Młodak, Andrzej

Article

An application of a complex measure to model–based imputation in business statistics

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Młodak, Andrzej (2021) : An application of a complex measure to model–based imputation in business statistics, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, Vol. 22, Iss. 1, pp. 1-28, https://doi.org/10.21307/stattrans-2021-001

This Version is available at:
https://hdl.handle.net/10419/236813

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 1–28, DOI 10.21307/stattrans-2021-001
Received – 02.03.2020; accepted – 17.07.2020

An application of a complex measure to model–based imputation in business statistics[1]

Andrzej Młodak[2]

ABSTRACT

When faced with missing data in a statistical survey or administrative sources, imputation is frequently used in order to fill the gaps and reduce the major part of bias that can affect aggregated estimates as a consequence of these gaps. This paper presents research on the efficiency of model–based imputation in business statistics, where the explanatory variable is a complex measure constructed by taxonomic methods. The proposed approach involves selecting explanatory variables that fit best in terms of variation and correlation from a set of possible explanatory variables for imputed information, and then replacing them with a single complex measure (meta–feature) exploiting their whole informational potential. This meta–feature is constructed as a function of a median distance of given objects from the benchmark of development. A simulation study and empirical study were used to verify the efficiency of the proposed approach. The paper also presents five types of similar techniques: ratio imputation, regression imputation, regression imputation with iteration, predictive mean matching and the propensity score method. The second study presented in the paper involved a simulation of missing data using IT business data from the California State University in Los Angeles, USA. The results show that models with a strong dependence on functional form assumptions can be improved by using a complex measure to summarize the predictor variables rather than the variables themselves (raw or normalized).

Key words: complex measure, ratio imputation, regression imputation, predictive mean matching, propensity score method.

1. Introduction

In this paper we aim to use imputation in order to obtain a data set that resembles the true data and that can be used for multiple purposes rather than a specific purpose.

[1] This paper is an extended and substantially modified version of results presented during the fourth European Establishment Statistics Workshop under the auspices of the European Network for Better Establishment Statistics (ENBES), held on 7th – 9th September 2015 in Poznań, Poland.

[2] Statistical Office in Poznań, Centre for Small Area Estimation, address: Statistical Office in Poznań, Branch in Kalisz, ul. Piwonicka 7–9, 62–800 Kalisz, Poland. E–mail: a.mlodak@stat.gov.pl and Calisia University – Kalisz, Poland. ORCID: https://orcid.org/0000-0002-6853-9163.

This is a common aim at national statistical institutes and other institutes that collect and disseminate statistical data (cf. e.g. De Waal *et al.* (2011)). For example, in order to protect the privacy of individual respondents and avoid disclosure of sensitive information, the actually collected data may not be released. Instead a national statistical institute may then opt to release imputed data that resemble the actually collected data as much as possible. Therefore, some tools of the statistical disclosure control (e.g. perturbative methods or construction of synthetic data) are, in fact, based on imputation models (cf. Hundepool *et al.* (2012)). The imputation is very important especially in the dissemination of microdata. It is because in official statistics (and not only here) a growing demand on detailed microdata has been observed in the last decades. Therefore, the production of maximally informative and secure microdata becomes crucial.

The main focus of this paper is to study the efficiency of the use of a complex measure as an auxiliary variable in some methods of model–based imputation, especially for business statistics. The complex measure reflects the diversification of objects (e.g. economic entities) in terms of a complex social or economic phenomenon, described by many variables. A measure of this kind is constructed in such a way as to ensure that information contained in the variables and mutual relationships between them are maximally exploited, which traditional models of dependency – i.e. regression function – can overlook (cf. e.g. Młodak (2014) or Malina and Zeliaś (1998)).

Using the proposed complex measure instead of using several auxiliary variables leads to a loss of less information. From a purely theoretical point of view, using the proposed complex measure instead of using several auxiliary variables cannot lead to better estimates. If it does, it means that the full potential of the imputation methods using several auxiliary variables is not used. However, from a practical standpoint, there are some compelling reasons for using the proposed complex measure, such as (1) for some imputation methods using a single complex measure instead of several auxiliary variables may be easier to implement in practice (this certainly holds for imputation methods that were designed for a single auxiliary variable, such as ratio imputation) and (2) in a single complex measure it may be easier to take outliers into account than in several auxiliary variables for which one would have to use multivariate outlier techniques. The additional motivation of the use of such an approach is that users of statistical information are looking now mainly for the provision of complex characteristics of macro–domains, such as, e.g. the labour market, infrastructure, environment, etc.

The proposed approach involves selecting from a set of possible explanatory variables for imputed information the best ones in terms of variation and correlation (called *diagnostic features*) and then replacing them with a single complex measure (called also the *meta–feature*) exploiting their whole informational potential.

This complex measure is constructed as a function of median distance of objects described by normalized diagnostic features from the benchmark of development, i.e. artificial object 'ideal' from the analysed point of view. The normalisation is performed using the Weber median (called sometimes also L_1*–median* or *geometric median*) being a point of multidimensional space minimizing the sum of Euclidean distances from the points representing given objects. In this way the resulting imputation models are simpler: they are less computationally demanding and easier to interpret, while being sufficiently efficient. Here, the utility of this solution will be verified using the following methods of model–based imputation: ratio imputation, regression imputation, regression imputation with iterative extension, predictive mean matching and propensity score method. Empirical analysis is conducted here using both simulated and real data. The assumptions of simulation account for circumstances most often observed in business statistics. That is, the values of variables were drawn from the multivariate log–normal distribution with relevant parameters such that the auxiliary variables are mutually correlated as little as possible and maximally – with the target variable. The missing data were modelled using the missing at random (MAR) approach taking into account the impact of auxiliary variables into lack of data for the target one. The efficiency of imputation is assessed using the estimates of precision (MSE of target parameter estimation using imputed data) decomposed into three components, including the term connected with the "pure" imputation effect. The MSE in the presented imputation should be not greater than when all original data are used and minimal as a measure of the precision of estimation.

This paper is structured as follows. In Section 2 we present the assumption and methods of selection of diagnostic variables in the taxonomic model as well as the construction of the complex measure on their basis using the Weber median and other ordinal statistics. Next, Section 3 provides a short description of the analysed model–based imputation methods and Section 4 discusses tools of quality assessment used in our investigation (i.e. approximate estimation of imputation precision). Section 5 contains methodology of conducted simulation study and its results. An empirical study which involved a simulation of missing data using IT business data from the California State University in Los Angeles, USA is presented in Section 6. Finally (Section 7) some conclusions are collected.

These results expand on some issues not included in the final version of relevant sections of „*Handbook on Methodology for Modern Business Statistics*", edited by L. Willenborg, S. Scholtus and R. van de Laar (Collaboration in Research and Methodology for Official Statistics), created in 2014 within the ESSnet (European Statistical System network) initiative MeMoBuSt (Methodology for Modern Business Statistics), but which were investigated during that project.

2. Construction of the complex measure

The complex measure is aimed at efficient creation of a single variable containing the information potential of many collected variables describing a composite social or economic phenomenon in given objects (e.g. economic entities). The construction of the complex measure consists of the following steps, described also (although in terms of interval data) by Młodak (2014).

Step 1. Choice of variables and data collection: one should use information which properly describes the subject of research. The collected variables containing such information should be measurable, complete and comparable. To improve data comparability, they should have the form of indices (i.e. need to be calculated per capita, per 1 km^2, per 1000 inhabitants, per enterprise, etc.). Keeping values expressed in absolute numbers (e.g. number of economic entities, total revenues, etc.) can lead to some distortion of results – some (often not numerous) objects are (by their nature or specific circumstances) characterized by values much greater than others (e.g. large cities versus rural areas). The use of indices (relatively small, from elsewhere) substantially reduces the scale–dependency of the whole procedure. Of course, if the final complex measure is used to impute or estimate values of some other target variable, the variable used to construct the complex measure should be strictly connected with the target variable.

Step 2. Verification of variables: firstly, the elimination of variables that are not efficient in discrimination of objects, i.e. dropping variables for which the absolute value of the coefficient of variation (CV) is smaller than an arbitrarily established threshold (usually 0.1 – cf. Młodak (2014)) is conducted. Such variables are regarded as not showing the diversification of the analysed objects and hence they are dropped. This procedure is justified by the assumption that taxonomic methods are applied to phenomena where the clear diversification of the analysed objects is expected and then complex measures should reflect such diversification. Next, variables are verified in terms of correlation – we eliminate variables that are too correlated with others (and, hence, carry similar information). Here the inverse correlation matrix method was used. Its diagonal entries belong to $[1, \infty)$ (cf. e.g. Neter, Wasserman and Kutner (1985)). If some of them are too large (more often greater than 10) then relevant variables are regarded as 'bad'. They can be eliminated, but not necessarily all. That is, if there are more than one 'bad' variable then one should exactly analyse correlations between such variables and on the basis of such a comparison make elimination which is as sparing as possible (i.e. as few variables as possible should be dropped) and simultaneously guarantees a sufficiently weak correlation of the remaining variables. The correlation verification requires then some subjectivity in taking decisions about elimination. In the case when the final complex measure will be used as an auxiliary

variable in the imputation or estimation model, the correlation verification is conducted usually taking into account also the correlation of possible diagnostic variables with the target one. That is, when the analysis of the correlation matrix analysis does not allow to take unambiguous decision about elimination of some 'bad' variables, the ones whose correlation with the target one is smaller than others are removed. This approach, proposed by Malina and Zeliaś (1998), exploits mutual connections between features. It is very important because the economy is a 'system of connected vessels' and therefore the variables should be perceived not separately, but rather jointly – as a whole. The set of variables which remains after verification is called the *set of diagnostic features*. Thus, each object i is described by values of diagnostic features $X_1, X_2, \dots, X_m$ and is represented by the point $\boldsymbol{\gamma}_i = (x_{i1}, x_{i2}, \dots, x_{im}) \in \mathbb{R}^m$, where x_{ij} denotes the value of diagnostic feature X_j for i–th object, $i = 1,2, \dots, n$, $j = 1,2, \dots, m$.

Step 3. Identification of the character of diagnostic features (variables after verification): considering the impact of variables on the situation of an entity with respect to a phenomenon of interest, we can distinguish three types of variables:

- *stimulants* – the higher the value, the better the situation of an object in this context (e.g. average monthly revenue or Gross Domestic Product per capita),
- *destimulants* – higher values indicate a deterioration of the entity's situation,
- *nominants* – variables which behave like stimulants below a certain critical point and may switch to being destimulants after crossing it. That is, below this point this feature has the characteristic of being a stimulant and above it – a destimulant. Or on the contrary – greater values (and simultaneously smaller than the optimum) are 'worse' whereas smaller (but greater than the optimum) are regarded as being 'better'.

The critical point for nominant can be identified by own experience or by consultation with famous experts. An alternative – and more formal – approach in this respect could be based on the Cramér–von Mises or Anderson–Darling test – the critical point will refer then to the extremum of theoretical distribution best adjusted to the empirical data (if it is U–shaped, of course).

Destimulants and nominants are converted into stimulants by taking their values with opposite signs (in the case of nominants this is done only to the part with destimulative properties).

Step 4. Normalization of features, aimed at obtaining a comparable form of diagnostic variables. There are many forms of normalization (see e.g. Zeliaś (2002)). To exploit all connections between them it is good to use the Weber median, i.e. the vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_m) \in \mathbb{R}^m$ minimizing the sum of Euclidean distance from points

$\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_n$ reflecting given objects (cf. Młodak (2006)). The normalisation formula is then as follows:

$$z_{ij} = \frac{x_{ij} - \xi_j}{1.4826 \cdot \underset{l=1,2,\ldots,n}{\text{med}} |x_{lj} - \xi_j|}, \quad i = 1,2,\ldots,n,\ j = 1,2,\ldots,m. \tag{1}$$

Recall that (cf. Rousseeuw and Leroy (1987)) a probabilistic premise for the use of the constant 1.4826 (approximatively equal to $1/(\varphi^{-1}(3/4))$, where φ is a distribution function of the normal distribution with an expected value of zero and a standard deviation equal to 1), is the fact that if $Y_1, Y_2, \ldots, Y_k$ are independent and identically distributed random variables having the normal distribution with a mean μ and a variance σ^2 ($\sigma > 0$), then $E\big(1.4826 \cdot \text{mad}(Y_1, Y_2, \ldots, Y_k)\big) \approx \sigma$ for sufficiently large natural k (which gives approximative standardization), where $\text{mad}(Y_1, Y_2, \ldots, Y_k) = \underset{i=1,2,\ldots,k}{\text{med}} |Y_i - \text{med}(Y_1, Y_2, \ldots, Y_k)|$ is the median absolute deviation of variables $Y_1, Y_2, \ldots, Y_k$. As we can see, the expression $\underset{i=1,2,\ldots,n}{\text{med}} |x_{ij} - \xi_j|$ used in (1) is a special modification of the median absolute deviation of X_j, i.e. $\text{mad}(X_j)$, where the classical median was replaced with the respective coordinate of the Weber median. To approximate the Weber median, Vandev (2002) proposed the iterative algorithm based on the Newton–Raphson procedure.

The normalisation (1) leads to minimization of scale–dependency of the final results of the procedure: the properties of Weber and classical median as well as relative character of the input variables ensures that outlying is usually strongly reduced without loss of information contribution.

Step 5. Definition and determination of the taxonomic benchmark of development – an artificial, ideal object is defined, with which others are compared. As it was noted by Młodak (2014), this object is usually described by the most desirable values of particular diagnostic features (in the normalized version). Because all diagnostic features are stimulants, one can assume that the benchmark is defined as being represented by the vector $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_m)$, where

$$\psi_j = \max_{i=1,2,\ldots,n} z_{ij}, \quad j = 1,2,\ldots,m.$$

Therefore, the object described by those values is regarded as being 'ideal'. This method can be perceived as being endogenic, because the benchmark is constructed on the basis of the internal properties of the analysed empirical model[3].

Step 6. Computation of distances of entities from the benchmark. A distance being a function of the absolute differences between the respective values for a given object

[3] Alternatively, the benchmark can be defined also in an exogenous manner, i.e. arbitrarily and independently from properties of the data and the model. Such approach can be justified by some commonly adopted standard occurring in some fields. For example, if we analyse some data concerning environmental protection, the values of the benchmark can be assumed as being represented by relevant thresholds of allowable pollution generated by factories established by the European Commission and valid in the European Union.

and for the benchmark is used here. Of course, it should be nonnegative, reflexive and symmetric to be well defined. There are many ways to define it. Now we use the median distance:

$$d_i = \underset{j=1,2,\dots,m}{\text{med}} \left| z_{ij} - \psi_j \right|, \quad i = 1,2,\dots,n.$$

Step 7. Determination of a synthetic measure. The synthetic measure $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)$ is constructed as a statistical function of distances of analysed objects from the benchmark. The central point of this construction is a postulate that this measure should be a continuous function of a position of the distance of a given object from the benchmark taking into account the general extreme values of such a distance in the model. In our studies it is based on median and median absolute deviation of distances. The synthetic measure enables then to better identify possible outliers, where the analysed situation is especially difficult. That is, we put

$$\eta_i = 1 - \frac{d_i}{\text{med}(\boldsymbol{d}) + 2.5 \cdot \text{mad}(\boldsymbol{d})}, \tag{2}$$

$i = 1,2,\dots,n$, $\boldsymbol{d} = (d_1, d_2, \dots, d_n)$, $\text{mad}(\boldsymbol{d}) = \underset{i=1,2,\dots,n}{\text{med}} |d_i - \text{med}(\boldsymbol{d})|$, where 2.5 is called the *robust threshold value*. It ensures that $[0, \text{med}(\boldsymbol{d}) + 2.5 \cdot \text{mad}(\boldsymbol{d})]$ represents approximatively the 90% confidence interval for $\boldsymbol{d}$. It allows for achieving sufficient robustness of $\boldsymbol{\eta}$ to outliers (cf. Rousseeuw and Leroy (1987)). The values belong usually to the interval $[0,1]$. Only in special extreme cases (i.e. when an object is a strong outlier) they can be negative. The highest the value of the index (2), the better the situation in the investigated context.

The aforementioned construction can be applied in many circumstances without strong general restrictions except for strict connection with the subject of interest (of which – for imputation and estimation) with the variable to be imputed/estimated and relevant quality of the input data. They should:

- be unambiguously and precisely defined,
- describe the analysed phenomenon as exhaustively as possible,
- maintain proportionality of representing partial phenomena,
- provide measurable, available and complete statistical information for all investigated objects.

Of course, the thresholds of proper variability and sufficiently small correlation should be carefully established. The use of statistics of observations (such as 'classical' median or the Weber median) allows for an increase in the resistance to extreme, but a few, outliers. Some difficulties with application of the Weber median may occur when the complex measure will be constructed for several consecutive periods in time. Irregular perturbations at only few point may result in a large change of the position of the Weber median (cf. Durocher and Kickpatrick (2009)). However, it concerns only

the situation on the plane ($\mathbb{R}^2$) when the change is neither along nor almost along the ray starting at previous Weber median and going through previous location on a given point (i.e. if change of $\boldsymbol{\gamma}_i$ into $\boldsymbol{\gamma}_i'$ is that $\boldsymbol{\gamma}_i'$ is located on – or very close to – the ray $\overrightarrow{\boldsymbol{\theta}\boldsymbol{\gamma}_i}$, where $\boldsymbol{\theta}$ is the Weber median of $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_n$). In practice, however, such a situation occurs very rarely: the taxonomy based on the Weber median is most efficient when it uses at least three diagnostic features (i.e. when $m \geq 3$); on the other hand, changes in the consecutive periods in time are usually relatively small. Moreover, the Weber median is unique for $n \geq 3$ and $m \geq 2$ (cf. e.g. Milasevic and Ducharme (1987)). Otherwise, it is assumed to be equal to the classical median.

The advantages of the construction are as follows:

- the complex measure provides clearly interpretable information about the whole multivariate phenomenon,
- owing to standardization and normalization it is independent of differences between diagnostic variables in the impact on the general situation of an object and in the scale of measurement,
- it exploits maximally possible connections between diagnostic features – even those which cannot be statistically quantified and – in the case of imputation or estimation – also with the target variable,
- it contains the maximum information potential of particular diagnostic variables,
- its distribution is – in some sense – a resultant of distributions of diagnostic variables (of which in terms of variation).

The last two properties can be arguments showing greater benefits coming from the use of the complex measure then, e.g. the Principal Component Analysis (PCA) – (cf. e.g. Jolliffe (2002)). The PCA generates usually several components each of them being often of the other quality expressed by the share in total common variance. In this case, the first of them, i.e. the one whose share is the greatest, is usually assumed as the complex measure. In practice, however, the loss of original variation borne in this way is often rather considerable. Moreover, in extremal situations the shares can be even equally distributed among components obtained by PCA. In contrary to PCA, the presented complex measure reflects a variation and shape of distribution of diagnostic features as maximally as possible. Similarly, the lasso regression (cf. e.g. Tibshirani (1996)), which – by the constraint for coefficients of regression – tends to marginalize some possible auxiliary variables, seems to be in the investigated context slightly doubtful: it poses a risk of omitting some important (although maybe statistically less significant) information. In examples given by Tibshirani (1996) sometimes even over a half of primarily considered auxiliary variables were omitted as their coefficients were zeros. The complex measure enables to avoid – to a large extent – such unnecessary loss of useful information.

3. Investigated methods of model–based imputation

Now, we describe briefly the methods of model–based imputation analysed in the paper and possibilities of implementation of the complex measure to them.

3.1. Ratio imputation

Ratio imputation consists in replacing missing values with the value of a known auxiliary variable multiplied by the ratio of some descriptive summary statistics of the variable with the missing value (e.g. mean, median or sum) and the relevant statistics for the auxiliary variable. It is tacitly assumed here that the ratio of the values of these variables for a given unit is the same as the ratio of some 'total' values of these two variables. For example, if data about the value of sales for an enterprise are missing, but its total expenditure amounts to €20,000, mean sales for the whole analysed group of enterprises which the given one belongs to is €30,000 and the mean expenditure is €21,000, then the predicted value of sales is computed as 20,000×(30,000/21,000) = 20,000×(10/7) = €28,571.43. Of course, depending on current circumstances, instead of summary statistics we can also use in this context relevant values for a higher level of data aggregation (i.e. the total value for a given NUTS unit). There are also some special cases of ratio imputation such as, e.g. its weighted option, suggested by Arcaro and Yung (2001).

If there are several variables which are strictly connected with the imputed one, we can optimize the choice of the variable to be used for the imputation, e.g. by analysing the distribution of the known values of the imputed variable and appropriate values of the possible auxiliary variable (e.g. using the Wilcoxon signed rank or Cramér – von Mises – in the version for two samples – test). The auxiliary variable, for which such a 'trimmed' distribution is closest to the distribution of the known value of the variable to be imputed, will serve as the basis for the ratio imputation. Other – and faster – possibility is to compute the Pearson's correlation coefficient and chose such a variable which is most correlated with the target one. These methods do not, however, guarantee an exploitation of the whole information potential concerning the connections between original variables. Therefore, the use of the complex measure $\boldsymbol{\eta}$ seems to be desirable in this situation.

The quality of this type of imputation depends, first of all, on the degree of association between imputed and auxiliary variables. The stronger it is, the better the adjustment of imputed values is. The usefulness of this attempt depends on the availability of an appropriate auxiliary variable. The variance is much less biased than in the case of mean imputation.

3.2. Regression imputation

The main idea of the *regression imputation* is that missing values are replaced with predicted values established using a specific regression equation constructed on the basis of available data for the variable with gaps (as the value of the dependent variable resulting from the regression model) and some fully available auxiliary variables treated as explanatory variables. This approach is aimed at predicting missing values in such a way that the imputed value should be as close as possible to the unknown value. In this context it is very important to include in the model as many explanatory variables as possible (provided, however, they are not strongly correlated). This action can significantly improve the quality of prediction. Such a construction seems to be technically sophisticated and its application requires much more time than many other imputation methods (e.g. in comparison with the situation when the regression equation is constructed separately for each variable to be imputed). Of course, this method does not guarantee that the implants will be fully plausible values, but one can expect that the deviation of implants from their appropriate expectations will be relatively at their lowest.

The basic regression model is given by:

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \boldsymbol{\varepsilon} \tag{3}$$

where $Y = (y_1, y_2, \dots, y_n)$ is the target variable with gaps, $X_1, X_2, \dots, X_m$ $(m \in \mathbb{N})$ – auxiliary variables and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ – the disturbance. OLS estimator of coefficients has the form $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_\mathrm{r}^\mathrm{T}\mathbf{X}_\mathrm{r})^{-1}\mathbf{X}_\mathrm{r}^\mathrm{T}\mathrm{Y}_\mathrm{r}$, where $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m)$, $\mathbf{X}_r$ and Y_r are matrix $\mathbf{X} = [x_{ij}]$, $x_{i0} = 1$, $i = 1,2,\dots,n$, $j = 0,1,2,\dots,m$ and vector Y restricted only to those units $r_1, r_2, \dots, r_q$, $r_l \in \{1,2,\dots,n\}$, $l = 1,2,\dots,q$, $q < n$, for which data on Y are available, respectively[4].

One can formulate now a question: how to choose the auxiliary variable? The basic criterion in this respect should be that the strict connection with the target variable must be retained. But to effectively conduct these types of imputation, each such variable should provide a unique and large information resource. Duplicating information should be avoided. It means that we have to establish such a set of variables which are mutually weakly correlated and simultaneously retain their mutual multivariate connection. This goal may be achieved by using the reversed correlation matrix and its analysis described in Section 2 (step 2). Instead of many auxiliary variables, we can use one, the complex measure (2) containing information provided

[4] Of course, if data on the analysed variable were collected in a sample survey and some population value is to be imputed/estimated, one can include survey weights to the OLS estimator of $\boldsymbol{\beta}$. It will be then of the form $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_r^\mathrm{T}\mathbf{W}_r\mathbf{X}_r)^{-1}\mathbf{X}_r^\mathrm{T}\mathbf{W}_r Y_r$, where $\mathbf{W}_r = \mathrm{diag}(w_{r_1}, w_{r_2}, \dots, w_{r_k})$ is the matrix of sampling weights restricted to those units for which data on Y are available.

by auxiliary variables and taking their connections into account. That is, the model (3) takes the form

$$Y = \beta_0 + \beta_1 \boldsymbol{\eta} + \boldsymbol{\varepsilon}. \tag{4}$$

In fact, the regression imputation is the generalized ratio imputation.

Next three methods have a specific form. They are based on iterative algorithms generating successive approximates of implants to obtain results of sufficient quality. Therefore, in fact, they are special cases of multiple imputation – an approach consisting of producing a number of complete data sets from the incomplete data by imputing the missing data finite number of times by some assumed model-based method. Then, each completed data set is analysed and the results are combined to achieve final imputed values and related inference (cf. Rubin (1987)).

3.3. Regression imputation with iterative extension

Regression imputation with iterative extension seems to be an efficient improvement of the relevant classical approach. Let $\hat{\sigma}^2\mathbf{V}$ (where $\mathbf{V} = (\mathbf{X}_r^{\mathrm{T}}\mathbf{X}_r)^{-1}$ and $\hat{\sigma}^2$ is the estimated variance of Y) be a covariance matrix for the model with Y being the explained variable in (3). Determination of imputed values for each imputation is performed such that we start from the model (3) and next new parameters $\boldsymbol{\beta}_* = (\beta_{*0}, \beta_{*1}, \dots, \beta_{*m})$ and $\hat{\sigma}_*^2$ are drawn from the posterior predictive distribution of the parameters. That is, they are simulated from $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m)$, $\hat{\sigma}^2$ and $\mathbf{V}$. The variance has the form $\sigma_*^2 = \hat{\sigma}^2(n - m - 1)/g$, where g is a random number from the χ^2_{n-m-1} (chi–square with $n - m - 1$ degrees of freedom) distribution and n is the number of non–missing data in Y. The regression coefficients are computed as $\boldsymbol{\beta}_* = \widehat{\boldsymbol{\beta}} + \sigma_*\mathbf{V}_{(c)}^{\mathrm{T}}Z$, where $\mathbf{V}_{(c)}^{\mathrm{T}}$ is the upper triangular matrix in the Cholesky decomposition of $\mathbf{V}$, i.e. $\mathbf{V} = \mathbf{V}_{(c)}^{\mathrm{T}}\mathbf{V}_{(c)}$ and Z is a vector of $k + 1$ independent random normal variables (cf. Yuan (2010)).

The missing values are then replaced by predictors obtained from the equation

$$Y_r = \beta_{*0} + \sum_{j=1}^{m} \beta_{*j} X_{rj} + z_r\sigma_*, \tag{5}$$

where X_{rj} are the values of covariates for such units for which data on Y are unavailable and z_r is a simulated normal deviate, $r = 1,2, \dots, n$. This operation can then be repeated starting from the formula (5) and so on. The number of iterations depends on the assumptions of the quality control (cf. Rubin (1987), Yuan (2010)). The synthetic measure (2) can be here also efficiently applied instead of the set of (sometimes numerous) covariates enabling a simplification of (5) to the form:

$$Y_r = \beta_{*0} + \beta_{*1}\boldsymbol{\eta}_r + z_r\sigma_*. \tag{6}$$

3.4. Predictive mean matching

Predictive mean matching is a method similar to the regression method with iterative extension except that instead of the main predictive equation for each missing value it imputes an observed value which is closest to the predicted value from the simulated regression model (cf. Yuan (2010), Horton and Lipsitz (2001)).

3.5. Propensity score method

Propensity score method is another way of applying regression imputation suggested by Little and Rubin (2002) and studied by Yuan (2010). The propensity score is understood as the conditional probability of assignment to a particular treatment, given a vector of observed covariates. In this method, the propensity score is generated for each variable with missing values to indicate the probability of that observation being missing. The observations are then grouped on the basis on these propensity scores and an approximate Bayesian bootstrap imputation (cf. Rubin (1987), p. 124) is applied to each group (Lavori *et al*. (1995))[5]. With a monotone missing pattern, the following steps described by Yuan (2010) are used to impute values for each variable Y with missing values:

1. Create an indicator variable Λ with the value 0 for observations with missing Y and 1 otherwise.
2. Fit a logistic regression model

$$\text{logit}(p) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \boldsymbol{\varepsilon}, \tag{7}$$

where $p = \Pr(\Lambda = 0 | X_1, X_2, \dots, X_m)$ and $\text{logit}(p) = \log(p/(1-p))$.

3. Create a propensity score for each observation to estimate the probability that it is missing.
4. Divide the observations into a fixed number of groups (typically assumed to be five) based on these propensity scores. This can be done by arbitrarily establishing some structure of intervals of propensity values and indicating observations whose propensity values belong to such particular intervals.
5. Apply approximate Bayesian bootstrap imputation to each group. That is, for a given group, suppose that Y_{obs} denotes the n_1 observations with nonmissing Y values and Y_{mis} denotes the n_0 observations with missing Y (where $n_1 > n_0$). Approximate Bayesian bootstrap imputation first draws n_1 observations randomly with replacement from Y_{obs} to create a new data set Y^*_{obs}. This is a nonparametric analogy of drawing parameters from the posterior predictive

[5] Of course, in the propensity score method not only Bayesian bootstrap can be used. This procedure is, however, most popular and seems to be efficient.

distribution of the parameters. The process then draws the n_0 values for Y_{mis} randomly with replacement from Y^*_{obs}. These values are implants.

Steps 1 through 5 are repeated sequentially for each variable with missing values. In our analysis also all auxiliary variables will be replaced with one synthetic measure (2), i.e. the formula (7) will take the form

$$\text{logit}(p) = \beta_0 + \beta_1 \boldsymbol{\eta} + \boldsymbol{\varepsilon}. \tag{8}$$

Yuan (2010) noted that the propensity score method was originally designed for a randomized experiment with repeated measures on the response variables. The goal was to impute the missing values to the response variables. The method uses only the covariate information that is associated with objects for which the imputed variable values are missing. It does not use correlations among variables. It is effective for inferences about the distributions of individual imputed variables, such as univariate analysis, but it is not appropriate for analyses involving relationships among variables, such as regression analysis (cf. Schafer (1999), p. 11). It can also produce badly biased estimates of regression coefficients when data for predictor variables are missing (cf. Allison (2000)).

4. MSE and its decomposition based on imputed data

The value of the imputation used should be evaluated using relevant methods of the quality control. That is, we have to assess whether the imputed values are best fitted and how the estimation precision of the population statistics using imputed data is influenced by adding or not adding disturbance terms to some models of imputations. This assessment can be done both at the stage of a preliminary simulation study or *ex post*, i.e. after performing the whole imputation process. Of course, in the latter case, it is much more difficult due to a shortage of material for efficient comparisons. Now, we will present how to estimate the Mean Squared Error (being a basic measure in this situation) for aggregated statistics based on imputed data.

Let A be the set of units in the sample, $\hat{\theta}_A$ be an estimator of θ computed using all sample data about the target variable. The variance estimation is strictly connected with θ. In general, Särndal (1992) showed that the total variance or – in terms of the theory of estimation – MSE of the estimator $\hat{\theta}$ of θ for the whole population[6], $\hat{V} = E\left(\hat{\theta} - \theta\right)^2$, can be decomposed in the sampling, imputation and mixed effect components:

$$\hat{V} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + 2\hat{V}_{\text{MIX}}, \tag{9}$$

[6] In many cases MSE coincides with variance of estimator. However, in a missing–data context due to a bias these two quantities cannot be equivalent. So, here we will use MSE as more informative index of quality.

where $\hat{V}_{\mathrm{SAM}} = E\left(\hat{\theta}_A - \theta\right)^2$, $\hat{V}_{\mathrm{IMP}} = E\left(\hat{\theta} - \hat{\theta}_A\right)^2$, $\hat{V}_{\mathrm{MIX}} = E\left(\left(\hat{\theta}_A - \theta\right)\left(\hat{\theta} - \hat{\theta}_A\right)\right)$ are the aforementioned components, respectively. The imputation term $\hat{V}_{\mathrm{IMP}}$ shows the part of variance resulting from expected deviation of imputed values from the true ones. Of course, in a simulation study, these expected values can be approximated by arithmetic means of relevant deviations obtained by the consecutive replications of sampling. J. K. Kim *et al.* (2006) analyse the problem of estimation of MSE in the complex sample design, when imputation is repeated q times, and derive a formula expressing the difference between the expected value of multiple imputation MSE estimation and the MSE of estimator $\hat{\theta}$. They have investigated such models for subpopulations (called also domains) and linear regression models. Arcaro and Yung (2001) propose approximately unbiased statistics for $\hat{V}_{\mathrm{SAM}}$, $\hat{V}_{\mathrm{IMP}}$ and $\hat{V}_{\mathrm{MIX}}$ using weighted mean and weighted ratio imputation with weights derived from traditional generalized regression estimator (GREG) for the mean based on relevant auxiliary data. Alternatively, they have analysed the MSE estimator for weighted ratio imputation with specially adjusted jackknife GREG weights as $\hat{V} = \sum_{j=1}^{h} \frac{n_j}{n_j - 1}\left(\hat{\theta}_j - \hat{\theta}\right)^2$ where $\hat{\theta}_j$ is the imputation estimator of θ, corrected using jackknife weights in the j–th stratum, n_j is the number of units belonging to this stratum, $j = 1,2, \ldots, h$, and $h \in \mathbb{N}$ is the number of strata in the sampling design. Of course, this algorithm is also efficient if instead of strata repetitions of simple random sampling in the simulation study are used, despite the fact that the samples could not be disjoint. Assuming that $\hat{\theta}$ is unbiased, Kim (2000) proposes unbiased unweighted MSE estimators for regression and ratio imputation models. Fuller and Kim (2005) proved the formula for fully efficient fractionally imputed MSE estimator based on the squared deviation of mean estimator and response probabilities, in particular imputation cells and subpopulations. Similar research for the balanced random imputation has been conducted by Chauvet *et al.* (2011).

In the case of the *ex post* quality control, we have a much more serious problem. Because there is no exact reference platform (the 'true' distribution of the target variable is unknown), it is necessary to rely only on an approximate estimation of imputation precision using approaches (cf. e.g. Andridge and Little (2010), who divide a sample into several complete data sets and, instead of repeated sampling, investigate a division of the population into complete and disjoint data sets and then estimate the MSE). This division might be based on an auxiliary variable (most preferably categorical) strictly connected with the target one. These classes should have approximately equal number of elements. Hence, we can obtain an estimate of error.

However, using decomposition (9) seems to be a better solution. That is, assuming that the units were sampled independently and $\hat{\theta}_A = \sum_{i \in A} y_i^* / |A|$, the MSE will be approximated by

$$\tilde{V} = \frac{1}{|A|^2} \sum_{i \in A} \left(y_i^* - \hat{\theta}_A\right)^2 = \tilde{V}_{\mathrm{SAM}} + \tilde{V}_{\mathrm{IMP}} + 2\tilde{V}_{\mathrm{MIX}}, \tag{10}$$

where the relevant components are estimated using the following statistics:

- sampling effects

$$\tilde{V}_{\text{SAM}} = \frac{1}{|A|^2}\sum_{i\in A}\left(\tilde{y}_i - \hat{\theta}_A\right)^2, \tag{11}$$

- imputation effects

$$\tilde{V}_{\text{IMP}} = \frac{1}{|A|^2}\sum_{i\in A}(y_i^* - \tilde{y}_i)^2, \tag{12}$$

- mixed effects

$$\tilde{V}_{\text{MIX}} = \frac{1}{|A|^2}\sum_{i\in A}(y_i^* - \tilde{y}_i)\left(\tilde{y}_i - \hat{\theta}_A\right), \tag{13}$$

where $y_i^* = py_i + (1-p)\hat{y}_i$ and $\tilde{y}_i = py_i + (1-p)\left(\left(n\hat{\theta}_A - |U|\hat{\theta}_U\right)/(n-|U|)\right)$, with $p = 1$ if the value of Y for i–th unit is available and $p = 0$ otherwise, $\hat{\theta}_U = \sum_{i\in U}\hat{y}_i/|U|$ is the estimator of θ obtained on the basis of imputation for the set of units for which data on Y are unavailable (U), where $|\cdot|$ denotes the cardinality of a given set. The value $\tilde{y}_i$ is equal to y_i if y_i is available and to mean of known values of Y otherwise. Thus, we can perform detailed diagnostics of our imputation method.

The formulas (10) – (13) are designed for single imputation. Instead, one can also use multiple imputation in which case the variance can be imputed using the well-known pooling rules by Rubin (1987). First of them is the within–imputation variance – the average of the mean of the within variance estimate, i.e. squared standard error – which reflects the sampling variance, i.e. the precision of the parameter of interest in each completed data set:

$$\hat{V}_w = \frac{1}{l}\sum_{i=1}^{l}\widehat{SE}_i^2,$$

where l is the number of imputed data sets and $\widehat{SE}_i$ – estimate of the sum of squared standard errors observed in i–th imputed data set, $i = 1,2,\dots,l$. The smaller sample, the larger the within–imputation variance.

The second rule concerns the between–imputation variance, which reflects the extra variance occurring due to the missing data and is estimated by taking the variance of the parameter of interest estimated over imputed data sets. It is computed using the formula

$$\hat{V}_b = \sqrt{\frac{\sum_{i=1}^{l}\left(\hat{\theta}_i - \hat{\bar{\theta}}\right)^2}{l-1}},$$

where $\hat{\theta}_i$ denotes the estimate of in i–th imputed data set and $\hat{\bar{\theta}} = \sum_{i=1}^{l}\hat{\theta}_i/l$ is the pooled estimate of θ, $i = 1,2,\dots,l$. The higher the level of missing data, the larger the between–imputation variance.

The total variance is then given by

$$\hat{V}_T = \hat{V}_w + \left(1 + \frac{1}{l}\right)\hat{V}_b.$$

5. Simulation study

To verify the efficiency of our method a simulation experiment was conducted. A sample consisting of 200 units was constructed to account for circumstances observed in business statistics. We have assumed that the target variable Y and three auxiliary variables X_1, X_2 and X_3 are considered. Their values were drawn jointly from the multivariate log–normal distribution. This sampling was realized by generating random vectors from the multivariate normal distribution with the vector of expected values $\mu = (3.5, 0.1, 1.3, -4.0)$ and covariance matrix of the form

$$\mathbf{\Sigma} = \begin{bmatrix} 2.5 & 1.0 & -1.7 & 2.4 \\ 1.0 & 1.3 & 0 & 0 \\ -1.7 & 0 & 3.4 & 0 \\ 2.4 & 0 & 0 & 6.7 \end{bmatrix}$$

and next exponentiating obtained results. Thus, Y is represented by first coordinate of any such vectors and X_1, X_2, X_3 – by second, third and fourth, respectively. Such an attempt is motivated by the following premises:

- most experts experienced in business statistics argue that the log–normal distribution is the best way to approximate the distribution of variable occurring in this field,
- the auxiliary variables should be uncorrelated each with other, but they all should be clearly correlated with the target variable; in the investigated case the expected Pearson correlation coefficients of X_1, X_2 and X_3 with Y are 0.5547, -0.5831 and 0.5864 respectively,
- in commonly used professional statistical software (SAS, R, etc.) the random number/vectors from log–normal distribution can be generated only by exponentiating values/vectors drawn only from the normal distribution. Moreover, in the multivariate case it is required that covariance matrix is positive definite[7].

In practice, the possibilities of establish and that all aforementioned conditions are simultaneously satisfied is slightly restricted. We have chosen the best possible solution in this situation.

Another problem was connected with the modelling of non–response. That is, it should be decided how to choose records for which data on Y are assumed to be missing and have to be imputed. Of course, the simplest way to do it seems to be taking a random subsample of the original data and drop values of Y for them (it is the so–called *Missing Completely at Random* – MCAR condition, according to the terminology

[7] It is not difficult to prove that the $n \times n$ covariance matrix is arrowhead (i.e. it has non zero diagonal and first row and first column entries), positive definite and produces Pearson's correlation matrix, whose entries in the first row and the first column are greater (in terms of absolute values) than 0.5 only if $n \leq 4$.

introduced by Little and Rubin (2002)). Unfortunately, this attempt is rarely used in practice – mainly due to the fact that its application leads usually to unbiased estimation. Thus, also differences between various methods in this context can be just random sampling fluctuations. Therefore, the MAR (*Missing at Random*) scenario, where it is assumed that the missing data mechanism depends on the auxiliary (X) variables, could be a more appropriate solution here. That is, the missingness can be explained by variables, for which full information is available. It makes MAR different than MCAR, where it is assumed that gaps in data result from random results. According to the relevant proposals, which can be found in literature (cf. e.g. Pampaka *et al.* (2016)), we use the logit model here, in which the importance of particular auxiliary variable is taken into account, i.e.

$$\text{logit}(p_i) = -0.5 + 0.3x_{i1} - 1.2x_{i2} + 0.2x_{i3},$$

where p_i is the probability that the value y_i is missing and x_{ij} denotes the value of X_j for i–th object, $j = 1,2,3$. The value y_i was regarded as missing if $\hat{p}_i \geq 0.5$ where $\hat{p}_i$ is the estimate of p_i obtained from this model, $i = 1,2,\dots,n$. The structural parameters (-0.5, 0.3, -1.2 and 0.2) in this model were established a priori so that to ensure various connections of auxiliary variables with possibility of lack of data on Y and simultaneously usually reasonable expected number of such missing items.

The following imputation methods were used:

- ratio imputation (denoted as R),
- ratio imputation with the complex measure (RM),
- regression imputation (RG),
- regression imputation with complex measure based on (4) (RGM),
- regression imputation with iteration (RGI),
- regression imputation with iteration based on complex measure using (6) (RGIM),
- predictive mean matching (PMM),
- predictive mean matching with the complex measure (PMMM),
- propensity score method (PSM),
- propensity score method with the complex measure using (8) (PSMM).

The classical ratio imputation was conducted using such an auxiliary variable which was best correlated with Y (in terms of available data). The complex measure used instead of the set of three independent auxiliary variables was determined using the formula (2) based on the normalization using the Weber median (1), taxonomic benchmark and distance of units from the benchmark indicated in steps 5 and 6 of the procedure described in Section 3. The experiment was conducted using especially constructed algorithm prepared in the SAS Enterprise Guide 4.3. software (and especially its IML environment). In the case of the regression, predictive mean

matching and propensity score method the mi procedure was used. In the case of RGI, RGIM, PMM and PMMM methods 10 iterations and for PSM and PSMM – 2 iterations were done. It was sufficient to ensure relevant quality of imputed data. The whole experiment was replicated 1000 times. It is worth noting that each replication covered both drawing new samples and generating the missing value.

Using the results of imputation we have computed the analysed measures of quality. That is, we have used the maximum distance from imputed values and the following measures of quality of estimation of expected value of Y using the empirical arithmetic mean of the available and imputed values: bias, MSE, estimated MSE (using only available data for Y supplemented with the imputed ones, formula (10)), components of the MSE (according to (10)) and the estimated MSE (formulas (11), (12), (13)).

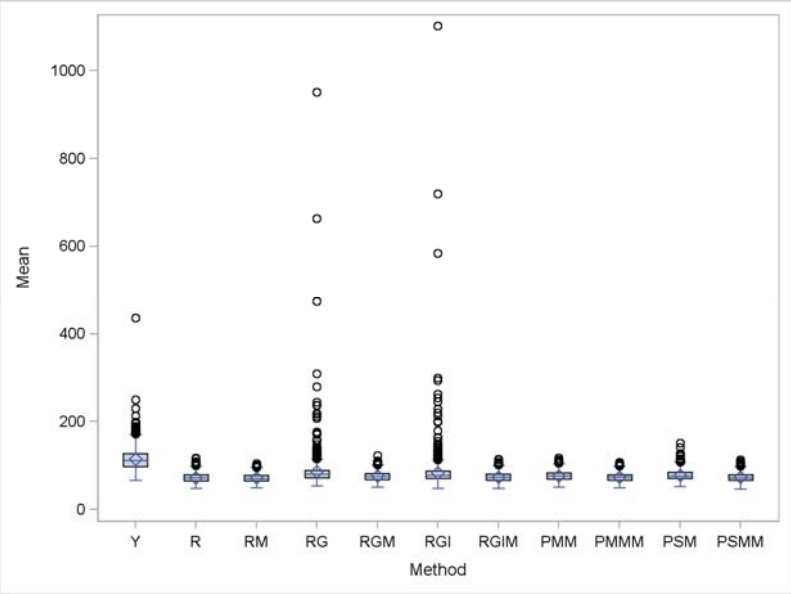

Figure 1. Box–and–whisker visualisation of distributions of the target variable Y means – original and with imputed values – by methods of imputation.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

Figure 1 presents box–and–whisker plots reflecting the distribution of means of original, complete, values of Y and means of modified Y where modelled missing values are replaced with implants generated using particular methods presented above. These means were computed for relevant units over 1000 replications. One can observe that using the complex measure instead of the simple collection of auxiliary variables a substantial reduction of outlying (expressed especially by extreme values of the means) results of regression imputation (also with iteration) is achieved. Such an improvement in this context is considerable also for the propensity score method. In any case, the population mean of Y seems to be slightly underestimated, but this bias is not great.

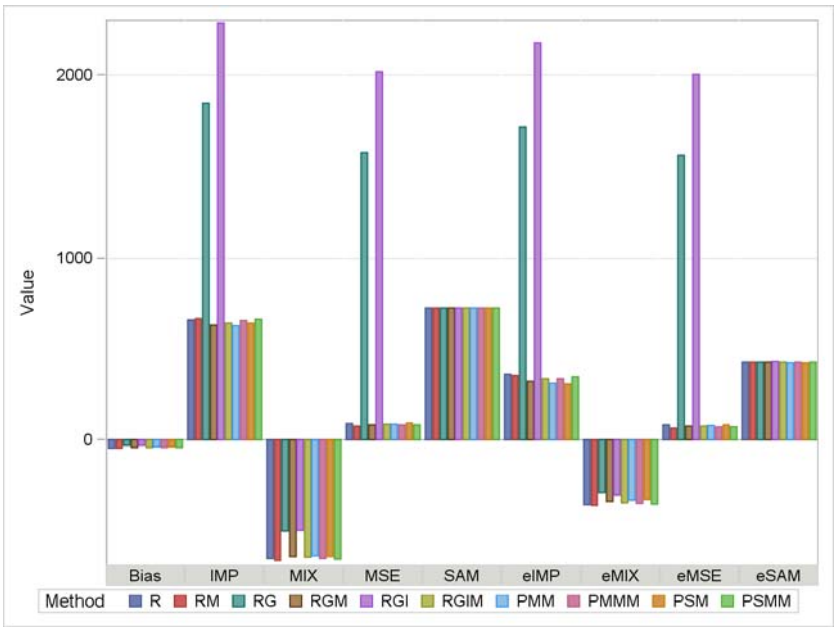

Figure 2. Comparison of quality indicators for estimation of mean of Y using original and imputed data by methods of imputation.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

The results of the experiment in terms of quality indicators of imputation are presented in Figure 2. It contains the average values of these indices computed over 1000 trials. MSE, SAM, IMP and MIX denote here the mean square error and its sampling, imputation and mixed effect components and eMSE, eSAM, eIMP and eMIX their estimators, computed using formulas (10), (11), (12) and (13), respectively. One can observe that using the complex measure substantially improves the quality of imputation for R, RG and RGI methods. Such an improvement is especially considerable in terms of MSE and its imputational component. For the remaining methods this advantage is lower, but also observable. The sampling component (SAM) does not depend on the imputation methods, by the assumption. Some very small increase in IMP and MIX (in absolute values) components after using the complex measure in R, PMM and PSM methods may be a result of slightly higher bias of imputed values in this case, which was observed already in Figure 1.

It would be also interesting to employ the imputation methods when the complex measure (2) is used and when all variables are used, but in application of imputation methods based on individual variables, these variables are normalized according using the formula (1). Figures 3 and 4 show distribution of target variable means and quality indicators obtained after 1000 replications made according to the above described principles.

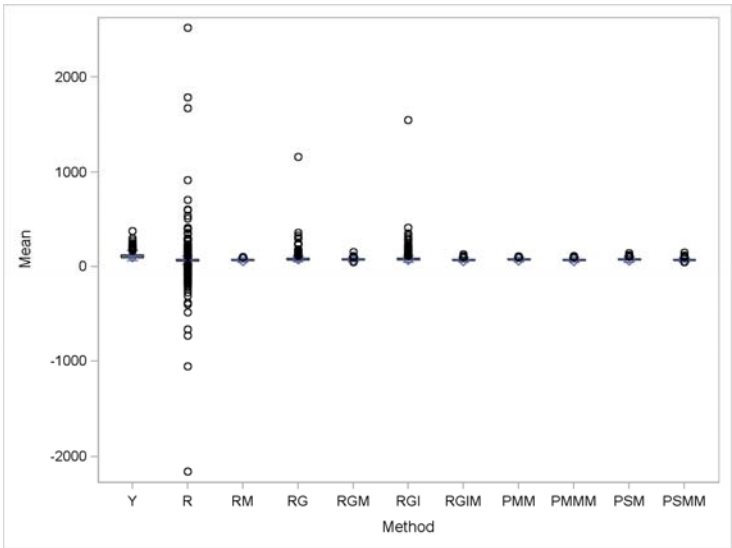

Figure 3. Box–and–whisker visualisation of distributions of the target variable *Y* means – original and with imputed values – by methods of imputation, variables normalized using the Weber median.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

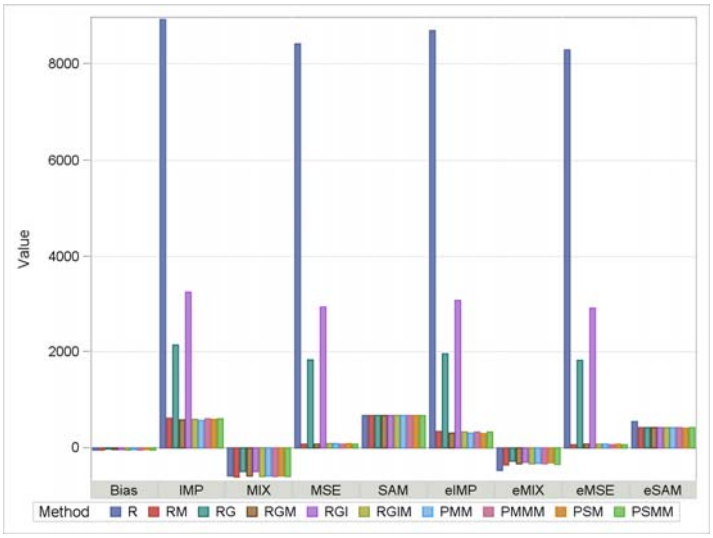

Figure 4. Comparison of quality indicators for estimation of mean of *Y* using original and imputed data by methods of imputation, variables normalized using the Weber median.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

One can observe that majority of our previous observations was confirmed. That is, the reduction of imputation error when using the complex measure instead of

variables normalized using the Weber median is especially considerable for R, RG and RGI methods. In other cases, the advantage of such replacement is much smaller. The imputation based on the complex measure allows also for reduction of distorting outliers in the imputed values. These results lead to the conclusion that for some methods the use of the complex measure is more efficient than when the individual variables in model-based imputation are used – independently from that whether these variables are normalized or not.

6. Empirical study

Our alternative study was based on data on 36 firms representing the IT sector in Bermuda, Canada, China, Denmark, Finland, Germany, Greece, Indonesia, Japan, Mexico, Russia, South Korea, Sweden, the UK and the USA, stored on Instructional Web Server of the California State University in Los Angeles, USA (http://instructional1.calstatela.edu/mfinney/Courses/491/hand/sas_exercise/tech3.xls). Here, the following five variables are recorded: Return on Equity (ROE, %), Revenues (in million $), Revenue Growth (RGR, %), Total Shareholder Return (TSR, %) and Profits (PR, in million $).

The set originally contained 39 firms, but due to missing data for ROE three of them had to be dropped. For the purposes of the study, the revenues were chosen as the variable to be imputed. To implement the non-response we use – similarly as in the case of the simulation study (Section 5) – the MAR condition for missing data. Two independent options of the logit model of missingness of data on revenues were applied. The first of them takes the correlation of available variables with revenues into account, the second one underlies the variability of ROE, RGR, TSR and PR. On the other hand, we had also to remember about upper limits of computational capability of the software (connected with exponentiating used in determination of estimates of the missingness probability). Finally, the analysed logit models had the following forms:

- option 1: $\text{logit}(p_i) = 0.001ROE_i + 0.5RGR_i - 1.1TSR_i - 0.004PR_i$,
- option 2: $\text{logit}(p_i) = 0.27ROE_i + 0.005RGR_i - 0.005TSR_i - 0.04PR_i$.

Again, y_i is regarded as missing when $\hat{p}_i \geq 0.5$, $i = 1,2, \dots, n$. The use of the option 1 generated 7 gaps in data whereas the option 2 resulted in 8 non–response items.

Similarly as in Section 5, we have now constructed the complex measure which will serve as a support for imputation. It was based on three indicators which are strongly diversified and weakly correlated with the revenues, selected according to Step 2 of the procedure described in Section 1, i.e. ROE, RGR and TSR. The variation of all possible variables was very high (i.e. the coefficient of variation amounted to 77.2% for ROE, 96.2% for RGR, 137.8% for TSR and 299.2% for PR). The diagonal entries of the inverse of correlation matrix are respectively: 1.5149, 1.9651, 1.6123 and 1.2122. Hence, the

possible auxiliary variables are mutually weakly correlated. Because PR is weaker correlated with the target variable than others (the relevant Pearson's correlation coefficient amounted to 0.1905, whereas for the remaining variables it exceeded 0.21 or – for TSR – even 0.3), we decided to remove it. Thus, we have obtained the set of diagnostic features. All of them are stimulants for revenues. We have computed the Weber median for them. Its coordinates amounted to 21.5131, 21.9906 and 26.9717 respectively. Next, – according to Step 4, we have normalized the diagnostic features using the formula (1). These normalized values were the basis of computation of the final complex measure by determination of the taxonomic benchmark of development (step 5), which was described by the vector (3.5726, 2.4161, 5. 3996), computation of distance of entities from the benchmark (Step 6) and the values of the complex measure (Step 7, formula (2)).

To impute simulated missing data, the same methods as in Section 5 were applied, i.e. ratio imputation (R), regression imputation (RG), regression imputation with iteration (RGI), predictive mean matching (PMM) and propensity score method (PSM), each of them also with the option based on the complex measure (RM, RGM – with formula (4), RGIM – formula (6), PMMM and PSMM – with formula (8), respectively). In classical ratio imputation the Total Shareholder Return was used as a reference variable, because it is the one most correlated with the target variable.

In Figure 5, the means of revenues and their relevant 95% confidence intervals for original (Y) and imputed data – when the option 1 is applied – are visualised. The vertical dashed line shows the true mean of Y. We can observe that the use of the complex measure in the case of ratio imputation substantially improves the precision of mean estimation. Better adjustment in this context is considerable also for RG and PMM when explanatory variables are replaced with the complex measure.

Table 1 shows the comparison of basic statistics describing the shape of original distribution of Y with complete data and distribution of Y where data gaps were filled up by imputed values obtained using a relevant imputation method. One can notice here that differences between respective extremes (minimums and maximums), median and quartiles of the original variable and its imputed version confirm our earlier conclusions to a very large extent. It is possible that many of them are caused by existing incidentally very small or very high values of imputed values, which is the integral part of risk connected with any imputation. The relatively high coefficient of variation seems to be a good justification of this view. Moreover, the skewness and kurtosis of distribution of 'complete' Y is better approximated by option based on the complex measure for the predictive mean matching (PMMM) and diversification (expressed by CV=136.2%, whereas for Y CV=130.4%) – by the propensity score method (PSMM).

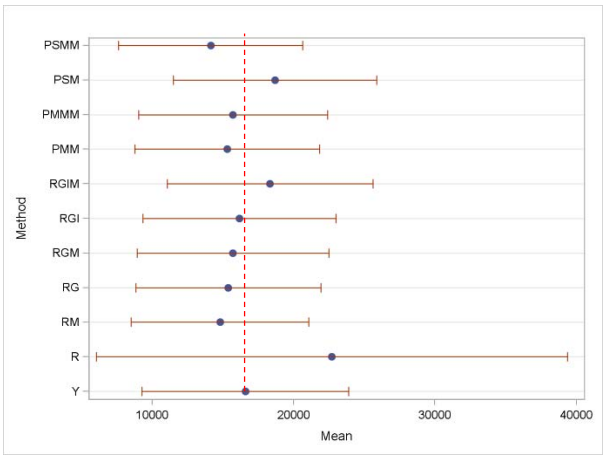

Figure 5. Means and their 95% confidence intervals for original and imputed data on revenues (in million $) – option 1 of missing data model.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

Respective results of the use of option 2 in modelling data missingness are presented in Figure 6 and Table 2. They show again that most of the imputation methods underestimate the mean of Y. The use of the complex measure improves the quality of estimation of the mean in the case of ratio, regression and propensity score methods. Also, in this case the PMMM approach better approximates skewness and kurtosis whereas PSMM – CV and skewness of the original values (Y) than PMM and PSM, respectively.

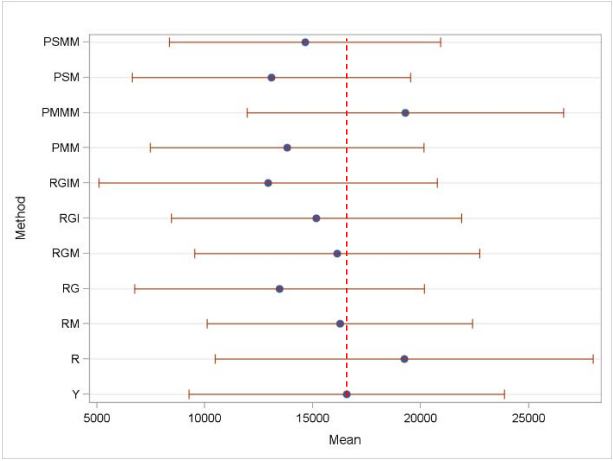

Figure 6. Means and their 95% confidence intervals for original and imputed data on revenues (in million $) – option 2 of missing data model.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

Of course, in contrast to the simulation study, data used here are arbitrary and fixed. Hence, the coefficients in the MAR condition formulas are established once and, in consequence, the imputation was done also once. Since most of the analysed imputation methods are stochastic and the result of one execution is random, it is difficult to formulate general conclusions on this basis. However, we have taken one more trial which allowed for an increase of random contribution to MAR condition and can to some extent verify the aforementioned observations.

In the second attempt for imputation based on individual variables we have taken their values normalized using the Weber median according to the formula (1). Moreover, we ensure that the number of removed (and imputed) values of the target variable will be relatively large, but not too large. More precisely, we assumed that the number of removed values should be not smaller than 11 and nor greater than 17. The MAR condition was still logit: $\text{logit}(p_i) = \beta_1 \cdot ROE_i + \beta_2 \cdot RGR_i + \beta_3 \cdot TSR_i - \beta_4 \cdot PR_i$, where – to take the correlation of auxiliary variables with the target one – β_1 and β_2 were sampled from the uniform $U\left(-\left(0.7 \cdot \frac{k}{1000}\right), 1 - \left(0.7 \cdot \frac{k}{1000}\right)\right)$ distribution whereas β_1 and β_1 – from $U\left(0.7 \cdot \frac{k}{1000}, 1 + \left(0.7 \cdot \frac{k}{1000}\right)\right)$ where k= 2349 and takes the following values: $\beta_1 = 0.0000399$, $\beta_2 = 3.7883793$, $\beta_3 = 2.44456$ and $\beta_4 = 0.5172775$. This way, we have obtained data with 15 gaps. Figure 7 shows the comparison of means and their confidence intervals in this case.

One can observe that the application of the complex measure in the model-based imputation allows often for tightening the confidence imputation for means. We have also repeated this simulation for some other k, which allows for satisfaction of all above described assumptions and the results were quite similar.

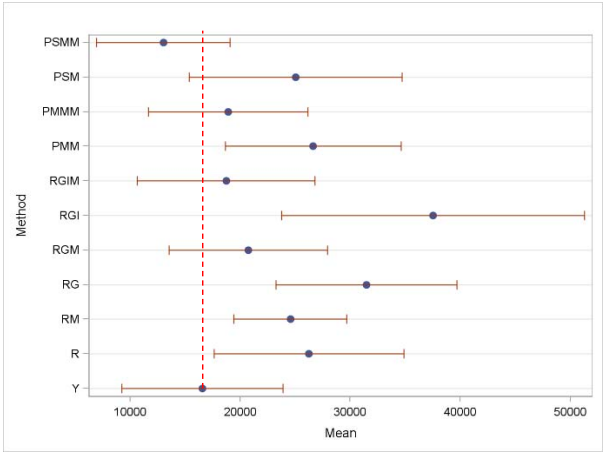

Figure 7. Means and their 95% confidence intervals for original and imputed data on revenues (in million $) – data normalized using the Weber median and random coefficients in MAR.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

7. Conclusions

The main conclusions which can be formulated on the basis of our studies are as follows. Construction of a complex measure using ordinal statistics (of which the Weber median) ensures a more efficient exploitation of mutual connections between possible auxiliary variables and therefore more informative imputation (cf. Młodak (2006, 2014)). Using the complex measure instead of one or more auxiliary variables is especially favourable for ratio and regression imputation, where the improvement in quality expressed by bias or Mean Squared Error is usually large. In many cases similar observation can be done for the regression imputation with iteration. For the predictive mean matching and propensity score method replacing the summarized predictor variables themselves by the complex measure gives not such considerable advantages – probably due to additional correction mechanism built into these procedures (i.e. indicating best 'donor' of imputed value in the former and grouping observations and bootstrap in the latter case). However, also in these cases using the complex measure can lead to better reflection of some important characteristics of distribution of the original variable, i.e. variation, skewness or kurtosis, which describe its shape.

It is worth noting that the efficiency of particular imputation algorithms in the simulation study depends, among other things, on the MAR condition being regarded as inherently strong. The logit model used in the presented simulation study assumes that the occurrence of data gaps of the target variable depends on all fully available variables. Such an attempt enables to exploit connections between analysed variables (both with incomplete and complete data) and, in consequence, the whole information potential of the database. It then proves to be more systematic than random reasons of the gaps.

The complex measure provides more stable results, i.e. gives substantially lower risk of excessive outliers. However, one should remember that the conditions for the efficient use of the complex measure are: proper choice of auxiliary variables on the basis of which it is constructed and the method of its construction. The choice of model–based imputation method which has to be applied with the complex measure depends on the main aim of the imputation: for estimation of mean or similar population statistics ratio or regression imputation is recommended; if a scientist is more interested in approximation of the shape of unknown distribution, the predictive mean matching or the propensity score method can provide better effects for him/her.

Acknowledgement

The author is very grateful to two anonymous reviewers for valuable comments and suggestions, which substantially contributed to improvement of quality of this study.

References

ALLISON, P. D., (2000). Multiple Imputation for Missing Data: A Cautionary Tale, Sociological Methods and Research, Vol. 28, pp. 301–309.

ANDRIDGE, R. R. and LITTLE, R. J. A., (2010). A Review of Hot Deck Imputation of Survey Non–response, International Statistical Review, Vol. 70, pp. 40–64.

ARCARO, C. and YUNG, W., (2001). Variance estimation in the presence of imputation, SSC Annual Meeting, Proceedings of the Survey Method Section, pp. 75–80.

CHAUVET, G., DEVILLE, J.–C. and HAZIZA, D., (2011). On Balanced Random Imputation in Surveys, Biometrika, Vol. 98, pp. 459–471.

DE WAAL, T., PANNEKOEK, J. and SCHOLTUS, S. (2011). Handbook of Statistical Data Editing and Imputation, Wiley Handbooks in Survey Methodology, John Wiley & Sons, Inc., Hoboken, New Jersey.

DUROCHER, S. and KICKPATRICK, D., (2009). The projection median of a set of points, Computational Geometry, Vol. 42, pp. 364–375.

HORTON, N. J. and LIPSITZ, S. R., (2001). Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables, Journal of the American Statistical Association, Vol. 55, pp. 244–254.

HUNDEPOOL, A., DOMINGO–FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K., DE WOLF, P.–P., (2012). Statistical Disclosure Control, Series: Wiley Series in Survey Methodology, John Wiley & Sons, Ltd.

JOLLIFFE, I. T. (2002). Principle Component Analysis. Second Edition. Springer – Verlag, New York, Berlin, Heidelberg.

KIM, K., (2000). Variance estimation under regression imputation model, Proceedings of the Survey Research Methods Section, American Statistical Association.

KIM, J. K., BRICK, M., FULLER, W. A. and KALTON, G., (2006). On the bias of the multiple-imputation variance estimator in survey sampling, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 68, pp. 509–521.

LAVORI, P. W., DAWSON, R. and SHERA, D., (1995). A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data, Statistics in Medicine, Vol. 14, pp. 1913–1925.

LITTLE, R. J. A. and RUBIN, D. B., (2002). Statistical Analysis with Missing Data. Second Edition, John Wiley & Sons, Inc., New York.

MALINA, A. and ZELIAŚ, A., (1998). On Building Taxonometric Measures on Living Conditions, Statistics in Transition, Vol. 3, No. 3, pp. 523–544.

MILASEVIC, P. and DUCHARME, G. R., (1987). Uniqueness of the Spatial Median, The Annals of Statistics, Vol. 15, No. 3, pp. 1332–1333.

MŁODAK, A., (2014). On the construction of an aggregated measure of the development of interval data, Computational Statistics, Vol. 29, pp. 895–929.

MŁODAK, A., (2006). Multilateral normalisations of diagnostic features, Statistics in Transition, vol. 7, pp. 1125–1139.

NETER, J., WASSERMAN, W. and KUTNER, M. H., (1985). Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs, 2nd edition, Homewood, IL: Richard D. Irwin, Inc., U.S.A.

PAMPAKA, M., HUTCHESON, G. and WILLIAMS, J., (2016). Handling missing data: analysis of a challenging data set using multiple imputation, International Journal of Research & Method in Education, vol. 39, No. 1, pp. 19–37.

ROUSSEEUW, P. J. and LEROY, A. M., (1987). Robust Regression and Outlier Detection, ed. by John Wiley & Sons, New York.

RUBIN, D. B., (1987). Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, New York.

SÄRNDAL, C. E. (1992). Methods for estimating the precision of survey estimates when imputation has been used, Survey Methodology, vol. 18, pp. 241–252.

SCHAFER, J. L., (1997). Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.

TIBSHIRANI, R., (1996). Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 58, No. 1, pp. 267–288.

VANDEV, D. L., (2002). Computing of Trimmed L1 – Median, Laboratory of Computer Stochastics, Institute of Mathematics, Bulgarian Academy of Sciences, (preprint), available at

http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/papers/aspap.pdf .

YUAN, Y. C., (2010). Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0), SAS Institute Inc, Rockville, MD, U.S.A.

ZELIAŚ, A., (20042). Some Notes on the Selection of Normalization of Diagnostic Variables, Statistics in Transition, vol. 5, No. 5, pp. 787–802.

APPENDIX

Table 1. Descriptive statistics of distribution of original and imputed revenues – option 1 of missing data model

Specification	Minimum (in million $)	Lower quartile (in million $)	Median (in million $)	Upper quartile (in million $)	Maximum (in million $)	Coefficient of Variation (%)	Skewness	Kurtosis
Y	623.10	2479.85	7928.40	22353.50	83221.00	130.4	1.7949	2.5318
R	-113226.99	1961.50	8270.11	28538.50	189397.87	216.9	1.2942	5.1495
RM	623.10	2528.20	8559.45	18095.47	83221.00	125.7	2.2671	5.5879
RG	-1356.38	1287.05	7928.40	25849.59	83221.00	125.9	1.9174	4.0467
RGM	-16694.32	1961.50	8559.45	26042.13	83221.00	127.4	1.6282	3.2586
RGI	-12566.21	1961.50	8559.45	28412.30	83221.00	125.0	1.6147	2.8658
RGIM	623.10	2528.20	8955.52	29907.79	83221.00	117.4	1.4907	1.5600
PMM	623.10	2528.20	7863.75	25702.00	83221.00	126.2	1.9797	4.1440
PMMM	623.10	1961.50	7863.75	28538.50	83221.00	125.7	1.8569	3.4739
PSM	623.10	2528.20	8559.45	31375.00	83221.00	113.9	1.4717	1.7915
PSMM	623.10	1128.60	6834.45	19005.00	83221.00	136.2	2.1422	4.8090

Source: Own work using the algorithm written in SAS Enterprise Guide 4.3 (with IML environment).

Table 2. Descriptive statistics of distribution of original and imputed revenues – option 2 of missing data model

Specification	Minimum (in million $)	Lower quartile (in million $)	Median (in million $)	Upper quartile (in million $)	Maximum (in million $)	Coefficient of Variation (%)	Skewness	Kurtosis
Y	623.10	2479.85	7928.40	22353.50	83221.00	130.4	1.7949	2.5318
R	623.10	4161.80	9485.78	20753.68	126578.72	134.4	2.7496	8.6040
RM	623.10	4161.80	9939.75	19445.90	83221.00	111.7	2.2160	5.5749
RG	-23310.83	2479.85	7863.75	18391.27	83221.00	147.4	1.8536	4.5729
RGM	-8927.63	2530.55	9276.10	26041.16	83221.00	120.8	1.7166	3.6362
RGI	-4920.05	1961.50	7928.40	26399.92	83221.00	131.1	1.8004	3.4871
RGIM	-28352.42	1106.80	7863.75	22353.50	83221.00	179.1	1.1556	1.8188
PMM	623.10	2479.85	7983.90	14172.45	83221.00	135.4	2.3786	5.8865
PMMM	623.10	4161.80	9939.75	28538.50	83221.00	112.3	1.7784	2.9581
PSM	623.10	1287.05	5168.95	14172.45	83221.00	145.6	2.3500	5.6783
PSMM	623.10	3244.15	8020.10	19005.00	83221.00	126.7	2.3127	5.7103

Source: Own work using the algorithm written in SAS Enterprise Guide 4.3 (with IML environment).