University of Tartu

Faculty of Mathematics and Computer Science

Institute of Mathematical Statistics

Cliona Georgia Dalberg

Imputation of inventories in Estonian Commercial Register

Bachelor thesis

Supervisors Mare Vähi,

Ebu Tamm

Tartu 2013

Table of contents

Introduction

Missing values is a problem which often troubles statisticians, because most of the analysis methods consider full data. Almost every dataset has unobserved values due to the unconsciousness of the respondents, technical errors and several other reasons. One way to deal with the missingness is trying to replace missing values – imputing. Imputing is a cost-effective measure, it allows to use data which otherwise would be discarded. Imputing also minimizes bias and makes using rectangular dataset and complete data analysis possible (Longford 2005, p. 38; Scheffer 2002, p. 156).

This bachelor thesis was written as a part in the project "Integrating annual bookkeeping reports into statistical production system", which was requested from Statistics Estonia by Eurostat. Data were collected from annual reports of Estonian Commercial Register. Main goal was to complete the section of inventories in the dataset of 2011.

The first part of the thesis concentrates on giving overview about missing patterns and applications and theory of selected methods. In the second part a simulation is carried out, the dataset is described and arranged and previously specified methods are tested. Code of the program is added to appendix.

Thesis is written in Microsoft Word 2007 and imputing is done is SAS Enterprise Guide and SAS 9.2 (using IVEware).

Author of this thesis wants to thank Statistics Estonia for offering opportunity to participate in this project and allowing to use data from Commercial Register.

1. The Missingness Mechanisms

Following overview is based on (Scheffer 2002, pp. 153-154; Longford 2005, pp. 28-46).

Most commonly there are three missingness mechanisms distinguishable: Missing Completely at Random (**MCAR**), Missing at Random (**MAR**) and Not Missing at Random (**NMAR**).

For describing the missingness more properly, a response indicator R is defined (R can also be noted as a nonresponse indicator). Missing values are indicated by 0 and recorded items are indicated by 1. X^* denotes the complete data, X the recorded part and X_{mis} the missing part.

MCAR refers to data where the missingness is completely random and does not depend on the actual value of the missing data nor any other variable. Conditional distribution of the response indicator R given the completely observed data X^* coincides with distribution of R.

$$(R|X^*) \sim (R).$$

MAR indicates that missing value of variable Y is affected by some other conditional variable's X value. Conditional distribution of the response indicator given the complete data X^* coincides with conditional distribution of R given the recorded data X, so that the missing data does not contain any information about R.

$$(R|X^*) \sim (R|X).$$

NMAR refers to data where the missing is caused by the actual value of variable itself. The response indicator depends on the missing data.

There are several methods to use for dealing with missing data. Most widespread are:

1. Case deletion – incomplete records are discarded:
    a. Listwise – if a subject is missing values on any of the variables, it is excluded completely (Williams 2012, pp. 2-4);
    b. Pairwise – each pair of variables is watched separately, if subject is missing value on one or on the both variables, it is excluded (Williams 2012, pp. 2-4).
2. Mean imputation – all missing values of variable X are replaced with the mean of the observed values of X (Longford 2005, pp. 40-41).

3. Hot Deck – a random subject similar to recipient is selected and his/her data will be used instead of missing value (Longford 2005, pp. 43-44).
4. Last Observation Carried Forward (LOCF) – missing value will be replaced with predefined substitute variable's value (Longford 2005, p. 41).
5. Regression imputation – missing values of variable X will be predicted using regression model which uses completely recorded variable Z: $X = f(Z) + \varepsilon$, where $f(z) = \beta_0 + \beta_1 z$ and ε is a random variable (Longford 2005, pp. 45-46).
6. Expectation-maximization algorithm – iterative procedure where each iteration consists of two steps: the E-step, which estimates the complete-data log-likelihood and the M-step where the likelihood function is maximized, using the assumption that missing data is known, the sufficient statistics are replaced by their estimates gathered from the E-step (Borman, S., 2004, p 5).

1.1. Single and multiple imputation

Single and multiple imputation are discerned. In multiple imputation first of all a model is fitted, then plausible values generated which is followed by analyzing each completed data set and finally an average of completed data estimators is found. Basically comparing to single imputation more datasheets are created and therefore the role of randomness decreases (Longford 2005, pp. 61-64).

1.2. MAR

The most common missingness mechanism assumed in practice is MAR. (Longford 2005, p. 62; Schafer 1997, ch. 2.2.1). The following overview about MAR is based on (Scheffer 2002).

Scheffer generated a sample of 1,000 cases with 3 explanatory variables and a dependent variable, last one was generated using combination of previous ones with added random component. Then she artificially created all of the three missingness mechanisms.

Eight different methods using various software were used by her for observing what happened to mean and standard deviation while dealing with missing data:

1. All value (Pairwise deletion) in SPSS;
2. Listwise in SPSS;
3. Group means in SOLAS software, single imputation (SI);
4. Hot deck in SOLAS software, SI;
5. Regression in SPSS MVA software, SI;
6. Expectation-Maximization algorithm in SPSS MVA software, SI;
7. Expectation-Maximization algorithm in SOLAS software, multiple imputation (MI);
8. Markov chain Monte Carlo (MCMC) in NORM software, MI.

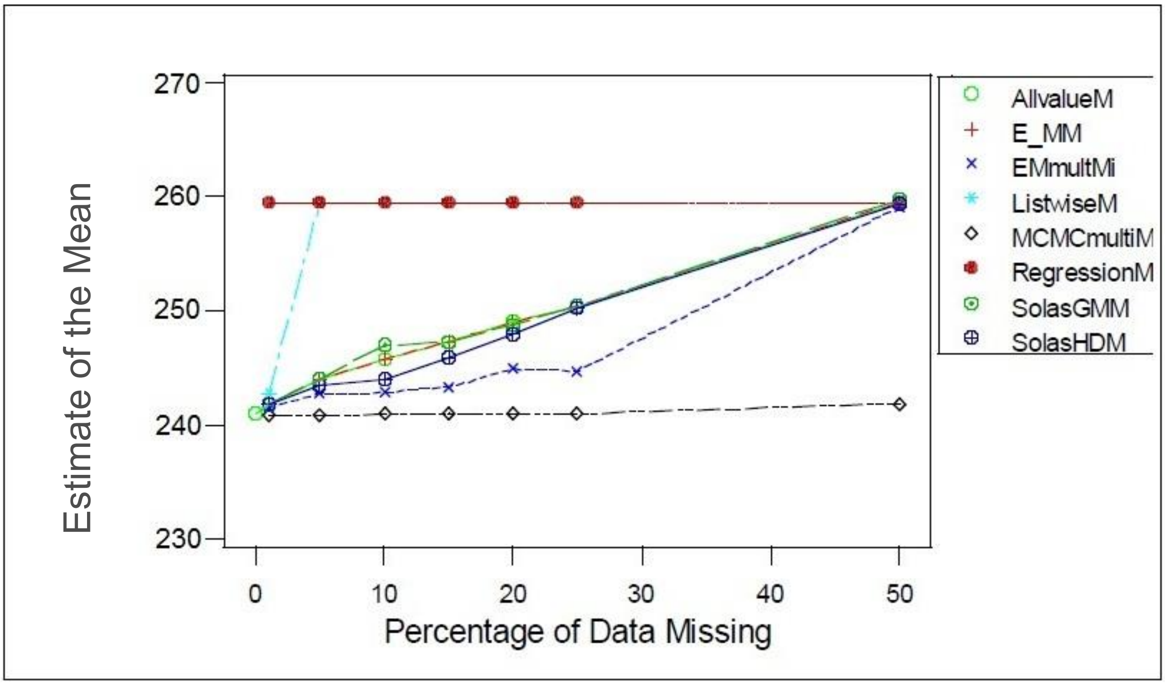

Figure 1. Plot of the mean for MAR imputed data by amount of the data missing (Scheffer 2002, p. 158).

The correct value of mean was 240.99. Up to 5 % of data missing all of these methods, except listwise and regression, estimated the mean quite well. SPSS MVA regression does not perform well due to the fact that regression parameters are biased because they are derived using case deletion and therefore estimates of the moments can be conditional (because only observed values are used) and may differ essentially from the unconditional moments (Hippel 2004, p. 160). Up to 10% of data missing hot deck and EM in SOLAS and MCMC in NORM estimate fine. When half of the data were missing then only MCMC gave rational result.

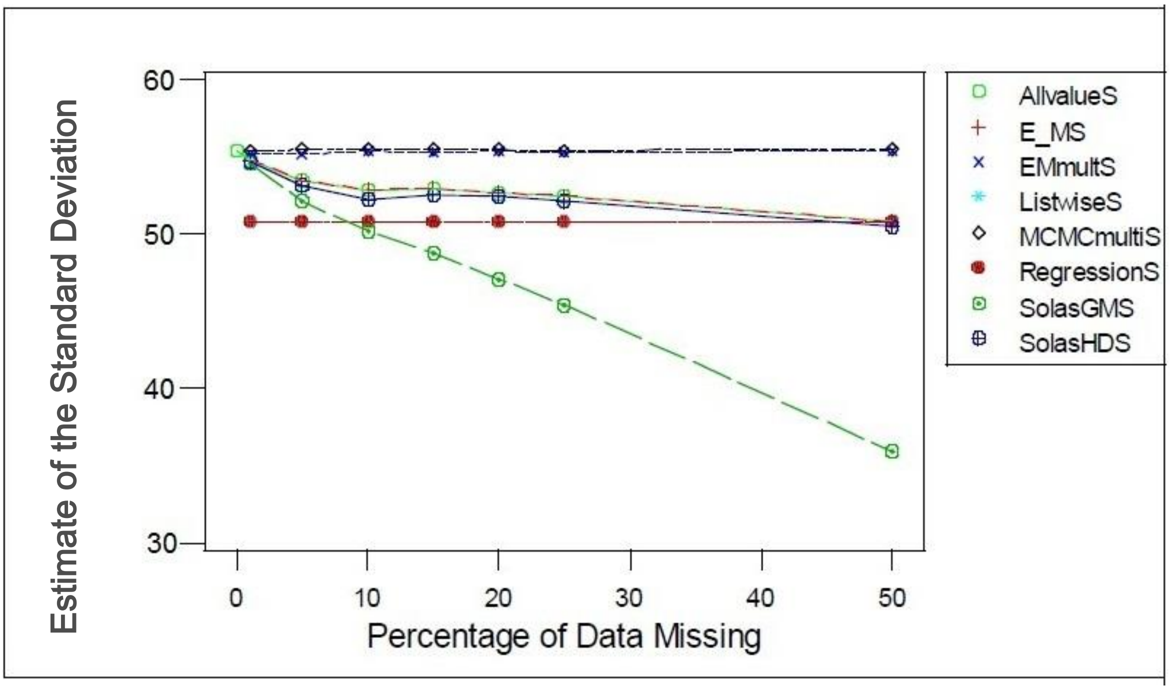

Figure 2. Plot of the standard deviation for MAR imputed data by amount of the data missing (Scheffer 2002, p. 158).

The precise value of the standard deviation was 55.39. Only two multiple imputations (EM in SOLAS and MCMC in NORM) did not fail to retain the structure of variance which was almost no change. The mean imputation underestimated standard deviation strongly.

2. Imputation Methods

Selection of methods is based on evaluations in (Scheffer 2002).

When missingness mechanism is MAR, then single imputation gives reasonable results up to 10% of data missing while imputing the mean value. However, when variance structure is vital, then no more than 5% of the data should be missing. Multiple imputation offers decent results up to 25% of data missing.

Which method to choose also depends on the missing pattern and the type of the variable with missing values (Yuan 2011, p. 3).

2.1. Missing patterns

Missing pattern can be **monotone** or **non-monotone** (Yuan 2011, p. 3; Longford 2005, pp. 26-28). Missing pattern is said to be monotone when each variable has less missing values than subsequent variables. When X_1 and X_2 are two vectors with same length, then $X_1 \succcurlyeq X_2$ means that the value of X_1 is at least as much as value of X_2 for every subject. For a dataset with columns $X_1 \dots X_k$ and response indicators $R_1 \dots R_k$ monotone response pattern is defined by $R_1 \succcurlyeq \cdots \succcurlyeq R_k$ (recorded at least as much as).

Non-monotone pattern is pattern which is not monotone. When notion “arbitrary” is used, then missing pattern can be any kind of.

2.2. Ignorable and non-ignorable missing

Missing data mechanism is called **ignorable,** when data model parameters θ and missing data indicators parameters ϕ are distinct, which means that knowing the values of either θ or ϕ does not deliver any additional information about the other one (MAR and MCAR).

Missing data is non-ignorable when the missing is dependent on the value of missing observation (data is NMAR) (Yuan 2011, p. 2; Marlin, Roweis, Zemel 2005, Introduction).

2.3. Regression imputation

Overview of regression imputation is given on the basis of (Yuan 2011, pp. 3-4; Käärik lecture materials 2012).

For a variable with missing values, a model is fitted using observed values for the variable. With this model, a new model is drawn and is used to impute missing values.

If Yj is the variable with missing values, then $Y_j \sim P\left(Y_j \middle| Y_1, Y_2, \dots, Y_{j-1}\right)$ is the distribution, where the values will be imputed from.

2.3.1. Monotone regression

Used for continuous variable when missing pattern is monotone.

Regression model is $Y_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$, where $X_1, \ldots, X_k$ are the covariates generated from preceding variables $Y_1, Y_2, \ldots, Y_{j-1}$ (Yuan 2011, p. 4)

Definition 1. The posterior predictive distribution is the distribution of unobserved observations (X_{mis}) conditional on the observed data (X_{obs}), θ is the parameter.

$$P(X_{mis}|X_{obz}) = \int P(X_{mis}, \theta|X_{obs}) d\theta$$

$$= \int P(X_{mis}|\theta, X_{obs}) P(\theta|X_{obs}) d\theta = \int P(X_{mis}|\theta) P(\theta|X_{obs}) d\theta$$

(Hitchcock, Posterior Predictive Distribution, pp. 1- 2).

Definition 2. If A $(n \times n)$ is a symmetric positive definite matrix, which means that $a_{ij} = a_{ji}$ for all $i, j = 1, \ldots, n$ and $x^T A x > 0$ for all column vectors x (n-dimensional), then the Cholesky decomposition is an upper triangular matrix U with strictly positive diagonal entries such that $A = U^T U$ (Weisstein, Eric W., "Cholesky Decomposition").

To impute the missing values for Y_j three steps are repeated at each imputation (Yuan 2011, p. 4, Käärik 2012):

1. The regression model for Y_j is fitted using observed values for the variable Y_j and covariates $X_1, X_2, \ldots, X_k$. This model includes the regression parameter estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k)$, $\widehat{\beta} = (X^T X)^{-1} X^T Y_j$, where X is the design matrix, and the associated covariance matrix $\hat{\sigma}_j^2 V_j$, where V_j is the usual $(X^T X)^{-1}$ matrix derived from the intercept and covariates $X_1, X_2, \ldots, X_k$.

2. New parameters $\hat{\beta}_* = \left(\hat{\beta}_{*0}, \hat{\beta}_{*1}, \ldots \hat{\beta}_{*k}\right)$ and σ_{*j}^2 are drawn from the posterior predictive distribution (definition 1) of the parameters, which are simulated from $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k)$, $\hat{\sigma}_j^2$ and V_j.

The variance is drawn as

$$\sigma_{*j}^2 = \hat{\sigma}_j^2(n_j - k - 1)/g$$

where g is a $\chi^2_{n_j-k-1}$ random variate and n_j is the number of observed values for Y_j . The regression coefficients are drawn as

$$\beta_* = \hat{\beta} + \sigma_{*j} V_{hj}^{'} Z$$

where $V_{hj}^{'}$ is the upper triangular matrix in the Cholesky decomposition (definition 2), $V_j = V_{hj}^{'} V_{hj}$ and Z is a vector of $k + 1$dependent random normal variates.

3. The missing values are then replaced by $\beta_{*0} + \beta_{*1}x_1 + \beta_{*2}x_2 + \cdots + \beta_{*(k)}x_k + z_i\sigma_{*j}$ where $x_1, x_2, \ldots, x_k$. are the values of the covariate and z_i is a simulated normal deviate.

2.3.2. Monotone logistic regression

Used for monotone missing patterns when imputed variable is ordinal classification variable (discrete variables with natural order) (Yuan 2011, p. 3)

Dependent variable has either binomial or Bernoulli distribution. Here we see a case, where the variable with missing values is a binary variable.

Main idea is similar to monotone regression, but here not a value of Y, but the probability of the value is predicted using function $\operatorname{logit} \pi = \ln\left(\frac{\pi}{1-\pi}\right)$, where π is the probability of "success" $P(Y = 1) = \pi$ (Käärik 2012).

2.4. Sequential regression multivariate imputation (SRMI)

Following overview on SRMI is given on the basis of (Ragunathan, Lepkowski, Hoewyk and Solenberger 2001, pp 85-88; Traat, lecture materials)

Sequential regression assumptions:

1. Population is essentially infinite.
2. Simple random sample.
3. Ignorable missing data.
4. Data types:
    a. Continuous;
    b. Binary;
    c. Categorical (more than two categories);
    d. Count;
    e. Mixed (firstly zero-non-zero status is discrete and secondly all of the values different from zero are continuous).

Usually survey data include many variables with very different distributions. Also restrictions may be necessary, because some of the variables may be measured only on certain subjects and in addition there might be logical bounds for some variables which need to be taken into account when imputing. For example components of inventories can not exceed inventories total. Using SRMI it is possible to handle complex data structure.

Definition 3. A prior distribution $p(\theta)$ of a parameter is the probability distribution that represents uncertainty about the parameter before the current data are examined. Prior distribution describes which values of θ are more likely and which are less likely to appear (Prior Distributions 2012, p. 1; Traat, p.2).

Definition 4. Multiplying the prior distribution and the likelihood function together leads to the posterior distribution $p(\theta|x)$ of the parameter, where $f(x|\theta)$ is the distribution of the observed data

$$p(\theta) = \frac{f(x|\theta)p(\theta)}{f(x)}$$

(Traat, p. 3).

Definition 5. A prior $p(\theta)$ is non-informative if it has minimal impact on the posterior distribution of θ (Prior Distributions, SAS/STAT(R) 9.2 User's Guide).

Definition 6. Flat prior is a prior distribution, which assigns equal likelihood on all of the parameter's values. In linear regression flat priors on the regression parameter are non-informative (Prior Distributions, SAS/STAT(R) 9.2 User's Guide).

Imputations are created through a sequence of multiple regressions, type of the regression model depends on the type of the imputed variable. All other variables observed or imputed for that individual are covariates. "The imputations are defined as draws from the posterior predictive distribution specified by the regression model with a flat (definition 6) or non-informative (definition 5) prior distribution for the parameters in the regression model" (Ragunathan, Lepkowski, Hoewyk and Solenberger 2001, p. 86). The sequence of imputing can be continued in a cyclical manner, each time replacing previously drawn values with new ones, creating complementary relationships between imputed values and exploiting the correlational structure among covariates. For multiple imputation every P^{th} set of imputed values can be used in the cycles or different random starting seeds can be used.

Let X denote a $n \times p$ design or predictor matrix including all the variables without any missing values, where n is a sample size. X consists of binary, continuous, count, mixed and dummy variables, last ones represent categorical variables. In addition it may also include column for intercept, offset and design variables.

Let $Y_1, \dots, Y_k$ denote k variables which have missing values. $Y_1, \dots, Y_k$ are ordered by the amount of missing values, from least to most. However, the pattern does not have to be monotone.

The joint conditional density of $Y_1, \dots, Y_k$ given X is

$$f(Y_1, Y_2, \dots, Y_k | X, \theta_1, \theta_2, \dots, \theta_k) = f_1(Y_1 | X, \theta_1)\, f_2(Y_2 | X, Y_1, \theta_2) \dots f_k\left(Y_k \middle| X, Y_{1,} Y_2, \dots, Y_{k-1}, \theta_k\right)$$

where $f_j = 1,2, \dots k$ are the conditional density functions and θ_j is a vector of parameters in the conditional distribution. All of the conditional densities are modeled through regression models with unknown parameters θ_j and draw from the predictive distribution of the missing values given observed values. We assume that the prior distribution (definition 3) for the parameters $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ is $\pi(\theta) \propto 1$ (equal probabilities to the likelihood).

Depending on the type of the variable, following models are used:

1. Y_j is continuous variable – a normal linear regression model on a suitable scale;
2. Y_j is binary variable – a logistic regression model;
3. Y_j is categorical variable – a polytomous or generalized logit regression model;
4. Y_j is count variable – a Poisson loglinear regression model;
5. Y_j is mixed:
    a. zero - non zero status – two stage model using logistic regression
    b. if status is non-zero, the values are imputed using normal linear regression model.

Each imputation has c rounds. Firstly, Y_1 which has the least values missing, is imputed on X. Considering that a flat prior is assumed for the regression coefficients, the Y_1 missing values' imputations are the draws from the corresponding posterior predictive distribution. Then X is updated by adding Y_1. Then the imputation process is repeated for Y_2 using updated X. This will be continued until all of the variables are imputed. Basically Y_1 is regressed on $U = X$, Y_2 is regressed on $U = (X, Y_1)$, where Y_1 has imputed values, Y_3 is regressed on $U = (X, Y_1, Y_2)$ and so on.

Then imputation is repeated in all other c-1 rounds including all of the Y variables in the predictor set (except the one which was previously used as the dependent variable). Therefore Y_1 is regressed on $U = (X, Y_2, \dots, Y_k)$, Y_2 is regressed on $U = (X, Y_1, Y_3, \dots, Y_k)$ and so on. This action is repeated a predestinated times or until stable imputed values appear.

Restrictions need to be taken into account for some variables, because distributions may include any kind of values which might not be suitable for current variables. For instance when enterprise does not have any inventories, then components of inventories should not be imputed. In some cases the imputation can be restricted with the value in sample. In that occasion it is possible that variable needs to be changed before adding it to covariates, possibly dummy variables may be created. For instance, usually finished goods and work in progress appear in industry or construction enterprises and thus imputing may be restricted with the dummy variable, which represents only suitable field of activity. Some variables require truncated regression models and the imputations are then drawn from the corresponding truncated distribution conditional on the drawn value of the parameters.

Drawing values of parameters directly from their posterior distribution with truncated likelihood can turn out to be rather difficult.

However, it can be easily done for a given parameter value, for example using Sampling-Importance-Resampling (SIR) algorithm. Firstly, the trial parameters are drawn without adding any bounds and then each trial value will be added an importance ratio. Importance ratio is defined as the ratio of the true posterior density with bounds to the trial density without bounds. Finally, a single parameter's value is resampled with probability proportional to the importance ratios. There are other possibilities as well according to the type of the variable and situation, which are the possible values of the variable.

At the end of round the first complete dataset is available. If the missing pattern is monotone, the imputations in the first round are approximate draws from the joint posterior predictive density of the missing values given the observed values. Approximations of the draws from the logistic, polytomous and count variables can be improved by using reject algorithms like SIR in each subsequent round. If the pattern is not monotone, then Gibbs algorithm can be used.

2.5. Markov Chain Monte Carlo method (MCMC)

Arbitrary missing pattern is allowed and multivariate normality is assumed. (Yuan 2011, p. 5).

Definition 7. If X_n is a Markov chain with state space S and transition function P, then π is called stationary distribution when $\pi(x)$ is probability distribution so that

$$\sum_{x \in S} \pi(x)P(x,y) = \pi(y), y \in S$$

which means that π does not depend on the time moment (Markov Chains: Stationary Distributions, p. 1).

Using Markov chain Monte Carlo method it is possible to generate pseudorandom draws from probability distribution using Markov chains. Purpose is to construct Markov chain, which stationary distribution is the distribution of our interest. When simulating steps of the Markov chain repeatedly, it is feasible to simulate draws from distribution of interest. (Schafer 1997, ch 1.2.2; Yuan 2011, p. 5).

The most popular Markov chain Monte Carlo methods are Gibbs sampling (Schafer 1997, ch 3.4.1) and Metropolis-Hastings algorithm (Schafer 1997, ch 3.4.4).

Through Markov chain Monte Carlo it is possible in many cases to simulate the entire joint posterior distribution of the unknown quantities. (Schafer 1997, ch 1.2.2).

Advantages:

1. implementation may be easier while dealing with complex problems;
2. may be the only method when high-dimensional parameters are unknown;
3. asymptotic approximations have not been made;
4. provides random draws from their joint posterior distribution instead of point estimate.

However, dealing with large datasets and complicated models requires fast computer and a lot of memory (Schafer 1997, ch 1.3).

Overview of process is given in (Yuan 2011, p. 5).

Assuming that the data is from multivariate normal distribution, data augmentation is applied to Bayesian inference with missing data by repeating these two following steps:

1. I-step (imputation): With the estimated mean vector and covariance matrix, the I-step simulates the missing values for each observation independently. If the variables with missing values are denoted by $Y_{i(mis)}$and the variables with observed values $Y_{i(obs)}$, then the I-step draws values for $Y_{i(mis)}$ from a conditional distribution of $Y_{i(mis)}$ given $Y_{i(obs)}$
$$Y_{i(mis)} \sim \left(Y_{i(mis)} \middle| Y_{i(obs)}\right).$$

2. P-step (posterior): This step simulates the posterior population mean vector and covariance matrix from the complete sample estimates. These new estimates are then used in the I-step. Without prior information about the parameters, a non-informative prior is used. Other informative priors can also be used.

3. Description of the dataset

Similar datasets were available for years 2009, 2010 and 2011. The main goal was to fix up and impute missing values for year 2011.

Data were collected from annual reports of Commercial Register. Annual report consists of several sections: four main parts (balance sheet, profit and loss account, cash flow statement and statement of changes in equity) and notes (additional information, not compulsory). This paper concentrates on the imputation of inventories, corresponding data derived from balance sheet, notes and profit and loss account. All together there were 33 variables in the primary dataset:

1. $JYKOOD$ - code of the enterprise in Commercial Register.
2. $TI\_valuuta$ - currency of the monetary variables.

   From balance sheet.

3. Bi_60_1 – Inventories total (at the end of the year).
4. Bi_60_2 – Inventories total (at the beginning of the year).
5. Bi_190_1 – Assets total (at the end of the year).
6. Bi_190_2 – Assets total (at the beginning of the year).
7. Bi_590_1 – Total of liabilities and equity (at the end of the year).
8. Bi_590_2 – Total of liabilities and equity (at the beginning of the year).

   From notes.

9. $L13\_10\_1$ – Raw materials and materials (at the end of the year) .
10. $L13\_10\_2$ – Raw materials and materials (at the beginning of the year).
11. $L13\_20\_1$ – Work in progress (at the end of the year).
12. $L13\_20\_2$ – Work in progress (at the beginning of the year).
13. $L13\_30\_1$ – Finished goods (at the end of the year).
14. $L13\_30\_2$ – Finished goods (at the beginning of the year).
15. $L13\_40\_1$ – Merchandise purchased for resale (at the end of the year).
16. $L13\_40\_2$ – Merchandise purchased for resale (at the beginning of the year).
17. $L13\_50\_1$ – Prepayments to suppliers (at the end of the year).
18. $L13\_50\_2$ – Prepayments to suppliers (at the beginning of the year).

19. $L13\_60\_1$ – Inventories total (at the end of the year).
20. $L13\_60\_2$ – Inventories total (at the beginning of the year).
21. $L51\_60\_1$ – Code of the field of activity (according to EMTAK, which is The Estonian Classification of Economic Activities).
22. $L51\_50\_1$ – Sales revenue according to field of activity (according to EMTAK).
23. $L51\_20\_1$ – Whether is primary activity or not (according to EMTAK).
24. $L51\_30\_1$ – Percentage of sales revenue (according to EMTAK).

25. $TYYP$ – type of data produced to Commercial Register (XBRL – electronic , PDF – on paper).
26. $LAADIMINE$ – date, when the data were taken from the Commercial Register.
27. $VERSIOON$- number of version.
28. MAJ_ALGUS- beginning of the accounting period.
29. MAJ_LOPP – end of the accounting period.

From profit and loss account.

30. Ka_90_1 – Change of work in progress and finished goods inventories remainders.
31. Ka_70_1 – Change of agricultural production inventories remainders.
32. Ka_50_1 – Sales revenue.
33. Ka_360_1 – Profit (loss) of financial year.

3.1. Describing the missingness and organizing the dataset

Following overview is based on (Schwartz, Chen, Duan 2011).

For describing the missingness in data a SAS macro %missingPattern was used. There were three parameters to specify when calling out the macro – a dataset, type of the missingess analysis (four different available) and an output dataset.

Available patterns in %missingPattern macro are.

1. $misspattern1$ - for each variable a missingness indicator is generated, where 1 presents the missing value and 0 the observed value. Indicators are named $m\_name - of - the - variable$. In output dataset each row represents one singular pattern of

the indicators. Also number of subjects with each pattern ($NObs$) is delivered as well as the proportion of the pattern ($missPattern\_prop$).

2. $misspattern2$ - the number of missing values and the percent of missing data in each variable is delivered
3. $mispattern3$ - the pairwise concordance between any two variables is provided (delivers percentages of data missing in the first variable when second is observed and vice versa, and percentages of two variables being observed or missing together). Examining this pattern allows us to decide more easily which variable should be used in models and analysis, for example when total of liabilities and equity equals to the assets total, then the one with fewer missing values should be taken into account.
4. $mispattern4$- checks the data for unit non-response – whether the most extreme missing pattern matches the theoretical pattern for unit non-response (design variables are still measured). If such a pattern is found, then the data, where it was, is outputted, otherwise there appears a remark in the log window.

The most difficult and important was to fill the missing values for the inventories in the notes accurate as possible.

Enterprises with more than one field of activity occupied one row for each field. At first there were 109,565 observations. Imputation was done only for the main field of activity. After removing observations where $L51\_20\_1$ stated that the field of activity was not primary, 76,167 observations were kept.

Secondly, there was a variable $regkood$ created with the value of the first number of $JYKOOD$. $Regkood$ represented the type of the enterprise. All of the variables where the value of $regkood$ was other than 1, were deleted. Enterprises whose code started with either 8 or 9 were non-profit enterprises and were not substantial fro the analysis in future. There were 12,203 values dismissed and 63,964 observations were still left in the dataset.

Thirdly, a variable $aegind$ was created to express whether the period of the accounting year was shorter, longer or exactly one year. There were 2,032 observations with accounting year longer and 4,783 with accounting year shorter than one year and due to that they were not included in further analysis. As a result 57,149 observations were left.

By reason of balance sheet being compulsory, all of the missing values in the total of inventories were replaced with zeros. Before replacing additional inspection was made. If the inventories total at the beginning of the year in 2011 was missing, it was replaced with the value of inventories total at the end of the year 2010, if possible. There were 507 values replaced (Bi_60_2 was renamed to $Bi\_60\_2\_uus$). After that the values of inventories total in the notes were synchronized with balance sheet values. There were 31,017 enterprises, where the inventories total at the beginning of the year and 31,454 at the end of the year were replaced by zero.

A variable arv was created using another dataset called *kogum_erilised*, which is statistical profile of Statistics Estonia. It is updated every year and all of the statistics of economy is based on this profile. Variable arv expressed the number of persons employed in the enterprise. 14,232 observations did not have the number of persons employed in the dataset *kogum_erilised.* Also a variable $tegevusala$ was created to express the field of activity (it was converted to numerical and called $tegnum$), using already existing code of field of activity.

One of the components – prepayments for suppliers – had also negative values. Thus case, where inventories total was zero and only the prepayments for suppliers was missing, was inspected. No observations that kind appeared.

Also some requirements needed to be fulfilled.

1. The inventories total in the balance sheet had to equal to the inventories total and the sum of all the components of inventories in the notes at the end of the year and at the beginning of the year (later indicated as the first condition).

$$\begin{aligned} Bi\_60\_1 &= L13\_60\_1 \\ &= L13\_10\_1 + L13\_20\_1 + L13\_30\_1 + L13\_40\_1 + L13\_50\_1 \\ Bi\_60\_2 &= L13\_60\_2 \\ &= L13\_10\_2 + L13\_20\_2 + L13\_30\_2 + L13\_40\_2 + L13\_50\_2 \end{aligned}$$

2. In balance sheet assets total had to equal to the total of liabilities and equities (later indicated as the second condition).

$$Bi\_190\_1 = Bi\_590\_1.$$

$$Bi\_190\_2 = Bi\_590\_2.$$

3. Change of work in progress and finished goods inventories remainders added change of agricultural production inventory remainders from profit and loss account had to equal to the change between the sum of the work in progress and finished goods at the end of the year and at the beginning of the year (later indicated as the third condition).

$$Ka\_70\_1 + Ka\_90\_1 = (L13\_20\_1 + L13\_30\_1) - (L13\_20\_2 + L13\_30\_2).$$

The first condition had two parts. The first part, where the inventories total in the balance sheet had to equal to the inventories total in the notes, was met, because previously the missing values were replaced with zeros and then synchronized. The second part, where the inventories total in the balance sheet had to equal to the sum of all components in the notes, had 8,966 observations, where the condition was not met at the end of the year and 8,688 observations at the beginning of the year. When this condition was met and some of the components were missing, they were substituted with zeros to avoid imputing some other value to them, which could have caused the condition not to met.

The second condition should have been met in all of the cases, because the balance sheet is compulsory. Although, at the end of the year, there were 4 cases with assets total missing and 3 cases with the total of liabilities and equities missing. Luckily none of the observations had both of them missing and due to that the missing values were replaced with each others value.

At the beginning of the year, it was not such an easy case. There were 255 observations with property and debts missing and 257 observations with total of liabilities and equities missing. There were 255 observations with both of them missing at the same time. Where the total of liabilities and equities was missing, the value of property and debts was used for replacing, which was done for 2 observations.

All of the cases at the beginning of the year still missing were tried to replace with values from the end of previous year. Both of these two variables had 76 replacements, which means that 179 values were still missing. After replacing as much as possible with real values, other missing values were changed to zeros due to the fact that balance sheet is compulsory.

The third condition was met in 10,864 cases.

Table 1 shows which variables needed imputing and how many values were missing and also the amount of values missing.

Table 1. Variables needed imputing

Variable	Number missing	Percent missing
Ka_360_1	12	0.02
Ka_50_1	513	0.90
Ka_70_1	56072	98.11
Ka_90_1	54852	95.98
$L13\_10\_1$	8701	15.22
$L13\_10\_2$	8420	14.73
$L13\_20\_1$	8859	15.50
$L13\_20\_2$	8587	15.02
$L13\_30\_1$	8829	15.44
$L13\_30\_2$	8556	14.97
$L13\_40\_1$	8649	15.13
$L13\_40\_2$	8372	14.65
$L13\_50\_1$	8735	15.28
$L13\_50\_2$	8461	14.80
$L51\_30\_1$	1324	2.32
$TARVANKP$	14232	24.90

4. Simulation

To observe how SRMI acted with current data, simulation was carried out. Three datasets were compared fugitively:

1. where only the condition $Bi\_60\_1 = lisasummalopp$ were met (first dataset) ($lisasummalopp$ was the sum of the components of inventories from notes at the end of the year and $lisasummalgus$ at the beginning of the year);
2. where $Bi\_60\_1 = lisasummalopp$ and $Bi\_60\_2\_uus = lisasummalgus$ (second dataset);
3. where $Bi\_60\_1 = lisasummalopp$ and $Bi\_60\_2\_uus = lisasummaalgus$ and $Ka\_70\_1 + Ka\_90\_1 = L13\_20\_1 + L13\_30\_1 - L13\_20\_2 - L13\_30\_2$ (third dataset).

First dataset had 48,183 observations, second 46,805 and third had 1,252 observations. Observing mean values of variables needed imputing, it occurred that the first and the second dataset had similar values, but the third one had values unlike the others. Due to that, simulation was carried out on two datasets: on the second and on the third.

4.1 Simulation of the second dataset

The second dataset had all of the variables included in conditions 1 and 2 (Bi_60_1, $Bi\_60\_2\_uus$ and all of the components on inventories) fully observed. Then artificially approximately 10%, 30% and 50% of observations were set to missing for those variables. Other variables, which were not part of the conditions 1 and 2, but had missing values, were not set any extra missingness.

After imputing table 5 was created. The first row shows the difference between $vahe1$, calculated after imputing, and the real value of $vahe1$, second row shows same thing for $vahe2$. Last column has real mean values. Other numbers in table were calculated as relations between the difference from real value and the real value, for example

$$(Ka\_360\_1 - true\ value)/true\ value.$$

Thus zero expresses the most accurate value. Closer to zero, the better result imputation gave. Negative operator means that the mean value after imputing was smaller than the true value, positive operator meaning is vice versa. Average in bottom row expresses the mean value of absolute values of relations previously calculated. Also cases where the type of components of inventories in notes is either mixed or continuous were compared.

When the type on components of inventories was mixed, there did not seem to be any regularity in results. For example method clearly imputed Ka_90_1 worse when 50% of data were missing than case where 10% of data were missing. Similar relation was noticeable with $vahe3$, $L13\_40\_1$ and $L13\_40\_2$. On the other hand $L13\_20\_1$ and $L13\_20\_2$ had best results, when 50% of the data were set to missing. Some variables had worst results when 30% of data were missing, which did not indicate to relation that method worked better with this specific dataset when amount of data missing was smaller. Also average difference was bigger when 30% of the data were set to missing than 50%.

When type of components of inventories from notes was continuous, then results varied less. Although there were also some cases, where best result was achieved when 50% of data were set to missing, trend was that less the data were missing, better the results were.

Comparing to the mixed-type components, continuous components had much less misleading results.

Table 2. Comparison of difference from true value correspondingly to the type of components of inventories by amount of data missing in the second dataset.

Variable	10% missing		30% missing		50% missing		True value
	Mixed	Continuous	Mixed	Continuous	Mixed	Continuous	
$vahe$1	– 1,856,218	–46,132	–52,167	-147,319	–8,262,160	–304,363	0
$vahe$2	– 2,381,702	–36,154	–858,738	-136,882	–1,516,298	–270,023	0
$vahe$3	–71.9	0,3	75.1	-0.9	1142.2	–3.1	2,954
Ka_360_1	0.0	0	0.0	0.0	0.0	0	36,751
Ka_50_1	0.0	0	0.0	0.0	0.0	0	509,100
Ka_70_1	–1.0	-1	–1.2	-0.9	5.3	–0.2	9,049
Ka_90_1	–14.6	-0.6	–42.5	-0.9	255.3	–1,2	13,160
L13_10_1	0.4	1.2	1.1	3.8	1.5	8.2	8,984
L13_20_1	413.9	2.1	6.0	6.7	20.7	12.2	4,431
L13_30_1	3.7	1.7	2.1	5.2	9.4	11.1	4,975
L13_40_1	0.0	0.6	0.1	1.8	304.5	–0.5	26,635
L13_50_1	–0.1	0.9	0.6	4.4	0.6	9.3	2,099
L13_10_2	1.3	1.3	3.9	4.2	6.8	8.7	7,971
L13_20_2	448.4	2.4	7.5	8	22.2	14.3	4,031
L13_30_2	3.6	1.8	177.9	5.1	–0.4	9.7	4,383
L13_40_2	23.9	0.6	0.7	1.9	59.8	4	22,979
L13_50_2	–0.4	-3.2	1.0	2.5	–0.6	5.1	1,859
L51_30_1	0.0	0	0.0	0	0.0	0	91
$TARVANKP$	0.0	0	0.0	0	0.0	0	6
Average	57.8	1.0	18.8	2.7	107.6	5.2	–

4.2 Simulation of the third dataset

When inspecting the third condition, missing values were treated as zeros by program and thus they were replaced with zeros before erasing values, because condition was met (concerning Ka_70_1 and Ka_90_1). When variable needed imputing, but was not part of one of the conditions, none of the replacements were made, because true values were not known. That meant $tarvankp$, L51_30_1 and Ka_50_1, which were not fully observed before deleting, may had had bigger or smaller percent of data missing than other values.

Similarly to table 5, three first rows in table 6 show mean values of $vahe1, vahe2$ and $vahe3$, rest of the numbers state how many times did the imputed mean differ from the true mean value. As well as in table 5, when type of the components of inventories was mixed, then, some of the variables had mean values after imputing more close to the real value when 10% of the were set to missing, others had opposite situation and mean computed after imputing was closer to the real value when 50% of the data were missing. Also mean values of $vahe1$, $vahe2$ and $vahe3$, calculated after imputing, which were suppose to be zeros, were strongly underestimated. When type of the components of inventories was continuous, then results were notably better, up to 10% of the data missing, none of the mean values were mistaken more than 60%.

Table 3. Comparison of difference from true value correspondingly to the type of components of inventories by amount of data missing in the third dataset.

Variable	10% missing		30% missing		50% missing		True value
	Mixed	Continuous	Mixed	Continuous	Mixed	Continuous	
$vahe1$	–9,579,095	–50,323	–31,138,071	–183,613	–43,860,356	–217,140	0
$vahe2$	–3,373,620	–47,144	–8,340,237	–156,679	–14,255,545	–227,624	0
$vahe3$	–1,524,906	3,596	–2,302,306	-515	5,415,163	–7,041	0
Ka_360_1	0.0	0	0.0	0.0	0.0	0.0	100,601
Ka_50_1	0.0	0	0.0	0.0	0.0	0.0	1,469,381
Ka_70_1	–0.1	–0.2	37.7	0.0	0.6	–0.2	5,245
Ka_90_1	–29.9	–0.6	605.1	1.5	1054.4	–1,0	13,308
$L13\_10\_1$	–9.8	0.3	0.0	1.2	0.8	1.7	83,043
$L13\_20\_1$	47.4	0.0	170.8	0.2	–0.4	0.1	32,165
$L13\_30\_1$	1.1	0.3	73.2	0.9	128	0.6	57,237
$L13\_40\_1$	525.9	0.5	1585.9	1.7	2683	2.5	13,529
$L13\_50\_1$	20.2	0.6	-0.5	0.9	60.0	2.5	3,061
$L13\_10\_2$	2.1	0.4	0.3	1.3	1.2	1.9	72,028
$L13\_20\_2$	–0.1	0.1	–0.3	0	0.4	0.6	27,754
$L13\_30\_2$	9.0	0.2	0.0	0.8	0.2	0.8	47,138
$L13\_40\_2$	199.9	0.5	602.9	1.8	1020	2.5	13,580
$L13\_50\_2$	29.1	0.4	45.3	0.1	97.9	0.2	3,030
$L51\_30\_1$	0.0	0.0	0.0	0	0.0	0	87
$TARVANKP$	–0.2	–0.1	–0.1	–0.1	–0.1	-0.1	23
Average	54.6	0.26	195.1	0.7	315.4	0.91	–

These results indicate that sequential regression functioned well when type of the components of inventories was continuous. This was taken into account for further imputations.

5. Practicing methods

5.1 Using MCMC

First of all, it became clear that the missing pattern was not monotone. Thus monotone regression was not possible option to use for imputing.

Secondly, MCMC was tested to impute all of the variables needed imputing. The number of burn-in iterations before the first imputation, which were later discarded, was 200. Number of iterations between imputations was 100 and a single chain was used for all imputations. Some warnings emerged, which declared that the covariance matrix computed in the EM process was singular and due to that linearly dependent variables for the observed data were excluded from the likelihood function and it might not have given appropriate results. SAS suggested increasing the number of iterations as one possible solution to assure convergence of the EM. Also increasing the value of the convergence criterion was recommended. The iterations are said to have converged when the maximum change in the parameter estimates between iteration steps is smaller than the value specified, which default setting was 10^{-4} (MCMC Method Specifications, SAS/STAT(R) 9.2 User's Guide, Second Edition).

Same problems were continual after increasing the maximum number of iterations of the EM algorithm from 200 to 500, number of burn-in iterations to 400 and changing the convergence criterion to 10^{-3}.

Those results indicated that some kind of grouping might be needed. BY statement was used for number of persons employed which meant that imputing was done separately in every size group of enterprise. Groups were divided as follows: 1 person employed, 2 to 9 persons employed, 10–19 persons employed and more than 20 persons employed. Some of the observations had number of persons employed missing and due to that it was previously imputed using MCMC. Only number of persons employed was generated, using fully observed variables ($tegnum$, Bi_60_1, $Bi\_60\_2\_uus$, Bi_190_1, $Bi\_190\_2\_uus$, $L51\_50\_1$). Boundaries were added as well, to ensure that only positive values were imputed. Maximum value of number of persons employed of observed values was 3,113 and due to that imputed number of persons employed was bounded from both sides accordingly with 0 and 3,735.

Additionally all of the values were bounded with the maximal and the minimal value of the observed values and approximately 20% was either added or subtracted to guarantee that the missing value fitted the interval. Arranging into groups and adding intervals was not sufficient enough. Program was not able to impute value from predefined interval with 100 tries.

Another solution might have been using MCMC imputing for only as much needed for making missing pattern monotone (this option was available in proc MI choosing impute= monotone) and after that using monotone regression method (Yuan 2011, p. 3, 11). This did not work due to the similar errors which occurred trying to impute the whole dataset.

Different tactics were tested: variables were imputed one by one based on the number of values missing – variable with the least missing values was done first. The first one imputed was Ka_360_1, the profit, which had only 12 observations missing. Then $L51\_30\_1$, the percent of sales revenue with 1,324 missing values was imputed, following $L13\_40\_1$, which represented merchandise purchased for resale at the end of the year. $L13\_40\_1$ had 11,350 missing values, 10 more than $L13\_40\_2$, but imputation of the beginning of the year did not succeed. Also all the other variables gave same error as when imputing $L13\_40\_2$ – an imputed variable value was not in the specified range after 100 tries. Without boundaries algorithms failed to converge.

5.2 Using IVEware

SAS macro IVEware was used for imputing all the missing values. A floating point error – overflow – occurred, which meant that computer had hardware limitations trying to fit infinite number to space of finite number (Montgomery, N. 2008).

Some of the possible solutions recommended, were setting boundaries to values, changing the random seed and not using too strongly correlated variables at the same time (Floating Point Errors and Overflows, SAS/STAT(R) 9.22 User's Guide).

All of the values were bounded with the maximal and the minimal value of the observed values and approximately 20% was either added or subtracted similarly to the previously tested MCMC method. Also random seed was changed and correlations were examined. Nothing was done about strongly correlated variables, because strong relations appeared mostly between variables, which measured same things at the different time moments or when variable was one part of the other.

Also imputing work in progress and finished goods were restricted with field of activity. Work in progress and finished goods are assumed only in fields of industry and building (code of field of activity starts with either 0, 1, 2, 3 or 4). Additionally those components may occur on following fields, which are not all covered in previous situation: manufacturing, construction, maintenance and repair of motor vehicles, publishing activities, motion picture, video and television programme production, sound recording and music publishing activities, programmes and broadcasting, telecommunications, computer programming, consultancy and other similar activities, architectural and engineering consulting; testing and analysis, research and development, market research and public opinion polls, security and investigation, maintenance of buildings and landscapes, office management, office support and other business support activities and repair of computers and personal and household goods.

Table 4 . Strongly correlated variables

Variable	Correlation more than 0.8				
Bi_60_1	$Bi\_60\_2\_uus$	$L13\_10\_1$	$L13\_40\_1$	$L13\_10\_2$	
$Bi\_60\_2\_uus$	Bi_60_1	$L13\_10\_1$	$L13\_40\_1$	$L13\_10\_2$	
Ka_50_1	$L51\_50\_1$	$L13\_10\_2$			
$L51\_50\_1$	Ka_50_1	$L13\_10\_1$	$L13\_10\_2$		
$L13\_10\_1$	$Bi\_60\_2\_uus$	Bi_60_1	$L51\_50\_1$	$L13\_10\_2$	
$L13\_20\_1$	$L13\_20\_2$				
$L13\_30\_1$	$L13\_30\_2$				
$L13\_40\_1$	Bi_60_1	$Bi\_60\_2\_uus$	$L13\_40\_2$		
$L13\_10\_2$	Bi_60_1	$Bi\_60\_2\_uus$	Ka_50_1	$L51\_50\_1$	$L13\_10\_1$
$L13\_20\_2$	$L13\_20\_1$				
$L13\_30\_2$	$L13\_30\_1$				
$L13\_40\_2$	$L13\_40\_1$				

SAS IVEware was used for imputing 10 times. Afterwards values of imputed variables were averaged (averaged variables were named $variable\_name\_kesk$). Variables $imputeeritudkokku1$ and $imputeeritudkokku2$, which represented sums of the inventories in notes correspondingly at the end and at the beginning of the year.

Variables

$$imputeeritudkokku1 = L13\_1\_1\_kesk + L13\_20\_1\_kesk + L13\_30\_1\_kesk + L13\_40\_1\_kesk + L13\_50\_1\_kesk$$

$$imputeeritudkokku2 = L13\_1\_2\_kesk + L13\_20\_2\_kesk + L13\_30\_2\_kesk + L13\_40\_2\_kesk + L13\_50\_2\_kesk,$$

were created. Also variables $vahe1$ and $vahe2$ were created, which were defined as differences between correct values of inventories total from balance sheet and sums of imputed components of inventories in notes:

$$vahe1 = Bi\_60\_1 - imputeeritudkokku1,$$

$$vahe2 = Bi\_60\_2\_uus - imputeeritudkokku2.$$

$Vahe3$ stood for difference between change of work in progress and finished goods inventories remainders added change of agricultural production inventory remainders from profit and loss account and change between the sum of the work in progress and finished goods at the end of the year and at the beginning of the year:

$$vahe3 = (Ka\_70\_1\_kesk + Ka\_90\_1\_kesk) - (L13\_20\_1\_kesk + L13\_30\_1\_kesk) - (L13\_20\_2\_kesk + L13\_30\_2\_kesk).$$

Table 5. Maximal and minimal value and mean of $vahe1$, $vahe2$ and $vahe3$ (EUR)

Variable	Maximum	Minimum	Mean
$vahe1$	1,828,693	–18,663,453	–85,405
$vahe2$	11,860,949	– 6,657,089	–53,819
$vahe3$	24,153,292	–3,465,807	–3,032

Results were not as good as expected. To get the inventories total values fit with correct values, coefficients $tegur1$ ($tegur2$) were calculated, dividing Bi_60_1 ($Bi\_60\_2\_uus$) with the $imputeeritukokku1$ ($imputeeritukokku2$):

$$tegur1 = Bi\_60\_1 / imputeeritudkokku1,$$

$$tegur2 = Bi\_60\_2\_uus / imputeeritudkokku2.$$

If Bi_60_1 ($Bi\_60\_2\_uus$) and $imputeeritukokku1$ ($imputeeritukokku2$) were equal, then $tegur1$ ($tegur2$) was fixed as 1 (without defining it separately, dividing with zero occurred). Afterwards all of the components of inventories were multiplied with $tegur1$ or $tegur2$, according to time when the variables were measured (variables were renamed to $variable\_name\_tegur$). Variables $vaheteg1$ and $vaheteg2$ represented the difference between the actual inventories total and the inventories total found after multiplying component with coefficient:

$$vaheteg1 = Bi\_60\_1 - (L13\_10\_1\_tegur + L13\_20\_1\_tegur + L13\_30\_1\_tegur + L13\_40\_1\_tegur + L13\_50\_1\_tegur),$$

$$vaheteg2 = Bi\_60\_2\_uus - (L13\_10\_2\_tegur + L13\_20\_2\_tegur + L13\_30\_2\_tegur + L13\_40\_2\_tegur + L13\_50\_2\_tegur).$$

As table 4 shows, there were not any notable misleading.

Table 6. Maximum, minimum and mean value of differences after multiplying with coefficient (EUR)

Variable	Mean	Minimum	Maximum
$vaheteg1$	$-4.5 \cdot 10^{-13}$	$-7.5 \cdot 10^{-9}$	$3.7 \cdot 10^{-9}$
$vaheteg2$	$-2.1 \cdot 10^{-13}$	$-3.7 \cdot 10^{-8}$	$1.5 \cdot 10^{-9}$

Next it was necessary to find coefficients for Ka_90_1 and Ka_70_1 as well, to get the third condition met, which was:

$$Ka\_70\_1 + Ka\_90\_1 = L13\_20\_1 + L13\_30\_1 - L13\_20\_2 - L13\_30\_2.$$

At the same time the first condition needed to stay satisfied. Right side of the third condition was considered as a correct value and coefficient $tegur3$ was calculated as following:

$$tegur3 = \frac{L13\_20\_1\_kesk + L13\_30\_1\_kesk - L13\_20\_2\_kesk - L13\_30\_2\_kesk}{Ka\_90\_1 + Ka\_70\_1}.$$

Similarly as before Ka_70_1 and Ka_90_1 were multiplied with $tegur3$ (and renamed accordingly $Ka\_70\_1\_tegur$ and $Ka\_90\_1\_tegur$) and thereafter left half of the condition was recalculated. Some additional measures were taken into account, associated with 1,907 observations, where $Ka\_90\_1\_kesk$ and $Ka\_70\_1\_kesk$ were imputed as zeros, but $L13\_20\_1\_tegur + L13\_30\_1\_tegur - L13\_20\_2\_tegur - L13\_30\_2\_tegur$ did not give zero. Due to the fact that Bi_60_1 had real values, it was not reasonable to change those. Also adding and subtracting between elements of the right side of the third condition without changing fit of the first condition was not possible, because both elements of one sum in the first condition (e. g $L13\_20\_1$ and $L13\_30\_1$) had same operators. Thus a change was needed in other components as well. Two cases were distinguished:

1. If the right side of the condition was bigger than zero.

$Vaheteg3$ was calculated as follows:

$$vaheteg3 = (Ka\_70\_1\_tegur + Ka\_90\_1\_tegur) - [(L13\_20\_1\_tegur + L13\_30\_1\_tegur) - (L13\_20\_2\_tegur + L13\_30\_2\_tegur)].$$

$L13\_20\_1$ and $L13\_30\_1$ were both added one fourth of absolute value of difference ($vaheteg3$) and other components of inventories at the end of the year were added one sixth of absolute value of difference each (variables were renamed to $variable\_teguruus)$, examples:

$$L13\_20\_1\_teguruus = L13\_20\_1\_tegur - \frac{|vaheteg3|}{4},$$

$$L13\_10\_1\_teguruus = L13\_10\_1\_tegur + \frac{|vaheteg3|}{6}.$$

$L13\_20\_2$ and $L\_13\_30\_2$ were subtracted same amount as $L13\_30\_1$ and $L13\_20\_1$, because they had opposite operators in condition and other components of the inventories total at the beginning of the year were subtracted one sixth of the absolute value of difference.

2. When right side of the condition was smaller than zero, everything was vice versa. No such cases were observed.

As a result both of the conditions were met at least with accuracy $-5.8 \cdot 10^{-11}$.

Summary

Thesis focused on imputing inventories section in annual reports of Commercial Register year 2011.

Firstly, missingness mechanisms were introduced and imputing methods presented.

Secondly, the dataset was described and fixed. All of the non-profit enterprises and ones with financial year different than one year were excluded. Also missing values were replaced from dataset of 2010 or set to zero, if possible. Additionally variables were synchronized and required conditions observed.

Thirdly, a simulation was carried out. Two datasets were created on the basis of the required conditions and were set 10%, 30% and 50% of missing values. Then sequential regression was carried out. Also case where components of inventories had mixed type was compared to the case where components of inventories were continuous.

Fourthly, Monte Carlo Markov chain method and sequential regression were practiced, because missingness was not monotone. MCMC was not successful. SRMI had better results. After that coefficients were calculated to get required conditions met. In the end a small simulation was carried out to observe how did the proportion of missingness affect meeting requested relations between variables.

As a result author suggests to add compulsory fields into annual bookkeeping report in Commercial Register, which represent whether or not enterprise has each and every component of inventory. Current situation allows enterprises to present their annual reports long after they are useful for statistical analysis and moreover many enterprises do not consider filling notes necessary, because it is voluntary, and thus a lot of useful information is not collected.

Varude imputeerimine Eesti Äriregistris 2011. aastal

Bakalaureusetöö

Cliona Georgia Dalberg

Kokkuvõte

Bakalaureusetöö tehti projekti “Metoodika väljatöötamine statistika tegemiseks kombineeritud administratiivsete andmeallikate ja uuringute andmete baasil” raames, mille tellis Eesti Statistikaameti käest Eurostat. Töös kasutati Äriregistrist pärit majandusaasta aruannete andmeid.

Bakalaureusetöö eesmärgiks oli imputeerida puuduvad väärtused varude osas 2011. aastal.

Töö esimeses osas anti lühiülevaade puudumisest ja selle mehhanismidest ning omadustest. Veel käsitleti mõningaid mitmese imputeerimise meetodeid nagu monotoonne regressioon, Monte Carlo Markovi ahelatega ja järjestikune regressioon.

Töö teises pooles räägiti läbiviidud simulatsioonist, kus võeti aluseks andmestikud, mis täitsid nõutud tingimusi. Neis andmestikes seati puuduvaks vastavalt ligikaudu 10%, 30% või 50% andmetest. Järgnes järjestikuse regresssiooni rakendamine ning saadud tulemuste keskmiste võrdlemine. Lisaks vaadeldi simulatsiooni aspektist, et milline tunnuse tüüp varude komponentidel lisaaruandes annaks korrektsema tulemuse. Seejärel tegeleti andmestiku korrastamisega, eemaldades mittehuvipakkuvad vaatlused ning asendades puuduvad väärtused olemasolevate andmetega varasemast aastast. Järgnevalt püüti varemkirjeldatud meetodeid rakendada. MCMC teostati SAS Enterprise Guide abil ning järjestikust regressiooni praktiseeriti makro IVEwarega, viimane osutus ainsana tulemuslikuks variandiks. Seejärel leiti mitmeid koefitsiente, et tagada andmestikus nõutud tunnustevahelised seosed.

Andmetöötlus tehti SAS Enterprise Guide’i ning SAS 9.2-ga, töö kirjutati Microsoft Word 2007-ga.

References

1. Borman, S., “The Expectation Maximization Algorithm. A short tutorial.” [pdf], available at <http://www.seanborman.com/publications/EM_algorithm.pdf>[Accessed 1 May 2013]
2. “Floating Point Errors and Overflows”, *SAS/STAT(R) 9.22 User's Guide* 2010, available at <http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_mcmc_sect037.htm> [Accessed 18 April 2013]
3. Hitchcock, D., “Posterior Predictive Distribution” [pdf], 2012, available at < http://www.stat.sc.edu/~hitchcock/stat535slidesday18.pdf> [Accessed 28 March 2013]
4. Käärik, E., Lecture materials “Andmeanalüüs II“ 2012 autumn semester
5. Longford N. T., *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*, Springer, 2005
6. “Markov Chains: Stationary Distributions”, available at <http://www.biostat.umn.edu/~sudiptob/pubh8429/MarkovChains7.pdf > [Accessed 27 February 2013]
7. Marlin, B.M., Roweis, S.T., Zemel, R.S., “Unsupervised Learning with Non-Ignorable Missing Data” [pdf], available at < http://www.cs.nyu.edu/~roweis/papers/aistat-lnimd.pdf > [Accessed 1 March 2013]
8. “MCMC Method Specifications”, *SAS/STAT(R) 9.2 User's Guide* 2009, Second Edition, available at: <http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mi_sect026.htm> [Accessed 22 April 2013]
9. Montgomery, N., Pharma, N., “Floating Point error – what, why and how to!!” [pdf] 2008, PHUSE 2008 Paper CS08, available at <http://www.phusewiki.org/docs/2008/PAPERS/CS08.pdf> [Accessed 25 April 2013]
10. “Posterior Predictive Distribution”, *SAS/STAT(R) 9.2 User's Guide* 2009, Second Edition, available at <http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introbayes_sect004.htm> [Accessed 22 March 2013]
11. “Prior Distributions”, *SAS/STAT(R) 9.2 User's Guide* 2009, Second Edition, available at

<http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introbayes_sect004.htm > [Accessed 22 March 2013]

12. Raghunathan, T. E., Lepkowski, J.M.,Van Hoewyk, J. and Solenberger, P., "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models", *Survey Methodology*, June 2001, Vol. 27, No. 1, pp. 85-95, available at < http://www.statcan.gc.ca/ads-annonces/12-001-x/5857-eng.pdf > [Accessed 14 April 2013]
13. Schafer J. L., *Analysis of Incomplete Multivariate Data*, Chapman and Hall/CRC, 1997
14. Scheffer J., "Dealing with missing data", *Research Letters in the Information and Mathematical Sciences*, 2002 (3), pp 153-160, available at <http://equinetrust.org.nz/massey/fms/Colleges/College%20of%20Sciences/IIMS/RLIMS/Volume03/Dealing_with_Missing_Data.pdf > [Accessed 17 February 2013]
15. Schwartz, T., Chen, Q., Duan, N., "Studying Missing Data Patterns Using a SAS® Macro", SAS Global Forum 2011, available at < http://support.sas.com/resources/papers/proceedings11/339-2011.pdf > [Accessed 01 March 2013]
16. Traat I., Lecture materials, "Bayesi statistika Markovi ahelatega"
17. von Hippel, P. T., "Biases in SPSS 12.0 Missing Value Analysis", *The American Statistician*, 2004, Vol. 58, No. 2, pp 160-164, available at <http://www.utexas.edu/lbj/sites/default/files/file/news/Biases%20in%20SPSS.pdf> [Accessed 12 March 2013]
18. Williams R., "Missing data" [pdf], 2012, available at < http://www3.nd.edu/~rwilliam/stats2/l12.pdf > [Accessed 22 February 2013]
19. Weisstein, E. W., "Cholesky Decomposition." From *MathWorld*--A Wolfram Web Resource, available at <http://mathworld.wolfram.com/CholeskyDecomposition.htm> [Accessed 15 May 2013]
20. Yuan, Y, "Multiple Imputation Using SAS Software", *Journal of Statistical Software*, December 2011, Volume 45, Issue 6, available at <http://www.jstatsoft.org/v45/i06/paper> [Accessed 17 January 2013]

Appendix

Fixing up the dataset

```
/* Creating dataset named yksteist to folder puud */
data puud.yksteist;
set test.maa_andmed_2011;
run;
/* Creating dataset named kymme to folder puud, changing some names of
variables to distuingish from year 2011 */
data puud.kymme;
set test.maa_andmed_2010;
rename Bi_60_1 = Bi_60_1_10 Bi_60_2 = Bi_60_2_10 jykood = jykood10 bi_190_2
= bi_190_2_10 bi_590_2 = bi_590_2_10;
run;
/* Taking only main field of activity*/
data puud.kymme;
set puud.kymme;
where L51_20_1 = 1;
run;
data puud.yksteist;
set puud.yksteist;
where L51_20_1 = 1;
run;
/* Creating variable regkood, which represents the type of the enterprise*/
data puud.yksteist;
set puud.yksteist;
regkood = substr (jykood, 1, 1);
run;
/* Checking, if there are any enterprises with other regkood than 1*/
proc freq data = puud.yksteist;
tables regkood;
run;
/* Leaving out non-profit enterprises */
data puud.yksteist;
set puud.yksteist;
if regkood = 8 or  regkood = 9 then delete;
run;
/* Creating variable aeg, which represents the time of the accounting
period */
data puud.yksteist;
set puud.yksteist;
aeg = maj_lopp-maj_algus;
aegaasta = aeg/60/60/24/364 ; /* converting to years */
run;
data puud.yksteist;
set puud.yksteist;
if aegaasta > 1 then aegind = 2;
else if aegaasta = 1 then aegind = 1;
else aegind = 0;
run;
/* Taking the enterprises whose accounting period is exactly one year
(aegind = 1) */
data puud.yksteist;
set puud.yksteist;
if aegind = 2 or aegind = 0 then delete;
run;
/*Changing missing inventories total to zeros in the balance sheet */
```

```
/* Replacing with value of 2010, creating new variable bi_60_2_uus, where
the result is held */
proc sql;
create table puud.asendatud as
select a.*, coalesce (Bi_60_2, Bi_60_1_10) as bi_60_2_uus
from puud.yksteist as a left join puud.kymme as b
on a.jykood = b.jykood10;
run;
/* Observing, how many replacements were made and how many values were
replaced with zeros*/
data puud.asendatud;
set puud.asendatud;
if Bi_60_1 = . then Bi_60_1_ind = 1;
if Bi_60_2 = . then Bi_60_2_ind = 1;
if Bi_60_2_uus = . then Bi_60_2_uusind = 1;
run;
proc freq data = puud.asendatud;
tables Bi_60_1_ind Bi_60_2_ind Bi_60_2_uusind;
run;
/* Replacing with zeros */
data puud.asendatud;
set puud.asendatud;
if (Bi_60_1 = .)
then Bi_60_1 = 0;
run;
data puud.asendatud;
set puud.asendatud;
if (Bi_60_2_uus = .)
then Bi_60_2_uus = 0;
run;
/* Adding variable tootajad, which represents the number of persons
employed */
/* Creating new variable jykoodnum, which is numerical instead of
character, then it is possible to compare */
data puud.asendatud;
set puud.asendatud;
jykoodnum = input(jykood, 8.);
run;
proc sql noprint;
create table puud.asendatuduus as
select a.*, b.ark, b.tarvankp
from  puud.asendatud  as a left join  puud.kogum_erilised as b
on a.jykoodnum = b.ark;
run;
/* Observnig, how many enterprises did not have number of employees in
dataset kogum_erilised */
data puud.asendatuduus;
set puud.asendatuduus;
if tarvankp = . then tootajadpuudu = 1;
run;
proc freq data = puud.asendatuduus;
tables tootajadpuudu;
run;
/* Creating variable tegevusala, which represents field of acivity */
data puud.asendatuduus;
set puud.asendatuduus;
tegevusala = SUBSTR(L51_60_1,1,3);
run;
/* Before imputing components are replaced with zeros, when total of
inventories is zero */
data puud.asendatuduus;
```

```
set puud.asendatuduus;
array polevarusid L13_10_1 L13_20_1 L13_30_1  L13_40_1  L13_50_1 ;
do over polevarusid;
if Bi_60_1 = 0 then polevarusid = 0;
end;
run;
/* Same thing at the beginning of the year */
data puud.asendatuduus;
set puud.asendatuduus;
array polevarusid L13_10_2  L13_20_2  L13_30_2  L13_40_2  L13_50_2;
do over polevarusid;
if Bi_60_2_uus = 0 then polevarusid = 0;
end;
run;
/* Calculating the sum of components of inventories at the beginning and in
the end of the year (lisasummlopp ja lisasummalagus)*/
data puud.asendatuduus;
set puud.asendatuduus;
lisasummalopp = sum(L13_10_1, L13_20_1, L13_30_1, L13_40_1, L13_50_1);
lisasummaalgus= sum(L13_10_2, L13_20_2, L13_30_2, L13_40_2, L13_50_2);
run;
/* Cheking for observations, where Bi_60_1 doesn't equal to L13_60_1 at the
time both observed */
data puud.asendatuduus;
set puud.asendatuduus;
if Bi_60_1 ~= . and L13_60_1 ~= . and Bi_60_1 ~= L13_60_1
then olemaseivorduind_1 = 1;
if Bi_60_2_uus ~= . and L13_60_2 ~= . and Bi_60_2_uus ~= L13_60_2
then olemaseivorduind_2 = 1;
run;
proc freq data = puud.asendatuduus;
tables olemaseivorduind_1 olemaseivorduind_2; /* there were not any */
run;
/* Substituting all L13_60_1 values with Bi_60_1 and L13_60_2 with
Bi_60_2_uus */
data puud.asendatuduus;
set puud.asendatuduus;
if L13_60_1 = . then L13_60_1 = Bi_60_1;
if L13_60_1 = . then L13_60_2 = Bi_60_2_uus;
run;
/* Comparing lisasummalopp and Bi_60_1 */
data puud.asendatuduus;
set puud.asendatuduus;
if Bi_60_1 = lisasummalopp
then lopuind = 1;
else lopuind = 0;
run;
/* Comparing lisasummaalgus and Bi_60_2_uus */
data puud.asendatuduus;
set puud.asendatuduus;
if  Bi_60_2_uus = lisasummaalgus
then alguseind = 1;
else alguseind = 0;
run;
proc freq data = puud.asendatuduus;
tables lopuind alguseind;
run;
/* Substituting missing values, where sums are equal with zeros, no need
for imputing those */
data puud.asendatuduus;
set puud.asendatuduus;
```

```
array klapib L13_10_2  L13_20_2  L13_30_2  L13_40_2  L13_50_2;
do over klapib;
if alguseind =1 and klapib= . then klapib = 0;
end;
run;
/* Substituting missing values similarly at the end of the year */
data puud.asendatuduus;
set puud.asendatuduus;
array klapib L13_10_1  L13_20_1  L13_30_1  L13_40_1  L13_50_1;
do over klapib;
if lopuind =1 and klapib= . then klapib = 0;
end;
run;
/* Checking whether some variables in balance sheet have missing values */
/* Replacing missing values of Bi_190_1 with Bi_590_1 */
data puud.asendatuduus;
set puud.asendatuduus;
if bi_190_1 = . then Bi_190_1ind = 1;
if bi_590_1 = . then Bi_590_1ind = 1;
run;
proc freq data = puud.asendatuduus;
tables Bi_190_1ind Bi_590_1ind;
run;
data puud.asendatuduus;
set puud.asendatuduus;
if Bi_190_1 ~= Bi_590_1 and Bi_190_1 =.
then Bi_190_1 = Bi_590_1;
else if Bi_190_1 ~= Bi_590_1 and Bi_590_1 =.
then Bi_590_1 = Bi_190_1;
run;
/* Observing the missing at the same time in Bi_190_2 and Bi_590_2*/
%missingPattern (datain = puud.asendatuduus,
                         varlist = Bi_190_2 Bi_590_2,
                         missPattern1 = 'TRUE',
                         dataout1 = puud.mis1asendatuduus);
/* Observing missing */
%missingPattern (datain = puud.asendatuduus,
                         varlist = Bi_190_2  Bi_590_2 B,
                         missPattern2 = 'TRUE',
                         dataout2 = mis2asendatuduus);
/* Replacing missing values of Bi_190_2 with Bi_590_2 -ga */
data puud.asendatuduus;
set puud.asendatuduus;
if Bi_190_2~= Bi_590_2 and Bi_190_2 =.
then Bi_190_2 = Bi_590_2;
else if Bi_190_2 ~= Bi_590_2 and Bi_590_2 =.
then Bi_590_2 = Bi_190_2;
run;
/* Replacing still missing values from previous year if possible */
proc sql;
create table puud.asendatuduus1 as
select a.*, coalesce (Bi_190_2, Bi_190_2_10) as bi_190_2_uus
from puud.asendatuduus as a left join puud.kymme as b
on a.jykood = b.jykood10;
run;
proc sql;
create table puud.asendatuduus2 as
select a.*, coalesce (Bi_590_2, Bi_590_2_10) as bi_590_2_uus
from puud.asendatuduus1 as a left join puud.kymme as b
on a.jykood = b.jykood10;
run;
```

```
/* Watching results */
%missingPattern (datain = puud.asendatuduus2,
                    varlist = Bi_190_2 Bi_190_2_uus Bi_590_2 Bi_590_2_uus,
                    missPattern2 = 'TRUE',
                    dataout2 = misasendatuduus2);
/*Observing condition number 3 */
data puud.asendatuduus2;
set puud.asendatuduus2;
if sum(Ka_70_1, Ka_90_1) = sum(L13_20_1,L13_30_1) - sum(L13_20_2,L13_30_2 )
then jaakideind = 1;
else jaakideind = 0;
run;
/* Changing field of activity to numerical and creating indicator to
restrict imputing Ka_70_1*/
data puud.asendatuduus2;
set puud.asendatuduus2;
emtak1=  substr(L51_60_1,1,1);
emtak2=  substr(L51_60_1,1,2);
tegnum = input(tegevusala, 8.);
emtak1num= input(emtak1, 8.);
emtak2num= input(emtak2, 8.);
run;
data puud.asendatuduus2;
set puud.asendatuduus2;
if emtak1 in (0,1,2,3,4) or emtak2 in
(0,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31
                       32,33,34,35,58,59,60,61,62,71,72,80,81,82) or
tegnum in (452,732,952)
then emtakind = 1;
else emtakind = 0;
run;
```

Testing MCMC

```
/* Finding max, min and mean for boundaries and initial values*/
proc means data = puud.asendatuduus2 max min mean;
var Ka_360_1 Ka_50_1 Ka_70_1 Ka_90_1 L13_10_1 L13_20_1 L13_30_1 L13_40_1
L13_50_1
L13_10_2 L13_20_2 L13_30_2 L13_40_2 L13_50_2 L51_30_1 tarvankp;
run;
/*Testing MCMC, iteration are increased*/ /* when Imputing until monotone
missing mcmc statement is added impute = monotone */
proc mi data = puud.asendatuduus2 seed=501462
mu0 =41044 592954 9027 26068 17550 7300 8477 33279 2762 14600 6313 7183
28047 2466 91 7
out = mcmc2;
em maxiter = 500 converge = 1E-3 ;
mcmc nbiter= 500;
var Ka_360_1 Ka_50_1 Ka_70_1 Ka_90_1 L13_10_1 L13_20_1 L13_30_1 L13_40_1
L13_50_1
L13_10_2 L13_20_2 L13_30_2 L13_40_2 L13_50_2 L51_30_1 TARVANKP Bi_60_1
Bi_60_2_uus Bi_590_1 Bi_590_2_uus L51_50_1;
run;
/* Adding variable suurusgrupp, which represents size of the enterprise */
data puud.tootajad;
set puud.tootajad;
if tarvankp = 1 then suurusgrupp = 1;
if tarvankp >= 2 and tarvankp <= 9 then suurusgrupp = 2;
if tarvankp >= 10 and tarvankp <= 20 then suurusgrupp = 3;
if tarvankp > 20 then suurusgrupp = 4;
run;
/* Imputing every size group separately, already increased iteartions and
added boundaries*/
proc mi data = puud.tootajad seed=501462 maximum = 79911600 1489512000
998877 29656081 178484400 18366962 17040936 70146000 6440400 135349200
19161933 10016611 26776800 6678247 100 minimum = -17740800 -776040 -275443
-4361803 0 0 0 0 -9600 0 0 0 0 -67200 0 1
mu0 =  41044 592954 9027 26068 16557 6659 7932 31535 2556 13931 5939 6806
26841 2215 91
out = puud.mcmcsuurus2;
em maxiter = 400 converge = 1E-4 ;
by suurusgrupp;
mcmc  nbiter=400 ;
var  Ka_360_1 Ka_50_1 Ka_70_1 Ka_90_1 L13_10_1 L13_20_1 L13_30_1 L13_40_1
L13_50_1
L13_10_2 L13_20_2 L13_30_2 L13_40_2 L13_50_2 L51_30_1 TARVANKP Bi_60_1
Bi_60_2_uus Bi_590_1 Bi_590_2_uus L51_50_1;
run;
/* Imputing one-by-one*/
%missingPattern (datain = puud.asendatuduus2,
                        missPattern2 = 'TRUE',
                        dataout2 = korraks);
/* Imptuting Ka_360_1, all the others similarly, maksimum, minimum, seed
hanged and previously imputed variable added to var if it was imputed
properly */
proc mi data = puud.asendatuduus2 nimpute = 1 seed=507262 maximum =
79911600 minimum = -17740800
mu0= 41044 out=puud.impI;
em;
mcmc ;
var  Ka_360_1 tegnum Bi_60_1 Bi_60_2_uus Bi_190_1 Bi_190_2_uus L51_50_1;
run;
```

Using SRMI

```
options set = SRCLIB "\\haug\statgroups\Metoodika\MST\Statistika
tarkvarad\IVEware"
sasautos = ('!SRCLIB' sasautos) mautosource;
data _null_;
  infile datalines;
  filename setup "\\haug\statgroups\Metoodika\MST\Statistika
tarkvarad\IVEware\TEMP\impute.set";
  file setup;
  input;
  put _infile_;
datalines4;
  title Multiple imputation;
  datain          puud.asendatuduus2;
  dataout         puud.imputedmaailma10 all;
  default         drop;
  continuous      Bi_60_1 Bi_60_2_uus Bi_190_1 Bi_190_2_uus Bi_590_1
                  Bi_590_2_uus Ka_50_1 Ka_360_1 L51_50_1 L51_30_1 L13_10_1
                  L13_10_2 L13_20_1 L13_20_2 L13_30_1 L13_30_2 L13_40_1
                  L13_40_2 L13_50_1 L13_50_2 Ka_90_1 Ka_70_1;
  count           tarvankp;
  transfer        jykood TI_valuuta TYYP LAADIMINE VERSIOON MAJ_ALGUS
                  MAJ_LOPP L51_60_1 L51_20_1 emtakind L13_60_1 L13_60_2;
  count           tarvankp;
  bounds          Ka_360_1 (> -17740800, < 79911600)
```

```
                  Ka_50_1 (> -776040, < 1489512000)
                  Ka_70_1 (> -275443, < 998877)
                  Ka_90_1 (> -4361803 < 29656081)
                  L13_10_1 (> 0, < 178484400)
                  L13_20_1 (> 0, < 18366962)
                  L13_30_1 (> 0, < 17040936)
                  L13_40_1 (> 0, < 70146000)
                  L13_50_1 (> -9600, <6440400)
                  L13_10_2 (> 0, < 135349200)
                  L13_20_2 (> 0, < 19161933)
                  L13_30_2 (> 0, < 10016611)
                  L13_40_2 (> 0, < 26776800)
                  L13_50_2 (> -67200, < 6678247)
                  L51_30_1 (> 0, < 100)
                  TARVANKP (>= 1, < 3735) ;
  restrict         Ka_90_1 (emtakind = 1)
                  Ka_70_1 (emtak1num = 0,1);
  MINRSQD         .01;
  iterations      10;
  multiples        10;
  seed             5876315;
  perturb         Sir;
  print           DETAILS;
  run;
;;;;

%impute(name=impute, dir='\\haug\statgroups\Metoodika\MST\Statistika
tarkvarad\IVEware\TEMP');

/* Separating all the multiples from each other, similarly all ten of
them*/
data puud.imputedmaailma10mult3;
set puud.imputedmaailma10;
where _mult_ = 3;
rename L13_10_1 = L13_10_1_3 L13_20_1 = L13_20_1_3 L13_30_1 = L13_30_1_3
L13_40_1 = L13_40_1_3 L13_50_1 = L13_50_1_3
         L13_10_2 = L13_10_2_3 L13_20_2 = L13_20_2_3 L13_30_2 = L13_30_2_3
L13_40_2 = L13_40_2_3 L13_50_2 = L13_50_2_3
         Ka_360_1 = Ka_360_1_3 Ka_50_1  = Ka_50_1_3  Ka_70_1  = Ka_70_1_3
Ka_90_1 = Ka_90_1_3  L51_30_1 = L51_30_1_3
         TARVANKP = tarvankp_3      ;
run;
/* Merging together */
data puud.imputedmaailma10kokku;
merge puud.imputedmaailma10mult1 puud.imputedmaailma10mult2
puud.imputedmaailma10mult3 puud.imputedmaailma10mult4
puud.imputedmaailma10mult5 puud.imputedmaailma10mult6
puud.imputedmaailma10mult7 puud.imputedmaailma10mult8
puud.imputedmaailma10mult9 puud.imputedmaailma10mult10;
by jykood;
run;
/* Averaging, all other variables similarly */
data puud.imputedmaailma10kokku;
set puud.imputedmaailma10kokku;
L13_10_1_kesk = mean(L13_10_1_1, L13_10_1_2, L13_10_1_3, L13_10_1_4,
L13_10_1_5, L13_10_1_6, L13_10_1_7, L13_10_1_8, L13_10_1_9, L13_10_1_10);
run;
/* Observing conditions and calculating sums and differences */
data puud.imputedmaailma10kokku;
set puud.imputedmaailma10kokku;
```

```
imputeeritudkokku1 = sum(L13_10_1_kesk, L13_20_1_kesk, L13_30_1_kesk,
L13_40_1_kesk, L13_50_1_kesk);
imputeeritudkokku2 = sum(L13_10_2_kesk, L13_20_2_kesk, L13_30_2_kesk,
L13_40_2_kesk, L13_50_2_kesk);
vahe1 = Bi_60_1 - imputeeritudkokku1;
vahe2 = Bi_60_2_uus - imputeeritudkokku2;
vasakpool = sum(Ka_70_1_kesk, Ka_90_1_kesk);
parempool = L13_20_1_kesk + L13_30_1_kesk - L13_20_2_kesk - L13_30_2_kesk;
vahe3 = vasakpool - parempool;
run;
/* Creating coefficient for condition one */
data tegur;
set puud.imputedmaailma10kokku;
if imputeeritudkokku1 = Bi_60_1  then tegur1 = 1;
else tegur1 = Bi_60_1 / imputeeritudkokku1;
if imputeeritudkokku2 = Bi_60_2_uus  then tegur2 = 1;
else tegur2 = Bi_60_2_uus / imputeeritudkokku2;
L13_10_1_tegur = L13_10_1_kesk*tegur1;
L13_10_2_tegur = L13_10_2_kesk*tegur2;
L13_20_1_tegur = L13_20_1_kesk*tegur1;
L13_20_2_tegur = L13_20_2_kesk*tegur2;
L13_30_1_tegur = L13_30_1_kesk*tegur1;
L13_30_2_tegur = L13_30_2_kesk*tegur2;
L13_40_1_tegur = L13_40_1_kesk*tegur1;
L13_40_2_tegur = L13_40_2_kesk*tegur2;
L13_50_1_tegur = L13_50_1_kesk*tegur1;
L13_50_2_tegur = L13_50_2_kesk*tegur2;
summateg1 = sum(L13_10_1_tegur, L13_20_1_tegur, L13_30_1_tegur,
L13_40_1_tegur, L13_50_1_tegur);
summateg2 = sum(L13_10_2_tegur, L13_20_2_tegur, L13_30_2_tegur,
L13_40_2_tegur, L13_50_2_tegur);
vaheteg1 = Bi_60_1-summateg1;
vaheteg2 = Bi_60_2_uus-summateg2;
run;
/* Fixing up third condition*/
data kolmastingimus;
set tegur;
vasakpool = Ka_70_1_kesk + Ka_90_1_kesk;
parempool = L13_20_1_tegur + L13_30_1_tegur - L13_20_2_tegur -
L13_30_2_tegur;
if vasakpool = 0 then tegur3 = 0;
else tegur3 =  parempool/vasakpool;
Ka_70_1_tegur = Ka_70_1_kesk*tegur3;
Ka_90_1_tegur = Ka_90_1_kesk*tegur3;
vasakpoolteg = Ka_90_1_tegur + Ka_70_1_tegur;
vaheteg3 = vasakpoolteg - parempool;
run;
/* Calculating coeffincient for the third condition*/
data kolmastingimus1;
set kolmastingimus;
if vasakpoolteg = 0 and parempool > 0 then do ; /* parempool < 0 did not
occur*/
      L13_20_1_teguruus = L13_20_1_tegur -(abs(vaheteg3)/4);
      L13_30_1_teguruus = L13_30_1_tegur -(abs(vaheteg3)/4);
      L13_10_1_teguruus = L13_10_1_tegur +(abs(vaheteg3)/2/3);
      L13_40_1_teguruus = L13_40_1_tegur +(abs(vaheteg3)/2/3);
      L13_50_1_teguruus = L13_50_1_tegur +(abs(vaheteg3)/2/3);
      L13_20_2_teguruus = L13_20_2_tegur +(abs(vaheteg3)/4);
      L13_30_2_teguruus = L13_30_2_tegur +(abs(vaheteg3)/4);
      L13_10_2_teguruus = L13_10_2_tegur -(abs(vaheteg3)/2/3);
      L13_40_2_teguruus = L13_40_2_tegur -(abs(vaheteg3)/2/3);
      L13_50_2_teguruus = L13_50_2_tegur -(abs(vaheteg3)/2/3);
summateg1uus = sum(L13_10_1_teguruus, L13_20_1_teguruus, L13_30_1_teguruus,
L13_40_1_teguruus, L13_50_1_teguruus);
summateg2uus = sum(L13_10_2_teguruus, L13_20_2_teguruus, L13_30_2_teguruus,
L13_40_2_teguruus, L13_50_2_teguruus);
vaheteg1uus = Bi_60_1-summateg1uus; /*used for checking whether second
condition stayed met */
vaheteg2uus = Bi_60_2_uus-summateg2uus;
run;
/* recalculating the third condition and observing the results */
data kolmastingimus1;
set kolmastingimus1;
parempooluus = sum(L13_20_1_teguruus, L13_30_1_teguruus) -
sum(L13_20_2_teguruus , L13_30_2_teguruus);
vaheteg3uus = vasakpoolteg - parempooluus;
run;
proc means data = kolmastingimus1 max min mean;
var vaheteg3uus;
run;
```

Simulation

```
/* second dataset, 10% set to missing */
/* similarly 30% and 50% and then repeated with the third dataset */
data puud.vordlus1_10;
set puud.vordlus1; /* vordlus1 has Bi_60_1 equal to the sum of components
and Bi_60_2_uus */

suvaline = uniform(22772);
if suvaline < 0.1 then kustutada = 1;
else kustutada = 2;
run;
/* deleting values */
data puud.vordlus1_10;
set puud.vordlus1_10;
array puudu L13_10_1 L13_10_2 L13_20_1 L13_20_2 L13_30_1 L13_30_2 L13_40_1
L13_40_2 L13_50_1 L13_50_2;
do over puudu;
if kustutada = 1 then puudu = .;
end;
run;
```

IVEware code similar to imputing puud.imputedmaailma10.

```
/* creating variables to observe met of the conditions, similarly for all
of the 6 datasets used in simulation */
data vordlus2_10S; /* third dataset, S- simulated */
set puud.vordlus2_10S;
vahe1 = Bi_60_1 - sum(L13_10_1, L13_20_1, L13_30_1, L13_40_1, L13_50_1);
vahe2= Bi_60_2_uus - sum(L13_10_2, L13_20_2, L13_30_2, L13_40_2, L13_50_2);
vahe3= Ka_70_1+Ka_90_1-L13_20_1-L13_30_1+L13_20_2+L13_30_2;
run;
proc means data = vordlus2_10S mean max min;
var vahe1 vahe2 vahe3 Bi_60_1 Bi_60_2_uus Ka_360_1 Ka_50_1 Ka_70_1 Ka_90_1
L13_10_1 L13_20_1 L13_30_1 L13_40_1 L13_50_1
L13_10_2 L13_20_2 L13_30_2 L13_40_2 L13_50_2 L51_30_1 tarvankp;
run;
```

Non-exclusive licence to reproduce thesis and make thesis public

I, Cliona Georgia Dalberg, (date of birth: 16 August 1991),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu,, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Imputation of inventories in Estonian Commercial Register

supervised by Mare Vähi and Ebu Tamm,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu/Tallinn 29.04.2013

Retsensioon

Cliona Georgia Dalbergi bakalaureusetööle

Varude imputeerimine Eesti Äriregistris 2011. aastal

(*Imputation of inventories in Estonian Commercial Register*)

Cliona Georgia Dalberg uurib oma töös meetodeid puuduvate väärtuste imputeerimiseks varude osas 2011. aastal.

Retsenseeritava töö maht on 44 lehekülge, millest 10 lehekülge on lisas toodud SASi kood. Töö on ingliskeelne. Töö koosneb järgmistest olulistest osadest.

- ***The Missingness Mechanism*** (4 lk), kus antakse lühiülevaade andmekao tekke mehhanismist.
- ***Imputation Methods*** (10 lk), kus käsitletakse selliseid meetodeid nagu monotoonne regressioon, MCMC ja järjestikune regressioon.
- **Rakendus reaalse andmestiku korral** (16 lk), mis koosneb kolmest peatükist ja kus kõigepealt kirjeldatakse andmeid, andmekao tekkimist ja kahe meetodi rakendust – MCMC ja järjestikune regressioon.

Töö on kirjutatud projekti „Metoodika väljatöötamine statistika tegemiseks kombineeritud administratiivsete andmeallikate ja uuringute baasil“ raames, mille tellis Eesti Statistikaameti käest Eurostat.

Kahtlemata, on läbitehtud töömaht väga suur, on nähtud palju vaeva nii andmete teisendamise ja simuleerimisega, kui ka teoorias vajalike artiklite uurimisega (viidete loetelu sisaldab 20 nimetust). Töös kasutatud SASi makrod puuduvate mustrite genereerimiseks ja järjestikuse regressiooni kasutamiseks on retsensendi jaoks uued, ilmselt ka Cliona Georgia jaoks samuti, nende selgeks tegemine on ka lisapluss tudengile.

Töös kasutatud inglise keel on ilmselt tingitud Eurostati tellimusega. Siiski enne Eurostatile esitamist soovitaksin tööd uuesti läbi lugeda, kuna esineb grammatikavigu.

Tööd lugedes tekkisid järgmised märkused:

- Kirjanduse loetelule vastab termin *References,* mitte *Citations*.
- Kontrollida veel üks kord kasutatud viiteid kirjandusele – kord puudub aastaarv kirjanduse loetelus, kord puudub viidatav allikas loetelust tervinisti, mõnikord on viidatud erinevates stiilides. Näiteid vigadest on lk. 5 (Borman), lk. 6 (Hippel), lk. 11 (Prior Distributions), lk 12 (Hoewyk and Solenberger), lk. 26 (Montgomery), jne.
- Lk. 6 kasutatud meetodite loetelus viimane on Monte Carlo. Eelnevad 7 on ilusti eelnevalt ära kirjeldatud, viimase kohta pole aga midagi. Võiks ikka samuti lisada.
- Graafikud 1 ja 2 on raskesti loetavad. Kohati kasutatud värv ei ole õnnestunud, lisaks legendis kasutatud meetodite tähistused ei ole eespool defineeritud, lugeja peab ise suutma aru saama.
- Lk. 9 Kordajate vektor – transponeerimismärk on puudu.
- Lk. 15 lause „*However, dealing with large datasest and complicated models requires fast computer and a lot of memory. (Schafer, 1997)*“. Seda lauset oleks võinud ikka kriitilise pilguga üle vaatama, arvestades, et praeguseks aastaarvuks on 2013.
- Makro %*impute* pole seletatud töös, ometi on tal palju võimalusi ja kasutatavaid argumente. Lisatud koodis pole ka vastavaid kommentaare leidnud, lisaks on seal kasutatud topeltrida *count*, mis tekitab veel rohkem segadust.

Siiski need märkused on retsensendi subjektiivne arvamus ja töö lugemist see otseselt ei seganud. Tööd lugedes tekkisid järgmised küsimused:

1. Lk. 3 Lause „*Imputing also minimizes bias*..“ Mida sellega on tahetud öelda? Kui näiteks, on vastanute hulgas rohkem naissoost pensionäre ja mittevastanud on enamasti noored mehed, siis imputeerimine vanemate inimeste põhjal ei vähenda nihet absoluutselt.

2. Lk. 5 Lause „*In multiple imputation first of all a model is fitted, then plausible values (are?) generated whis is (are?)…“.* Mida taheti öelda selle lausega? Mis väärtustest on juttu?

3. Lk. 5 on öeldud, et kasutati 8 erinevat meetodit ja erinevat tarkvara andmete analüüsimiseks. Miks erinevat tarkvara?

4. Lk. 5 ja 6 on vaadeldud ühte uuritavat tunnust. Kas sel juhul ei peaks *pairwise deletion* ja *Listwise deletion* andma ühe ja sama tulemuse?

5. Lk. 8 seletatakse regressioonijärgset omistust. Mida taheti öelda lauses „*With this model, a new model is drawn and is used to impute missing values.*“
6. Lk. 9 *three steps*. Kust järsku tekkisid näitajad ? Mida nad tähendavad ja mille jaoks seal olulised on?

7. Lk. 16, andmete kirjeldus. Millised tunnused on võetud *X* tunnusteks, ehk kõikseteks andmeteks.

8. Kas puuduvateks andmeteks on terved ettevõtted (*Unit nonresponse*) või uuritavate tunnuste üksikud väärtused (*Item nonresponse*), sest just esimene on sagedasti esinev olukord.

9. Lk. 22-23. Kas kasutatavates tabelites on leitud keskmiste vahe? Seda ei loe tekstist välja. Kas poleks parem kasutada suhtelist vahet?

Vaatamata märkustele ja küsimustele vastab Cliona Georgia Dalbergi töö bakalaureusetöödele esitatavatele nõuetele ja üliõpilane väärib loodusteaduse bakalaureusekraadi.

Natalja Lepik

Lektor, TÜ MSI

7. juuni 2013

Vastused retsensioonis esitatud küsimustele (retsensendile vastatud suuliselt)

1. Kui vastanute ning mittevastanute grupid on omavahel sarnased ning teineteisest oluliselt erinevad, siis tõepoolest ei pruugi nihe väheneda. Üldiselt on mitmeses imputeerimise meetodid disainitud nihet vähendama. Konkreetse väite aluseks oli võetud lause artiklist "Dealing with missing data " (Scheffer lk 156) ning seda toetavad veel „SAS Macros Useful in Imputing Missing Survey Data" (Carlson, Cox, Bandeh lk 1089) ning „Imputation Explanation. Find the best way to handle missing data in surveys " (Allen, Seaman).
2. In multiple imputation first of all a model is fitted, then plausible values generated which is followed by analyzing each completed data set …

Näitena vaatame lihtsat lineaarset regressiooni

$$Y = \beta_0 + \beta_1 Z + \varepsilon$$

Regressioonparamteerid

$$\beta = (\beta_0, \beta_1)^T$$

hinnatakse vähimruutude meetodil täielikult vaadeldud tunnuste põhjal:

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T,$$

kus parameetrite hinnangu dispersioon on

$$\sigma^2 (Z^T Z)^{-1}$$

ning

$$\sigma^2 = var(\varepsilon)\ ja\ Z = (1, z)$$

Z on vektor, kuhu on lisatud ühtedest koosnev vabaliikmete veerg.

Seejärel tõmmatakse juhuslik suurus u jaotusest

$$\chi^2_{n-2}$$

ning leitakse uue “plausible” mudeli dispersioon.

$$\tilde{\sigma}^2 = \frac{(n-2)\hat{\sigma}^2}{u}$$

Seejärel hinnatakse “plausible” mudeli parameetrid jaotusest:

$$\tilde{\beta} \sim N\left\{\hat{\beta},\ \tilde{\sigma}^{\,2}\ (Z^T Z)^{-1}\right\}$$

“Plausible” mudel omandab kuju

$$Y_{mis} \sim \tilde{\beta}_0 + \ \tilde{\beta}_1 Z_{(Y,mis)} + \varepsilon,$$

kus

$$Z_{(Y,mis)}$$

on Z-i väärtuste vektori, mis on seotus nende kirjetega, kus Y-l on puudumised ning

epsilon on juhuslik viga jaotusest

$$N(0, \tilde{\sigma}^2).$$

Kokkuvõtvalt on selle lausega tahetud öelda seda, et lisaks esialgsele vaadeldud andmete pealt saadud mudelile, luuakse juhuslikkuse abil ka teisi natuke erinevaid mudeleid (plausible models), et saaks teha imputeerimist mitu korda ning tulemusi pärast keskmistada. Kui juhuslikkuse sammu ei oleks, saaksime me igakord sama mudeli ning seega teeksime sisuliselt ühekordset imputeerimist (single imputation).

3. Tõenäoliselt tahtis artikli autor (Judy Scheffer) testida erinevaid olemasolevaid imputeerimsimeetodeid ning ühes tarkvaras ei leidunud piisavalt palju võimalusi.

4. Peaksid andma küll sama tulemuse. Jääb oletada, et autor kasutas *pairwise deletion'* i korral ka mõnes teises tunnuses puudumised lisatud.

5. Kattub küsimusega 2, mõeldud on "plausible" mudelit.

6. Nagu öeldud leheküljel 9:

   Juhusliku vea dispersiooni hinnang j-l tunnusel

   $$\hat{\sigma}^2{}_j$$

   Maatriks

   $$V_j = (X^T X)^{-1}$$

   ning dispersioonimaatriks

   $$\hat{\sigma}^2{}_j V_j.$$

   Kõik need on vajalikud, et saaks konstrueerida uue mudeli.

7. X - tunnusteks ehk kõikseteks andmeteks olid esialgu võetud varud kokku, kokku varad ning kokku kohustused ja omakapital nii aasta alguses kui lõpus (kõik täielikult vaadeldud tunnused v.a. taustunnused). Seejärel lisandus X tunnuseks esimesena imputeeritud aruandeaasta kasum, kuna sellel tunnusel oli kõige vähem puudumisi. Sama põhimõttega lisati järjest tunnuseid vastavalt tabelile 1 lk 21.

8. Item nonresponse. Kuna bilansiaruanne oli kohustuslik, oli kõikide ettevõtete kohta midagi teada.

9. Arvud tabelis vastavad lk 22 toodud valemile mille järgi lahutati pärast imputeerimist arvutatud tunnuse keskmisest tõene keskmine ning jagati seejärel tõese keskmisega. Viimases tabeli teas on keskmiste keskmised (veergude keskmised).

Continuation

First of all some corrections were made. It became clear that due to the fact that the profit and loss account was compulsory, missing values in variables Ka_70_1, Ka_50_1 and Ka_360_1 also were replaced with zeros.

The same rule did not apply to Ka_90_1, which was thought previously. Some background variables were added, to discern balance sheet schemes (two possibilities) and missing values were replaced with zeros only when the enterprise had filled the first balance sheet scheme.

In addition it turned out that no negative values were allowed in notes and inventories of enterprises including negative values were deleted.

Another simulation was carried out because there was a reason to believe that enterprises with similar field of activity have similar division of inventories and maybe imputing based only on enterprises from same field offers more accurate results.

Groups of enterprises were formed as follows:

1. Modules B – C (according to Statistical Classification of Economic Activities in the European Community (NACE Rev.2))
2. D – E
3. F 411
4. F 412 – F 439
5. G 452
6. G 451 and H 453 – G 478
7. H
8. I
9. L
10. J and M - S
11. K
12. T
13. U
14. A

Groups 12 and 14 did not have any enterprises, group 13 included only fully observed enterprises.

Table 7. Missingness in groups

Group	Inventories missing
1	502
2	27
3	14
4	497
5	59
6	688
7	274
8	116
9	424
10	1740
11	75
13	0

Basis of this simulation was the dataset, where inventories total in balance sheet was equal to the sum of the components of the inventories in notes at the beginning and at end of the year. That dataset had 46, 802 observations. Using uniform distribution inventories of 4,400 enterprises were deleted randomly, which was approximately 9.4%. After that IVEware was used with 5 imputations performed and then datasets were averaged. Similarly as before 4,416 observations were set to missing randomly and then IVEware was used separately in each group, then averaged and joined together.

Calculating coefficients to ensure that required conditions were met was also changed.

Firstly variable $tegur$ was calculated, which was assigned values as follows:

$$Left\ side\ (LS) = \ Ka\_70\_1\ +\ Ka\_90\_1$$

$$Rigth\ side\ (LS) = L13\_20\_1\ +\ L13\_30\_1 - L13\_20\_2 - L13\_30\_2$$

$$tegur = \left\{ \begin{array}{c} \frac{LS}{RS} \qquad ,when\ RS\ \neq 0 \\ 0\ \ , \qquad when\ LS = RS = 0 \\ .\ ,when\ RS = 0\ and\ LS\ \neq 0 \end{array} \right\}$$

If $tegur$ was assigned value **.** (missing), then all components on the right side were attributed one fourth of the difference between left and right side. In other cases components of the right side were multiplied with $tegur$. After that two other conditions were taken care of :

$$Bi\_60\_1 \quad = L13\_10\_1 + L13\_20\_1 + L13\_30\_1 + L13\_40\_1 + L13\_50\_1$$

$$Bi\_60\_2 \quad = L13\_10\_2 + L13\_20\_2 + L13\_30\_2 + L13\_40\_2 + L13\_50\_2$$

$Vahe1$ was calculated as difference between sum on the right and total of inventories in the left. Then $L13\_10\_1$, $L13\_40\_1$ and $L13\_50\_1$ were added/subtracted one third of the difference so that afterwards difference between sides was zero.

Assuring that necessary conditions are met in this way comparing to previous one, has advantage, that none of the correct values in balance sheet or profit and loss account were changed.

In table 8 an example of simulation result is displayed. First column states the variable, the second field of activity, third how many enterprises there were, fourth size of enterprise by number of employees, “Real sum” shows the sum in fully observed dataset, “Imputed sum “ shows the sum after deleting and imputing, “difference 1” expresses the difference between previous two columns and “relation 1” expresses quotient of real sum and imputed sum. Last three columns are calculated similarly to previous three. Last row shows the mean value of column.

Table 8. Example of simulation results

	Field	How many						Groups	Difference	
Variable	code	enterprises	Employees	Real sum	Imputed sum	Difference 1	Relation 1	separately	2	Relation 2
L13_10_1	10	154	1-9	1,985,299	2,389,296	-403,997	0.83	5,315,974	-3330675	0.37
L13_20_1	10	154	1-9	329,539	240,769	88,769	1.37	364,031	-34492	0.91
L13_30_1	10	154	1-9	418,525	55,522	363,002	7.54	162,393	256131	2.58
L13_40_1	10	154	1-9	449,428	1,578,225	-1,128,797	0.28	-508,054	957482	-0.88
L13_50_1	10	154	1-9	149,553	-931,469	1,081,022	-0.16	-2,002,001	2151554	-0.07
L13_10_2	10	154	1-9	1,318,893	1,626,251	-307,358	0.81	4,476,800	-3157907	0.29
L13_20_2	10	154	1-9	467,570	397,875	69,694	1.18	517,737	-50167	0.9
L13_30_2	10	154	1-9	564,921	137,200	427,720	4.12	247,471	317449	2.28
L13_40_2	10	154	1-9	442,630	1,285,959	-843,329	0.34	-478,468	921098	-0.93
L13_50_2	10	154	1-9	90,575	-562,697	653,272	-0.16	-1,878,951	1969526	-0.05
L13_10_1	10	33	10-19	1,118,498	962,361	156,136	1.16	1,223,178	-104680	0.91
L13_20_1	10	33	10-19	131,546	184,714	-53,168	0.71	130,715	831	1.01
L13_30_1	10	33	10-19	315,726	415,894	-100,168	0.76	386,324	-70598	0.82
L13_40_1	10	33	10-19	350,560	694,534	-343,974	0.5	302,529	48030	1.16
L13_50_1	10	33	10-19	25,470	-315,704	341,174	-0.08	-100,947	126417	-0.25
L13_10_2	10	33	10-19	1,236,247	102,7243	209,003	1.2	1,321,495	-85248	0.94
L13_20_2	10	33	10-19	105,673	160,763	-55,090	0.66	104,583	1090	1.01
L13_30_2	10	33	10-19	338,980	405,901	-66,921	0.84	378,512	-39532	0.9
L13_40_2	10	33	10-19	251,658	554,197	-302,539	0.45	224,631	27026	1.12
L13_50_2	10	33	10-19	27,932	-187,616	215,548	-0.15	-68,732	96664	-0.41
						0.00	1.11		0.00	0.63

It was interesting that in all cases imputed sum was either strongly over or under estimated, but errors in both ways compensated and mean value was always zero or very close to it. Due to that deciding, which way - imputing over all observations or imputing inside groups – is better was based on mean values of relations. The closest the mean was to one, the better it was. If difference between the mean values of relations of different imputation ways was smaller than 0.2, then it was considered as non-important. It became clear that mean of relation was closer to one more often when imputing was done over all observations than done separately in groups.

Also cases where enterprise did not have any inventories and the values should have been zeros, had more accurate results when imputing over all values. When imputing separately in groups there were strongly overestimated values instead of zeros.

Addditional conditions

There were some additional conditions suggested by the accounting team members, which should apply to inventories:

1. If code of field is between 452000 and 453000 and Ka_90_1 = 0;
2. If code of field is between 451000 and 452000 and Ka_90_1 = 0 and number of employees is larger or equal than 10 and less or equal than 19;
3. If code of field is between 453000 and 458000 and Ka_90_1 = 0 and number of employees is larger or equal than 10 and less or equal than 19;
4. If code of field is between 550000 and 557000 and Ka_90_1 = 0 and $L51\_60\_1$ is larger than 350000;
5. If code of field is between 490000 and 550000 and Ka_90_1 = 0;

   Then:

   - $L13\_20\_1$ = 0;
   - $L13\_20\_2$ = 0;
   - $L13\_30\_1$ = 0;
   - $L13\_30\_2$ = 0;

6. If $L51\_60\_1$ is smaller than 45000 or $L51\_60\_1$ is between 45200 and 45300 or $L51\_60\_1$ is larger than 48000;

Then:

- $L13\_40\_1$ = 0;
- $L\_13\_40\_2$ = 0;

7. If $L51\_60\_1$ is between 4500 and 45200 or between 45300 and 48000 and code of field is between 451000 and 452000 or between 453000 and 480000 and number of employees is bigger than 1 and smaller than 9;

   Then:

   - $L13\_40\_1$ = Bi_60_1;
   - $L13\_40\_2$ = $Bi\_60\_2\_uus$;

Before practicing additional conditions, fully observed dataset had 46,802 observations and after conditions were put into practice, there were 41,251 observations, which was contradictory. That was the reason, why all of the conditions were monitored separately.

Table 9. Comparison before and after practicing the first additional condition

	Missing before	Missing after.	Difference	Zeros	Zeros after	Difference
$L13\_20\_1$	8,858	8,658	200	46,953	47,155	202
$L13\_20\_2$	8,587	8,401	186	47,273	47,462	189
$L13\_30\_1$	8,828	8,628	200	46,523	46,726	203
$L13\_30\_2$	8,556	8,370	186	46,847	47,035	188

First additional condition changed accordingly 200- 186- 200- 186 missing values to zeros. Comparing number of changed observations to number of zeros, it is seen that accordingly 2- 3- 3- 2 observations different from zeros were also changed.

Table10. Comparison before and after practicing the second additional condition

	Missing before	Missing after.	Difference	Zeros	Zeros after	Difference
L13_20_1	8,658	8,651	7	47,155	47,162	7
L13_20_2	8,401	8,395	6	47,462	47,468	6
L13_30_1	8,628	8,620	8	46,726	46,734	8
L13_30_2	8,370	8,363	7	47,035	47,042	7

Second additional condition did not change any observations to zeros, which already had some other value.

Table11. Comparison before and after practicing the third additional condition

	Missing before	Missing after.	Difference	Zeros	Zeros after	Difference
L13_20_1	8,651	8,649	2	47,162	47,164	2
L13_20_2	8,395	8,393	2	47,468	47,470	2
L13_30_1	8,620	8,618	2	46,734	46,736	2
L13_30_2	8,363	8,361	2	47,042	47,044	2

Third additional condition changed two missing observations to zeros in each variable. Practicing fourth additional condition did not change anything.

Table12. Comparison before and after practicing the fifth additional condition

	Missing before	Missing after.	Difference	Zeros	Zeros after	Difference
L13_20_1	8,649	8,269	380	47,164	47,552	388
L13_20_2	8,393	8,016	377	47,470	47,853	383
L13_30_1	8,618	8,239	379	46,736	47,130	394
L13_30_2	8,361	7,985	376	47,044	47,430	386

Fifth condition changed accordingly 8- 6- 5- 1 observations to zeros, which were not previously missing.

Table13. Comparison before and after practicing the sixth additional condition

	Missing before	Missing after.	Difference	Zeros	Zeros after	Difference
L13_40_1	8,651	2,483	6,168	39,741	49,824	10,083
L13_40_2	8,375	2,403	5,972	39,741	49,824	10,083

Fifth condition changed accordingly 3,915- 4,111 observations to zeros, which were not previously missing.

Table14. Comparison before and after practicing the seventh additional condition

	Missing before	Missing after.	Difference	Zeros	Zeros after	Difference
L13_40_1	2483	436	2047	49824	49824	0
L13_40_2	2403	416	1987	49824	49824	0

Seventh condition changed accordingly 2,047- 1,987 observations to zeros, which were not previously missing.

As a result there were less completely observed observations and due to that practicing additional conditions were not considered as effective as imputation without them.

Bakalaureusetöö jätk

Cliona Georgia Dalberg

Kokkuvõte

Esmalt parandati bakalureusetöös tekkinud vead - asendati kasumiaruande tunnuste puudumised nullidega, kustutati negatiivseid varusid omavate ettevõtete väärtused ning tehti mõningaid täpsustusi.

Teisena viidi läbi simulatsioon, mis annab ülevaate imputeerimisest täpsemalt - kahekohalise EMTAKi koodi tasemel. Simulatsiooni läbiviimiseks jagati ettevõtted vastavalt tegevusaladele 14 gruppi, millest kahte ei sattunud ühtegi väärtust. Simulatsiooni aluseks võeti täielikult vaadeldud varudega kirjeid sisaldav andmestik, seejärel seati juhuslikult ühtlase jaotuse abil puuduvaks 10% ning teostati 5 kordusega mitmene imputatsioon, kasutades järjestikust regressiooni SASi makros IVEware. Simulatsiooni tehes muudeti võrreldes varasemaga ka nõutavate seoste kehtimise tagamiseks koefitsientide leidmist.

Tulemuseks saadi, et imputeerimine gruppides eraldi ei olnud efektiivsem imputeerimisest üle kõikide kirjete. Huvitavaks osutus see, et kõikides kahekohalise EMTAKi järgi vaadeldud rühmades tuli keskmine summade erinevus tõelisest summast null.

Kolmandana prooviti rakendada lisatingimusi enne imputeerimist, mis peaks tagama rohkem täielikke kirjeid ja täpsema tulemuse. Kõikide lisatingimuste mõju vaadeldi eraldi, lugedes muutused andmestikus. Selgus, et nii mõnigi tingimus muutis selliseid kirjeid, mis olid vaadeldud. Seega otsustati lisatingimuste rakendamisest loobuda.