1. Purpose of the method

Imputations (predictions for missing values) are derived directly from the values that are observed in the same record, using derivation rules that do not contain any parameters to be estimated such as is the case in models.

2. The related scenarios

2.1. The method applies to a single data set composed of microdata

The data set can be the result of a combination of several data sets. It mainly refers to data configurations 1 to 5 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used on an equal footing.

2.2. Statistical tasks: Data editing and imputation

3. Description of the method

Many deductive imputations can be performed using simple rules in 'if-then' form, for example:

if (total labour costs = 'missing' and employees on the payroll = 0)

then total labour costs := 0.

These rules are compiled by subject-matter experts.

Variable restrictions may help to determine deductive imputations. A particularly rich source for deductive imputations is formed by the extensive systems of equations that may be hold for quantitative variables (e.g., in case of Structural Business Statistics). A typical survey may involve around 100 variables with 30 equality restrictions. Most of these *equality restrictions* have the general form

Total = Subtotal_1 + Subtotal_2 + … + Subtotal_s.

If, in such a case, one of the subtotals or the total is missing, it is immediately clear with which value the missing variable should be imputed: there is a single equation with a single unknown, so a unique solution exists. More generally, we may encounter several variables with missing values that are involved in several inter-related equality restrictions. This means we have a system of equations with multiple unknowns, for which it is not immediately clear whether the values of some missing variables are uniquely determined by this system, and, if so, what these unique values would be. In such a context, deductive imputations may found by using techniques from linear algebra (Pannekoek, 2006). For categorical variables, similar methods are discussed in De Waal et al. (2011, Section 9.2.4).

4. Examples

Consider a fictitious survey with eleven variables that should satisfy five equality restrictions:

$$\begin{cases} y_1 + y_2 = y_3 \\ y_2 = y_4 \\ y_5 + y_6 + y_7 = y_8 \\ y_3 + y_8 = y_9 \\ y_9 - y_{10} = y_{11} \end{cases}$$

Suppose that we want to use deductive imputation to treat as many missing values as possible in the following incomplete record (where '–' indicates a missing value):

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}
154	–	166	–	–	–	–	25	–	204	–

By using the algorithm in Pannekoek (2016), we obtain the following partially imputed record:

y_1	$\tilde{y}_2$	y_3	$\tilde{y}_4$	y_5	y_6	y_7	y_8	$\tilde{y}_9$	y_{10}	$\tilde{y}_{11}$
154	12	166	12	–	–	–	25	191	204	-13

5. Input data (characteristics, requirements for applicability)

A data set containing microdata with missing values.

6. Output data (characteristics, requirements).

A data set containing partially imputed microdata, which is an updated version of the first input data set.

7. Tools that implement the method

R package deducorrect.

8. Appraisal

Deductive imputation is most effective when it is applied at the very beginning of the imputation process, after the removal of erroneous values, but before other forms of imputation have been used. In this way, other imputation methods have more nonmissing auxiliary variables available, e.g., to estimate model parameters.

The method should be used, in principle, only for imputing values that can be derived with certainty from the observed values. In all other cases, it is usually better to use non-deductive methods, such as model-based imputation or donor imputation.

9. References

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

Memobust (2014). Deductive Imputation, in *Memobust Handbook on Methodology of Modern Business Statistics.* https://ec.europa.eu/eurostat/cros/content/memobust_en

Pannekoek, J. (2006), Regression Imputation with Linear Equality Constraints on the Variables. Working Paper, UN/ECE Work Session on Statistical Data Editing, Bonn.

Method: Model based imputation

1. Purpose of the method

The objective in model-based imputation is to find a predictive model for each target variable in the data set that contains missing values. The model is fitted on the observed data and subsequently used to generate imputations for the missing values.

2. The related scenarios

2.1 The method applies to a single data set composed of microdata. The data set can be the result of a combination of several data sets. It mainly refers to the data configurations 1 to 5 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used on an equal footing.

2.2 Statistical tasks: Data editing and imputation

3. Description of the method

The objective in model-based imputation is to find a predictive model for each target variable in the data set that contains missing values. The model is fitted on the observed data and subsequently used to generate imputations for the missing values. This is a valid approach when the missing data mechanism is Missing at Random (MAR), in a sense the missingness can be fully accounted for by variables where there is complete information (Little and Rubin, 2002). Many practical applications use a separate model for each variable in the data set. Some multivariate extensions are discussed in the section. By the definition, it is clear that all methods based on models available in the statistical literature may be used, provided that we are able to estimate their parameters in the presence of missing values. As an illustrative example, we described one of the model most applied in practice: the linear regression.

A standard linear regression model for the prediction of y given a set of auxiliary variables $x_1,\ldots,x_q$ may be expressed as

$$y=\alpha+\beta_1x_1+\cdots+\beta_qx_q+\varepsilon\ , \qquad (1)$$

with $\alpha,\beta_1,\ldots,\beta_q$ unknown parameters and ε a disturbance term, where it is assumed that the disturbances for all units are drawn independently from the same normal distribution with mean 0 and variance σ^2.

The parameters in model (1) are estimated – usually through ordinary least squares – from the records for which both y and the auxiliary variables are observed. This results in a prediction for y given the auxiliary variables:

$$\hat{y} = a + b_1 x_1 + \cdots + b_q x_q \, , \tag{2}$$

with $a, b_1, \ldots, b_q$ denoting the least squares estimates of $\alpha, \beta_1, \ldots, \beta_q$. Assuming that the auxiliary variables are always observed, this predicted value can be computed for both item respondents and item nonrespondents on y.

There are now two generic ways to obtain an imputation $\tilde{y}_i$ from the regression model: without a disturbance term or with a disturbance term. In the first case, the predicted value from (2) is substituted directly for the missing value:

$$\tilde{y}_i = \hat{y}_i = a + b_1 x_{1i} + \cdots + b_q x_{qi} \, . \tag{3a}$$

This results in a deterministic imputation. In the second case, we add a disturbance to the predicted value, i.e., we impute:

$$\tilde{y}_i = \hat{y}_i + e_i = a + b_1 x_{1i} + \cdots + b_q x_{qi} + e_i \, . \tag{3b}$$

The disturbance e_i can be a random draw from the normal distribution with mean 0 and variance σ^2, to be in line with the posited regression model (1). (Actually, σ^2 is unknown in practice and is often estimated by the residual error of the fitted model). Adding a disturbance results in a stochastic imputation.

It should be noted that mean imputation can be seen as a special case of regression imputation, namely in the absence of auxiliary variables. Similarly, ratio imputation can be seen as a special case of regression imputation with one auxiliary variable and with the constant term fixed to 0. In this case, model (1) reduces to

$$y = \beta x + \varepsilon \, .$$

Under the alternative assumption that the variance of the disturbances equals $\sigma^2 x$ rather than σ^2, the weighted least squares estimate for β is just the observed ratio $\hat{R} = \sum_{k \in obs} y_k \Big/ \sum_{k \in obs} x_k$ with obs denoting the set of item respondents for variable y.

The model-based so far described impute a data set on a variable-by-variable basis. There are methods that adopt a multivariate approach to imputation. Although these multivariate methods are more complex to use, they have some theoretical advantages, e.g., the relationships between and all other variables in the data set are better preserved.

Raghunathan et al. (2001) proposed a method, known as *sequential regression imputation*. Under this approach, one models the distribution of each target variable separately, conditional on the values of the other variables. This yields a set of single-variable regression models, which have to be estimated in an iterative manner. To do this, the following procedure can be used:

1. Initialise the procedure by imputing each missing value in the original data set by a simple method (e.g., mean imputation).
2. For each variable in turn:
    a. Estimate the parameters of the conditional regression model using all records in the current data set for which this variable was originally observed.
    b. Use the estimated conditional model to impute the originally missing values for this variable. This updates the current data set for the next iteration.
3. Repeat Step 2 until 'convergence'.

Although not frequently used in the NSIs, multiple imputation is an approach particularly useful to deal with missing data. It consists of imputing several times the data set, and then combining the estimates of each imputed data to compute a final single estimate and the accuracy of the estimator by using appropriate formulas. Multiple imputation allows to take into account in the evaluation of the precision of estimates the additional variability due to the imputation procedure (Rubin, 1986).

More details on model based imputations can be found in Memobust (2014).

4. Examples

An illustrative example is taken from De Waal et al., (2011). Each year, Statistics Netherlands receives a version of the so-called Municipal Base Administration (MBA). The MBA contains, for each address, data on the people living at the address, including the family relations. Information on how the households living at the address are exactly composed is, however, lacking. For the annual Household Statistics it is essential to know which persons living at the same address constitute a household according to the definition applied at Statistics Netherlands. From 1999 on, the Municipal Base Administration is used to determine the main variables *Number of households* and *Household composition* from the structure of the family or families living at the address. For more than 90% of the addresses in the MBA the information for these derived variables can be constructed. For the remaining addresses, however, neither the number of households nor the exact household composition can be derived. For these addresses, imputation is used, with separate imputation models for different situations. Here we discuss the simplest type of addresses with unknown Household composition: addresses with two unrelated persons—that is, two persons that are not married or registered as each other's partner and who are not family of each other. For these addresses it is unknown whether the two persons together constitute one household or are both single and each has its own household. First, deductive imputation is applied, by means of a deductive rule: 'when both persons started to live at the address on the same date according to the MBA, then they are considered to constitute one household'. The remaining addresses are linked to the Labor Force Survey (LFS) and 1662 addresses with two persons were obtained. Based on these data an imputation model was built. By means of information obtained from the interviewers of the LFS and the actual data collected by means of the LFS, for each of the 1662 addresses it was determined whether the address contained one or two households. A logistic regression model was developed with age of persons as auxiliary variables. The result is used for each address with two unrelated persons in the MBA that did not link to the LFS to estimate the probability that the address contains two households. The estimated probability is used to draw either "two households" or "one household."

5. Input data (characteristics, requirements for applicability)

A data set containing microdata with missing values

6. Output data (characteristics, requirements)

A data set containing imputed microdata, which is an updated version of the first input data set

7. Tools that implement the method

Regression imputation with common types of models (e.g., linear regression, logistic regression) is provided as a standard feature in tools such as SPSS, SAS, Stata and R. Specialised packages are available for sequential regression imputation, such as IVEware (in SAS), and mice and mi (in R).

8. Appraisal

Imputation on variable-by-variable basis may give biased estimates of the relationships among variables. A multivariate approach better preserves those relationships.

The main practical advantage of the sequential regression approach lies in the flexibility provided by the use of separate, conditional regression models.

9. References

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

Little R.J.A., Rubin D.B. (2002). Statistical analysis with missing data, 2nd edition, John Wiley & Sons.

Memobust (2014). Model based imputation, in *Memobust Handbook on Methodology of Modern Business Statistics.* https://ec.europa.eu/eurostat/cros/content/memobust_en

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

Rubin D. B. (1986). Multiple imputation for nonresponse in surveys. New York, John Wiley & Sons.

1. Purpose of the method

The objective of donor imputation is to fill in the missing values for a given unit by copying observed values of another unit, the donor. Typically, the donor is chosen in such a way that it resembles the imputed unit as much as possible on one or more background characteristics.

2. The related scenarios

2.1. The method applies to a single data set composed of microdata

The data set can be the result of a combination of several data sets. It mainly refers to data configurations 1 to 5 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used on an equal footing, and when administrative data are used in a predictive setting.

2.2. Statistical tasks: Data editing and imputation.

3. Description of the method

The objective in donor imputation is to fill in the missing values for a given unit (the *recipient*) by copying the corresponding observed values of another unit (the *donor*). The term *hot deck* donor imputation applies when the donor comes from the same data set as the recipient.

Letting y_i denote the score of the i^{th} unit on the target variable y and using the index d for a donor, we can write the generic formula for hot deck donor imputation as:

$$\tilde{y}_i = y_d \,. \tag{1}$$

Typically, one searches for a donor that resembles the recipient as much as possible on one or more auxiliary variables. There exist different ways to select a donor, in the following we describe *random hot deck imputation*, *nearest-neighbour imputation*, and *predictive mean matching.* More methods can be found in de Waal et al (2011) and Memobust (2014).

In random hot deck imputation, imputation classes are formed based on categorical auxiliary variables. For each recipient unit i in a given imputation class, the group of potential donors consists of the units within the same class with y observed. Of these potential donors, one is selected at random and used to impute the recipient.

In nearest-neighbour imputation, the auxiliary variables are used to define a distance function $D(i,k)$ between units i and k, where i is the recipient and k is a potential donor. The *nearest neighbour* of unit i is defined as the respondent d that minimises a distance function. Many distances cab be used, an example is the 'minimax', assuming that the auxiliary variables ($x_1,\ldots,x_q$) are all quantitative

$$D_\infty(i,k) = \max_{j=1,\ldots,q} \left| x_{ji} - x_{jk} \right| ,$$

in this case, the nearest neighbour should not deviate strongly from the recipient on any auxiliary variable x_j.

Little (1988) described a variant of donor imputation known as predictive mean matching. In this imputation method, a linear regression is first performed of the target variable y on some auxiliary variables $x_1,\ldots,x_q$. The regression model is fitted on the data of units without item non-response. Next, the resulting regression equation is used to obtain predicted values $\hat{y}$ for all records. For item non-respondent i with predicted value $\hat{y}_i$, we select as donor the item respondent d for which the predicted value $\hat{y}_d$ is as close as possible to $\hat{y}_i$. Finally, the *observed* value y_d of the donor is imputed. The latter feature makes this method a form of donor imputation rather than model-based imputation.

4. Examples

In Istat, the main variables of Structural Business Statistics (SBS) for small and medium enterprises are based on integrated administrative data. Financial Statement, Studi di settore (Fiscal Authority survey that aims at evaluating the capacity of enterprises to produce income and at indirectly assessing whether they pay taxes correctly), and Tax Return are used to build a microdata file composed of the main economic variables. Since not all the variables are available in all the data sources, and the sources cover only subsets of the target population, the microdata file is a result of an imputation process. The imputation procedure is based on a combination of different techniques that are introduced to comply with requirements given by constraints, such as statistical relationships among main variables, balance edits, and presence of zero-inflated variables. The procedure is based on a combination of different techniques. According to simulations carried out to see which are the best imputation methods for the data at hand, the entire imputation process is composed by 4 sequential steps:

- deterministic imputation based on the guidelines of subject matter experts;
- imputation of the variables 'Turnover', 'Purchases of goods', 'Purchases of services' and 'Total Change in Stocks', through Predictive Mean Matching (PMM);
- imputation of the variables 'Changes in contract work in progress', 'Changes in internal work capitalized under fixed assets', 'Use of third party assets', 'Other operating charges' and 'Other income and earnings', through Nearest Neighbor Donor (NND);
- imputation of the variables 'Changes in stocks of raw materials and for resale', 'Changes in stock of finished and semi-finished products' through a two-step procedure composed by a logistic and a linear regression model.

The choice of each imputation method for different groups of variables is due to: the variable distribution characteristics (only positive, zero-inflated, etc.), the presence of a (weak/strong) linear relationship between variables and the presence of balance edits.

In this context, both the PMM and NND approaches have the advantage to recover live values from donors. Since the PMM technique relies on a multivariate normal model, it has been used to treat variables having a genuine continuous distribution. On the contrary, the NND method has been used to treat variables with distribution characterized by 0 inflation and a non-linear relation. For more details see Di Zio et al., 2016.

Statistics Belgium used a "cascade" of imputation methods, from the preferred one to the less desired one for imputing environmental expenditures by companies. The question "by environmental media" was not mandatory within a mandatory Structural Business Survey (thus a high level of mandatory auxiliary data was available).The different "preferred methods" depended on the strata, mainly of the sector. There were selected based on a previous study of correlations in order to verify that the method was robust enough. For example, temporal imputation (data from the same company, previous years) is preferred, NACE4 digits donors are preferred to NACE2 digits donors etc., but it might depend on the sector and the item (for example, investments can generally not be imputed because of seldom occurrence). Statistics Belgium used (by order of most frequent preference) the following methods (Kestemont, 2004):

1) (manual) Editing

2) Deterministic imputation

3) Serial imputation from 2000-2001 to 2002 (same respondent)

4) Serial imputation from 1999-2001 to 2002 (same respondent)

5) Serial imputation from 1999-2000 to 2002 (same respondent)

6) Temporal imputation using donors from 2001 to 2002

7) Temporal imputation using donors from 2000 to 2002

8) Temporal imputation using donors from 1999 to 2002

9) Trend imputation from 2001 to 2002

10) Trend imputation from 1999 to 2002

11) Sector imputation using factors (turnover, employment, wages, env. taxes)

12) Stratum imputation & imputing value of the nearest previous year

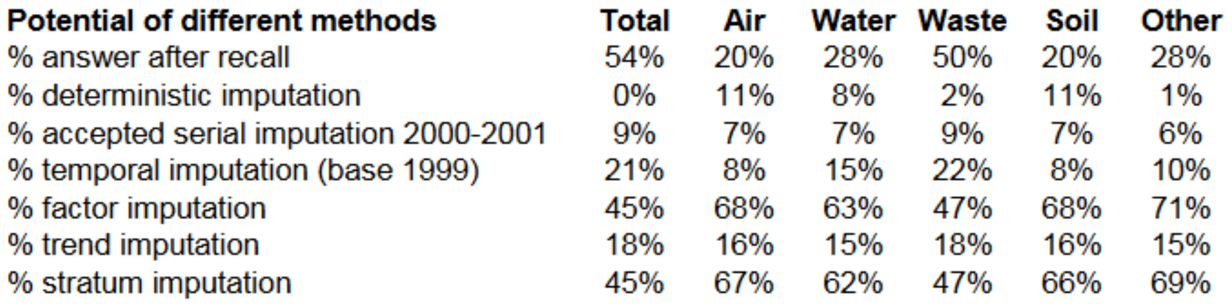
13) Imputation of environmental taxes on total current PAC exp

14) Secondary deterministic imputation

(15) Simple extrapolation)

External donors (from same stratum) where used in steps 6-8 (the donors give the trend to be applied to the missing item), 9 & 10 (trend of big companies applied to year-missing small companies), 11 (using correlation to other available data to impute item missing values) and 12 (simply taking the mean of the responses of the same stratum). Case 12 differs from simple extrapolation by the fact that it does not oblige to calculate a post-weight only for one missing item.

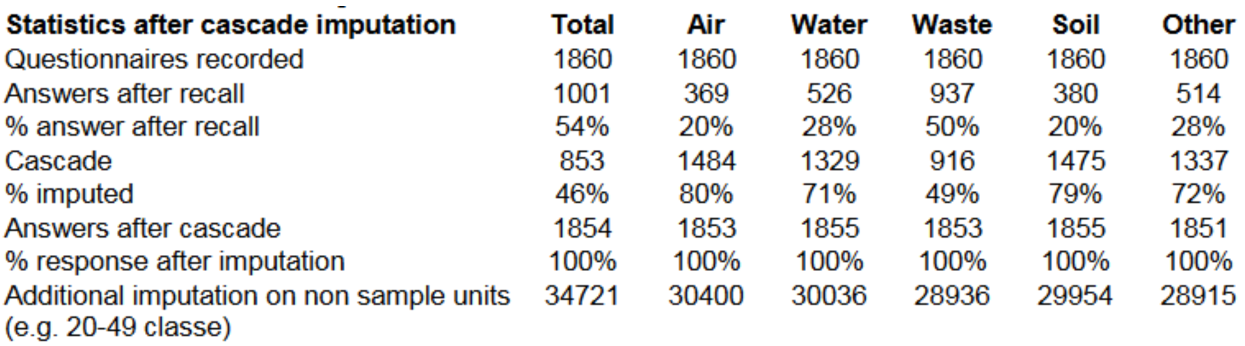
The potential of each method was the following:



Potential of different methods	Total	Air	Water	Waste	Soil	Other
% answer after recall	54%	20%	28%	50%	20%	28%
% deterministic imputation	0%	11%	8%	2%	11%	1%
% accepted serial imputation 2000-2001	9%	7%	7%	9%	7%	6%
% temporal imputation (base 1999)	21%	8%	15%	22%	8%	10%
% factor imputation	45%	68%	63%	47%	68%	71%
% trend imputation	18%	16%	15%	18%	16%	15%
% stratum imputation	45%	67%	62%	47%	66%	69%

Of course, these methods are not cumulative because a same data can be imputed from different methods. They can be used in cascade of preference.

The contributions from the all "cascade" of methods in the results was as follows:



Statistics after cascade imputation	Total	Air	Water	Waste	Soil	Other
Questionnaires recorded	1860	1860	1860	1860	1860	1860
Answers after recall	1001	369	526	937	380	514
% answer after recall	54%	20%	28%	50%	20%	28%
Cascade	853	1484	1329	916	1475	1337
% imputed	46%	80%	71%	49%	79%	72%
Answers after cascade	1854	1853	1855	1853	1855	1851
% response after imputation	100%	100%	100%	100%	100%	100%
Additional imputation on non sample units (e.g. 20-49 classe)	34721	30400	30036	28936	29954	28915

The resulting file was microdata without missing item values, with weights and ready to be extrapolated.

5. Input data (characteristics, requirements for applicability)

A data set containing microdata with missing values.

6. Output data (characteristics, requirements)

A data set containing imputed microdata, which is an updated version of the first input data set.

7. Tools that implement the method

Several R packages, among them StatMatch and mice. Banff (Statistics Canada) , CANCEIS (Statistics Canada).

8. Appraisal

In practice, one often encounters records with several missing values. In that case, the standard approach is to impute all missing values in a record from the same donor. This helps to preserve the multivariate relations between the imputed variables.

9. References

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

Di Zio M., Guarnera U., Varriale R. (2016), Estimation of the main variables of the economic account of small and medium enterprises based on administrative data. Rivista di Statistica Ufficiale n. 1.

Kestemont, B. (2004), Environmental expenditures by the Belgian industries in 2002, Statistics Belgium Working paper n°9, Direction générale Statistique et information économique, Brussels, http://statbel.fgov.be/fr/binaries/p009n009%5B1%5D_tcm326-34514.pdf

Little, R. J. A. (1988), Missing-Data Adjustments in Large Surveys. Journal of Business & Economic Statistics 6, 287–296.

Memobust (2014). Model based imputation, in Memobust Handbook on Methodology of Modern Business Statistics. https://ec.europa.eu/eurostat/cros/content/memobust_en.

1. Purpose of the method

The objective of imputation is to find a prediction at micro level for each target variable in the data set that contains missing values. We refer to longitudinal data when the same variables of the same units are measured several times at different moments.

2. The related scenarios

2.1. The method applies to micro and macro data set composed of longitudinal microdata

The data set can be the result of a combination of several data sets. It mainly refers to the data configuration 8 (deliverable 1). Statistical usages are mainly Direct tabulation, Substitution and supplementation for Direct collection, editing and imputation, indirect estimation where administrative and statistical data are used.

2.2. Statistical tasks: Data editing and imputation.

3. Description of the method

Imputation methods for longitudinal data usually take into account the historical information of each unit to define any type of imputation method. Let y_{it} be a missing value of unit *i* at period *t* on variable *y*. Then *y*-values of unit *i* at previous and subsequent periods can be used to create an imputed value $\tilde{y}_{it}$. In official statistics simple methods are generally used such as the *last observation carried forward* (the last observed value of a unit is used for the values of the later periods that must be imputed). Also modified version of the usual imputation methods that takes into account of past observations are applied (see for details Memobust, 2014a). An original method in this setting is the one illustrated in Little and Su (1989). It can be used for quantitative variables, which can be modeled as a combination of period effect and an individual effect and for which stochastic imputation is desired. It is a nearest neighbour technique, that takes into account both cross-sectional and longitudinal information in defining the nearest neighbours. The column effect c_t gives the mean change of the variable *y* over time and is estimated by:

$$c_t = \frac{\bar{y}_t}{\frac{1}{M}\sum_{t=1}^{M}\bar{y}_t}$$

where $\bar{y}_t$ is the mean of the observed y_{it} at period *t*, *M* is the number of periods (or waves) for which the average is considered to be significant. The row effect r_i for unit *i* is represented by:

$$r_i = \frac{1}{m_i}\sum_t \frac{y_{it}}{c_t}$$

where the sum is calculated over the m_i available y_{it} for unit *i* over all the periods it is observed.

The residual is derived considering all the units for which the periods, missing for unit *i*, are observed. All these units are sorted according to the row effect value and, among them, the one presenting a row effect closest to that of unit *i*, i.e. unit *j*, is selected.

The residual of unit *j* is represented by:

$$e_{jt} = \frac{y_{jt}}{r_j c_t} \qquad (1)$$

In the case of a multiplicative model, the final estimation is:

$$\tilde{y}_{it} = r_i c_t e_{jt} \qquad (2)$$

In this case, a zero row effect will result in a zero imputed value.

The three terms represent, respectively, the row, column, and residual effects. In particular the first two terms estimate the predicted mean, and the last term is the component of the imputation from the matched case

Considering (1), the (2) can be also written as:

$$\tilde{y}_{it} = r_i c_t \frac{y_{jt}}{r_j c_t} = \frac{r_i}{r_j} y_{jt} \qquad (3)$$

From (3) it can be derived that, the final estimation is proportional to the y_{jt} value, adjusted by the ratio between the row effects of the units *i* and *j*.

4. Examples

Consider the following small sample of fictitious responses to wages and salaries (Table 1) in three waves.

Table 1. Sample of 10 observations

OBS	Wages & salaries		
	Wave 1	Wave 2	Wave 3
1		400	420
2	675	235	700
3	345	690	800
4	200	480	210
5	200		
6	350	370	
7	400	450	470
8	0	790	790
9	360	450	600
10	135	130	200

Empty cells denote missing items, so for instance observation 1 do not respond to the wages and salaries questions in wave 1, but provides responses in subsequent waves. The first step consists in calculating the column effects based on complete cases only, i.e., units responding in all 3 waves. Column effects incorporates trend information into the imputed amounts: here the wave 1 column effect of 0.70 indicates that the mean current wages and salaries in wave 1 is 30% lower than the overall mean current wages and salaries, and the means in waves 2 and 3 are 6% and 24% higher than the overall mean, respectively. In the second step, the row effects are calculated: for each unit the row effect is the mean (computed on the number of recorded cases) of the reported values divided by the correspondent column effect. In our example, the row effect for unit 1 is ((400/1.06+420/1.24)/2). Then, the sample is ordered by increasing row effects (see Table 2). For each observation to be imputed, the closest donor as the closest complete case is identified:

Table 2 Ordered sample according to the row effects

OBS	Wages & salaries			
	Wave 1	Wave 2	Wave 3	
10	135	130	200	**159**
5	200			**287**
4	200	480	210	**303**
1		400	420	**357**
6	350	370		**425**
7	400	450	470	**458**
8	0	790	790	**460**
9	360	450	600	**475**
2	675	235	700	**585**
3	345	690	800	**596**
	0.70	**1.06**	**1.24**	

Finally, imputations are obtained by multiplying the actual value for the variable of interest of the donor with the row effect of the recipient divided by the row effect of the donor. That is:

- Obs1 - Wave 1: 200*357/303 = 235.64 ~ 236
- Obs5 - Wave 2: 480*287/303 = 454.65 ~ 455
- Obs5 - Wave 3: 210*287/303 = 198.91 ~ 199
- Obs6 - Wave 3: 470*425/459 = 435.18 ~ 435.

5. Input data (characteristics, requirements for applicability)

A data set containing longitudinal microdata with missing values.

6. Output data (characteristics, requirements)

A data set containing imputed longitudinal microdata, which is an updated version of the first input data set.

7. Tools that implement the method

8. Appraisal

Advantages of the Little-Su method are that:

a) imputed values incorporate information about trend at macro and micro level;
b) it does not require separate modelling for different pattern of missing data.

9. References

Little, R. J. A. and Su, H.-L. (1989), Item Non-response in Panel Surveys. In: D. Kasprzyk, G.

Duncan, and M. P. Singh (eds.), Panel Surveys, John Wiley and Sons, 400–425.

Memobust (2014a). Imputation for longitudinal data, in Memobust Handbook on Methodology of Modern Business Statistics. https://ec.europa.eu/eurostat/cros/content/memobust_en

Memobust (2014b). Little and Su method, in Memobust Handbook on Methodology of Modern Business Statistics. https://ec.europa.eu/eurostat/cros/content/memobust_en.

1. Purpose of the method

Imputation is often used to deal with missing data. When imputation is used, the missing data are estimated and filled in into the data set. In many cases the imputation problem is further complicated owing to the existence of constraints in the form of edit restrictions, or edits for short, that have to be satisfied by the data. Examples of such edits are that the profit of an enterprise equals its turnover minus its costs, that the turnover of an enterprise should be at least zero, and that males cannot be pregnant. Records that do not satisfy these edits are inconsistent, and are hence considered incorrect. While imputing missing data, National Statistics Institutes (NSIs) preferably take these edits into account, and thus ensure that the imputed data satisfy all edits.

2. The related scenarios

2.1. Imputation under edit constraints can be used for a single data set composed of microdata.

It can be used in data configurations 1 to 5 (see Deliverable 1). Statistical usage is "imputation".

2.2. Statistical tasks: Data editing and imputation.

2.3. Related methods are imputation methods preserving known totals.

3. Description of the method

Adjustment of imputed values has been studied by Pannekoek and Zhang (2015). Their adjustment approach consists of two steps. In a first step, the imputation step, nearest neighbour hot deck imputation is used to find pre-imputed values. In a second step, the adjustment step, these pre-imputed values are adjusted so the resulting record satisfies all edits. This is done by minimizing the total adjustment subject to the constraint that all edits are satisfied. De Waal and Coutinho (2016) explored this approach in some more detail. An adjustment approach has also been examined and evaluated in Coutinho, De Waal and Remmerswaal (2011). In their case, the pre-imputed values were generated from a multivariate normal distribution that was fitted on the observed data.

Geweke (1991), Tempelman (2007), Holan et al. (2010) and Kim at al. (2014) have developed imputation methods satisfying edits that impute all variables with missing data simultaneously. These methods use a standard statistical distribution as a starting point. Next, that statistical distribution is truncated to the feasible region described by the edits. Geweke (1991) and Tempelman (2007) have proposed to use the truncated multivariate normal distribution. Holan et. al (2010) use a multiple imputation technique based on normally distributed random variables with singular covariance matrices. Kim et al. (2014) have developed a Bayesian multiple imputation approach, using a Dirichlet process mixture of truncated multivariate normal distributions to generate the imputations.

De Waal and Coutinho (2017) also used a simultaneous imputation approach, but they explored the reverse idea. Instead of starting with a statistical distribution that is truncated to the feasible region defined by the edits, they start with the feasible region defined by the edits and then build a statistical distribution with support on that region. Any point drawn from the constructed distribution will then automatically satisfy all edits.

Tempelman (2007) has also proposed to use a model based on the Dirichlet distribution. However, that imputation method can only be used for a very specific situation, namely when all variables are non-negative, sum up to a known constant, and no other edit restrictions are defined for these variables.

The approach proposed by Raghunathan, Solenberger and Van Hoewyk (2002) can be used for both numerical and categorical data. In this approach, for each variable a separate imputation model is constructed, where in principle all other variables may be used as auxiliary data. These imputation models are applied iteratively. Only edit restrictions involving at most one variable with missing data can be handled.

Besides an adjustment approach as mentioned above, Coutinho, De Waal and Remmerswaal (2011) also developed and studied a sequential imputation method. When imputing a certain variable in a certain record, they first eliminate all variables to be imputed subsequently from the edits for that record by means of Fourier-Motzkin elimination. This leads to an interval of allowed values for the current variable to be imputed. To generate potential imputation values Coutinho, De Waal and Remmerswaal (2011) assume that the data can be approximated by a multivariate normal distribution. An imputation value is only accepted if it lies in the allowed interval, otherwise a new value is drawn from the estimated multivariate normal distribution, conditional on the observed values in the record under consideration. De Waal and Coutinho (2012) used the same general approach, but used hot-deck donor imputation instead of the (approximate) multivariate normal model.

For contingency tables, Winkler (2003) proposed to use a parametric model for the data, and estimate its parameters by maximizing its likelihood function while taking into account the structural zeroes that follow from the edit restrictions. Winkler (2008) provided more details on how this can be done for log-linear models.

For numerical data, Beaumont (2005) has developed a calibrated imputation approach that in principle can deal with edit restrictions. The approach by Beaumont (2005) starts with preliminary imputed values. These preliminary imputed values are then calibrated to obey constraints arising from the known totals and the specified edits. This is achieved by solving a mathematical optimization problem, where an objective function measuring the differences between the preliminary imputations and the adjusted values is minimized and the constraints are satisfied.

The problem of imputation of missing categorical data having to satisfy edits and to preserve totals is examined in Favre, Matei and Tillé (2005). In their approach only one categorical variable is to be imputed subject to edit restrictions and known totals. Their approach consists of four steps. In the first step, edit restrictions are used to find structural zeroes for the variable to be imputed, i.e. for each record in the data set the categories that are not allowed are determined. In the second step, for each record the probabilities of imputing the categories that are allowed are estimated. In the third step, these probabilities are calibrated so that, for each category, they sum up to the corresponding total and, for each record, to 1. In the fourth step, a weighted version of Cox' controlled rounding algorithm is used to fix one of the probabilities for the allowed categories per record to 1 and the probabilities of the other allowed categories to 0. The category for which the probability is set to 1 for a certain record is imputed in that record.

Pannekoek, Shlomo and De Waal (2013) developed three imputation methods for numerical satisfying edits and preserving known or previously estimated totals. Two of those imputation methods are sequential ones. Both methods use Fourier-Motzkin elimination to derive allowed intervals for that variable for all records in which its value was missing. Any value in such an interval can lead to an imputed record satisfying all edits.

The first method Pannekoek, Shlomo and De Waal (2013) developed is based on a modified regression approach that incorporates the known total for the current variable to be imputed. Next, an optimization problem is solved in which the imputed values are adjusted as little as possible while satisfying the interval restrictions and preserving the total. In a second method proposed by Pannekoek, Shlomo and De Waal (2013) a random residual is added to the vales obtained from the modified regression approach before the imputed values are adjusted. Finally, a third method uses an imputed data set satisfying edits and preserving totals as starting point. Such a data set may be obtained from the first or second imputation method proposed by Pannekoek, Shlomo and De Waal (2013). Next, the third method uses a Markov Chain Monte Carlo approach to draw pairs of records with at least one common variable with missing values in both records. The values for the common variables in those records are then re-imputed.

De Waal, Coutinho and Shlomo (forthcoming) also propose a sequential imputation for numerical data. For each variable to be imputed, they derive the same allowed intervals as in Pannekoek, Shlomo and De Waal (2013). They derive a simple check for the preservation of the total. They generate possible imputation values by means of hot-deck donor imputation. Only a drawn value that lies in the allowed interval for the record at hand and that passes the check for preservation of the total is actually used for imputation. Otherwise, a new possible imputation value is drawn.

Coutinho, De Waal and Shlomo (2013) developed an imputation method for categorical data that takes edits and known frequencies into account. This is a sequential approach where all variables are imputed in turn. The categories that are imputed are obtained from univariate hot-deck donor imputation. While imputing a missing value, it is ensured that the imputed record can satisfy all edits. This is achieved by calculating so-called implied edits by means of an elimination method proposed by Fellegi and Holt (1976).

If a category obtained by univariate hot-deck donor imputation is acceptable with respect to the edits, it is checked whether also the known totals can be preserved. Coutinho, De Waal and Shlomo (2013) show that this check can be formulated as a so-called Harem problem known from combinatorial mathematics.

For more on imputation methods satisfying edit constraints see De Waal (2017).

4. Examples

Giving examples for all imputation methods satisfying edit constraints would take too much space due to the large number of such methods. We therefore restrict ourselves to giving a simple example for an adjustment approach (see, e.g. Pannekoek and Zhang 2015).

Suppose we have only 3 variables x, y and z, and only one edit $x + y = z$. Suppose $z = 100$ and x and y are both missing. In an adjustment approach we first pre-impute x and y. Say, we impute 40 for x and 50 for y. In the second step of an adjustment approach we would then adjust these pre-imputed values as little as possible (according to some target function) subject to the constraint that the edit would be satisfied. Depending on the target function, this might for instance lead to imputation values of 45 for x and of 55 for y.

5. Input data (characteristics, requirements for applicability)

Imputation under edit constraints uses microdata as input data.

6. Output data (characteristics, requirements)

Imputation under edit constraints produces microdata as output data.

7. Tools that implement the method

Some software packages developed by NSIs, such as GEIS (Kovar and Whitridge, 1990), SPEER (Winkler and Draper, 1997), SLICE (De Waal, 2001) and Banff (Statistics Canada 2009), ensure that edits are satisfied after imputation. These software packages apply relatively simple imputation approaches, such as pro-rating and adjustment of the imputed values so that edits are satisfied.

8. Appraisal

The strength of imputation methods satisfying edit constraints is that they preserve logical relationship in the data, such as "males cannot be pregnant" and "the profit of an enterprise equals its turnover minus its costs". A possible problem of such imputation methods is that statistical relationships might be distorted.

9. References

Coutinho, W. and T. de Waal (2012), *Hot Deck Imputation of Numerical Data under Edit Restrictions*. Discussion paper 201223, Statistics Netherlands.

Coutinho, W., T. de Waal and M. Remmerswaal (2011), Imputation of Numerical Data under Linear Edit Restrictions. *Statistics and Operations Research Transactions 35*, pp. 39-62.

Coutinho, W., T. de Waal and N. Shlomo (2013), Calibrated Hot Deck Imputation Subject to Edit Restrictions. *Journal of Official Statistics 29*, pp. 299-321.

De Waal, T. (2001), SLICE: Generalised Software for Statistical Data Editing. *Proceedings in Computational Statistics* (ed. J.G. Bethlehem and P.G.M. Van der Heijden), Physica-Verlag, New York, pp. 277-282.

De Waal, T. (2017), *Imputation Methods Satisfying Constraints*. UN/ECE Work Session on Statistical Data Editing, The Hague.

De Waal, T. and W. Coutinho (2012), *Hot Deck Imputation under Edit Restrictions*. Discussion paper, Statistics Netherlands.

De Waal, T. and W. Coutinho (2016), *Adjusted Nearest-Neighbour Imputation Satisfying Edit Restrictions*. Report, Statistics Netherlands.

De Waal, T. and W. Coutinho (2017*), Imputation of Numerical Data under Edit Restrictions: The Vertices Approach*. Discussion paper, Statistics Netherlands.

De Waal, T., W. Coutinho and N. Shlomo (forthcoming). Calibrated Hot Deck Imputation for Numerical Data under Edit Restrictions. To be published in *Journal of Survey Statistics and Methodology*.

Favre, A.-C., A. Matei and Y. Tillé (2005), Calibrated Random Imputation for Qualitative Data. *Journal of Statistical Planning and Inference 128*, pp. 411-425.

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

Geweke, J. (1991), *Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities*. Report, University of Minnesota.

Holan, S.H., D. Toth, M.A.R. Ferreira and A.F. Karr (2010), Bayesian Multiscale Multiple Imputation with Implications for Data Confidentiality. *Journal of the American Statistical Association 105*, pp. 564-577.

Kim, H.J., J.P. Reiter, Q. Wang, L.H. Cox and A.F. Karr (2014), Multiple Imputation of Missing or Faulty Values under Linear Constraints. *Journal of Business and Economic Statistics 32*, pp. 375-386.

Kovar, J. and P. Whitridge (1990), Generalized Edit and Imputation System; Overview and Applications. *Revista Brasileira de Estadistica 51*, pp. 85-100.

Liu, T.-P. and E. Rancourt (1999), *Categorical Constraints Guided Imputation for Nonresponse in Survey*. Report, Statistics Canada.

Pannekoek, J., N. Shlomo and T. de Waal (2013), Calibrated Imputation of Numerical Data under Linear Edit Restrictions. *Annals of Applied Statistics 7*, pp. 1983-2006.

Pannekoek, J. and L.C. Zhang (2015), Optimal Adjustments for Inconsistency in Imputed Data. *Survey Methodology 41*, pp. 127-144.

Raghunathan, T.E., J.M. Lepkowski, J. Van Hoewyk and P. Solenberger (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology 27*, pp. 85-95.

Statistics Canada (2009), *Functional Description of Banff - the Generalized Edit and Imputation System*. Technical Report, Statistics Canada.

Tempelman, C. (2007), *Imputation of Restricted Data*. Doctorate thesis, University of Groningen.

Winkler, W.E. (2003), *A Contingency-Table Model for Imputing Data Satisfying Analytic Constraints*. Research Report Series 2003-07, Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Winkler, W.E. (2008), *General Methods and Algorithms for Modeling and Imputing Discrete Data under a Variety of Constraints*. Research Report Series 2008-08, Statistical Research Division, U.S. Census Bureau, Washington, D.C.

Winkler, W.E. and L.A. Draper (1997), The SPEER Edit System. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva.

1. Purpose of the method

Different estimates for the same phenomenon could lead to confusion among users of these figures. Many other NSIs, such as Statistics Netherlands, have therefore adopted a one-figure policy. According to this one-figure policy, estimates for the same phenomenon in different tables should be equal to each other, even if these estimates are based on different underlying data sources. We call this univalency and say that estimates should be univalent.

When using a mix of administrative data sources and surveys to base estimates upon, the one-figure policy becomes problematic as for different (combinations of) variables data on different units, e.g. different persons, may be available. This means that different estimates concerning the same variable may yield different results, if one does not take special precautions.

2. The related scenarios

2.1. Repeated weighting can be used for data sets composed of microdata. It can be used in data configuration 5 (see Deliverable 1). Statistical usage is "Direct tabulation".

2.2. Statistical tasks: "Integrate data".

2.3. Alternative methods are repeated imputation, mass imputation and macro-integration.

3. Description of the method

Univalency between tables can be enforced by means of repeated weighting (see e.g. Houbiers 2003). When repeated weighting is used, a separate set of weights is assigned to sample units for each table of population totals to be estimated. In this approach population tables are estimated sequentially and each table is estimated using as many sample units as possible in order to keep the sample variance as low as possible. The combined data from the data sources are divided into rectangular data blocks. Such a block consists of a maximal set of variables for which data on the same units has been collected. The data blocks are chosen such that each table to be estimated is covered by at least one data block.

Data from a data source covering the entire population can simply be counted. Data only available from surveys are weighted by means of regression weighting. In that case starting weights need to be assigned to all units in the block to be weighted. For a survey one usually starts with the inverse inclusion probabilities of the sample units, corrected for response selectivity. For a data block containing the overlap of two surveys, one usually begins with the product of the standard survey weights from each of the surveys as starting weight for each observed unit.

When estimating a new table, all cell values and margins of this table that are known or have already been estimated for previous tables are kept fixed to these known or previously estimated values. This is achieved by using regression weighting, where the starting weights are adjusted by calibrating to known or previously estimated values. This ensures univalency of the cell values and margins of the new table and previous estimates.

4. Examples

Suppose we have a population of 10,000 persons, and two surveys: one with a sample size of 2,000 with observations on the kind of job of persons (fulltime versus part-time) and one with a sample size of 1,000 with observations of the educational level of persons (low, middle, high). We assume that all survey weights in Survey 1 are equal to 5, and in Survey 2 to 10. The observed numbers in these surveys are given in Tables 1 and 2 below.

Table 1. Observations Survey 1

Kind of job	Number
Fulltime	1.200
Part-time	800

Table 2. Observations Survey 2

Educational level	Number
Low	175
Middle	500
High	325

To obtain estimates for the kind of jobs in the population we weight Survey 1. The results are given in Table 3.

Table 3. Population estimates for "kind of job"

Kind of job	Number
Fulltime	6.000
Part-time	4.000

To obtain estimates for the education level we weight Survey 2. The results are given in Table 4.

Table 4. Population estimates for "educational level"

Educational level	Number
Low	1.750
Middle	5.000
High	3.250

We assume that the two surveys have an overlap, allowing us to estimate the relation between the kind of job and the educational level of people. Suppose the observations in the overlap are given by Table 5. We assume that for all units in the overlap the starting weight is equal to 50.

Table 5. Observations in the overlap of the two surveys

	Fulltime	Part-time	Total
Low	30	20	50
Middle	50	40	90
High	30	30	60
Total	110	90	200

When standard weighting using the starting weights were used, we would obtain the results of Table 6. Note that the marginals of Table 6 differ from the numbers in Table 3 and 4.

Table 6. Population estimates for "kind of job x educational level" after standard weighting

	Fulltime	Part-time	Total
Low	1.500	1.000	2.500
Middle	2.500	2.000	4.500
High	1.500	1.500	3.000
Total	5.500	4.500	10.000

When repeated weighting is used, the estimates in Tables 3 and 4 are kept fixed. This is achieved by adjusting the starting weights in the overlap of Surveys 1 and 2. The resulting estimates after repeated weighting may, for instance, be given by Table 7. Note that Table 7 indeed reproduces the earlier results in Tables 3 and 4.

Table 7. Population estimates for "kind of job x educational level" after repeated weighting

	Fulltime	Part-time	Total
Low	1.200	550	1.750
Middle	2.800	2.200	5.000
High	2.000	1.250	3.250
Total	6.000	4.000	10.000

At Statistics Netherlands, repeated weighting has been applied to produce univalent estimates for the Dutch Population and Housing Census.

5. Input data (characteristics, requirements for applicability)

Repeated weighting uses microdata as input data.

6. Output data (characteristics, requirements)

Repeated weighting produces univalent macrodata, in particular univalent tables.

7. Tools that implement the method

At Statistics Netherlands a tool called VRD has been developed for repeated weighting.

8. Appraisal

Repeated weighting does achieve univalency.

Repeated weighting has several technical drawbacks, for instance when there are no units with certain characteristics in (the overlap of) surveys, but we know that there are units with those characteristic in the population (see De Waal 2016 for more on these technical drawbacks).

9. References

De Waal, T. (2016), Obtaining Numerically Consistent Estimates from a Mix of Administrative Data and Surveys. Statistical Journal of the IAOS 32, pp. 231–243

Houbiers, M. (2004), Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. Journal of Official Statistics 20, 55-75.

1. Purpose of the method

Different estimates for the same phenomenon could lead to confusion among users of these figures. Many other NSIs, such as Statistics Netherlands, have therefore adopted a one-figure policy. According to this one-figure policy, estimates for the same phenomenon in different tables should be equal to each other, even if these estimates are based on different underlying data sources. We call this univalency and say that estimates should be univalent.

When using a mix of administrative data sources and surveys to base estimates upon, the one-figure policy becomes problematic as for different (combinations of) variables data on different units, e.g. different persons, may be available. This means that different estimates concerning the same variable may yield different results, if one does not take special precautions.

2. The related scenarios

2.1. Mass imputation can be used for a single data set composed of microdata. It can be used in data configurations 3 to 5 (see Deliverable 1). Statistical usage is "Direct tabulation".

2.2. Statistical tasks: "Direct tabulation" and "Integrate data".

2.3. With respect to the statistical task "Integrate data", alternative methods are repeated weighting, repeated imputation and macro-integration.

3. Description of the method

In the mass imputation approach, one imputes all variables for which no value was observed for all population units (see Whitridge, Bureau and Kovar 1990, Whitridge and Kovar 1990, Shlomo, De Waal and Pannekoek 2009). This leads to a rectangular data set with values for all variables and all population units. After imputation, estimates of totals can be obtained by simply counting or adding the values of the corresponding variables.

The approach relies on the ability to capture all relevant variables and relevant relations between them in the imputation model, and to estimate the model parameters sufficiently accurately. Given that all relevant variables and relevant relations among them can be captured accurately by the imputation model, the approach is very straightforward.

4. Examples

At Statistics Netherlands, mass imputation is examined for the Educational Attainment. By mass imputing educational attainment, an estimate for the highest educational level of all people in the Dutch population will become available. Those estimates can subsequently be used to compile part of the Dutch Population and Housing Census. If results of mass imputation are satisfactory, this mass imputation approach is planned to be used for the 2021 Census.

5. Input data (characteristics, requirements for applicability)

Mass imputation uses microdata as input data.

6. Output data (characteristics, requirements)

The output of mass imputation consists of microdata.

7. Tools that implement the method

Tailor-made programs and scripts are used.

8. Appraisal

A fundamental problem with a mass-imputed data set is that it may be used for purposes for which it was never intended. Moreover, is most application of mass imputation it is hard to tell from the imputed data set itself that one is using it for unintended purposes.

An example is combining the amount of money spent per month on dog food, (which may be known from a Budget Survey), with whether or not people have a dog as pet (which may be known from a Survey on Living Conditions). Including these two variables – the amount of money spent per month on dog food and whether one has a dog as pet or not – in an imputation model is, except in very exceptional cases, not deemed important enough. Including information on their correlation in an imputation model is even more unlikely.

If these variables and information on their correlation are not included in the imputation model and one is not aware of this, one may decide to analyse and publish the relation between these variables. In this case one may come to the unjustified conclusion that many people who do not have a dog as pet spent money on dog food, and that conversely many people who do have a dog a pet do not buy dog food.

The situation is different for the Educational Attainment data, as here mass imputation is used for a clear purpose and all relevant variables can be included in the imputation model.

In order to alleviate the problem of not knowing from the imputed data set itself that one is using it for unintended purposes, one could consider releasing additional information about which (combinations of) variables are included in the imputation model and are hence controlled for. For instance, for a data set with three variables x_1, x_2 and x_3, one could use the notation [x_1 | x_2 x_3] (Zhang 2017) to express that x_1 and the combination x_2 and x_3 are included in the imputation model. As the user now knows that, for example, the combination $x1$ and x_2 is not included in the imputation model, he/she will also know that estimates involving the combination of x_1 and x_2 will not be valid.

9. References

Shlomo, N., T. de Waal and J. Pannekoek (2009), *Mass Imputation for Building a Numerical Statistical Database*. UN/ECE Work Session on Statistical Data Editing, Neuchâtel, Switzerland.

Whitridge, P., M. Bureau and J. Kovar (1990), Mass Imputation at Statistics Canada. In: *Proceedings of the Annual Research Conference*, U.S. Census Bureau, Washington D.C., pp. 666-675.

Whitridge, P. and J. Kovar (1990), Use of Mass Imputation to Estimate for Subsample Variables. In: *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, pp. 132-137.

Zhang. L.-C. (2017), *Personal communication*.