Discussion Papers

586

Carsten Kuchler
Martin Spiess

The Data Quality Concept of Accuracy in the Context of Public Use Data Sets

Berlin, May 2006

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

The Data Quality Concept of Accuracy in the Context of Public Use Data Sets

Carsten Kuchler

Santander Consumer Bank, Riskmanagement Retail, Methodology & Projects

Berliner Platz 12 / Fliethstraße 67, D-41061 Mönchengladbach

Martin Spiess

SOEP, DIW Berlin, Königin-Luisestr. 5,

D-14195 Berlin, Germany, mspiess@diw.de

and

International Institute of Management

University of Flensburg, Auf dem Campus 1, D-24943 Flensburg

The Data Quality Concept of Accuracy in the Context of Public Use Data Sets

Abstract

Like other data quality dimensions, the concept of accuracy is often adopted to characterise a particular data set. However, its common specification basically refers to statistical properties of estimators, which can hardly be proved by means of a single survey at hand. This ambiguity can be resolved by assigning 'accuracy' to survey processes that are known to affect these properties. In this contribution, we consider the sub-process of imputation as one important step in setting up a data set and argue that the so called 'hit-rate' criterion, that is intended to measure the accuracy of a data set by some distance function of 'true' but unobserved and imputed values, is neither required nor desirable. In contrast, the so-called 'inference' criterion allows for valid inferences based on a suitably completed data set under rather general conditions. The underlying theoretical concepts are illustrated by means of a simulation study. It is emphasised that the same principal arguments apply to other survey processes that introduce uncertainty into an edited data set.

1 Introduction

(1) At the latest since the mid-1990's, Official Statistics is more and more gripped by what is occasionally called the quality revolution. In the triangle of severe budget cuts, increasing user demands and the competition with a growing number of scientific and other non-official data providers, Official Statistics is pushed to optimise the range and particularly the quality of services offered. At any rate, out of this needful process a variety of survey quality definitions emerged. Current approaches decompose this rather general term into handier subordinated concepts, like relevance, timeliness, coherence, accessability, etc. (e.g. in Brackstone, 1999; Eurostat, 2003; Biemer und Lyberg, 2003). While the emphasis of these sub-concepts may vary from approach to approach, they all have in common to consider accuracy as a major quality objective.

(2) The definitions of accuracy given by Brackstone (1999) and Eurostat (2003) basically refer to the deviation of an arbitrary statistic from its respective population value. However, since the population parameter is generally unknown and the statistic is computed from a randomly selected sample, this approach can hardly be given a clear statistical sense. Instead, adopting, for example, the so called frequentist point of view, allows for reasoning about this deviation in terms of the statistical properties of the underlying estimator over hypothetically infinite cycles through the survey and the subsequent estimation processes. This is regularly done in terms of the well-known Mean Squared Error (MSE) or its

systematic squared bias and random variance component

$$\text{MSE}\left(\hat{\theta}\right) := \text{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \left(\text{E}\,\hat{\theta} - \theta\right)^2 + \text{Var}\,\hat{\theta},$$

where $\hat{\theta}$ denotes the estimator of a population parameter θ. However, the concept of accuracy, like any of the data quality dimensions, is intended to be applied in assessing particular surveys. Thus, the question arises as to how a single survey at hand can be declared 'accurate' by referring to statistical properties of estimators, which, then again, cannot be proved by means of this single survey.

(3) An answer arises from shifting the focus from the particular data set towards the survey processes it emanates from. Biemer und Lyberg (2003) call this the *process view* of survey quality, where "one has to assure quality by using dependable processes, processes that lead to good product characteristics. The basic thought is that product quality is achieved through process quality" (Biemer und Lyberg, 2003, p. 14). From this point of view, survey quality may rather be considered as a general objective to be accomplished throughout the entire survey process by applying in a sense appropriate methods in each process step of data production. In taking up this as a starting point, it is crucial to clarify, what is meant by 'good product characteristics' in terms of the overall survey process, its derived sub-processes and particularly with respect to the different sub-concepts of survey quality.

(4) For the accuracy objective, 'product characteristics' of a survey can, in a very general sense, be considered to arise from its intended context of use, that on his

3

part is determined by the estimators applied – and this is where the circle closes: Given an estimator $\hat{\theta}$ that is to be applied to a particular survey, then this survey can be said to be accurate (with respect to $\hat{\theta}$) if the preceding data production processes keep the assumptions associated with $\hat{\theta}$. 'Accuracy' then primarily refers to the realisation of specific methodological requirements of sub-processes like sampling, editing, imputation, etc., that are known to affect the statistical properties of estimators computed from the processed data. Thus the rather abstract accuracy objective resolves into the methodological problem of how to appropriately perform the related sub-process in order to assure the preconditions for valid inferences – and this can be done irrespective of the particular data set to be assessed.

(5) In practical terms this means to tell the accuracy story right from the end: In the first place one has to have a notion of what kind of estimators will be fitted to the data and particularly of the assumptions they depend on. In addition one might take into account the conditions for assessing the statistical properties of the estimators in terms of their estimated variance, the computation of confidence intervals, etc. Only then, one can seek survey methods that meet all these assumptions and thus ensure the statistical properties of the estimators to be applied. That is, whether or not a data set can be considered accurate, solely depends on the analyst's purposes which are reflected in the selection of estimators applied: a particular data set may be accurate for a specific analysis and

may not for another. For an example from the design-based context, consider the Horvitz-Thompson estimator for weighted totals (Horvitz und Thompson, 1952), that essentially depends on the appropriate realisation of the random sampling process from a finite population, and in case of violations is not guaranteed to result in valid inferences.

(6) Providing valid inferences from in this sense accurate data sets should not be an issue in the classical field of application in Official Statistics, where the processes of data production and analysis are in one hand, and the analytical purposes are restricted to rather straightforward estimators like totals or means. An entirely new situation emerges when Official Statistics provides micro data, either for internal use like for fitting complex National Account models or by releasing public and scientific use files. In both cases, the data producer is not a priori aware of which estimators will be applied to the data, and thus has to perform the survey processes, such that the resulting data set allows for a preferably wide range of analyses.

(7) In this contribution we address the problem of how the term 'accuracy' can be assigned a clear meaning within the sub-process of imputation, i.e. the substitution of unobserved or erroneous values during data editing. Special attention is paid to the requirements of imputing public use data sets. For that purpose, the specific methodological requirements of imputation methods need to be examined, that potentially lead to accurate data in the above sense. This is subject

to the second section that deals with three basic evaluation approaches for impu-
tation methods, shedding some light on the methodological requirements under
discussion. It should be noted however, that imputing for non-observed values
is just one component in the process of setting up a public use data base. The
same principal arguments apply to any sub-process embedded into the survey
process, as long as they affect inferences by introducing uncertainty to the data.
In section three we discuss different imputation methods and illustrate possible
problems associated with them via a simulation study. Conclusions and some
discussion can be found in section four.

2   Assessment of imputation methods

(8) Consider a data set $\mathbf{Y}$ with observed elements in $\mathbf{Y}_{obs}$ and values declared
to be missing in $\mathbf{Y}_{mis}$. Then any imputation method is intended to compute
in a sense reasonable substitutions $y_{ij} \leftarrow \hat{y}_{ij} \in \mathbf{Y}_{imp}$ for any missing element
$y_{ij} \in \mathbf{Y}_{mis}$. Subsequent inferences are based on the corresponding completed data
set $\widehat{\mathbf{Y}}$ consisting of $\mathbf{Y}_{obs}$ and the imputed values in $\mathbf{Y}_{imp}$. In the remainder we
occasionally refer to the (n, p) data matrix $\mathbf{Y}$ and the corresponding completed
matrix $\widehat{\mathbf{Y}}$ as illustrated in figure 1. In order to keep notation simple, $\mathbf{Y}$ is assumed
to have an univariate missing pattern, i.e. the variables Y_1 to Y_{p-1} are completely
observed and only the values $y_{i'p}$ with $i' = n_{cc} + 1, ..., n$ of the p-th variable
are declared to be missing. Within the observed part of $\mathbf{Y}$ one can distinguish
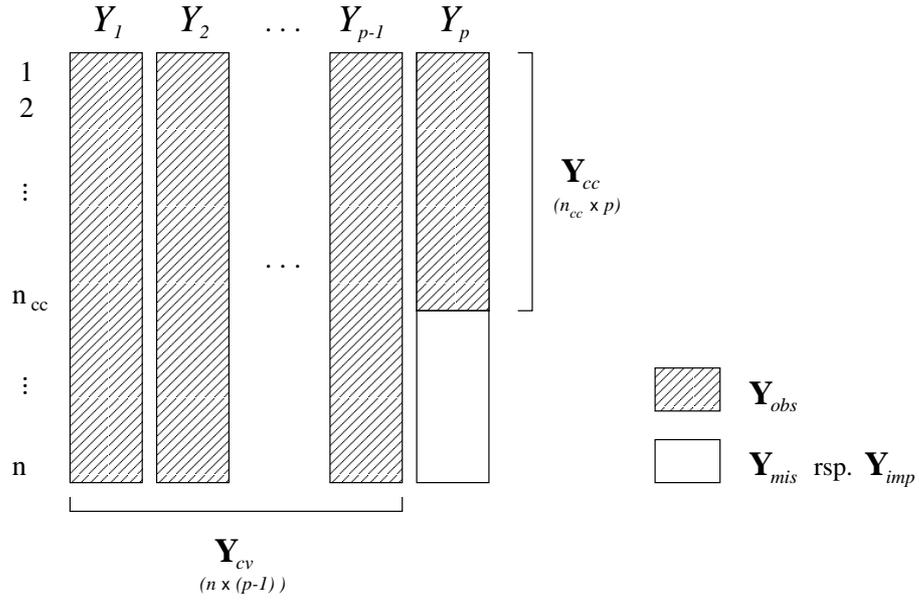
6

Figure 1: Incomplete data matrix $\mathbf{Y}$ with a univariate missing pattern, consisting of observed elements in $\mathbf{Y}_{obs}$ and values declared to be missing in $\mathbf{Y}_{mis}$. Substituting $\mathbf{Y}_{mis}$ by imputations in $\mathbf{Y}_{imp}$ yields the completed matrix $\widehat{\mathbf{Y}}$.

between the (n_{cc}, p) matrix of the complete observations $\mathbf{Y}_{cc}$ and the $(n, p-1)$ matrix of the complete variables $\mathbf{Y}_{cv}$.

(9) Assume there are two alternative imputations at hand for substituting a missing value, both of them within the co-domain of the incomplete variable and satisfying the predefined editing rules. To decide for one value and rejecting the other remains arbitrary, unless additional information about the (statistical) properties of the underlying imputation methods is taken into account. These properties can be determined analytically or at least in appropriate simulation studies. Reviewing the literature reveals a number of evaluation criteria for impu-

tation methods, which can in general be subsumed under three basic approaches:

(a) *The inference criterion*: Provided an arbitrary estimator applied to the complete data set $\mathbf{Y}$ results in a valid inference with respect to the corresponding estimand. Then an imputation method should complete a data set such that applying the same estimator to the completed data set $\widehat{\mathbf{Y}}$ results in a valid inference as well (cf. Rubin, 1996).

(b) *The hit-rate criterion*: Each imputed value $\hat{y}_{i'p} \in \mathbf{Y}_{imp}$ should lie as close as possible to the corresponding unobserved value $y_{i'p} \in \mathbf{Y}_{mis}$, such that an arbitrary distance measure $d\left(\cdot\right)$ is minimised over $\mathbf{Y}_{imp}$,

$$\sum_{i'=n_{cc}+1}^{n} d\left(y_{i'p}, \hat{y}_{i'p}\right) \overset{!}{=} \min .$$

(c) *The plausibility criterion*: Each imputed value $\hat{y}_{i'p} \in \mathbf{Y}_{imp}$ needs to be covered by the co-domain $\mathcal{Y}_p$ of the random variable Y_p and causes no inconsistencies with observed or other imputed values in $\widehat{\mathbf{Y}}$. When $\hat{y}_{i'p}$ is generated during data editing, the i'-th observation should finally pass all specified data checks.

(10) The inference criterion directly assigns the general accuracy specification to the sub-process of imputation. Hence it can be assumed to be appropriate, but still has to prove its theoretical foundation and practical realisation, which is subject to the following section. However, the approaches (b) and (c) need to be examined with respect to the central question of how they can be related to the

8

general accuracy specification, i.e. whether or not a data set completed with an imputation method that fulfils at least one of these criteria ensures the statistical properties of potentially applied estimators.

(11) The hit-rate criterion refers to the intuitive idea, that an optimal imputation directly substitutes a missing value by its corresponding unobserved counterpart. Provided this ideal case, it even covers the inference criterion, since imputing true values obviously results in consistent estimators and hence in an accurate data set in the proposed statistical sense. In addition, the hit-rate criterion is straightforward to implement and provides results which are easy to interpret. These features make it a common choice for evaluation studies on imputation methods applied to data sets with generated missing values. Chambers (2001) compiles a number of measures that among others implement the hit-rate criterion for simulation studies of this kind. Even though it is "hardest to achieve", for him the hit-rate criterion is particularly relevant when the edited data set is publicly released by the data supplier or is internally used to determine prediction models (Chambers, 2001, 11).

(12) The hit-rate criterion is based on the implicit assumption, that the unobserved values of Y_p follow a specific function that can to some extent be approximated by an appropriate imputation method. Thus, unless this function is deterministic and known, imputations are in fact (uncertain) point estimates that are considered to be optimal if a predefined loss function $d\left(\cdot\right)$ is minimal.

9

However, evaluating the value of $d\left(\cdot\right)$ for a given data set and a given set of imputations ignores that $\mathbf{Y}_{imp}$ is a random variable, with its values following a stochastic instead of a deterministic function. That is, the deviations measured by the distance function $d\left(\cdot\right)$ are subject to the variability of the random variable $\mathbf{Y}_{imp}$ and $\mathbf{Y}_{obs}$ (in a model-based view), or $\mathbf{Y}_{imp}$ and the selection indicators (in a design-based view). Hence generating imputed values that are as close as possible to their unobserved counterparts is restricted by conceptual bounds that are not reflected and are not even reflectable by the hit-rate approach.

(13) Reinforcing these conceptual objections, a simple example by Rubin (1996) shows, that the hit-rate criterion even fails in producing valid estimators. Given a biased coin with the probability of realising 'head' being 0.6. The prediction model that both sides of the coin are 'heads' yields a hit rate of $0.6\cdot1.0+0.4\cdot0.0 = 0.6$. In contrast, applying the true model results in the lower hit rate $0.6 \cdot 0.6 + 0.4 \cdot 0.4 = 0.52$. With respect to the hit-rate criterion, the first model is to be selected as imputation model, even though it results in seriously biased estimators by systematically overestimating the fraction of 'heads' from the completed data set. For conceptual reasons the deterministic hit-rate criterion obviously fails the crucial demand of ensuring the statistical properties of estimators within the imputation process, and thus is inappropriate to found the general accuracy objective.

(14) The plausibility criterion might be applied individually or in combination

with one of the other criteria. The first way is regularly chosen by data suppliers and particularly National Statistical Agencies that are editing their data manually by fitting corrections such that none of some predefined editing rules are violated. Since accounting for several editing rules (and their combination) is a profoundly complex problem, it is regularly resolved in an inscrutable process of trial-and-error accompanied by expert-knowledge of clerical staff. Of course, the data quality and particularly the accuracy of the resulting data set cannot be assessed in statistical terms, since the underlying process remains arbitrarily. In the second case, the plausibility criterion is applied in a two-step process. Firstly, missing values are imputed in accordance with, for example, the inference criterion, and secondly, $\widehat{\mathbf{Y}}$ is subject to a subsequent pass through the specified checks in order to avoid implausible (combinations of) values. In doing so, marginal distributions tend to be distorted and bias is likely to be introduced in corresponding estimators (for example raising low or zero incomes to a minimum living wage increases the averaged incomes computed from the completed data set). Thus the absurd situation emerges that assuring plausibility in terms of the associated criterion may cause a loss of accuracy in terms of the inference criterion. However, in both cases the usually adopted deterministic view of 'correcting' values due to the plausibility criterion ignores the uncertainty inherent in (a) the decision as to which values are treated as 'incorrect' if combinations of values fail a data check, and, (b) the new values imputed.

3 Computing valid inferences from $\widehat{\mathbf{Y}}$

(15) Evaluation of an imputation method by means of the inference criterion has to focus on distributional characteristics of the resulting estimator of interest. For our discussion we will concentrate on the first two (central) moments of this distribution. Thus, considering an estimand θ, an appropriate imputation method particularly needs to provide (asymptotically) unbiased estimators $\hat{\theta}$ as well as valid estimators of the variance of $\hat{\theta}$ by taking into account all sources of variability. In the following paragraphs both topics will briefly be addressed in terms of the multiple imputation theory by Rubin (1987).

(16) Unless the so called *Missing at Random* (MAR) assumption fails, the goal of arriving at (asymptotically) unbiased estimators for the mean of a variable with missing values can even be achieved with rather unsophisticated conditional mean imputation procedures (Little und Rubin, 2002). The MAR assumption states, that the probability of observing a value in $\mathbf{Y}$ is independent from its unobserved counterpart, i.e. generally

$$P\left(y_{ij} \text{ is observed } | \mathbf{Y}\right) = P\left(y_{ij} \text{ is observed } | \mathbf{Y}_{obs}\right)$$

for all possible values in $\mathbf{Y}_{mis}$ (Little und Rubin, 2002, 12). On the other hand, it is straightforward to show, that applying statistical standard procedures to data sets with imputed unconditional or conditional means causes a substantial underestimation of the variance and co-variances of the incompletely observed
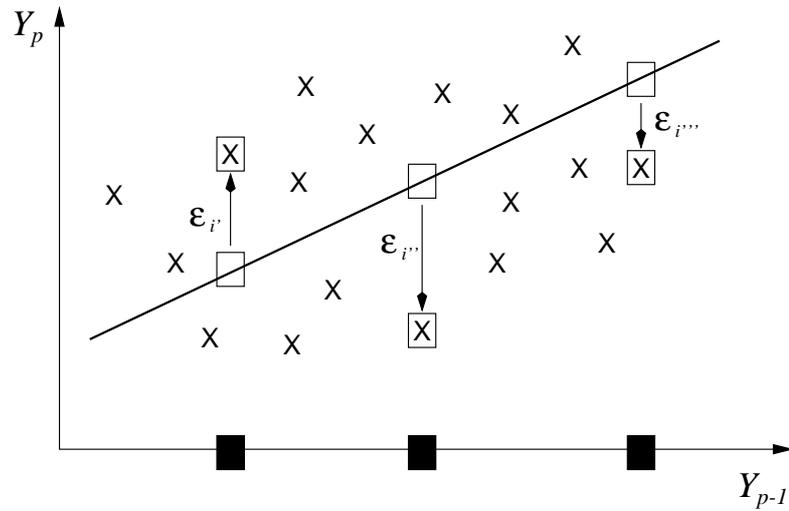
Figure 2: Example of an *Stochastic Regression Imputation* in a bivariate linear setting with Y_p regressing on Y_{p-1}. Complete observations are denoted by "X" and the incomplete observations with missing values in Y_p by "■". The associated imputed values are raised from the conditional means ("□") on the regression line by adding white noise ("ε").

variable, that moreover increases with the fraction of missing values (Little und Rubin, 2002, 61). This is due to the fact that imputed means just inflate the sample by adding values right at the centre of the (conditional) distributions and thus have no impact on the calculation of statistics that are based on the deviation from the mean.

(17) This shrinkage of the scatter-plot can be avoided by adding white noise to the imputed values. For example, assuming that an incomplete variable Y_p is regressed on the completely observed variables $Y_1, ..., Y_{p-1}$, the regression impu-

tation model is

$$y_{i'p} \leftarrow \hat{y}_{i'p} = \hat{\beta}_0\left(\mathbf{Y}_{cc}\right) + \sum_{j=1}^{p-1} \hat{\beta}_j\left(\mathbf{Y}_{cc}\right) y_{i'j} + \varepsilon_{i'} \, ,$$

with the regression parameters computed from $\mathbf{Y}_{cc}$ (see figure 2 for this linear variant of *Stochastic Regression Imputation*). Within this approach, a natural choice for the distribution of ε is the Gaussian noise model

$$\varepsilon \sim \mathrm{N}\left(0 \, ; \, \frac{1}{n_{cc} - 2} \sum_{i=1}^{n_{cc}} (\hat{y}_{ip} - y_{ip})^2\right) \, ,$$

with the variance term representing the estimated variance of the residuals computed in $\mathbf{Y}_{cc}$ (other noise models are examined by Schenker und Taylor, 1996). Provided the selected noise model is appropriate and the MAR assumption holds, stochastic regression imputation results in unbiased estimators of the mean, the variance, co-variances and even the parameters for regressing Y_p on $Y_1, ..., Y_{p-1}$, and vice versa (Little und Rubin, 2002, 65).

(18) However, the second issue of evaluating the variance of estimators due to the inference criterion is still untreated. In the previous paragraphs and particularly when rejecting the hit-rate criterion, imputed values were considered as estimates of unobserved values rather than straightforward insertions. Thus a valid estimation of the variance of an (asymptotically unbiased) estimator needs to reflect two sources of variation: (a) the general variance fraction due to the observed values, and (b) the variance fraction due to the prediction of the missing values during imputation.

14

(19) This point is easily demonstrated by means of a simple model-based simulation study that compares different imputation strategies with respect to the statistical properties of estimators applied to completed data sets. Point of reference is the (unbiased) estimation of the mean of Y_p by the sample mean in the complete data case (COM), i.e. with all values observed. For this purpose we generated 100 values from a normally distributed variable with mean $\theta = 2$ and variance $\sigma^2 = 4$. From the complete data set 30% of the values were deleted such that the missing values were *missing completely at random (MCAR)*, i.e.

$$P\left(y_p \text{ is observed } | \mathbf{Y}\right) = P\left(y_p \text{ is observed}\right).$$

for all possible values in $\mathbf{Y}$ (Little und Rubin, 2002, 12). After having imputed for the deleted values, the completed data set was analysed, i.e. we estimated the mean by the sample mean of the completed data set, computed an estimate of the variance of the estimator and finally tested the hypothesis $H_0 : \theta = 2$ with $\alpha = 0.05$. The simulation cycle of generating a data set, deleting values, imputing for $\mathbf{Y}_{mis}$ and analysing the completed data set was repeated 20000 times. Finally, over the 20000 cycles, the mean of the estimates (m), the square root of the mean of estimated variances (SD_E), the standard deviation of the estimates (SD) and the proportion of rejections of the null hypothesis was calculated for the complete date case as well as for the imputation strategies applied (cf. table 1).

(20) There were four imputation strategies applied in the simulation study. We will first focus on two so called single imputation strategies: A missing value

15

is replaced by the sample mean of the observed values in $\mathbf{Y}_{cc}$ (MEAN) or by random draws with replacement from $\mathbf{Y}_{cc}$ (simple random draw imputation, SR). Inferences are based on standard methods for completely observed data sets, i.e. imputed values are treated as if they were observed, thus ignoring the uncertainty due to their prediction. The results in Table 1 illustrate the points discussed above. Since the missing values are MCAR, the means of the estimates $\hat{\theta}$ are close to the true values, regardless of the imputation strategy. The crucial point is that both single imputation strategies (MEAN and SR) grossly underestimate the variances of the estimators, leading to anti-conservative inferences, i.e. the null hypothesis is rejected too often. In the simple case considered here, the estimated variance of the estimator of the mean could of course be easily corrected to account for this additional variation, but in most real-life situations a correction term for the variance estimator is not available. However, these results show that when applying standard methods, the MEAN and SR strategy fail in providing valid inferences even in a straightforward estimation situation under the relaxed MCAR condition.

(21) To account for these problems and at the same time keep the analyses of incompletely observed data sets simple, Rubin (1987) developed the so called multiple imputation method, where each missing value is replaced by $D > 1$, in some sense 'proper' values or predictions (Rubin, 1987). One major characteristic of multiple imputations being proper is that all the uncertainty in the predictions

Table 1: Mean (m), estimated standard errors (SD$_E$), standard deviation (SD), proportion of rejection of the null H$_0 : \theta = 2$ ($\alpha = 0.05$) over 20 000 simulations.

	COM	MEAN	SR	SR-MI	ABB
m	2.00	2.00	2.00	2.00	2.00
SD$_E$	0.200	0.167	0.199	0.228	0.239
SD	0.200	0.237	0.260	0.238	0.239
rej	0.05	0.17	0.14	0.06	0.05

is reflected in their variation. Going back to the simulation study, each observed value can be decomposed in the 'true' mean and a random error. However, neither the 'true' mean nor the error is observable. Thus, to generate imputations, both need to be estimated. For drawing valid inferences from a completed data set, one has to account for the randomness in both estimators. In terms of the multiple imputation approach this is easily done by imputing D combined predictions of the mean and the error, and thus generating D completed data sets, that can in turn be analysed with standard methods for completely observed data sets. Finally, the resulting D estimators are combined according to simple rules given by Rubin (1987) or Little und Rubin (2002).

(22) In the simulation study, a first multiple imputation strategy was to repeat simple random draw imputation. This is equivalent to fixing $\hat{\theta}$ at the sample mean and repeatedly draw from the residual distribution with variance fixed at

its sample value (SR-MI). Note that this strategy ignores the uncertainty inherent in the estimated mean and variance. According to the second strategy (ABB), we first drew a naive bootstrap sample from the observed part of the sample and then, for each missing value, randomly drew a value (with replacement) from this bootstrap sample. This is equivalent to calculating the sample mean for the bootstrap sample and then draw values for the error terms. Thus, repeated application of this procedure leads to multiple imputations that in addition to the uncertainty in the predictions given the mean of the observed part of the sample, also reflect the uncertainty inherent in the sample mean itself. Rubin (1987) introduced this imputation strategy as 'Approximative Bayesian Bootstrap'. For both multiple imputation strategies, we generated $D = 20$ imputations for each missing value. Table 1 shows that in contrast to the estimator based on the ABB strategy, the estimator based on the SR-MI approach slightly underestimates the variance of the estimators.

(23) Thus, following usual practice to analyse singly imputed data sets as if the imputations were 'recovered' true values and treating them as being observed can be seriously misleading. Unfortunately, this strategy is reinforced by data suppliers that release singly imputed data sets without giving information on how to draw proper inferences or even flagging the imputations.

18

4   Conclusions

(24) In this paper we examined the concept of 'accuracy' applied to data sets as provided, e.g., by official statistical agencies. We argue that since the ultimate goal usually is to draw valid inferences about a population and not about the data set itself, the term 'accuracy' should not refer to a single data set, but rather should be some kind of 'process accuracy'. According to this view, a survey can be said to be accurate (with respect to $\hat{\theta}$) if the preceding data production processes keep the assumptions associated with $\hat{\theta}$. This implies that a data set can be accurate with respect to one analysis but inaccurate with respect to another. These points were discussed for the imputation process as a sub-process of data production, where missing values are replaced by somehow 'plausible' values. We argued that, for example, the 'hit-rate' criterion is misleading in assessing the accuracy of a single data set: If the unobserved values can be 'recovered' by some known deterministic function, then such a measure is unnecessary. Otherwise, the imputations are uncertain predictions and this uncertainty has to be taken into account. Ignoring this uncertainty leads to seriously biased inferences as illustrated in section 3. This point is particularly relevant if a potential user is not aware of the imputation process and thus is unable to account for the uncertainty in the subsequent inferences.

(25) It should be noted, that although we only consider the imputation step, the same principal arguments apply to the entire survey process. For example,

in a design-based context, a sample should be a probability sample (cf. Särndal et al., 1992), otherwise it is hard to justify inferences based on, e.g., the Horvitz-Thompson estimator. This particularly holds for the so called error-localisation problem, where for a set of values violating some editing rules, a decision has to be made which combination of values is erroneous and which values have to be set to missing for subsequent imputation. To illustrate this problem, consider a data set consisting of the two variables age and marital status. Observing a 14 year old widow may violate a predefined editing rule that says if age is lower than sixteen the martial status must show "single". The question arises, which of the observed values is erroneous and needs to be corrected: age, martial status or both. In principle there are $2^i - 1$ possible combinations, where i is the number of variables involved in the violated editing rule. Considering a variety of editing rules that are additionally related by shared variables (like age regularly approaches in a multitude of editing rules) makes the problem of deciding for one combination profoundly complex. A comprehensive description of the problem in terms of propositional logic and an approach for linking it with the subsequent imputation process is given by Fellegi und Holt (1976). Since the underlying problem can be shown to be NP-complete, there are only approximate solutions available (for a promising branch-and-bound algorithm see de Waal und Quere, 2003). However, regardless of whether the decision which values are set to missing is done by more or less complex deterministic decision algorithms or clerical staff (which makes it incomprehensible for statistical approaches), the basic drawback

remains untreated: Since a "true" solution of the error-localisation problem is unavailable, in any case and for any approach, the price to be paid for a decision is additional uncertainty entering the survey process, which needs to be revealed for and included in inferences based on the suitably edited data set.

References

Biemer, Paul P. und Lars E. Lyberg (2003). *Introduction to Survey Quality*. Hoboken: Wiley.

Brackstone, Gordon (1999). *Managing Data Quality in a Statistical Agency*. Survey Methodology, 25(2): 139–149.

Chambers, Ray (2001). *Evaluation Criteria for Statistical Editing and Imputation*. National Statistical Methodological Series, 28.

Eurostat (2003). *Standard Report*. , Working Group on Assessment of Quality in Statistics.

Fellegi, Ivan P. und D. Holt (1976). *A Systematic Approach to Automatic Edit and Imputation*. Journal of the American Statistical Association, 71(353): 17–35.

Horvitz, D. G. und D. J. Thompson (1952). *A Generalisation of Sampling without Replacement from a Finite Universe*. Journal of the American Statistical Association, 47(260): 663–685.

Little, Roderick J. A. und Donald B. Rubin (2002). *Statistical Analysis with Missing Data*. New York: Wiley, 2. Auflage.

Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

— (1996). *Multiple Imputation after 18+ Years*. Journal of the American Statistical Association, 91(434): 473–489.

Särndal, C.-E., B. Swensson und J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.

Schenker, Nathaniel und Jeremy M. G. Taylor (1996). *Partially Parametric Techniques for Multiple Imputation*. Computational Statistics & Data Analysis, 22: 425–446.

de Waal, Ton und Ronan Quere (2003). *A Fast and Simple Algorithm for Automatic Editing of Mixed Data*. Journal of Official Statistics, 19(4): 383–402.