UDK 519.2:005.5
DOI: 10.7251/FIN2104031M

ORIGINALNI NAUČNI RAD

Elvis Mujkić*, Jelena Poljašević**

Višestruka imputacija kao metod eliminacije nedostajućih podataka u SPSS-u

Multiple imputations as a method of elimination of missing data in SPSS

Rezime

Nedostajući podaci javljaju se u svim oblastima istraživanja, a naročito u oblasti društvenih nauka. Kao takvi, mogu smanjiti statističku moć istraživanja i proizvesti pristrasne procjene, što može rezultirati neadekvatnim zaključcima. U ovom radu dat je pregled obrazaca i mehanizama po kojima nedostajući podaci mogu da nedostaju u istraživanju. Pored navedenog, u radu su predstavljene tradicionalne i savremene metode koje se mogu koristiti za eliminaciju nedostajućih podataka i ukazuje se na prednosti i nedostatke jedne i druge grupe metoda. Zbog visoke pristrasnosti pri procjeni parametara koju izazivaju tradicionalne metode tretmana nedostajućih podatka, kao što su metode brisanja nedostajućih podataka u cijelini ili u paru, metode jednostruke imputacije, preporučuje se primjena savremenih metoda, kao što je metoda višestruke imputacije. S obzirom na to, u radu je dat primjer sprovođenja višestruke imputacije u SPSS programu.

Ključne riječi: *nedostajući podaci, MCAR, MAR, NMAR, višestruka imputacija, SPSS.*

Abstract

Missing data appear in all areas of research, and the client is in the field of social sciences. As such, they can reduce the statistical power of research and produce biased estimates, which can result in inadequate conclusions. This paper provides an overview of the patterns and mechanisms by which missing data may be missing in research. In addition to the above, the paper presents traditional and modern methods that can be used to eliminate missing data and points out the advantages and disadvantages of both groups of methods. It is recommended to use modern methods - such as multiple imputation methods, due to the high bias in the assessment of parameters caused by traditional methods of missing data treatment, such as methods of deleting missing data in whole or in pairs or single imputation methods.With that in mind, this paper gives an example of conducting multiple imputation in the SPSS program.

Keywords: *missing data, MCAR, MAR, NMAR, multiple imputation, SPSS*

* Molson Coors BH d.o.o Banja Luka, e-mail: elvismujkic73@gmail.com
** Vanredni profesor Ekonomskog fakulteta Univerziteta u Banjoj Luci, e-mail: jelena.poljasevic@efbl.org

UVOD

U različitim oblastima istraživanja prikupljaju se različiti podaci u cilju donošenja određenih zaključaka na osnovu analize prikupljenih podataka u uzorku. Zajednički sadržalac svih istraživanja, bez obzira na to o kojoj naučnoj oblasti se radi, jesu nedostajući podaci. Nedostajući podaci najčešće se definišu kao vrijednosti koje nisu evidentirane ili zabilježene za određenu posmatranu varijablu. Kao takvi, u velikoj mjeri otežavaju obradu i analizu prikupljenih podataka, a time posljedično utiču i na zaključke koji se donose na osnovu uzorka. Uzimajući u obzir navedeno, predmet ovog rada jeste problem nedostajućih podataka u istraživanju. Cilj rada je predstaviti praktičnu primjenu metode višestruke imputacije, kao jedne od savremenih metoda za rješavanje problema nedostajućih podataka, u SPSS programu.

1. PREGLED LITERATURE

1.1. Pojam, obrasci i mehanizmi nedostajućih podataka

Nedostajući podaci (engl. Missing data) definišu se kao vrijednosti podataka koji nisu pohranjeni za datu posmatranu varijablu. Kao takvi, uobičajna su pojava u gotovo svim oblastima istraživanja, a naročito u oblasti društvenih nauka (Kang, 2013). Nedostajući podaci se, prema Graham i saradnicima (Graham, Cumsile, Elek-Fisk, 2003), javljaju iz dva razloga. Jedan razlog je da ispitanici koji učestvuju u određenom istraživanju ne odgovaraju iz neodređenog razloga na neka pitanja. Drugi razlog je taj što zbog osipanja uzorka u longitudinalnim istraživanjima dolazi do pojave nedostajućih čitavih kompleta podataka.

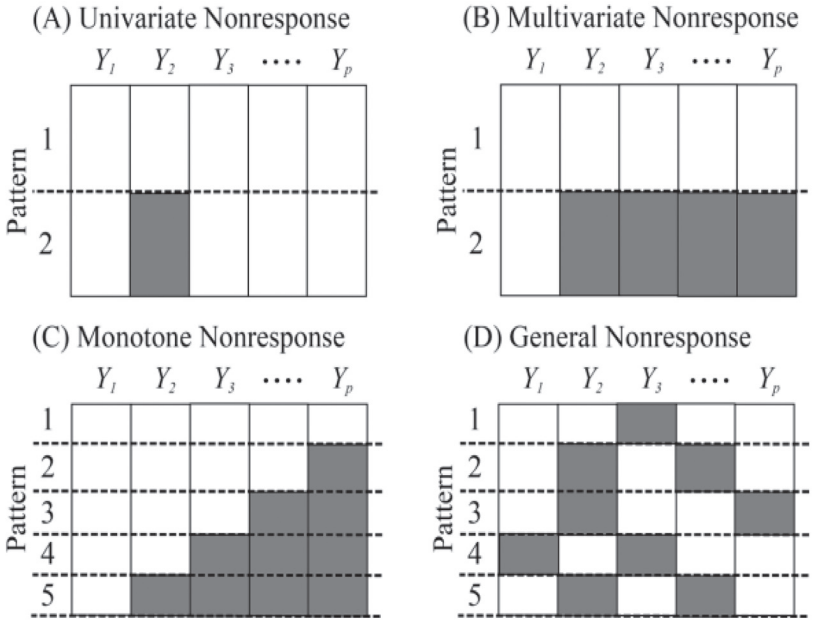
Problematika nedostajućih podataka zavisi od samog tipa podataka. Ukoliko su podaci nominalne kategorije, tim je teže predvidjeti, zamijeniti ili nadomjestiti nedostajuće podatke. Pored samog oblika distribucije i tipa nedostajućih podataka, važan faktor je i veličina uzorka u istraživanju, a time i količina nedostajućih podataka, broj i tip ostalih prikupljenih podatka itd. (Oblaković, Sokolovska, Dinić, 2015).

Kada je riječ o količini nedostajućih podataka, u literaturi ne postoji utvrđena granica u pogledu prihvatljivog postotka podatka koji nedostaju u uzorku, a koji ne bi uticao na statističke zaključke izvedene iz tog uzorka. Shodno tome, Shcafer (1999) je utvrdio da procenat nedostajućih podatka u uzorku od 5% ili manje ne utiče na zaključke koji se donesu na osnovu takvog uzorka. Sa druge strane, Bennett (2001) ističe da je statistička analiza najvjerovatnije pristrasna ukoliko u uzorku nedostaje više od 10% podataka. Tabachnick i Fidell (2012) utvrdili su da mehanizmi i obrasci nedostajućih podataka imaju veći uticaj na rezultate istraživanja nego udio podataka koji nedostaje u uzorku.

Obrasci podataka koji nedostaju mogu se podijeliti na:

– jednovarijantne i multivarijantne obrasce,
– monotone i nemonotone obrasce,
– opšti obrazac.

Slika 1. *Obrasci nedostajućih podataka*



Izvor: Howard, 2013.

Jednovarijantni obrasci nedostajućih podataka označavaju situaciju u kojoj se podaci koji nedostaju pojavljuju samo na jednoj varijabli, što je predstavljeno panelom A na slici broj 1. Ukoliko se nedostajući podaci pojavljuju na više varijabli, tada govorimo o multivarijantnom obrascu nedostajućih podataka, što se vidi sa panela B. Ako podatak za određenu varijablu nedostaje ne samo u određenoj vremenskoj tački nego i u svim kasnijim vremenskim prilikama, za obrazac podataka koji nedostaju za datu varijablu kaže se da je monotoni, što se vidi na panelu C. Sa druge strane, ukoliko se nedostajući podatak uoči na određenoj varijabli u datom vremenskom trenutku, a zatim se u nekom kasnijem vremenskom trenutku pojavi, odnosno ne bude više nedostajući podatak, tada za takav obrazac kažemo da je nemonotoni (Xian, 2016).

Kada je riječ o mehanizmima nedostajućih podatka, prvi koji je ukazao na mehanizme nedostajućih podataka bio je Rubin, koji je definisao tri osnovna mehanizma nedostajanja podataka (Rubin, Little, 2002):

1. potpuno slučajno nedostajući podaci (engl. missing completely at random – MCAR),

INTRODUCTION

Different data are collected in different areas of research in order to draw certain conclusions based on the analysis of the data collected in the sample. The common denominator of all research, regardless of the scientific field, is the missing data. Missing data are most often defined as values that are not recorded for a particular observed variable. As such, they greatly complicate the processing and analysis of the collected data, and thus consequently influence the conclusions made on the basis of the sample. Taking into account the above, the subject of this paper is the problem of lack of data in the research.The aim of this paper is to present the practical application of the multiple imputation method, as one of the modern methods for solving the problem of missing data, in the SPSS program.

1. LITERATURE REVIEW

1.1. The concept, patterns and mechanisms of missing data

Missing data is defined as the value of data that is not stored for a given observed variable. As such, they are common in almost all areas of research, especially in the social sciences (Kang, 2013). According to Graham et al. (Graham, Cumsile, & Elek-Fisk, 2003), missing data appear for two reasons.

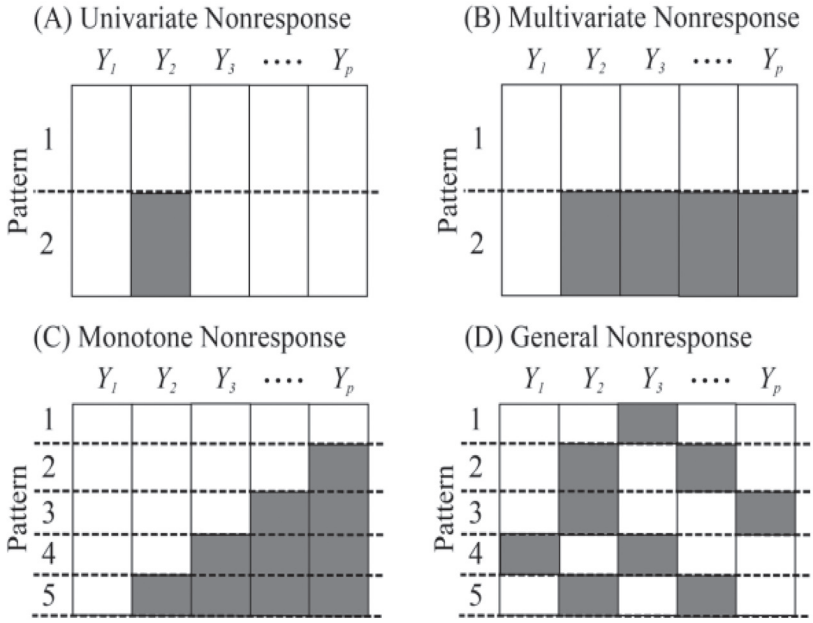
One reason is that the respondents who participate in a certain research do not answer some of the questions for an indefinite reason. Another reason is the lack of missing complete data sets due to sample scattering in longitudinal studies.

The problem of missing data depends on the type of data itself. If the data is of a nominal category, it is more difficult to predict, replace or substitute the missing data. In addition to the form of distribution and the type of missing data, an important factor is the size of the sample in the research, and thus the amount of missing data, the number and type of other collected data, etc. (Oblaković, Sokolovska, & Dinić, 2015). When it comes to the amount of missing data, there is no established limit in the literature regarding the acceptable percentage of missing data in a sample that would not affect the statistical conclusions derived from that sample.Accordingly, Shcafer (1999) found that the percentage of missing data in a sample of 5% or less does not affect the conclusions drawn from such a sample. On the other hand, Bennett (2001) points out that statistical analysis is most likely biased if more than 10% of data are missing from a sample. Tabachnick and Fidell (2012) found that the mechanisms and patterns of missing data have a greater impact on research results than the share of missing data in a sample.

Missing data forms can be divided into:

– single and multivariate forms,
– monotonous and non-monotonous patterns,
– general pattern.

Picture 1 *Patterns of missing data*



Source: Howard, 2013.

One-variant missing data patterns indicate a situation in which missing data appear on only one variable, as presented with panel A in Figure 1. If the missing data appear on more than one variable, then we are talking about a multivariate pattern of missing data, which can be seen in panel B. If the data for a certain variable is missing not only at a certain time point, but also in all subsequent time conditions, the missing data pattern for a given variable is said to be monotonous, as seen in panel C. On the other hand, if the missing data is observed on a certain variable at a given time and then at a later time it appears, or is no longer missing data, then we say that such a pattern is non-monotonic (Xian, 2016).

When it comes to missing data mechanisms, the first to point out missing data mechanisms was Rubin, who defined three basic mechanisms of missing data (Rubin & Little, 2002):

1. missing completely at random (MCAR),
2. missing at random (MAR) and
3. not missing at random data (NMAR).

The Accidentally Missing Data Mechanism (MCAR) implies that the probability of missing data on an observed variable is not related to other variables or to the values of a given variable Y

2. slučajno nedostajući podaci (engl. missing at random – MAR) i
3. podaci koji ne nedostaju slučajno (engl. not missing at random – NMAR).

Mehanizam potpuno slučajno nedostajućih podataka (MCAR) podrazumijeva da vjerovatnoća nedostatka podataka na posmatranoj varijabli nije povezana s drugim varijablama niti sa vrijednostima date varijable Y (Enders, 2010).

Podaci koji nedostaju po MCAR mehanizmu ne prate nikakav obrazac na osnovu koga bi se mogla predvidjeti vrijednost koja nedostaje, odnosno, kako navodi Kang (2013), podaci koji nedostaju po MCAR mehanizmu nisu povezani ni sa specifičnom vrijednošću koja bi trebalo da se dobije ni sa podacima iz skupa posmatranih varijabli. U nastavku je dat primjer mehanizma potpuno slučajno nedostajućih podataka (MCAR).

Tabela 1. *MCAR mehanizam*

A	B	C	D
-	1	0	1
1	0	-	1
0	-	1	0
1	0	-	1
0	0	1	-

Izvor: prilagođeno prema Warnes, 2021.

Kažemo da podaci nedostaju po NMAR mehanizmu, odnosno da podaci ne nedostaju slučajno, ukoliko distribucija nedostajućih podataka zavisi od samih nedostajućih podataka (Schafer, Graham, 2002). Za razliku od MAR mehanizma, kod NMAR mehanizma postojeći podaci u skupu podataka ne pružaju dovoljno informacija da se na osnovu njih izvrši adekvatna aproksimacija podataka koji nedostaju (Lang, Little, 2018). U nastavku je dat primjer NMAR mehanizma.

Tabela 2. *NMAR mehanizam*

A	B	C	D
1	1	0	1
1	0	0	1
0	1	-	0
1	0	0	1
0	0	-	1

Izvor: prilagođeno prema Warnes, 2021.

Za razliku od MCAR mehanizma, nedostajanje podataka po MAR mehanizmu podrazumijeva da nedostajući podaci nisu pod uticajem neke varijable u kojoj ima nedostajućih podataka, ali jesu pod uticajem neke druge varijable u skupu podataka (Oblaković, Sokolovska, Dinić, 2015). Dakle, MAR mehanizam ne pretpostavlja da se na osnovu drugih varijabli u posmatranom skupu podataka ne mogu predvidjeti podaci koji nedostaju.

Naime, kako navodi Oblaković i dr. (2015), nedostajući podaci u slučaju MAR mehanizma mogu se objasniti na osnovu raspoloživih podataka, jer su varijable koje su povezane sa uzorkom izmjerene i mogu se uključiti u model u cilju dobijanja nepristrasnih procjena parametara. U nastavku je dat primjer MAR mehanizma.

Tabela 3. *MAR mehanizam*

A	B	C	D
1	1	0	-
1	0	0	-
0	1	1	-
1	0	0	1
0	0	1	1

Izvor: prilagođeno prema Warnes, 2021.

Na osnovu distribucije postojećih podatka na varijablama A, B i C može se predvidjeti vrijednost nedostajućih podataka na varijabli D. Uzimajući u obzir navedeno, vrijednosti nedostajućih podataka na varijabli D mogle bi biti 0, 1, 0.

1.2. Tretmani nedostajućih podataka

Postoje različiti tretmani nedostajućih podataka, koji se u osnovi mogu svrstati u dvije grupe:

– tradicionalni tretmani i
– moderni tretmani.

Tradicionalni tretmani obuhvataju metode kao što su isključivanje nedostajućih podatka i jednostruke imputacije, a moderni tretmani obuhvataju metode kao što su metode zasnovane na maksimalnoj vjerodostojnosti, metode višestruke imputacije i dr. (Oblaković, Sokolovska, Dinić, 2015).

1.2.1. Tretmani temeljeni na brisanju nedostajućih podataka

Tretmani temeljeni na brisanju nedostajućih podatka javljaju se u dvije varijante (Kang, 2013):

– brisanje, odnosno isključivanje nedostajućih podataka u cjelini (engl. listwise deletion), tzv. pametno brisanje sa liste, i

(Enders, 2010). Missing data according to the MCAR mechanism do not follow any pattern on the basis of which the missing value could be predicted, i.e. according to Kang (2013), the missing data according to the MCAR mechanism are not related to the specific value to be obtained, nor to the data from set of observed variables. The following is an example of a missing completely at random data (MCAR).

Table 1 *MCAR Mechanism*

A	B	C	D
-	1	0	1
1	0	-	1
0	-	1	0
1	0	-	1
0	0	1	-

Source: adjusted according to Warnes, 2021.

We say that data is missing by the NMAR mechanism if the distribution of missing data depends on the missing data themselves (Schafer & Graham, 2002). Unlike the MAR mechanism, the NMAR mechanism does not provide enough data in the data set to adequately approximate the missing data (Lang & Little, 2018). An example of the NMAR mechanism is given below.

Table 2 NMAR mechanism

A	B	C	D
1	1	0	1
1	0	0	1
0	1	-	0
1	0	0	1
0	0	-	1

Source: adjusted according to Warnes, 2021.

Unlike the MCAR mechanism, the lack of data according to the MAR mechanism implies that missing data are not influenced by a variable in which there is missing data, but are influenced by some other variable in the data set (Oblaković, Sokolovska, & Dinić, 2015). Thus, the MAR mechanism does not assume that the missing data cannot be predicted on the basis of other variables in the observed data set.

Namely, as stated by Oblaković et al. (2015) missing data in the case of the MAR mechanism can be explained on the basis of available data, as the variables associated with the sample have been modified and can be included in the model in order to obtain unbiased parameter estimates. An example of the MAR mechanism is given below.

Table 3 *MAR mechanism*

A	B	C	D
1	1	0	-
1	0	0	-
0	1	1	-
1	0	0	1
0	0	1	1

Source: adjusted according to Warnes, 2021.

Based on the distribution of existing data on variables A, B and C, the value of missing subtactans on variable D can be predicted. Taking this into account, the value of the missing data on variable D could be 0,1,0.

1.2. Missing data treatments

There are different treatments for missing data, which can basically be divided into two groups:

– traditional treatments and

– modern treatments.

Traditional treatments include methods such as exclusion of missing data and single imputation, and modern treatments include methods such as methods based on maximum credibility, multiple imputation methods, etc. (Oblaković, Sokolovska, & Dinić, 2015).

1.2.1. *Treatments based on deleting missing data*

Treatments based on deleting missing data occur in two variants (Kang, 2013):

– deletion, i.e. exclusion of missing data in its entirety (listwise deletion), the so-called smart delete from list and

– deletion, i.e. exclusion of missing data by pairs (pairwise deletion).

Listwise deletion means deleting observations on which at least one variable has missing data. This means that further analysis

– brisanje, odnosno isključivanje nedostajućih podataka po parovima (engl. pairwise deletion).

Isključivanje nedostajućih podataka u cjelini (engl. listwise deletion) podrazumijeva brisanje opažanja na kojima barem jedna varijabla ima nedostajući podatak. To znači da se dalja analiza vrši na osnovu onih opažanja koja nemaju nijedan nedostajući podatak. Sa druge strane, isključivanje nedostajućih podataka po parovima podrazumijeva isključivanje iz analize samo onih vrijednosti varijabli koje nedostaju za pojedina opažanja, uz zadržavanje ostalih vrijednosti na datoj varijabli za dato opažanje (Ilić, 2012).

Tabela 4. *Isključivanje u cjelini i u parovima*

Pol	Br. zaposlenih	Prodaja	Pol	Br. zaposlenih	Prodaja
M	25	343	M	25	343
~~Ž~~		~~280~~	Ž	----	280
M	33	332	M	33	332
~~M~~		~~272~~	M	---	272
Ž	25		Ž	25	---
M	29	326	M	29	326
	~~26~~	~~259~~	---	26	259
M	32	297	M	32	297

Izvor: prilagođeno prema Sunil, 2016.

Primjena metode brisanja nedostajućih podataka u cjelini (engl. listwise deletion) opravdana je u slučaju da se utvrdi da nedostajući podaci nedostaju po MCAR mehanizmu (Kang, 2013). Međutim, ukoliko nedostajući podaci ne nedostaju po MCAR mehanizmu, tada primjena metode brisanja u cjelini može uzrokovati pristrasnost u procjeni parametara (Donner, 1982). Kang (2013) navodi da brisanje u paru izaziva manji nivo pristrasnosti ukoliko podaci nedostaju po MCAR ili MAR mehanizmu, međutim, ako nedostaje veliki procenat podataka, analiza će biti manjkava. Graham i saradnici (Graham i dr., 2003) navode da su oba načina eliminacije nedostajućih podataka bazirana na brisanju generalno neprihvatljiva. U prilog tome idu rezulitati koji pokazuju da se isključivanjem gubi na snazi testa. Ovo iz razloga što se smanjuje veličina uzorka čak i pod pretpostavkom da podaci nedostaju po MCAR mehanizmu, s obzirom na to da su t-testovi funkcije veličine uzorka (Ilić, 2012). Iako tretmani temeljeni na brisanju nedostajućih podataka imaju očigledne manjkavosti, u okviru statističkih paketa često su automatski odabrana opcija (Oblaković, Sokolovska, Dinić, 2015).

1.2.2. Tretmani temeljeni na imputaciji

Imputacija je proces zamjene podataka koji nedostaju procijenjenim vrijednostima. Umjesto brisanja nedostajućih podataka, ovaj pristup čuva sve slučajeve zamjenom podataka koji nedostaju vjerovatnom vrijednošću procijenjenom drugim dostupnim informacijama. Nakon što su ovim pristupom zamijenjene sve vrijednosti koje nedostaju, skup podataka analizira se korištenjem standardnih tehnika za potpune podatke (Kang, 2013).

Tretmani nedostajućih podataka temeljeni na imputaciji mogu se podijeliti u dvije grupe:

– tretmani nedostajućih podataka bazirani na jednostrukoj imputaciji i

– tretmani nedostajućih podataka bazirani na višestrukoj imputaciji.

Jedna od metoda jednostruke imputacije jeste metoda zamjene nedostajućih podataka srednjom vrijednošću. Zamjena srednjom vrijednošću jeste postupak pri kojem se vrši zamjena svih nedostajućih podataka date varijable srednjom vrijednošću te varijable (Ilić, 2012). Zamjena se vrši srednjom vrijednošću ukoliko su varijable numeričke, međutim, ukoliko su varijable kategorijalne, tada se zamjena vrši modalnom vrijednošću, dok se u slučaju ordinalnih varijabli nedostajući podaci zamjenjuju medijanom (Oblaković, Sokolovska, Dinić, 2015). Primjenom tretmana zamjene nedostajućih podataka dolazi do suženja varijanse varijabli s nedostajućim podacima, što može rezultirati uticajem na visinu korelacija s ostalim vrijablama. Kako bi se izbjegao problem suženja varijanse, nedostajući podatak može se zamijeniti srednjom vrijednošću koja je izračunata za poduzorak, a ne na cijelom uzorku za datu varijablu (Tabachnick, Fidell, 2001). Poduzorak mogu biti dva ili tri podatka koji prethode nedostajućem podatku i koji slijede nakon nedostajućeg podatka. Na ovaj način varijansa varijable koja sadrži nedostajuće podatke i dalje je sužena, ali je to suženje ograničeno na poduzorak.

U okviru jednostruke imputacije svrstava se i metoda imputacije pomoću regresije. Imputacija pomoću regresije podrazumijeva da se nedostajući podaci zamjenjuju predviđenim vrijednostima koje su utvrđene na osnovu regresione jednačine u kojoj se kao nezavisne varijable koriste varijable na kojima su zabilježeni podaci. Ovaj tretman ima niz prednosti, jer se zadržava veličina uzorka u odnosu na brisanje u cjelini ili parovima i izbjegava se značajna promjena standardne devijacije ili oblika distribucije (Kang, 2013).

Pored navedenih prednosti, ovaj tretman nedostajućih podataka ima i svoje nedostatke. Oni se ogledaju u činjenici da su nedostajući podaci zamijenjeni na osnovu postojećih podatka zabilježenih na drugim varijablama, što može rezultirati visokim stepenom korelacije. Pored toga, treba pretpostaviti da postoji linearni odnos između varijabli korištenih u regresionoj jednačini, kada ga možda i nema (Swalin, 2018).

Tretman jednostruke imputacije obuhvata i metod slučajne imputacije. Metoda slučajne imputacije podrazumijeva zamjenu nedostajuće vrijednosti vrijednošću koju je prijavio drugi ispitanik (donator), koji je ili odabran slučajno iz uzorka ili odabran na osnovu sličnosti „primatelja“ u smislu vrijednosti prijavljenih za druge varijable (Newman, 2003). Dakle, slučajna imputacija podrazumijeva imputaciju nedostajućih podataka izmjerenom vrijednošću nekog drugog slučaja sa sličnim obrascem odgovora na ostalim izmjerenim varijablama ili vrijednošću date varijable nasumično odabranog slučaja iz uzorka (Oblaković, Sokolovska, Dinić, 2015).

1.2.3. Metode zasnovane na modelu

Metod višestruke imputacije, kao i metode maksimalne vjerodostojnosti spadaju u grupu metoda zasnovanih na modelu, čija je osnovna prednost u tome što daju nepristrasne procjene u slučaju MCAR i MAR mehanizama (Oblaković, Sokolovska, Dinić, 2015).

Metoda višestruke imputacije proizvodi višestruke vrijednosti za imputaciju jedne vrijednosti koja nedostaje korišćenjem različitih

is performed on the basis of those observations that do not have any of the missing data. On the other hand, excluding missing data by pairs means excluding from the analysis only those values of variables that are missing for individual observations, while retaining other values on a given variable for a given observation (Ilić, 2012).

Table 4 *Exclusion as a whole and in pairs*

Gender	No. employees	Sales	Gender	No.employees	Sales
M	25	343	M	25	343
Ž	–	280	Ž	---	280
M	33	332	M	33	332
M		272	M	---	272
Ž	25		Ž	25	---
M	29	326	M	29	326
	26	259	---	26	259
M	32	297	M	32	297

Source: adjusted according to Sunil,2016

The application of the listwise deletion method is justified in case it is determined that the missing data are missing according to the MCAR mechanism (Kang, 2013). However, if the missing data are not missing by the MCAR mechanism, then the application of the deletion method as a whole can cause bias in parameter estimation (Donner, 1982). Kang (2013) states that deleting in pairs causes a lower level of bias if data is missing under the MCAR or MAR mechanism; however, if a large percentage of data is missing, the analysis will be deficient. Graham et al (Graham i dr., 2003) state that both ways of eliminating missing data based on deletion are generally unacceptable. This is supported by the results which show that the exclusion loses the effect of the strength test. This is due to the fact that the sample size decreases even on the assumption that data are missing by the MCAR mechanism, since t-tests are a function of sample size (Ilić, 2012). Although treatments based on deleting missing data have obvious shortcomings, within statistical packages, the option is often automatically selected (Oblaković, Sokolovska, & Dinić, 2015).

1.2.2. Treatments based on imputation

Imputation is the process of replacing the missing data with estimated values. Instead of deleting any case that has any missing value, this approach preserves all cases by replacing the missing data with a probable value estimated by other available information. After all missing values have been replaced by this approach, the data set is analyzed using the standard techniques for a complete data (Kang, 2013).

Imputation-based treatments for missing data can be divided into two groups:

- treatments of missing data based on single imputation and
- treatments of missing data based on multiple imputations.

One of the methods of single imputation is the method of replacing missing data with a medium value. Replacement with a medium value is a procedure in which all missing data of a given variable are replaced with the medium value of that variable (Ilić, 2012). Replacement is performed with a medium value if the variables are numerical, however, if the variables are categorical then the replacement is performed with a modal value, while in the case of ordinal variables the missing data are replaced with a median (Oblaković, Sokolovska, & Dinić, 2015). The application of the missing data replacement treatment narrows the variance of variables with missing data, which may result in an impact on the amount of correlations with other variables. To avoid the problem of narrowing the variance, the missing data can be replaced by the medium value calculated for the subsample rather than on the whole sample for a given variable (Tabachnick & Fidell, 2001). A subsample can be two or three data that precede the missing data and that follow the missing data. In this way, the variance of the variable containing the missing data is still narrowed, but this narrowing is limited to the subsample.

The method of imputation by regression is also included in the framework of single imputation. Imputation by regression implies that the missing data are replaced by predicted values determined on the basis of the regression equation in which the variables on which the data are recorded are used as independent variables. This treatment has a number of advantages because it retains the sample size relative to the deletion as a whole or in pairs and avoids a significant change in the standard deviation or form of distribution (Kang, 2013).

In addition to the above advantages, this treatment of missing data also has its drawbacks. They are reflected in the fact that missing data have been replaced on the basis of existing data recorded on other variables, which may result in a high degree of correlation. In addition, it should be assumed that there is a linear relationship between the variables used in the regression equation, even when that relationship may not exist (Swalin, 2018).

The treatment of single imputation includes the method of random imputation. The random imputation method involves replacing the missing value with the value reported by another respondent (donor), who is either randomly selected from the sample or selected based on the similarity of the "recipient" in terms of values reported for other variables (Newman, 2003). Thus, random imputation implies imputation of missing data by the measured value of another case with a similar response pattern on other measured variables or the value of a given variable of a randomly selected case from a sample (Oblaković, Sokolovska, & Dinić, 2015).

1.2.3. Metode zasnovane na modelu

The multiple imputation method as well as the maximum reliability method belong to the group of model-based methods, whose main advantage is that they give impartial estimates in the case of MCAR and MAR mechanisms(Oblaković, Sokolovska, & Dinić, 2015).

The multiple imputation method produces multiple values to impute a single missing value using different simulation models. This method introduces the variability of imputed data to find a number of convincing answers for missing data . In multiple imputations, each missing data is replaced with m values obtained from m iterations, where $m > 1$ and m are usually in the range of 3 to 10 (Khan & Hoque, 2020).

simulacijskih modela. Ova metoda uvodi varijabilnost imputiranih podataka kako bi se pronašao niz uvjerljivih odgovora za nedostajuće podatke. U višestrukoj imputaciji, svaki podatak koji nedostaje zamjenjuje se s m vrijednostima dobijenim iz m iteracija, pri čemu je m > 1 i m se obično nalazi u intervalu od 3 do 10 (Khan, Hoque, 2020).

Metoda višestruke imputacije sprovodi se kroz tri koraka. Prvi korak je sama imputacija. Preferirana metoda imputacije je ona koja odgovara obrascu nedostajućih podataka. Naime, u zavisnosti od toga da li se radi o jednovarijantnom ili monotonom obrascu, nedostajući podaci mogu se imputirati pomoću metode regresije ili metode prediktivnog srednjeg podudaranja ako je varijabla na kojoj podaci nedostaju kontinuirana. Ukoliko podaci nedostaju proizvoljno koristi se MCMC (engl. Markov chain Monte Carlo) metoda (Dong, Peng, 2013).

Primjena logističke regresije u okviru metode višestruke imputacije podrazumijeva da se najprije odredi dihotomna varijabla koja sadrži podatke o tome da li na njoj ima ili nema nedostajućih podataka. U ovom prvom koraku stvara se nekoliko slučajno odabranih poduzoraka koji sadrže kompletirane podatke, odnosno koji ne sadrže nedostajuće vrijednosti niti na jednoj varijabli, kako bi se identifikovala distribucija varijable koja sadrži nedostajuće podatke. Rezultat prvog koraka jesu predviđene vrijednosti nedostajućih podataka. U sljedećem koraku ponovo se kreira nekoliko slučajnih poduzoraka, s tom razlikom što sada poduzorak uključuje i varijable sa nedostajućim podacima.

Zatim se sve nedostajuće vrijednosti zamjenjuju procjenama na osnovu rezutata regresije iz prvog koraka. Na osnovu dobijenih setova podataka računaju se pojedinačne procjene, a do konačnih procjena modela dolazi se uprosječavanjem parametara procjene iz kreiranih multiplih setova (Tabachnick, Fidell, 2001).

MCMC metod je za numeričke varijable, dok se za kategorijalne varijable primjenjuju druge metode, kao što je, na primjer, MIC (engl. multiple imputation for categorical data) ili FCS metod (engl. fully conditional specification).

Prednosti ovog tretmana su brojne. Jedna od njih ogleda se u činjenici da se može primijeniti na longitudinalnim podacima. Pored toga, dobijaju se manje greške parametara u odnosu na, na primjer, brisanje, bilo u paru ili u cjelini (Oblaković, Sokolovska, Dinić, 2015).

Kao što je već navedeno, metode maksimalne vjerodostojnosti spadaju u metode zasnovane na modelima i podrazumijevaju iterativni postupak. To znači da se u cilju eliminacije nedostajućih podataka koriste svi raspoloživi podaci u uzorku, i kompletirani i nedostajući, kako bi se utvrdile vrijednosti koje imaju najveću vjerovatnoću pojavljivanja u posmatranim podacima (Alison, 2002). Jedna od metoda maksimalne vjerodostojnosti je EM algoritam.

EM algoritam (engl. expectation maximization algorithm) predstavlja dvostepeni iterativni postupak, koji se sastoji od dva koraka:

– E koraka i

– M koraka.

Prvi korak, korak E, dobio je ime od prvog slova engleske riječi expectation, što u prevodu znači očekivanje, dok je drugi korak, korak M, dobio ime od prvog slova engleske riječi maximization, što u prevodu znači maksimizacija (Vasić, 2018).

Prvim korakom, E korakom, procjenjuju se parametri distribucije na osnovu raspoloživih podataka, dok se u drugom koraku, M koraku, računaju parametri za nedostajuće podatke maksimiziranjem vjerovatnoće dobijanja očekivanih vrijednosti iz E koraka. Ova dva koraka ponavljaju se u nizu iteracija sve dok se ne postigne konvergencija, odnosno dok se procjene parametara ne počnu razlikovati iz iteracije u iteraciju (Oblaković, Sokolovska, Dinić, 2015).

Kod primjene EM algoritma, u koraku E se za svaki jedinstveni obrazac nedostajućih podataka kreira model višestruke linearne regresije za svaku promjenljivu koja ima nedostajuće podatke. U datom modelu, zavisna ili objašnjena promjenljiva predstavlja promjenljivu sa nedostajućim podacima, a nezavisne ili objašnjavajuće promjenljive su one koje su u datom jedinstvenom obrascu kompletno raspoložive. Na taj način se nedostajući podaci ocjenjuju pomoću linearne povezanosti sa kompletno raspoloživim promjenljivim. Kod datih modela višestruke linearne regresije, nepoznati parametri se ocjenjuju pomoću elemenata vektora sredina promjenljivih, kao i kovarijacione matrice promjenljivih (Vasić, 2018).

Tabachnick i Fidell (2001) ističu da je prednost EM algoritma, kao metode maksimalne vjerodostojnosti, u tome što je jednostavan i što postoji manja opasnost od previsoke saglasnosti između podataka i modela, odnosno situacije u kojoj testirani model izgleda bolje nego što zaista jeste. Dong i Peng (2013) ističu kao nedostake EM algoritma to što je primjena ograničena na linearne modele i podatke koji su normlano distribuirani. Drugi nedostatak koji navedeni autori navode vezan je za softvere za analizu nedostajućih podataka. Naime, softveri uglavnom ne obezbjeđuju procjenu standardnih grešaka parametara, zbog čega se EM ne preporučuje kada je primarni cilj statističko testiranje intervala povjerenja procijenjenih parametara.

FIML metod (engl. full information maximum likelihood) jeste metod koji se obično predstavlja kao kovarijansna matrica promjenljive i vektora očekivanja. Prednost u odnosu na EM algoritam je u tome što omogućava direktno izračunavanje odgovarajućih standardnih grešaka i testiranje statistike. Postupak zahtijeva da podaci budu bar tipa MAR ili MCAR (Ilić, 2012).

2. METODOLOGIJA I REZULTATI

U ovom poglavlju biće dat primjer kako primijeniti metodu višestruke imputacije u SPSS (Statistical Package for Social Sciences) programu, kao jednom od, danas, najčešće korišćenih programa za obradu podataka.

U primjeru se analiziraju 23 finansijska pokazatelja za 226 preduzeća iz oblasti trgovine na veliko i malo, koja su registrovana na teritoriji Republike Srpske, a koja su, prema podacima APIF-a, u periodu od 2017. do 2020. godine imala blokiran račun. Svrha analize je kreiranje modela za predviđanje platežne nesposobnosti preduzeća iz navedene oblasti.

S obzirom na to da se radi o finansijskim pokazateljima, kao varijablama koje se analiziraju, isti su identifikovani kao varijable numeričkog tipa (Scale). Pravilno označavanje tipa varijabli, prije same analize podataka, vrlo je značajno kako se ne bi desilo da se dobiju neočekivane vrijednosti, na primjer, pol u decimalnom zapisu i slično.

Prvi korak u provođenju višestruke imputacije u SPSS programu jeste analiza mehanizma po kojem podaci nedostaju. Identifikovanje mehanizma nedostajućih podatka vrši se tako što se sa padajućeg menija izabere opcija Analyze, a zatim se iz navedene opcije bira Multiple Imputation / Analyze Patterns..., nakon čega se otvara prozor za dijalog u okviru kojeg je potrebno u polje Analyze Across Variables iz polja Variables prebaciti varijable na kojima želimo izvršiti eliminaciju nedostajućih podataka. U okviru prozora Output potrebno je čekirati sve tri ponuđene opcije.

The multiple imputation method is implemented in three steps. The first step is imputation itself. The preferred method of imputation is that which corresponds to the pattern of missing data. Namely, depending on whether it is a single-variant or monotonic pattern, missing data can be imputed using the regression method or the predictive medium matching method if the variable on which the data are missing is continuous. If data are missing, the MCMC (Markov chain Monte Carlo) method is used arbitrarily (Dong & Peng, 2013).

The application of logistic regression within the multiple imputation method implies that the dichotomous variable is first determined, which contains data on whether or not there is missing data on it. In this first step, several randomly selected subsamples are created that contain complete data, ie that do not contain missing values on any of the variables, in order to identify the distribution of the variable that contains the missing data. The result of the first step is the predicted values of the missing data. In the next step, several random subsamples are re-created, with the difference that the subsample now includes variables with missing data.

Then, all missing values are replaced by estimates based on the regression results from the first step. Based on the obtained data sets, individual estimates are calculated, and the final estimates of the model are obtained by averaging the estimation parameters from the created multiple sets (Tabachnick & Fidell, 2001).

The MCMC method is for numerical variables, while other methods are used for categorical variables, such as MIC (multiple imputation for categorical data) or FCS method (fully conditional specification).

The benefits of this treatment are numerous. One of them is reflected in the fact that it can be applied to longitudinal data. In addition, there are fewer parameter errors compared to, for example, deletion either in pairs or as a whole (Oblaković, Sokolovska, & Dinić, 2015).

As already mentioned, maximum plausibility methods belong to model-based methods and imply an iterative procedure. This means that in order to eliminate the missing data, all available data in the sample are used, both completed and missing, in order to determine the values that have the highest probability of appearing in the observed data (Alison, 2002). One of the methods of maximum credibility is the EM algorithm.

EM expectation maximization algorithm is a two-step iterative procedure, which consists of two steps:

– E step and

– M step.

The first step, step E, got its name from the first letter of the English word expectation, which means expectation, while the second step, step M, got its name from the first letter of the English word maximization, which means maximization (Vasić, 2018).

The first step, E step, estimates the distribution parameters based on the available data, while the second step, M step, calculates the missing data parameters by maximizing the probability of obtaining the expected values from E step. These two steps are repeated in a series of iterations until convergence is achieved, i.e. until the parameter estimates begin to differ from iteration to iteration (Oblaković, Sokolovska, & Dinić, 2015).

When applying the EM algorithm, in step E, a multiple linear regression model is created for each unique missing data pattern for each variable that has missing data.

In a given model, a dependent or explained variable is a variable with missing data, and independent or explanatory variables are those that are completely available in a given unique pattern. In this way, missing data is evaluated using a linear relationship with completely available variables. In the given models of multiple linear regression, unknown parameters are estimated using the elements of the vector of the medium variables, as well as the covariance matrix of the variables (Vasić, 2018).

Tabachnick i Fidell (2001) point out that the advantage of the EM algorithm, as a method of maximum reliability, is that it is simple and there is less danger of too much agreement between data and models, i.e. a situation where the tested model looks better than it really is. Dong i Peng (2013) point out as shortcomings of the EM algorithm that the application is limited to linear models and data that are normally distributed. Another shortcoming cited by these authors is related to software for analyzing missing data.

Namely, software generally does not provide an estimate of standard parameter errors, which is why EM is not recommended when the primary goal is statistical testing of the confidence interval of the estimated parameters. The FIML (full information maximum likelihood) method is a method that is usually presented as a covariance matrix of a variable and an expectation vector. The advantage over the EM algorithm is that it allows direct calculation of appropriate standard errors and testing of statistics. The procedure requires that the data be at least of the MAR or MCAR type (Ilić, 2012).

2. METHODOLOGY AND RESULTS

This chapter will give an example of how to apply the method of multiple imputation in the SPSS (Statistical Package for Social Sciences) program, as one of the most commonly used data processing programs today.

The example analyzes 23 financial indicators for 226 companies in the field of wholesale and retail trade, which are registered in the territory of Republika Srpska, and which, according to APIF data, had a blocked account in the period from 2017-2020. The purpose of the analysis is to create a model for predicting the insolvency of companies in this area. Since these are financial indicators, as variables that are analyzed, they are identified as variables of numerical type (Scale). Proper labeling of the type of variables, before the analysis of the data itself, is very important so that it does not happen that unexpected values are obtained, for example, gender in decimal notation and the like.

The first step in conducting multiple imputations in the SPSS program is to analyze the mechanism by which data is lacking. Identifying the mechanism of missing data is done by selecting the Analyze option from the drop-down menu, then selecting Multiple Imputation / Analyze Patterns ... from the specified option, after which a dialog window opens in which you need to transfer variables from the Variables field (on which we want to eliminate the missing data) in the Analyze Across Variables field. Within the Output window, you need to check all three offered options.

The minimum percentage of missing data for the variables we want to display is determined in the field Minimum precentage missing for variable to be displayed, within which 10 is automatically offered.

In this example, this percentage was reduced to 0.1 to include a larger number of variables with missing data.

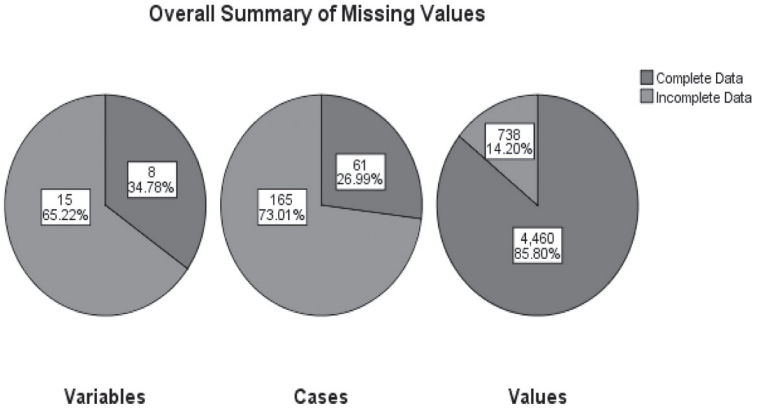
Pressing OK gives the results of the analysis, which consists of 4 parts. The first part of the analysis, Overall Summary of Missing Values - refers to the frequency and percentage of variables, cases and cells in the matrix that contain missing data. Based on Figure

Minimalni procenat nedostajućih podataka za varijable koje želimo prikazati određuje se u polju Minimum precentage missing for variable to be displayed, u okviru kojeg je automatski ponuđeno 10.

U ovom primjeru je ovaj procenat smanjen na 0,1 kako bi se obuhvatio veći broj varijabli sa nedostajućim podacima. Pritiskom na OK dobijaju se rezultati analize koji se sastoje od četiri dijela.

Prvi dio analize, Overall Summary of Missing Values, odnosi se na frekvencu i procenat varijabli, slučajeva i ćelija u matrici koji sadrže nedostajuće podatke. Na osnovu slike 2. vidi se da 15 od 23 varijable, odnosno 65,22% unesenih varijabli sadrži nedostajuće podatke. Nedostajući podaci javljaju se kod 165 od 226 posmatranih preduzeća u uzorku, što čini 73,01%. Treći grafički prikaz sa slike broj 3. govori da 738 ćelija u matrici sadrži nedostajuće podatke, a da 4.460 ćelija ima kompletirane podatke.

Slika 2. *Nedostajući podaci – varijable, slučajevi, ćelije*



Izvor: izrada autora u SPSS programu

U tabeli 5. dat je drugi dio analize, koji se odnosi na pregled frekvenci i procenata nedostajućih podataka po varijablama, po opadajućem redoslijedu. Na osnovu tabele može se vidjeti da najveći procenat nedostajućih podataka imaju varijable ROE, RZ i KP, odnosno prinos na kapital, racio zaduženosti (46%) i kvalitet prihoda (39,8%). Kako su prediktorske varijable korišćene u ovom radu numeričke, u tabeli je dat i pregled aritmetičke sredine i standardne devijacije.

Tabela 5. *Pregled nedostajućih podataka po varijablama*

Variable Summary[a,b]

	Missing		Valid N	Mean	Std. Deviation
	N	Percent			
ROE	104	46.0%	122	-.04374	4.387070
RZ	104	46.0%	122	15.83432	57.961009
KP	90	39.8%	136	1.94204	3.636423
RPK	72	31.9%	154	-88.75668	898.142236
KOZ	53	23.5%	173	9.28087	29.162227
BPM	52	23.0%	174	-10.05555	67.014697
NPM	52	23.0%	174	-10.09716	67.045458
NP_II	45	19.9%	181	4.35534	11.477801
NP_I	45	19.9%	181	2.74051	10.006440
NOKUS	45	19.9%	181	-7.98816	79.315459
ROP	29	12.8%	197	17.43561	70.495075
logGGE	19	8.4%	207	7.43478	3.004106
KFS	16	7.1%	210	-7324.09819	54922.277027
ROOD	9	4.0%	217	2.75444	5.693066
KOOI	3	1.3%	223	1.13151	1.995308

a. Maximum number of variables shown: 25

b. Minimum percentage of missing values for variable to be included: 0.1%

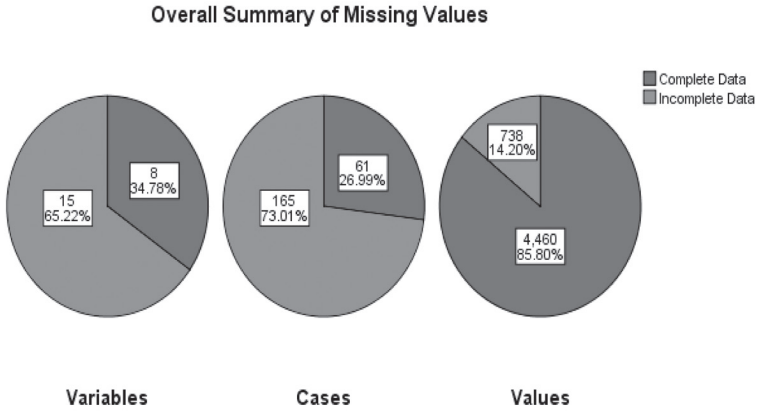
Izvor: izrada autora u SPSS programu

Treći dio analize odnosi se na obrazac nedostajućih podataka, koji je predstavljen slikom 3. Na osnovu grafikona, koji je dat slikom 3, može se zaključiti da preovladava monotoni obrazac nedostajućih podataka i da podaci ne nedostaju po slučajnom rasporedu. Drugim riječima, nije priustan MCAR (engl. Missing completely at random) mehanizam nedostajućih podataka.

Prikaz koji je dat slikom 3. služi vizuelnoj inspekciji mehanizma po kojem podaci nedostaju. Prvi red predstavlja, u našem primjeru, grupu slučajeva na kojima nema nedostajućih podataka. Drugi obrazac, koji se nalazi u drugom redu, obuhvata slučajeve, odnosno preduzeća kojima nedostaje samo racio obrata obaveza prema dobavljačima i tako svaki sljedeći red predstavlja novi obrazac u

2, it can be seen that 15 of the 23 variables, i.e. 65.22% of the entered variables contain missing data. Missing data appear in 165 of the 226 observed companies in the sample, which is 73.01%. The third graph from Figure 3 shows that 738 cells in the matrix contain missing data, and that 4460 cells have completed data.

Figure 2 *Missing data - variables, cases, cells*



Source: author's work in the SPSS program

Table 5 gives the second part of the analysis, which refers to the review of frequencies and estimates of missing data by variables, in descending order. Based on the table, it can be seen that the largest percentage of missing data have the variables ROE, RZ and KP, i.e. return on capital, debt ratio (46%) and income quality (39.8%). As the predictor variables used in this paper are numerical, the table also provides an overview of the arithmetic medium and standard deviation

Table 5 *Overview of missing data by variables*

Variable Summary[a,b]

	Missing		Valid N	Mean	Std. Deviation
	N	Percent			
ROE	104	46.0%	122	-.04374	4.387070
RZ	104	46.0%	122	15.83432	57.961009
KP	90	39.8%	136	1.94204	3.636423
RPK	72	31.9%	154	-88.75668	898.142236
KOZ	53	23.5%	173	9.28087	29.162227
BPM	52	23.0%	174	-10.05555	67.014697
NPM	52	23.0%	174	-10.09716	67.045458
NP_II	45	19.9%	181	4.35534	11.477801
NP_I	45	19.9%	181	2.74051	10.006440
NOKUS	45	19.9%	181	-7.98816	79.315459
ROP	29	12.8%	197	17.43561	70.495075
logGGE	19	8.4%	207	7.43478	3.004106
KFS	16	7.1%	210	-7324.09819	54922.277027
ROOD	9	4.0%	217	2.75444	5.693066
KOOI	3	1.3%	223	1.13151	1.995308

a. Maximum number of variables shown: 25

b. Minimum percentage of missing values for variable to be included: 0.1%

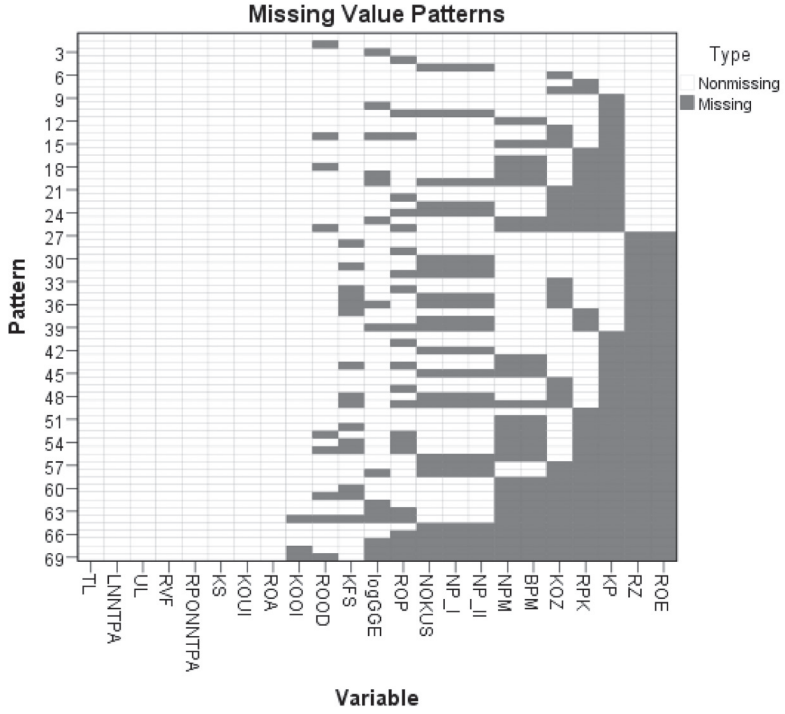
Source: author's work in the SPSS program

The third part of the analysis refers to the missing data pattern, which is presented in Figure 3.Based on the graph given in Figure 3, it can be concluded that a monotonous pattern of missing data prevails and that data is not missing at random. In other words, the Missing Completely At Random (MCAR) mechanism is missing.

The representation given in Figure 3 serves to visually inspect the mechanism by which the data is missing. The first row represents, in our example, a group of cases where there is no missing data. The second pattern, which is in the second row, includes cases, i.e. companies that lack only the ratio of turnover of liabilities to suppliers, and so each subsequent line represents a new pattern in which an increasing number of cases and variables are missing. The last, 69th line, i.e. the pattern is the pattern that includes companies that, as can be seen, lack the largest amount of data or financial ratio indicators. So, the rows of this graph have numbered patterns of missing data, and the columns have variables. The variables are sorted by the percentage of missing data, so the variable TL (current liquidity) has no missing data, while the largest percentage of missing data has the variable ROE (return on equity).

kojem nedostaje sve veći broj slučajeva i varijabli. Posljednji, 69. red, odsnosno obrazac, jeste obrazac koji obuhvata preduzeća kojima, kao što se vidi, nedostaje najveći broj podataka odnosno finansijskih racio pokazatelja. Dakle, u redovima ovog grafikona su numerisani obrasci nedostajućih podataka, a u kolonama varijable. Varijable su poredane prema procentu nedostajućih podataka, pa tako varijabla TL (tekuća likvidnost) nema nedostajućih podataka, dok najveći procenat nedostajućih podataka ima varijabla ROE (prinos na kapital).

Slika 3. *Obrazac nedostajućih podataka*

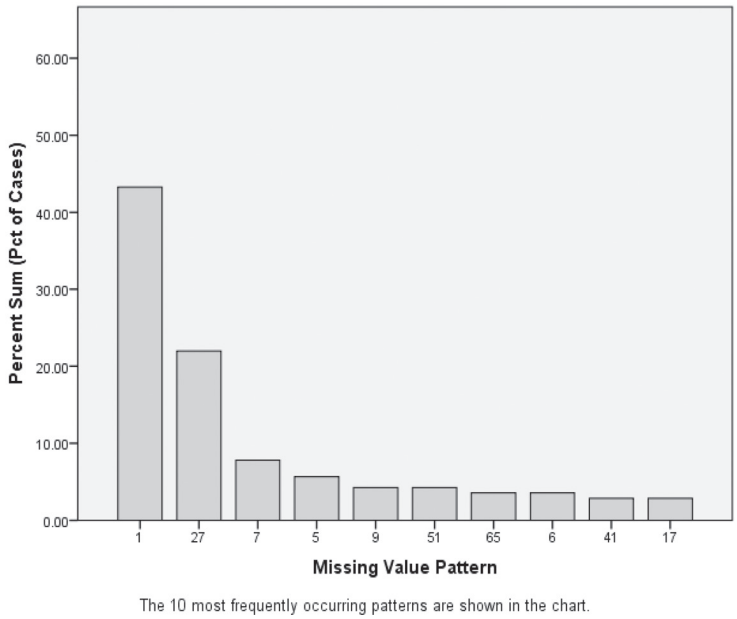


Izvor: izrada autora u SPSS programu

Kada je riječ o mehanizmu nedostajućih podataka, ukoliko se uoči neka pravilnost ili koncentracija u osjenčenim poljima, tada se može konstatovati da postoji određeni obrazac po kojem podaci nedostaju. U takvoj situaciji podaci ne nedostaju slučajno, tj. nije riječ o MCAR mehanizmu. Ukoliko se na grafikonu uoči porast ili smanjenje osjenčenih površina, to ukazuje da nedostajući podaci nedostaju po monotnom obrascu. Monotonicitet je samo jedan od NMAR mehanizama nedostajućih podataka (Oblaković, Sokolovska, Dinić, 2015). U našem primjeru se na osnovu grafikona, koji je dat slikom broj 3, može vidjeti da postoji koncentracija osjenčenih polja u donjem desnom uglu, što nas navodi na zaključak da je u podacima najvjerovatnije prisutan monotonicitet.

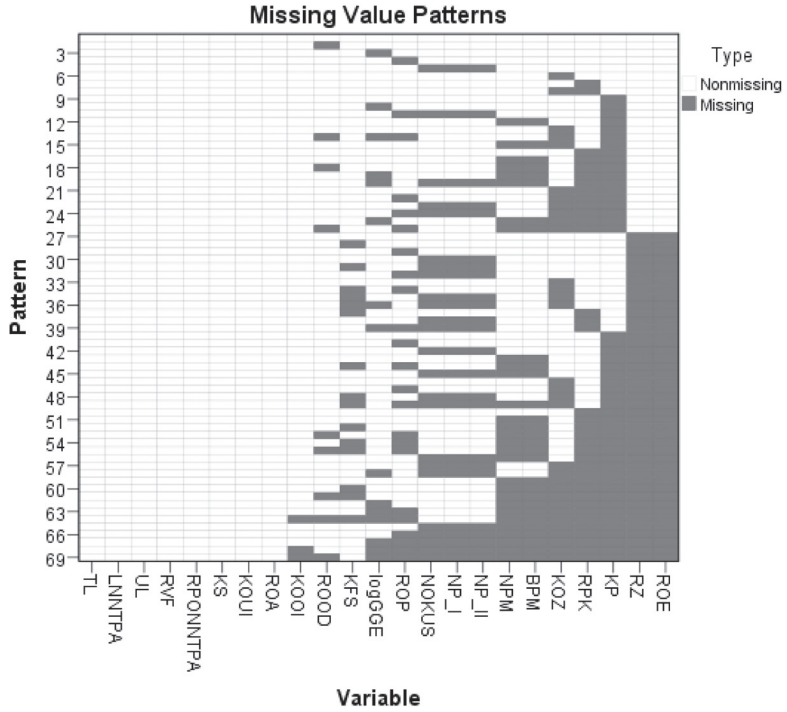
Na osnovu četvrtog dijela analize može se vidjeti koji se obrazac nedostajućih podataka najčešće ponavlja.

Slika 4. *Procenat slučajeva s određenim obrascem nedostajanja podataka*



Izvor: izrada autora u SPSS programu

Figure 3 *Missing data pattern*



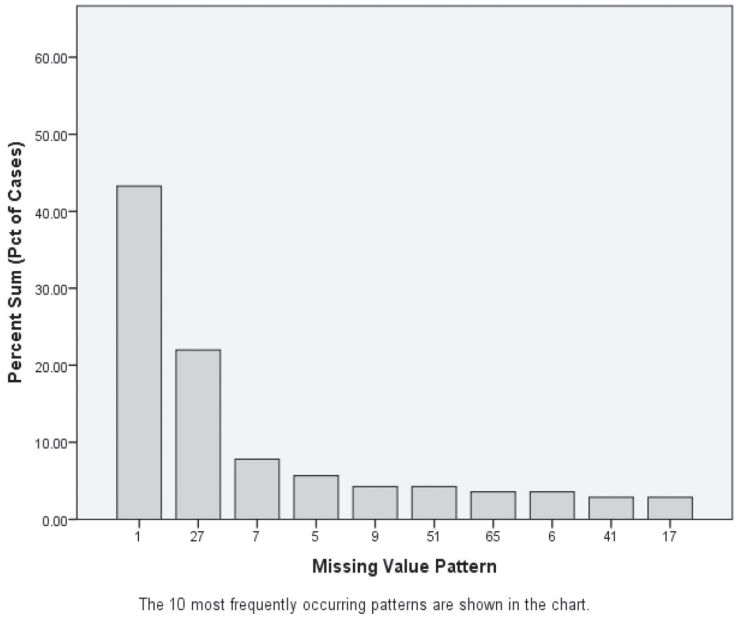
Source: author's work in the SPSS program

When it comes to the mechanism of missing data, if some regularity or concentration is observed in the shaded fields, then it can be stated that there is a certain pattern according to which the data is missing. In such a situation, the data are not missing by accident, i.e. it is not an MCAR mechanism. If the chart shows an increase or decrease in shaded areas, it indicates that the missing data are missing according to the monotonic pattern. Monotonicity is only one of the NMAR mechanisms of missing data (Oblaković, Sokolovska, & Dinić, 2015).

In our example, based on the graph given in Figure 3, it can be seen that there is a concentration of shaded fields in the lower right corner, which leads us to the conclusion that monotonicity is most likely present in the data.

Based on the fourth part of the analysis, it can be seen which pattern of missing data is most often repeated.

Figure 4 *Percentage of cases with a certain pattern of lack of data*



Source: author's work in the SPSS program

The graph, given in Figure 4, shows that the most common is the first pattern, the pattern according to which data is missing. However, if we go back to the chart from Figure 3, we see that there is actually no missing data according to this pattern. The next pattern according to which data are most often missing is form 27.

After analyzing the missing data, we can approach their replacement by applying multiple imputations. The application of multiple imputation in the SPSS program means that the Multiple Imputation / Impute Missing Data Values option is selected from the Analyze drop-down menu ... After that, a dialog window opens in which it

Grafikon koji je dat slikom 4. prikazuje da je načešći prvi obrazac, obrazac po kojem nedostaju podaci. Međutim, ukoliko se vratimo na grafikon sa slike 3, vidimo da zapravo po tom obrascu nema nedostajućih podataka. Sljedeći obrazac po kojem podaci najčešće nedostaju jeste obrazac 27.

Nakon analize nedostajućih podataka, možemo pristupiti njihovoj zamjeni primjenom višestruke imputacije. Primjena višestruke imputacije u SPSS programu podrazumijeva da se iz padajućeg menija Analyze odabere opcija Multiple Imputation / Impute Missing Data Values... Nakon toga otvara se prozor za dijalog u okviru kojeg je potrebno da se iz polja Variables prebace varijable nad kojima želimo da izvršimo višestruku imputaciju u polje Variables in Model. Broj imputacija može se mijenjati u polju Imputations, ali smo u primjeru zadržali pet imputacija, koliko se inače u literaturi preporučuje. Matrica sa zamijenjenim nedostajućim podacima može se sačuvati kao nova matrica ili u okviru postojeće matrice. U ovom primjeru matrica sa zamijenjenim nedostajućim podacima sačuvana je kao nova sa imenom impute.matrica.

Na kartici Method može se izvršiti odabir metode imputacije, o kojima je više riječi bilo na prethodnim stranicama. SPSS nudi dvije metode: prvu – MCMC, koja se primjenjuje kada podaci nedostaju po slučajnom mehanizmu, i drugu, koja se primjenjuje kada postoji monotonicitet. U ovom primjeru zadržana je automatska procjena metode, jer se u tom slučaju podaci skeniraju u odnosu na monotonicitet i ukoliko se on detektuje primjenjuje se drugi metod. Na kartici Constraints možemo podesiti da u okviru rezultata analize dobijemo i procenat nedostajućih podataka po varijabli i deskriptivne pokazatelje varijabli klikom na dugme Scan Data. Output kartica je posljednja u ovom prozoru i na njoj možemo čekirati sve tri ponuđene kućice. Ukoliko čekiramo opciju za kreiranje istorije iteracija u novoj matrici, toj novoj matrici moramo dati ime, na primjer iteracija.istorija.

U okviru ispisa rezultata, u tabeli Imputation Models može se vidjeti koji tip modela je primijenjen za koje varijable, a u skladu sa definisanim nivoom mjerenja, da li je primijenjena logistička ili linearna regresija. Pored navedenog, u tabli Imputation Models može se vidjeti i koji su se prediktori koristili za predikciju nedostajućih podataka u okviru svake varijable. Dio table Imputation Models iz našeg primjera dat je slikom broj 5.

Slika 5. *Imputation Models*

Imputation Models

	Model		Missing Values	Imputed Values
	Type	Effects		
KOOI	Linear Regression	TL,LNNTPA, UL,RVF, RPONNTPA, KS,KOUI, ROA,ROOD, KFS,logGGE, ROP,NOKUS, NP_I,NP_II, NPM,BPM, KOZ,RPK,KP, RZ,ROE	3	15
ROOD	Linear Regression	TL,LNNTPA, UL,RVF, RPONNTPA, KS,KOUI, ROA,KOOI, KFS,logGGE,	9	45

Izvor: izrada autora u SPSS-u

Nova matrica sa zamijenjenim vrijednostima, u gornjem desnom uglu, sadrži polje Orginal data sa padajućom listom koja daje pregled zamijenjenih podataka u svakoj imputaciji od 5 sprovedenih. Ćelije sa zamijenjenim podacima su osjenčene. Dalja analiza i obrada podataka radi se nad zamijenjenim podacima dobijenim u petoj imputaciji.

U novijim istraživanjima preferiraju se metode koje se svrstavaju u grupu savremenih metoda, kao što su metode višestruke imputacije i maksimalne vjerodostojnosti. Ovo iz razloga što savremene metode daju manje pristrasne procjene u odnosu na tradicionalne, čak i kada podaci ne nedostaju po NMAR mehanizmu.

ZAKLJUČAK

Na osnovu metoda koje su predstavljene u ovom radu, a koje se primjenjuju za eliminaciju nedostajućih podataka, može se zaključiti da je u savremenim uslovima istraživanja primjena metoda baziranih na brisanju nedostajućih podataka neadekvatna i da se kao takva ne preporučuje. Metode jednostruke imputacije, koje se takođe svrstavaju u grupu tradicionalnih metoda, zbog svojih nedostataka, u novijim istraživanjima o nedostajućim podacima rijetko se preporučuju. Međutim, treba istaći da nedostaci ovih metoda mogu da budu zanemarljivi u situacijama kada su ispunjeni određeni uslovi, kao što je, na primjer, situacija kada se utvrdi da nedostajući podaci nedostaju po MCAR mehanizmu i kada je broj nedostajućih podataka mali.

IZVORI

1. Alison, P. (2002). *Missing data*. Sage University papers series on quantitative applications in the social sciences. Preuzeto sa: https://assets.thalia.media/doc/f3/a7/f3a7a631-469a-4907-8740-6b16b2763f1d.pdf
2. Bennett, A. (2001). How can I deal with missing data in my study? *Australian and New Zeland jouranl of public health*, 463–469. doi:https://doi.org/10.1111/j.1467-842X.2001.tb00294.x
3. Dong, Y., Peng, C. (2013). Principled missing data methods for researchers. *SpringerOpen Journal*, 2–17. Preuzeto sa: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/

is necessary to transfer the variables over which we want to perform multiple imputations from the Variables field to the Variables in Model field. The number of imputations can be changed in the Imputations field, but we have kept 5 imputations, as recommended in the literature. A matrix with replaced missing data can be saved as a new matrix or within an existing matrix. In this example, the matrix with the missing data replaced is saved as new with the name impute.matica

On the Method tab, you can select the imputation method, which was discussed more on the previous pages. SPSS offers two methods:1. MCMC applicable when data are missing by random mechanism and 2. which is applied when there is monotony.

In this example, the automatic estimation of the method is retained, because in that case the data are scanned in relation to the monotonicity, and if it is detected, another method is applied. On the Constraints tab, we can set the percentage of missing data by variables and descriptive indicators of variables by clicking on the Scan Data button. The output card is the last in this window and we can check all three offered boxes on it. If we check the option to create an iteration history in a new matrix, we must give that new matrix a name, for example iteration.history.

Within the printout of the results, in the table Imputation Models you can see which type of model was applied for which variables, and in accordance with the defined level of measurement - whether logistic or linear regression was applied. In addition to the above, in the Imputation Models panel you can see which predictors were used to predict the missing data within each variable. Part of the Imputation Models panel from our example is given in Figure 5.

Figure 5 *Imputation Models*

Imputation Models

	Model		Missing Values	Imputed Values
	Type	Effects		
KOOI	Linear Regression	TL,LNNTPA, UL,RVF, RPONNTPA, KS,KOUI, ROA,ROOD, KFS,logGGE, ROP,NOKUS, NP_I,NP_II, NPM,BPM, KOZ,RPK,KP, RZ,ROE	3	15
ROOD	Linear Regression	TL,LNNTPA, UL,RVF, RPONNTPA, KS,KOUI, ROA,KOOI, KFS,logGGE,	9	45

Source: author's work in the SPSS program

The new matrix with the replaced values, in the upper right corner, contains the Original data field with a drop-down list that gives an overview of the replaced data in each imputation of the 5 implemented. Cells with replaced data are shaded. Further analysis and data processing is performed on the replaced data obtained in the 5th imputation.

CONCLUSION

Based on the methods presented in this paper, which are used to eliminate missing data, it can be concluded that in modern research conditions the application of methods based on deleting missing data is inadequate and is not recommended as such.

Single imputation methods, which are also classified as traditional methods, are rarely recommended in recent research on missing data due to their shortcomings. However, it should be noted that the shortcomings of these methods can be negligible in situations where certain conditions are met, such as when the missing data are found to be missing under the MCAR mechanism and when the number of missing data is small.

Recent research has preferred methods that fall into the group of modern methods, such as methods of multiple imputation and maximum reliability. This is because modern methods give less biased estimates compared to traditional ones, even when data are not missing under the NMAR mechanism.

LITERATURE

1. Alison, P. (2002). *Missing data*. Sage University papers series on quantitative applications in the social sciences. Retrieved https://assets.thalia.media/doc/f3/a7/f3a7a631-469a-4907-8740-6b16b2763f1d.pdf
2. Bennett, A. (2001). How can I deal with missing data in my study? *Australian and New zeland jouranl of public health*, 463-469. doi:https://doi.org/10.1111/j.1467-842X.2001.tb00294.x
3. Dong, Y., & Peng, C. (2013). Principled missing data methods for researchers. *SpringerOpen Journal,* 2-17. Retrieved https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/
4. Donner, A. (1982). The Relative Effectiveness of Procedures Commonly Used in Multiple Regression Analysis for Dealing with Missing Values. *The American Statistician*, 378-381. Retrieved https://www.tandfonline.com/doi/abs/10.1080/00031305.1982.10483055
5. Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: The Guilford Press. Retrieved https://books.google.ba/books?hl=hr&lr=&id=MN8ruJd2tvgC&oi=fnd&pg=PA1&dq=missing+data+&ots=dKiBrV_js3&sig=J9Xjqsmca1MAlyCENUkg9t0F6JM&redir_esc=y#v=onepage&q=missing%20data&f=false
6. Graham, J., Cumsile, P., & Elek-Fisk, E. (2003). Methods for hadnling missing data. *Handbook of Psychology*, 87-144. doi:10.1002/0471264385.wei0204

4. Donner, A. (1982). The Relative Effectiveness of Procedures Commonly Used in Multiple Regression Analysis for Dealing with Missing Values. *The American Statistician,* 378–381. Preuzeto sa: https://www.tandfonline.com/doi/abs/10.1080/00031305.1982.10483055
5. Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: The Guilford Press. Preuzeto sa: https://books.google.ba/books?hl=hr&lr=&id=MN8ruJd2tvgC&oi=fnd&pg=PA1&dq=missing+data+&ots=dKiBrV_js3&sig=J9Xjqsmca1MAlyCENUkg9t0F6JM&redir_esc=y#v=onepage&q=missing%20data&f=false
6. Graham, J., Cumsile, P., Elek-Fisk, E. (2003). Methods for hadnling missing data. *Handbook of Psychology*, 87–144. doi:10.1002/0471264385.wei0204
7. Howard, J. (2013). *Using principal component analysis (PCA) to obtain auxiliary variables for missing data estimation in large data sets*. Lawrence: ProQuest LLC. Preuzeto sa: https://www.researchgate.net/publication/263547743_Using_principal_component_analysis_PCA_to_obtain_auxiliary_variables_for_missing_data_estimation_in_large_data_sets
8. Ilić, D. (2012). *Ocenjivanje indeksa repa raspodele korišćenjem nekompletnh uzoraka*. Beograd: Matematički fakultet, Univerzitet u Beogradu. Preuzeto sa: https://nardus.mpn.gov.rs/handle/123456789/2849?show=full
9. Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 402–406. doi: 10.4097/kjae.2013.64.5.402
10. Khan, S., Hoque, L. (2020). SICE: an improved missing data imputation technique. *Journal of Big Data*, 7–37. Preuzeto sa: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00313-w
11. Lang, M., Little, D. (2018). Principled Missing Data Treatments. *Prevention Science*, 284–294. doi:https://doi.org/10.1007/s11121-016-0644-5
12. Newman, D. (2003). Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. *Organizational Research Methods*, 328–362. Preuzeto sa https://journals.sagepub.com/doi/10.1177/1094428103254673
13. Oblaković, M., Sokolovska, V., Dinić, B. (2015). Tretmani nedostajućih podataka. *Primenjena psihologija*, 289–309. doi:10.19090/str.2015.3.289-309
14. Rubin, D., Little, R. (2002). *Statistical Analysis with Missing Data*, Second Edition. New Jersey: John Wiley & Sons, Inc. Preuzeto sa: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119013563
15. Schafer, J., Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*. doi:10.1037/1082-989X.7.2.147
16. Schafer, L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 3–15. doi:https://doi.org/10.1177/096228029900800102
17. Sunil, R. (2016). *A Comprehensive Guide to Data Exploration*.
18. Swalin, A. (2018). *How to Handle Missing Data*. Preuzeto sa: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
19. Tabachnick, B., Fidell, L. (2001). *Using multivariate statistic* (4th edition). Allyn and Bacon, Boston.
20. Tabachnick, G., Fidell, L. (2013). *Using Multivariate Statistics*. Pearson Education.
21. Vasić, V. (2018). Rešavanja problema multivarijacionih nedostajućih anketnih podataka primenom EM algoritma. *Ekonomske ideje i praksa*, 35–50. Preuzeto sa: http://www.ekof.bg.ac.rs/wp-content/uploads/2014/10/003.pdf
22. Warnes, Z. (2021, July 6). *Missing Value Handling – Missing Data Types*. Preuzeto sa: Towards Data Science: https://towardsdatascience.com/missing-value-handling-missing-data-types-a89c0d81a5bb
23. Xian, L. (2016). *Methods and Applications of Longitudinal Data Analysis*. Academic Press. doi:https://doi.org/10.1016/C2013-0-13082-6

7. Howard, J. (2013). *Using principal component analysis (PCA) to obtain auxiliary variables for missing data estimation in large data sets*. Lawrence: ProQuest LLC. Retrieved https://www.researchgate.net/publication/263547743_Using_principal_component_analysis_PCA_to_obtain_auxiliary_variables_for_missing_data_estimation_in_large_data_sets
8. Ilić, D. (2012). *Ocenjivanje indeksa repa raspodele korišćenjem nekompletnh uzoraka*. Beograd: Matematički fakultete, Univerzitet u Beogradu. Preuzeto sa https://nardus.mpn.gov.rs/handle/123456789/2849?show=full
9. Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 402-406. doi: 10.4097/kjae.2013.64.5.402
10. Khan, S., & Hoque, L. (2020). SICE: an improved missing data imputation technique. *Journal of Big Data*, 7-37. Retrieved https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00313-w
11. Lang, M., & Little, D. (2018). Principled Missing Data Treatments. *Prevention Science*, 284-294. doi:https://doi.org/10.1007/s11121-016-0644-5
12. Newman, D. (2003). Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. *Organizational Research Methods*, 328-362. Retrieved https://journals.sagepub.com/doi/10.1177/1094428103254673
13. Oblaković, M., Sokolovska, V., & Dinić, B. (2015). Tretmani nedostajućih podataka. *Primenjena psihologija*, 289-309. doi:10.19090/str.2015.3.289-309
14. Rubin, D., & Little, R. (2002). *Statistical Analysis with Missing Data*, Second Edition. New Jersey: John Wiley & Sons, Inc. Retrieved https://onlinelibrary.wiley.com/doi/book/10.1002/9781119013563
15. Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*. doi:10.1037/1082-989X.7.2.147
16. Schafer, L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 3-15. doi:https://doi.org/10.1177/096228029900800102
17. Sunil, R. (2016). *A Comprehensive Guide to Data Exploration*.
18. Swalin, A. (2018). *How to Handle Missing Data*. Retrieved https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
19. Tabachnick, B., & Fidell, L. (2001). *Using multivariate statistic* (4th edition). Allyn and Bacon, Boston.
20. Tabachnick, G., & Fidell, L. (2013). *Using Multivariate Statistics*. Pearson Education.
21. Vasić, V. (2018). Rešavanja problema multivarijacionih nedostajućih anketnih podataka primenom EM algoritma. *Ekonomske ideje i praksa*, 35-50. Retrieved http://www.ekof.bg.ac.rs/wp-content/uploads/2014/10/003.pdf
22. Warnes, Z. (2021, July 6). *Missing Value Handling — Missing Data Types*. Retrieved Towards Data Science: https://towardsdatascience.com/missing-value-handling-missing-data-types-a89c0d81a5bb
23. Xian, L. (2016). *Methods and Applications of Longitudinal Data Analysis*. Academic Press. doi:https://doi.org/10.1016/C2013-0-13082-6