WP.40
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (vii): New and emerging methods

Partial (donor) imputation with adjustments

Key Invited Paper

Prepared by Jeroen Pannekoek, Statistics Netherlands and Li-Chun Zhang, Statistics Norway

I. Introduction

1. We are concerned with the task of reconciling conflicting information in statistical micro data that may arise due to partial (donor) imputation. The missing values are imputed either by the corresponding values of a suitable donor or by statistical estimation. The imputed record contains then two parts of data from different sources. One part contains the observed values from the original record and the other the imputed values. Edit rules that involve variables from both parts of the record will often be violated. For instance in business statistics we may have that *turnover* must be equal to the sum of *profit* and *costs*, *costs* is again the sum of costs for material, personnel, housing etc. and all variables except profit are non-negative. If some of the variables are missing, the imputed values taken from a donor may not satisfy the various restrictions, together with the observed values of the original record.

2. One strategy to remedy this problem is to make adjustments to the imputed values that are minimal in some sense, such that a record consistent with the edit rules results. In this paper, the edit rules are specified as linear equality-/inequality-constraints on the variables. The minimal adjustments are then obtained by minimizing a chosen distance metric subjected to these constraints.

3. We consider two different approaches to the distance metric. In the first case one sets out to minimize the changes directly. It will be shown that different distance functions under this approach lead to adjustments that preserve different aspects of the structure of the data. In the second case one sets out to minimize the differences in the changes, so that the adjustments are most ‘uniform’ in some sense. Under this latter approach, even the values that are not explicitly involved in any constraints will be adjusted because of the changes made to the variables that are directly constraint-bound. The properties and interpretations of the different approaches are illustrated numerically based on empirical data in business statistics.

II. Numerical illustration of the problem

A. Imputation of a business record with missing data

4. To illustrate the problem, we consider a small part of a record from a structural business survey with missing data that is to be imputed. The data for this record are shown in Table 1. For this record two response patterns are assumed; one with only Turnover observed and one were also Employees and Wages are observed. There are a number of common ways to impute the missing values in such a record. One possibility is the use of the values from a donor record to impute the missing values in the recipient record. This donor can, for instance, be the “nearest neighbour” donor record, from the same category of economic activity and closest to the recipient record in some metric based on some common observed

variables, for instance Turnover for response pattern (I) and Employees, Turnover and Wages for response pattern (II). Imputation then entails the replacement of the missing values by the corresponding values from the donor record, we call this partial donor imputation because not all the values of the donor are transferred to the recipient.

Tabel 1. Data, missing data and donor values for some variables in a business record.

Variable	*Name*	*Response pattern 1*	*Response pattern 2*	*Donor Values*
x_1	Profit			330
x_2	Employees (Number of employees)		25	20
x_3	Turnover main (Turnover main activity)			1000
x_4	Turnover other (Turnover other activities)			30
x_5	Turnover (Total turnover)	950	950	1030
x_6	Wages (Costs of wages and salaries)		550	500
x_7	Other costs			200
x_8	Total costs			700

B. The consistency problem

5. Business records generally have to adhere to a number of accounting and logical constraints. These constraints are widely employed for checking the validity of a record and are, in this context, referred to as edit-rules. For the example record above, the following three edit-rules are formulated:

a1: $x_1 - x_5 + x_8 = 0$ (Profit = Turnover – Total Costs)
a2: $x_5 - x_3 - x_4 = 0$ (Turnover = Turnover main + Turnover other)
a3: $x_8 - x_6 - x_7 = 0$ (Total Costs = Wages + Other costs)

6. Partial donor imputation for either response pattern in Table 1 leads to violation of these edit-rules, which we refer to as the consistency problem. In particular, for response pattern (I), the first two edit-rules involving Turnover are violated and, for response pattern (II), all three edit-rules are violated. To obtain a consistent record some of the values in thes record have to be changed or “adjusted”. Often, the imputed values are the candidates for adjustment while the actually observed values are not changed. However, other choices of adjustable and non-adjustable or fixed values can be made.

7. 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 950/1030 to make them add up to the observed Turnover, (2) then 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, also with the same factor to make them add up to the previously adjusted value of Total costs. Indeed, one may be tempted to extend this rescaling to imputed variables that are not in edit-constraints (only Employees in this case), which is not necessay for consistency with the specified edit-rules but can be justifiable if it is assumed that these variables are related to Turnover in approximately the same way as in the donor record. This last option is further discussed in section IV.

8. This easy and intuitive solution becomes more complicated for the response pattern (II). Whereas the first two steps may be carried out as before, the third step shows some difficulties of this approach. Total costs appears in two edit-rules: *a1* and *a2*. In both edit-rules one variable is observed (Turnover and Wages, respectively) but Total costs is only adjusted to satisfy *a1* and the resulting adjusted value is irrespective of the observed value of Wages, thereby ignoring relevant information on the Total costs. Indeed, in such cases it can even happen that Total costs is adjusted downwards to the extend that it becomes smaller than Wages and hence there is no acceptable non-negative solution for Other costs. Adjusting a variable that appears in multiple edit-rules to just one of them is not only suboptimal in the sense described above, but also leads to rather arbitrary choices of the order in which edit-rules should be handled.

9. Another problem that the second response pattern illustrates is that a simple proportional adjustment is more plausible when variables have to be adjusted such that their *sum* equals a constant than when variables have to be adjusted in order to render their *difference* equal to a constant. For instance, for edit rule *a3*, formulated as Wages = Total costs – Other costs, with values $550 \neq 700 - 200$ and 550 fixed, a proportional adjustment of Total costs and Other costs would result in values of 770 and 220. However, , much smaller adjustments can be obtained by *increasing* Total costs to 740 while *decresing* Other costs to 190. Such cases are therefore mostly excluded from prorating schemes.

10. For the further analysis of edit-rules and adjustment methods it is convenient to express the restrictions in matrix notation. as $\mathbf{Ax} = \mathbf{0}$, where $\mathbf{A}$ is the *constraint* or *restriction* matrix. For the restrictions a1 – a3 this matrix is given by

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

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

11. Successive moves from one non-zero element to another either in the same row or in the same column of a restriction matrix generates a *path*. A set of variables are *connected* (to each other) if there is a path between any two of them. A variable that is not connected with any other variables is an *isolated* variable. In the accounting matrix $\mathbf{A}$ above, the variable Employees (x_2) is an isolated variable, and the rest of the variables are connected. An example of a path between Profit (x_1) and Turnover main (x_3) is ($x_1 \rightarrow x_5 \downarrow x_5 \rightarrow x_3$). Given a set of connected variables, a *joint* among them is a variable that has more than one non-zero element in the corresponding column of the restriction matrix. Different constraints are connected to each other through the joints. Indeed, two subsets of variables are *separated* from each other by a set of joints if any path between two variables, i.e. one form each subset, must pass through the set of joints. The joints of the matrix $\mathbf{A}$ here are (x_5, x_8). Moreover, (x_3, x_4) are separated from all the other variables by x_5, and (x_6, x_7) are separated from the others by x_8, and x_1 by (x_5, x_8).

12. We observe the following given the restriction matrix.

(a) An isolated variable, such as Employees (x_2) in Table 1, can be imputed freely without causing consistency problems.

(b) Provided *all* the joints are observed or given by external sources, such as (x_5, x_8) in Table 1, the consistency problem among the set of connected variables can be resolved by dealing with one constraint at a time, e.g. by separate prorating for each constraint.

(c) Adjustments of a subset of variables do not cause consistency problem for the remaining connected variables given the joints that separate these variables. In Table 1, for instance, (x_3, x_4) can be adjusted freely given x_5 without causing consistency problem for the other variables.

(d) The imputation (or adjustment) of any variable may potentially cause consistency problems for all the connected variables that are not separated by the given joints. In both response patterns of Table 1 the joint x_5 is given. However, only (x_3, x_4) are separated from the other variables by x_5. The consistency problem among the rest of the connected variables (x_1, x_6, x_7, x_8) can be resolved using a traditional method in two steps: first, adjust the remaining joints (i.e. x_8) in a consistent manner given the observed joints (i.e. x_5); next, consider x_8 to be fixed and adjust the rest of the variables as in situation (a) and (b). Thus, for response pattern (I), one might first impute x_8, say, proportionally to x_5. The remaining variables can be adjusted with regard to one-constraint at a time. For the response pattern (II), however, x_6 is also observed, such that it no longer seems desirable if one is to impute x_8 *without* taking into account x_6, because the two are connected. There arises therefore a need to deal with all the constraints that are connected to x_8 simultaneously, which requires an approach beyond the realm of traditional single-constraint adjustment methods such as prorating.

(e) Constraints for which it is optimal to adjust the variables in the same direction (either an increase or a decrease) can be identified from the restriction matrix as rows in which the entries corresponding to adjustable variables have the same sign.

C. An optimization approach

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

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

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

$$\begin{array}{l} \tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}_0)\,. \\ s.t.\;\; \mathbf{A}_1\tilde{\mathbf{x}} = \mathbf{0} \;\; and \;\; \mathbf{A}_2\tilde{\mathbf{x}} \le \mathbf{0} \end{array} \qquad (1)$$

For ease of exposition we will write the constraints in (1) more compactly as $\mathbf{A}\tilde{\mathbf{x}} \le \mathbf{0}$.

III. Minimum adjustment approaches

A. Loss-functions and adjustment models

14. The conditions for a solution to the problem formulated in (1) can be found by inspection of the Lagrangian for this problem, which can be written as

$$L(\mathbf{x}, \boldsymbol{\alpha}) = D(\mathbf{x}, \mathbf{x}_0) + \sum\nolimits_k \alpha_k \left(\sum\nolimits_i a_{ki} x_i\right), \qquad (2)$$

with $\boldsymbol{\alpha}$ a vector of Langrange multipliers, one for each of the constraints *k*, a_{ki} the element in the *k*-th row (corresponding to constraint *k*) and *i*-th column (corresponding to variable x_i) of the restriction matrix A and *D*(**x,** **x**$_0$) a loss-function measuring the distance or discrepancy between **x** and **x**$_0$. From optimisation theory it is well known that for a convex function *D*(**x,** **x**$_0$) and linear (in)equality constraints, the solution vector $\tilde{\mathbf{x}}$ must satisfy the so-called Karush-Kuhn-Tucker (KKT) conditions (e.g. Luenberger, 1984). One of these conditions is that the gradient of the Lagrangian w.r.t. **x** is zero, i.e.

$$L'_{x_i}(\tilde{x}_i, \boldsymbol{\alpha}) = D'_{x_i}(\tilde{x}_i, \mathbf{x}_0) + \sum\nolimits_k \alpha_k a_{ki} = 0, \qquad (3)$$

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

15. **Least Squares** (LS). First, we consider the least squares criterion to find an adjusted x-vector that is closest to the original unadjusted data, that is: $D(\mathbf{x}, \mathbf{x}_0) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0)$, is the Least Squares (LS) criterion, $D'_{x_i}(\tilde{x}_i, \mathbf{x}_0) = \tilde{x}_i - x_{0,i}$, and we obtain from (3)

$$\tilde{x}_i = x_{0,i} + \sum\nolimits_k a_{ki}\alpha_k\,. \qquad (4)$$

This shows that the least squares criterion results in an additive structure for the adjustments: the total adjustment to variable $x_{o,i}$ is the sum of adjustments to each of the constraints *k*. These adjustments consist of an adjustment parameters α_k that describes the amount of adjustment due to constraint *k* and variables a_{ki} (with values 1,-1 or 0) that describe whether variable $x_{o,i}$ is adjusted by α_k, $-\alpha_k$ or not at all. The adjustment

16. **Weighted Least Squares** (WLS). For the weighed least squares criterion, $D(\mathbf{x},\mathbf{x}_0)=\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T\,Diag(\mathbf{w})(\mathbf{x}-\mathbf{x}_0)$, with $Diag(\mathbf{w})$ a diagonal matrix with a vector with weights along the diagonal, we obtain from (3)

$$\tilde{x}_i = x_{0,i} + \frac{1}{w_i}\sum\nolimits_k a_{ki}\alpha_k\,. \tag{5}$$

Contrary to the least squares case where the amount of adjustment is to a constraint is equal in absolute value (if it is not zero) for all variables in that constraint, the amount of adjustment now varies between variables according to the weights: variables with large weights are adjusted less than variables with small weights. The weighted least squares approach to the adjustment problem has been applied by Thomson et al. (2005). They used weights of 10,000 for observed values and weights of 1 for imputed values. Effectively, this means that if a consistent solution can be obtained by changing only imputed variables, this solution will be found. Otherwise (some of the) observed variables will also be adjusted.

17. One specific form of weights that is worth mentioning is obtained by setting the weight w_i equal to $1/x_{0,i}$ resulting, after dividing by $x_{0,i}$ in the adjustment model

$$\frac{\tilde{x}_i}{x_{o,i}} = 1 + \sum\nolimits_k a_{ki}\alpha_k\,, \tag{6}$$

which is an additive model for the *ratio* between the adjusted an unadjusted values. It may be noticed that the expression on the right-hand side of (6) is the first-order Taylor expansion (i.e. around 0 for all the α_k 's) to the multiplicative adjustment given by

$$\frac{\tilde{x}_i}{x_{o,i}} = \prod\nolimits_k (1 + a_{ki}\alpha_k) \tag{7}$$

From (6) we see that the α_k 's determine the difference from 1 of the *ratio* between the adjusted and unadjusted values, which is usually much smaller than unity in absolute value (e.g. an effect of 0.2 implies a 20% increase due to adjustment which is large in practice). The products of the α_k 's are therefore often much smaller than the α_k 's themselves, in which cases (6) becomes a good approximation to (7), i.e. the corresponding WLS adjustments is roughly given as the product of the constraint-specific multiplicative adjustments.

18. **Kullback-Leibler divergence** (KL). The Kullback-Leibler divergence measures the difference between **x** and **x₀** by the function $D_{KL} = \sum\nolimits_i x_i(\ln x_i - \ln x_{0,i} - 1)$. It can be shown that for this discrepancy measure, the adjustment model takes on the following form

$$\tilde{x}_i = x_i \times \prod\nolimits_k \exp(-a_{ik}\alpha_k). \tag{8}$$

In this case the adjustments have a multiplicative form and the adjustment for each variable is the product of adjustments to each of the constraints. The adjustment factor $\gamma_k = \exp(-a_{ik}\alpha_k)$ in this product represents the adjustment to constraint *k* and equals 1 for a_{ik} is 0 (no adjustment), $1/\gamma_k$ for a_{ik} is 1 and the inverse of this factor, γ_k, for a_{ik} is -1.

B. The successive projection algorithm

19. The optimization problem (1) can be solved explicitely if the objective function is the (weighted) least squares function and there are only equality constraints. For other convex functions and/or inequality constraints, problem (1) can be solved by several optimization algorithms. In this section we briefly review a very simple such algorithm that is easy to implement and contains as a special case the – among survey methodologists well known – Iterative Propoprtional Fitting (IPF) algorithm for adjusting contingency tables to new margins. Algorithms of this type are extensively discussed in Censor and Zenios (1997) and applications to adjustment problems are described in De Waal et al. (2011). The algorithm is an iterative procedure in which the edit-constraints are used one at a time. It starts by minimally adjusting the original inconsistent vector $\mathbf{x}_0$ to one of the constraints. The resulting solution is then updated such that a next constraint is satisfied and the difference with the previous solution is minimized and so on. In this way, if there are K constraints, K minimal adjustment problems with a

single constraint each need to be solved, which is much easier than a simultaneous approach. After the K-steps one iteration is completed and a next iteration starts that will again sequentially adjust the current solution to satisfy each of the constraints.

20. To describe this algorithm we must make the distinction between adjustable variables and fixed variables more explicit. Without loss of generality we can separate the adjustable variables from the fixed values in $\mathbf{x}$ by the partitioning $\mathbf{x} = (\mathbf{x}_m^Y, \mathbf{x}_o^T)^T$ where $\mathbf{x}_m$ denotes the subvector of $\mathbf{x}$ containing the imputed values (or more generally, the adjustable values) and $\mathbf{x}_o$ the subvector containing the remaining observed (or, more generally fixed) values. The restriction matrix $\mathbf{A}$ can then be partitioned conformably as $\mathbf{A} = (\mathbf{A}_n, \mathbf{A}_o)$. From $\mathbf{Ax} \leq \mathbf{0}$ we then obtain as constraints for the adjustable variables: $\mathbf{A}_\mathbf{m}\mathbf{x}_m \leq -\mathbf{A}_o\mathbf{x}_o = \mathbf{b}$, say.

21. In an iteration t the algorithm cycles through the constraints adjusting the current $\mathbf{x}$-vector to each of them. For equality constraints this adjustment solves the minimization problem

$$\begin{aligned}\mathbf{x}_m^{t,k} &= \arg\min_\mathbf{x} D(\mathbf{x}, \mathbf{x}_m^{t,k-1}) \quad \text{s.t. } \mathbf{a}_{m,k}\mathbf{x}_m^{t,k} - b_k = 0 \\ &= P_{D,k}(\mathbf{x}_m^{t,k-1}),\end{aligned}$$

with $\mathbf{a}_{m,k}$ the row of $\mathbf{A}_m$ corresponding to constraint k and b_k the corresponding element of $\mathbf{b}$. The equality constraint $\mathbf{a}_{m,k}\mathbf{x}_m^{t,k} = b_k$ defines a hyperplane and $\mathbf{x}_m^{t,k}$ is the vector on this hyperplane closest to $\mathbf{x}_m^{t,k-1}$. Therefore, it is the (generalized) projection with respect to the distance D of $\mathbf{x}_m^{t,k-1}$ on this hyperplane, which is denoted above by $P_{D,k}(\mathbf{x}_m^{t,k-1})$. For the least-squares criterion this is the usual orthogonal Euclidean projection. As the algorithm cycles through the constraints the $\mathbf{x}_\mathbf{m}$-vector is projected successively on each of the corresponding hyperplanes. and converges to the solution which is on the intersection of these hyperplanes.

22. For the least squares criterion the solution this projection step is given by

$$\mathbf{x}_m^{t,k} = \mathbf{x}_m^{t,k-1} + \bar{r}^{t,k} a_{m,k}^T \text{ with } \bar{r}^{t,k} = (b_k - \mathbf{a}_{m,k}\mathbf{x}_m^{t,k-1})/(\mathbf{a}_{m,k}\mathbf{a}_{m,k}^T)\,.$$

Note that $\bar{r}^{t,k}$ is a kind of "average" residual for constraint k since, if the values of $\mathbf{a}_{m,k}$ are confined to 0, 1 or - 1, then $\mathbf{a}_{m,k}\mathbf{a}_{m,k}^T$ is the number of adjustable values in constraint k.

23. For the KL-criterion, the projection cannot be expressed in closed form for general $\mathbf{a}_k$. However, if the elements of this vector are all either zero or one, which occurs when the constraint is that a sum of $\mathbf{x}_m$-values equals a fixed value b_k, the adjustment to an equality constraint k can be expressed as

$$\begin{aligned}x_{m,i}^{t,k} &= x_{m,i}^{t,k-1}\rho^{t,k} \quad \text{if } a_{m,k,i} = 1 \\ &= x_{m,i}^{t,k-1} \qquad\quad \text{if } a_{m,k,i} = 0\end{aligned}$$

where the adjustment factor $\rho^{t,k}$ is given by the rate of violation of constraint k: $\rho^{t,k} = b_k / (\sum_i a_{m,k,i} x_{m,i})$. In this case the resulting algorithm is equivalent to the IPF-algorithm that, when applied to a rectangular contingency table, adjust the counts in the table to new row- and column-totals by multiplying, successively, the counts in each row by a factor such that they add up to the new row-total and similarly for the columns.

24. For inequality constraints, the constraint can be satisfied with "slack", i.e. $\mathbf{a}_{m,k}\mathbf{x}_m^{t,k}$ is strictly smaller than b_k. In that case it may be possible to improve on the objective function by removing (some of) the adjustment to constraint k to the extend that either all adjustment is removed or the constraint becomes satisfied with equality. To accomplish this we first undo the adjustment made in the previous iteration to this constraint. If the constraint becomes violated, the projection step is performed with the result that the constraint becomes satisfied with equality, which is the minimum feasible adjustment. If after undoing the previous adjustment the constraint is not violated, no adjustment is performed.

IV. Generalized ratio adjustments

25. The distance metrics considered above can be characterized as *decomposable*, in the sense that the overall distance between two vectors is given as a (weighted) sum of 'distances' between the corresponding components. A consequence is that a variable that does not stand in any constraints will retain the initial (donor) value under the minimum adjustment approach. But this might not always seem reasonable. For instance, for the response pattern 1 where Total Turnover is the only observed variable, it seems natural if one chooses to adjust all the donor values by the ratio between this observed Turnover value and the corresponding donor value, motivated by a common ratio-model-like assumption.

26 It is important to distinguish between plausible adjustments based on a statistical assumption and necessary adjustments based on an edit constraint. While a ratio-model assumption can be used to motivate a proportionality constraint, it is a 'soft' one that the edited record does not necessarily have to satisfy, in contrast to a 'hard' proportionality edit constraint such as "Value = Price * Quantum" that must be imposed. Notice also that some range restrictions may be statistical and soft in the same way, such as a range restriction for the ratio between Total Wage Cost and Number of Employees.

27. A general question is therefore how to take into account the various plausible 'soft' statistical assumptions in addition to the edit constraints? One possibility is first to carry out a prediction of the missing values based on appropriate statistical assumptions and utilizing both the observed and the donor values. Since the predicted values do not necessarily satisfy the edit constraints, one may apply the minimum adjustment approach afterwards to the predicted values instead of the donor values. Two difficulties may easily arise. Firstly, two steps of processing are required, which complicates both the implementation and the assessment of the statistical properties of a chosen procedure. Secondly, it may be difficult to formulate an explicit statistical model for prediction that is able to accommodate the various partial missing patterns that may occur. For instance, a straight-forward ratio adjustment may be intuitive in response pattern 1. But how should it be adapted to pattern 2, where the observed values do not share a common ratio, and there is no obvious choice of the size variable among the observed ones?

28. As a second possibility, one may add to the distance metric some suitable penalty terms based on statistical assumptions. Take as an example the ratio assumption between Total Wage Cost and Number of Employees. Let z be, say, the difference between the ratio based on the donor values and that on the adjusted values, and one would like it to be as close to zero as possible. One can then add to the distance metric a penalty term $\omega_z z^2$, where ω_z is a tuning parameter for controlling the weight of the penalty. Obviously, there are other ways to set up the penalty. What is important is to be able to bring into the minimization problem the variables that do not stand in any edit constraints. But it is not difficult to envisage that such a remedy would complicate the practice considerably, because suitable penalties must be constructed for all the variables that do not directly stand in any edit constraints. Indeed, for each one of them, one needs to formulate plausible statistical assumptions that necessarily involve some variables that do stand in at least one edit constraint, as well as to set the tuning parameters.

29. For a potential general approach suitable for automated data processing, consider the following alternative *loss function*. Assume component-wise multiplicative adjustments given by

$$x_j = x_{0,j}\delta_j$$

for $j = 1,...,J$. Put

$$\Delta = \frac{1}{2}\sum_j (\delta_j - \bar{\delta})^2 , \quad \text{where} \qquad \bar{\delta} = \frac{1}{J}\sum_j \delta_j$$

Minimizing Δ subjected to the edit constraints yields what we call the *generalized ratio adjustments*.

30. The first thing to notice is that, unlike the least-square-type of distance metrics considered earlier, the empirical variance of the component-wise adjustments is a *non-decomposable* loss function, where each adjustment is dependent on the other adjustments. Given non-decomposability as such, even the values that are not explicitly involved in any constraints will be adjusted because of the changes made to the variables that are constraint-bound. Here lies the potential of using a non-decomposable loss function

to handle all the unconstrained variables in a practical production setting, without a detailed analysis of the various response patterns and their interactions with the edit constraints.

31. Next, we notice that, provided a single observed value such as in response pattern 1, the generalized ratio adjustments are reduced to a global proportional adjustment according to the ratio between this observed value and the corresponding donor value, confirming to the ratio-adjustment intuition in this case. More generally, we consider the empirical-variance loss function to aim at a kind of most-uniform adjustment solution as a generalization of the ratio-model adjustments. For instance, for response pattern 2, where there are three observed values that do not share a common ratio towards the corresponding donor values, the generalized ratio adjustments are given by the component-wise ratios that deviate least from each other. To formulate an explicit statistical model as an extension of the simple ratio model in order to handle exactly this response pattern is not as practical in a production setting.

32. Thirdly, we notice that there is no need for additive generalized ratio adjustments. Put

$$x_j = x_{0,j} + \delta_j^A = x_{0,j}\delta_j^M$$

for $j = 1,...,J$, where each additive adjustment δ_j^A corresponds to a multiplicative one δ_j^M, provided the $x_{0,j}$'s are non-zero. The component-wise ratio adjustment can equivalently be given as

$$x_j / x_{0,j} = 1 + \delta_j^A / x_{0,j} = \delta_j^M$$

In other words, the same generalized ratio adjustments can be obtained by minimizing the empirical variance of δ_j^M (i.e. over j) in the case of multiplicative adjustments or, equivalently, by minimizing the empirical variance of $\delta_j^A / x_{0,j}$ in the case of additive adjustments. The additive adjustments can only make a difference if the initially zero donor values are allowed to be non-zero through the adjustment. But since ratio adjustment of a zero value is not well defined, it would no longer make sense in such cases to adopt a ratio-type of motivation for the distance metric.

33. Finally, there is an intrinsic difference between an approach for most-uniform adjustments and one for direct minimum adjustments. In the special case where a unit has no observed value at all, the minimum adjustment approach would lead to imputation of all the donor values *without* any adjustment, whereas the most-uniform adjustments are apparently not well defined since e.g. any global proportional adjustment of the donor record minimizes the empirical-variance loss function above. In this sense the generalized ratio adjustments is a genuine partial imputation method, because some constraints towards observed values are necessary in order to identify a unique solution. In practice, this is hardly a problem for donor imputation, because it is always possible to turn a unit imputation problem into a partial one, by incorporating the information for donor identification as part of the data vector to be dealt with. More specifically, let x be the statistical variables of interest as above. Let z contain the variables that are used for donor identification, which may or may not contain variables that are subjected to missing, but must contain some variables that are always known, because otherwise a donor may not be identifiable. Let (x, z) be the combined vector of variables, with possible overlap between x and z, which must be 'observed' for the variables that are always known for donor identification. Regardless of whether these values actually match between the donor and the recipient, every instance of unit imputation of x can now be treated as an instance of partial imputation of (x, z). In this way unit imputation without adjustment becomes a special case of partial imputation, which can be motivated in cases where the donor and the recipient exactly match on all the values used for donor identification.

V. Example revisited

34. The different methods (LS, WLS and KL) have been applied to the example record of section II. For the WLS method we used as weights the inverse of the $\mathbf{x}_0$-values so that the relative differences between $\mathbf{x}$ and $\mathbf{x}_0$ are minimized and the adjustments are proportional to the size of the $\mathbf{x}_0$-values. For this choice of weights, the KL- and WLS-methods lead to results that are equal up to the first decimal. The results of the different methods, for both response patterns in Table 1, are given in Table 2. The observed values that are treated as fixed are shown in bold, the (other) imputed values are adjustable.

35. For both response patterns, the LS adjustment procedure leads to one negative value for *Turnover other*, which is not allowed for this variable. Therefore the LS-procedure was run again with a non-negativity constraint added for the variable *Turnover other*. This results simply in a zero for that variable and a value of 950 for *Turnover main* to ensure that *Turnover = Turnover main + Turnover other.* Without the non-negativity constraint, the LS-results clearly show that for variables that are part of the same constraints (in this case the pairs of variables x_3, x_4 and x_6, x_7. that are both appearing in one constraint only), the adjustments are equal: -40 for x_3, x_4 and -16 for x_6, x_7. *Total costs* (x_8) is part of two constraints and therefore the total adjustment to this variable consists of two additive components. One component to adjust to the constraint *a1*:x_1 - x_5 + x_8 = 0 (*Profit = Turnover – Total Costs*) and one component to adjust to.*a3*: x_8 - x_6 - x_7 = 0 (*Total Costs = Wages + Other costs*). For response pattern 1, the first component is minus 48 - which is also the single adjustment component for *Profit* - and the second component is 16 – which is also the single adjustment component for *Wages* and *Other costs* (with opposite sign). These two components add up to the adjustment of -32.

Table 2. Example business record and adjusted values.

Variable	*Name*	*Response pattern 1*				*Response pattern 2*			
		Imputed Unadj.	LS	WLS/ KL	Gen. Ratio	Imputed Unadj.	LS	WLS/ KL	Gen. Ratio
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	1000	960	922	922	1000	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

Imputed Unadj. = Imputed unadjusted values.
LS = adjustred values according to the LS criterion.
WLS/KL = adjusted values according to the WLS or KL criterion.
Gen. Ratio = adjusted values with the Generalized Ratio method.

36. The results for the WLS/KL solution show that for this weighting scheme the adjustments are larger, in absolute value, for large values of the imputed variables 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 (as KL-type adjustments) also clearly shows since 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 1 is also equal (to 0.94) because these variables are in the same single constraint and so the ratio between these variables is unaffected by this adjustment. However the ratio of each of these variables to *Total Costs* is not unaffected 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 subject to two adjustment factors.

37. As expected, the generalized ratio adjustments reduce to a global proportional adjustment of all the imputed values by a ratio of 0.922 (=950/1030) for response pattern 1, including the variable Employee. This is a main difference from the minimum-adjustment approaches that are based on decomposable loss functions. For response pattern 2, the generalized ratio adjustments are closer to the WLS/KL solution. The empirical variance of the multiplicative factors (i.e. proportional to the loss function Δ) is 0.0270 by the generalized ratio adjustments, it is 0.0276 for the WLS/KL solution, but is increased to 0.1434 for the LS solution. The relative sum of squared changes for all the variables, i.e. twice the loss function for the WLS solution, is 50.6 for the WLS/KL solution, it is 51.6 for the generalized ratio adjustment, and is increased to 78.0 for the LS solution. Finally, the un-weighted sum of squared change, i.e. twice the loss function for the LS solution, is 20925 for the LS solution, it is 23976 for the WLS/KL solution and 25090 for the generalized ratio adjustment. In terms all the three loss functions, therefore, the generalized ratio adjustments are closer to the WLS/KL solution.

VI. Conclusion

38. Imputation is generally used as a method to compensate for partially missing values in many surveys. Often, especially in business surveys, the data have to satisfy many carefully specified edit-rules, derived from logical relations or accounting equations. Many imputation methods will not ensure that these edit-rules are satisfied and, as a consequence, inconsistencies in the imputed micro-data will arise. Using the information from the edit-rules to adjust the imputed records such that they conform to these edit-rules will often enhance the data quality and will assure the consistency between estimates of target parameters at any level of aggregation.

39. In this paper we have formulated an optimization approach to this adjustment problem. Two variations of this approach have been considered. The first one seeks to minimize the adjustments needed to ensure consistency. In this approach only variables that appear in edit-rules will be adjusted because other variables will not cause inconsistency problems. The second is called the "generalized ratio" approach. In this approach all imputed values are adjusted and the adjustments are as uniform across variables as possible. The inconsistency is seen as an indication that there are systematic differences between the donor values and the observed values and it is therefore plausible to adjust all imputed variables.

40. The optimization approach to the inconsistency problem provides a general methodology that extends beyond the traditional single-constraint adjustment methods such as prorating. This approach handles all constraints simultaneously and, if variables appear in more than one constraint then they are adjusted according to all of them. Besides being an optimal method in the sense according to the chosen distance metric or loss function, this simultaneous approach also has the practical advantage that there is no need to specify the order in which the constraints are to be handled.

41. For the minimum adjustment approach several distance metrics have been analysed. It is shown that (weighted) least-squares loss leads to additive adjustments and that minimizing the Kullback-Leibler information criterion (KL) leads to multiplicative adjustments. It is also shown that for a specific choice of weights the weighted least-squares solution is an approximation to the KL-solution.

42. In most cases we expect that multiplicative adjustments obtained by the minimum adjustment approach and the generalized ratio adjustments to yield similar results. The main difference seems to be the following. The minimum adjustment solution can be decomposed into the adjustments that correspond to each relevant constraint in a straight-forward manner. However, the same decomposability also means that special care needs to be given to the unconstrained (or isolated) variables in the data set. The generalized-ratio adjustments can be motivated as a generalization of the global ratio-adjusted imputation, which implicitly achieves the effects of corresponding ratio-model-like statistical assumptions, including the isolated variables, which can be practical in a production setting.

VII. References

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

Cenzor, Y., and S.A. Zenios (1977), *Parallel Optimization. Theory, Algorithms, and Applications.* Oxford University Press, New York.

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

Luenberger, D. G. (1984), *Linear and Nonlinear programming, second edition.* Addison-Wesley, Reading.

Thomson, K., J. T. Fagan, B. L. Yarbrough and D. L. Hambric (2005), *Using a Quadratic Programming Approach to Solve Simultaneous Ratio and Balance Edit Problems*. Working paper 32, UN/ECE Work Session on Statistical Data Editing, Ottawa.