INFO-2158 UDK: 574:004:004.6:519.213
Primljeno / Received: 2015-02-12 Izvorni znanstveni rad / Original Scientific Paper

# MISSING DATA PROBLEMS IN NON-GAUSSIAN PROBABILITY DISTRIBUTIONS

# PROBLEMI NEDOSTAJUĆIH PODATAKA U DISTRIBUCIJAMA VJEROJATNOSTI KOJE NISU GAUSSOVE

*Lovorka Gotal Dmitrović[1], Vesna Dušak[2], Jasminka Dobša[2]*

*University North, University Centre Varazdin, Croatia[1]; Faculty of Organization and Informatics, University of Zagreb, Varazdin, Croatia[2]*

*Sveučilište Sjever, Sveučilišni centar Varaždin, Hrvatska[1]; Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin, Hrvatska[2]*

*Abstract*

Ecology as a scientific discipline has been developing rapidly and becoming the interdisciplinary science based on Information and Communication Technologies (ICT). Discovering, integrating and analyzing a huge amount of heterogeneous data is crucial in exploring complex ecological issues. Ecoinformatics offers tools and approaches for the management of environmental data which it transforms further into information and knowledge. The development of Information Technologies with the special emphasis on the research methods of gathering and analyzing data, their storage and data access, has significantly enhanced the laboratory methods and their reports. The above, influences the data quality, as well as the research itself. Moreover, it provides a stable base for the development and the replacement of missing data. The improper missing data handling can lead to invalid conclusions. Therefore, it is important to use the adequate methods for handling the missing data. This paper compares The Deleting Rows Method (Listwise Deletion Method) and six single imputation methods, namely: Last Observation Carried Forward (LOCF), Hot-deck Imputation, Group Mean Imputation, Estimated Mean Value Imputation (Regression), Mode Imputation and Median Imputation. For the purposes of this study, the actual, empirical data was collected and used from the non- Gaussian probability distribution of the observed technical system. Mostly, these are asymmetric probability distributions with a tail. Data sets with missing data were created by deleting values with a random number generator. The experiment was repeated three times for each

*Sažetak*

Ekologija kao znanstvena disciplina brzo se razvija i postaje interdisciplinarna znanost koja se temelji na informacijsko komunikacijskim tehnologijama (IKT). Otkrivanje, integriranje i analiza ogromnih količina heterogenih podataka je ključno u istraživanju složenih ekoloških pitanja. Ekoinformatika nudi alate i pristupe za upravljanje okolišnim pokazateljima i pretvara ih u informacije i znanje. Razvoj informacijskih tehnologija s posebnim naglaskom na metode istraživanja prikupljanja i analizu podataka, njihovu pohranu i pristup podacima znatno poboljšava laboratorijske metode i njihova izvješća. Sve to utječe na kvalitetu podataka, uključujući istraživanja i pruža stabilnu bazu za njihov razvoj i zamjenu podataka koji nedostaju. Nepravilno rukovanje s „nedostajućim podacima“ može dovesti do pogrešnih zaključaka. Dakle, važno je koristiti odgovarajuće metode za upravljanje podacima koji nedostaju. U ovom radu će se usporediti metoda brisanja reda te šest metoda jednostruke metode imputacije: metoda posljednjeg provedenog promatranja, metoda Hot-deck imputacije, metoda imputacije srednje vrijednosti grupe, metoda imputacije procijenjene srednje vrijednosti (regresija), metoda imputacije moda i metoda imputacije medijana. Za potrebe ovog istraživanja, prikupljeni su empirijski podaci tehničkog sustava kod kojih se podaci ne raspoređuju prema Gaussovim distribucijama vjerojatnosti. Uglavnom su to asimetrične distribucije s repom. Skupovi s nedostajućim podacima stvoreni su brisanjem vrijednosti koristeći generator slučajnih brojeva. Eksperiment je ponovljen tri puta za svaku ispitivanu varijablu

100%, 95% and 75% sets of the collected data. Experiments have shown that the best imputation data results were provided by Hot-Deck Method, especially when there was a larger number of missing data, which has been confirmed by the Tests of Goodness. The same results, regardless of the set size, were provided by Listwise Deletion Method, which is simpler.

nad skupovima od: 100%, 95% i 75% prikupljenih podataka. Eksperimenti su pokazali da je najbolje rezultate imputacije podataka dala Hot-deck metoda, naročito kad nedostaje veći broj podataka što su potvrdili i testovi slaganja. Iznenađujuće je to da skoro jednako dobre rezultate, neovisno o veličini skupa, daje metoda brisanja redaka koja je puno jednostavnija.

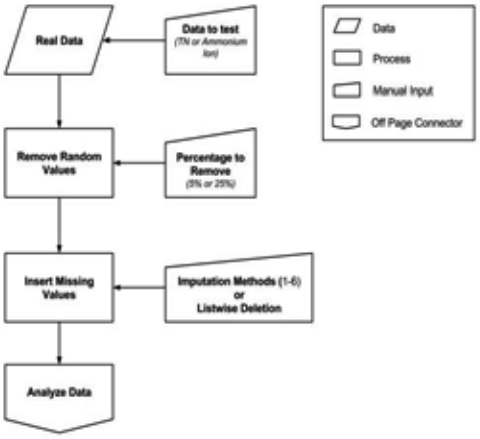
**Introduction**

Data collection and cleaning is a fundamental requirement for high quality research commission, and requires a significant amount of working hours. The available data is often incomplete and sometimes indistinguishable. Ecological data could be collected and organized in many different ways, for example, by observations and manual recording in the laboratory, by collecting field data via hand-written sheets, tape recorders, and computers, as well as by automated data collection via laboratory and field instrumentation. According to C. Strasser /**1**/, data management planning is often underappreciated in a project design. Nevertheless, it can save time, enhance research efficiency and, most importantly, it can fulfil the obligations to the research sponsors that increasingly require explicit data management plans as part of research proposals. Presently, there are many different approaches and tools used for data organization and management. They are ranging from spreadsheets and statistical software to relational database management systems, geographic information systems, etc. It should be pointed out that wide knowledge is required to recognize which approach is suitable for different eco types systems and research objects. Every approach, has advantages and disadvantages, but has to follow the unique data life cycle including: plan and experimental design; collection and data assurance (i.e. quality assurance and quality control); description (i.e. ascribing metadata) data preservation (i.e. data deposit in a secure data repository); discovery ovo bi promijenila u detection (i.e. identifying data that might be needed to answer a question); integration (i.e. merging data from multiple data sources), as well as, results analysis and interpretation (i.e. statistical analysis, visualization). Ecology has been evolving rapidly and changing increasingly into a more open, accountable, interdisciplinary, and collaborative data-intensive science based on Information and Communication Technologies (ICT). Ecoinformatics offers tools and approaches for managing ecological data and transforming the data into information and knowledge, according to 'Journal of Ecoinformatics Mission'. Manually collected data transferred into spreadsheets often results with errors, since diverse types of data could be mixed within a single column, while data summaries frequently conflate with raw data. Relational databases allow the employment of constraints and determine data types that can be entered (e.g. data typing), in order to assure the data quality. So, relational databases are crucial for analyzing and meta data preparation as well as for facilitating simulation models to establish conceptual models. Statistical software tools support many of the functions available through spreadsheet programs and provide the benefit of supporting robust calculations, data analysis, quality assurance, visualization and data sub-setting.

The development of Information Technologies with the special emphasis on research methods of gathering and analyzing data, their storage and data access has significantly enhanced the laboratory methods and their reports. This development, together with the computer supported laboratory devices allows performing more sophisticated and precise analysing and easier data management. All these affect the quality of data including the research itself and provide a stable base for the development and replacement of missing data. The improper missing data handling can lead to incorrect conclusions. Therefore, it is important to use adequate methods to handle missing data. This paper compares The Deleting Rows Method (Listwise Deletion Method) and six Single Imputation Methods, namely: Last Observation Carried Forward (LOCF), Hot-deck Imputation, Group Mean Imputation, The Method of Estimated Mean Value Imputation (regression), The Method of Imputation Mode, and The Meth-

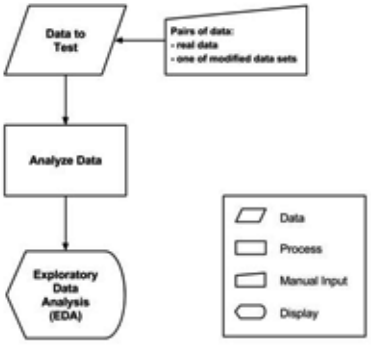
od of Imputation Median. For the purposes of this research, actual and empirical data was collected. Several experimental sets were created in a way that some values had been eliminated from the set by using a random number generator. The experiment was done on the three sets. The first set of 100 % data (contains empirical data), another set of 95 % data (5% of the data from the actual set is missing) and a third set of 75 % data (25% of the actual data set is missing) (Figure 1). The research has included all the variable components of the studied complex engineering system.

Figure 1: Data preparation flowchart




The matching sets with imputed values versus the real datasets, display the quality of imputation methods. Furthermore, providing the exploratory statistics, the theoretical distributions are derived from the actual dataset. In addition to single imputation methods, the Listwise Deleting Method would be used, to make comparison between mentioned methods with the actual dataset (Figure 2).

Figure 2: Statistical analysis flowchart




**Missing data**

Missing data statistically occur when all values of the observed variables are not present. In the social sciences, for example, refusing to answer a question in a survey would result with missing data. During data collection process by either a laboratory analysis, or collecting data by measurements, data could be lost over a specific period. This may occur because either the examiner drops some measured data, or data is not collected over some period, or the measurement was not performed at all (e.g. holidays, Sundays ...)

Missing data reduces the representativeness of the sample and may distort the conclusions. Therefore, it is necessary to take more action to prevent the missing of the actual value from the collected data.

G.B. Durrant /**2**/ groups the missing data according to the reasons why data is missing:

- missing completely at random (MCAR),

- missing at random (MAR) and

- not missing at random (NMAR).

Knowing and understanding reasons why data is missing may help the analysis of the remaining data. If missing values are missing at random, the sample is still representative. However, if the values are systematically missing, the results may be incorrect. It is therefore important to determine the type of data missing according to the reason. The values in the set of data are missing completely at random (MCAR), if the events are due to occur completely independently, they do not depend on the observed variables and parameters of interest, and if they appear completely randomly /**3**/. The values are missing at random (MAR) is an alternative, and occurs when the missing data is relating to a particular variable, for example, accidentally skipped answer from the questionnaire /**4**/. Do not missing at random (NMAR) is data that is missing for a reason, i.e. the value of the variable is missing due to the reason that it does not exist. Such are intentionally skipped questions in the questionnaire by the participants with particular characteristics, for example (Heitjan, D.F., 1994; Schafer, J.L., 1997; Little, R.J.A., Rubin, D.B., 2002). In addition, the missing data can be *univariate*, which means that the missing data occurs only in one variable of response, or *multivariate* if the missing data appears in more than one variable /**5**/.

**Imputation methods**

When one or more values are missing in a set of numbers, most software packages use Listwise Deletion Method. It is a simple method, most commonly used for missing data treatment. This method deletes rows containing gaps, and uses only the complete ones /**6**/, /7/. Although this method has been widely used, it is biased and therefore limited and eligible to errors. Moreover, the variance estimate cannot be indirect, so relations between variables could be distorted and the representativeness of the sample lost. Compensation methods for recovering data sets are: Case Method, The Imputation Method, The Pondering Method and Model-based Procedures such as maximum likelihood estimates. Imputation is the process of replacing missing data with substituted values of variables. By the substitution of missing data imputation, a complete set of data is obtained and so the standard techniques for integrated data sets could be applied. The main reason for using the imputation methods is to reduce the insufficient response bias, which occurs due to the distribution of missing values gap. As a result of using the imputation method, a representative set could be obtained. Contrary to the methods of deletion, in the imputation methods, the size of the samples stays the same as real dataset, resulting in potentially greater efficiency. Imputation methods commonly use the observed auxiliary information (variables), from which the missing data is indirectly obtained /**8**/. There are different methods of imputation. The main method of classification is: single and multiple imputations. This paper compares Deleting Rows Method (Listwise Deletion Method) and six Single Imputation Methods:

- Last Observation Carried Forward (LOCF),
- Hot-Deck Imputation,
- Group Mean Imputation,
- Estimated Mean Value Imputation (Regression),
- Mode Imputation Method,
- Median Imputation Method

Last Observation Carried Forward method (LOCF) is used when the data is longitudinal i.e. the measurements are repeated for the same variable on the same sample type. The values of Last Observation Carried Forward are used to fill in missing values later in the research. The method assumes that the value of the measurement remains constant (unchanged) up to the last measurement. This assumption may be biased if the values change in time. By Estimated Mean Value Imputation (Regression), the total average of numerical variable for each item that is missing in this variable, is being imputed. A variation of this procedure is Class Mean Imputation, where classes are defined on the basis of independent variables. The disadvantage of this procedure is that the distribution of the studied variables increases and the relation between variables could be distorted (Kalton, G., 1983; Lessler, J.T., Kalsbeek, W.D., 1992, Little, R.J.A., Rubin, D.B., 2002). Such simple methods of imputation have been commonly used in the social sciences (Jinn, J.H., Sedransk, J., 1989; Allison, P.D., 2001). However, these methods are often inadequate for handling the missing data, therefore more sophisticated methods should be used. Mode Imputation Method and Median Imputation Method are similar to this method. Median Imputation Method replaces the missing data by imputing value of median or other central value. Half of the value set is located above the median and another half below it. The median is less sensitive to extreme values than the mean value, and is especially suitable for asymmetrical distributions. Mode Imputation Method replays the missing data imputing mode (the value that occurs most frequently i.e. the value of the maximum frequency). This method replaces the missing data and artificially increases the maximum of probability distribution. The assigned value is a predicted value, with or without additions, and not really the observed value by the Hot-Deck Method /**9**/. That is to say, the Hot-deck Method complemented by comparing data from the same type of another sample. It is a commonly used method of imputation in practice, and is suitable when ranked data is observed. The advantage of the method is imputing the values that were used. The method is usually non-parametric (or semi-parametric) and is suitable in the instances when data components are distorted or show certain features, such as truncation and rounding effects, as is often the case in social scientific research. In Hot-Deck Imputation, imputed values have the same shape as the distribution of the observed data /**10**/ and the method is suitable for large samples. Predictive Mean Matching Imputation Method is used when Regression is provided. This is a deterministic method, and randomization can be introduced to define a set of

values the closest to the predicted value, as well as the random value from the closest set, in order to apply imputation (Schenker, N., Taylor, J.M.G., 1996; Nordholt, E.C., 1998, Little, R.J.A., Rubin, D.B., 2002). In the case of single imputation, this method uses Least Squares Method. Missing values are implemented by using the predicted value providing linear regression model. The method is used to insert all continuous variables in the overall data set. Another form of Estimated Mean Value Matching Imputation is t Hot -deck Imputation with classes defined on a range of estimated values. The method can achieve even more donoring imputation values within a class, which reduces the variance of imputation estimators. Donors imputation values within the class can be prepared with or without substitutions whereas the later is expected to lead to further reduction of variance (Durrant, G.B., Skinner, C., 2005; Kim, J.K., Fuller, W., 2004). Estimating the mean matching is a composite method, combining elements of Regression, Nearest Neighbour and Hot -Deck Imputation. Since the semi-parametric method uses the imputation models but does not fully rely on them, it is less sensitive to model misspecifications than, for example, the imputation regression /**11**/.

**Experiment I - concentration of ammonium ion**

Data set 1 - analysis of baseline variables

The actual data was obtained from the laboratory analysis of communal waste water. For the purposes of this study the data was collected daily during the period of 100 days. It considers the value of the measured concentration of ammonium ions. Descriptive statistics (Table 1) and theoretical probability distributions (Table 2) for the data of concentration values are obtained using *Stat: Fit* applications *(Servicing Model v.4)*.

Data set 2 - 5% data missing from a range of real value variables

From the total data on the concentration values of ammonium ions, 5% of the collected data has been ejected by using a random number generator. Subsequently, the empty fields were imputed by using single imputation methods. The comparison was made using the actual values and Listwise Deletion.

The values of the descriptive statistics for sets of variables are presented in Table 1. The values of theoretical probability distribution are presented in Table 2. Figure 1 shows the theoretical probability distributions (Pearson type 6 distributions) for each of the methods, and for a series of the original data.

The Pearson 6 distribution is a continuous distribution bounded on the low side. Its function is:

$$f(x) = \frac{(\frac{x-min}{\beta})^{p-1}}{\beta\left[1+\left(\frac{x-min}{\beta}\right)\right]^{p+q} B(p,q)}$$

where is:
x - random number; x > min,
min - minimum x; min ϵ (-∞,∞),
β - scale parameter; β > 0,
p - shape parameter/vector of probabilities, p > 0 and q - shape parameter/vector of quantiles, q > 0.

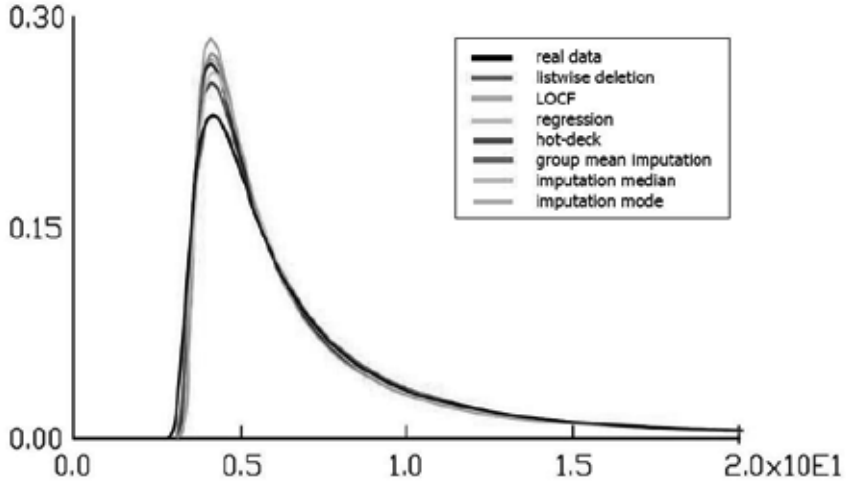



Figure 1: The theoretical probability distributions of ammonium ion concentration values where 5% of the data from the actual set is missing

Mode Imputation Method, Regression, Last Observation Carried Forward (LOCF) Method, Hot-Deck, Group Mean Imputation and Median Imputation Method, greatly increase the probability value with the highest probability (peak) (Figure 1).

The best similarity results with real data, in descriptive statistics (Table 1), have been obtained by Hot-deck imputation. All parameters of data set which was supplemented using Hot-Deck Method are very similar to real data. The worst results, in descriptive statistics, have been obtained by using Mode Imputation Method and Median Imputation Method. The parameters of skewness and kurtosis are disturbed.

Table 1: Descriptive statistics of ammonium ion concentration values where 5% of the data from the actual set is missing

| | REAL DATA | LISTWISE DELETION | GROUP MEAN IMPUTATION | LOCF | METHODS OF MPUTATION MEDIAN | METHODS OF MPUTATION MODE | HOT DECK MPUTATION | REGRESSION |
|---|---|---|---|---|---|---|---|---|
| data points | 100 | *95* | 100 | 100 | 100 | 100 | 100 | 100 |
| Minimum | 2,7 | 3,1 | 3,1 | 3,1 | 3,1 | 3,1 | 3,1 | 3,1 |
| Maximum | 31,8 | 31,8 | 31,8 | 31,8 | 31,8 | 31,8 | 31,8 | 31,8 |
| Mean | 8,744 | 8,82 | 8,82 | 8,77 | 8,684 | *8,589* | 8,8103 | **8,73343** |
| Median | 6,1 | **6,1** | *6,3* | **6,1** | **6,1** | 5,9 | **6,1** | 6 |
| Mode | 4,2 | **4,2** | **4,2** | *4,1* | **4,2** | 4,15 | **4,2** | **4,2** |
| standard deviation | 6,74574 | 6,66162 | *6,57319* | 6,68036 | 6,60013 | 6,65063 | **6,77276** | 6,65914 |
| Variance | 45,505 | 44,3772 | *43,2068* | 44,6272 | 43,5618 | 44,2309 | **45,8703** | 44,3441 |
| coefficient of variance | 76,4823 | 76,1851 | *74,5259* | 76,1728 | 76,0034 | 77,432 | 76,8732 | **76,2488** |
| Skewness | 1,49173 | 1,50702 | 1,5317 | **1,49914** | *1,57073* | 1,56525 | 1,48404 | 1,51507 |
| Kurtosis | 1,01938 | 1,10524 | 1,23543 | **1,05586** | *1,29506* | 1,26351 | 0,958748 | 1,11704 |

Note: the value closest to the actual value is shown in **bold**, and the furthest from the actual value in *italics*

The best similarity with real data results, in the theoretical probability distribution parameters (Table 2) were given by using Regression. All methods showed p-parameter significant error, which indicates the goodness of fit regardless to the level of significance. All methods increased β-parameter of distribution.

Table 2.: Theoretical probability distributions parameters for ammonium ion concentration values where 5% of the data from the actual set is missing

| | | REAL DATA | LISTWISE DELETION | GROUP MEAN IMPUTATION | LOCF | METHODS OF IMPUTATION MEDIAN | METHODS OF IMPUTATION MODE | HOT DECK IMPUTATION | REGRESSION |
|---|---|---|---|---|---|---|---|---|---|
| Theoretical probability distribution | | Pearson 6 (2,7, 0,194, 19,9, 1,48) | Pearson 6 (3,1, 0,643, 4,81, 1,35) | Pearson 6 (3,1, 1,07, 3,59, 1,49) | Pearson 6 (3,1, 0,298, 8,97, 1,26) | Pearson 6 (3,1, 0,408, 7,35, 1,35) | Pearson 6 (3,1, 0,254, 10,1, 1,27) | Pearson 6 (3,1, 0,254, 10,1, 1,27) | Pearson 6 (3,1, 0,251, 10,4, 1,26) |
| Minimum | | 2,7 | 3,1 | 3,1 | 3,1 | 3,1 | 3,1 | 3,1 | 3,1 |
| p | | 19,9164 | 4,80958 | *3,59135* | 8,97256 | 7,34846 | 10,0505 | 6,01526 | **10,4434** |
| q | | 1,48142 | 1,34947 | **1,49235** | 1,26274 | 1,34979 | 1,26787 | 1,27235 | *1,25921* |
| β | | 0,194172 | 0,642548 | *1,06821* | 0,297545 | 0,407811 | 0,25359 | 0,457237 | **0,251438** |
| $\chi^2$-test | total classes | 6 | | | | | | | |
| | interval type | equal probable | | | | | | | |
| | $\chi^2$ | 18,8 | 20,4 | 16,8 | 23,1 | *29* | 10,2 | 24,6 | **19,3** |
| | degrees of freedom | 5 | | | | | | | |
| | A | 0,05 | | | | | | | |
| | $\chi^2$ (5, 0,05) | 11,1 | | | | | | | |
| | p-value | 0,00209 | **0,00106** | 0,00498 | 0,00032 | *$2{,}32 \cdot 10^{-5}$* | 0,0708 | 0,000169 | 0,0017 |
| | Results | REJECT | REJECT | REJECT | REJECT | REJECT | DO NOT REJECT | REJECT | REJECT |
| Kolmogorov-Smirnov test | data points | 100 | 95 | 100 | 100 | 100 | 100 | 100 | 100 |
| | ks stat | 0,106 | **0,104** | *0,093* | 0,103 | 0,0975 | 0,0994 | 0,101 | 0,103 |
| | A | 0,05 | | | | | | | |
| | ks stat (100, 0,05) | 0,134 | *1,137* | 0,134 | 0,134 | 0,134 | 0,134 | 0,134 | 0,134 |
| | p-value | 0,194 | 0,242 | *0,332* | 0,226 | 0,28 | 0,259 | 0,242 | **0,223** |
| | Results | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT |
| Anderson-Darling test | data points | 99 | *94* | 99 | 99 | 99 | 99 | 99 | 99 |
| | ad stat | 3,2 | 3,04 | 2,9 | 3,14 | 3,07 | **3,21** | 3,05 | *3,02* |
| | A | 0,05 | | | | | | | |
| | ad stat(0,05) | 2,49 | | | | | | | |
| | p-value | 0,0217 | 0,026 | *0,0308* | 0,0232 | 0,0251 | **0,0214** | 0,0258 | 0,0269 |
| | Results | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT |

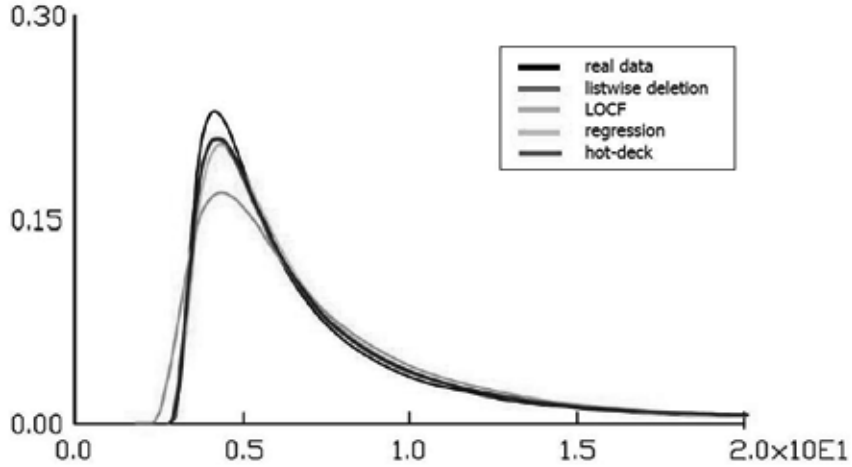
Note: the value closest to the actual value is shown in **bold**, and the furthest from the actual value in *italics*

Data set 3 - 25% data missing from a range of real value variables

From the total data on the concentration values of ammonium ions, 25% of the collected data has been ejected by using a random number generator. Subsequently, the empty fields were imputed by using single imputation methods. The comparison was made using the actual values and Listwise Deletion.

The values of the descriptive statistics for sets of variables are presented in Table 3. The values of theoretical probability distribution are presented in Table 4. Figure 2 shows the theoretical probability distributions for each of the methods, and for a series of the original data.

Figure 2: The theoretical probability distributions of ammonium ion concentration values where 25% of the data from the actual set is missing




Mode Imputation Method, Group Mean Imputation and Median Imputation Method, impute values which do not match tests in data set of valid Pearson 6 probability distribution (Figure 2). The best results were given by the following methods: Hot-Deck, LOCF and Listwise Deletion. It should be pointed out that, Regression significantly decreased the highest probability values (peak).

Table 3: Descriptive statistics concentration values of ammonium ion where 25% of the data from the actual set is missing

| | REAL DATA | LISTWISE DELETION | GROUP MEAN IMPUTATION | LOCF | METHODS OF MPUTATION MEDIAN | METHODS OF MPUTATION MODE | HOT DECK MPUTATION | REGRESSION |
|---|---|---|---|---|---|---|---|---|
| data points | 100 | *75* | 100 | 100 | 100 | 100 | 100 | 100 |
| Minimum | 2,7 | 2,7 | 2,7 | 2,7 | 2,7 | 2,7 | 2,7 | *1,82759* |
| Maximum | 31,8 | 31,8 | 31,8 | 31,8 | 31,8 | 31,8 | 31,8 | 31,8 |
| Mean | 8,744 | 8,88533 | 8,88533 | **8,743** | 8,189 | *7,714* | 8,819 | 8,83108 |
| Median | 6,1 | 6,3 | 7,15 | 6,4 | **6,1** | *4,9* | 6,3 | 6,3 |
| Mode | 4,2 | 4,3 | 8,88533 | 4,25 | *6,1* | **4,2** | **4,2** | 4,3 |
| standard deviation | 6,74574 | **6,65137** | *5,75055* | 6,37461 | 5,87692 | 6,10135 | 6,39777 | 6,6335 |
| Variance | 45,505 | **44,2407** | *33,0688* | 40,6356 | 34,5382 | 37,2265 | 40,9315 | 44,0033 |
| coefficient of variance | 76,4823 | 74,8579 | *64,7196* | 72,911 | 71,766 | 79,0945 | 72,5453 | **75,1154** |
| Skewness | 1,49173 | **1,4452** | 1,67723 | 1,52454 | *1,8983* | 1,86829 | 1,38748 | 1,52383 |
| Kurtosis | 1,01938 | **0,992302** | 2,3591 | 1,23202 | *2,74154* | 2,53023 | 0,910649 | 1,28115 |

The best similarity results on empirical (real) data, using descriptive statistics was given by Listwise Deletion (Table 3). All parameters of data set using Listwise Deletion match the empirical (real) data. The worst results, in descriptive statistics, were obtained by using Group Mean Imputation and Median Imputation methods. Parameters of skewness and kurtosis are disturbed for all data set except Listwise deletion.

The best similarity with real data results, in theoretical probability distributions (Table 4) were obtained by using Regression. All methods, (Figure 1 – Listwise deletion, LOCF, Hot-deck Imputation and Regression), show p-parameter significant error, which indicates the goodness of fit regardless to the level of significance. All methods increased the distribution´s β-parameter.

Table 4: Theoretical probability distributions of ammonium ion concentration values where 25% of the data from the actual set is missing

| | | REAL DATA | LISTWISE DELETION | LOCF | HOT DECK IMPUTATION | REGRESSION |
|---|---|---|---|---|---|---|
| Theoretical probability distribution | | Pearson 6 (2,7, 0,194, 19,9, 1,48) | Pearson 6 (2,7, 0,599, 7,58, 1,57) | Pearson 6 (2,7, 0,628, 8,14, 1,71) | Pearson 6 (2,7, 0,49, 9,14, 1,56) | Pearson 6 (1,83, 0,309, 26,3, 2,07) |
| Minimum | | 2,7 | 2,7 | 2,7 | 2,7 | *1,82759* |
| P | | 19,9164 | *7,5769* | 8,13617 | 9,14154 | **26,2516** |
| Q | | 1,48142 | 1,5769 | 1,70972 | **1,55583** | *2,07168* |
| B | | 0,194172 | 0,598904 | *0,627621* | 0,489549 | **0,309062** |
| $\chi^2$-test | total classes | 6 | | | | |
| | interval type | equal probable | | | | |
| | $\chi^2$ | 18,8 | **17,2** | 20,5 | *30,7* | 13,9 |
| | degrees of freedom | 5 | | | | |
| | A | 0,05 | | | | |
| | $\chi^2$ (5, 0,05) | 11,1 | | | | |
| | p-value | 0,00209 | 0,00407 | **0,00102** | $1{,}08 \cdot 10^{-5}$ | 0,0164 |
| | results | REJECT | REJECT | REJECT | REJECT | REJECT |
| Kolmogorov-Smirnov test | data points | 100 | 75 | 100 | 100 | 100 |
| | ks stat | 0,106 | **0,104** | 0,0987 | 0,11 | 0,112 |
| | A | 0,05 | | | | |
| | ks stat (100, 0,05) | 0,134 | 0,154 | 0,134 | 0,134 | 0,134 |
| | p-value | 0,194 | *0,366* | 0,266 | **0,167** | 0,153 |
| | results | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT |
| Anderson-Darling test | data points | 99 | 74 | 99 | 98 | 99 |
| | ad stat | 3,2 | 2,85 | 3,11 | *5,32* | **3,25** |
| | A | 0,05 | | | | |
| | ad stat (0,05) | 2,49 | | | | |
| | p-value | 0,0217 | 0,0327 | 0,0242 | *0,00203* | **0,0204** |
| | results | REJECT | REJECT | REJECT | REJECT | REJECT |

Note: the value closest to the actual value is shown in **bold**, and the furthest from the actual value in *italics*

**Experiment II - concentration of total nitrogen**

Data set 1 - analysis of baseline variables

The experiment was repeated. The data was collected daily during a period of 100 days, but this time for the concentration of total nitrogen. Statistical analysis of concentration values, using *Stat: Fit* applications *(Servicing Model v.4)* obtained data descriptive statistics (Table 5) and theoretical probability distributions (Table 6).

Data set 2 - 5% data missing from a range of real value variables

From the total data on the concentration values of nitrogen ions, 5% of the collected data was ejected by using a random number generator. Subsequently, the empty fields were imputed by using single imputation methods. The comparison was made by using the actual values and Listwise Deletion.

The values of the descriptive statistics for sets of variables are presented in Table 5. The values of theoretical probability distribution are presented in Table 6. Figure 3 shows the theoretical probability distributions for each of the methods, and for a series of the original data.

The Log-logistic distribution is a continuous distribution bounded on the lower side. Its mathematical model is:

$$f(x) = \frac{p(\frac{x-min}{\beta})^{p-1}}{\beta\left[1+\left(\frac{x-min}{\beta}\right)\right]^{p}}$$

where is:
- x - random number, x>min,
- min - minimum x, min $\epsilon$ (-∞,∞),
- $\beta$ - scale parameter, $\beta > 0$ and
- p - shape parameter, p > 0

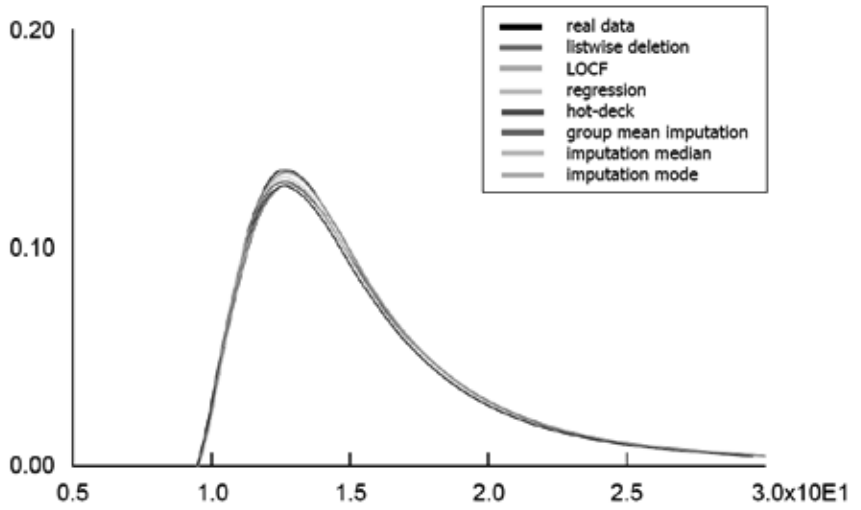



Figure 3: The theoretical probability distributions of total nitrogen concentration values where 5% of the data from the actual set is missing

Mode Imputation Method, Regression, Last Observation Carried Forward (LOCF) Method, Hot-Deck and Median Imputation Method, greatly increase the probability value with the highest probability (peak) (Figure 1).

The best similarity results with real data, in descriptive statistics (Table 1), have been obtained by Hot-deck imputation, Regression and LOCF. All parameters of data set which was supplemented using all methods are very similar to real data. The worst results, in descriptive statistics, have been obtained by using Group Mean Imputation Method, but the parameters aren't disturbed.

Table 5: Descriptive statistics of total nitrogen concentration values where 5% of the data from the actual set is missing

| | REAL DATA | LISTWISE DELETION | GROUP MEAN IMPUTATION | LOCF | METHODS OF MPUTATION MEDIAN | METHODS OF MPUTATION MODE | HOT DECK MPUTATION | REGRESSION |
|---|---|---|---|---|---|---|---|---|
| data points | 100 | 95 | 100 | 100 | 100 | 100 | 100 | 100 |
| Minimum | 9,5 | 9,5 | 9,5 | 9,5 | 9,5 | 9,5 | 9,5 | 9,5 |
| Maximum | 34,1 | 34,1 | 34,1 | 34,1 | 34,1 | 34,1 | *34,08* | *34,08* |
| Mean | 16,4824 | *16,6558* | 16,6471 | 16,514 | 16,538 | 16,538 | 16,6279 | **16,5104** |
| Median | 14,3 | 14,4 | *14,5* | 14,35 | **14,3** | **14,3** | 14,4 | 14,4 |
| Mode | 14,25 | **14,25** | 14,2 | **14,25** | *14,5* | *14,5* | 14,345 | 14,3698 |
| standard deviation | 6,12747 | *6,23633* | 6,07693 | 6,11316 | 6,09868 | 6,09868 | **6,11975** | 6,11703 |
| Variance | 37,5459 | *38,8919* | 36,9291 | 37,3707 | 37,1939 | 37,1939 | **37,4513** | 37,4181 |
| coefficient of variance | 37,1768 | 37,4424 | *36,5044* | 37,018 | 36,8768 | 36,8768 | 36,8041 | **37,0496** |
| Skewness | 1,45704 | *1,38163* | 1,42281 | 1,45539 | **1,45793** | **1,45793** | 1,40563 | 1,45019 |
| Kurtosis | 1,05129 | *0,80968* | 1,02225 | **1,05584** | 1,06871 | 1,06871 | 0,931929 | 1,04541 |

Note: the value closest to the actual value is shown in **bold**, and the furthest from the actual value in *italics*

The best similarity with real data results, in the theoretical probability distribution parameters (Table 2) were given by using Regression. All the methods have shown good matching in the scale and shape parameters.

Table 6: The theoretical probability distributions parameters of total nitrogen concentration values where 25% of the data from the actual set is missing

| | | REAL DATA | LISTWISE DELETION | GROUP MEAN IMPUTATION | LOCF | METHODS OF IMPUTATION MEDIAN | METHODS OF IMPUTATION MODE | HOT DECK IMPUTATION | REGRESSION |
|---|---|---|---|---|---|---|---|---|---|
| Theoretical probability distribution | | Log-logistic (9,5, 2,17, 4,98) | Log-logistic (9,5, 2,11, 5,1) | Log-logistic (9,5, 2,17, 5,21) | Log-logistic (9,5, 2,18, 5,03) | Log-logistic (9,5, 2,19, 5,07) | Log-logistic (9,5, 2,19, 5,07) | Log-logistic (9,5, 2,15, 5,15) | Log-logistic (9,5, 2,16, 5,02) |
| Minimum | | 9,5 | 9,5 | 9,5 | 9,5 | 9,5 | 9,5 | 9,5 | 9,5 |
| p | | 2,16591 | 2,11415 | *2,17428* | 2,1782 | 2,19331 | 2,19331 | 2,15148 | **2,15519** |
| β | | 4,97902 | 5,10469 | *5,21472* | 5,0283 | 5,07253 | 5,07253 | 5,15024 | **5,0249** |
| $\chi^2$-test | total classes | 6 | | | | | | | |
| | interval type | equal probable | | | | | | | |
| | $\chi^2$ | 15,4 | 13,7 | 11,7 | 16,2 | *18,1* | 18,1 | 18,6 | **16,2** |
| | degrees of freedom | 5 | | | | | | | |
| | A | 0,05 | | | | | | | |
| | $\chi^2$ (5, 0,05) | 11,1 | | | | | | | |
| | p-value | 0,00864 | **0,0177** | 0,0388 | 0,0064 | *0,00285* | *0,00285* | 0,00232 | 0,0064 |
| | results | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT |
| Kolmogorov-Smirnov test | data points | 100 | 95 | 100 | 100 | 100 | 100 | 100 | 100 |
| | ks stat | 0,097 | **0,0933** | *0,0864* | 0,0923 | 0,107 | 0,107 | 0,0951 | 0,101 |
| | A | 0,05 | | | | | | | |
| | ks stat (100, 0,05) | 0,134 | *1,137* | 0,134 | 0,134 | 0,134 | 0,134 | 0,134 | 0,134 |
| | p-value | 0,285 | 0,358 | *0,42* | 0,341 | 0,188 | 0,188 | 0,307 | **0,239** |
| | results | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT |
| Anderson-Darling test | data points | 99 | *94* | 99 | 99 | 99 | 99 | 99 | 99 |
| | ad stat | 3,25 | 3,1 | 2,87 | 3,23 | 3,35 | **3,35** | *3,19* | *3,19* |
| | A | 0,05 | | | | | | | |
| | ad stat(0,05) | 2,49 | | | | | | | |
| | p-value | 0,0205 | 0,0242 | *0,032* | 0,0208 | 0,0183 | **0,0183** | 0,0218 | 0,022 |
| | results | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT | REJECT |

Note: the value closest to the actual value is shown in **bold**, and the furthest from the actual value *italic*

## Data set 3 - 25% data missing from a range of real value variables

From the total data on nitrogen ions concentration values, 25% of the collected data was ejected by using a random number generator. Subsequently, the empty fields were imputed by using single imputation methods. The comparison was made by using the actual values and Listwise Deletion.

The values of the descriptive statistics for sets of variables values are presented in Table 7.The values of theoretical probability distribution are presented in Table 8. Figure 4 shows the theoretical probability distributions for each of the methods, and for a series of the original data.

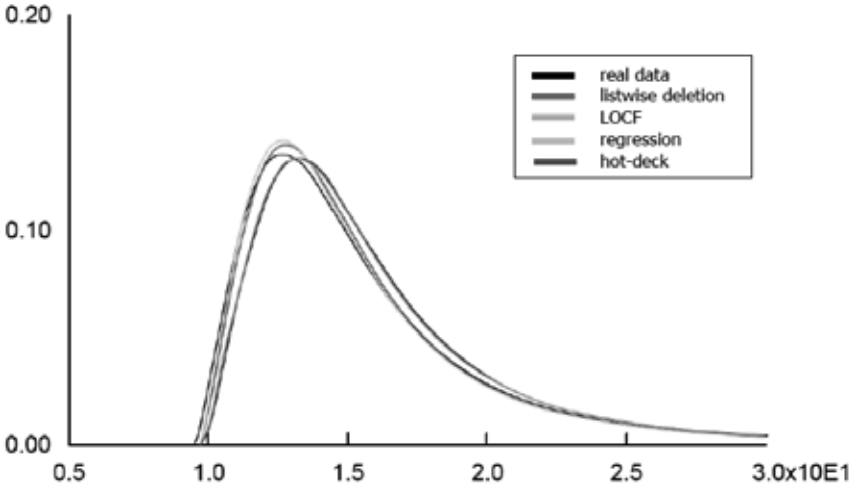



Figure 4: The theoretical probability distributions of total nitrogen concentration values where 25% of the data from the actual set is missing

Mode Imputation Method, Group Mean Imputation and Median Imputation Method, impute values which do not match tests in data set of valid Log-Logistic probability distribution (Figure 4). The best results were given by the following methods: LOCF and Listwise Deletion. Regression and Hot-Deck method significantly shift the peak of probability values.

The best similarity results with real data, in descriptive statistics (Table 7), have been obtained by Hot-deck imputation and Regression. All parameters of data set which was supplemented using Listwise deletion, LOCF, Hot-deck imputation and Regression are very similar to real data. The worst results, in descriptive statistics, have been obtained by using Group Mean Imputation, Median Imputation and Mode Imputation methods, whereas all the parameters are slightly disturbed.

The best similarity with real data results, in the theoretical probability distribution parameters (Table 8) were given by using Listwise deletion. All the methods (Figure 4 – Listwise Deletion, LOCF, Hot-Deck Imputation and Regression) have shown good matching in the scale and shape parameters.

Table 7: Descriptive statistics concentration values of total nitrogen where 25% of the data from the actual set is missing

| | REAL DATA | LISTWISE DELETION | GROUP MEAN IMPUTATION | LOCF | METHODS OF IMPUTATION MEDIAN | METHODS OF IMPUTATION MODE | HOT DECK IMPUTATION | REGRESSION |
|---|---|---|---|---|---|---|---|---|
| data points | 100 | *75* | 100 | 100 | 100 | 100 | 100 | 100 |
| Minimum | 9,5 | 9,7 | 9,7 | 9,7 | 9,7 | 9,7 | 9,68 | 9,68 |
| Maximum | 34,1 | 33,8 | 33,8 | 33,8 | 33,8 | 33,8 | 33,8 | 34,5291 |
| Mean | 16,4824 | 16,2333 | 16,2956 | 16,192 | *15,75* | *15,75* | 15,9573 | **16,6492** |
| Median | 14,3 | 14,4 | *15,6* | 14,35 | **14,3** | **14,3** | **14,3** | 14,5 |
| Mode | 14,25 | 14,2 | 16,4824 | *13,05* | 14,3 | 14,3 | **14,25** | 13,325 |
| standard deviation | 6,12747 | 5,54223 | *4,79285* | 5,55721 | 4,86493 | 4,86493 | 5,18118 | **5,74965** |
| Variance | 37,5459 | 30,7163 | *22,9714* | 30,8826 | 23,6676 | 23,6676 | 26,8446 | **33,0585** |
| coefficient of variance | 37,1768 | 34,141 | *29,4119* | 34,3207 | 30,8885 | 30,8885 | 32,469 | **34,5341** |
| Skewness | 1,45704 | **1,4688** | 1,66477 | 1,44505 | *1,90908* | *1,90908* | 1,55254 | 1,53563 |
| Kurtosis | 1,05129 | 1,32037 | 2,70604 | **1,21806** | *3,16392* | *3,16392* | 1,78117 | 1,5509 |

Table 8: The theoretical probability distributions parameters of total nitrogen concentration values where 25% of the data from the actual set is missing

| | | REAL DATA | LISTWISE DELETION | LOCF | HOT DECK IMPUTATION | REGRESSION |
|---|---|---|---|---|---|---|
| Theoretical probability distribution | | Log-logistic (9,5, 2,17, 4,98) | Log-logistic (9,7, 2,19, 4,84) | Log-logistic (9,7, 2,14, 4,72) | Log-logistic (9,68, 2,34, 4,8) | Log-logistic (9,68, 2,3, 5,23) |
| Minimum | | 9,5 | 9,7 | 9,7 | 9,68 | 9,68 |
| p | | 2,16591 | **2,18853** | 2,13734 | *2,34375* | 2,30394 |
| β | | 4,97902 | **4,83744** | *4,7214* | 4,80016 | 5,22967 |
| $\chi^2$-test | total classes | 6 | | | | |
| | interval type | equal probable | | | | |
| | $\chi^2$ | 15,4 | 10 | **13,4** | *8,36* | 12,9 |
| | degrees of freedom | 5 | | | | |
| | A | 0,05 | | | | |
| | $\chi^2$ (5, 0,05) | 11,1 | | | | |
| | p-value | 0,00864 | 0,0741 | **0,0199** | *0,137* | 0,0241 |
| | Results | REJECT | REJECT | REJECT | REJECT | REJECT |
| Kolmogorov-Smirnov test | data points | 100 | 75 | 100 | 100 | 100 |
| | ks stat | 0,097 | 0,0844 | *0,0807* | **0,0923** | 0,0875 |
| | A | 0,05 | | | | |
| | ks stat (100, 0,05) | 0,134 | *0,154* | 0,134 | 0,134 | 0,134 |
| | p-value | 0,285 | *0,628* | 0,388 | **0,341** | 0,406 |
| | Results | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT | DO NOT REJECT |
| Anderson-Darling test | data points | 99 | 74 | 99 | 98 | 99 |
| | ad stat | 3,25 | 2,68 | 2,87 | 4,95 | 2,87 |
| | A | 0,05 | | | | |
| | ad stat (0,05) | 2,49 | | | | |
| | p-value | 0,0205 | 0,0398 | 0,0321 | *0,00305* | **0,0318** |
| | Results | REJECT | REJECT | REJECT | REJECT | REJECT |

Note: the value closest to the actual value is shown in **bold**, and the furthest from the actual value in *italics*

**Conclusion**

In this paper, the results of applying Imputation Method and Listwise Deletion Method on the samples of ammonium ions concentration values and total nitrogen concentration values have been presented. Missing data are missing at random (MAR). For this purpose, two experiments were carried out. The first experiment was performed on the ammonium ions concentration values, while the other experiment was done on total nitrogen concentration values. The experiments were conducted for the three instances, i.e. a complete set of data, a set with 5% missing data and a set with 25% missing data. Statistical analysis showed that in the first experiment, the values of concentration of ammonium matched Pearson distribution 6, while in the second experiment; the total nitrogen concentration corresponded to Log-logistic distribution. Both theoretical distributions are asymptotic. Subsequently, the experiments were performed over the same set of data, or with 5% and 25% missing values of ammonium and total nitrogen concentrations. In the first part of experiment, 5% of data is missing and the missing values were imputed by using the imputation methods and Listwise Deletion. In the second part of the experiment, 25% of data is missing. Test results of goodness imputation methods and Listwise Deletion methods refer to the percentages of missing data and goodness-stream concentrations:

1) When a small number of values is missing, all methods show the good matching of probability distributions according to descriptive statistics. However when ¼ of values (25%) are missing, the methods of mean, median or mode imputation show a big deviation. Therefore, the probability distribution of the obtained data does not correspond to the empirical distribution.

2) The interesting fact is that the Listwise Deletion Method, which is the simplest method, provides very good matching results with the probability distributions, followed by Hot-Deck Imputation Method and Last Observation Carried Forward method. Regression method strongly levels the probability distribution by decreasing the maximum probability, as well as, increasing the minimum probability values distributions. Last Observation Carried Forward (LOCF), as well as Listwise Deletion methods, closely follow the observed distributions, with the exception of a deviation from the actual value at the peak of the distribution.

*References*

/1/ Strasser, C. et al., DataONE promoting data stewardship through best practices, In Proceedings of the Environmental Information Management Conference 2011 (EIM 2011), 126–131, University of California, 2011.
/2/ Durrant, Gabriele B., Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review, ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute, (S3RI), University of Southampton, 2005.
/3/ Rubin, D. B., Multiple Imputation for Nonresponse in Surveys, New York, Chichester, 1987.
/4/ Little, R.J.A., A Test of Missing Completely at Random for Multivariate Data with Missing Values, Journal of the American Statistical Association, 83, 404, 1198-1202, 1988.
/5/ Durrant, G.B. and Skinner, C., Using Missing Data Methods to Correct for Measurement Error in a Distribution Function, Survey Methodology, forthcoming, 2005.
/6/ Allison, P. D., Missing Data, Sage University Papers Series on Quantitative Applications in the Social Sciences, series no. 07-136, Thousand Oaks, 2001.
/7/ Ibrahim, J.G., Chen, M.H. Lipsitz, S.R. and Herring, A.H., Missing-Data Methods for Generalised Linear Models: A Comparative Review, Journal of the American Statistical Association, 100, 469, 332-346. 2005.
/8/ Schafer, J.L., Graham, J.W., Missing Data: Our View of the State of the Art, Psychological Methods,7, 2, 147-177, 2002.
/9/ Schenker, N., Taylor, J.M.G., Partially Parametric Techniques for Multiple Imputation, Computational Statistics and Data Analysis, 22, 425-446, 1996.
/10/ Rubin, D. B., Multiple Imputation for Nonresponse in Surveys, New York, Chichester, 1987.
/11/ Schenker, N., Taylor, J.M.G., Partially Parametric Techniques for Multiple Imputation, Computational Statistics and Data Analysis, 22, 425-446, 1996.

*Literature*

1. Bevanda, V., Sinković, G., Standardi za informacijsko-komunikacijsku tehnologiju (ICT), Informatologia (4). pp. 295-300, 2007.
2. Heitjan, D. F., Ignorability in General Incomplete-Data Models, Biometrika, 81, 4, 701-708, 1994.

3. Jinn, J.H., Sedransk, J., Effect on Secondary Data Analysis of Common Imputation Methods, Sociological Methodology, 19, 213-241, 1989.
4. Kalton, G., Compensating for Missing Survey Data, Michigan, 1983.
5. Kim, J.K., Fuller, W., Fractional Hot Deck Imputation, Biometrika, 91, 3, 559-578, 2004.
6. Lessler, J.T., Kalsbeek W. D., Nonsampling Error in Surveys, New York, Chichester, 1992.
7. Little, R.J.A., Rubin, D.B. Statistical Analysis with Missing Data, New York, 2002.
8. Michener, W. K., Jones, M. B., Ecoinformatics: supporting ecology as a data-intensive science, Trends in Ecology and Evolution, February 2012, Vol. 27, No. 2, http://www.cell.com/trends/ecology-evolution/pdf/S0169-5347(11)00339-9.pdf, *download: 12.01.2015.*
9. Nordholt, E.S., Imputation: Methods, Simulation, Experiments and Practical Examples, International Statistical Review, 66, 2, 157-180, 1998.
10. Pearson, K., Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material, Philosophical Transactions of the Royal Society, 1895.
11. Schafer, J.L., Analysis of Incomplete Multivariate Data, London, 1997.
12. Šehanović, J., Prikaz Zbornik radova V. međunarodnog simpozija "Informacijski sustavi 94", Informatologia, 26 (3-4). pp. 83-88, 1994.
13. Šehanović, J., Žugaj, M., Međunarodni simpozij "informacijski sustavi" (1989-1996), Informatologia, 28 (1-2). pp. 57-61, 1996.
14. Tomčić, Z., Šehanović, J., Turistički hidrometereološki informacijski sistem, Informatologia, 25 (3-4). pp. 73-76, 1993.