FACTA UNIVERSITATIS (NIŠ)
SER. MATH. INFORM. Vol. 36, No 1 (2021), 191-204
https://doi.org/10.22190/FUMI201118016I

Original Scientific Paper

MISSING DATA SAMPLES: SYSTEMATIZATION AND CONDUCTING METHODS-A REVIEW

Ivana D. Ilić[1], Jelena M. Višnjić[1], Branislav M. Randjelović[2] and Vojislav V. Mitić[2]

[1]**Faculty of Medical Sciences, Department of Mathematics and Informatics, Bulevar dr. Zorana Djindjića 81, 18 000 Niš, Serbia**
[2]**Faculty of Electronic Engineering, Department of Mathematics and Informatics, Aleksandra Medvedeva 14, 18 000 Niš, Serbia**

Abstract. This paper investigates the phenomenon of the incomplete data samples by analyzing their structure and also resolves the necessary procedures regularly used in missing data analysis. The research gives a crucial perceptive of the techniques and mechanisms needed in dealing with missing data issues in general. The motivation for writing this brief overview of the topic lies in the fact that statistical researchers inevitably meet missing data in their analysis. The authors examine the applicability of regular approaches for handling the missing data situations. Based on several previously published results, the authors provide an example of the incomplete data sample model that can be implemented when confronting with specific missing data patterns.

Keywords: Missing data, EM algorithm, Listwise deletion, Missing data analysis.

1. Introduction

One important issue which affects almost all datasets, despite major advances in the design and collection of data is the incompleteness. This situation appears when no data value is stored for some feature or an attribute in the dataset. The incompleteness may occur for different reasons. For instance, missing data in a survey may arise when there are no data for a respondent or when some variables for a respondent are unknown because of refusal to provide or failure to collect the

Received November 18, 2020; accepted December 14, 2020.
Corresponding Author: Ivana D. Ilić, Faculty of Medical Sciences, Department of Mathematics and Informatics, Bulevar dr. Zorana Djindjića 81, 18 000 Niš, Serbia | E-mail: ivanailicc3@gmail.com
2010 *Mathematics Subject Classification.* Primary 62D10; Secondary 62D05

response. Also, missing data may occur if the data collection was not done properly or if the mistakes were made with the data entry caused by the researchers themselves. Nevertheless, the problem of an adequate conduction of missing data remains, regardless of whether missing data result from a participant disintegration, a nonresponse item, or an irregular availability of respondents. See [10] or [20] for a summarization of these questions. In addition, we must point out a significant difference between "the item nonresponse" and "the unit nonresponse". The item nonresponse situation indicates that the respondent skipped one or more questions in the analysis. On the other hand, the unit nonresponse appears when the respondent refused to cooperate and consequently, all the resulting data are missing for this respondent. Trough the existing literature we conclude that the methods used for the item nonresponse and the unit nonresponse have been completely different.

In the last few years many articles devoted to the problem where practical missing data issues are discussed have appeared in various domains such as: economy, politics, biomedical research, social sciences, medicine and engineering. Giannone et al.(see [8]) developed a formal method for evaluating gross domestic product (GDP) growth using the large datasets with missing observations monitored by central banks. Schumacher and Breitung (see [34]) used a novel real-time dataset with missing values for the German economy in the empirical application of forecasting the GDP growth. For more practical applications with incomplete samples in various domains see for instance: [16] and [22] in economy and finance, [9], [3], [17] and [26] in biomedical field, [5] in social sciences and [29] in astrophysics.

The prosperity of the missing data procedures available to scientists often produces uncertainty regarding to the choice of the eventual implemented method. Our purpose is to discuss the applicability of general methods for dealing with missing data and to review current advances associated with specific missing data techniques. An additional intention of this paper is to propose a mathematical model (Chapter 4) that can be used in certain missing data situations under specified conditions.

2. An overview of the missing data classification

The task of classification of the data incompleteness type is a complex phenomena and its attaintment depends upon several factors that need to be taken under consideration. In the results obtained in [27] each data has certain likelihood of being missing. Based on that assumption he classified the incomplete sample problems into three categories. The data are said to be missing completely at random (MCAR) if the probability of being missing is the same for all cases. This practically means that the reasons of the data missingness are unrelated to the data, meaning that the missingness has nothing to do with the person being questioned. For example, a questionnaire might be lost in the post, or a blood sample might be ruined in the laboratory for an unknown reason, so that certain portion of the data will be missing simply because of some bad coincidences. An example which describes clearly this type of the data is when we take a random sample of a population. In this situation, each member of the population has an equal chance of being

included in the sample. So, the unobserved data of members in the population that were not included in the sample are MCAR. Basically, we may conclude that the points that are missing in the MCAR case present a random subset of the data. There is no systematic mechanism that makes some data more likely to be missing than others. Although in the MCAR pattern we may consequently neglect many of the difficulties that come about the data are missing, we must have in mind that the MCAR model is a bit rare in the real life statistical researches. If we denote a full matrix of the data in the analysis with $\mathbb{X}$, it is obvious that it can be written in the form $\mathbb{X} = \{X, \widetilde{X}\}$, where X are the observed and $\widetilde{X}$ the missing data. Let us define R as a matrix with the identical dimensions as $\mathbb{X}$ where:

$$R_{i,j} = \begin{cases} 1, & \text{if the data is missing} \\ 0, & \text{otherwise.} \end{cases}$$

Now, mathematical simplification of MCAR data type can be formulated as:

$$P(R\big|X, \widetilde{X}) = P(R),$$

meaning that the probability of the realization of R matrix will not depend neither on the observed nor on the unobserved data.

The second structure of the incompleteness is missing at random (MAR) and it covers much wider class of the statistical survey settlements. In this case, the probability of being missing is the same only within groups defined by the observed data. As an example of this situation is the case of a survey where only younger people have missing values measuring IQ. This fact indicates that the probability of missing data referring to IQ is clearly related to age. Another example might be the missing answers considering the body weight only in the women's respondents, so that we may consequently conclude that in this case missingness is related to sex. Such data obviously are not MCAR. But, if however, we know the sex of the respondents and if we can assume MCAR within the particular gender, then the data are MAR. Another example of MAR is when we take a sample from a population, where the probability of the data being inserted depends on some known property. Basically, missing data are missing at random (MAR) when the likelihood of missing data on a variable depends on some other measured variable in the model, but not to the value of the variable with missing values itself. Nevertheless, the assumption that the pattern is MAR is in practice very difficult to prove, so it is crucial to implement the correlates of missingness into the chosen missing data procedure in order to reduce bias and enhance the chances of satisfying the MAR assumption. Definitely, MAR is more general situation and therefore more realistic than MCAR. The largest number of the modern incomplete data tools generally start from the MAR hypothesis. Mathematically reduced, this data type can be express as follows:

$$P(R\big|X, \widetilde{X}) = P(R\big|X),$$

meaning that the realization of the R matrix will depend on the observed data only.

The third concept is called missing not at random (MNAR), although in the literature we can often notice the term NMAR (not missing at random) for the same model. MNAR indicates that the data likelihood of being missing differs for some unknown reasons. The fact is that in this particular case the missing values on a variable are dependent on the values of that variable itself, even after controlling all other variables. MNAR is the most complicated case for the researches. Approaches to overcome the MNAR situation are to reveal more detailes about the causes of the missingness or to carry out what-if analyses in order to evaluate the measure of subtleness of the results. The example which illustrates this type of the data is when the answers refer to IQ are missing only at the respondents with low IQ. Another illustration of this structure is that when the survey participants with serious depression are more likely to refuse to fulfill the answers referring to the depression severity. More, in public opinion research the MNAR appears when persons having infirm opinions answer less frequently. The difficulty with the MNAR structure is that it is unfeasible to prove that outcomes are MNAR without recognize the values that are missing. So, the trouble lies in the fact that the data incompleteness is totally related to the unobserved data, meaning to the incidences or components that are not evaluated and registered by the researcher.

The differences between these structures that are firmly described in [27] are crucial for realize why some techniques will offer better results against the others. His basic hypothesis lays in the fact that the researcher needs to provide the conditions under which a missing data method can produce valid statistical interpretations. Basic methods settle only the restrictive and sometimes implausible MCAR premise. Therefore, in this case we must have in mind that there is a substantial probability of obtaining biased estimates. Mostly, missing data are neither MCAR nor MNAR. Instead, the probability that an observation is missing commonly depends on information for that subject that is present, meaning that the reason for missingness is based on other observed respondent characteristics. This situation defines obviously the MAR model. For the additional description and comparison of the three basic patterns of the missing data see [33].

In order to illustrate an example taken from the real data, we used the result [18] given by Lai, who created the regression line and predict the voting intention by using peoples' age. Please see Figure 2.1 of scatter plots for the comparison of different types of the missing data. The model that we define in the Chapter 4 can be implemented on the MCAR type of the data.

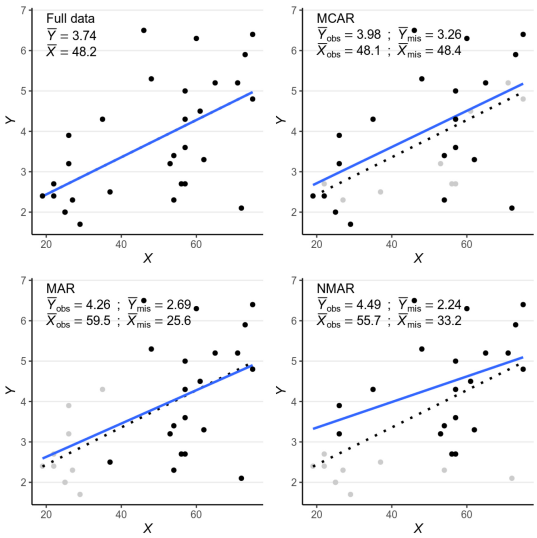

FIG. 2.1: Scatter plots of different types of missing data

3. The analysis of the incomplete sample regulation techniques

The crucial strategy in dealing with the missing data problem is to apply the data analysis techniques which are robust to the deviations caused by the incompleteness of the data set. This robustness of the technique practically means that there exists reliance that some smooth and tolerable violations of the premises and starting hypothesis will result in almost no bias or misinterpretation in the resulting outcomes based on the population under analysis. On the other hand, it needs to be pointed out that it is not achievable to use such methods in every situation. That is why a large number of different handling procedures for the missing data issues has been established.

According to [10], the methods for dealing with missing values can be evaluated by three means: it should yield to an unbiased parameter estimate, one should be able to obtain reasonable estimates of the standard error of confidence intervals and it should have good statistical power. Traditional missing data methods such as complete case analysis often produce bias and inaccurate conclusions. Similar problems extend to single imputation techniques commonly thought of as improvements over complete case methods. Research demonstrates that procedures such as multiple imputation, which incorporate uncertainty into estimates for missing data, often provide significant improvements over traditional methods.

Generally, the most commonly used procedures can be divided into three main groups which are explained thoroughly in next paragraphs: Deletion methods, Single Imputation Methods and Multiple Imputation methods.

3.1. Deletion methods

Listwise deletion stands for the basic method in overcoming the possible com-

plications caused by the incompleteness of the data set. This procedure is also called the Complete-case analysis. The conducting mechanism simply ignores all the cases which obtain one or more missing values recognizing the variables that are under examination and it is an inevitable part of many statistical softwares such as STATA, SPS, SAS etc.

The advantage of the listwise deletion method is its reliability, accuracy and its availability. Under hypothesis of MCAR data type, the listwise deletion produces the standard errors and significance levels absolutely acceptable referring to the reduced subset of data. But, we must note that these values are often higher when implement this technique using all possible data.

In real life situations various challenges occur. For instance, when the number of variables is huge and when more than a half of the original sample is obscured and vanished. More, dealing with structures that are not MCAR, the listwise deletion can severely bias the evalution of means, regression coefficients and correlations. It is showed in the study of Little and Rubin (see [20]) that the bias of the estimated mean grows together with the disparity among means of the observed and missing variables. Also, the bias grows with the higher percentage of the data that are missing. Interesting investigation on the subject was performed by Schafer and Graham (see [33]), where the bias of the complete-case analysis under MAR and MNAR premises was analyzed. It is important to imply that there are settlements in which listwise deletion can give better estimates than even the most refined and smooth statistical mechanisms. Miettinen (see [21]) indicates that this method states for the only access that guaranties that no bias is possible under any conditions. If we go further trough literature, Enders (see [7]) claims that in most settlements, the discommodities of listwise deletion far exceed its conveniences. Schafer and Graham (see [33]) show that only if the incompleteness problem can be solved by eliminating only a small part of the sample, then the technique may be solidly efficient. Vach (see [35]) claims that "there exists something like a critical missing rate up to which missing values are not too dangerous".

Another method, known as the Pairwise deletion (often called the available-case analysis) tries to improve the waist data problem of listwise deletion. In listwise deletion a case is ignored from a survey for the reason that it consists of one or more missing values within the variables under analysis. Pairwise deletion appears in the situations when statistical method accepts cases that involve some missing data. The technique cannot include the specific variable with a missing value into analysis, but it can still exploits the incomplete case when investigating other variables with complete values. The advantage of this procedure is that it increments a power of the survey. On the other hand, it has certain deficiencies. It presumes that the incomplete sample is MCAR.

The illustration for understanding the mechanism of the method of pairwise deletion is to take a dataset having following variables: age, gender, education, income, and political affiliation. For each case in the dataset, the values of some of the variables are more likely to be missing than others depending on the surveyee's sensitiveness to the survey questions. Let's say we are interested in knowing if

there is a correlation between age and political affiliation. Using pairwise deletion, any given case may contribute to certain analysis but not to others, depending on whether the needed data are available. Hence for our analysis in this example, all cases with available data on age and political affiliation will be included regardless of the missing values for other variables like gender, income or education. The pairwise deletion is an alternative to the listwise deletion to mitigate the loss of data.

3.2. Imputation methods

Other routine way that is frequently practiced among the statisticians is imputation. This method basically replaces the missing values with certain estimated values and then it analysis the complete data set such that it treats the imputed estimates as the original observed values. The procedures for the best choice of these estimates differ and in this paragraph we describe the most exploited imputations that are used in surveys. The imputation procedures are divided in two groups: single imputation methods and multiple imputation methods.

3.2.1. Single imputation

In single imputation, missing values are replaced by a value defined by a certain rule. For example, Mean imputation is a smooth and simple method which evaluates the mean of the observed values for the particular variable in all cases that are not missing. Conceivably, the preference of this technique is that it retains the same sample size and the same mean. On the other hand, mean substitution reduces the variation of analyzed scores and this reduction in separate variables is proportional to the number of missing data. Further, mean substitution may significantly transform the values of correlations. The regression imputation is a procedure which utilizes the values of other variables in order to forecast the missing values in a variable. That is achieved by applying a regression model. Usually the regression model is structured by using the observed data and eventually related to the regression weights the missing values are projected and restored.

Next example of the single imputation is the Hot-deck imputation, the technique which inserts a missing value from a randomly selected similar data set. The part of the expression "deck" suggests that the contributed values arrive from the identical set as the initial data-set. The term "hot" in the above phrase is for the reason of data being instantly employed.

On the contrary to the last method, the cold-deck imputation chooses contributors data belonging to a different data-set. It is a term for a technique that fills a missing values with values from some outward origin, such as some previous similar survey. According to the above explanation, the reason for the expression "cold-deck" is evident.

3.2.2. Multiple imputation

Multiple imputation methods use the distribution of the observed data in order to estimate multiple values that catch the oscillations around the true value. The

idea of multiple imputation (MI) was first introduced by Rubin (see [28]), in which each missing value is replaced with $m > 1$ simulated values prior to analysis. In multiple imputation, there are three operational steps: imputation or fill-in phase, the analysis phase and pooling phase. First phase constitute the complete data set by filling in the missing values with the estimated values (using some of the convenient statistical methods). This process of fill-in repeats several times. The analysis phase, studies each of the obtained complete data sets by using a suitable statistical method. Finally, in the third step the parameter estimates resulted from each of the considered data set are then connected and analyzed so that the best conclusions can are accomplished. Final phase aggregates all the results and reveals the best summary estimate of the missing data. Clearly, it is obvious that the method of multiple imputation is more unbiased that the single imputation method, because of the use of multiple sets. That way we kind of "washing" out the coincidences that might occur. The disadvantage of this approach is the greater expanse of time and effort comparing to single imputation.

The most familiar and widely exploited model-based method is the EM algorithm described thoroughly by Dempster et al (see [4]). Also, high influential articles given by Rubin (see [27]) and Little and Rubin (see [19]), gave the formulation of EM algorithm and the dominated framework for dealing with missing data. Many examples of EM algorithm were provided by Little and Rubin (see [20]) and Schafer (see [30]).This iterative technique involves the expectation (E-part) and the maximization (M-part). It replaces missing data with estimated values, evaluates the parameters, repeatedly estimates the missing values, re-estimates the parameters and iterates until convergence (see [20]). Over the repetitions until convergence, we conclusively obtain the missing values.

To simplify this approach, let us assume that the complete data-set consists of $\mathbb{X} = \{X, \widetilde{X}\}$ but that only X is observed. The complete-data log likelihood function is then denoted by $l(\theta; X, \widetilde{X})$ where θ is the unknown parameter vector for which we need to find the MLE (which is based on EM algorithm). Further, let $t = 1, 2, ...$ represents all parameters of distribution and $f_{\theta_t}(X)$ and $f_{\theta_t}(\widetilde{X})$ are the assumed probability distributions at t-th iteration. First, the E-part is activated and evaluates the expected value of $l(\theta; X, \widetilde{X})$ given the observed data X and the current iteration parameter estimate θ.

Principally, we define

$$Q(\theta; \theta_t) := E[l(\theta; X, \widetilde{X})|X, \theta_t] = \int l(\theta; X, \widetilde{X}) p(\widetilde{x}|X, \theta_t) dx, \tag{3.1}$$

where $p(\cdot|X, \theta_t)$ is the conditional density of $\widetilde{X}$ given the observed data X and assuming $\theta = \theta_t$.

Next, the M-part of the analysis starts and it maximizes the expectation (3.1) over θ. That is we put:

$$\theta^t := \max_{\{\theta\}} Q(\theta; \theta_t).$$

We then set $\theta_t = \theta^t$. The two steps are iterated until the sequence of θ^t converges.

Recent work implies that multiple imputation and specialized modeling procedures offer universal methods for handling the missing data. It is proven that they perform fine over many types of missing data structures. There are different EM algorithms for different applications. Although this method provides excellent parameter estimates, EM is not particularly good for hypothesis testing.

Nevertheless, the development of informational technology and the advances in relevant statistical software make these methods available to the researchers in various fields. For example, multiple imputation procedures under the normal model are implemented in Schafer's NORM program [30]. Detailed, step-by-step instructions for running NORM are available in [12] (also see [11], [31], [32]). ML methods, often called FIML (full information maximum likelihood) methods deal with the missing data, do parameter estimation, and estimate standard errors all in a single step. Available software for running this procedure are AMOS: [1], LISREL: [15]; also see Mplus: [24]; and Mx: [25]. Basically, in 1987. Little and Rubin published their classical book *Statistical Analysis With Missing Data* (see [19]), and they established the groundwork for missing data software to be developed over the next 20 years and beyond. See also [13] for recent review of software handling missing data.

4. Mathematical model generated for the MCAR type of data

Let $X_1, X_2, ...$ be independent identically distributed random variables and let us assume that only observations at certain points are available. Denote the observed random variables among $\{X_1, \ldots, X_n\}$ by $\widetilde{X}_1$, ..., $\widetilde{X}_{M_n}$. Here the random variable M_n represents the number of the registrated random variables among the first n terms of the sequence (X_n). Incomplete sample may be obtained, for example, if every term of (X_n) is observed with probability p, independently of other terms, and in this case M_n is binomial random variable. This refers to MCAR type of missing data distribution. Now, let:

$$E\left(X_j\right) = m, \quad D\left(X_j\right) = \sigma^2 \quad \textit{and} \quad S(n) = \sum_{j=1}^{M_n} \widetilde{X}_j.$$

We obtain the following results straightforward:

$$\begin{aligned} E\left(S(n)\right) &= \sum_{k=0}^{\infty} E\left(S(n)\big|M_n = k\right) \cdot P\{M_n = k\} \\ &= \sum_{k=0}^{\infty} E\left(\sum_{j=1}^{M_n} \widetilde{X}_j \big| M_n = k\right) \cdot P\{M_n = k\} \\ &= \sum_{k=0}^{\infty} E\left(\sum_{j=1}^{k} \widetilde{X}_j\right) \cdot P\{M_n = k\} = \sum_{k=0}^{\infty} k \cdot m \cdot P\{M_n = k\}. \end{aligned}$$

Conclusively we have:

$$E\left(S(n)\right) = m \cdot E\left(M_n\right) = E(X_1) \cdot E\left(M_n\right). \tag{4.1}$$

Further we have that:

$$\begin{aligned}
D\left(S(n)\right) &= E\left(S(n)\right)^2 - \left(E(S(n)\right)^2 = E\left(S(n)\right)^2 - m^2\left(E\left(M_n\right)\right)^2 \\
&= E\left(\sum_{j=1}^{M_n} \widetilde{X}_j\right)^2 - m^2\left(E\left(M_n\right)\right)^2 \\
&= \sum_{k=0}^{\infty} E\left\{\left(\sum_{j=1}^{M_n} \widetilde{X}_j\right)^2 \Big| M_n = k\right\} \cdot P\{M_n = k\} - m^2\left(E\left(M_n\right)\right)^2 \\
&= \sum_{k=0}^{\infty} E\left(\sum_{j=1}^{k} \widetilde{X}_j\right)^2 \cdot P\{M_n = k\} - m^2\left(E\left(M_n\right)\right)^2 \\
&= \sum_{k=0}^{\infty} \left\{D\left(\sum_{j=1}^{k} \widetilde{X}_j\right) + \left(\sum_{j=1}^{k} E(\widetilde{X}_j)\right)^2\right\} \cdot P\{M_n = k\} - m^2\left(E\left(M_n\right)\right)^2 \\
&= \sum_{k=0}^{\infty} (k\sigma^2 + k^2 m^2) \cdot P\{M_n = k\} - m^2\left(E\left(M_n\right)\right)^2 \\
&= \sigma^2 E(M_n) + m^2 E(M_n)^2 - m^2\left(E\left(M_n\right)\right)^2 \\
&= \sigma^2 E(M_n) + m^2 D(M_n).
\end{aligned}$$

Since we assumed that $X_1, X_2, \ldots$ are identically distributed, the last equality we can write as:

$$D\left(S(n)\right) = D(X_1)E(M_n) + E(X_1)^2 D(M_n). \tag{4.2}$$

If M_n has a binomial distribution with parameters n and p where p is the probability of a successful outcome, i.e the probability of a variable to be observed. If we put $q = 1 - p$ the probability of failure, that is the probability of a variable to be missing we have the equations (4.1) and (4.2) written in the form:

$$E\left(S(n)\right) = mnp$$

and

$$D\left(S(n)\right) = \sigma^2 np + m^2 npq = np(\sigma^2 + m^2 + q).$$

Further, it is possible to extend the application of the proposed model in the case when the observed random variables are determined by a general point process

and when only conditions on M_n are imposed. It may be interesting to see the implementation of the proposed mathematical model based on a strictly stationary sequence of random variables $(X_n)_{n\geqslant 1}$ with "short range" dependence. This problem was considered and analyzed by Mladenovic and Piterbarg (see [23]) where consistency of Hill's estimator was proved. The main presumed condition in the paper means that the finite dimensional distributions of (X_n) are invariant under shifts and the dependence between observations from (X_n) becomes weaker as time separation becomes larger. More, under additional conditions this model of incompleteness was considered by Ilic and Mladenovic (see [14]), where the asymptotic behavior of the Pareto index estimator, proposed by Bacro and Brito (see [2]), was analyzed. Also it it can be proved that in the case when the number of observed variables M_n has the binomial distribution the sequence $\widetilde{X}_1$, ..., $\widetilde{X}_{M_n}$ of observed variables is asymptotically stationary (according to the definition from [6]). The proposed model can be used for various practical situations where more thorough theoretical tool is necessary in order to describe the incompleteness of the data. It can be interesting for the researchers in this area for the mathematical establishment of certain incomplete structures in the surveys.

Finally, we give the necessary conditions that are used in the above research papers in order to enhance the mathematical approach in confronting with the missing data in stationary sequences.

Assumption A. The sequence X_1, X_2, ... does not depend on M_n and

$$\frac{M_n}{n} \xrightarrow{p} c_0 > 0 \quad \text{as } n \to +\infty.$$

Suppose β_n is a sequence of real numbers such that

$$\lim_{n\to\infty} \beta_n = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{\beta_n}{n} = 0.$$

Let

$$K_n = \left[\frac{M_n}{\beta_n}\right] \quad \text{and} \quad B_n = \begin{cases} 0, & M_n = 0 \\ \frac{K_n}{M_n}, & M_n \geq 1 \end{cases}$$

where the floor function [.] denotes the largest previous integer. Define $\widetilde{Y}_i = (\ln \widetilde{X}_i - \ln F^{-1}(1-B_n))_+$ and $\widetilde{Y}_i^{\zeta} = I\left\{\ln \widetilde{X}_i - \ln F^{-1}(1-B_n) > \frac{\zeta}{\sqrt{K_n}}\right\}$ where $\zeta \in R$.

Assumption B. For any $h \in N$ and $\theta \in R$

$$Var\left\{\sum_{j=1}^{h}\left((\widetilde{Y}_{j+k} - E\widetilde{Y}_{j+k}) + \theta(\widetilde{Y}_{j+k}^{\zeta} - E\widetilde{Y}_{j+k}^{\zeta})\right)\right\}$$

does not depend on k.

Remark 4.1. In the case when the number of observed variables M_n has the binomial distribution both the **Assumption A** and **Assumption B** are satisfied. In this case the sequence $\widetilde{X}_1$, ..., $\widetilde{X}_{M_n}$ of observed variables is asymptotically stationary, according to the definition from [6].

5. Conclusion

Missing data is an intermittent issue in many areas such as: market research, database analysis, social analysis, medical research and generally in survey research. Even a small percent of missing data can produce significant problems in the statistical analysis possibly leading to wrong conclusions. The purpose of this article is to identify the problem, to recognize the missing data pattern and to choose the proper methodology for dealing with the incomplete sample. Further intention of this paper is to indicate the possibility of the potential implementation of the proposed mathematical formulation in statistical researches having the MCAR data structure. Prospective research will undeniably derive further improvements and expansions of the proposed mathematical models and practical techniques in order to achieve higher efficiency in situations in which missing data appear.

Acknowledgements

Authors were supported in part by Ministry of Education, Science and Technological Development, Republic of Serbia (Grant No. 174025, 174007, III-43007, TR-32012, 144)

REFERENCES

1. J. L. Arbuckle and W. Wothke: *Amos 4.0 User Guide*. Smallwaters, Chicago IL, 1999.
2. J. N. Bacro and M. Brito: *Strong limiting behaviour of a simple tail Pareto-index estimator*. Statistics and Decisions, **3** (1993) 133–134.
3. S. R. Browning and B. R. Browning: *Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering*. The American Journal of Human Genetics, **81** (2007) 1084–1097.
4. A. P. Dempster, N. M. Laird, D. B. Rubin: *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*. J. R. Stat. Soc., **B39** (1977) 1–38.
5. S. Dilmaghani, H. Isaacc, P. Soonthornnonda, E. Christensen, C. H. Ronald: *Harmonic Analysis of Environmental Time Series with Missing Data or Irregular Sample Spacing*. Environ. Sci. Technol., **41** (2007) 7030–7038.
6. W. Dunsmuir and P. M. Robinson: *Asymptotic theory for time series containing and amplitude modulated observations*. The Indian Jornal of Statistics, Series A, **43(3)** (1981) 260–281.
7. C. K. Enders: *Missing not at random models for latent curve analyses*. Manuscript submitted for publication, 2010.
8. D. Giannone, L. Reichlin, D. Small: *Nowcasting: The real-time informational content of macroeconomic data*. Journal of Monetary Economics No., **55** (2008) 665–676.

9. M. Gottschalk, M. E. Tinetti, D. I. Baker, M. King, E. M. Terrence, D. Acampora, L. Leo-Summers, H. G. Allore : *Effect of Dissemination of Evidence in Reducing Injuries from Falls.* New England Journal of Medicine, **359** (2008) 252–261.
10. J W. Graham: *Missing data analysis: Making it work in the real world.* Annu. Rev. Psychol., **60** (2009) 549–576.
11. J. W. Graham, S. M. Hofer: *Multiple imputation in multivariate research.* In Modeling Longitudinal and Multiple-Group Data: Practical Issues, Applied Approaches, and Specific Examples, Eds. T. D. Little,K. U. Schnabel, J. Baumert. Hillsdale, NJ: Erlbaum, 2000, pp. 201–218.
12. J. W. Graham, P. E. Cumsille, E. Elek-Fisk: *Methods for handling missing data.* In Research Methods in Psychology, Eds. J. A. Schinka, W. F. Velicer, (pp. 87–114). Volume 2 of Handbook of Psychology, Ed. IB Weiner. New York:Wiley 2003.
13. N. J. Horton, K. P. Kleinman: *Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models.* Am. Stat. **61** (2007) 79–90.
14. I. Ilić and P. Mladenović: *On tail index estimation using a sample with missing observations.* Statistics and Probability Letters, **82** (2012) 949–958.
15. K. G. Jöreskog and D. Sörbom: *LISREL 8 Users Reference Guide.* Sci. Software, Chicago, 1996.
16. P. Kline and A. Santos: *Sensitivity to Missing Data Assumptions: Theory and An Evaluation of the U.S. Wage Structure.* In: Proceedings of a 2010 Seoul Summer Economics Conference, 2010.
17. L. Koopman, J. M. G. van der Heiden, D. E. Grobbee and M. M. Rovers: *Comparison of Methods of Handling Missing Data in Individual Patient Data Meta-analyses: An Empirical Example on Antibiotics in Children with Acute Otitis Media.* American Journal of Epidemiology, **167(5)** (2007) 540-545.
18. M. Lai: Course Handouts for Bayesian Data Analysis Class, 2020.
19. R. J. A. Little and D. B. Rubin: *Statistical Analysis with Missing Data.* Wiley, New York, 1987.
20. R. J. A. Little and D. B. Rubin: *Statistical Analysis with Missing Data.* Wiley, New York, 2n ed., 2002.
21. O. S. Miettinen: *Theoretical Epidemiology: Principles of Occurence Research in Medicine.* John Wiley and Sons, New York, 1985.
22. P. Mladenović and Z. Petrovic: *Cagan's paradox and money demand in hyperinflation:Revisited at daily frequency.* Journal of International Money and Finance, **29** (2010) 1369–1384.
23. P. Mladenović and V. Piterbarg: *On estimation of the exponent of regular variation using a sample with missing observations.* Statist. Prob. Letters. **78** (2008) 327–335.
24. L. K. Muthén and B. O. V. Muthén: *Mplus User's Guide.* CA: Muthén and Muthén, Los Angeles, 4th ed., 2007.
25. M. C. Neale, S. M. Boker, G. Xie, H. H. Maes: *Mx: Statistical Modeling.* Virginia Commonwealth Univ. Dept. Psychiatry., Richmond, 5th ed., 1999.

26. F. J. Prevosti and M. A. Chemisquy: *The impact of missing data on real morphological phylogenies: influence of the number and distribution of missing entries.* Cladistics, **26** (2009) 326–339.

27. D. B. Rubin: *Inference and missing data.* Biometrika, **63** (1976) 581–592.

28. D. B. Rubin: *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York, 1987.

29. R. J. Ryder and G. K. Nicholls: *Missing data in a stochastic Dollo model for cognate data, and its application to the dating of Proto-Indo-European.* The Smithsonian/NASA Astrophysics Data System,eprint arXiv:0908.1735, 2009.

30. J. L. Schafer: *Analysis of Incomplete Multivariate Data.* Chapman and Hall, New York, 1997.

31. J. L. Schafer: *Multiple imputation: a primer.* Stat. Methods Med. Res. **8** (1999) 3–15.

32. J. L. Schafer, M. K. Olsen: *Multiple imputation for multivariate missing data problems: a data analyst's perspective.* Multivar. Behav. Res., **33** (1998) 545–571.

33. J. L. Schafer, J. W. Graham: *Missing data: our view of the state of the art.* Psychol. Methods, **7** (2002) 147–177.

34. C. Schumacher, J. Breitung: *Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data.* International Journal of Forecasting, **24** (2008) 386–398.

35. W. Vach: *Logistic Regression with Missing Values in the Covariates.* Springer, New York, 1994.