Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology 41-1

Optimal adjustments for inconsistency in imputed data

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
VOLUME 31
NUMBER 1

by Jeroen Pannekoek and Li-Chun Zhang

Release date: June 29, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^{s}	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 127-144
Statistics Canada, Catalogue No. 12-001-X

Optimal adjustments for inconsistency in imputed data

Jeroen Pannekoek and Li-Chun Zhang[1]

Abstract

Imputed micro data often contain conflicting information. The situation may e.g., arise from partial imputation, where one part of the imputed record consists of the observed values of the original record and the other the imputed values. Edit-rules that involve variables from both parts of the record will often be violated. Or, inconsistency may be caused by adjustment for errors in the observed data, also referred to as imputation in Editing. Under the assumption that the remaining inconsistency is not due to systematic errors, we propose to make adjustments to the micro data such that all constraints are simultaneously satisfied and the adjustments are minimal according to a chosen distance metric. Different approaches to the distance metric are considered, as well as several extensions of the basic situation, including the treatment of categorical data, unit imputation and macro-level benchmarking. The properties and interpretations of the proposed methods are illustrated using business-economic data.

Key Words: Edit-rules; Consistent micro-data; Optimization; Benchmarking.

1 Introduction

We are concerned with the task of reconciling conflicting information in imputed micro data. To illustrate, consider a small part of a record from a structural business survey given in Table 1.1. Two response patterns are postulated; one with only Turnover observed and one where also Employees and Wages are observed. There are many ways to impute the missing values in such a *recipient* record and the proposed adjustment methods apply irrespective of the imputation method used. The use of partial donor imputation is shown in Table 1.1, where the *donor* record is the 'nearest neighbour' from the same category of economic activity and closest to the recipient record with respect to Turnover for response pattern (I) and Employees, Turnover and Wages for response pattern (II). The imputation is said to be partial because a value of the donor is transferred to the receptor if and only if the corresponding one is missing in the recipient record.

Business records generally have to adhere to a number of accounting and logical constraints. For checking of the validity of a record these are referred to as edit-rules. For the example record here, suppose the following three edit-rules are formulated:

$$a1: \quad x_1 - x_5 + x_8 = 0 \quad (\text{Profit} = \text{Turnover} - \text{Total Costs})$$

$$a2: \quad x_5 - x_3 - x_4 = 0 \quad (\text{Turnover} = \text{Turnover main} + \text{Turnover other})$$

$$a3: \quad x_8 - x_6 - x_7 = 0 \quad (\text{Total Costs} = \text{Wages} + \text{Other costs}).$$

Partial donor imputation leads to violation of these edit-rules, which we refer to as the (*micro-level*) *consistency problem*: for response pattern (I), the first two edit-rules involving Turnover are violated; for response pattern (II), all three edit-rules are violated. To obtain a consistent record, *some* of the eight values (i.e., including both the observed and imputed ones) have to be changed. Now, in the two cases

[1] Jeroen Pannekoek, Statistics Netherlands, Henri Faasdreef 312, 2492 JP Den Haag, The Netherlands. E-mail: j.pannekoek@cbs.nl; Li-Chun Zhang, University of Southampton, Social Statistics and Demography, Highfield SO17 1BJ, Southampton, UK and Statistics Norway, Kongensgate 6, Pb 8131 Dep, 0033 Oslo, Norway. E-mail: L.Zhang@soton.ac.uk.

here, it is possible to change only the imputed values to satisfy all the edit-rules, so let us consider adjustments of the imputed values for the moment.

Table 1.1
Data, missing data and donor values for variables in a business record. Employees (Number of employees); Turnover main (Turnover main activity); Turnover other (Turnover other activities); Turnover (Total turnover); Wages (Costs of wages and salaries)

Variable	Name	Response (I)	Response (II)	Donor Values
x_1	Profit			330
x_2	Employees		25	20
x_3	Turnover Main			1,000
x_4	Turnover Other			30
x_5	Turnover	950	950	1,030
x_6	Wages		550	500
x_7	Other Costs			200
x_8	Total Costs			700

Traditional adjustment methods, such as the prorating method implemented in Banff (Banff Support Team 2008), are designed to handle one constraint at a time. In response pattern (I), the prorating method could proceed as follows: (1) adjust the imputed values for Total costs and Profit with a factor 950/1,030 so that they add up to the observed Turnover, (2) adjust the imputed values for Turnover main and Turnover other with the same factor to satisfy the second edit, and (3) adjust the imputed values of Wages and Other costs, again with the same factor to make them add up to the previously adjusted value of Total costs.

For response pattern (II): step (1) and (2) may be carried out as before, but step (3) needs to be modified unless the observed Wages is to be ‘over-written’. Notice that Total costs appears in two edit-rules: $a1$ and $a3$. When the imputed Total costs is only adjusted according to $a1$ in step (1), the relevant information in the observed Wages is ignored. Indeed, depending on the values available it can even happen that Total costs is adjusted downwards in step (1) to the extend that there is no acceptable non-negative solution left for Other costs at step (3). In general, adjusting a variable that appears in multiple edit-rules according to only one of them is not only suboptimal in theory, it also requires an arbitrary choice of the order in which the edit-rules are to be handled, and it may unnecessarily cause a break-down of the procedure.

Under the assumption that the inconsistency is not due to systematic errors, we propose an optimization approach that treats all the constraints simultaneously. To this end it is convenient to express the edit restrictions in matrix notation, as $\mathbf{Cx} = \mathbf{d}$, where $\mathbf{C}$ is the *constraint* (or *restriction*) matrix, and $\mathbf{d}$ a constant vector. For the restrictions $a1 - a3$, we have

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 \end{pmatrix} \text{ and } \mathbf{d} = \mathbf{0}.$$

The non-zero elements in a *row* of the constraint matrix identify all the variables that are involved in the corresponding edit constraint, and the non-zero elements in a *column* of the constraint matrix identify all the edit constraints that involve the corresponding variable.

In addition, there are often linear inequality constraints. The simplest case is the non-negativity of most economic variables. The constraints can then be formulated as $\mathbf{C}_{\text{eq}}\mathbf{x} = \mathbf{d}_{\text{eq}}$ and $\mathbf{C}_{\text{ineq}}\mathbf{x} < \mathbf{d}_{\text{ineq}}$, corresponding to the equality and inequality constraints. For ease of exposition we shall, without noting otherwise, adopt the compact expression $\mathbf{C}\mathbf{x} \le \mathbf{d}$.

As mentioned earlier, not all the values need or should be adjusted. We therefore make a general distinction between *free* (or adjustable) and *fixed* (not adjustable) variables. This includes as a special case the situation where all the data values are considered adjustable. We emphasize that the distinction is not necessarily that between the imputed and observed variables, and imputation may have been carried out for missing values as well as erroneous observed ones. For instance, some imputed values may be held fixed because they are derived by logical reasoning as in deductive imputation, or they may have been obtained from external sources that are considered more reliable. Whereas some observed values may be considered unreliable and are allowed to be changed. Given the absence of systematic errors, a general approach is to identify the adjustable variables by "error localization" (e.g., de Waal, Pannekoek and Scholtus 2011), treating the imputed and observed values as equally error-prone. Nevertheless, in much of the text below we shall treat the imputed values as adjustable and the observed ones as fixed for ease of elaboration.

Given the free and fixed variables, the complete data record is accordingly partitioned into sub-vectors $\mathbf{x}_{\text{free}}$ and $\mathbf{x}_{\text{fixed}}$, and the constraints matrix into $\mathbf{C}_{\text{free}}$ and $\mathbf{C}_{\text{fixed}}$, containing the columns of $\mathbf{C}$ that correspond to $\mathbf{x}_{\text{free}}$ and $\mathbf{x}_{\text{fixed}}$, respectively. The constraints for the adjustable variables are then given by $\mathbf{C}_{\text{free}}\mathbf{x}_{\text{free}} \le \mathbf{d} - \mathbf{C}_{\text{fixed}}\mathbf{x}_{\text{fixed}}$ or, equivalently,

$$\mathbf{A}\mathbf{x}_{\text{free}} \le \mathbf{b} \tag{1.1}$$

where the matrix $\mathbf{A}$ represents the constraints on the free variables and will be called the *accounting* matrix and $\mathbf{b}$ the constant vector for these constraints. Notice that, while the constraint matrix $\mathbf{C}$ is derived a priori from the edit-rules alone, without reference to the actual data, and is the same for all the records, the accounting matrix $\mathbf{A}$ is generally different from one record to another, since the distinction between free and fixed variables varies across the units.

Our strategy to remedy the micro inconsistency problem in imputed data is to make adjustments to the adjustable values that are minimal according to some chosen distance (or discrepancy) measure, such that the adjusted record satisfies all the edit-rules. All the constraints are simultaneously handled assuming the absence of systematic errors.

The rest of the paper will contain the following. The optimization approach will be outlined in Section 2. We consider different distance (or discrepancy) measures, the adjustments they generate, and illustrate their properties and interpretations using the example record above. In Section 3 we discuss possible extensions of the basic approach to adjustments based on statistical assumptions in addition to logical constraints, treatment of categorical data, unit imputation with adjustments, and adjustments for macro-level benchmarking constraints in combination with micro-level consistency. In Section 4 we examine the pasture area data from the Norwegian Agriculture Census 2010, including an approach to the assessment of uncertainty due to editing. A final short summary is provided in Section 5.

2 The minimum adjustment approach

2.1 The optimization problem

We propose to resolve the consistency problem outlined above by adjusting the free variables simultaneously and as little as possible, such that all the edit-rules are satisfied. Let the *adjustable* part of the record *before* adjustment be denoted by a J-vector $\mathbf{x}_0$ and by $\tilde{\mathbf{x}}$ the corresponding J-vector *after* the adjustment. The optimization problem can be formulated as:

$$\begin{aligned} \tilde{\mathbf{x}} &= \arg\min_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}_0) \\ \text{s.t.} \quad & \mathbf{A}\tilde{\mathbf{x}} \le \mathbf{b}, \end{aligned} \tag{2.1}$$

where $D(\mathbf{x}, \mathbf{x}_0)$ is a function measuring the distance (or discrepancy) between $\mathbf{x}$ and $\mathbf{x}_0$, and $\mathbf{A}$ the $K \times J$ accounting matrix associated with the K constraints on $\tilde{\mathbf{x}}$ given in (1.1). We will consider different functions D in Section 2.2.

The conditions for a solution to the minimization problem (2.1) can be found by inspection of the Lagrangian for this problem, which can be written as

$$L(\mathbf{x}, \boldsymbol{\alpha}) = D(\mathbf{x}, \mathbf{x}_0) + \boldsymbol{\alpha}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) \tag{2.2}$$

where $\boldsymbol{\alpha}$ is a K-vector of Lagrange multipliers, or *dual* variables, with components α_k, one for each of the K constraints, and $\mathbf{a}_k$ the k^{th} row (corresponding to constraint k) of the accounting matrix $\mathbf{A}_{K \times J}$. Notice that an additional non-negativity restriction needs to be applied to each α_k corresponding to an inequality constraint, but not the α_k of an equality constraint.

From optimization theory it is well known that for a convex function $D(\mathbf{x}, \mathbf{x}_0)$ and linear constraints, the solution to (2.1) is given by vectors $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}$ that satisfy the so-called Karush-Kuhn-Tucker (KKT) conditions (see, e.g., Luenberger 1984; Boyd and Vandenberghe 2004). One of them is that the gradient of the Lagrangian w.r.t. $\mathbf{x}$ is zero when evaluated at $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}$, i.e.,

$$L'_{x_j}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}) = D'_{x_j}(\tilde{\mathbf{x}}, \mathbf{x}_0) + \sum_k a_{kj} \tilde{\alpha}_k = 0, \tag{2.3}$$

where a_{kj} is the (k, j)-element of $\mathbf{A}$, and $L'_{x_j}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}})$ the gradient of L w.r.t. x_j evaluated at $\tilde{\mathbf{x}}$ and $\tilde{\boldsymbol{\alpha}}$, and D'_{x_j} that of D. From (2.3), we can see how different choices for D lead to different solutions to the adjustment problem, which we will refer to as the *adjustment models*.

2.2 Distance functions and adjustment models

A widely used distance function in many areas of statistics is the weighted least squares (WLS) function given by $D(\mathbf{x}, \mathbf{x}_0) = 1/2 (\mathbf{x} - \mathbf{x}_0)^T \mathbf{W} (\mathbf{x} - \mathbf{x}_0)$, where $\mathbf{W}$ is a diagonal matrix with diagonal elements w_j, for $j = 1, \ldots, J$. We then obtain, from (2.3), the adjustment model

$$\tilde{x}_j = x_{0,j} - \frac{1}{w_j}\sum_k a_{kj}\tilde{\alpha}_k. \tag{2.4}$$

The WLS-criterion thus results in additive adjustments: the total adjustment to the *initial* value $x_{0,j}$ is the weighted sum of the adjustments that correspond to each of the K constraints. The adjustment due to the k^{th} constraint depends on the following:

- The adjustment parameter (i.e., the dual variable) $\tilde{\alpha}_k$ that describes the amount of adjustment. A smaller value for $\tilde{\alpha}_k$ (in absolute sense if k refers to an equality constraint) corresponds to a smaller adjustment; a zero value for $\tilde{\alpha}_k$ means that no adjustment due to that constraint takes place.
- The constant a_{kj} (i.e., an element of the accounting matrix) describes the direction and size of the adjustment to variable j. Often, a_{kj} is 1, -1 or 0 and then describes whether $x_{0,j}$ is adjusted by $\tilde{\alpha}_k, -\tilde{\alpha}_k$ or not at all.
- The weight w_j : variables with larger weights are adjusted less than those with smaller weights. The special case of $w_j \equiv 1$ yields the ordinary least squares (LS) criterion, where the amount of adjustment due to each constraint is the same for all the relevant variables.

A specific choice of the weights is $w_j = 1/x_{0,j}$, for $j = 1, \ldots, J$, in which case the squared relative adjustments are minimized and a larger initial value (i.e., $x_{0,j}$) is adjusted more than a smaller one *in absolute sense*. Dividing (2.4) by $x_{0,j}$ we obtain

$$\frac{\tilde{x}_j}{x_{0,j}} = 1 - \sum_k a_{kj}\tilde{\alpha}_k, \tag{2.5}$$

which is an additive adjustment model for the *ratio* between the adjusted and unadjusted values. It may be noticed that this is the first-order Taylor expansion (i.e., around 0 for all the $\tilde{\alpha}_k$'s) to the multiplicative adjustment given by

$$\frac{\tilde{x}_j}{x_{0,j}} = \prod_k \left(1 - a_{kj}\tilde{\alpha}_k\right). \tag{2.6}$$

From (2.5) we see that $\tilde{\alpha}_k$ determines the relative change from the initial $x_{0,j}$ to the adjusted $\tilde{x}_j$, which in absolute sense is usually much smaller than unity. For instance, $\tilde{\alpha}_k = \pm 0.2$ implies $|20\%|$ adjustment of $x_{0,j}$ if $a_{kj} = \pm 1$, which is large in practice. The products of the $\tilde{\alpha}_k$'s are therefore often much smaller than the α_k's themselves, in which case (2.5) becomes a good approximation to (2.6), and one may regard the WLS adjustment to be roughly given as the product of all the constraint-specific multiplicative adjustments.

Multiplicative adjustment by (2.6) may change the sign of $x_{0,j}$ if $a_{kj}\tilde{\alpha}_k > 1$ for some k. Multiplicative adjustments that preserve the sign of the initial $x_{0,j}$ can be obtained using the

Kullback-Leibler (KL) divergence measure (not formally a distance function), given by $D_{KL} = \sum_j x_j \left(\ln x_j - \ln x_{0,j} - 1\right)$. We then have, from (2.3), the adjustment model

$$\tilde{x}_j = x_{0,j} \prod_k \exp\left(-a_{kj}\tilde{\alpha}_k\right). \tag{2.7}$$

The adjustment due to constraint k is equal to 1 if a_{kj} is 0 (i.e., no adjustment), it is $\exp\left(\tilde{\alpha}_k\right)$ if a_{kj} is 1 and it is $1/\exp\left(\tilde{\alpha}_k\right)$ if a_{ik} is -1. Since $1 - a_{kj}\tilde{\alpha}_k$ is the first-order approximation of $\exp\left(-a_{kj}\tilde{\alpha}_k\right)$ around $\tilde{\alpha}_k = 0$ if $a_{kj} \pm 1$, the WLS and KL criteria can be expected to yield similar adjustments as long as these are small or moderate.

2.3 Methods for solving the minimum adjustment problem

The general convex optimization problem (2.1) can be solved explicitly if the objective function is the weighted least squares and there are only equality constraints. In this case, the Lagrangian is $L(\mathbf{x}, \boldsymbol{\alpha}) = 1/2\,(\mathbf{x} - \mathbf{x}_0)^T\, \mathbf{W}\,(\mathbf{x} - \mathbf{x}_0) + \boldsymbol{\alpha}^T\,(\mathbf{A}\mathbf{x} - \mathbf{b})$, and the equations to be solved are

$$L'_{\mathbf{x}}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{W}(\mathbf{x} - \mathbf{x}_0) + \mathbf{A}^T\boldsymbol{\alpha} = \mathbf{0} \tag{2.8}$$

$$L'_{\boldsymbol{\alpha}}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}. \tag{2.9}$$

Solving (2.8) for $\mathbf{x}$ and substituting the result in (2.9) we obtain

$$\tilde{\boldsymbol{\alpha}} = \left(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T\right)^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{b})$$

and then, on back substitution in (2.8), we obtain explicitly

$$\tilde{\mathbf{x}} = \mathbf{x}_0 - \mathbf{W}^{-1}\mathbf{A}^T\left(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T\right)^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{b}). \tag{2.10}$$

For other objective functions and with inequality constraints in general, there are no explicit solutions to (2.1). However, there are many free or commercial algorithms for the convex optimization problem. For the application in this paper we used the R programming language and applied the so-called row-action or Successive Projection Algorithms (SPA) - see e.g., Censor and Zenios (1997). The SPA is an iterative algorithm that uses the constraints (rows of the accounting matrix) one by one. In one iteration the $\mathbf{x}$-vector is sequentially adjusted to each of the constraints. The operation of adjusting to a single constraint requires only to update the elements of the $\mathbf{x}$-vector that are involved in that constraint (corresponding to the non-zero elements of the currently processed row of the accounting matrix). After all constraints are visited one iteration is completed and the next one is started. For the WLS criterion, an R-package is available that implements the SPA and is especially designed for the adjustment problem (van der Loo 2012).

2.4 Example revisited

Table 2.1 shows the minimum adjustments of the example record in Table 1.1, using the LS-, WLS- and KL-criterion, respectively. The observed values are treated as fixed and shown in bold, the imputed values are adjustable. For the WLS method we use $w_j = 1/x_{0,j}$, giving results that are equal to the KL-criterion up to the first decimal.

For both response patterns, the LS adjustment procedure leads to a negative value for Turnover other which is not acceptable (Table 2.1). When the LS-procedure is rerun with a non-negativity constraint for the variable Turnover other, the result is simply a zero for that variable and 950 for Turnover main due to constraint $a2$. Without the non-negativity constraint, the LS-adjustments are -40 for x_3 and x_4, and -16 for x_6 and x_7, i.e., same adjustment for each pair of variables that appear in the same constraint. The variable Total costs (x_8) is part of two constraints and the total adjustment to this variable consists of two additive components. One component is due to constraint $a1$, and the other due to $a3$. For response pattern (I), the first component is -48 and the second component is 16, and the two add up to -32 in Table 2.1.

Table 2.1
Imputation and adjustment of business record in Table 1.1. DI: Partial donor imputation without adjustment; LS: Least-squares distance; WLS: Weighted least-squares distance; KL: Kullback-Leibler divergence measure; GR: Generalized ratio adjustments

Variable	Name	Response (I)				Response (II)			
		DI	**LS**	**WLS/KL**	**GR**	**DI**	**LS**	**WLS/KL**	**GR**
x_1	Profit	330	282	291	304	330	260	249	239
x_2	Employees	20	20	20	18	**25**	**25**	**25**	**25**
x_3	Turnover Main	1,000	960	922	922	1,000	960	922	921
x_4	Turnover Other	30	-10	28	28	30	-10	28	29
x_5	Turnover	**950**	**950**	**950**	**950**	**950**	**950**	**950**	**950**
x_6	Wages	500	484	470	461	**550**	**550**	**550**	**550**
x_7	Other costs	200	184	188	184	200	140	151	161
x_8	Total costs	700	668	658	646	700	690	701	711

The WLS/KL adjustments are larger, in absolute sense, for larger imputed values than for smaller ones. 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 can be observed as the adjustment *factor* for both these variables is 0.92 (for both response patterns). The adjustment factor for Wages and Other costs in response pattern (I) is equally 0.94 because these variables are in the same constraint $a3$, such that the ratio between their initial values is unaffected by this adjustment. However, the initial ratio of each of these variables to Total Costs is not preserved because Total Costs has a different sign in the constraint $a3$ and, moreover, Total Costs is also part of constraint $a1$ so that it is subjected to two adjustment factors.

3 On possible extensions to related adjustment problems

3.1 Generalized ratio adjustments

The ratio model is routinely used for case weighting in business surveys under the assumption that the economic variables can all be related proportionally to a common measure-of-size of the business unit, see e.g., Särndal, Swensson and Wretman (1992). Motivated by the ratio model one could multiply all the donor values by 950/1,030 to obtain the imputed values for the example record under response pattern (I), including the variable Employees (x_2) for which the initial imputed value 20 does not formally violate any constraints. This shows that there may be situations where, in addition to the logical and accounting constraints, adjustments may be introduced based on statistical assumptions.

For response pattern (II), the observed Employees (x_2), Turnover (x_5) and Wages (x_6) can all potentially be used as the measure-of-size variable in a ratio model, so that a single ratio adjustment does not present itself. However, we may postulate the existence of a *common* ratio between the recipient and donor records under the ratio model, and regard the observed ratios (i.e., 20/25 for Employees, 950/1,030 for Turnover and 550/500 for Wages) as its random manifestations. Then, it seems that a plausible approach is to identify this common ratio as the value that minimizes the variance, or any other dispersion measure that is deemed suitable, of the three individual ratios. Finally, insofar as the common ratio pertains to the other variables, it becomes possible to adjust them using the following *generalized ratio* (GR) approach.

Assume the multiplicative adjustment model $\tilde{x}_j = x_{0,j}\delta_j$, where each δ_j is a random manifestation of a theoretical common ratio. Put the distance function

$$D(\tilde{\mathbf{x}}, \mathbf{x}_0) = 1/2\left(\boldsymbol{\delta}^T\boldsymbol{\delta} - \bar{\delta}^2\right) \tag{3.1}$$

where $\boldsymbol{\delta}$ is the vector of δ_j's and $\bar{\delta}$ the mean of them. For all the variables subjected to the common ratio, including both free and fixed ones, we now carry out the adjustment in two steps. The first step is a conceptual one, where we imagine that an adjustment $\tilde{x}_j / x_{0,j}$ is made to the fixed variables: if $\tilde{x}_j = x_j$ is observed and fixed, then $\delta_j = x_j / x_{0,j}$, whereas $\delta_j = 1$ if $\tilde{x}_j$ is the imputed value $x_{0,j}$ but to be held fixed from 'further' adjustment. At the second step, adjustments are made to the initial values of the free variables by solving the optimization problem (2.1) with (3.1) as the distance function. This yields the GR adjustments of the free variables involved.

An important condition of the GR approach is that at least one of the δ_j's must relate to a fixed variable. Otherwise, $\tilde{x}_j \equiv x_{0,j}$ would be the trivial solution because this always yields $D = 0$. Notice that we have suppressed the denotation J in (3.1), and slightly abused the denotations $\mathbf{x}_0$ and $\tilde{\mathbf{x}}$ introduced for (2.1). Take response pattern (I) in Table 1.1, the fixed value $x_5 = 950$ needs to be included in (3.1), yielding $\delta_5 = \tilde{x}_5 / x_{0,j} = x_5 / x_{0,j} = 950/1,030$. Solving (2.1) for all the other variables yields then $\delta_j \equiv 950/1,030$ and $D = 0$. Whereas, without including δ_5, one would have merely obtained $D = 0$ at $\delta_j = 1$ and $\tilde{x}_j = x_{0,j}$ for $j \neq 5$.

The GR adjustments for response pattern (II) are given in Table 2.1. All the three observed δ_j's, for $j = 2, 5$ and 6, are included in (3.1) and held fixed for the optimization problem. The results are seen to be close to the WLS/KL adjustments. The empirical variance of the multiplicative factors is 0.0270 for the GR adjustments, 0.0276 for WLS/KL and 0.1434 for LS. The relative sum of squared changes, i.e., twice the WLS distance, is 50.6 for the WLS/KL adjustments, 51.6 for GR and 78.0 for LS. Finally, the unweighted sum of squared changes, i.e., twice the LS distance, is 20,925 for the LS adjustments, 23,976 for WLS/KL and 25,090 for GR. Thus, in terms all the three distance functions, the GR adjustments are closer to WLS/KL than LS.

Now, the distance (or discrepancy) measures considered in Section 2.2 may be characterized as *decomposable*, since the overall distance between two vectors is given as a (weighted) sum of the 'distances' between the corresponding components. A consequence is that a variable that does not stand in any constraints will retain the initial value under the minimum adjustment approach. In contrast, the distance (3.1) is *non-decomposable*, where each adjustment is dependent on the other adjustments. As a result even the values that are not explicitly involved in any constraints will be adjusted as long as they are included in the distance function, because of the changes made to the variables that are constraint-bound. The variable Employees provides an example in Table 2.1. The GR approach provides thus a possibility for adjustments based on statistical assumptions in addition to logical and accounting constraints. Indeed, with a single fixed variable included in (3.1), the GR adjustments are reduced to a common proportional adjustment, in accordance with the ratio-adjustment intuition in this case. With multiple fixed variables included, the GR approach aims at a kind of most-uniform adjustments as a generalization of the single-ratio model. For response pattern (II) in Table 1.1, the approach at once takes into account all the three observed ratios. To achieve the same by formulating an explicit statistical model for exactly this response pattern is not as practical in a production setting.

3.2 Adjustments involving categorical data

A categorical variable carries different constraints from a continuous one. It is worth considering the extent to which categorical variables may be incorporated in the optimization approach. We shall distinguish three types of categorical data that are common in practice.

Firstly, we call a categorical/discreet variable *pseudo-continuous* if in practice it can be dealt with as if it were a continuous variable. Typical examples of pseudo-continuous variables are age, number of employees, household size, etc. Pseudo-continuity can affect the choice of adjustment model and distance function. For instance, both additive and proportional adjustments may be acceptable for the number of employees, whereas a proportional adjustment of household size or age seems unnatural. Still, having chosen the adjustment model and distance function, one may handle a pseudo-continuous variable just like a real one. Rounding is necessary afterwards and its effect needs to be monitored.

Secondly, what we call a *nominal* categorical variable indicates whether a unit falls into a particular category. A nominal variable with M categories, labelled $x = 1, 2, \ldots, M$, carry with it the constraint

$$\prod_{m=1}^{M} (\tilde{x} - m) = 0. \tag{3.2}$$

However, the labels (e.g., 1 = tomatoes, 2 = beans, 3 = cucumbers) are not suitable for operations such as addition, multiplication or rounding. Neither is a nominal value 3 more distant to 1 than 2. Therefore, the constraint (3.2) can not be taken into account under the minimum adjustment approach which assumes interval scale measurements. The adjustment of an observed value that does not satisfy (3.2) must be handled by marking it as missing and, then, imputing some admissible as well as suitable value, i.e., just like in case the value is missing to start with.

Thirdly, a variable may be defined to have value zero for the units that are not eligible. Depending on whether the measure is pseudo-continuous or nominal when the unit is eligible, we have a *semi-continuous/-nominal* variable that has a non-zero probability of being zero. The difference to pseudo-continuity above is that a semi-continuous variable may require an additional non-negativity constraint in the accounting matrix. Consider then a semi-nominal variable. In practical questionnaire design, such a variable is often split in two, say, X_1 and X_2. Let $X_1 = 1$ if the unit is engaged in a certain activity, say, production of greenhouse vegetables, and let $X_1 = 0$ otherwise. Let X_2 be a nominal measure of activity when $X_1 = 1$, and $X_2 = 0$ otherwise. Formally, the logical constraint can be given as

$$(1 - \tilde{x}_1)\,\tilde{x}_2 + \tilde{x}_1 \prod_{m=1}^{M} (\tilde{x}_2 - m) = 0 \tag{3.3}$$

Consider all the possible data patterns, including when a value is missing (indicated by "$-$"):

- $(x_1, x_2) = (-, x_2)$: The value $\tilde{x}_1$ can be deduced provided admissible x_2, i.e., x_2 is either 0 or satisfies (3.2), otherwise the situation turns into case $(x_1, x_2) = (-, -)$ below.
- $(x_1, x_2) = (x_1, -)$: If $x_1 = 0$ then $\tilde{x}_2 = 0$; if $x_1 = 1$ then (3.3) reduces to (3.2) above.
- $(x_1, x_2) = (-, -)$: Both values need to be imputed by values that satisfy (3.3).
- (x_1, x_2) : Violation of (3.3) is e.g., the case if $(x_1, x_2) = (1, 0)$ or if $x_1 = 0$ and $x_2 > 0$. We have case $(-, x_2)$ above if x_2 is fixed, $(x_1, -)$ if x_1 is fixed, or $(-, -)$ if neither is fixed.

To summarize, the constraints (3.2) and (3.3) can not be handled by the minimum adjustment approach with linear constraints considered before. Instead, they need to taken care of by the imputation method. Often, donor-based imputation (e.g., Statistics Canada's CANCEIS software that implements the Nearest Neighbour Imputation Methodology, NIM) can be designed to impute categorical data such that user specified constraints are satisfied, see e.g., Bankier, Lachance and Poirier (2000).

3.3 Adjustment of donor-based unit imputation

In donor-based unit imputation the whole record of values are taken from the chosen donor. This has advantages over joint modelling of all the target variables if there are many of them. Chen and Shao (2000) establish the consistency of survey estimator based on nearest neighbour imputation (NNI) under mild conditions. The key assumption is that the difference in the conditional expectations of any target variable between a donor and a receptor, given the variables on which the distance metric is calculated, is bounded by the "distance" between them. That is, they have the same expectations for all the statistical variables if the "distance" between them is zero.

There is thus a need for adjusting donor-based unit imputation when the "distance" between the receptor and the donor is not zero. To illustrate with the example record in Table 1.1, suppose Turnover (x_5) is always known from the administrative source and is used for donor identification, so that partial imputation under response pattern (I) becomes unit imputation. Since Turnover of the receptor differs from that of the donor, the distance between them is not zero, and it seems natural that the donor values should be adjusted to take this difference into account. Indeed, now that there are constraints involving Turnover, adjustments are necessary in any case.

Let $\mathbf{x}$ contain the variables that may be missing. Let $\mathbf{z}$ contain the known variables that are used for donor identification. Let $\mathbf{x}^* = \left(\mathbf{x}^T, \mathbf{z}^T\right)^T$ be the combined vector of variables. Unit imputation (giving $\mathbf{x}_0$) can be regarded as partial imputation of the missing sub-vector $\mathbf{x}$ of $\mathbf{x}^*$. The need for adjustment of unit imputation may arise if there are edit-rules that involve both values of $\mathbf{x}$ and $\mathbf{z}$, and/or if the $\mathbf{z}$-values do not match exactly between the donor and receptor. Indeed, unit imputation without adjustment may rather be considered exceptional in practice.

3.4 Macro-level benchmarking in addition to micro-level constraints

A business census requires imputation and editing in order to arrive at a complete dataset for statistical production. Or, a statistical register may be constructed based on a combination of administrative data and one or several sample surveys. Editing and imputation are again necessary. A common feature is that, unlike survey sampling, no case weighting is needed.

When processing such data, macro-level *benchmark* constraints are frequently imposed due to concerns for statistical efficiency and/or macro-level consistency with external sources. A benchmark constraint is satisfied if the complete data add up to the given benchmark total, which may refer to different aggregation levels, i.e., containing both population and sub-population totals. For instance, certain key national totals may be estimated by some suitable method and imposed as benchmark constraints afterwards. Or, a set of domain-level benchmark constraints may be derived by some small area estimation technique. Also benchmark constraints from external sources are common in structural business statistics - an example from the Norwegian Agriculture Census 2010 will be described in Section 4.

Methods for imputation under benchmark constraints have been studied by Beaumont (2005), Chambers and Ren (2004), Zhang (2009) and Pannekoek, Shlomo and de Waal (2013). The approach taken here is similar to the one taken in the first two papers. In both these papers a weighted least squares distance between initial imputed values (or outlying values in the case of Chambers and Ren 2004) and adjusted imputed values is minimized subject to the constraint that sample-weighted totals based on the adjusted data are equal to the benchmark totals. Here, we assume that some suitable imputation method has been applied to yield the initial complete population dataset, which may or may not be benchmarked. The inconsistency problem on the micro-level implies that adjustments of the initial complete data set will be necessary in general.

Denote by $\mathbf{X}$ the complete dataset of interest, where each row corresponds to a unit-level record as the one in Table 1.1, and each column corresponds to a particular variable. Let $\mathbf{X}_0$ be the initial complete dataset after imputation and $\tilde{\mathbf{X}}$ the adjusted dataset. Each benchmark constraint applies to a particular column vector of $\mathbf{X}$ and over the units that fall under its domain. That is, it can be expressed generically as $\mathbf{r}^T \mathrm{col}(\mathbf{X}) = t$, where $\mathrm{col}(\mathbf{X})$ is the column vector of concern, and $\mathbf{r}$ is the indicator vector for whether a unit belongs to the domain of concern, and t the benchmark total. In this way all the benchmark constraints may be summarized as

$$[\mathbf{r}]^T [\mathrm{col}(\mathbf{X})] = \mathbf{t} \tag{3.4}$$

where each column of $[\mathrm{col}(\mathbf{X})]$ corresponds to a benchmark constraint, and each column of $[\mathbf{r}]$ the corresponding indicator vector, and $\mathbf{t}$ the vector of all the benchmark totals. Notice the similarity between (3.4) and (1.1). A minimum adjustment approach follows on specifying the adjustable and fixed values and the distance (or discrepancy) function.

Both the benchmark constraints and the micro-level constraints can be seen as linear constraints on the very long vector containing all elements of $\mathbf{X}$, $\operatorname{vec}(\mathbf{X})$, say. Conceptually, all constraints together can therefore be expressed in the form (1.1). The restriction matrix of this formulation is, however, huge and very sparse. The rows corresponding to the micro-level constraints contain possibly non-zero values corresponding to the values in the record they apply to and zeros for all other values of $\operatorname{vec}(\mathbf{X})$ and the rows corresponding to the benchmark constraints contain non-zero elements only corresponding to the values in $\operatorname{vec}(\mathbf{X})$ that contribute to that benchmark total. In practice, the optimization problem generated by (3.4) in addition to the micro-level constraints can be handled using the SPA, i.e., one constraint at a time and operating only on the elements of $\operatorname{vec}(\mathbf{X})$ corresponding to the non-zero elements in that constraint, without actually forming this huge and sparse constraint matrix. For the benchmark constraints we only need to process the columns of $[\operatorname{col}(\mathbf{X})]$ one by one and for the micro-level constraints we process each unit-level record one at a time. These iterative minimum adjustments along the columns and rows of $\mathbf{X}$ resemble the iterative proportional fitting (or raking) algorithm for fitting log-linear models to contingency table data and for adjusting (contingency) tables to new margins, which is formally identical to a SPA with the KL-divergence and equality constraints only.

4 Case study

4.1 Imputation and adjustment of pasture data

The population for the "main questionnaire" of the Norwegian Agriculture Census 2010 contains about 45,000 units. Questions 22 - 24 deal with pasture area:

- Question 22 inquires the units that possess productive pasture.
- Question 23 inquires the total productive pasture area in 2010.
- Question 24 inquires the composition of pasture area by the last time it was seeded: (1) 2006 - 2010, (2) 2001 - 2005, and (3) 2000 or earlier.

Denote by $x_{0,1}, x_{0,2}$ and $x_{0,3}$ the three reported categories of pasture area in Question 24. Let $x_0 = \sum_{j=1}^{3} x_{0,j}$ be the sum that is the subject of Question 23. Now, this total is available from the government agency that administers the relevant subsidy. In editing the reported x_0 is overwritten by the administrative figure, denoted by $\tilde{x}$, and held as fixed afterwards. Next, Question 22 can be inferred given $\tilde{x}$ and held as fixed afterwards, so that only Question 24 remains to be handled.

Below we describe the treatment of the 34,480 units that have productive pasture area according to their respective observation patterns (Table 4.1, where the unit index i of all the variables was omitted for ease of presentation).

- 10,378 units reported a total pasture area that is consistent with the administrative source: these are the potential donors; no adjustment is needed.
- 11,827 units have a reported total that is greater than the known value: these have a micro-level inconsistency problem. Of course, missing values can also be the case if $\sum_j r_j < 3$, but the

chance is small, so we shall assume that there are no missing values among these units. All the observed values are adjustable, such that the accounting equation is given by

$$\sum_{j;r_j=1} \tilde{x}_j = \tilde{x}.$$

The GR approach simply yields the proportional adjustment $\tilde{x} \big/ \sum_{j;\, r_j=1} x_{0,j}$. The same adjustment is given by the WLS-approach with $w_j = 1/x_{0,j}$ if $r_j = 1$, as well as by the KL approach. We notice that there is no particular motivation for considering additive adjustments for these data.

- 3,876 units have *no* reported pasture area of any kind, despite they have productive pasture area according to the administrative source: these constitute unit-missing records. The nearest-neighbour (NN) donor is found according to $\tilde{x}$, within each of the 12 "farming forms", which is a classification known for the whole population. In the case of multiple NN donors, we choose the one with the shortest physical distance, which make the NN-imputation completely deterministic, given all the $\tilde{x}$- values. Finally, a proportional adjustment of the donor values is carried out in order to satisfy the accounting equation

$$\sum_{j;r_j^*=1} \tilde{x}_j = \tilde{x}$$

  where r_j^* is the observation/reporting indicator associated with the donor.

- 3,019 units have reported pasture areas of *all* the three kinds, but their sum is less than the known total: these have a micro-level inconsistency problem. A proportional adjustment is applied to all the reported values w.r.t. the accounting equation $\sum_{j=1}^{3} \tilde{x}_j = \tilde{x}$.

- The last two groups are the 2,703 units with one kind of reported pasture area and the 2,677 units with two kinds of reported pasture area. Obviously, that the reported total is less than the known value here may be caused by inconsistency and/or missing values. To avoid introducing systematic pattern through editing, we let the decision depend on the donor. Take a unit with only one reported pasture area. Firstly, the potential donors are limited to those from the same "farming form", as well as having *at least* the same kind of pasture area. The NN donor is then selected among these to minimize

$$\max\left(\left|\tilde{x}^*/\tilde{x} - 1\right|, \left|x_j^*/\tilde{x}^* - x_{0,j}/\tilde{x}\right|_{j;r_j=1}\right)$$

  where (x_1^*, x_2^*, x_3^*) and $\tilde{x}^*$ are the values of the potential donor. In other words, the NN donor is selected both w.r.t. the relative difference between the total pasture area as well as the proportion of the reported kind of pasture area to the corresponding total. Let the NN donor be associated with $\mathbf{x}^*$ and $\mathbf{r}^*$. If $\sum_j r_j^* > 1 = \sum_j r_j$, then we assume that there are missing values where $r_j^* = 1$ but $r_j = 0$; whereas, if $\sum_j r_j^* = \sum_j r_j$, then we assume that there is only an inconsistency problem. The remaining imputation and adjustment actions are straightforward. The same treatment is applied to the units with two reported pasture areas, with obvious modifications due to $\sum_j r_j = 2$.

Table 4.1
Observation pattern among units with productive pasture area: $r_j = 1$ if $x_{0,j}$ is reported, $r_j = 0$ otherwise; $j = 1, 2, 3$ for three categories of pasture area

Total	$\sum_j r_j x_{0,j} = \tilde{x}$	$\sum_j r_j x_{0,j} > \tilde{x}$	$\sum_j r_j x_{0,j} < \tilde{x}$			
			$\sum_j r_j = 0$	$\sum_j r_j = 1$	$\sum_j r_j = 2$	$\sum_j r_j = 3$
34,480	10,378	11,827	3,876	2,703	2,677	3,019

The sub-population and population totals based on imputation with adjustments are given in Table 4.2, in comparison with raw data totals and the census file totals. We notice the following. (a) The census file had been edited in a 'traditional' way that involves much clerical work (about 1.5 man-year in total). In contrast, the editing procedures here are fully automated, and everything (i.e., exploratory analysis, decision of the treatments, programming and processing) was done in less than two days. Although the questions concerning pasture areas are only 3 out of a total of 36 questions of the "main questionnaire", it is obvious that the potential saving in time could be enormous. (b) The differences between the imputed totals and the census totals are small for all sub-populations, compared to those between the raw data and the census totals. All the changes from the raw data are in the 'right' direction, judged by the census results. One may conclude that the automated editing procedures have achieved most of the census editing results. (c) It is possible to introduce benchmark constraints in addition. An an illustration, we used the census file sub-population totals for the 3,876 unit-missing records, in addition to the known pasture area total for each of them. Convergence was reached in 23 iterations with the WLS criterion. (d) For the 5,380 units where partial missing may be the case, imputation of 'missing' values was carried for about 25% of them in the census processing, whereas it is about 75% by the editing procedure here. The number of cases for partial missing is probably under-estimated in the census file because it is based on selective manual checks. In any case, not withstanding the differences in the individual treatments, the edited totals are fairly close to each (Table 4.2, under $0 < \sum_j r_j < 3$).

Table 4.2
Sub-population and population pasture area totals based on raw data, imputation with adjustments and census production data. (All figures ×10^5)

	$\sum_j r_j x_{0,j} > \tilde{x}$			$\sum_j r_j x_{0,j} < \tilde{x}$					
				$\sum_j r_j = 3$			$0 < \sum_j r_j < 3$		
Raw	8.20	6.95	12.76	1.40	1.45	1.53	1.33	0.86	3.05
Impute & Adjust	5.24	4.34	8.71	1.72	1.81	1.88	2.01	1.87	3.51
Census	5.47	4.37	8.45	1.73	1.85	1.84	2.04	1.54	3.80
	$\sum_j r_j = 0$			$\sum_j r_j > 0$			Total		
Raw	-	-	-	14.0	12.4	21.9	-	-	-
Impute & Adjust	1.20	1.06	1.93	12.2	11.3	19.3	13.43	12.38	21.17
Census	1.31	1.23	1.66	12.6	11.0	19.1	13.95	12.25	20.79

4.2 Approximate mean squared error estimation

As the measure of uncertainty for the pasture area data here, we use the mean squared error of prediction (MSEP) given by

$$\mathrm{MSEP}_j = E\left\{\left(\tilde{X}_j - X_j\right)^2 \middle| \mathbf{R}_U, \tilde{\mathbf{X}}_U\right\}$$

where $X_j = \sum_{i \in U} x_{ij}$ is the target population total and $\tilde{X}_j = \sum_{i \in U} \tilde{x}_{ij}$ is the corresponding total based on imputation with adjustments, for $j = 1, 2, 3.$ Moreover, $\tilde{\mathbf{X}}_U = (\tilde{x}_i)_{i \in U}$ contains the known pasture area totals in the population, and $\mathbf{R}_U$ is the matrix of missing indicators whose i^{th} row is given by $(r_{i1}, r_{i2}, r_{i3}).$

Now, while it is customary that adjustments due to inconsistency in the micro data are referred to as imputation in statistical data editing, the eventual uncertainty associated with this is generally 'ignored' afterwards. This amounts to assume that $\tilde{x}_{ij} = x_{ij}$ if $r_{ij} = 1.$ What remains to be accounted for is the uncertainty associated with the imputation of the missing values and the subsequent adjustment of the donor values, under the assumption that neither imputation nor adjustment introduces bias to the final value. This amounts to assume that $E\left(\tilde{x}_{ij} - x_{ij}\right) = 0$ if $r_{ij} = 0.$ Under these two assumptions, we have

$$\begin{aligned} \text{MSEP}_j &= E\left\{\left(\sum_{i \in U}(1 - r_{ij})\,\tilde{x}_{ij} - \sum_{i \in U}(1 - r_{ij})\,x_{ij}\right)^2\right\} \\ &= V\left(\sum_{i \in U; \mathbf{r}_i = \mathbf{1}, d_{ij} \geq 1} d_{ij}\delta_{ij}x_{ij}\right) + V\left(\sum_{i \in U; r_{ij} = 0} x_{ij}\right) \\ &\approx \sum_{i \in U; \mathbf{r}_i = \mathbf{1}, d_{ij} \geq 1} d_{ij}^2 V\left(\delta_{ij}x_{ij}\right) + \sum_{i \in U; r_{ij} = 0} V\left(x_{ij}\right) \end{aligned}$$

where d_{ij} is the number of times x_{ij} is used as a donor value for imputation of missing data, and the decomposition of variance holds provided the distributions of the units are independent of each other. Moreover, provided $d_{ij} \geq 1,$

$$\delta_{ij} = \sum_{k \in U; x_{kj}^* = x_{ij}} \tilde{x}_{kj} \Big/ \left(d_{ij}x_{ij}\right)$$

where $x_{kj}^* = x_{ij}$ means that x_{ij} is used as the donor value for $x_{kj},$ and $\tilde{x}_{kj}$ is the final value after adjustment. In other words, δ_{ij} is the combined adjustment made to $d_{ij}x_{ij},$ where $d_{ij}x_{ij}$ would have been the contribution of x_{ij} to $\tilde{X}_j$ through imputation if it had been donor imputation *without* adjustment. Notice that d_{ij} can be treated as a constant in the last (approximate) equation as long as the donor identification depends only on $\mathbf{R}_U$ and $\tilde{\mathbf{X}}_U.$ This is true for the 3,876 unit-missing records, but not exactly for the 5,380 units that may have partial missing. As explained in Section 4.1, the NN-identification in fact also depends on the observed x_{ij} - values. For this reason, the last equation holds only approximately.

A ratio model for the conditional variance of x_{ij} seems natural here, i.e.,

$$x_{ij} = \beta_j x_i + \varepsilon_{ij} \text{ where } E\left(\varepsilon_{ij}\right) = 0 \text{ and } V\left(\varepsilon_{ij}\right) = \sigma_j^2 x_i^{\alpha_j}$$

where $\left(\beta_j, \sigma_j^2, \alpha_j\right)$ may vary according to the *composition* of the pasture areas, denoted by $\mathbf{q} = (1,1,1), (1,1,0), (1,0,1)$ and $(0,1,1),$ where $q_{ij} = 1$ if unit i has the j^{th} type pasture and 0 otherwise. Notice that, in the case of $\sum_j q_{ij} = 1,$ we have $x_{ij} = \tilde{x}$ if $q_{ij} = 1,$ so that the conditional variance is zero. The parameters of this ratio model can be estimated from the 10,378 potential donors satisfying $\sum_j r_j x_{0,j} = \tilde{x}.$ Exploratory data analysis shows that $\alpha_j = 2$ is a reasonable choice in all the cases, so that in the calculations below only β_j and σ_j^2 vary according to the observation pattern, denote by $\left(\beta_{j;h}, \sigma_{j;h}^2\right)$ for $h = 1, \ldots, 4.$ Notice that, as a result of $\alpha_j \equiv 2,$ the same $\hat{\sigma}_{j;h}^2$ will be obtained regardless of j whenever $\sum_j q_{ij} = 2.$ Take e.g., $\mathbf{q} = (1, 1, 0)^T$, we have $\hat{\beta}_1 + \hat{\beta}_2 = 1,$ such that the 'standardized' fitted residuals are given by $\hat{\varepsilon}_{i1} / \tilde{x}_i = x_{i1} / \tilde{x}_i - \hat{\beta}_1$ and $\hat{\varepsilon}_{i2} / \tilde{x}_i = x_{i2} / \tilde{x}_i - \hat{\beta}_2 = \left(\tilde{x}_i - x_{i1}\right) / \tilde{x}_i - \left(1 - \hat{\beta}_1\right) = -\hat{\varepsilon}_{i1} / \tilde{x}_i$. In any case, we obtain $\hat{V}_h\left(x_{ij}\right) = \hat{\sigma}_{j;h}^2 \tilde{x}_i^2$ for unit i with composition $h.$

The adjustment factor δ_{ij} seems difficult to model in advance. But its mean and variance can be estimated empirically *after* imputation and adjustment have been carried out, denoted by $\mu_\delta = E\left(\delta_{ij}\right)$ and $\sigma_\delta^2 = V\left(\delta_{ij}\right),$ respectively. Moreover, we assume δ_{ij} to be independent of x_{ij} conditional on $\tilde{x}_i$. This seems a plausible assumption, since the former depends mostly on how x is distributed in the 'neighbourhood' of $x = \tilde{x},$ whereas the latter depends on the variation across j given that the sum is equal to $\tilde{x}.$ For instance, asymptotically as the chance of finding a donor in any arbitrarily close neighbourhood tends to unity, the adjustment factor δ_{ij} tends to 1 in probability, irrespective of the values of x_{ij}. It now follows that, given composition $h,$ an estimate of the corresponding $V_h\left(\delta_{ij} x_{ij}\right)$ is given by

$$\hat{V}_h\left(\delta_{ij} x_{ij}\right) = \hat{\sigma}_{j;h}^2 \tilde{x}_i^2 \hat{\sigma}_\delta^2 + \left(\hat{\beta}_{j;h} \tilde{x}_i\right)^2 \hat{\sigma}_\delta^2 + \hat{\sigma}_{j;h}^2 \tilde{x}_i^2 \hat{\mu}_\delta^2.$$

Finally, combining all the above, we obtain an approximate MSEP estimate as

$$\widehat{\text{MSEP}}_j \approx \sum_h \sum_{i \in U_h; \mathbf{r}_i = \mathbf{1}} d_{ij}^2 \hat{V}_h\left(\delta_{ij} x_{ij}\right) + \sum_h \sum_{i \in U_h; r_{ij} = 0} \hat{V}_h\left(x_{ij}\right).$$

The results of approximate variance estimation are given in Table 4.3. We know in advance that the regression coefficient of the ratio model must vary according to the composition of pasture area, but the estimates of $\sigma_{j;h}^2$ suggest that it has been sensible to allow the variance parameter to depend on $h.$ The estimated mean of δ_{ij} is close to unity for all the pasture area types, making no indications that the assumptions regarding the adjustment factors are unreasonable. The variance of δ_{ij} is clearly the largest for $j = 2,$ which is also reflected in the fact that the estimated MSEP here has the largest increase compared to NN-imputation without adjustment. The relative root MSEPs are too small to account for the actual differences between the census totals and the imputed totals (given in Table 4.2). This serves to illustrate the following general impression regarding the assessment of uncertainty due to editing. Systematic effects in terms of the first-order moments of the resulting statistics usually dominate the overall uncertainty due to editing. But they are also more difficult to quantify compared to the second-order variance properties. In the case here, this concerns the two 'first-order' assumptions made in the beginning, i.e., $\tilde{x}_{ij} = x_{ij}$ if $r_{ij} = 1$ and $E\left(\tilde{x}_{ij} - x_{ij}\right) = 0$ if $r_{ij} = 0.$ More sophisticated assumptions about the error-mechanism of consistency adjustments in editing are needed in order to progress beyond such an 'optimistic' approach.

Table 4.3
Approximate variance estimation for imputation with adjustment. RMSEP: Root MSEP. RMSEP by NN-imputation without adjustment in parentheses

		$j = 1$	$j = 2$	$j = 3$
$\hat{\beta}_j$	$\mathbf{q} = (1, 1, 1)$	0.312	0.359	0.329
	$\mathbf{q} = (1, 1, 0)$	0.346	0.654	-
	$\mathbf{q} = (1, 0, 1)$	0.407	-	0.593
	$\mathbf{q} = (0, 1, 1)$	-	0.567	0.433
$\hat{\sigma}_j^2$	$\mathbf{q} = (1, 1, 1)$	0.0248	0.0511	0.0364
	$\mathbf{q} = (1, 1, 0)$	0.0478	0.0478	-
	$\mathbf{q} = (1, 0, 1)$	0.0464	-	0.0464
	$\mathbf{q} = (0, 1, 1)$	-	0.0798	0.0798
	$\left(\hat{\mu}_\delta, \hat{\sigma}_\delta^2\right)$	(0.992, 0.0248)	(1.020, 0.0994)	(1.003, 0.0236)
	$\widehat{\text{RMSEP}}$	3,267 (3,134)	4,190 (3,530)	3,111 (2,925)
	$\widehat{\text{RMSEP}} \Big/ \sum_{i; r_{ij}=0} \tilde{x}_{ij}$	1.41%	1.79%	0.93%
	$\widehat{\text{RMSEP}} \Big/ \tilde{X}_j$	0.24%	0.34%	0.15%

5 Summary

In this paper we have formulated an optimization approach to the micro-level inconsistency problem that may be caused by measurement errors and/or imputation of missing values. This provides a general methodology that extends beyond the traditional single-constraint adjustment methods such as prorating. All constraints are handled simultaneously; if a variable appears in more than one constraint then it is adjusted according to all of them. Besides being optimal according to the chosen distance (or discrepancy) function, the approach also has the practical advantage that there is no need to specify the order in which the constraints are to be applied.

Several distance (or discrepancy) functions are analysed. It is shown that minimizing the weighted least squares leads to additive adjustments and minimizing the Kullback-Leibler divergence measure leads to multiplicative adjustments. However, for a specific choice of weights the WLS solution of the optimization problem is an approximation to the KL solution.

Adjustments based on statistical assumptions in addition to the logical constraints is introduced under the generalized ratio approach. The GR adjustments can be considered as a generalization of the single-ratio adjustment under a ratio model. All the observed variable-specific ratios between the receptor and donor records are utilized; a variable that does not stand in any constraint can also be adjusted if it is included in the distance function.

Also discussed are adjustments involving categorical data, unit-missing records and macro-level benchmark constraints in addition to the micro-level consistency constraints. Taken together, the proposed optimization approach is applicable to continuous data in a number of situations.

Acknowledgements

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

References

Banff Support Team (2008). *Functional Description of the Banff System for Edit and Imputation*. Technical Report, Statistics Canada.

Bankier, M., Lachance, M. and Poirier, P. (2000). *2001 Canadian Census Minimum Change Donor Imputation Methodology*. Working paper 17, UN/ECE Work Session on Statistical Data Editing, Cardiff.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society, Series B* (Statistical Methodology), 67, 445-458.

Boyd, S., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Censor, Y., and Zenios, S.A. (1997). *Parallel Optimization*. Theory, Algorithms, and Applications. Oxford University Press, New York.

Chambers, R.L., and Ren, R. (2004). Outlier robust imputation of survey data. In *Proceedings of the Survey Research Methods Section,* American Statistical Association, 3336-3344.

Chen, J., and Shao, J. (2000). Biases and variances of survey estimators based on nearest neighbour imputation. *Journal of Official Statistics,* 16, 113-132.

de Waal, T., Pannekoek, J. and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. New Jersey: John Wiley & Sons Inc., Hoboken.

Luenberger, D.G. (1984). *Linear and Nonlinear Programming, Second Edition*. Addison-Wesley, Reading.

Pannekoek, J., Shlomo, N. and de Waal, T. (2013). Calibrated imputation of numerical data under linear edit restrictions. *Annals of Applied Statistics,* 7, 1983-2006.

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

van der Loo, M. (2012). rspa: Adapt numerical records to (in)equality restrictions with the Successive Projection Algorithm. R package version 0.1-5. Available at: http://cran.r-project.org/web/packages/rspa/index.html.

Zhang, L.-C. (2009). *A Triple-Goal Imputation Method for Statistical Registers*. Working paper 28, UN/ECE Work Session on Statistical Data Editing, Neuchâtel, Switzerland.