Research on edit and imputation methodology: the throughput programme

Jeroen Pannekoek

Discussion paper (09022)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2007-2008	= 2007 to 2008 inclusive
2007/2008	= average of 2007 up to and including 2008
2007/'08	= crop year, financial year, school year etc. beginning in 2007 and ending in 2008
2005/'06–2007/'08	= crop year, financial year, etc. 2005/'06 to 2007/'08 inclusive

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

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Facility Services

Cover
TelDesign, Rotterdam

Information
Telephone .. +31 88 570 70 70
Telefax .. +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax .. +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008309022 X-10

Research on edit and imputation methodology: the throughput programme

Jeroen Pannekoek

Summary: One of the methodological research programmes of the Division of Methodology and Quality, is the "throughput-programme". This research programme focuses on methodology for the statistical process between input (data gathering and storage) and output (estimation, analysis and publication). This "throughput process" consists of procedures for the detection of errors, correction of errors and imputation (filling in estimates) of missing values. These throughput procedures are also referred to as editing and imputation (E&I) procedures. The goal of our throughput research programme is to improve the E&I procedures both in cost-effectiveness and in quality. The way this can be accomplished is by developing new or improved E&I methods based on mathematical models or algorithms, that can be implemented in software and applied automatically.

Keywords: Throughput-programme, Edit and imputation methodology, automatic E&I procedures.

Contents

1. Introduction

Users of statistical information are nowadays demanding high quality data on aspects of society with a great level of detail and produced within a short span of time. National statistical institutes (NSIs) fulfil a central role in providing such high quality statistical information. A major complicating factor is that the data sources, both traditional surveys as well as administrative data inevitably contain errors and missing values. In order to prevent substantial biases and inconsistencies in publication figures, NSIs carry out an extensive process of checking the collected data and correcting them if necessary.

One of the methodological research programmes of the Division of Methodology and Quality, the "throughput-programme" is especially targeted to this checking and correction process. This research programme focuses on methodology for the statistical process between input (data gathering and storage) and output (estimation, analysis and publication). The goal of the throughput process is to transform, as far as possible, raw input data with missing values and errors to complete and errorless (clean) statistical data, ready for the estimation of publication figures and secondary analysis. In practice, the errorless "true" data cannot be retrieved but a well designed throughput process strives to approximate these true data as close as possible with methods and procedures that are as cost-effective as possible.

Such a throughput process consists of procedures for the detection of errors, correction of errors and imputation (filling in estimates) of missing values. The detection and correction of errors is also commonly called "editing" and the throughput procedures are also referred to as editing and imputation (E&I) procedures.

The goal of our throughput research programme is to improve the methodology of E&I procedures both in cost-effectiveness and in quality. The way this can be accomplished is by developing new or improved E&I methods based on mathematical models or algorithms, that can be implemented in software and applied automatically.

This paper describes our current and intended research projects in this area. The paper is organized as follows. To give some background for the research projects of the throughput programme, and for reference in the remainder of the report, a short overview of the statistical E&I processes is given in section 2. This section begins by introducing edit rules. These rules play a crucial role in both the detection and the correction of errors since they describe when (combinations of) values of variables are incorrect. Next, the overall E&I process is presented as a sequence of sub-processes in which different error detection and correction methods are applied. The third section describes 6 research projects on E&I methodology that are currently being carried out or planned to be started during the next year.

2. Overview of the edit and imputation process.

2.1 Edit rules

Edit rules (or edits for short) are constraints on the variables defining the admissible values and combinations of values of the variables in each record. Records must satisfy all edit rules in order to qualify as a valid record. Records that violate one or more edit rules are considered inconsistent and it is deduced that some variable(s) in such a record must be in error (at least in case of fatal edits, as is explained below).

Common types of edits for business data are expressed as restrictions on the range of variables, as linear equations or inequalities in two or more variables or as bounds on ratios of variables. Furthermore, edits are often divided into hard or fatal edits and soft or query edits. Examples of these different edit types are given below.

An example of an edit in the form of a linear equation is

total turnover = total profit + total costs, (1.1)

stating that the turnover of an enterprise should equal the sum of its profit and costs. Such equalities are referred to as *balance edits*. In business statistics there are many balance edits and they are often connected by a common variable, resulting in a hierarchical system of edit rules. For example, the *total costs* variable in the edit rule above can itself be expressed as a sum:

total costs = personnel costs + capital costs + transport costs + other costs (1.2)

and, *personnel costs* can again be broken down as

personnel costs = salary costs + insurance costs + training costs + other costs. (1.3)

Many variables both in business and in social statistics must be non-negative (turnover, number of employees, number of children and age, for example) leading to so-called non-negativity edits expressed as $x \geq 0$. Inequalities between variables also occur as edits. For example, each of the component variables that sum up to a total variable, such as in the breakdown of total costs, cannot be larger than the total.

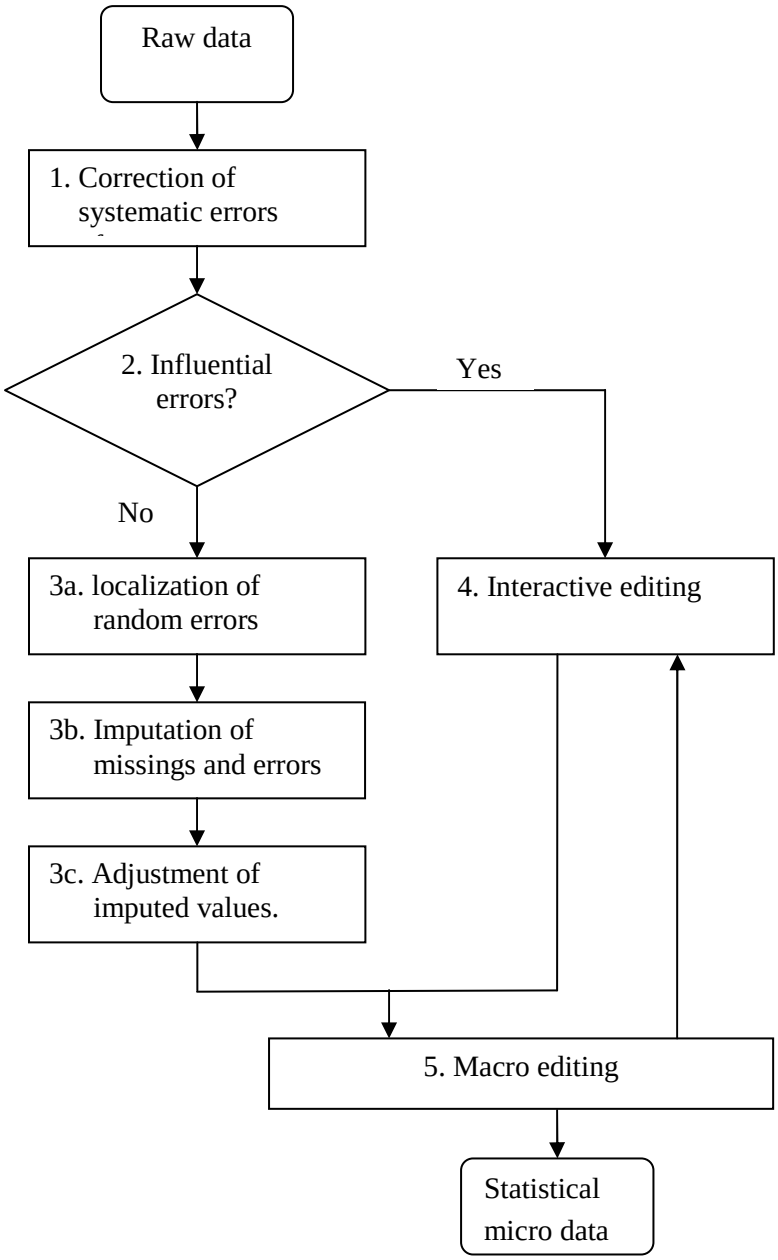
The equality and inequality edits exemplified above are called *fatal* or *hard* edits: they must hold true for a correct record. This class of edits is opposed to the so-called *non-fatal* or *soft* edits whose violation points to highly unlikely or anomalous (combinations of) values that are suspect to be in error although this is not a logical necessity. Many soft edits take the form of what is known as a “ratio edit”.

A ratio edit specifies that the ratio of two variables should be within certain bounds, for example the ratio of turnover and the number of employees. Ratio edits are often used to compare data on the same units from different sources, such as values reported in the current survey with values for the same variables reported in last year’s survey or values of variables from a tax register with similarly defined variables from a survey.

2.2 Process steps

The overall E&I process can be broken down into a number of sub-processes as shown in figure 1. An explanation of these different sub-processes follows below.

Figure 1. Breakdown of the throughput process in sub-processes



Step 1. Correction of systematic errors.

A systematic error is an error that occurs frequently between responding units. This type of error can occur when a respondent misunderstands or misreads a survey question. A well-known type of systematic error is the so-called unity measure error which is the error of, for example, reporting financial amounts in Euros instead of the requested thousands of Euros. Systematic errors can lead to substantial biases in

aggregates. Once detected, systematic errors can easily be corrected because the underlying error mechanism is known. Systematic errors that can be detected with very high accuracy by a specific search algorithm are called “obvious” errors. Detection and correction of obvious systematic errors is an important first step in an editing process. It can be done automatically and reliably at virtual no costs and hence will improve both the efficiency and the quality of the E&I process.

Step 2. Selection

Errors that cannot be resolved in the previous step will be taken care of either manually, by subject-matter specialists, or automatically by specialized edit and imputation algorithms. Manual editing is very time consuming and therefore expensive and adversely influencing the timeliness of publications. Moreover, when manual editing involves re-contacting the respondents, it will also increase the response burden. Therefore, SN, as most other statistical offices, has adopted selective editing strategies. This means that only records that potentially contain influential errors are edited manually, whereas the remaining records are edited automatically. In this way, manual editing is limited to those errors where the editing has substantial influence on publications figures, Hidiroglou and Berthelot (1986), Granquist (1995), Hoogland (2006).

The extend to which a record potentially contains influential errors is measured by a *score function,* Farwell and Raine (2000). This function is constructed such that records with high scores likely contain errors that have substantial effects on estimates of target parameters. For the selection step, a threshold value for the score has been set and all records with scores above this threshold are directed to manual reviewers whereas records with scores below the threshold are treated automatically.

Step 3a. Localization of erroneous values (random errors)

The first step in the automatic treatment of errors is the localization of errors. Once the fatal edits defined and implemented, it is straightforward to check whether the values in a record are inconsistent in the sense that some of these edits are violated. It is, however, not so obvious how to decide which variables in an inconsistent record are in error. This activity is called error localization. SN, as most other statistical offices, uses the paradigm of Fellegi and Holt (1976) as a guiding principle to locate the errors that cause violation of edit rules. This paradigm reads “the data in each record should be made to satisfy all edits by changing the fewest possible items (fields)”. Using this principle, a number of fields equal to the minimum required number is designated to be erroneous, such that there exists values for these fields that satisfy all edit rules. Often this principle leads to many solutions with the same minimum number of fields to change. This situation can be improved by generalizing the rule to: “the data in a record should be made to satisfy all edits by changing the fewest possible weighted number of fields”. Here each variable in a record is given a weight, the so-called reliability weight of this variable. A reliability weight is a measure of confidence in the value of the variable, the higher the weight, the more reliable the variable is considered to be. Currently, subject-matter specialists fix the reliability weights in advance. Algorithms that

locate errors based on this principle have been developed and implemented at SN in a software package called SLICE, see De Waal (2002, 2003) and De Waal and Quere (2003).

Step 3b. Imputation

In imputation, predictions from (parametric) models are derived for values that are missing or flagged as erroneous. An imputation model predicts a missing value using a function of auxiliary variables, the predictors. The auxiliary variables may be obtained from the current survey or from other sources such as historical information (the value of the missing variable in a previous period) or, increasingly important, administrative data. The most common types of imputation models are variants of regression models with parameters estimated from the observed data. However, especially for categorical variables, donor methods are also frequently used. Donor methods replace missing values in a record with the corresponding values from a complete and valid record. Often a record is chosen such that it resembles as much as possible the record with missing values.

Step 3c. Consistency adjustment of imputed values

In most cases, the edit rules are not taken into account by the imputation methods. As a consequence, the imputed records are in general inconsistent with the edit rules. This problem is currently solved by the introduction of an adjustment step in which adjustments are made to the imputed values such that the records satisfies all edit rules and the adjustments are as small as possible. This problem is formulated as a linear programming problem, minimizing the sum of the absolute values of the adjustments under the constraint that the resulting adjusted imputations satisfy the edit rules. An algorithm to solve this problem is implemented in SLICE.

Step 4. Interactive editing

Substantial mistakes by somewhat larger enterprises that can have an appreciable influence on publication aggregates are not considered suitable for the generic procedures of automatic editing. These records are treated by subject-matter specialists that can make use of specific information such as the scanned version of the original questionnaire, values of a previous version of the survey for the same respondent, stratum means or medians and other indicators. These editors use interactive software like the Blaise system developed at SN that can check the prespecified edit rules. If necessary the data may immediately be corrected using the available information and the subject-matter knowledge. This process is called interactive editing.

Step 5. Macro-editing

The steps considered so far all used micro-editing methods, i.e. methods that use the data of a single record and related auxiliary information to check and correct it. Micro-editing processes can be conducted from the start of the data collection phase, as soon as records become available. In contrast, macro-editing techniques use information from other records and can only be applied if a substantial part of the data is collected. Macro-editing techniques are also selective editing techniques in

the sense that they aim to direct the attention only to possibly influential erroneous values. An important macro-editing technique simply consists of verifying whether figures to be published seem plausible (Granquist, 1995). This can be accomplished by comparing quantities in publication tables with the same quantities in previous publications, with quantities based on register data, or with related quantities from other sources. Only if an unusual aggregate quantity is observed, a micro-editing procedure is applied to the individual records and fields that contribute to it. Since the majority of the records is available, it is possible for this micro-editing procedure to use graphical or model based outlier detection methods to identify possible anomalous and suspect values in the micro data.

The process flow suggested in figure 1 resembles very much the process flow in SNs Structural Business Surveys. For other surveys the process flow can be different, not all steps are always carried out and also the order of steps may be different. For social surveys, selective editing is not very important because the contributions of individuals to a publication total are not so much different, contrary to contributions of small and large enterprises in business surveys. Consequently, there is less need for manual editing of influential records and step 4 need not be performed. Often, in social surveys, due to a lack of hard edit rules the main type of detectable error is the missing value and process steps 3a and 3c are not performed either. For administrative data the collection of all records or a large part of it, is often available at once. This is different from the situation for surveys where the data are collected during a period of time. For administrative data it is therefore possible to form preliminary estimates immediately and to start with macro-editing as a tool for selective editing, and a process could start with step 1, followed by step 5 then step 4 and possibly step 3.

3. Research projects

3.1 Overview of research projects

In this section, 6 research projects are described that all aim at further developing E&I methodology in order to improve the efficiency and transparency of the E&I process and the quality of the resulting statistical micro data. The projects are targeted at the automated E&I procedures, since improvements in the automated methods will not only enhance the data quality but will also reduce the need for manual editing. In particular, errors that are resolved automatically in step 1 will never arrive at the manual editing step. Furthermore, if the automatic procedures in step 3 are improved, a larger proportion of the data can be entrusted to these automated procedures and less records need to be processed by the manual step 4.

The projects described in this paper are:

Correction of errors with detectable cause. This project concerns the methods applied in step 1 and seeks to improve and extend the current methodology for automatic detection and correction of systematic and obvious errors.

Model assisted localization of random errors. This project is related to process step 3a. The aim is to improve on the Fellegi-Holt methodology for localization of erroneous fields by supplementing this methodology with model-based outlier detection techniques to uniquely single out erroneous fields.

Imputation of complex economic data. The purpose of this project is to improve the accuracy of imputations for business surveys by developing models that incorporate the information contained in edit rules, especially the systems of connected balance edits as described in (1.1)-(1.3). These methods will be applied in step 3b but are also relevant for step 3c since the imputed values are intended to already satisfy all or a major part of the edit rules.

Calibrated imputation. In this project imputation methods will be developed that not only satisfy edit rules but also the constraint that the total of (some of) the imputed variables should be equal to known population totals. This last constraint occurs in situations where estimates of population parameters are made based on a combination of data from registers and surveys.

Improving imputation by estimated response probabilities. In this project the strength and weaknesses of imputation and weighting as methods to correct for missing data are investigated with the purpose to combine the approaches to arrive at improved imputation models.

Indicators for the effects of E&I on the quality of data and estimates. The purpose of the E&I process is to enhance the quality of the micro data and resulting estimates of population parameters. In order to evaluate the effectiveness of the E&I process standard evaluation measures, or indicators for each process step will be developed. This will include measures for the accuracy of estimates based on imputed values.

3.2 Deductive correction of errors with detectable cause

In the first phase of the E&I process, systematic errors for which the error mechanism is (assumed to be) known, are corrected. The classical example of such an error is the thousand-error mentioned in section 2.1. The occurrence of this error and ways to detect it are frequently described in the literature (Al-Hamad et al., 2008). The literature on systematic errors is in fact dominated by the unity error and little research has been reported about the detection of other types of systematic errors.

Detecting and resolving errors with a known cause leads to (almost) perfect corrections (if the assumed error mechanism is correct) at virtually no costs and prevents unnecessary manual editing. Because this is the ideal type of error correction, research is being carried out at SN aiming to extend the number of errors that can be corrected in this way. So far this research has succeeded in methods for

detection and correction of sign errors, interchange of returns and costs errors and rounding errors.

Sign errors occur in balance variables that are differences between corresponding cost and returns items. For example *operational result = operational returns - operational costs* and equivalent splitting up of *capital result* and *exceptional result*.

Interchange of returns and costs errors occur when in the above equalities costs and returns are interchanged. Rounding errors occur when, in additions such as the balance edits in section 2.1 or the equations above, the sum variable is only one unit smaller or larger than the sum of the component items. We call such inconsistencies rounding errors, because they may be caused by values being rounded to multiples of 1000 Euros.

For these three types of errors and combinations thereof, we have succeeded in finding detection and correction algorithms, see Scholtus (2008). In a test it has been shown that sign errors could be detected and resolved in more than 20% of the records and the correction of rounding errors reduced the number of violated balance edits by one third. This test, however, was limited to data of the structural business statistics for the wholesale branch in 2001.

This research will be extended in two ways. Firstly, we are looking for other types of errors for which the error mechanism can be detected. For instance: misplacement of the decimal separator, interchange of digits and double-keystroke or missed-keystroke errors. Secondly, we are going to broaden the investigation of the occurrence of these types of errors to recent data for other branches of economic activity and to electronic questionnaires.

3.2.1 Research questions

1. Can additional types of errors be identified for which the error mechanism can be detected? For instance: misplacement of the decimal separator or interchange of digits.
2. The successful results of the methods that deductively correct errors were obtained by a test with older data on particular branches of economic activity. Further experimentation should answer the question to what extent these errors occur in various types of data:
    - Recent data of other branches of economic statistics.
    - Electronic versus paper and pencil versions of the same questionnaire.
    - Administrative data.

3.3 Model assisted localization of random errors

SNs current methodology for the automatic detection of random errors is based on the Fellegi-Holt paradigm (see section 2.2). This method finds the minimum number of fields that must be changed in order to obtain a consistent record in the sense that it satisfies all fatal edit rules. This method often yields multiple solutions, each having the same (weighted) number of fields to change. When all weights are

different, multiple solutions should hardly be possible. However, subject matter specialists can only differentiate between the reliability of a few groups of variables and therefore many weights are equal and multiple solutions occur. In practice, a single solution has to be chosen and without additional information, a random choice seems a logical one. However, the problem of multiple solutions can be turned into an advantage if additional information can be exploited to choose the solution with the fields that are most likely in error. Two possible approaches along these lines are the following:

One possibility is to use model-based outlier detection techniques to identify the fields that are suspect because they are far from their expected values. From the multiple solutions, the solution with the least suspect values could than be singled out as the final solution.

Another approach is to determine the weights empirically. A reliability weight for a variable can, for instance, be based on how often edit rules are violated by that variable or, as in the first approach, on model-based outlier techniques. Empirically determined reliability weights will show more variation than weights that are determined a priori by subject matter specialists, thus leading almost surely to a unique solution.

3.3.1 Research questions

1. Which statistical approaches to error localization can be combined with Fellegi-Holt methodology to localize errors?
2. How effective are these methods to identify errors in practical applications?

3.4 Imputation methods for complex business data

Data sets of the Structural Business Statistics (SBS) are complex because the variables are related via a hierarchical system of balance edits and a number of inequality edits must hold as well (see section 2.1). Currently applied imputation methods at SN and at other NSI's do not take edit constraints into account. If the information contained in the constraints can be incorporated in the imputation model, it is expected that the imputations become more accurate and the process can be simplified because the adjustment step 3c will become superfluous, see Pannekoek (2006).

The current imputation method for the SBS-data uses a regression model with a single predictor for each target variable and is applied to each target variable separately. These imputations are not very accurate for a number of variables and often violate edit rules.

Research has been carried out to solve these problems by using multivariate models that allow to impute all variables simultaneously and take their relations defined by the edit-constraints into account, Tempelman (2007). A generic method based on a truncated multivariate normal distribution has been developed and tested. This method uses all observed variables as predictors and takes the constraints into

account. A drawback, however is that the method requires complex iterative algorithms for estimating the parameters and convergence is occasionally problematic.

An alternative is sequential imputation, see Raghunathan et al. (2001). This method uses a separate model for the conditional distribution of each variable given the other variables. The imputation process sequentially imputes each variable given the others with draws from the conditional distribution. This process is repeated until convergence is achieved. This method is attractive because the univariate conditional distributions can be modelled much easier and more flexible (not necessarily with the same normal linear regression model) than the multivariate distribution and constraints are easily taken into account. Moreover, the algorithms involved are relatively simple. Preliminary tests with this method show encouraging results but also make clear that some problems still have to be solved. One problem is that the models used should yield fairly accurate predictions because these predicted values will be used as predictors for the next variable. A grossly misspecified imputation model for one variable will therefore mitigate the prediction error to the next variable and so on. Improvements over the standard linear regression model may be found in transformations of variables and specific models for instance for semi-continuous variables (see e.g. Pannekoek and de Waal, 2005).

The sequential approach need not be limited to univariate conditional distributions. It is also possible to model the conditional distributions of strongly related groups of variables. For instance, the components of a balance edit as (1.2) may be modelled simultaneously. Some preliminary research indicates that it is advantageous to transform these components to proportions of the total. Such proportions, adding up to one, are called compositional data for which special modelling techniques exists (Aitchison, 1982).

3.4.1 Research questions

1. The general research question for the problem of imputing complex business data is: what is an effective imputation model, or strategy consisting of the application of a number of models, that is consistent with the edit rules, leads to accurate predicted values and is easily applicable?
2. Is imputing with a simple model that does not take edit-constraints into account and adjusting them later, always inferior to imputation models that incorporate the constraints? Both empirically and theoretically.

3.5 Calibrated imputation

Statistics Netherlands, as other statistical agencies, relies more and more on information obtained from registers.

Register-based statistics usually require more than one data source, primary and secondary, as a single source does not always contain all the necessary variables and/or population units. The individual sources often do not cover all population units. Especially primary data usually covers only a subset of the population units.

For the remaining population units the variables measured in the primary source are missing.

Several methodologies may be used to estimate tables that are consistent (the same marginal totals), and that are estimated from multiple (incomplete) sources. Among them are:

- Repeated weighting;
- Repeated imputation.

The method of repeated weighting has been developed already (Knottnerus and Van Duin, 2006; Kroese and Renssen 2000) and is applied at SN to estimate tables concerning social statistics. Repeated weighting can be hard when all kinds of logical relations between variables exist in the data sets (edit rules). In that case, the method can even result in inconsistent tables. Another disadvantage of repeated weighting is that it generally leads to a complex process that could easily result in mistakes (multiple data sets, each covering a subset of the population). Furthermore, the conceptual meaning of the technique is not always evident because the weight of a record differs for each estimated table. If those weights are interpreted as 'a record represents an *x*-number of population units', it is uncertain whether a population can be constructed that fits the set of estimated tables.

Repeated imputation may be an alternative for repeated weighting. Repeated imputation starts with a list of all population units with identifying variable(s). Next, the different data sources with their variables are linked to this list. Repeated imputation uses an imputation model for a group of related variables. The resulting, estimated totals are used later, when estimating a second group of related variables, as given totals that need to be preserved. The imputation model needs to be specified in such a way that besides the given totals, it also accounts for a set of edit rules, Pannekoek et al. (2008).

Repeated imputation works with a data set in which each record corresponds to one population unit. This implies that, at each moment, a full data set is available with all population units. After each estimation of a group of related variables, the data file is updated, with the restriction that earlier estimated totals are preserved. For each estimation procedure, the data set is refined and completed with new estimates. Repeated imputation does not aim for a consistent data set at micro level, but it is a means to achieve a set of tables with consistent totals. If a user is interested in strata that differ from the ones used in the imputation model, inconsistencies may occur.

3.5.1 Research questions

1. What are the limits on the number of restrictions in the form of edit rules that give good outcomes in the case of repeated weighting versus repeated imputation?
2. Do these methods assume error-free data sets and what are the consequences of using data sets that still contains errors?
3. What are the practical limitations when large numbers of tables are to be estimated in a consistent way?

3.6 Improving imputation by estimated response probabilities

A problem common to virtually all surveys is the occurrence of non-response. If the non-respondents differ systematically with respect to the target variable(s), estimates of population parameters based on the respondents alone will be biased to an unknown extend. The two commonly applied methods to deal with non-response are imputation and weighting.

Weighting to correct for non-response is similar to weighting to correct for unequal inclusion probabilities in unequal probability sampling. Unequal inclusion probabilities are corrected for by weighting with the inverse of the inclusion probabilities to obtain approximately unbiased estimates of population totals. Weighting to correct for non-response uses the inverse of the response probabilities to achieve the same result. Unfortunately, response probabilities are unknown. However, when register information on both respondents and non-respondents is available, the probability of responding can be modelled as a function of these covariates and the response probabilities can be estimated. If these estimated response probabilities are accurate (consistent), the resulting weighted estimators of population parameters will approximately be unbiased.

Imputation corrects for missing values by filling in expected values or draws according to some models for the missing values. If the residuals of the model have expectation zero for each combination of values of the predictors, then the imputation model is correctly specified and estimates of totals based on imputed data are approximately unbiased.

A third option is to somehow combine weighting with imputation. The idea is that the resulting imputations should lead to estimators without non-response bias when either the response model (but not necessarily the imputation model) is correct *or* the imputation model (but not necessarily the response model) is correct. Estimators with this property are called double robust because they are protected against the misspecification of either of the models as long as the other model is correctly specified. In practice, of course, neither of the models will hold exactly and some non-response bias will remain. However, by combining methods, hopefully, the non-response bias will be reduced more than by using either of the methods alone. Recently, methods combining response- and imputation models have received considerable interest in the literature, mainly in the biostatistics literature (e.g. Bang & Robins, 2005) but also in general statistical publications (Kang and Schafer, 2007, with discussion) and, most interestingly for our purposes, also in the survey statistics literature (Haziza and Rao, 2003).

3.6.1 Research questions

1. Can the use of estimated response probabilities in combination with imputation models in practical applications at SN indeed lead to more accurate estimation of (sub)population totals (in a Mean Squared Error sense)? Can anything be said, in general, about the circumstances under which this method leads to reductions in MSE?

2. What is the best way of modelling response probabilities: are there better, perhaps semi-parametric, models for response probabilities than the usual logistic regression model?
3. When response probabilities are small, the corresponding weights will be large which leads to large variances of estimators of model parameters and resulting imputations and no improvement in MSE of estimators of target parameters. How can we reduce the influence of large weights in order to minimize the MSE?

3.7 Process and quality indicators

An E&I process consists of procedures to detect and correct errors (including missing values). Although the procedure may be set up just to "clean up" the data in a current survey, the detection of errors can also give valuable information on the quality of the data. If particular edit rules are often violated, this may be caused by systematic errors originating from the misinterpretation of certain questions or definitions. If such systematic errors are detected, automated correction procedures can be added to the E&I process to resolve these problems. In addition, the questionnaire can be improved to prevent such errors for future versions of the same survey. Another reason why edits can be violated often, at least for soft edits such as ratio edits, is that the bounds are set too tight. Such cases should show up by a routine application of indicators for edit failures, detected erroneous fields or detected obvious errors.

In the phase of correcting errors, imputation models are used to replace erroneous and missing values. The quality of these corrections is determined by the accuracy of the predictions of the model. This accuracy is hard if not impossible to assess analytically because the true values are unknown. However, simulation approaches can be used to investigate the quality of the imputation model. One such approach is to start with a fully observed and valid subset of the data, then create missing values in these data, impute these missing values and compare the imputations with the now known true values. The problem here is how to create missing data in a realistic fashion such that the conclusions of the simulation can be generalized to real life applications. A particular instance of this simulation approach is cross validation. This is a well known method to investigate specifically the predictive accuracy of models, and is therefore also relevant to the evaluation of imputation models.

In the estimation phase, estimates of population totals are calculated based on imputed data. For fully observed data it is customary to calculate standard errors and confidence intervals for these estimates. The formulas for standard errors for fully observed data are not valid for data that contain imputed values. For simple cases analytic formulas have been derived for imputed data. However, for more complex problems involving several different imputation models with constraints and possibly applied in a sequential scheme, the analytic approach seems infeasible. An alternative approach may be to resort to bootstrap or Jacknife techniques to devise a general variance estimation routine applicable to different complex imputation procedures.

3.7.1 Research questions

1. Can a basic set of standard and useful process indicators be defined for different steps of an E&I process?
2. Can a standard simulation procedure be designed with which different imputation strategies can be tested and evaluated? An example of such an approach is given in Rancourt et al. (2003).
3. What is the best practical way of estimating variances for estimates based on imputed data? Can a general bootstrap or Jacknife procedure be identified that is suitable for complex imputation strategies with a mix of imputation methods?

4. References

Aitchison, J., 1982, The statistical analysis of compositional data (with discussion), *Journal of the Royal Statistical Society*, B. 44, pp. 139-177.

Al-Hamad, A., D., Lewis and P. L. do Nascimento Silva, 2008, Paper presented at the UN/ECE Work Session on Statistical Data Editing, Vienna.

Bang, H. and J.M. Robins, 2005, Double Robust estimation in missing data and causal inference models, *Biometrics*, 61, 962-972.

De Waal, T., 2002, Algorithms for Automatic Error Localisation and Modification. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Helsinki.

De Waal, T., 2003, *Processing of Erroneous and Unsafe Data*. Ph.D. Thesis, Erasmus University, Rotterdam.

De Waal, T. and R. Quere, 2003, A Fast and Simple Algorithm for Automatic Editing in Mixed Data. Journal of Official Statistics, 19, 383-402.

Fellegi, I. P. and D. Holt, 1976, A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71, 17—35.

Farwell, R. K. and M. Raine, 2000, Some Current Approaches to Editing in the ABS. In: *Proceedings of the Second International Conference on Establishment Surveys*, pp. 529-538.

Granquist, L., 1995, Improving the Traditional Editing Process. In: *Business Survey Methods* (eds. Cox, Binder, Chinnappa, Christianson and Kott), John Wiley & Sons, New York, pp. 385-401.

Haziza, D. and J.N.K. Rao, 2003, A Non-response Model Approach to Inference Under Imputation for Missing Survey Data, *Survey Methodology*, 32, 1, 53-64.

Hoogland, J., 2006,. Selective Editing using Plausibility Indicators and SLICE. In *Statistical Data Editing, Volume 3: Impact on Data Quality*, pp.106-130. United Nations, New York and Geneva.

Hidiroglou, M.A. and J.-M. Berthelot, 1986, Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology*, 12, pp. 73-83.

Kang, J.D.Y. and J.L. Schafer, 2007, Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data, *Statistical Science*, 22, 4, 523-539

Knottnerus, P. and C.N.A. van Duin, 2006, Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics* Volume 22 (3), pp. 565 – 584.

Kroese, A.H. and R.H. Renssen, 2000, New applications of old weighting techniques - Constructing a consistent set of estimates based on data from different sources. *International Conference on Establishment Surveys II, Buffalo, New York, 17–21 June,* pp. 831–840.

Little, J.A. and D.R. Rubin, 2002, *Statistical Analysis with Missing Data, second edition*, Wiley, New York.

Pannekoek. J., 2006, Regression imputation with linear equality constraints on the variables. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Bonn.

Pannekoek, J. and T. de Waal, 2005, Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project, *Journal of Official Statistics*, 21, pp. 257-286.

Pannekoek, J., N. Shlomo and T. de Waal, 2008, Calibrated imputation of numerical data under linear edit restrictions, Concept-report, CBS.

Rancourt, E., J-F Beaumont, D. Haziza and C. Mitchell, 2003, Statistics Canada's new software to better understand and measure the impact of nonresponse and imputation. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Madrid.

Raghunathan T.E, J.M. Lepkowski, J. van Hoewyk and P. Solenberger, 2001, A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, 27, 85-95.

Scholtus, S., 2008, Algorithms for detecting and resolving obvious inconsistencies in business survey data. DMV-2008-04-25-SSHS, Discussion paper, CBS.

Tempelman, C., 2007, *Imputation of Restricted Data*. Ph.D. Thesis, University of Groningen.