MISSING VALUE IMPUTATION USING CONTEMPORARY COMPUTER CAPABILITIES: AN APPLICATION TO FINANCIAL STATEMENTS DATA IN LARGE PANELS

ALEŠ GORIŠEK[1]
MARKO PAHOR[2]

Received: October 5, 2016
Accepted: December 13, 2016

ABSTRACT: *This paper addresses an evaluation of the methods for automatic item imputation to large datasets with missing data in the setting of financial data often used in economic and business settings. The paper aims to bridge the gap between purely methodological papers concerned with individual imputation techniques with their implementation algorithms and common practices of missing value treatment in social sciences and other research. Historical methods for handling the missing values are rendered obsolete with the rise of cheap computing power. Regardless of the condition of input data, various computer programs and software packages almost always return some results. Despite this fact, item imputation in scientific research should be executed only to reproduce reality, not to create a new one. In the review papers comparing different methods we usually find data on performance of algorithms on artificial datasets. However, on a simulated dataset that replicates a real-life financial database, we show, that algorithms different from the ones that perform best on purely artificial datasets, may perform better.*

Keywords: *missing data, imputation, regression imputation, EMB algorithm, panel data, big data*

JEL Classification: C18, C65

DOI: 10.15458/85451.38

INTRODUCTION

Methods and procedures concerned with missing values in scientific datasets have been well documented and described. To gain some insight into ad-hoc methods such as complete case analysis[3], available case analysis[4] and single imputation methods like hot

[1] University of Ljubljana, Faculty of Economics, PhD candidate, Ljubljana, Slovenia, e-mail: gorisek@gmail.com

[2] University of Ljubljana, Faculty of Economics, Ljubljana, Slovenia, e-mail: marko.pahor@ef.uni-lj.si

[3] Also known as the Listwise Deletion method. Only cases with complete data are used in analysis.

[4] Also known as the Pairwise Deletion method. This method tries to maximize the use of available data in each step of analysis. Even if some data points in a row are missing, but are not needed in the current step, the data that is present is used in the current step of analysis.

deck imputation[5] and mean imputation[6], one could start with Pigott (2001), Tanguma (2000) and Peugh and Enders (2004). These methods are easily implemented, but they require assumptions about the data that rarely hold in practice (Pigott, 2001). Sloppy use of aforementioned techniques can lead to biased or outright wrong results of scientific analysis. Imputation of missing values increases in complexity with the introduction of a regression model[7], stochastic regression model and multiple imputation methods, such as bootstrapped stochastic regression. More complex imputation procedures in general also yield much better imputed values. Thus, the amount of work included, pays dividends. With the wide availability of powerful computers, model based methods like Expectation Maximization (EM), and multiple imputation (MI) methods like Expectation Maximization Bootstrap (EMB)[8] and Approximate Bayesian Bootstrap (ABB) are gaining prominence (Siddique and Belin, 2008). Multiple imputation techniques use bootstrapping to calculate missing value sets with Bayesian or regression imputation (Honaker, King and Blackwell, 2011). Another group are algorithms for autoregressive spectral estimation of lost sample values in discrete-time signals, which can be described with AR and ARMA[9] models (Kazlauskas and Pupeikis, 2014). Genetic Algorithm based, Kernel based, Multi-Layer Perceptron and other Neural Networks based methods have also been evaluated (Andrew and Selamat, 2012).

In the literature, a number of studies exist that compare the effectiveness of different missing value imputation mechanisms in various settings (e.g. Olinsky, Chen and Harlow, 2003; Parwoll and Wagner, 2012; Yesilova, Kaya and Almali, 2011). In these studies authors test different mechanisms of missing data processes, they do however assume some theoretical distribution of the underlying variables, usually the normal distribution. They also assume at least missing at random[10] pattern of missing values. Although these are fair assumptions, corresponding to the standard assumptions of the widely used statistical methods, they do not correspond to empirically observed distributions in social sciences in general and in particular in financial statements data. We show that a purpose-built algorithm on a real-life dataset can outperform the state-of-the-art algorithms that work very well under the assumptions of normal distribution of variables (Allison, 2011). Procedure is tested on a dataset of approximately log-normally distributed variables and a nonrandom pattern of missing data, as the one found in financial reports databases[11].

[5] Missing value is imputed with an observed response of similar unit. Historically, the term hot deck originates from the era, when punch cards were used for computer storage. The deck of the cards that was currently being processed was »hot« (Andridge and Little, 2010).

[6] Missing value is replaced by the mean of available values.

[7] Missing values in one variable are imputed using a regression model based on other variables.

[8] EMB algorithm combines EM algorithm with a resampling procedure provided by bootstrapping. A rather mathy derivation of EM algorithm can be found in Dempster, Laird and Rubin (1977). The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

[9] AR and ARMA stand for Auto Regressive and Auto Regressive Moving Average, respectively.

[10] Missing patterns are explained in the beginning of section 2 of this paper.

[11] Special properties of yearly financial reports data are described in subchapter 3.2 of this paper below.

The aim of this paper is not to review all the data imputation techniques and list all possible methods with their assumptions. A good resource for that is Little and Rubin (2014). Missing data imputation is usually a means to an end of a broader research process. The aim of this paper is to show one possible pragmatic approach to research with data that has missing values. The content and the meaning of the data in the encompassing research project are taken into consideration. With the use of the best imputation procedure considering the properties of the data, imputed values are much closer to the true values than with simple or out of the box solutions. Despite the computational complexity of more elaborate techniques, with the right choice of imputation algorithm, much better results and considerable speed gains can be achieved. Speed gains are most notable when using parallel processing capabilities of contemporary computers and other big data technologies.

We compare the performance of the imputation procedures first on an artificially created dataset that follows the conventional normal distribution of variables on two different missing value mechanisms. Then we move to a more realistic case of a large panel dataset of financial statement data for six industries in fifteen countries in ten years from Amadeus[12] database. We use the database to extract the distribution and relations among a set of commonly used variables in economic research. Following these distributions and relations we build a simulated dataset with the same distribution and correlation properties. Finally, we proceed to simulation of different missing value mechanisms on the simulated dataset.

In chapter 2 we continue this paper with a short review of the missing value mechanisms and description of the nature of a practical problem our paper aims to solve. Chapter 3 describes the reasoning behind the derivation and provides the description of the customized two-step imputation algorithm. Chapter 4 describes the real life dataset, used as the basis for our experiments. Chapter 5 is the core of this paper presenting the comparison of performance of various imputation methods. We first check the performance of different imputation methods on the artificial, normally distributed dataset in subchapter 5.1 and then on the simulated dataset that follows the empirically observed distributions and relations in subchapter 5.2. We end the paper with conclusions and suggestion for further research in chapter 6.

1 PROBLEM DESCRIPTION

Missing values are not just blank spaces waiting to be filled with imputed data or somehow removed from the analysis. The pattern of the missing data can contain valuable information. When imputing missing values, one must be most concerned with the so-called missing data mechanism (Eekhout, 2014; Rubin and Little, 2002). Data imputation methods have different assumptions regarding missing data mechanism. If these assumptions do not match the situation with the data, the results of the imputation method may not reflect the real situation. A new reality can be created, which is wrong. Missing data mechanisms can be classified into one of the three categories:

[12] Amadeus is a database prepared by Bureau van Dijk. Amadeus contains information on around 21 million companies across Europe.

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

MCAR data are missing totally randomly. One could test for MCAR missing data mechanism using Little's test or some other procedures found in cited literature. Data following MAR pattern are missing at random, conditionally. That is, we know of some variable that influences the amount of missing values and we can control for that variable. MNAR pattern is the most troublesome of all. Missing values are related to some variable for which we cannot control. When deducing the missing value pattern, knowledge of the data and the field of research are of great help.

A typical setting in economic research is to use a panel data structure, e.g. data for a cross-section of companies for several years. If the same cross-section is present in all observed years, we talk about a balanced panel. If companies are entering and leaving the set, we have an unbalanced one. Let us assume that in the final analysis we need k interval variables X_k. These interval variables are analyzed separately for each possible combination of values in l categorical variables C_l. One of the categorical variables C_τ for which $l = \tau$ can also serve as a time series index in panel dataset.

In Amadeus dataset under consideration, observations (companies) were entering and leaving our problem space (the economy). Since we wanted to assess the influence of all available data, we opted for unbalanced panel. Choice of balanced panel would simplify the process, but the analysis would lose a lot of its power due to removal of observations that did not exist at all τ values. As an example, in Table 1 are descriptive statistics on data for four selected industries in Austria.

Table 1: Analysis of the amount of useful data - Amadeus, Austria, selected industries

Dataset	Num. of observations	Valid triplets	Complete cases
Source	217194	510427 (32,5%)	11708 (5,4%)
Imputed where possible	148768	771727 (51,9%)	22119 (14,9%)

Source: Own measurement

Valid data triplet is a tuple of sales, number_of_employees and assets for one company for one year. If any of these three data points is missing, other values are useless in our analysis. Since valid triplets are calculated per year and our dataset has data for 10 years, number of valid triplets must be divided with number of years to be compared to amount of complete cases with data for all years. With "imputed where possible" solution opting for balanced panel (using only 14,9% of all data) would leave us with just a quarter of available data (51,9%).

Another reason for using the unbalanced panel lies in the fact that we are not aware of the missing value mechanism. Choosing a balanced panel on available data could thus introduce bias into the analysis, due to removal of observations that is not random, but follows some existing but uncontrolled for pattern. To check, whether data is valid for certain observation at value τ , we used a control variable X_τ , which was complete_ year in the case of Amadeus dataset. If the data on X_τ is missing or the value of X_τ is indicating an invalid set of values for observation n at τ , then the data is not used in further data imputation process. Such subset of data is invalidated. It is prudent to assume, that observations at such singular conditions exhibit different characteristics than under ordinary circumstances, e.g. companies behave differently in years when they are entering or leaving the economy than in years of normal business activities.

Let $D_{[n,(l+k-1)]}$ be the matrix of data observations[13] . D is combined as a block matrix from matrix $C_{[n,l\setminus\{\tau\}]}$ representing the data points with categorical data and matrix $X_{[n,k]}$ representing the data points with interval data.

$$D_{[n,(l+k-1)]} = \left[C_{[n,l\setminus\{\tau\}]} X_{[n,k]} \right]$$

In our case study, data is acquired on the basis of a query to a database, which listed valid values of observed categorical variables C_l as a condition for selection. Thus, a record in the database with a missing value on the observed C_l is automatically excluded from our dataset. This is a clear case where MAR assumption has to be evaluated. MAR is the underlying assumption of many out-of-the-box data imputation algorithms, software packages and programs. If a pattern of missing observations can be suspected, data should be treated accordingly.

Up to this point we know enough about data, that we could brute force execute any out-of-the-box data imputation method listed in the introduction of this paper.

2 CUSTOMIZED MISSING DATA IMPUTATION

In this chapter we describe a custom two-step method for missing data imputation that can be used in contexts of unbalanced panel data, as the one usually found in financial statements databases. We later proceed to show that this method is superior to off-the-shelf methods implemented in contemporary software.

[13] n does not represent the number of companies, but rather $number\ of\ companies * Card(\tau)$. In our case τ contains the year of observations and $Card(\tau)$ is the number of all years. Other categorical variables $C_{l\setminus\{\tau\}}$ like country and industry are mere descriptors and do not require special attention.

3 IMPUTATION PREPARATION – MINIMIZING DURATION OF COMPUTATION

Values of X_k interval variables have different covariance matrices depending on the combinations in values of C_I. Because of this our original dataset gets partitioned into $\Pi_{i=1}^{I} Card(C_i)$ [14] independent datasets, some of which may be empty. From the viewpoint of data imputation procedure, computation of independent datasets can be solved in decoupled processes. Such problems are called embarrassingly parallelizable. This fact plays a key role in the employment of big data and other parallel capabilities of IT technology. Usage of parallel computing technology can result in substantial time savings (Wilkinson and Allen, 1999; Fox, Williams and Messina, 2014).

Selection from Amadeus dataset used in this paper contains data for ten (10) years, for six (6) industries in fifteen (15) countries. There are 217194 companies in the dataset. For observation to be valid in a particular year, data is needed for three variables.

The two-step imputation procedure presented below in this article runs a calculation of the mean value of available data for each of three needed variables for each company in the first step. In step two, 10*15*6=900 linear models are estimated. Each of these 217194 * 3 + 900 = 652482 imputation blocks are independent of each other and can be calculated in parallel.

Similar reasoning is employed for multiple imputation methods such as EMB algorithm, also used for comparison following in this article. Well-programmed software packages use the independent partitions in the data, if provided as function call parameters to parallelize the computations.

4 SETTING THE STAGE FOR CUSTOM TWO-STEP IMPUTATION METHOD

With the analysis of the structure and relationships in datasets, taking into account the subject matter of the broader research topic, tailor made data imputation procedures can be prepared. Using the knowledge about the structure of the Amadeus dataset and the expected properties of data on yearly financial reports, following is a derivation of such procedure.

Long dataset where each row contains data for only one observation for one point in time is rewritten to a wide-panel-type of block matrix W. A group of observations where all values of C_I are equal, the only varying categorical column being C_T is rewritten to a wide form as:

$$W_{\left[\frac{n}{\tau},(I-1+k*\tau)\right]} = \left[C_{\left[\frac{n}{\tau},I-1\right]} X_{\left[\frac{n}{\tau},k*\tau\right]} \right]$$

[14] $\Pi_{i=1}^{I} Card(C_i)$ is a product of number of elements of all categorical variables.

Each set of values $X_{k,min(\tau)} \cdots X_{k,max(\tau)}$ represents a time series. In the data with imputed values, we want the relationship between variables X_k to stay unbiased. With the use of regression imputation or various multiple regression imputation techniques, we may increase the correlation between X_k variables, thus introducing bias to our research findings. In our example, we want the relationships between data on sales, number_of_employees and assets to remain clean, i.e. imputation of missing values should not make these variables appear more correlated to each other than they are in reality. Even companies from the same industry are organized differently and create value using different mix of resources. That means that even naïve use of Bayesian imputation methods can give us bad results.

Financial statements data of companies are submitted with a well-defined frequency, once a year in our case. The frequency of data sampling is low and transcends seasonal anomalies. The cycles of strong changes in national economic conditions span several decades. It is easy to extract short term trends from the data. In a decade, a zig-zag curve of rapid swings on any of variables from the set X_k for any company is not likely.

The profitability of individual company is in large part dependent on its own, business specific effects (McGahan and Porter, 1997). Thus, we can assume, that existing data about the company is carrying more information about its own missing values, than the data about the rest of the industry in a certain country in a certain year, that we have for other companies. Under such assumptions, mean value imputation is a viable method.

Where for any point in time no data about a variable exists for a company, there is no basis for mean imputation from company's own data. In such case, data about the rest of the industry in a certain country in a certain year can be used in combination with existing data about the company. If such data is present, regression can be used to impute missing data.

As an example of good, context dependent missing value imputation method, below described two-stage procedure is used.

5 CONTEXT DEPENDENT TWO-STEP IMPUTATION METHOD DESCRIPTION

Step 1

If there is enough data present for any partition $C_{\Lambda\{\tau\}}$ in any of the time series from X_k , it makes sense to impute the missing values from this data. Since correlation among time series X_k for individual company is not important in our research, we opt for a simple mean imputation[15]. For all data points where `complete_year` variable is valid, the potential missing value is predicted from neighboring two cells. If no valid values are

[15] We are not interested in correlation between time series within one company. That is why attenuation of correlation between variables, which is a consequence of mean value imputation, is not problematic in our case.

available on one side of the time series, a trend deduced from former/latter data points is used. At least two valid data points are needed for such imputation to take place. If there is no data for certain observation in a particular time series, or if there is only one data point, regression imputation described in step 2 is used.

Step 2

From data in the source sample[16], based on our domain specific knowledge, we try to find a variable or combination of variables X_k as regressors in linear model for regression estimation of missing values for particular $X_p \subset X_k$. Financial statement data provide us with several variables X_k that are a superset of X_k. Thus, some are not included in the research model, i.e. are not in the set X_k. These variables are more or less correlated with the variables in the set X_k and can be used as regressors, i.e. inputs into the regression imputation procedure.

$$X_p \subset X_k$$

$$X_p = \overrightarrow{X_r}\vec{\beta} + \vec{\varepsilon}$$

We aim to keep the relationships between variables X_k that are of interest in our final research to be as similar to the true relationships as possible. Using subset of X_k as predictors X_r for one of $X_{j \in k}$ would result in increased correlation between the variables X_k. It is thus desirable that:

$$X_r \cap X_k = \varnothing$$

It is possible, that the linear model from equation $X_p = \vec{X}_r\vec{\beta} + \vec{\varepsilon}$, obtained from regression analysis has insignificant p-values for any β or insignificant F-statistic. Such cases can happen, if there are not enough observations with valid data to successfully estimate a model, if there are nonlinear properties in the data, etc. It is thus necessary, to check for non-significance of coefficients or linear model as a whole and prevent imputation of values for X_p, computed from unreliable regression coefficients[17].

Final data assembly

If a value is present in the original dataset, that value is used. If it is possible to impute the missing values from each observation's own data, mean imputation is used. As a last

[16] Another option would be to use the dataset, obtained after execution of imputation in step 1.

[17] In our case, exploratory data shows that estimating number_of_employees from costs_of_employees yields strange results if companies with less than 10 employees are taken into account. Since the focus of our broader research problem is on companies with more than 50 employees, we are able to discard observations with number_of_employees value being less than 10. Still, there are combinations of year, industry, country, where no reliable regression model could be estimated.

resort, domain adjusted regression imputation is used, if the obtained linear model has statistically significant coefficients and F-statistic. If none of these options provides a value, data point is left empty (missing value is kept) and is accounted for in subsequent analysis.

6 DATA

The simulations are conducted on two different datasets that are labeled artificial dataset and simulated dataset. Artificial dataset refers to a randomly created dataset where data follow multivariate normal distribution, created purely for testing the results of imputation procedures, accounting for their possible assumptions. This dataset assumes only one period cross-section and simple correlation among variables. Simulated dataset is an artificially created dataset. Distribution and trends in individual time series follow the empirically observed ones found in Amadeus real dataset of financial statements. The missing value mechanism is controlled within the simulation for both datasets.

The missing data mechanism in the observed real financial dataset is unknown; we do know that it is not MCAR due to several reasons. E.g., when observing the percentage of missing values in individual years, more data is missing in earlier years of observations. Thus, data is MAR at best. If missing values are in any way correlated with a value of some variable (e.g.: smaller companies are less likely to report some datum), data is MNAR. If data is MNAR, it violates the basic assumption of some out-of-the-box missing value imputation techniques.

Data about companies (observations) in Amadeus dataset consist of a set of categorical variables C and a set of interval variables X. From Amadeus dataset with financial statements, let us choose set C to consist of country of origin, industry in terms of NACE rev. 2 classification[18], year and complete-year. The element complete-year is telling us, whether the data for a certain company represent the whole year or maybe just some fraction of it. Financial statements for individual companies consist of several tens of more or less correlated data points[19]. For brevity, let us only focus on sales, number_of_employees, costs_of_employees and assets, which are represented in a set of interval variables X.

[18] Statistical Classification of Economic Activities in the European Community, Rev. 2 (2008)

[19] Before analyzing the empirical data for distribution and relations the data was treated to ensure consistent representation of decimals, missing value identifiers, etc. Data treatment methods are not the focus of this article. Interested readers might want to refer to any introductory text on data analysis. Another important issue in the data preparation process is the decision on detection and treatment of outliers. Readers interested in this topic may refer to Aggarwal (2013) or any other text about outlier analysis. Ignoring or mistreating of outliers can have strong influence on data imputation accuracy (Quintano, Castellano and Rocca, 2010).

Table 2: Correlations between variables from over 3.5M observations of Amadeus data

	Assets	Num. of employees	Costs of employees	Sales
Assets	1.00000	0.95954	0.96216	0.95451
Num. of employees	0.95954	1.00000	0.90290	0.89875
Costs of employees	0.96216	0.90290	1.00000	0.97925
Sales	0.95451	0.89875	0.97925	1.00000

Source: Own measurement

Empirical properties of the real-life dataset

Since we are dealing with panel data, we almost always find clear trends observing particular variable for particular observed subject through time. Variables are also quite strongly correlated. Large companies are in general larger than small companies as measured in all variables: number of employees, costs of employees, assets and sales. Correlations vary depending on industry, country and year. Correlations between variables in the original dataset were calculated using pairwise complete observations[20] approach, to keep as much information about original data as possible. Due to assumptions on which imputation methods are based on, knowing the nature of the dataset is of utmost importance when choosing missing value imputation method.

Due to vast differences in sizes of the companies and the fact that no company has less than zero employees, the distribution of variables is not normal. It is a standard procedure to log the variables, assuming they are log-normally distributed. We find that three variables: sales, costs_of_employees and assets can be approximated by log-normal distribution. It is obvious from the Figure 1 that this assumption is not mathematically exact, but can be applied for the sake of brevity.

7 IMPUTATION METHODS ANALYSIS

We want to guarantee reproducible results, which are not dependent on particular dataset. Thus, we need the capability to control parameters of data and be able to create several different datasets with the same set of parameters. First, we execute a simplified experiment. We create two normally distributed variables, introduce correlation and apply various missing data patterns and imputation techniques. To be able to control the parameters of data distributions, remove noise and control the missing values mechanism, we prepare a simulation procedure, to create a simulated dataset.

[20] Available case analysis

8 ARTIFICIAL DATA, TWO VARIABLES, CORRELATION = 0.7

We simulate a series of datasets with two normally distributed random variables, each consisting of 10000 observations and correlation between variables set to 0.7. The simulated datasets are created using random number generator and Cholesky root of desired covariance matrix. On average, the measured correlation in the artificial datasets is 0.699 with a standard deviation of 0.007.

Missing pattern: MCAR

The algorithm is set to randomly remove approximately 20% of data points. After removal some of the cases are missing one and some both variables. The procedure leaves on average 6691.7 complete cases in the dataset, with a standard deviation of just above 18 cases. The average measured correlation of complete cases in the MCAR corrupt data set is 0.700 (s.d. 0.009). On average, the MCAR missing data process does not induce bias in the data, although we do observe an increased variability, probably due to smaller datasets.

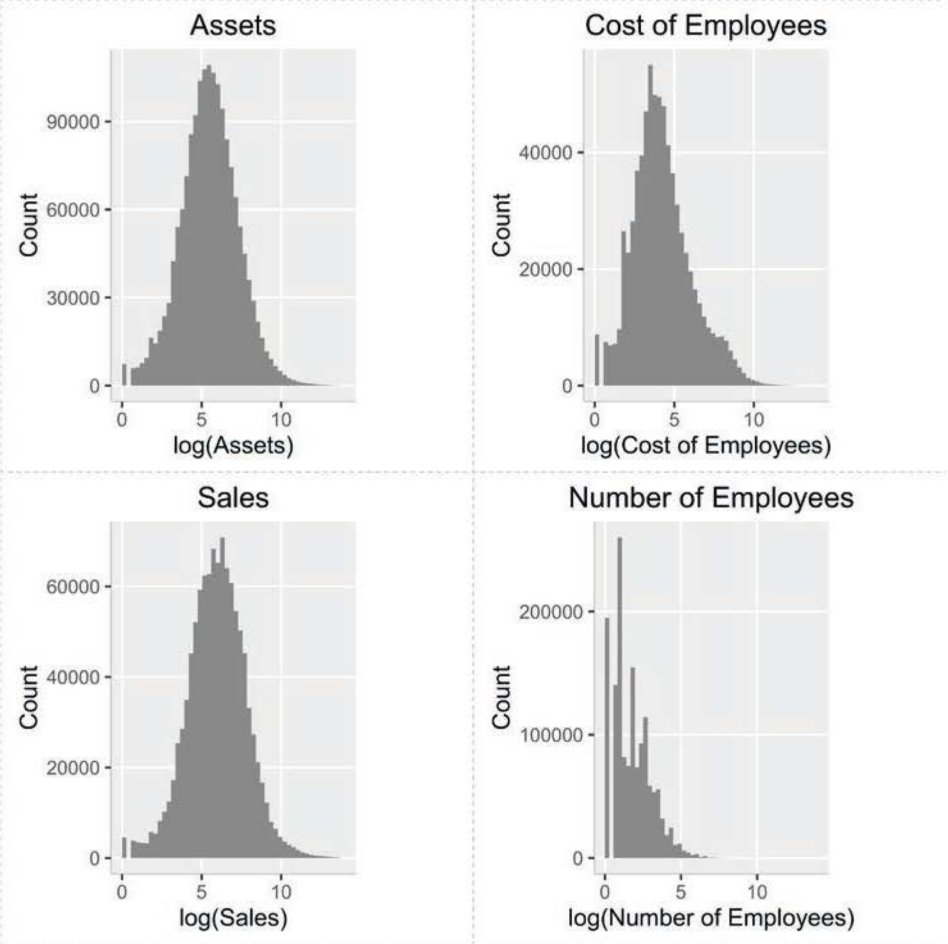
Table 3: Results: MCAR missing pattern, two normally distributed variables

Imputation method	mean(Corr.)	mean(Corr. diff.)	sd(Corr. diff.)	mean(% miss. left)
Mean	0.520	0.179	0.009	0.000
Regression	0.760	- 0.061	0.003	0.032
Amelia (EMB)	0.747	- 0.049	0.003	0.032

Source: Own measurement

The table is showing mean measured correlation between two variables after using each imputation procedure on MCAR pattern in second column. In column showing "mean(Corr. diff.)", the difference between initial correlation (0.7) and measured correlation is shown. Fourth column reports the standard deviation of measured correlation after imputation across several runs of experiment. Last column reports the percentage of values that are left missing.

Figure 1: Distribution of variables in the observed Amadeus dataset

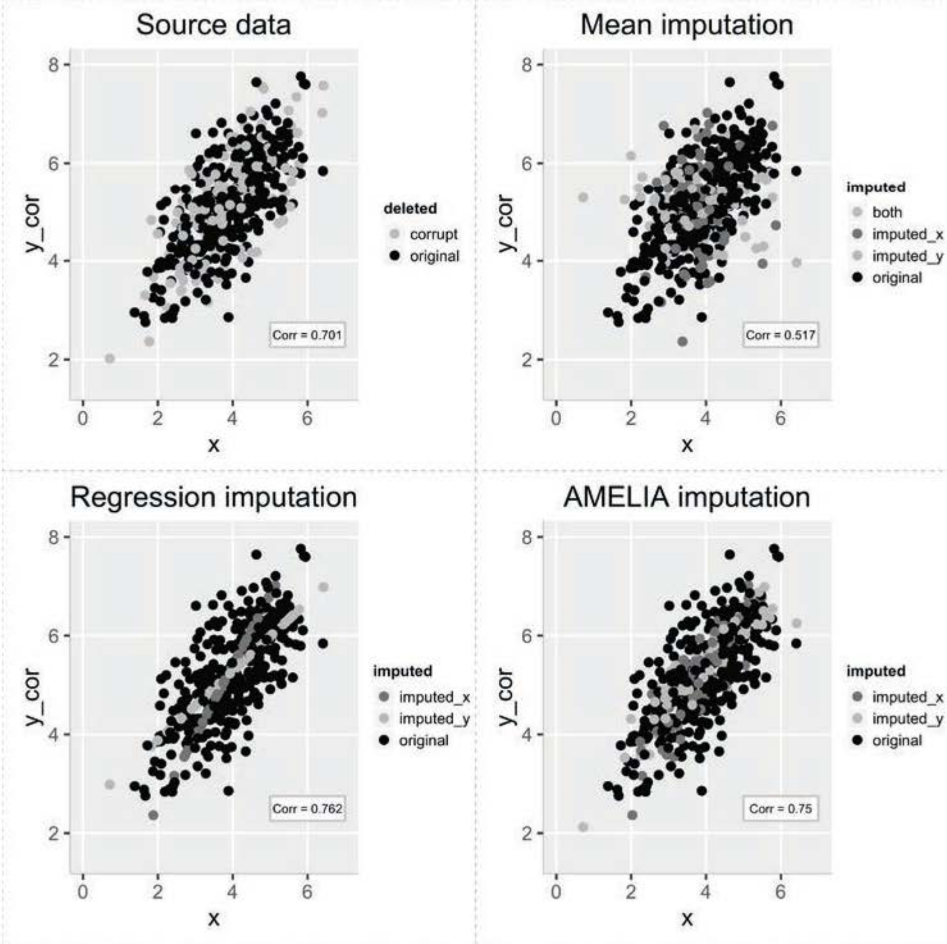


Source: Own measurements and visualization

For the sake of brevity, we can assume that assets, cost of employees and sales are log-normaly distributed. From the picture it is obvious, that this is not exactly true. However, number_of_employees evades the efforts to be molded into log-normal using the same number of bins as for other observed variables. Many companies have very small number of employees and the log function applied to discrete small natural numbers starting with 1 returns values 0, 0.69, 1.10, 1.39, 1.61, etc. Frequencies of these low numbers are high relative to numbers in other observed variables. With low number of bins in a histogram, cumulative distribution function starts to resemble a cumulative distribution function of Binomial distribution, but further analysis of this phenomenon exceeds the scope of this text. To further complicate the matters, companies are reporting rounded numbers. Other mechanisms influencing the distributions may exist, e.g. Amadeus may not include data on all companies from one country, but a certain sample, which may introduce selection bias.

The results of the simulations are presented in Table 3. As expected, mean imputation attenuates the correlation. Both regression imputation and EMB method used in AMELIA II increase the correlation. EMB imputation is showing slightly less biased results, since its initial assumptions are satisfied. Visual representation of results is in Figure 2.

Figure 2: MCAR missing pattern, two normally distributed variables



Source: Own measurements and visualization

The top right picture is showing how mean imputation spreads the imputed points and attenuates the correlation. Increase of correlation as a consequence of regression imputation clearly seen on bottom left picture. The picture on bottom right is produced by EMB algorithm and is less clear, but measured correlation is increased.

Missing pattern: MNAR

Setting the data to simulate the MNAR missing data process is just slightly more complicated. Data points should be missing according to some pattern in the data itself, such that we cannot control for that with another variable. In our case, there is a probability 0.7 for a data point to get corrupted if it matches a condition and zero probability otherwise. The condition is matched if the value in the first column in the row is in bottom 4 deciles of the first column's values. After this procedure we are left with an average of 7125.5 complete cases and a standard deviation of 816.7 cases. Measured correlation of complete cases in the MNAR corrupt data set is on average 0.642 with a standard deviation of 0.011. From the results depicted in Figure 3 we can see that a MNAR process like the one we simulate can introduce some bias to the imputed values, making the correlation between variables presented in Table 4 somewhat weaker.

Table 4: *Results: MNAR missing pattern, two normally distributed variables*

Imputation method	mean(Corr.)	mean(Corr. diff.)	sd(Corr. diff.)	mean(% miss. left)
Mean	0.486	0.212	0.024	0.000
Regression	0.713	- 0.015	0.015	0.096
Amelia (EMB)	0.723	- 0.025	0.018	0.096

Source: Own measurement

The table is showing mean measured correlation between two variables after using each imputation procedure on MNAR pattern in second column. In column showing "mean(Corr. diff.)", the difference between initial correlation (0.7) and measured correlation is shown. Fourth column reports the standard deviation of measured correlation after imputation across several runs of experiment. Last column reports the percentage of values that are left missing.

Again, mean imputation further attenuates the correlation. As in the MCAR case, both the regression imputation and the EMB method used in AMELIA II software increase the correlation. However, in the MNAR case, the regression imputation yields slightly better results than EMB method. We can explain this difference with EMB algorithm's assumption that the missing data pattern is MCAR. This assumption is violated by design of the experiment.

Artificial data, conclusion

Despite the fact, that mean imputation leaves no missing values in the final dataset, significant drop in correlation between the variables can be observed. Both regression and EMB imputation methods yield similar results with the correlation between variables only slightly off target. When the assumptions underlying the EMB method are met, this method proved superior. On the other hand, regression method is proven to be more robust to violations of the MCAR assumption.

9 SIMULATED DATA

Creating simulated dataset from parameters

To simulate correlated random variables resembling real Amadeus dataset given a correlation matrix, we could use the following procedure:

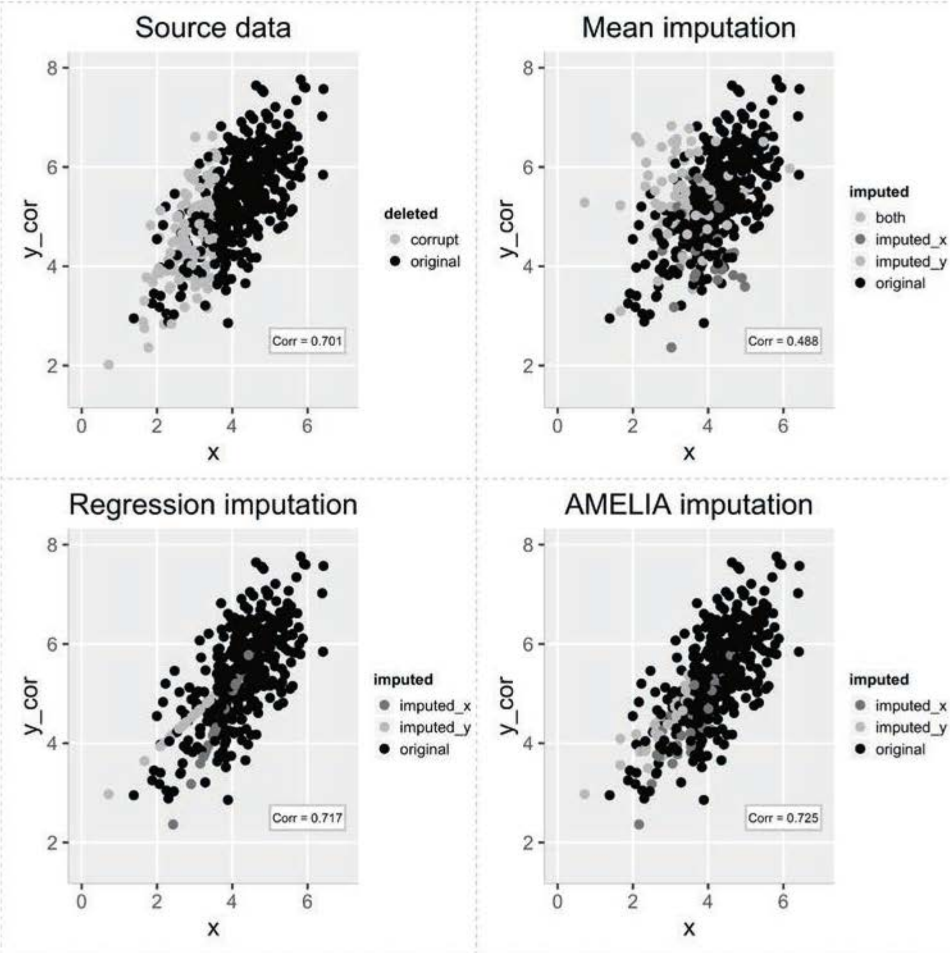
- Calculate Cholesky decomposition of correlation matrix, obtained from Amadeus data for particular year, industry and country
- Generate an $n * k$ matrix of standard normals, Z
- Calculate $X = LZ$ to get correlated normals
- Multiply the columns by $\sigma_i$and add μ_i to get correlated nonstandard normal

In the above procedure, n represents the number of observations we want to create. k represents number of variables, X is the final simulated dataset, L is the left Cholesky factor of the decomposition and Z is an individual variable with standard normal distribution. $\sigma_i$and μ_i are the parameters of target normal distribution of each variable $i \in \{1 \ldots k\}$. This procedure was used to introduce the correlation between the variables in artificial dataset in subchapter 5.1.

Such procedure cannot reproduce trends that are present in original financial statements data. We opted for a less elegant but simpler algorithm, that produces the data retaining the gist of the phenomenon, i.e. somewhat correlated groups of variables with trends:

- Estimate parameters of log-normal distribution of number_of_employees as D_e
- Estimate parameters of log-normal distribution of assets as D_a
- Randomly choose a trend t_e for number_of_employees from uniform distribution, chosen to lie between 0 and 1.5
- Randomly choose a trend $t_a$for assets from uniform distribution, chosen to lie between 0 and 1.2
- Create a random number $rand_{emp}$ from log-normal distribution with parameters from estimated D_e
- Create a vector of number_of_employees values for one row using $rand_{emp}$ and t_e, number of elements represents the number of years
- Create a random number $rand_{as}$ from log-normal distribution with parameters D_a
- Create a vector of assets values for one row using $rand_{as}$ and t_a
- Correlate assets to number_of_employees
- Find by how much does number_of_employees deviate from sample mean
- Apply the attenuated deviation to assets, we can choose attenuation as parameter
- Create sales which is in linear relationship with number_of_employees and assets, linear coefficients can be chosen as parameters

Figure 3: MNAR missing pattern, two normally distributed variables



Source: Own measurements and visualization

The top left picture shows MNAR missing pattern in the data. The top right picture is showing how mean imputation spreads the imputed points and attenuates the correlation. Increase of correlation as a consequence of regression imputation can be observed seen on bottom left picture. The picture on bottom right is produced by EMB algorithm and is again less clear. However, under MNAR missing pattern its results are more off in comparison to regression imputation as in Figure 2, which depicts imputation with MCAR missing pattern.

- Create costs_of_employees vector that is in linear relationship with number_of_employees, linear coefficient can be chosen
- Introduce some noise, parameters and distribution of noise can be controlled
- Repeat steps from third bullet onwards for as many times as there are rows in the

simulated data set you are creating

Such procedure gives us total control over parameters of the data. With controlled application of missing values using MCAR, MAR and MNAR patterns, we can measure the success rates of imputation methods, depending on all the parameters, with reproducible results.

Simulated data - MAR missing data pattern

Using a real data controlled simulation procedure described in subchapter 5.2.1, we create a series of datasets with 1000 observations of 4 variables in 10 time periods each. To simulate MAR missing pattern, we choose to delete 20% of points in all rows, where first column has a value greater than five. First column is left untouched, so imputation methods can use it. Such criterion results in an average of 2909.1 (s.d. 17.35) deleted data points and 589.2 (s.d. 17.35) complete cases left out of 1000 in initial simulated dataset. First, we would like to know, how closely do the imputed results resemble the ones that were deleted using the missing data process. We thus develop a simple metric to measure the difference between the original and the imputed data that takes account of both: the share of imputed values as well as the quality of imputation. As the metric, we use the sum of differences between the imputed value and the original (deleted value). Results of the simulations are presented in Table 5.

Table 5: Simulated data – MAR missing pattern

	mean(% miss. filled)	mean(% miss. left)	mean(Σ Abs(residuals))	sd(Σ Abs(residuals))
Mean	99.0	1.0	59525	17507
Regression	59.0	41.0	118933	80617
Two step	99.6	0.4	61238	16757
Amelia (EMB)	99.5	0.5	20497329	3338778

Source: Own measurements and calculations

The table is showing in the second column the percentage of data points still missing after each imputation procedure when data has MAR missing pattern. The third column is showing mean sum of absolute residuals (differences between real and imputed value) and the fourth column the standard deviation of absolute residuals after imputation across several runs of experiment.

From the Table 5 we can see that in terms of the share of imputed data, the regression method performs the worst. On average it is only able to replace less than 60 percent of missing data. Mean value imputation replaces 99 percent of missing data. Our two-step method and the EMB method both replace approximately 99.5 percent of missing data. In terms of the quality of imputation, mean imputation and two-step approach yield similarly good results. The two-step method is slightly worse but more consistent. Regression imputation is somewhat worse and much less consistent. EMB imputation proved to be completely inappropriate for this kind of data, as its imputed values deviate greatly from the deleted originals.

Simulated data - MNAR missing data pattern

Again, using a real data controlled simulation procedure described in subchapter 5.2.1, we create a dataset with 1000 observations of 4 variables in 10 time periods. To simulate MNAR missing pattern, we choose to delete 20% of points in all rows, where 23rd column has value greater than some quantile of itself. All columns are corrupt with missing values, so imputation methods are unable to find any pattern in missing value mechanism. Such criterion results in an average of 2975.7 (s.d. 117.9) deleted data points and 589.3 (s.d. 17.49) complete cases left out of 1000 in initial simulated dataset. Results are given in Table 6.

Table 6: Simulated data – MNAR missing pattern

	mean(% miss. filled)	mean(% miss. left)	mean(Σ Abs(residuals))	sd(Σ Abs(residuals))
Mean	98.6	1.4	66359	19999
Regression	59.0	41.0	105235	53220
Two step	99.4	0.6	68005	20471
Amelia (EMB)	99.6	0.4	20504011	3370479

Source: Own measurements and calculations

The table is showing in the second column the percentage of data points still missing after each imputation procedure when data has MNAR missing pattern. The third column is showing mean sum of absolute residuals (differences between real and imputed value) and the fourth column the standard deviation of absolute residuals after imputation across several runs of experiment.

From the Table 6 we can see that in terms of the share of imputed data, once more, the regression method performs worst. On average it is only able to replace less than 60 percent of missing data. Mean value imputation replaces 98.6 percent of missing data. Two-step method replaces 99.4 percent of missing data and the EMB performs best replacing on average 99.6 percent of missing data. In terms of the quality of imputation, mean imputation and two-step approach yield similarly good results. The two-step method is slightly worse but more consistent. EMB imputation manages to impute values to most data points. However, as in the MAR case, the EMB imputation performs worst in terms of imputation quality. Its sum of errors is several orders of magnitude higher than the next best method. Once again, in terms of deviation from true values, mean imputation and two-step imputation perform similarly well. The regression method lags behind both, but beats EMB imputation.

We have shown that in terms of getting missing data close to the "originals", both mean imputation and two-step procedure perform well, regardless of the missing data pattern. However, getting values on average close to the original ones is not yet indicative of whether there will be any bias in the relationships between the variables. As we have seen in the simple simulation in the subchapter 5.1, mean imputation is prone to introducing bias, consistently undershooting the original correlation. Thus, we continue the testing by checking the consistency of a common economics relation, namely a Cobb-Douglas type production function after imputation.

Estimating Cobb-Douglas type production function against imputed data

To test the effects of an imputation method on a well-known estimation problem, we estimate the α , β and A[21] of a Cobb-Douglas type production function.

$$Y = A * L^{\alpha} * K^{\beta}$$

For consistency with real-life datasets the observations in the simulated dataset are allowed to have a value zero. That makes the estimation using least squares regression on logged values impossible. We use an upgraded model that allows for the production function to be consistently estimated even with some values being zero (Battese, 1997):

$$\log(Y) = A + \alpha * \log(L) + \beta * \log(K) + \\ \kappa_1 * Y_0 + \kappa_2 * L_0 + \kappa_3 * K_0$$

$Y0$, $L0$ and $K0$ are dummy variables representing the cases, when Y, L or K have value zero. With such augmentation of the estimated model, we get unbiased results for the three coefficients we are looking for: A, α and β . Obtained values for the estimation on the MAR data are shown in Table 7 and for the MNAR in Table 8.

In our simulation, mean imputation and two-step imputation give the best results in both cases: MAR and MNAR. In both scenarios mean imputation outperforms the two- step procedure in the accuracy of the estimation of regression coefficient. Mean imputation performs somewhat worse in the estimation of the intercept. Complete cases estimation returns estimates that are relatively consistent with non-missing estimation in the slopes but greatly misses the mark for the intercept. Results of the regression imputation and the EMB algorithm are completely biased and as such useless.

[21] A represents total factor productivity

Table 7: *Estimated values of Cobb-Douglas production function: MAR*

Data set		A	α	β	$\lvert A-A'\rvert$	$\lvert\alpha-\alpha'\rvert$	$\lvert\beta-\beta'\rvert$
Simulated set	mean	0.222	0.667	0.581	0.000	0.000	0.000
	(sd)	(0.040)	(0.018)	(0.018)	(0.000)	(0.000)	(0.000)
Complete cases	mean	-0.622	0.696	0.618	-0.843	0.030	0.037
	(sd)	(0.077)	(0.016)	(0.023)	(0.097)	(0.007)	(0.010)
Mean imp.	mean	0.223	0.663	0.585	0.002	-0.003	0.003
	(sd)	(0.041)	(0.017)	(0.018)	(0.015)	(0.002)	(0.002)
Regression imp.	mean	-0.487	-0.049	0.099	-0.708	-0.715	-0.482
	(sd)	(0.584)	(0.032)	(0.037)	(0.612)	(0.032)	(0.039)
Two-step imp.	mean	0.222	0.663	0.584	0.000	-0.004	0.003
	(sd)	(0.041)	(0.017)	(0.018)	(0.015)	(0.002)	(0.002)
AMELIA imp.	mean	-0.487	-0.049	0.099	-0.708	-0.715	-0.482
	(sd)	(0.584)	(0.032)	(0.037)	(0.612)	(0.032)	(0.039)

Source: Own measurements and calculations

The table is showing the effects of the choice of missing values imputation method on estimated Cobb-Douglas productivity function coefficients. Pattern of missing values is MAR.

Table 8: *Estimated values of Cobb-Douglas production function: MNAR*

Data set		A	α	β	$\lvert A-A'\rvert$	$\lvert\alpha-\alpha'\rvert$	$\lvert\beta-\beta'\rvert$
Simulated set	mean	0.222	0.667	0.581	0.000	0.000	0.000
	(sd)	(0.040)	(0.018)	(0.018)	(0.000)	(0.000)	(0.000)
Complete cases	mean	-0.622	0.696	0.618	-0.844	0.030	0.037
	(sd)	(0.077)	(0.016)	(0.023)	(0.096)	(0.007)	(0.010)
Mean imp.	mean	0.216	0.665	0.583	-0.006	-0.001	0.001
	(sd)	(0.048)	(0.017)	(0.018)	(0.018)	(0.002)	(0.004)
Regression imp.	mean	-0.621	-0.036	0.084	-0.843	-0.703	-0.497
	(sd)	(0.702)	(0.045)	(0.043)	(0.711)	(0.042)	(0.049)
Two-step imp.	mean	0.217	0.665	0.583	-0.004	-0.002	0.002
	(sd)	(0.051)	(0.017)	(0.019)	(0.019)	(0.002)	(0.004)
AMELIA imp.	mean	-0.621	-0.036	0.084	-0.843	-0.703	-0.497
	(sd)	(0.702)	(0.045)	(0.043)	(0.711)	(0.042)	(0.049)

Source: Own measurements and calculations

The table is showing the effects of the choice of missing values imputation method on estimated Cobb-Douglas productivity function coefficients. Pattern of missing values is MNAR.

Discussion of the results for simulated data

As expected the situation with simulated data is more complex than with the clean artificial dataset. While the off-the-shelf EMB procedure performs quite well in the artificial, normally distributed case, it completely misses the mark for a dataset simulated to resemble the real-life financial reports data. Caution is thus required in the use of such procedures on real life data. The same caution should be applied to some other model-based imputation methods, one of them being the regression imputation that is also presented in this paper.

Simple approaches as complete-case approach introduce considerable bias in the estimates. However, simple mean substitution performs surprisingly well on individual variables from financial reports data. It is beating all other methods in the consistency of model estimations, save for our proposed two-step method. The tailor made two-step method comes close to and partially beats the mean imputation. The main advantage of our proposed method is in the fact that it is able to more than halve the share of non-imputed missing cases on average. This is an achievement comparable to the EMB, but without sacrificing too much of the consistency of results.

10 CONCLUSION AND SUGGESTIONS FOR FURTHER RESEARCH

From the results we can see that the described two-step imputation method yields better results than brute force use of available off-the-shelf algorithms. Assumption that data is missing completely at random or less strict assumption that data is missing at random is often wrong. The brute force use of existing data imputation algorithms can lead to invalid research conclusions.

In order to develop a good data imputation method, suited for particular data and research problem, profound knowledge of the dataset and research topic is of utmost importance. It makes sense to spend time assessing the expedience of different data imputation methods for the problem at hand. We may encounter some sort of consistency vs. efficiency tradeoff, as is the case with the two-step method proposed in this paper or as noted by Kmenta (1997).

The two-step method presented in this article is a tailor made missing value imputation procedure, suited for imputation of missing values into periodic financial reports. The method is far superior to naïve methods with regard to the amount of missing data points restored, while sacrificing small amount of consistency.

An idea for further research is a possible improvement of the two-step method with the use of some multiple imputation method instead of regression in second step. With such a measure, it would be possible to add another bit of stochastic properties to the procedure and perhaps attenuate the already small loss of consistency or further improve the rate of recovered values.

REFERENCES

Aggarwal, C. C. (2013). *Outlier Analysis.* New York: Springer.

Allison, P. D. (2011). Multiple Imputation for Missing Data: A Cautionary Tale. In *Quantitative Research Methods* (pp. 259–288). London: Sage.

Andrew, B. & Selamat, A. (2012). Systematic Literature Review of Missing Data Imputation Techniques for Effort Prediction. In *International Proceedings of Computer Science & Information Tech* (pp. 222–226).

Andridge, R. R. & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review, 78*(1), 40-64.

Battese, G. E. (1997). A note on the estimation of Cobb-Douglas production functions when some explanatory variables have zero values. *Journal of Agricultural Economics, 48*, 250–252.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.

Eekhout, I. (2014). *Don't Miss Out! Incomplete data can contain valuable information*. PhD Dissertation.

Fox, G. C., Williams, R. D. & Messina, G. C. (2014). *Parallel Computing Works!* Morgan Kaufmann.

Honaker, J., King, G. & Blackwell, M. (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1–47.

Kazlauskas, K. & Pupeikis, R. (2014). Missing Data Restoration Algorithm. *Informatica, 25*(2), 209–220.

Kmenta J. (1997). *Elements of econometrics.* Michigan: University Press.

Little R. J. A., Rubin D. B. (2014). *Statistical Analysis with Missing Data.* London: John Wiley & Sons.

McGahan, A.M. & Porter, M.E. (1997). How much does Industry matter, really? *Strategic Management Journal, 18*(Sum.Spec. Iss.), 15–30.

Olinsky, A., Chen, S. & Harlow, L. (2003). The comparative efficacy of imputations methods for missing data in structural equation modeling. *European Journal of Operational Research, 151*(1), 53–79.

Parwoll, M. & Wagner, R. (2012). The impact of missing values on PLS model fitting. In W. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, J. Kunze (EDs.) *Challenges at the interface of data analysis, computer science, and optimization* (pp. 537–544). Berlin: Springer.

Peugh, J. L. & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research, 74*(4), 525–556.

Pigott, T. D. (2001). A Review of Methods for Missing Data. *Educational Research and Evaluation, 7(*4), 353–383.

Quintano, C., Castellano, R. & Rocca, A. (2010). Influence of Outliers on Some Multiple Imputation Methods. *Metodološki zvezki, 7*(1), 1–16.

Rubin, D. B. & Little, R. J. (2002). *Statistical analysis with missing data.* Hoboken, NJ: J Wiley & Sons.

Siddique, J. & Belin, T. R. (2008). Using an Approximate Bayesian Bootstrap to Multiply Impute Nonignorable Missing Data. *Computational Statistics & Data Analysis, 53*(2), 405–415.

Tanguma, J. (2000). *A Review of the Literature on Missing Data.* Paper presented at the Annual Meeting of the Mid-South Educational Research Association, November, 2000.

Wilkinson, B. & Allen, M. (1999). *Parallel Programming, Techniques and Applications using Networked Workstations and Parallel Computers.* Upper Saddle River: Prentice Hall.

Yesilova, A., Kaya, Y. & Almali, M. N. (2011). A comparison of hot deck imputation and substitution methods in the estimation of missing data. *Gazi University Journal of Science,24*(1), 69–75.