{1_{n_j}}$ is a column with n_j ones. In this example there are no units that score on the third target variable. When this matrix is premultiplied by $\mathbf{Y'\Pi^{-1}}$ we obtain $\mathbf{Y'\Pi^{-1}Y} = Diag(\hat{\mathbf{x}}_{y.ht})$ and $\mathbf{B}$ can be expressed as

$$\mathbf{B} = Diag(\hat{\mathbf{x}}_{y.ht})\mathbf{C}' \left(\mathbf{C}Diag(\hat{\mathbf{x}}_{y.ht})\mathbf{C}'\right)^{-1}. \tag{7.11}$$

Substituting this value for $\mathbf{B}$ in (7.8) and using $\mathbf{C}\hat{\mathbf{x}}_{y.ht} = \hat{\mathbf{x}}_{z.ht}$ we obtain

$$\hat{\mathbf{x}}_{y.greg} = \hat{\mathbf{x}}_{y.ht} + Diag(\hat{\mathbf{x}}_{y.ht})\mathbf{C}' \left(\mathbf{C}Diag(\hat{\mathbf{x}}_{y.ht})\mathbf{C}'\right)^{-1} \left(\mathbf{x}_{z.pop} - \mathbf{C}\hat{\mathbf{x}}_{y.ht}\right), \tag{7.12}$$

which is equal to (7.4) with the initial unadjusted vector ($\mathbf{x}$) equal to the Horwitz-Thompson estimators for the cell-totals, the weighting matrix ($\mathbf{A}^{-1}$) a diagonal matrix with the initial vector along the diagonal and the values of the constraints (b) equal to the known population totals of the margins of the contingency table that are used as auxiliary variables.

7.2.3 Extension to time series data

The optimization problem described in 7.2.1 can be extended to time series data of the form x_{it} $(i = 1, \ldots, N,\ t = 1, \ldots, T)$. In this case the total number of the variables x_{it} is $N \cdot T$ and the constraint matrix will have $N \cdot T$ columns. The number of rows will be equal to the number of constraints as before. The matrix A will be a symmetric, $NT \times NT$ nonsingular matrix.

For this data we want to find adjusted values $\widehat{x}_{it}$ that are in some metric ς (for example Euclidean metric) close to the original time series. For this purpose we consider the following objective function

$$\min_{\widehat{x}} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{w_{it}} \varsigma(\widehat{x}_{it}, x_{it}), \tag{7.13}$$

where w_{it} denotes the variance of the i^{th} time series at time t. We minimize this function over all $\widehat{x}_{it}$ satisfying the constraints

$$\sum_{i=1}^{N}\sum_{t=1}^{T} c_{rit}\widehat{x}_{it} = b_r, \quad r = 1, \ldots, C. \tag{7.14}$$

In (7.14), r is the index of the restrictions and C is the number of restrictions. Furthermore, c_{rit} is an entry of the restriction matrix and b_r are fixed constants. Most economic variables cannot have negative signs. To incorporate this (and other) requirement(s) in the model, inequality constraints are included. A set of inequalities is given by

$$\sum_{i=1}^{N}\sum_{t=1}^{T} a_{rit}\widehat{x}_{it} \leq z_r, \quad r = 1, \ldots, I, \tag{7.15}$$

where I stands for the number of inequality constraints.

In Bikker et al. (2010) this model was extended by soft linear and ratio restrictions. A soft equality constraint is different from the hard equality constraints (7.14), in that the constants b_r are not fixed quantities but are assumed to have a variance and an expected value. This means that the resulting $\widehat{x_{it}}$ need not match the soft constraints exactly, but only approximately. A soft linear constraint similar to (7.14) is denoted as follows:

$$\sum_{i=1}^{N}\sum_{t=1}^{T} c_{rit}\widehat{x}_{it} \sim (b_r, w_r), \quad r = 1, \ldots, C. \tag{7.16}$$

By the notation $\sim$ in (7.16) we define b_r to be the expected value of the sum $\sum_{i=1}^{N}\sum_{t=1}^{T} c_{rit}\widehat{x}_{it}$ and w_r its variance. In the case that ς is the Euclidean metric the linear soft constraints can be incorporated in the model by adding the following term to the objective function in (7.13):

$$+\sum_{r=1}^{C}\frac{1}{w_r}\left(b_r - \sum_{i=1}^{N}\sum_{t=1}^{T} c_{rit}\widehat{x}_{it}\right)^2. \tag{7.17}$$

Another important extension of the model in Bikker et al. (2010) is the ratio constraint. The hard and soft ratio constraints that can be added to the model, are given by

$$\frac{\widehat{x}_{nt}}{\widehat{x}_{dt}} = v_{ndt} \quad \text{and} \quad \frac{\widehat{x}_{nt}}{\widehat{x}_{dt}} \sim (v_{ndt}, w_{ndt}), \tag{7.18}$$

where $\widehat{x}_{nt}$ denotes the numerator time series, $\widehat{x}_{dt}$ denotes the denominator time series, v_{ndt} is some predetermined value and w_{ndt} denotes the variance of

a ratio $\frac{\widehat{x}_{nt}}{\widehat{x}_{dt}}$. In order to add the soft ratio constraints to the objective function these are first linearized. The soft constraints in (7.18) can be rewritten as:

$$\widehat{x}_{nt} - v_{ndt}\widehat{x}_{dt} \sim (0, w^*_{ndt}). \tag{7.19}$$

The variance of the constraint will be different, we denote it as w^*_{ndt}. Soft linearized ratios are incorporated in the model in case when ς is an Euclidean metric, by adding the following term to the objective function

$$+ \sum_{n,d=1}^{N} \sum_{t=1}^{T} \frac{(\widehat{x}_{nt} - v_{ndt}\widehat{x}_{dt})^2}{w^*_{ndt}}. \tag{7.20}$$

The extensions of the constraints that can be handled beyond the traditional linear (in)equality constraints, opens up a number of applications to reconciliation problems in several areas. An example of one such application is described in section 7.4.

7.3 Reconciliation of census tables

In this section we describe the Dutch Census data and formulate the reconciliation of census tables as a macro-integration problem.

The aim of Dutch Census 2011 is to produce over 40 crosstables about demographics and occupation of Dutch population. For this task, data from many different sources and different structures are combined. The majority of the variables are obtained from the GBA (municipality data bases), however quite a few other sources (samples and registers) are used as well, such as the labor force survey (LFS).

Each table consists of up to 10 variables. We call these high dimensional crosstables hypercubes. Most of the variables are included in many hypercubes. The hypercubes have to be consistent with each other, in a sense that all marginal distributions that can be obtained from different crosstables are the same. Consistency is required for one dimensional marginals, e.g. the number of men, as well as for multivariate marginals, e.g. the number of divorced men aged between 25 and 30 year. Multivariate marginal crosstables are hypercubes as well.

In different hypercubes one variable may have a different category grouping (classification). For example, the variable age can be requested to be included in different hypercubes aggregated in different ways: groups of ten years, five

years and one year. Still, the marginal distributions of age obtained from different hypercubes should be the same for each level of aggregation.

In general, the data that are collected by SN involve many inconsistencies; the cause of this varies: different sources, differences in population coverage, different time periods of data collection, nonresponse, or modeling errors.

Currently at SN, the method of repeated weighting (see Houbiers, 2004) is used to combine variables from different sources and to make them consistent. Using repeated weighting, tables are reconciled one by one. Assuming that the tables 1 till t are correct, these figures are fixed. Then, the method of repeated weighting adjusts table $t + 1$, so that all marginals of this table become consistent with the marginals of all previous tables, 1 till t. The method of repeated weighting was successfully used for the last census in 2001. However, the number of the tables has increased since and with the number of tables the number of restrictions also increased. As a consequence, it is not obvious that the method of repeated weighting will work for the Census 2011.

The method of macro-integration has some advantages over repeated weighting. Firstly, the method of macro-integration reconciles all tables simultaneously, meaning that none of the figures need to be fixed during the reconciliation process. By doing so, there are more degrees of freedom to find a solution than in the method of repeated weighting. Therefore a better solution may be found, which requires less adjustment than repeated weighting. Secondly, the results of repeated weighted depend on the order of weighting the different tables, while the macro-integration approach does not require any order. Thirdly, the method of macro-integration allows inequality constraints, soft constraints and ratio constraints, which may be used to obtain better results.

A disadvantage of macro-integration is that a very large optimization problem has to be solved. However, by using up-to-date solvers of mathematic optimization problems, very large problems can be handled. The software that has been built at Statistics Netherlands for the reconciliation of National Accounts tables is capable of dealing with a large number of variables (500 000) and restrictions (200 000).

We should emphasize that reconciliation should be applied on the macro level. First the imputation and editing techniques should be carried out for each source separately on the micro level. The aggregated tables should then be produced, containing variables at the publication level. Furthermore, for each separate aggregated table, a variance of each entry in the table

should be computed, or at least an indication of the variance should be given. For example, an administrative source will in general have the most reliable information, and hence have a very small or zero variance. Each aggregated table separately can be optimized to get variance as small as possible, by imputation or other means. During the reconciliation process, each entry of all tables will be adapted in such a way that the entries with the highest variance will be adapted the most, until all constraints are met.

The procedure that we propose here is as follows:

1. For each data source define the variables of interest;
2. Use imputation and editing techniques to improve data quality on a micro level;
3. Aggregate the data to produce the tables, and calculate the variances of each entry;
4. Use reconciliation to make the tables consistent. Calculate the covariance matrix for the reconciled table.

We have identified different kinds of reconciliation problems for census data:

I For some variables we will have different classifications, for example the variable Age can be in years, or five year intervals or ten year intervals. It is required that number of persons obtained from the hypercube with the variable Age with one year intervals for example from 10 till 20 years should add up to the number of persons of this age interval obtained from any other hypercube, where Age is measured in five or ten years intervals. The objective function and the constraints can be modified in order to handle this problem.

II In the macro-integration approach presented in this paper, the reconciliation is carried out at the macro level. It is assumed that an initial estimate for each hypercube can be made. However the estimation of these hypercubes is not always straightforward. This is especially the case for hypercubes that include variables from different data sources: for example a register and a sample. An example of such a case is considered in Appendix E In this example we have three variables (Province, Gender and Age) obtained from the register and a sample that contains these three variables and the additional variable: Occupation. In Appendix E we combine these two data sets using macro-integration

notations, to obtain an initial estimate of one hypercube. This reconciliation problem is very simple and macro-integration might not be the first choice to solve the problem. However it immediately becomes complex when the number of hypercubes and the number of variables in these hypercubes increase.

III A problem that has to be solved in any method is the lack of information. Part of the source information is based on samples. However, these samples may not cover each of the categories of the variables in the hypercubes. For instance, a sample may not include any immigrant from Bolivia, while this sample may be the only source for some of the tables in the census.

7.3.1 The objective function

We distinguish two steps while making the census hypercubes:

1. At first the hypercubes should be made from all available sources;
2. Then for all hypercubes we should make the same marginals equal;

Building of the census hypercubes from different sources could be carried out using many different methods, like weighting or post-stratification. In Appendix E we present a simple example of making a hypercube using two different data. In this section we will not discuss these methods. From the macro-integration point of view the second step of making the hypercubes is of our interest.

Using the notation from the previous section we can now apply the macro-integration method for reconciliation of the hypercubes by their common marginals. In the previous section we defined the objective function (7.13) using an arbitrary metric. Here we use an Euclidean metric.

We introduce the notations especially for census data. A hypercube is defined by $H^{(j)}$, $j = 1, \ldots, N$ and any of this marginal hypercube is defined by $M^{(j)}$. A variable in the hypercube $H^{(j)}$ is defined by $x_i^{(j)}$, where the subindex i is the identity of the variable, for example Province or Age and the super index (j) identifies the hypercube where the variable is in. For example, if we have two hypercubes $H^{(1)}$ and $H^{(2)}$, the variables from $H^{(1)}$ will be defined by $x_1^{(1)}, x_2^{(1)}, \ldots, x_m^{(1)}$, assuming that the hypercube $H^{(1)}$ consists of m variables. Suppose now that the hypercube $H^{(2)}$ consists of n variables and it has three variables $x_1^{(2)}, x_2^{(2)}$ and $x_4^{(2)}$ in common with the hypercube $H^{(1)}$. Denote the marginal hypercube of $H^{(1)}$ consisting of these variables by $M_{1,2,4}^{(1)}$:

$$M^{(1)}_{1,2,4} = x^{(1)}_1 \times x^{(1)}_2 \times x^{(1)}_4.$$

Reconciling the hypercubes $H^{(1)}$ and $H^{(2)}$ so that their common marginal hypercubes are same will mean the finding of hypercubes $\widehat{H}^{(1)}$ and $\widehat{H}^{(2)}$ such that:

$$\varsigma(H^{(1)}, \widehat{H}^{(1)}) + \varsigma(H^{(2)}, \widehat{H}^{(2)}) \tag{7.21}$$

reaches its minimum under the condition that:

$$\widehat{M}^{(1)}_{1,2,4} = \widehat{M}^{(2)}_{1,2,4}. \tag{7.22}$$

In the case when the first marginal hypercube $M^{(1)}_{1,2,4}$ consists of the variables from a register, that are fixed and should not be reconciled, then instead of the condition in (7.22) we will have the following

$$\widehat{M}^{(2)}_{1,2,4} = M^{(1)}_{1,2,4}. \tag{7.23}$$

We can now define the objective function for the reconciliation of the hypercubes $H^{(j)}$, $j = 1, \ldots, N$. We want to find the hypercubes $\widehat{H}^{(j)}$, $j = 1, \ldots, N$ such that:

$$\min_{\widehat{H}} \sum_j \varsigma(H^{(j)}, \widehat{H}^{(j)}), \tag{7.24}$$

under the restriction that, all common marginal hypercubes are the same

$$\widehat{M}^{(j_1)}_{i,k,\ldots,l} = \cdots = \widehat{M}^{(j_k)}_{i,k,\ldots,l} \tag{7.25}$$

and the marginal hypercubes consisting of register data that are fixed, will not be changed:

$$\widehat{M}^{(j_1)}_{i,k,\ldots,l} = \cdots = M^{(j_n)}_{i,k,\ldots,l}. \tag{7.26}$$

If we transform the hypercube $H^{(j)}$ into a vector $\mathbf{h^{(j)}} = (\mathbf{h_1^{(j)}}, \ldots, \mathbf{h_{c_j}^{(j)}})'$ we can rewrite the objective function in (7.24) using the notation of the previous section. For all $\mathbf{h^{(j)}}$, $j = 1, ..., N$, we want to find vectors $\widehat{h}^{(j)}$, $j = 1, ..., N$ such that:

$$\min_{\widehat{\mathbf{h}}} \sum_{j=1}^{N} \sum_{i=1}^{c_j} \frac{1}{w_{ij}} \left(\widehat{h}_i^{(j)} - h_i^{(j)}\right)^2, \tag{7.27}$$

where w_{ij} is the weight of $h_i^{(j)}$.

7.3.2 Reconciliation of two hypercubes

Suppose we want to create two hypercubes, each with three variables. Hypercube one $H^{(1)}$ consists of variables Gender, Age and Occupation and the second hypercube, $H^{(2)}$ of the variables Gender, YAT (year of immigration) and Occupation. For convenience, we combine the original categories of these variables and consider the coding as it is presented in Table 7.1.

Table 7.1. Categories of the variables

Gender	1	Male
	2	Female
Age	1	<15 years
	2	15-65 years
	3	>65 years
Occupation	0	Not manager
	1	Manager
YAT	0	Not immigrant
	1	Immigrated in 2000 or later
	2	Immigrated before 2000

From these variables the only one that is observed in the survey is Occupation, the other three variables are obtained from the register and are therefore assumed to be fixed. The survey we use here is the LFS (labor force survey) and the register is the GBA (municipality data bases). As we mentioned already we assume that the figures obtained from GBA are exogenous, what means that these values should not be changed.

We aim to find the hypercubes $\widehat{H}^{(1)}$ and $\widehat{H}^{(2)}$ such that

$$\varsigma(H^{(1)}, \widehat{H}^{(1)}) + \varsigma(H^{(2)}, \widehat{H}^{(2)}) \tag{7.28}$$

is minimized under the restrictions that the marginal hypercubes of $\widehat{H}^{(1)}$ and $\widehat{H}^{(2)}$ coincide with the corresponding marginal hypercubes of the register. Hence we want to achieve that:

$$\widehat{M}^{(1)}_{\text{Gender, Age}} = M^{\text{register}}_{\text{Gender, Age}} \tag{7.29}$$

and

$$\widehat{M}^{(2)}_{\text{Gender, YAT}} = M^{\text{register}}_{\text{Gender, YAT}}. \tag{7.30}$$

In addition, the hypercubes should be reconciled with each other:

$$\widehat{M}^{(1)}_{\text{Gender, Occupation}} = \widehat{M}^{2}_{\text{Gender, Occupation}}; \tag{7.31}$$

Table 7.2. Hypercube 1

Sex	Age	Occup	0	I	II	III	IV	V
1	1	0	1761176	1501748	1501748	1501748	1501748	1501748
1	2	0	5181009	5065650	5065650	4924068	4907253	4916858
1	2	1	674373	507128	507128	648710	665525	655920
1	3	0	584551	831315	831315	1016430	1016072	1016276
1	3	1	13011	207889	20788	22774	23132	22928
2	1	0	1661478	1434236	1434236	1434236	1434236	1434236
2	2	0	5755370	5521997	5484427	5254234	5247781	5251467
2	2	1	241251	-37570	0	230193	236646	232960
2	3	0	534231	976868	986261	1370781	1370724	1370757
2	3	1	2037.85	399226	389833	5313	5370	5337

The first step before the actual reconciliation process is weighting up the sample to the population. The total number of GBA persons is $N_{GBA} = 16\,408\,487$ and the total number of LFS persons is $N_{LFS} = 104\,674$. The initial weight is

$$w = \frac{16\,408\,487}{104\,674}.$$

Table 7.3. Hypercube 2

Sex	YAT	Occup	0	I	II	III	IV	V
1	0	0	6723037	6505428	6505428	6378041	6362791	6371502
1	0	1	609945	444221	444221	571608	586858	578147
1	1	0	179174	213134	213134	291188	290865	291049
1	1	1	12697	98543	98543	20489	20812	20628
1	2	0	624524	680151	680151	773017	771417	772331
1	2	1	64741	172253	172253	79387	771417	772331
2	0	0	6965385	6889146	6879753	6870198	6864427	6867723
2	0	1	215699	184908	194301	203856	209627	206331
2	1	0	232472	253743	244350	319060	318945	319010
2	1	1	4232	70951	80344	5634	5749	5684
2	2	0	753222	790213	780820	869994	869369	869726
2	2	1	23357	105796	115189	26015	26640	26283

The results of the weighting are presented in Tables 7.2 and 7.3 under the column 0. Since we consider these figures as the starting figures before the

reconciliation process, we call these model 0. These figures have marginals consistent with each other but not with the register data, see Table 7.4. For example, the total number of men is 8214119 from Table 7.2 and 7.3 and 8113730 in Table 7.4.

We applied the optimization solver XPRESS for the problem defined in (7.28-7.31) using the Euclidean distance for ς and applying the weight 1 for all figures. The results of this reconciliation are presented in Tables 7.2 and 7.3 under the column I. We observed negative figures after the reconciliation, therefore we added the restriction that all figures have to be nonnegative to the previous setting and applied the solver. Results of this optimization problem are presented in Tables 7.2 and 7.3 under the column II. Next we used weights equal to the initial value of each figure. The results of this execution are to be found under the column III in Tables 7.2 and 7.3. Applying more realistic weights led to different results, compared with models I and II, the figures with smaller values are adjusted less and the figures with bigger values are adjusted more.

Table 7.4. Register

Sex	Age	YAT	Total
1	1	0	1437385
1	1	1	48553
1	1	2	15810
1	2	0	4619201
1	2	1	255335
1	2	2	698242
1	3	0	893063
1	3	1	7789
1	3	2	138352
2	1	0	1369468
2	1	1	49026
2	1	2	15742
2	2	0	4502083
2	2	1	267916
2	2	2	714428
2	3	0	1202503
2	3	1	7752
2	3	2	165839

Since we want to preserve the initial marginal distribution of the variable Occupation, the next step is to add a ratio restriction. We only added one ratio restriction, that is the relation between the managers and non managers for the whole population. At first we added this restriction as a hard con-

straint and afterwards as a soft constraint to the model. The results of these reconciliation problems are presented in columns IV and V of Tables 7.2 and 7.3. For the soft restrictions the weight we choose is equal to 707405400, which is 100 times the largest register value. This value is found by trial and error. By choosing this value the ratio constraints significantly influences the results, but its effect is clearly less than that of a hard ratio constraint.

Table 7.5. Ratio restriction

Model scenario	Ratio
Target value	16.631
Model outcome: no ratio (III)	17.091
Model outcome: hard ratio (IV)	16.631
Model outcome: soft ratio (V)	16.891

In Table 7.5 the ratios of the number of 'not manager' over the number of 'manager' is calculated for the models III, IV and V. The target value of the ratio is the ratio observed in LFS. As we could expect the value is best achieved in model IV, when the hard ratio restriction has to be fulfilled.

To compare the results of the models with each other we calculated the weighted quadratic difference between the reconciled values of models III, IV and V and the values of model 0, the hypercubes after the weighting, see Table 7.6.

Table 7.6. Weighted square difference

Model scenario	Difference
Model 0 - Model III	1955390
Model 0 - Model IV	1956704
Model 0 - Model V	1956696

The weighted squared difference in Table 7.6 is calculated as follows

$$\sum_{j=1}^{2} \sum_{i=1}^{c_j} \frac{1}{w_{ij}} \left(\widehat{h}_i^{(j)} - h_i^{(j)} \right)^2, \tag{7.32}$$

here we sum over two hypercubes, $\widehat{h}_i^{(j)}$ are the reconciled figures of model III, IV or V and $h_i^{(j)}$ are the values of model 0. The weighted square difference is smallest for model III, which implies that without the ratio restriction reconciled figures are closer to the original figures. We could anticipate this

result since the ratio restriction (as any additional restriction would do) forces the original figures towards the distribution of the ratio and therefore the outcome of the model with the hard ratio restriction differs most from the initial values.

7.4 Early estimates for labor market

The second application of macro-integration methods that we want to study is making early estimates for the labor market variables. This is a very complex task, mainly caused by the variety of data sources, that contain variables with almost equal, but not exactly the same definitions. The sources include one or more labor market variables. The main data sources for labor market variables are the tax office data and the Dutch Labor Force Survey (LFS). The tax office data are updated on a quarterly basis and LFS is a rotating panel design producing monthly figures. Among others we want to combine these two data to construct the early estimates of statistics that are at the moment based only on LFS or tax register data. The difficulties that should be resolved are of a different nature:

- First of all we have series of data of different frequency (in time);
- Definitions of the variables in both sources are often very different;
- Population coverage is not same;
- Other such as survey vs register data, nonresponse, etc.

Because of the different definitions the problem of labor market data in comparison with the National Accounts data is that, , it will not be possible to combine variables on a macro level, without first studying thoroughly the data structure.

We give here a simple example using the macro-integration method for combining two different sources. Suppose we have a labor force population of 60000 persons. We want to conduct a monthly labor forse population survey to find out an unemployment rate. Suppose for simplicity that the auxiliary variables in the population register of our interest are: Sex and Age. We will know the distribution of these variables according to our register;

Suppose for convenience that the total number of respondents in the survey we want to conduct is 6000. We divide 6000 over all cells of Age and Sex according to the register data, see Table 7.8;

Table 7.7. Population

Age	Sex		Total
	Woman	Man	
20-30	6600	6000	12600
30-40	6000	7200	13200
40-50	9600	8400	18000
≥ 50	9000	7200	16200
Total	31200	28800	60000

Table 7.8. Survey

Age	Sex		Total
	Woman	Man	
20-30	660	600	1260
30-40	600	720	1320
40-50	960	840	1800
≥ 50	900	720	1620
Total	3120	2880	6000

Table 7.9. Survey unemployment data

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20-30	35 (25)	34 (27)	36 (25)	35 (29)	37 (33)	33 (30)
30-40	40 (38)	35 (32)	42 (37)	36 (32)	42 (35)	37 (35)
40-50	60 (50)	56 (50)	58 (51)	56 (49)	61 (58)	58 (55)
≥ 50	42 (30)	38 (25)	42 (31)	40 (31)	43 (35)	40 (38)

In the survey we observe two variables: whether a person has a job and if not whether she/he is registered at the unemployment office (CWI). Suppose for simplicity that we do not have nonresponse and that the Table 7.9 of unemployment numbers is the result of the survey for three months, January, February and March. In parenthesis are the numbers of respondents registered at CWI;

From Table 7.9 we can estimate the number of unemployed persons in each group of the population, see Table 7.10. On the other hand from the unem-

Table 7.10. Weighted unemployment data

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20-30	350 (250)	340 (270)	360 (250)	350 (290)	370 (330)	330 (300)
30-40	400 (380)	350 (320)	420 (370)	360 (320)	420 (350)	370 (350)
40-50	600 (500)	560 (500)	580 (510)	560 (490)	610 (580)	580 (550)
≥ 50	420 (300)	380 (250)	420 (310)	400 (310)	430 (350)	400 (380)

ployment office (CWI) we have the number of persons that were registered as unemployed at the end of the quarter, see Table 7.11. Suppose that we do not have the timeliness issues for these survey and register and both data are available for us at around the same time.

Table 7.11. Unemployment office data at the end of the first quarter

Age	Sex	
	Woman	Man
20-30	350	330
30-40	390	360
40-50	600	570
≥ 50	370	395

The estimated registered unemployment figures from the survey and from the CWI are not consistent with each other. For example there are 330 women of age 20-30 registered according to the survey and 350 women according to the register at the end of march. Let us suppose that the register has a variance equal to zero, which means that 350 is a fixed figure.

Now, we want to achieve consistency between the labor force survey and the unemployment office register, in such a way that

1. The weighted survey data may be adjusted, while the CWI register data are fixed.

2. The numbers of persons registered at the CWI at the end of the third month in the reconciled survey data exactly matches the corresponding numbers in the unemployment register.

3. The ratios between the unemployment numbers and the numbers of persons registered at the CWI have to be as close as possible to their initial values as observed in the weighted labour force survey. For instance, for women of age 20-29 this ratio is 37/33 at the end of March. These ratios do not have to hold exactly, as they are observed in the sample.

4. All monthly changes of the number of unemployed persons are as close as possible to their initial value as observed in the weighted survey.

5. All monthly changes of the number of persons registered at the CWI are as close as possible to their initial value as observed in the weighted survey.

We will use a macro-integration model to reconcile the labor force survey with the register of the unemployment office. Define the weighted survey figures of unemployment by x_{ijt} and the number of persons registered at the CWI by y_{ijt}, where t stands for the month and i and j denote the entries of the matrix Age×Sex. The ratios between x_{ijt} and y_{ijt} will be denoted by d_{ijt} (i.e. $d_{ijt} = x_{ijt}/y_{ijt}$). Then, we want to find estimates $\widehat{x}_{ijt}$ of x_{ijt} and $\widehat{y}_{ijt}$ of y_{ijt} that satisfy the properties (1)-(5) listed above. The formulation of the model is

$$\begin{aligned} \min_{\widehat{y}\widehat{x}} \sum_{t=2}^{T} \sum_{ij} & \frac{((\widehat{x}_{ijt} - \widehat{x}_{ijt-1}) - (x_{ijt} - x_{ijt-1}))^2}{v_{1ij}} \\ + & \frac{((\widehat{y}_{ijt} - \widehat{y}_{ijt-1}) - (y_{ijt} - y_{ijt-1}))^2}{v_{2ij}} \\ + & \frac{(\widehat{x}_{ijt} - d_{ijt}\widehat{y}_{ijt})^2}{w^*_{ij}}, \end{aligned} \tag{7.33}$$

with

$$\widehat{y}_{ijt} = y^{CWI}_{ijk}, \quad \text{for all} \quad i, j \quad \text{and} \quad t = 3, k = 1. \tag{7.34}$$

where v_{1ij} denotes the variance of x_{ijt}, v_{2ij} the variance of y_{ijt} and w^*_{ij} is the variance of the linearized ratio $\widehat{x}_{ijt} - d_{ijt}\widehat{y}_{ijt}$.

The first term of (7.33) keeps the differences $\widehat{x}_{ijt} - \widehat{x}_{ijt-1}$ as close as possible to their initial values $x_{ijt} - x_{ijt-1}$ (the aforementioned property 4) and the second term does the same for y_{ijt} (property 5). The third term describes the soft ratio restrictions for the relation between unemployment and registering

at the CWI (property 3). They are similarly defined as the ratio constraints in (7.20). Here, we assume that the variance of the linearised ratios do not depend on t. The hard constraints in (7.34) ensure that the estimates of y_{ijt} of the last month ($t = 3$) are equal to the quarterly unemployment number of the first quarter ($k = 1$), as obtained from the CWI register y_{ijt}^{CWI} (property 2). Note that the quarterly unemployment numbers of the CWI y_{ijt}^{CWI} are included in the model as parameters only. They are not specified as free variables, because these figures are fixed (property 1).

The results of the model (7.33) - (7.34) where we have taken all variances $v_{1ij}, v_{2ij}, w_{ij}^{*}$ to be equal to 300, are shown in Table 7.12. These show that the number of persons registered at the CWI (the numbers between parenthesis) at the end of march are indeed consistent with the unemployment office register (as depicted in Table 7.11).

Table 7.12. Reconciled unemployment data

Age	January		February		March	
	Woman	Man	Woman	Man	Woman	Man
20-29	375.1 (268.0)	375.2 (298.3)	385.0 (268.2)	393.7 (319.0)	393.7 (350.0)	363.8 (330.0)
30-39	445.2 (422.0)	360.8 (329.9)	466.3 (410.9)	467.1 (329.8)	467.1 (390.0)	380.7 (360.0)
40-49	622.4 (519.0)	581.9 (519.5)	602.0 (529.4)	631.5 (509.5)	631.5 (600.0)	601.5 (570.0)
≥ 50	446.0 (318.8)	398.3 (262.5)	445.7 (329.2)	455.2 (323.7)	455.2 (370.0)	416.7 (395.0)

To illustrate the preservation of changes and ratios we focus, as an example, on the number of women in the age 20-29.

Figure 7.4 shows that the initial monthly changes are preserved quite accurately and from Table 7.13 it can be seen that the same holds true for the ratios between the number of unemployed and CWI registered persons. Figure 7.4 also shows that the data reconciliation raises both the number of CWI registered persons and the number of unemployed people at each time period. The explanation is that number of CWI registered persons in the survey at the end of month 3 is below the register figure at the end of the first quarter. Since the survey has to exactly match the register and since all monthly changes of the survey have to be preserved as much as possible, all monthly survey figures on the number of CWI registered persons are increased. The same occurs to the number of unemployed persons, which can be explained from the preservation of the ratios between the number of unemployed and CWI registered people at each time period.

Now, suppose that we decrease the variance of the monthly changes from 300 to 100, but we do not change the variance of the ratios between unemployed

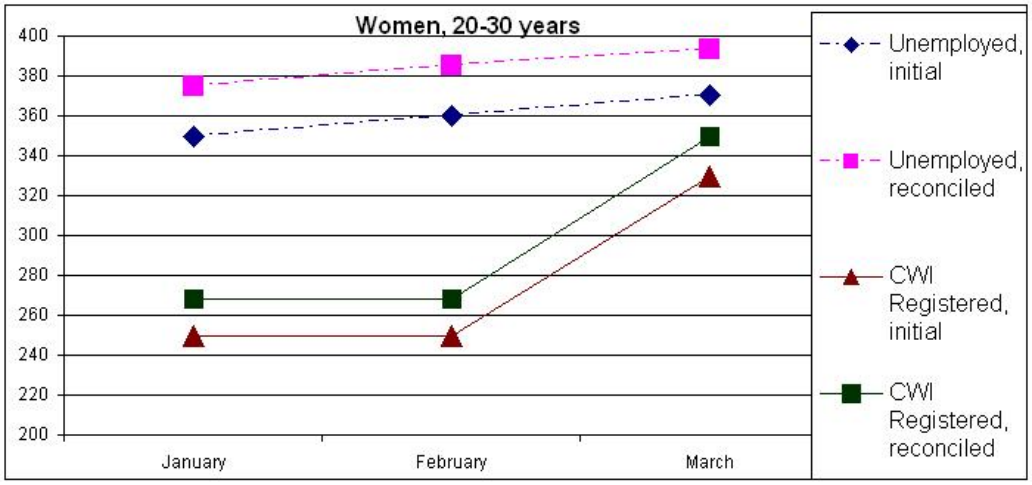


Figure 7.1. Women of 20-29 years; initial and reconciled unemployment and CWI Registered.

Table 7.13. Women of 20-29 years; ratio between unemployed and CWI registered persons

	January	February	March
Initial data	1.400	1.440	1.121
Reconciled data 1	1.400	1.436	1.125

and CWI registered persons. As a result, the initial quarterly changes are preserved better at the expense of the ratio between the number of unemployed and CWI registered persons, which becomes clear by comparing the results in Table 7.14 with the results in Table 7.13.

Table 7.14. Women of 20-29 year; ratio between unemployed and CWI registered persons, scenario 2

	January	February	March
Initial data	1.400	1.440	1.121
Reconciled data 2	1.396	1.432	1.130

This is of course only a simple example where we only have two data sources and did not take into account many issues that will occur while combining different sources. However the method can be extended further.

7.5 Conclusions

Reconciliation of tables on a macro level can be very effective, especially when a large number of constraints should be fulfilled. Combining data sources of different structures, on a macro level is often easier to handle than on a micro-level. When data are very large and many sources should be combined macro-integration could be the only technique that can be used. Macro-integration is also more versatile than (re-)weighting techniques using GREG-estimation in the sense that inequality constraints and soft constraints can be incorporated easily.

Bibliography

[1] Antal E., Tillé Y. (2011), *A Direct Bootstrap Method for Complex Sampling Designs from a Finite Population*, Journal of the American Statistical Association 106, pp. 534–543.

[2] Bakker B.F.M. (2011), *Micro integration: the state of the art*, in. Chapter 5 in: Report on WP1 of ESSnet on data integration.

[3] Ballin M., Di Zio M., D'Orazio M., Scanu M., Torelli N. (2008), *File concatenation of survey data: a computer assisted approach to sampling weights estimation*, Rivista di Statistica Ufficiale, Volume 2008/2-3, 5–12.

[4] Bethlehem J., Cobben F., Schouten B. (2011), *Handbook of Nonresponse in Household Surveys*, John Wiley&Sons, New Jersey.

[5] Bickel P. J., Freedman D. A. (1984), *Asymptotic Normality and the Bootstrap in Stratified Sampling*, The Annals of Statistics 12, pp. 470—482.

[6] Bikker R., Buijtenhek S. (2006), *Alignment of Quarterly Sector Accounts to annual data*, Statistics Netherlands.

[7] Bikker R., Daalmans J., Mushkudiani N. (2010), *A multivariate Denton method for benchmarking large data sets*, Statistics Netherlands.

[8] Bishop Y. M. M., S. E Feinberg., P.W. Holland (1975), *Discrete Multivariate Analysis*, MIT Press, Cambridge, Massachusetts.

[9] Boonstra H.J. (2004), *Calibration of tables of estimates*, Technical report, Statistics Netherlands.

[10] Boonstra H.J., Blois C. de., Linders G.J. (2010), *Macro-integration with inequality constraints-an application to the integration of transport and trade statistics*, Statistics Netherlands.

[11] Booth J. G., Butler R. W., Hall P. (1994), *Bootstrap Methods for Finite Populations*, Journal of the American Statistical Association 89, pp. 1282—1289.

[12] Canty A. J., Davison A. C. (1999), *Resampling-Based Variance Estimation for Labour Force Surveys*, The Statistician 48, pp. 379-–391.

[13] Cenzor Y., S.A. Zenios (1977), *Parallel Optimization. Theory, Algorithms, and Applications*, Oxford University Press, New York.

[14] Chao M.-T., Lo S.-H. (1985), *A Bootstrap Method for Finite Populations*, Sankhyā Series A 47, pp. 399—405.

[15] Chauvet G. (2007), *Méthodes de Bootstrap en Population Finie*, PhD Thesis, L'Université de Rennes.

[16] Chen J., Sitter R.R. (1999), *A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys*, Statistica Sinica, 9, 385–406.

[17] Coli A., Tartamella F., Sacco G., Faiella I., D'Orazio M., Di Zio M., Scanu M., Siciliani I., Colombini S., Masi A. (2006), *Construction of a microdata archive on Italian households through integration of Istat's household budget survey and Bank of Italy's survey on household income and wealth*, Documenti ISTAT, n.12/2006 (in Italian).

[18] De Heij V. (2011), *Samples and Registers*, Internal memo (in Dutch), Statistics Netherlands, The Hague.

[19] De Waal T., J. Pannekoek., S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*, John Wiley & Sons Inc., Hoboken, New Jersey.

[20] D'Orazio M., Di Zio M., Scanu M. (2006a), *Statistical matching for categorical data: Displaying uncertainty and using logical constraints*, JOS, Volume 22, 137–157.

[21] D'Orazio M., Di Zio M., Scanu M. (2006b), *Statistical matching: Theory and practice*, Wiley, Chichester.

[22] D'Orazio M., Di Zio M., Scanu M. (2010), *Old and new approaches in statistical matching when samples are drawn with complex survey designs*, Proceedings of the 45th Riunione Scientifica della Societŕ Italiana di Statistica, Padova 16–18 June 2010.

[23] EC (1993), *Council Regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community.*

[24] EC (2006), *Commission regulation (EC) No 1503/2006 of 28 September 2006. implementing and amending Council Regulation (EC) No 1165/98 concerning short-term statistics as regards definitions of variables, list of variables and frequency of data compilation.*

[25] EC (2008), *Regulation (EC) 177/2008 of the European Parliament and of the council of 20 February 2008 establishing a common framework for business registers for statistical purposes and repealing Council regulation (EEC) 2186/93.*

[26] Efron B., Tibshirani R. J. (1993), *An Introduction to the Bootstrap*, Chapman&Hall/CRC, London.

[27] EuroStat (2008), *STS requirements under NACE Rev. 2.*, 28 August 2008. STS requirements based on CR 1165/98, amended by CR 1158/2005.

[28] Fisher H., Oertel, J. (2009), *Konjunkturindikatoren im Dienstleistungsbereich: das mixmodell in der praxis Statistisches Bundesamt*, Wirtschaft und statistiek 3, 232–240.

[29] Fortini M., Liseo B., Nuccitelli A., Scanu M. (2001), *On Bayesian record linkage*, Research In Official Statistics, 4, 185–191.

[30] Gazzelloni S., Romano M.C., Corsetti G., Di Zio M., D'Orazio M., Pintaldi F., Scanu M., Torelli N., (2007), *Time use and Labour Force surveys: a proposal of data integration by means of statistical matching*, In: I tempi della vita quotidiana: un approccio multidisciplinare all'analisi dell'uso del tempo, (M. C. Romano ed.), collana Argomenti, n. 32, chap. 9, 375–403.

[31] Green P. J., Mardia K. V. (2006), *Bayesian alignment using hierarchical models, with application in protein bioinformatics*, Biometrika, 93, 235–254.

[32] Gnoss R (2010), *Powerpoint sheets with a method from Germal Federal Statistical Office Destatis*, as obtained from Ronald Gnoss.

[33] Gross S. (1980), *Median Estimation in Sample Surveys*, Proceedings of the Section on Survey Research Methods of the American Statistical Association pp. 181—184.

[34] Hoogland J. (2009), *Detection of potential influential errors in VAT turnover data used for short-term statistics*, Paper for the Work Session on Statistical Data Editing in Neuchâtel, Switzerland, 5–7 October 2009.

[35] Hoogland J. (2011), *Editing of mixed source data for turnover statistics*, Supporting paper for the Work Session on Statistical Data Editing in Ljubljana, Slovenia, 9-11 May 2011.

[36] Houbiers M.(2004), *Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands*, Journal of Official Statistics, 20, 55-75.

[37] Jaro M.A. (1989), *Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa*, Florida. Journal of the American Statistical Association, 84, 414–420.

[38] Kadane J.B. (1978), *Some statistical problems in merging data files*, In: Compendium of tax research, Department of Treasury, U.S. Government Printing Office, Washington StateD.C., 159-179 (1978). Reprinted in 2001, JOS, Volume 17, 423–433.

[39] Kiesl H., Raessler S.(2008), *The Validity of Data Fusion*, CENEX-ISAD workshop, Wien 29–30 May 2008. Available at http://cenex-isad.istat.it.

[40] Knottnerus P. (2003), *Sample Survey Theory: Some Pythagorean Perspectives*, Springer-Verlag, New York.

[41] Koskinen V. (2007), *The VAT data in short term business statistics*, Seminar on Registers in Statistics - methodology and quality 21–23 May, 2007 Helsinki.

[42] Kuijvenhoven L., Scholtus S. (2010), *Estimating Accuracy for Statistics Based on Register and Survey Data*, Discussion Paper 10007, Statistics Netherlands, The Hague.

[43] Lahiri P., Larsen M. D. (2005), *Regression analysis with linked data*, Journal of the American Statistical Association, 100, 222–230.

[44] Linder F., Van Roon D. (2011), *Combining Data from Administrative Sources and Sample Surveys; the Single-Variable Case. Case Study: Educational Attainment*, Report for Work Package 4.2 of the ESSnet project Data Integration.

[45] Luenberger D. G. (1984), *Linear and Nonlinear programming, second edition*, Addison-Wesley, Reading.

[46] Manski C.F.(1995), *Identification problems in the social sciences*,Harvard University Press, Cambridge, Massachusetts.

[47] McCarthy P. J., Snowden C. B. (1985), *The Bootstrap and Finite Population Sampling*, Technical Report, National Center for Health Statistics.

[48] Moriarity C., Scheuren F. (2001), *Statistical matching: a paradigm for assessing the uncertainty in the procedure*, JOS, Volume 17, 407–422.

[49] Moriarity C., Scheuren F. (2003), *A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation*, J. Bus. Econ. Stat., Volume 21, 65–73.

[50] Norberg J. (2005), *Turnover in other services*, Statistics Sweden, Final Technical implementation Report. 8 June 2005.

[51] Okner B.A. (1972), *Constructing a new data base from existing microdata sets: the 1966 merge file*, Ann. Econ. Soc. Meas., Volume 1, 325–342.

[52] Orchard C., Moore K., Langford A. (2010), *National practices of the use of administratice and accounts data in UK short term business statistics*, Deliverable of WP4 the ESSnet 'Use of administrative and accounts data in business statistics'.

[53] Orjala H. (2008), *Potential of administrative data in business statistics - a special focus in improvements in short term statistics*, IAOS Conference on Reshaping Official Statistics, Shanghai, 14–16 Octotber 2008.

[54] Pannekoek J (2011), *Models and algorithms for micro integration*, Chapter 6 in Report on WP2 (this volume).

[55] Raessler S. (2002) *Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches*, Springer Verlag, New York.

[56] Rao J. N. K., Wu C. F. J. (1988), *Resampling Inference with Complex Survey Data*, Journal of the American Statistical Association 83, pp. 231—241.

[57] Rao J.N.K, Wu C. (2008), *Empirical Likelihood Methods*, In: D. Pfeffermann and C.R. Rao (eds.) Handbook of Statistics, Vol. 29. Sample Surveys: Theory, Methods and Inference, pp. 189–207.

[58] Renssen R.H. (1998), *Use of statistical matching techniques in calibration estimation*, Survey Methodology, Volume 24, 171–183.

[59] Rubin D.B. (1986), *Statistical matching using file concatenation with adjusted weights and multiple imputations*, J. Bus. Econ. Stat., Volume 4, 87–94.

[60] Särndal C., Swensson B., Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

[61] Scheuren F., Winkler W.E. (1993), *Regression analysis of data files that are computer matched – Part I*, Survey Methodology, 19, 39–58.

[62] Shao J., Sitter R. R. (1996), *Bootstrap for Imputed Survey Data*, Journal of the American Statistical Association 91, pp. 1278—1288.

[63] Shao J., Tu D. (1995), *The Jackknife and Bootstrap*, Springer, New York.

[64] Seljak R. (2007), *Use of the tax data for the purposes of the short-term statistics*, Statistical Office of the Republic of Slovenia. Seminar on Registers in Statistics - methodology and quality 21–23 May, 2007 Helsinki.

[65] Sitter R. R. (1992a), *A Resampling Procedure for Complex Survey Data*, Journal of the American Statistical Association 87, pp. 755—765.

[66] Sitter R. R. (1992b), *Comparing Three Bootstrap Methods for Survey Data*, The Canadian Journal of Statistics 20, pp. 135—154. Statistics Finland (2009), *STS mini-workshop on 8th–9th of June 2009*, Meeting with experts from Statistics Netherlands, Statistics Sweden, UK

National Statistics, Statistics Estonia, Statistics Finland and National Board of Customs as part of the MEETS project.

[67] Stone J.R.N., Champerowne D.A., Maede J.E. (1942), *The Precision of National Income Accounting Estimates*, Reviews of Economic Studies 9, pp. 111–125.

[68] Tancredi A., Liseo B. (2005), *Bayesian inference for linked data*, Proceedings of the conference SCO 2006: 203–208.

[69] Tancredi A., Liseo B. (2011), *A hierarchical Bayesian approach to Record Linkage and size population problem*, Annals of Applied Statistics, in print.

[70] Thomson K., J. T. Fagan B. L. Yarbrough., D. L. Hambric (2005), *Using a Quadratic Programming Approach to Solve Simultaneous Ratio and Balance Edit Problems*, Working paper 32, UN/ECE Work Session on Statistical Data Editing, Ottawa.

[71] Torelli N., Ballin M., D'Orazio M., Di Zio M., Scanu M., Corsetti G. (2008), *Statistical matching of two surveys with a nonrandomly selected common subset* Workshop of the ESSnet-ISAD project, Vienna 29–30 May 2008.

[72] Vaasen A.M.V.J., Beuken, I.J. (2009), *New Enterprise Group delineation using tax information*, UNECE/Eurostat/OECD BR seminar, Luxembourg 6–7 October 2009 (Session 2).

[73] Van Delden A. (2010), *Methodological challenges using VAT for turnover estimates at Statistics Netherlands*, Conference on administrative simplification in official statistics (Simply2010), 2–3 December Ghent 2010, Belgium.

[74] Van Delden A., Bemmel K.H.J., (2011), *Handling incompleteness after linkage to a population frame: incoherence in unit types, variables and periods*, Chapter 4 in Report on WP2 (this volume).

[75] Van Delden A., de Wolf P.P., Banning R., Van Bemmel K.J.H., de Boer A.R., Carolina T., Hoogland J.J., Van der Loo M.P.J., Slootbeek M.H., Ouwehand P., Van der Velden H., (2010), *Methodology description DRT (In Dutch)*, Internal report, Statistics Netherlands.

[76] Van Delden A., Hoogland J. (2011), *Editing after data linkage*, Chapter 2 in Report on WP2 (this volume).

[77] Wagner I. (2004), *Schätsung fehlender Umsatzangaben für Organschaften im Unternemensregister*, Destatis, Statistisches Bundesambt, Wiesbaden.

[78] Wu C. (2004), *Combining information from multiple surveys through the empirical likelihood method*, Can. J. Stat., Volume 32, 1–12.

[79] Wu C. (2005), *Algorithms and R codes for the pseudo empirical likelihood method in survey sampling*, Survey Methodology, 31, 239–243.

[80] Wu C., J.N.K. Rao (2006), *Pseudo empirical likelihood ration confidence intervals for complex surveys*, The Canadian Journal of Statistics, 34, 359–375.

[81] Zhang L.C. (2011), *Topics of statistical theory for register-based statistics*, Paper for the ISI conference, Dublin 22–26 August 2011.

Appendix A

The next steps illustrate the use of the Liseo-Tancredi approach by means of R codes provided by the authors. As an example, the two data sets on newborns described before are linked: the A file corresponds to the P4 archive, while B corresponds to the CEDAP archive.

In order to apply the R codes, it is necessary to have the data sets already preprocessed and harmonized. Furthermore, blocks should already have been created. The following steps refer to the application of specific blocks of the two data sets: the P4 block that will be analyzed is named *p615.csv* while the CEDAP block is named *cedap615.csv* (blocks correspond to the newborns whose mother have Chinese nationality).

Both files have already been formatted so that the key variables are represented in the first three columns of each file. The first key variable represent the year of birth from 1964 to 1987. The other two key variables are month and day of birth, respectively.

This program starts reading the file function.R containing some other functions:

- gibbs.step=function(mu,n,k,ng,p,F,ftilde): this function simulates the distribution of the true values (μ),
- rpostbeta=function(shape1,shape2,lb): This function simulates the probability of error (β),
- rpostenne=function(na,nb,T,Ntronc): this function simulates the distribution of the real size of the population (N).

Furthermore it reads the two data files A and B
Note that the previous commands read only the first three columns of the two data sets. Different options could be used according to the nature of the files.

```
source("functions.R")
xA=read.csv("p615.csv")[,1:3]
xB=read.csv("cedap615.csv")[,1:3]
```

Coding key variables

Key variable categories should be coded with **integer numbers**, from 1 to k_i where k_i is the number of categories for the i-th key variable

```
xA[,1]=xA[,1]-1963
xB[,1]=xB[,1]-1963
k1=23
k2=12
k3=31
```

In this example, the first column represents a year (the birth year), starting from 1964. For this reason, the code has been created subtracting 1963 from the actual year. The whole set of the number of years contains 23 items (k1). The second and third variables are already coded with integer numbers starting from 1, hence only the number of categories should be declared (12 for the second - the month of birth - and 31 for the third - the day of birth).

Construction of the design matrix for the multivariate distribution of all the key variables

The matrix V should be read by row. Each row describes a cell of the multivariate contingency table of the key variables. It contains the Cartesian product of the categories of the key variables.

```
V=expand.grid(1:k3,1:k2,1:k1)[,c(3,2,1)]
```

Declaration on the number of units in the two files

Size of the first file is n_A; size of the second file is n_B; the number of key variables is h

```
na=nrow(xA); nb=nrow(xB); h=ncol(xA)
```

Data type format issues

This command transforms V columns in factor type objects This vector gives the number of categories in each key variable.

```
for (i in 1:h) V[,i]=as.factor(V[,i])
kappa=c()
for (i in 1:h) kappa[i]=nlevels(factor(V[,i])
```

Size of V

The number of cells in the key variables multivariate contingency table (i.e. the number of rows in V) is k

```
k=prod(kappa)
```

Some operational issues

The following commands allow to connect each unit observed in the two data files to the corresponding cell in V.

```
xa=as.numeric(apply(xA,FUN=category,MAR=1,kappa=kappa))
xb=as.numeric(apply(xB,FUN=category,MAR=1,kappa=kappa))
```

Fixing the initial parameters

The initial values for θ (probabilities of association in the superpopolation) and β (parameters describing the measurement error) are fixed with these command lines

```
theta.init=rep(1,k)
theta.init=theta.init/sum(theta.init)
beta.init=rep(0.9,h)
```

Analysis

The library MCMCpack analyzes the output and contains some of the main program functions {"B.CAT.matching.cenex.R"}

```
library(MCMCpack)
source("B.CAT.matching.cenex.R")
```

Arguments of the function B.CAT.matching

These are the arguments expected by the function B.CAT.matching

- xa = vector where every unit in file A is connected to the corresponding cell in V

- nMCMC=number of simulations
- burnin=number of initial simulation iterations to discard
- ng= number of updates of the matrix C in every iteration
- C.agg specifies if C should be computed or not (true e false respectively)
- beta.init=beta.init,theta.init=theta.init,mua.init=xa,mub.init=xb (initial values of the parameters)

This function computes the two matrices: "out" and "C.med". The matrix "out" reports the simulation of the model parameters but the matrix C. The matrix "C.med" gives the probability of being a link for each pair of units (a,b), $a \in A$ and $b \in B$.

Appendix B

Test of methodology to correct for frame errors in size class

To get an idea about the threshold value for L_ℓ and L_u, we conducted a small test with real data. We took VAT data of Q1 2008 – Q2 2010 (10 quarters) and selected the domestic, non-topX VAT units and computed their quarterly turnover.

For each unit within size class (sc) 0 and 1 we determined whether they get a newly imputed size class or whether they keep their original size class when we take L_u to be 5, 6, 7 or 8 and when we compute the median quarterly turnover of a size class at 2- or 3-digit NACE code.

Next, we considered the size class of a VAT unit to be wrong (i.e. a frame error) when its quarterly turnover was larger than a threshold value O_{tr}. We computed results for three threshold values, namely 1, 5 and 10 million euros. Based on the threshold, for each unit we determined whether its size class in the population frame was considered to be correct or incorrect.

Finally, for each quarter, we counted the number of units classified according to 'new size class' versus 'original size class' crossed by 'correct size class' versus 'incorrect size class'. We took the total outcome for 10 quarters. The results give an *indication* for false and true negatives and false and true positives at different values of L_ℓ and L_u with the median computed at two NACE levels.

Table B.1 shows an example of the test results for size class 0, with the median computed at 2-digit NACE code and L_ℓ=0. Table B.1 shows that the smaller L_u the more units were assigned to a new size class. Also, the smaller the value for O_{tr}, the more units were considered to be incorrect. Assuming that the number of size class errors in the frame is limited, we considered O_{tr} = 10 million euros to be most realistic value. When the median is computed

at 3-digit NACE level (not shown) the number of VAT units with a new size class is slightly smaller, but then we have more cases where we do not have the minimum of 10 units to compute the median.

We considered a false negative (cell 'old sc' × 'incorrect') to be much more severe than a false positive (cell 'new sc' × 'correct'), because assigning a new size class will probably lead to a reasonable imputation value. Based on the results we selected L_u=6: this leads to a limited number of false negatives and we avoid that the number of false positives becomes much larger than the number of true positives. In size class 1 (not shown) there was also a considerable number of units considered to be incorrect, namely 229 (O_{tr}=10 million euros), with only 1 false negative. We therefore selected L_ℓ=1.

Table B.1. Test of methodology to correct for errors in size class 0

	L_u=8		L_u=7		L_u=6		L_u=5	
	Incorrect[1]	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct
O_{tr}^{k-1}=10 × 10^6								
New sc	581	290	759	913	767	2776	767	7403
Old sc	191	470736	13	470113	5	468250	5	463623
Total	772	471026	772	471026	772	471026	772	471026
O_{tr}^{k-1}= 5 × 10^6								
New sc	753	118	1083	589	1358	2185	1389	6781
Old sc	651	470276	321	469805	46	468209	15	463613
Total	1404	470394	1404	470394	1404	470394	1404	470394
O_{tr}^{k-1}= 1 × 10^6								
New sc	869	2	1574	98	2541	1002	4328	3842
Old sc	4535	466392	3830	466296	2863	465392	1076	462552
Total	5404	466394	5404	466394	5404	466394	5404	466394

[1] Incorrect under certain assumptions, as explained in the text.

Appendix C

Derivation of Expressions for the Bias and Variance of $\hat{\theta}_{2y}$

We begin by evaluating the bias of $\hat{\theta}_{2y}$. From $\mathrm{E}(\hat{\theta}_{2y}) = \mathrm{E}_s[\mathrm{E}(\hat{\theta}_{2y} \mid s)]$ and expression (5.2), we obtain:

$$\begin{aligned}
\mathrm{E}(\hat{\theta}_{2y}) &= \mathrm{E}_s\Big[\mathrm{E}\big(\sum_{k\in\mathcal{U}_R} z_k + \sum_{k\in s_R}(y_k - z_k) + \sum_{k\in s_{NR}} w_{2k}y_k \;\big|\; s\big)\Big] \\
&= \mathrm{E}_s\Big[\sum_{k\in\mathcal{U}_R}\mathrm{E}(z_k) + \sum_{k\in s_R}\big(y_k - \mathrm{E}(z_k)\big) + \sum_{k\in s_{NR}} w_{2k}y_k\Big] \\
&= \mathrm{E}_s\Big[\sum_{k\in\mathcal{U}_R}(y_k + \lambda_k\mu_k) - \sum_{k\in s_R}\lambda_k\mu_k + \sum_{k\in s_{NR}} w_{2k}y_k\Big] \\
&= \sum_{k\in\mathcal{U}_R}(y_k + \lambda_k\mu_k) + \mathrm{E}_s\Big(-\sum_{k\in s_R}\frac{\pi_k\lambda_k\mu_k}{\pi_k} + \sum_{k\in s_{NR}} w_{2k}y_k\Big) \\
&\doteq \sum_{k\in\mathcal{U}_R}(y_k + \lambda_k\mu_k) - \sum_{k\in\mathcal{U}_R}\pi_k\lambda_k\mu_k + \sum_{k\in\mathcal{U}_{NR}} y_k \\
&= \theta_y + \sum_{k\in\mathcal{U}_R}(1-\pi_k)\lambda_k\mu_k,
\end{aligned}$$

In the second last line, it is used that $\sum_{k\in s_R} x_k/\pi_k$ is an unbiased estimator for $\sum_{k\in\mathcal{U}_R} x_k$, for any variable x. It is also used that $\sum_{k\in s_{NR}} w_{2k}y_k$ is an asymptotically unbiased estimator for $\sum_{k\in\mathcal{U}_{NR}} y_k$. Expression (5.6) now follows.

Next, we evaluate the variance of $\hat{\theta}_{2y}$ by means of the decomposition

$$\mathrm{V}(\hat{\theta}_{2y}) = \mathrm{E}_s[\mathrm{V}(\hat{\theta}_{2y} \mid s)] + \mathrm{V}_s[\mathrm{E}(\hat{\theta}_{2y} \mid s)].$$

Using the assumption that the z_k are independent, it follows from expression (5.3) that

$$\begin{aligned}
\mathrm{E}_s[\mathrm{V}(\hat{\theta}_{2y} \mid s)] &= \mathrm{E}_s\Big[\mathrm{V}\Big(\sum_{k\in\mathcal{U}_R\setminus s_R} z_k \,\Big|\, s\Big)\Big] \\
&= \mathrm{E}_s\Big[\sum_{k\in\mathcal{U}_R\setminus s_R} \mathrm{V}(z_k)\Big] \\
&= \mathrm{E}_s\Big[\sum_{k\in\mathcal{U}_R\setminus s_R} \lambda_k(\sigma_k^2 + \mu_k^2(1-\lambda_k))\Big] \\
&= \sum_{k\in\mathcal{U}_R} (1-\pi_k)\lambda_k[\sigma_k^2 + \mu_k^2(1-\lambda_k)]. \qquad \text{(C.1)}
\end{aligned}$$

The proof of the last line is analogous to the last four lines in the evaluation of $\mathrm{E}(\hat{\theta}_{2y})$.

For the second component, we find

$$\begin{aligned}
\mathrm{V}_s[\mathrm{E}(\hat{\theta}_{2y} \mid s)] &= \mathrm{V}_s\Big[\sum_{k\in\mathcal{U}_R} (y_k + \lambda_k\mu_k) - \sum_{k\in s_R} \lambda_k\mu_k + \sum_{k\in s_{NR}} w_{2k}y_k\Big] \\
&= \mathrm{V}_s\Big(-\sum_{k\in s_R} \lambda_k\mu_k + \sum_{k\in s_{NR}} w_{2k}y_k\Big). \qquad \text{(C.2)}
\end{aligned}$$

Combining (C.1) and (C.2) yields expression (5.7).

It is interesting to examine expression (C.2) in more detail. The weights w_{2k} have been found by fitting a regression model to the observations from $\mathcal{U}_{NR}$, say, $y_k = \boldsymbol{\beta}_2'\mathbf{x}_{2k} + \varepsilon_{2k}$. Denote the vector of fitted regression coefficients by $\hat{\boldsymbol{\beta}}_2$.

By a standard argument, it holds that

$$\begin{aligned}
\sum_{k\in s_{NR}} w_{2k}y_k &= \sum_{k\in s_{NR}} \frac{y_k}{\pi_k} + \hat{\boldsymbol{\beta}}_2'\Big(\sum_{k\in\mathcal{U}_{NR}} \mathbf{x}_{2k} - \sum_{k\in s_{NR}} \frac{\mathbf{x}_{2k}}{\pi_k}\Big) \\
&\doteq \sum_{k\in s_{NR}} \frac{y_k}{\pi_k} + \boldsymbol{\beta}_2'\Big(\sum_{k\in\mathcal{U}_{NR}} \mathbf{x}_{2k} - \sum_{k\in s_{NR}} \frac{\mathbf{x}_{2k}}{\pi_k}\Big) \\
&= \boldsymbol{\beta}_2' \sum_{k\in\mathcal{U}_{NR}} \mathbf{x}_{2k} + \sum_{k\in s_{NR}} \frac{\varepsilon_{2k}}{\pi_k},
\end{aligned}$$

since the discarded term $(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'\Big(\sum_{k\in\mathcal{U}_{NR}} \mathbf{x}_{2k} - \sum_{k\in s_{NR}} \frac{\mathbf{x}_{2k}}{\pi_k}\Big)$ is asymptotically irrelevant. Hence, for sufficiently large samples, we have

$$\begin{aligned}
\mathrm{V}_s\Big(-\sum_{k\in s_R}\lambda_k\mu_k+\sum_{k\in s_{NR}}w_{2k}y_k\Big) &\doteq \mathrm{V}_s\Big(-\sum_{k\in s_R}\lambda_k\mu_k+\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big)\\
&= \mathrm{V}_s\Big(\sum_{k\in s_R}\lambda_k\mu_k\Big)+\mathrm{V}_s\Big(\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big)\\
&\quad -2\mathrm{Cov}_s\Big(\sum_{k\in s_R}\lambda_k\mu_k,\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big).
\end{aligned}\tag{C.3}$$

Note that $\lambda_k\mu_k$ is only defined for $k\in\mathcal{U}_R$, while ε_{2k} is only defined for $k\in\mathcal{U}_{NR}$. For convenience, define $\lambda_k\mu_k=0$ for $k\in\mathcal{U}_{NR}$, and define $\varepsilon_{2k}=0$ for $k\in\mathcal{U}_R$. The first variance term may now be evaluated as follows:

$$\begin{aligned}
\mathrm{V}_s\Big(\sum_{k\in s_R}\lambda_k\mu_k\Big) &= \mathrm{V}_s\Big(\sum_{k\in s}\frac{\pi_k\lambda_k\mu_k}{\pi_k}\Big)\\
&= \sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}}(\pi_{kl}-\pi_k\pi_l)\frac{\pi_k\lambda_k\mu_k}{\pi_k}\frac{\pi_l\lambda_l\mu_l}{\pi_l}\\
&= \sum_{k\in\mathcal{U}_R}\sum_{l\in\mathcal{U}_R}(\pi_{kl}-\pi_k\pi_l)\lambda_k\mu_k\lambda_l\mu_l,
\end{aligned}$$

where we have used a standard formula for the variance of a Horvitz-Thompson estimator; see e.g. Särndal et al. (1992, p. 43). Similarly, the second variance term yields

$$\mathrm{V}_s\Big(\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big)=\sum_{k\in\mathcal{U}_{NR}}\sum_{l\in\mathcal{U}_{NR}}(\pi_{kl}-\pi_k\pi_l)\frac{\varepsilon_{2k}}{\pi_k}\frac{\varepsilon_{2l}}{\pi_l}.$$

Finally, the covariance term may be evaluated as follows:

$$\begin{aligned}
\mathrm{Cov}_s\Big(\sum_{k\in s_R}\lambda_k\mu_k,\sum_{k\in s_{NR}}\frac{\varepsilon_{2k}}{\pi_k}\Big) &= \mathrm{Cov}_s\Big(\sum_{k\in s}\frac{\pi_k\lambda_k\mu_k}{\pi_k},\sum_{k\in s}\frac{\varepsilon_{2k}}{\pi_k}\Big)\\
&= \sum_{k\in\mathcal{U}}\sum_{l\in\mathcal{U}}(\pi_{kl}-\pi_k\pi_l)\frac{\pi_k\lambda_k\mu_k}{\pi_k}\frac{\varepsilon_{2l}}{\pi_l}\\
&= \sum_{k\in\mathcal{U}_R}\sum_{l\in\mathcal{U}_{NR}}(\pi_{kl}-\pi_k\pi_l)\lambda_k\mu_k\frac{\varepsilon_{2l}}{\pi_l}.
\end{aligned}\tag{C.4}$$

In the second last line, use is made of a standard formula for the covariance of two Horvitz-Thompson estimators; see e.g. Särndal et al. (1992, p. 170).

In general, expression (C.4) may be non-zero. There exist, however, a few special cases where the covariance term always vanishes. For simple random sampling, we have $\pi_k = \pi_l = \frac{n}{N}$, $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$ (for $k \neq l$), and hence

$$\sum_{k \in \mathcal{U}_R} \sum_{l \in \mathcal{U}_{NR}} (\pi_{kl} - \pi_k \pi_l) \lambda_k \mu_k \frac{\varepsilon_{2l}}{\pi_l} = \left[\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \right] \sum_{k \in \mathcal{U}_R} \lambda_k \mu_k \frac{N}{n} \sum_{l \in \mathcal{U}_{NR}} \varepsilon_{2l} = 0,$$

since the sum of the residuals over $\mathcal{U}_{NR}$ equals zero by construction. Similarly, the covariance term also vanishes for stratified simple random sampling, provided that a separate regression model is fitted for each stratum.

Appendix D

The successive projection algorithm

The optimization problem (6.1) can be solved explicitly if the objective function is the (weighted) least squares function and there are only equality constraints. For other convex functions and/or inequality constraints, problem (6.1) can be solved by several optimization algorithms. In this section we briefly review a very simple such algorithm that is easy to implement and contains as a special case the – among survey methodologists well known – Iterative Proportional Fitting (IPF) algorithm for adjusting contingency tables to new margins (see, e.g Bishop et al., 1975). Algorithms of this type are extensively discussed in Censor and Zenios (1997) and applications to adjustment problems are described in De Waal et al. (2011). The algorithm is an iterative procedure in which the edit-constraints are used one at a time. It starts by minimally adjusting the original inconsistent vector x_0 to one of the constraints. The resulting solution is then updated such that a next constraint is satisfied and the difference with the previous solution is minimized and so on. In this way, if there are K constraints, K minimal adjustment problems with a single constraint each need to be solved, which is much easier than a simultaneous approach. After the K-steps one iteration is completed and a next iteration starts that will again sequentially adjust the current solution to satisfy each of the constraints.

To describe this algorithm we must make the distinction between adjustable variables and fixed variables more explicit. Without loss of generality we can separate the adjustable variables from the fixed values in $\mathbf{x}$ by the partitioning $x = (x_m^Y, x_o^T)^T$ where x_m denotes the subvector of $\mathbf{x}$ containing the adjustable values and x_o the subvector containing the remaining, fixed values. The restriction matrix $\mathbf{A}$ can then be partitioned conformably as $A = (A_m, A_o)$. From $Ax \leq 0$ we then obtain as constraints for the adjustable variables: $A_m x_m \leq -A_o x_o = b$, say.

In an iteration t the algorithm cycles through the constraints adjusting the current $\mathbf{x}$-vector to each of them. For equality constraints this adjustment solves the minimization problem

$$\begin{aligned} x_m^{t,k} &= \arg\min_x D(x, x_m^{t,k-1}) \quad \text{s.t. } a_{m,k} x_m^{t,k} - b_k = 0 \\ &= P_{D,k}(x_m^{t,k-1}) \ , \end{aligned}$$

with $a_{m,k}$ the row of A_m corresponding to constraint k and b_k the corresponding element of $\mathbf{b}$. The equality constraint $a_{m,k} x_m^{t,k} = b_k$ defines a hyperplane and $x_m^{t,k}$ is the vector on this hyperplane closest to $x_m^{t,k-1}$. Therefore, it is the (generalized) projection with respect to the distance D of $x_m^{t,k-1}$ on this hyperplane, which is denoted above by $P_{D,k}(x_m^{t,k-1})$. For the least-squares criterion this is the usual orthogonal Euclidean projection. As the algorithm cycles through the constraints the $\mathbf{x}_m$-vector is projected successively on each of the corresponding hyperplanes and converges to the solution which is on the intersection of these hyperplanes.

For the least squares criterion the solution of this projection step is given by

$$x_m^{t,k} = x_m^{t,k-1} + \bar{r}^{t,k} a_{m,k}^T \text{ with } \bar{r}^{t,k} = (b_k - a_{m,k} x_m^{t,k-1}) / (a_{m,k} a_{m,k}^T).$$

Note that $\bar{r}^{t,k}$ is a kind of "average" residual for constraint k since, if the values of $a_{m,k}$ are confined to 0, 1 or - 1, then $a_{m,k} a_{m,k}^T$ is the number of adjustable values in constraint k.

For the KL-criterion, the projection cannot be expressed in closed form for general a_k. However, if the elements of this vector are all either zero or one, which occurs when the constraint is that a sum of x_m-values equals a fixed value b_k, the adjustment to an equality constraint k can be expressed as

$$\begin{aligned} x_{m,i}^{t,k} &= x_{m,i}^{t,k-1} \rho^{t,k} \quad \text{if } a_{m,k,i} = 1 \\ &= x_{m,i}^{t,k-1} \quad \text{if } a_{m,k,i} = 0 \end{aligned}$$

where the adjustment factor $\rho^{t,k}$ is given by the rate of violation of constraint k: $\rho^{t,k} = b_k / (\sum_i a_{m,k,i} x_{m,i})$. In this case the resulting algorithm is equivalent to the IPF-algorithm that, when applied to a rectangular contingency table, adjust the counts in the table to new row- and column-totals by multiplying, successively, the counts in each row by a factor such that they add up to the new row-total and similarly for the columns.

For inequality constraints, the constraint can be satisfied with "slack", i.e. $a_{m,k} x_m^{t,k}$ is strictly smaller than b_k. In that case it may be possible to improve on the objective function by removing (some of) the adjustment to

constraint k to the extend that either all adjustment is removed or the constraint becomes satisfied with equality. To accomplish this we first undo the adjustment made in the previous iteration to this constraint. If the constraint becomes violated, the projection step is performed with the result that the constraint becomes satisfied with equality, which is the minimum feasible adjustment. If after undoing the previous adjustment the constraint is not violated, no adjustment is performed.

Appendix E

Census data example

In this section we will construct a simple hypercube using two data sources. Consider two data sets: one is obtained from GBA (municipality data bases) register and the other is from LFS (labor forse survey). The first data set consists of three variables: Province, Sex and Age and the second data set contains one additional variable: Occupation.

Table E.1. Categories of variable Province

Unknown	1
Groningen	2
Friesland	3
Drenthe	4
Overijssel	5
Flevoland	6
Gelderland	7
Utrecht	8
Noord-Holland	9
Zuid-Holland	10
Zeeland	11
Noord-Brabant	12
Limburg	13

For simplicity assume that the three common variables have the same categories in both data sets. Province has 13 categories, see Table E.1. The variable age is grouped in five year intervals and has 21 categories: $0- <5, 5- <10, ..., 95- <100, 100+$. Sex has 2 categories and occupation 12 categories, see Table E.2.
The data are initially available on the micro level. The total number of GBA persons is $N_{GBA} = 16\,408\,487$ and the total number of LFS persons is

Table E.2. Categories of variable occupation

Not stated	1
Armed forces occupations	2
Managers	3
Professionals	4
Technicians and associate professionals	5
Clerical support workers	6
Service and sales workers	7
Skilled agricultural, forestry, and fishery workers	8
Craft and related trades workers	9
Plant and machine operators, and assemblers	10
Elementary occupations	11
Not applicable	12

$N_{LFS} = 104\,674$. Both data sets were aggregated up to the publication level. The cross tables obtained are three and four dimensional hypercubes. The values of hypercube obtained from the second sample is then adjusted using the same weights for each cell. The initial weight is then defined as follows:

$$w = \frac{16\,408\,487}{104\,674}.$$

We assume that the figures of the first data set (obtained from the GBA) are exogenous. That means these values will not be changed.

Suppose that in the variables defined by $x_i^{(j)}$ the subindex i will define the identity of the variable for example Province and the super index will define the data set where the variable will originate from. In our example we have two data sets, hence $j = 1$ or 2. For convenience, the variables Province, Sex and Age are numbered by 1, 2 and 3. In the first data set these variables are defined by $x_1^{(1)}, x_2^{(1)}$ and $x_3^{(1)}$. Similarly, in the second data set the variables Province, Sex, Age and Occupation are defined as $x_1^{(2)}, x_2^{(2)}$, $x_3^{(2)}$ and $x_4^{(2)}$. We define the marginal distribution of the variable $x_i^{(j)}$ as follows:

$$x_{i,1}^{(j)}, \ldots, x_{i,r_i}^{(j)},$$

the second index here defines the categories of the variable. For example, the variable Province x_1 has 13 categories, $r_1 = 13$.
Each hypercube will have a crosstable of variables, containing the values

$$x_{1,j}^{(1)} \times x_{2,k}^{(1)} \times x_{3,l}^{(1)}, \quad j = 1, \ldots, 13, \quad k = 1, 2, \quad l = 1, \ldots, 21.$$

For example, when $j = 2$, $k = 2$ and $l = 8$ we have that

$$x_{1,2}^{(1)} \times x_{2,2}^{(1)} \times x_{3,8}^{(1)} = 20422$$

Table E.3. A part of the second hypercube

Province	Sex	age	Occupation	Number of persons
2	2	8	12	51
2	2	8	3	12
2	2	8	4	22
2	2	8	5	23
2	2	8	6	22
2	2	8	7	18
2	2	8	8	1
2	2	8	9	2
2	2	8	10	1
2	2	8	11	9

this means that there live 20422 women of age between 35 and 40 in the province Groningen. In the second data set we also have the extra variable Occupation. In case when $j = 2$, $k = 2$ and $l = 8$ the number of persons in each category of the variable Occupation are presented in Table E.3. Note that it is the part of the hypercube consisting of four variables. Observe that there are no persons in this hypercube with the categories 1 and 2 for the variable Occupation.

$$x^{(2)}_{1,2} \times x^{(2)}_{2,2} \times x^{(2)}_{3,8} \times \sum_{i=1}^{12} x^{(2)}_{4,i} = 161$$

We want to combine these two data sets into one. We can do this using the macro-integration method. For the simple example it is similar to post stratification methods. However, for the complete model, when we will have to make more than 50 hypercubes consistent with each other, the macro integration method is easier to generalize.

The reconciliation problem is defined as follows: We have variables $x^{(1)}_1, x^{(1)}_2$ and $x^{(1)}_3$ and $x^{(2)}_1, x^{(2)}_2$, $x^{(2)}_3$ and $x^{(2)}_4$. We want to find the estimates $\widehat{x}^{(2)}_1, \widehat{x}^{(2)}_2$, $\widehat{x}^{(2)}_3, \widehat{x}^{(2)}_4$ of $x^{(2)}_1, x^{(2)}_2$, $x^{(2)}_3$ and $x^{(2)}_4$, such that:

$$\sum_{k,l,h,i} \left(\widehat{x}^{(2)}_{1,k} \times \widehat{x}^{(2)}_{2,l} \times \widehat{x}^{(2)}_{3,h} \times \widehat{x}^{(2)}_{4,i} - x^{(2)}_{1,k} \times x^{(2)}_{2,l} \times x^{(2)}_{3,h} \times x^{(2)}_{4,i} \right)^2 \tag{E.1}$$

is minimized, under the restriction that the marginal distributions of the same variables from the sets 1 and 2 are the same:

$$(\widehat{x}^{(2)}_{i,1}, \ldots, \widehat{x}^{(2)}_{i,r_i}) = (x^{(1)}_{i,1}, \ldots, x^{(1)}_{i,r_i}), \quad \text{for} \quad i = 1, 2, 3. \tag{E.2}$$

Here we only require that the estimates $\widehat{x}_1^{(2)}, \widehat{x}_2^{(2)}, \widehat{x}_3^{(2)}, \widehat{x}_4^{(2)}$ should be as close as possible to the original values for each cell of the hypercube and the marginal distributions of the first three variables should be equal to the marginal distributions of these variables obtained from the first hypercube (register data).

We could make the set of restrictions heavier if we would add the restriction on the marginal distribution of the fourth variable to (E.2);

$$(\widehat{x}_{4,1}^{(2)}, \ldots, \widehat{x}_{4,r_4}^{(2)}) = (x_{4,1}^{(2)}, \ldots, x_{4,r_4}^{(2)}). \tag{E.3}$$

By this restriction we want to keep the marginal distribution of the variable occupation as it was observed in LFS.