PRIMENJENA PSIHOLOGIJA, 2015, Vol. *8*(3), STR. 289-309

UDK 159.9:519.21
Originalni naučni rad

Mirjana Oblaković

Centar za primenjenu statistiku, Univerzitet u Novom Sadu

Valentina Sokolovska

Odsek za sociologiju, Filozofski fakultet, Univerzitet u Novom Sadu

Bojana Dinić[1]

Odsek za psihologiju, Filozofski fakultet, Univerzitet u Novom Sadu

TRETMANI NEDOSTAJUĆIH PODATAKA[2]

U radu je dat kritički osvrt na najčešće korišćene tretmane nedostajućih podataka: tradicionalne, kao što su isključivanje nedostajućih podataka i jednostruke imputacije (zamena nedostajućih podataka srednjom vrednošću, imputacija pomoću regresije, slučajna imputacija), moderne, kao što su tretmani zasnovani na maksimalnoj verodostojnosti (npr. EM algoritam i FIML metod) i metodi višestruke imputacije. Ukazano je na prednosti i mane svakog od ovih tretmana i preporuke u vezi sa odabirom tretmana u odnosu na mehanizam nedostajanja podataka, tip i nivo merenja varijable, veličinu uzorka i slično. Takođe, dat je pregled prakse tretmana kategorijalnih i numeričkih nedostajućih podataka u psihologiji u objavljenim radovima u vrhunskim psihološkim časopisima. Zaključeno je da je najčešće korišćen tradicionalni metod izbacivanja slučajeva sa nedostajućim vrednostima, a potom se u nešto manjem broju koristi metod multiple imputacije. S obzirom na to, u radu je dat primer sprovođenja multiple imputacije u SPSS-u.

Ključne reči: analiza nedostajućih podataka, multipla imputacija, kategorijalne varijable, numeričke varijable

[1] Adresa autora:
bojana.dinic@ff.uns.ac.rs.

Primljeno: 11. 02. 2015.
Primljena korekcija:
29. 04. 2015.
Prihvaćeno za štampu:
14. 05. 2015.

[2] Rad je nastao u okviru projekta „Nasledni, sredinski i psihološki činioci mentalnog zdravlja“ koji finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije (projekat broj 179006).

U društvenim istraživanjima se često suočavamo sa problemom nedostajućih podataka. Graham i saradnici (Graham, Cumsille, & Elek-Fisk, 2003) razlikuju dva razloga nastajanja nedostajućih podataka. Prema jednom, ispitanici učestvuju u istraživanju ali iz neodređenog razloga ne odgovore na neka pitanja. Prema drugom, zbog osipanja uzorka u longitudinalnim studijama dolazi do nedostajućih čitavih kompleta podataka. Pored ova dva osnovna razloga, mogu se razlikovati i specifični razlozi, kao što su izostajanje odgovora na pitanja s početka ili s kraja primenjenih upitnika, kao i izostajanje ispitanika iz poslednje ili prve faze longitudinalnog istraživanja.

Iako bi bilo veoma poželjno utvrditi razlog nedostajanja podataka u cilju precizne analize podataka, često se moramo zadovoljiti samo opisima nedostajućih podataka. Naime, po završetku prikupljanja podataka, potrebno je utvrditi mehanizam po kojem podaci nedostaju odnosno njihovu distribuciju i zavisnost distribucije od izmerenih karakteristika uzorka. Obrazac nedostajanja podataka mnogo je važniji od same količine nedostajućih podataka. Ove obrasce, odnosno mehanizme nedostajanja podataka, možemo svesti na tri osnovna (Little & Rubin, 1987): 1. potpuno slučajno distribuirani nedostajući podaci (engl. *missing completely at random – MCAR*), 2. slučajno distribuirani nedostajući podaci (engl. *missing at random – MAR*) i 3. podaci koji ne nedostaju po slučajnom rasporedu (engl. *missing non at random – MNAR*). Razlika između MCAR i MAR je u tome što izostajanje podataka u okviru MCAR nije pod uticajem nijedne varijable u datom setu. S druge strane, izostajanje podataka po MAR mehanizmu nije pod uticajem varijable u kojoj ima nedostajućih podataka, ali jeste pod uticajem neke druge varijable u setu podataka. Zamislimo da nam je cilj istraživanja da ispitamo stavove prema politici u odnosu na socio-demografske karakteristike. Vrlo je moguće, na primer, da ispitanici s određenim stavom ne žele da daju odgovor na pitanje o visini primanja, i u tom slučaju postoji pristrasnost u davanju odgovora na pitanje o primanjima (MAR mehanizam). Isključivanjem uzroka nedostajućih podataka u opisanom slučaju ne bismo dobili nepristrasne procene parametara u modelu (Graham et al., 2003). Ukratko, ne bismo mogli da generalizujemo rezultate. Ipak, ukoliko podaci nedostaju po MAR mehanizmu, nedostajući podaci se mogu objasniti raspoloživim podacima, jer su varijable koje su povezane sa ovim uzrokom izmerene i mogu se uključiti u model u cilju dobijanja nepristrasnih procena parametara. Ukoliko, pak, distribucija nedostajućih podataka zavisi od samih nedostajućih podatka, kažemo da podaci ne nedostaju po slučajnom rasporedu, odnosno da se distribuiraju po MNAR obrascu (Schafer & Graham, 2002). U ovom slučaju, nedostajući podatak je povezan s razlogom zašto nedostaje. U skladu sa prethodnim primerom, to bi bila tendencija ispitaniká s većim primanjima da ne izveštavaju o svojim primanjima. Analize nad ovakvim podacima daju pristrasne procene parametara (Graham, 2009; Schafer & Graham, 2002).

Proveru pretpostavke o potpuno slučajnoj raspodeli nedostajućih podataka možemo dobiti na primer preko Littleovog MCAR testa koji je dostupan u SPSS-u. Ukoliko ovaj test nije značajan, možemo zaključiti da nedostajući podaci ne odstupaju značajno od potpuno slučajne raspodele. Ponekad je korisno proveriti rela-

cije nedostajanja podataka i varijabli u modelu. To se jednostavno može testirati. Potrebno je napraviti dihotomnu varijablu u kojoj će se napraviti razlika u odnosu na to da li nedostajući podaci postoje ili ne, i testirati da li se ova dva poduzorka razlikuju u odnosu na ostale varijable iz modela (Tabachnick & Fidell, 2001).

Podaci na numeričkim, ali i kategorijalnim varijablama, mogu nedostajati po ma kom od navedenih mehanizama. Ipak, možemo reći da nedostajući podaci na kategorijalnim varijablama ponekad predstavljaju teži problem. Naravno, problematika zavisi od samog tipa podataka. Ukoliko su u pitanju nominalne kategorije, tim je teže predvideti, zameniti ili nadomestiti nedostajuću vrednost. Pored oblika distribucije i tipa nedostajućih podataka, u obzir treba uzeti i količinu nedostajućih podataka, veličinu uzorka istraživanja, pouzdanost primenjenih instrumenata ukoliko su korišćeni, broj i tip ostalih prikupljenih podataka itd. S obzirom na ovako veliki broj činilaca, valja naglasiti da je svako istraživanje specifično i da je pre ikakve analize potrebno razumevanje podataka i uzimanje u obzir svih činilaca koji potencijalno ugrožavaju njihovu pouzdanost i validnost. Ipak, kao i kod svih statističkih procedura, potrebno je postaviti ciljeve, pa prema njima slediti opšte preporuke o tome kako se i pod kojim uslovima mogu tretirati nedostajući podaci ukoliko želimo dobiti validne rezultate analiza.

Tretmani nedostajućih podataka i preporuke

Postoje brojni načini tretiranja nedostajućih podataka koji uključuju tradicionalne i moderne tretmane. Tradicionalni su npr. isključivanje nedostajućih podataka i jednostruke imputacije. Moderni tretmani su tzv. metodi zasnovani na modelu, kao što su metodi zasnovani na maksimalnoj verodostojnosti i metodi višestruke imputacije. Većina ovih metoda koristi se kako za zamenu nedostajućih vrednosti na numeričkim, tako i na kategorijalnim varijablama. Ipak, neke od njih nisu primerene za korišćenje na kategorijalnim varijablama, dok se druge preporučuju.

Isključivanje nedostajućih podataka

Postoje dva načina isključivanja nedostajućih podataka. Prvi način se odnosi na isključivanje nedostajućih podataka u celini (engl. *listwise deletion*) tzv. pametno brisanje sa liste, a drugi podrazumeva tzv. isključivanje nedostajućih podataka po parovima (engl. *pairwise deletion*). Ovi tretmani nedostajućih podataka su među starijim metodama, no njihovo korišćenje je još uvek veoma rasprostranjeno.

Isključivanje nedostajućih podataka u celini podrazumeva jednostavno rad samo sa onim slučajevima koji nemaju nijedan nedostajući podatak. Graham i saradnici (Graham et al., 2003) navode da su oba načina isključivanja nedostajućih podataka, generalno, neprihvatljiva. Tome u prilog idu rezultati koji pokazuju da se isključivanjem slučajeva gubi na snazi testa (Graham et al., 2003; Ilić, 2012).

Rezultati u vezi sa isključivanjem su prilično konzistentni – isključivanje podataka u celini može dati nepristrasne procene parametara ukoliko su nedostajući podaci MCAR i ako nedostaje manje od 5% slučajeva, ali ne i ukoliko su podaci MAR tipa (Allison, 2002; Arbuckle, 1996; Brown, 1994; Chan, 1998; Muthén, Kaplan, & Hollis, 1987; Wothke 2000). Ipak, rezultati navedenih istraživanja pokazuju da tretiranje nedostajućih podataka na ovaj način rezultira manje efikasnim procenama parametara, čak i kad su podaci MCAR tipa. Iako ova procedura ima očiglednih manjkavosti, u okviru statističkih paketa često je automatski odabrana opcija. King i saradnici (King, Honaker, Joseph, & Scheve, 2001) nalaze da 94% analiza anketnih istraživanja koriste upravo potpuno isključivanje nedostajućih podataka, i da analitičari gube prosečno trećinu podataka na taj način.

Isključivanje po parovima podrazumeva isključivanje iz analize samo onih slučajeva koji imaju nedostajuće podatke na varijablama na kojima se vrši analiza. Na primer, ukoliko analiza počiva na korelacionoj matrici, svaka pojedinačna korelacija iz matrice će biti računata posebno, na svim raspoloživim podacima. Iako se ovim načinom iskorišćavaju svi podaci kojima raspolažemo, ovaj način ima još veće manjkavosti u odnosu na prethodno opisani. Naime, vidimo i iz prethodnog primera korelacione matrice, da će svaka korelacija biti računata na različitom skupu podataka. Šta više, na ovaj način moguće je čak generisati interkorelacionu matricu čija determinanta nije pozitivno definitna (Graham et al., 2003; Ilić, 2012). To bi značilo da matrica rezultira negativnim vrednostima karakterističnih korenova, odnosno da je izolovana varijansa nekih komponenti negativna. Dakle, ova procedura može dati pristrasne procene parametara, upravo zbog različitih uzoraka na kojima su računate pojedinačne vrednosti (Graham et al., 2003). Kako ne postoji jedinstvena veličina uzorka, ne postoji ni osnov za računanje standardnih grešaka parametara.

Jednostruke imputacije

Zamena nedostajućih podataka srednjom vrednošću. U slučaju numeričkih varijabli, ovaj tretman podrazumeva da se nedostajući podatak zameni aritmetičkom sredinom dobijenom na slučajevima koji imaju podatak na datoj varijabli. U slučaju kategorijalnih varijabli, zamena bi se morala vršiti modalnom vrednošću, a u slučaju ordinalnih – medijanom. Većina autora se slaže da je zamena nedostajućih vrednosti srednjom vrednošću neprihvatljiva i da daje pristrasne procene parametara (Fichman & Cummings, 2003; Graham et al., 2003). Ovim tretmanom je varijansa varijable s nedostajućim podacima sužena, što može uticati na visinu korelacija s ostalim varijablama u modelu (Tabachnick & Fidell, 2001). Eventualni kompromis može se napraviti ukoliko se nedostajući podatak ne zameni srednjom vrednošću na celom uzorku za datu varijablu, već srednjom vrednošću za poduzorak s karakteristikama slučaja koji ima nedostajući podatak. Ipak, i ova varijanta rezultira suženom varijansom varijable koja sadrži nedostajuće podatke, ali je ovaj sužen varijabilitet ograničen na specifični poduzorak slučajeva (Tabachnick & Fidell, 2001).

Imputacija pomoću regresije. Ova vrsta imputacije podrazumeva predviđanje nedostajućih podataka na osnovu ostalih izmerenih varijabli. Ovaj tretman je, takođe, dosta kritikovan. Na primer, Little (1992) tvrdi da će procenjene greške regresionih koeficijenata prilikom imputacija biti potcenjene, pošto se greška imputacije ne uzima u obzir prilikom njihovog računanja. Graham i saradnici (Graham et al., 2003) tvrde da jednostruka imputacija putem regresije stvara znatnu pristrasnost. Kako se nedostajući podaci predviđaju na osnovu postojećih podataka, ovaj metod se ne preporučuje pri analizi kovarijanse ili korelacija, jer se njime precenjuje povezanost između varijabli, odnosno obrazac korelacija biva veštački konzistentniji (Tabachnick & Fidell, 2001). Takođe, navodi se još nekoliko nedostatka ovog tretmana. Kao i u slučaju zamene srednjom vrednošću, nedostatak se odnosi na redukovanu varijansu varijable s nedostajućim podacima, jer je njena procena verovatno bliža srednjoj vrednosti. Drugi nedostatak odnosi se na to da varijable u modelu na osnovu kojih se vrši imputacija moraju biti dobri prediktori varijable na kojoj postoje nedostajući podaci, kako bi ovaj tretman nedostajućih podataka imao smisla (Tabachnick & Fidell, 2001).

Slučajna imputacija (engl. *hot deck imputation*). Slučajna imputacija podrazumeva imputaciju izmerene vrednosti nekog drugog slučaja sa sličnim obrascem odgovora na ostalim izmerenim varijablama (uglavnom na osnovu najmanje Euclideanske distance) ili vrednost date varijable nasumično odabranog slučaja iz uzorka (Newman, 2003). Iako se imputacija pomoću regresije možda i češće koristi od slučajne imputacije, slučajna imputacija pokazuje nekoliko prednosti. Najpre, metod slučajne imputacije ne zahteva parametrijski model, odnosno pogodniji je za kategorijalne podatke od imputacije regresijom. Zatim, njome se bolje uspostavlja gubitak varijabiliteta odnosno varijansa varijable nije potcenjena, kao u slučaju prethodno opisanih tretmana. Potom, pristrasnosti u procenjenim parametrima su manje ukoliko su podaci MCAR tipa, iako se pokazuje da dolazi do precenjivanja greške tipa I (Brown, 1994). Ipak, neka istraživanja pokazuju da i ovaj postupak može da daje iskrivljene procene korelacija i ostalih mera asocijacije, bez obzira na mehanizam nedostajanja podataka (Schafer & Graham, 2002). Uprkos nedostacima, čini se da je slučajna imputacija uveliko zastupljena u praksi tj. u anketnim istraživanjima od strane vladinih agencija kako u svetu (Chen & Astebro, 2003), tako i kod nas (Ilić, 2012).

Postavlja se pitanje koji je od ovih tretmana najbolji. Rezultati istraživanja u vezi s ovim problemom prilično su oprečni. U jednom istraživanju je vršeno poređenje 8 tretmana, ali bez uključivanja modernih metoda (Switzer, Roth, & Switzer, 1998). U ovom istraživanju je pokazano da je zamena aritmetičkom sredinom dala najlošije rezultate, a isključivanje nedostajućih podataka je dalo rezultate s najmanjom pristrasnošću, dok za njima slede imputacija pomoću regresije i slučajna imputacija. S druge strane, neka istraživanja pokazuju da slučajna imputacija i zamena aritmetičkom sredinom daju bolje rezultate u odnosu na imputaciju regresijom (Mundfrom & Whitcomb, 1998). Newman (2003), na primer, navodi da sve metode imputacije poseduju fundamentalnu manu potcenjivanja standardnih grešaka parametara, samim tim što dodaju vrednosti u nepotpun set podataka

i time precenjuju stvarnu veličinu uzorka. Schafer i Graham (Schafer & Graham, 2002) dolaze do zaključka da ni jedna od navedenih metoda nije adekvatna, tj. ne dovodi do tačne tj. nepristrasne i efikasne procene parametara i njihovih standardnih grešaka. Ovakav zaključak je prilično obeshrabrujući, ali treba imati u vidu da u ovim istraživanjima nisu obuhvaćeni svi tretmani nedostajućih podataka. Novija istraživanja (Arbuckle, 1996; Baraldi & Enders, 2010; Finch, 2010; Graham, 2009; Graham, Hofer, & MacKinnon, 1996; Muthén et al, 1987; Schafer & Graham, 2002; Wothke, 2000) ukazuju na to da metodi tretiranja nedostajućih podataka koji su zasnovani na modelu daju bolje rezultate, te smo njih, kao i njihove prednosti i mane, detaljnije prikazali u nastavku.

Metodi višestruke imputacije

Metodi višestruke imputacije, zajedno s metodima zasnovanim na maksimalnoj verodostojnosti, spadaju u grupu metoda zasnovanih na modelu. U okviru ovih metoda se pretpostavlja kakva je udružena distribucija svih varijabli u modelu (Pigott, 2001). Njihova osnovna prednost je u tome što daju nepristrasne procene u slučaju MCAR i MAR mehanizama nedostajanja podataka.

U okviru metoda višestruke imputacije najpre se primeni logistička regresija u kojoj je kriterijum dihotomna varijabla koja sadrži podatke o tome da li na toj varijabli ima ili nema nedostajućih podataka. U ovom koraku se dobijaju predviđene vrednosti nedostajućih podataka i normalno distribuirani reziduali predikovanih vrednosti. Dakle, u ovom koraku se stvara nekoliko nasumično odabranih poduzoraka slučaja koji imaju kompletirane podatke, kako bi se identifikovala distribucija varijable koja sadrži nedostajuće podatke. Nakon toga, u narednom koraku ponovo se stvara nekoliko nasumično odabranih poduzoraka, ali ovoga puta uključujući i slučajeve s nedostajućim podacima. U okviru ovih poduzoraka se sve nedostajuće vrednosti zamenjuju procenama na osnovu rezultata regresije iz prvog koraka. Iz ovako dobijenih setova, računaju se pojedinačne procene, a do konačnih procena modela se dolazi uprosečavanjem parametara procene iz ovako kreiranih multiplih setova (Tabachnick & Fidell, 2001). Pri tome, zadaje se određeni broj imputacija tj. broj ponavljanja celog postupka i u svakoj imputaciji se računaju konačne procene. Najjednostavniji metod dobijanja konačnih ocena iz više setova je Rubinov metod skalarnih (jednodimenzionalnih) parametara (Rubin, 1987). Detaljan opis konkretnih operacija dobijanja konačnih parametara može se naći u Rubin (1987), Schafer i sar. (Scafer et al., 2002) i Finch (2010).

Da bi konačne procene bile adekvatne, potrebno je doneti nekoliko odluka. Prva se odnosi na izbor metoda na kojem će biti zasnovana imputacija. Metodi višestruke imputacije uglavnom se zasniva na multivarijatnoj normalnosti, pri čemu je najčešće korišćen MCMC (engl. *Markov chain Monte Carlo*). Međutim, izgleda da su ovi metodi prilično robusni na kršenje ovog uslova (Allison, 2000; King et al., 2001; Schafer, 1997). Peng i Zhu (Peng & Zhu, 2008) navode da metodi višestruke imputacije pokazuju prednost u primeni nad kategorijalnim podacima, u odnosu, na primer, na metode maksimalne verodostojnosti. Pomenuti MCMC

metod je za numeričke varijable, a ukoliko su varijable kategorijalne, postoje drugi metodi imputacije kao, na primer, MIC (engl. *multiple imputation for categorical data*) ili FCS metod (engl. *fully conditional specification*) zanovani na multinominalnoj distribuciji. Ipak, ti metodi mogu da postanu veoma komplikovani u slučaju da postoji više varijabli sa više nivoa, i njihovo opisivanje podrazumeva veliki broj interakcija, pa zahtevaju i veoma velike uzorke da bi procene bilo moguće izvesti. Stoga Schafer (1997) preporučuje upotrebu višestruke imputacije za numeričke varijable sa zaokruživanjem zamenjenih vrednosti tako da uzmu vrednosti koje su moguće za datu kategorijalnu varijablu. Finch (2010) nalazi da za varijable ordinalnog tipa, višestruka imputacija za numeričke varijable sa zaokruživanjem daje efikasnije i manje pristrasne procene od višestruke imputacije za nominalne varijable. Međutim, drugi istraživači ukazuju na oprez prilikom ovakvog postupka i to da treba uzeti u obzir mehanizam nedostajanja podataka, kompleksnost modela, distribuciju kategorijalnih podataka i slično (Dong & Peng, 2013).

Druga odluka odnosi se na to koje varijable treba uvrstiti u model kao prediktore i nju donosi sam istraživač na osnovu poznavanja fenomena koji ispituje. Ovaj aspekt, između ostalog, predstavlja razliku u odnosu na metode zasnovane, na primer, na maksimalnoj verodostojnosti. Preporuka je da se u prediktivni model uvrste varijable koje su od teorijskog značaja, koje su povezane sa mehanizmom nedostajanja i koje koreliraju sa varijablama s nedostajućim podacima (Scafer, 1997; van Buuren et al., 1999).

Treća odluka odnosi se na broj slučajnih uzoraka tj. imputacija zarad dobijanja dobrog rešenja. Rubin (1996) navodi da je 5 imputacija dovoljno, a Fichman (2003) tvrdi da je dovoljno do 10, s tim što je retko kada potrebno više od 5 imputacija. Međutim, broj imputacija zavisi, pre svega, od količine nedostajućih podataka. Bodner (2008), na primer, preporučuje da se sprovede onoliko imputacija koliko ima procenata nedostajućih podataka. Schafer i Olsen (Schafer & Olsen, 1998) daju pregled broja imputacija u zavisnosti od količine nedostajućih podataka. Na primer, ako je proporcija nedostajućih informacija .5, procene parametara sa 5, 10 i 20 imputacija su efikasne, redom, u 91%, 95%, i 98%. Proporcija nedostajućih informacija (engl. *fraction of missing information*) jeste procena odnosa kompletnih i nedostajućih informacija, i računa se kao odnos varijanse kompletnih podataka i varijanse podataka sa nedostajućim vrednostima. Ako ima nedostajućih podataka, uzorak je u stvari manji, i samim tim je varijansa veća, kao i proporcija nedostajućih informacija (Liu, 1994). S druge strane, novija istraživanja pokazuju da je, u cilju očuvanja statističke moći testova da detektuje male efekte, potrebno i više imputacija, nego što je to ranije tvrđeno. Na primer, Graham i saradnici (Graham, Olchowski, & Gilreath, 2007) pokazuju da korišćenjem 5 imputacija u slučaju kada ima 50% nedostajućih podataka, moć višestruke imputacije opada za 13% u poređenju sa, na primer, FIML metodom zasnovanom na maksimalnoj verodostojnosti, dok u slučaju kada ima 30% nedostajućih podataka – opada za 7%. Stoga, ovi autori preporučuju korišćenje čak 40 imputacija u slučaju kada ima 50% nedostajućih podataka. Generalno, kada postoji veći procenat nedostajućih podataka i kada je model kompleksniji, preporučuje se broj imputacija koji je svakako veći od 5 (Dong & Peng, 2013).

Prednosti ovog tretmana su brojne. Varijabilnost je mnogo više verna pravoj varijabilnosti, pošto se uzima u obzir varijabilnost tokom uzorkovanja i tokom imputacije. Pored toga, prednost je u mogućnosti primene na longitudinalnim podacima (tj. nacrtima s ponovljenim faktorima ili s vremenskim serijama) i na jednoajtemskim merama odnosno na podacima s jednom opservacijom na varijabli (Tabachnick & Fidell, 2001). Još jedna prednost je u tome što se dobijaju manje greške parametara u odnosu na, na primer, isključivanje ili imputacije pomoću regresije, a posebno kod sistematski nedostajućih podataka (Newman, 2003). Takođe, pokazano je da multipla imputacija daje validne procene čak i kada je multivarijatna normalnost narušena (Demirtas, Freels, & Yucel, 2008; Schafer, 1997, 1999) i kada se testira model s velikim brojem varijabli (Dong & Pong, 2013).

Metodi maksimalne verodostojnosti

Ovi metodi takođe spadaju u metode zasnovane na modelima, i podrazumevaju iterativni postupak. Drugim rečima, koriste se svi raspoloživi podaci, i kompletirani i nedostajući, kako bi se identifikovale vrednosti parametara koje imaju najveću verovatnoću pojavljivanja u posmatranim podacima (Allison, 2001). Postoji nekoliko algoritama u okviru ove porodice metoda, te ćemo u ovom radu objasniti one koji se najčešće koriste u okviru komercijalnih softvera.

EM algoritam (engl. *expectation maximization algorithm*). Ovaj algoritam podrazumeva procenu parametara kroz iterativno ponavljanje dva koraka. Prvi korak je tzv. E-korak (od engl. *expectation*) koji podrazumeva računanje funkcije verovatnoće očekivanih vrednosti podataka na osnovu poznatih (izmerenih) vrednosti i trenutne procene parametara (Enders, 2010). Dakle, prvim E-korakom se procenjuju parametri distribucije na osnovu raspoloživih podataka. U drugom, M-koraku (od engl. *maximization*) računaju se parametri za nedostajuće podatke maksimiziranjem verovatnoće dobijanja očekivanih vrednosti iz E-koraka (Newman, 2003). Ova dva koraka se ponavljaju u nizu iteracija dok se ne postigne konvergencija, odnosno dok se procene parametara ne počnu razlikovati iz iteracije u iteraciju. Procena parametara dobijena ovim metodom je nepristrasna i efikasna na velikim uzorcima, čak i ukoliko podaci nedostaju po MAR obrascu (Fichman, 2003; Schafer & Graham, 2002). Prednost ovog tretmana je što je jednostavan i što postoji manja opasnost od tzv. overfita tj. previsoke saglasnosti između podataka i modela odnosno situacije u kojoj se dobija da testirani model izgleda bolje, nego što to zaista jeste (Tabachnick & Fidell, 2001). Svojevrsni katalog EM algoritama može se naći kod Littlea i Rubina (Little & Rubin, 1987). Iako je sama procedura u statističkim paketima namenjena uglavnom numeričkim varijablama (recimo u SPSS-u), postoje i opcije za primenu na kategorijalnim varijablama (npr. u Mplus-u).

S druge strane, nedostatak ovog tretmana je što je primena ograničena na linearne modele i podatke koji su normalno distribuirani. Drugi nedostatak je povezan sa kreiranjem softvera za analizu nedostajućih podataka i odnosi se na to što softveri uglavnom ne obezbeđuju procene standardnih grešaka parametara

(postoje posebni skripti za njihovo računanje, ali su se pokazali kao komplikovani za primenu), te se EM ne preporučuje kada je primarni cilj statističko testiranje intervala poverenja procenjenih parametara (Dong & Peng, 2013).

FIML metod (engl. *full information maximum likelihood* ili *raw maximum likelihood*). Ovaj metod se često koristi u strukturalnom modelovanju i nalazi, na primer, u okviru paketa AMOS. FIML metod je robusniji na odstupanja od normalnosti ili misfit modela, te u ovim uslovima pokazuje prednost u odnosu na druge metode (Arbuckle, 1996; Enders, 2001). Međutim, ako je uslov multivarijatne normalnosti zadovoljen, FIML i EM često daju slične rezultate (Dong & Peng, 2013). No, iako FIML daje nepristrasne procene parametara u slučajevima kada je narušena normalnost, dešava se da su standardne greške precenjene i da se češće odbaci model koji u stvari ima dobar fit (Enders, 2001). Poboljšanje se može dobiti kada se u strukturalni model uvrste varijable koje su povezane sa varijablama s nedostajućim podacima ili s mehanizmom nedostajanja (Graham et al., 2003). Takođe, iako se u simulacionim studijama pokazalo da FIML metod ima dobru moć detektovanja interakcija između varijabli u modelima s relativno malim uzorcima, kada su u pitanju binarne i ordinalne varijable, potreban je mnogo veći uzorak, a ponekad i veći broj indikatora da bi se dostigla ista moć detektovanja interakcija (Huang & Bentler, 2009). Ono što je bitna prednost u odnosu na EM jeste to što FIML u softerskim paketima daje procene standardnih grešaka parametara. Pored toga, u studiji Newmana (Newman, 2003) je pokazano da samo metod višestruke imputacije i FIML metod obezbeđuju adekvatne standardne greške procenjenih parametara, dok se one potcenjuju, na primer, EM metodom.

Na osnovu navedenog može se zaključiti da se isključivanje nedostajućih podataka na bilo koji način nikako ne preporučuje. Takođe, u novijim istraživanjima retko se preporučuju tradicionalni metodi zamene putem jednostruke imputacije. Nedostaci ovih metoda mogu da budu zanemarljivi u slučajevima kada su ispunjeni određeni uslovi (npr. ukoliko su nedostajući podaci slučajno distribuirani i ukoliko je broj nedostajućih podataka zanemarljivo mali). Međutim, ponekad metod zamene nedostajućih podataka može da stvori pogrešnu sliku o podacima, kao u situaciji kada bismo kao metod imputacije koristili regresiju, a potom te iste varijable koristili za proveru nekih korelacionih modela – korelacije bi bile veštački više nego što to zaista jesu. Stoga je vrlo važno razmotriti koji tretman treba koristiti s obzirom na istraživački problem, a potom i obrazac, tip i količinu nedostajućih podataka. U novim preporukama preferiraju se moderni metodi kao što su metode višestruke imputacije i maksimalne verodostojnosti. Ti metodi daju manje pristrasne procene u odnosu na tradicionalne, čak i kada se podaci ne distribuiraju slučajno (tj. po MNAR mehanizmu). Međutim, FIML metod i metod višestruke imputacije u odnosu na npr. EM metod pokazuju dve važne prednosti. Jedna prednost je u tome što su robusniji na odstupanje od multivarijatne normalnosti, a druga što se njima mogu dobiti nepristrasnije standardne greške na osnovu kojih se mogu računati intervali poverenja. Ovo su ujedno i ključni argumenti za primenu modernih u odnosu na tradicionalne metode.

Praksa tretmana nedostajućih podataka u radovima vrhunskih časopisa

Iako u ovom radu nije dat pregled svih tehnika tretmana nedostajućih podataka, cilj je bio prikazati najstarije, najčešće korišćene i najviše preporučivane tehnike. No, neke tehnike, izgleda, uprkos teorijskim i empirijskim potvrdama kvaliteta i činjenici da su već neko vreme uključene u statističke pakete, nisu još uvek naišle na učestalost korišćenja kao neki stariji postupci. Cilj ovog dela rada je eksploracija trenutno korišćenih tretmana nedostajućih podataka u vrhunskoj nauci, konkretno u okviru oblasti psihologije. Pri tome, ispitaće se tretman nedostajućih podataka kategorijalnih i numeričkih varijabli posebno, budući da postoje različite preporuke u slučaju tretmana ovih tipova varijabli.

Metod

Uzorak članaka. U cilju odabira radova iz oblasti psihologije, pretraživana je baza podataka *Science Direct.* Pritom, uzeti su u obzir članci koji su izdati u prethodnih 5 godina, tj. od 2009. do sadašnjeg trenutka (bazi je pristupljeno maja 2014. godine). U slučaju odabira radova u kojima su nedostajući podaci s nominalne ili ordinalne skale, pretraga je podrazumevala da reč *logistic* bude u okviru rezimea rada i fraza *missing data* u okviru teksta rada. U slučaju odabira radova u kojima su nedostajući podaci s intervalne ili racio skale, pretraga je podrazumevala da fraze *linear regression* ili *multiple regression* budu u okviru rezimea rada i fraza *missing data* u okviru teksta rada. Dodatne ključne reči pretrage su uvedene iz razloga što korelacione studije čine većinu istraživanja u društvenim naukama, pa tako i u psihologiji, te smo na ovaj način hteli da steknemo uvid u uobičajenu praksu zamene nedostajućih podataka. U uzorak članaka smo ciljno hteli uključiti samo visoko kvalitetne članke iz oblasti psihologije. U skladu s tim, pretragu smo ograničili na članke iz uticajnijih časopisa sa impakt faktorom višim od 3, a dodatni kriterijum bio je da u datom časopisu postoji bar 5 rezultata ovako postavljene pretrage. Pretpostavka je da će se ovakvim izborom časopisa smanjiti mogućnost uključivanja nepreporučljivih tehnika tretiranja nedostajućih podataka u konkretnom članku koji se analizira.

Pretragom su izdvojena 34 članka koja su zadovoljavala postavljene kriterijume u slučaju kategorijalnih i ordinalnih podataka. Najveći broj članaka je bio iz 2012. godine (41%), zatim iz 2014. (21%), 2013. (18%), 2010. (12%) i u 2011. je pronađeno 9% članaka. U slučaju članaka u kojima su nedostajući podaci s intervalne ili racio skale, izdvojeno je 38 članaka. Najveći broj članaka je bio iz 2014. godine (26%), zatim iz 2013. (21%), 2012. i 2011. (18%), a iz 2010. je 16% pronađenih članaka.

Rezultati i diskusija

Iako su u većem broju izdvojenih članaka na neki način opisane varijable s nedostajućim podacima, u oko 30% radova nedostaje opis ovih varijabli, što se ne čini zanemarljivim. Pri tome, kada je naveden, opis varijabli i količine nedostajućih podataka prilično je šarenolik: varijable se nekad navode samo opisno (npr. navodi se da li su varijable sa nedostajućim podacima kategorijalne ili numeričke, navodi se da postoje nedostajući podaci na nezavisnim varijablama, i slično), nekad se imenuju samo neke varijable, nekad se imenuju varijable i navodi se količina nedostajućih podataka za svaku, a nekad je navedena samo ukupna količina nedostajućih podataka (Tabela 1). Raspon nedostajućih podataka je veliki, pri čemu u većem broju radova (38% radova sa NP na kategorijalnim varijablama i 45% radova sa NP na numeričkim varijablama) nedostaje više od 5% podataka, što ukupno, što po pojedinačnim varijablama. Takođe, mehanizam po kojem nedostaju podaci je veoma retko opisivan i testiran.

U radovima dominira tradicionalni tretman nedostajućih podataka tj. isključivanje slučajeva sa nedostajućim vrednostima (Tabela 1). Tek u manjem broju radova korišćen je metod multiple imputacije, dok, na primer, EM algoritam i FILM metod uopšte nisu korišćeni u selektovanim radovima. S obzirom na to da je isključivanje dominantni tretman nedostajućih podataka, detaljnije smo ispitali opis i količinu nedostajućih podataka u radovima u kojima je korišćen samo ovaj tretman. U okviru studija s nedostajućim podacima na varijablama nominalne i ordinalne skale, u 8 od 15 studija je navedena ukupna količina nedostajućih podataka. U preostalim studijama je navedena količina nedostajućih podataka po varijablama (4), po talasu ispitivanja (jedna longitudinalna studija) i u okviru 2 studije nije navedena količina nedostajućih podataka. Od ovih 15 studija, u 8 studija je navedena količina podataka bila veća od 5%. U okviru studija s nedostajućim podacima na varijablama intervalne i racio skale, od 16 samo u 2 studije nije naveden nikakav opis nedostajućih podataka. U 7 studija je navedena samo ukupna količina nedostajućih podataka, u 6 studija je, pored ukupne količine nedostajućih podataka, dat još neki opis količine nedostajućih podataka za neke ili sve varijable, dok je u jednoj studiji dat samo raspon nedostajućih podataka u okviru varijabli. Od ovih 16 studija, u 8 studija je navedena količina nedostajućih podataka bila više od 5%. Dakle, može se zaključiti da detaljniji opis nedostajućih podataka manjka i u studijama u kojima je korišćeno isključivanje, koje predstavlja najkorišćeniji tretman. Takođe, može se primetiti da nema upadljivih razlika u opisu nedostajućih podataka, ni primenjenim tretmanima u zavisnosti od tipa podataka.

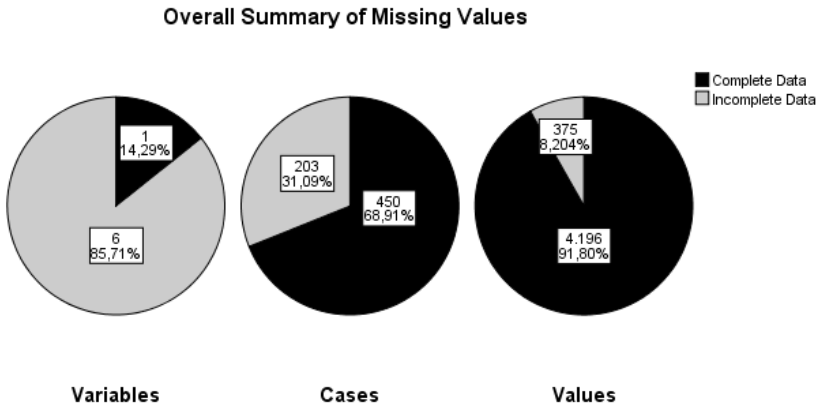
Tabela 1
Prikaz nedostajućih podataka u studijama sa kategorijalnim i numeričkim varijablama

	Studije s NP na varijablama nominalne i ordinalne skale (34 članka)	Studije s NP na varijablama intervalne i racio skale (38 članaka)
Opisane varijable sa NP	24 (71%)	26 (68%)
Imenovane varijable s NP	21 (62%)	22 (58%)
Navedena količina NP	26 (76%)	30 (79%)
Raspon ukupne količine NP	0.75–30.78% (12 članaka)	2–35% (12 članaka)
Navedena količina NP po varijablama	11(32%)	16 (42%)
Raspon količine NP po varijablama	0.3–59% (11 članaka)	0.1–40% (6 članaka)
Navedeni mehanizam NP	6 (18%)	4 (11%)
MCAR	1 (3%)	2 (5%)
MAR	3 (9%)	/
MNAR	2 (6%)	/
Nespecifikovano		2 (5%)
Tretman NP		
Isključivanje	15 (44%)	16 (42%)
Višestruka imputacija	7 (21%)	8 (21%)
Isključivanje i još neki tretman	6 (18%)	8 (21%)
Nespecifikvana imputacija	3 (9%)	3 (9%)
Zamena AS	2 (6%)	3 (9%)
Imputacija pomoću regresije	1 (3%)	/

Kako uraditi multiplu imputaciju u SPSS-u?

S obzirom na to da je višestruka imputacija metod koji se, posle isključivanja podataka, češće primenjuje kao tretman nedostajućih podataka u radovima vrhunskih časopisa, biće dat primer kako sprovesti ovaj metod u SPSS-u. U ovom prime-

ru, koristićemo matricu s 653 ispitanika koja sadrži nekoliko uobičajenih podataka koje dobijamo od učenika, kao što su pol, razred, školski uspeh, obrazovanje oca i majke, procenu materijalnog stanja, i pored toga imamo njihov skor na upitniku stavova prema nasilju. Navedene varijable su kategorijalne (*Nominal* ili *Ordinal*), osim stavova prema nasilju koji su numeričkog tipa (*Scale*). Vrlo je važno pravilno označiti varijable u samoj matrici, pre početka analize, da se ne bi desilo da se dobiju neočekivane vrednosti, na primer, pol u decimalnom zapisu i slično. Najpre treba analizirati mehanizam po kojem podaci nedostaju preko izbora opcije *Analyze/ Multiple Imputation/Analyze Patterns*... U okviru ovog prozora za dijalog u okvir *Analyze Across Variables* treba prebaciti varijable u kojima želimo zameniti nedostajuće podatke i čekirati sve tri ponuđene opcije u Outputu. Kao minimalni procenat nedostajućih podataka za varijable koje želimo prikazati, možemo staviti 0,1 umesto 10, kako bi obuhvatili veći broj varijabli s nedostajućim podacima. Pritiskom na OK dobijamo ispis koji je organizovan u 4 dela. Prvi deo, *Overall Summary of Missing Values* (Slika 1), odnosi se na frekvencu i procenat varijabli, slučajeva-ispitanika i ćelija u matrici koji sadrže nedostajuće podatke. Na osnovu Slike 1 se može videti da u našem primeru 6 od unetih 7 varijabli (tj. 85.71% unetih varijabli) sadrži nedostajuće podatke, 203 (31.09%) ispitanika ima nedostajuće podatke i 375 ćelija u matrici ima nedostajuće vrednosti, što čini 8.204% od ukupnog broja podataka.



Slika 1. Opšti pregled nedostajućih podataka.

Drugi deo, tabela *Variable Summary*, daje prikaz frekvenci i procenta nedostajućih podataka po varijablama, po opadajućem redosledu. Na osnovu Slike 2 se može videti da najveći procenat nedostajućih podataka ima varijabla školski uspeh (28.2%) i obrazovanje oca (24.3%), dok je procenat nedostajućih podataka ostalih varijabli manji od 5%. Kao što se može primetiti, varijabla koja se odnosi na razred učenika nije data u tabeli jer je to varijabla u kojoj nema nedostajućih podataka. S obzirom na to da je varijabla koja se odnosi na stavove prema nasilju numerička, tj. označena kao *Scale* prilikom definisanja varijable u matrici, za nju smo dobili i podatak o aritmetičkoj sredini i standardnoj devijaciji na dostupnim podacima.

Variable Summary[a,b]

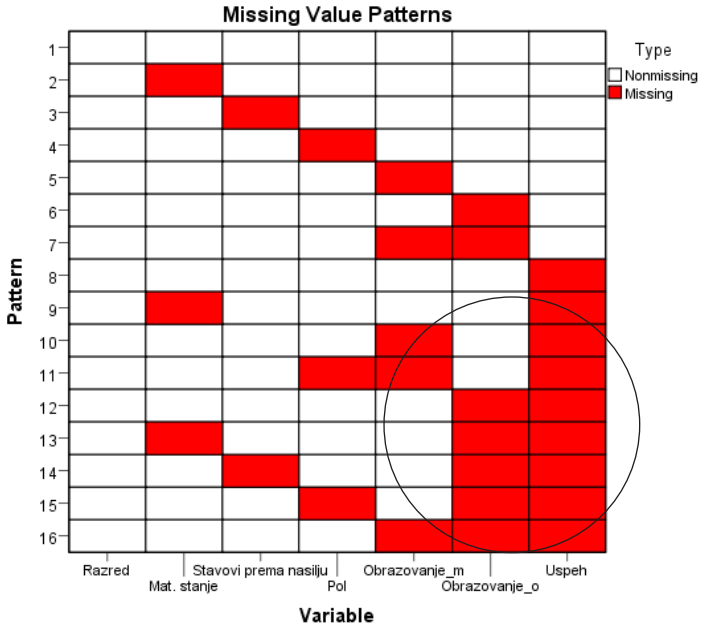
	Missing		Valid N	Mean	Std. Deviation
	N	Percent			
Uspeh	184	28,2%	469		
Obrazovanje_o	159	24,3%	494		
Obrazovanje_m	12	1,8%	641		
Pol	10	1,5%	643		
Stavovi prema nasilju	5	0,8%	648	37,5787	8,92974
Mat. stanje	5	0,8%	648		

a. Maximum number of variables shown: 25

b. Minimum percentage of missing values for variable to be included: 0,0%

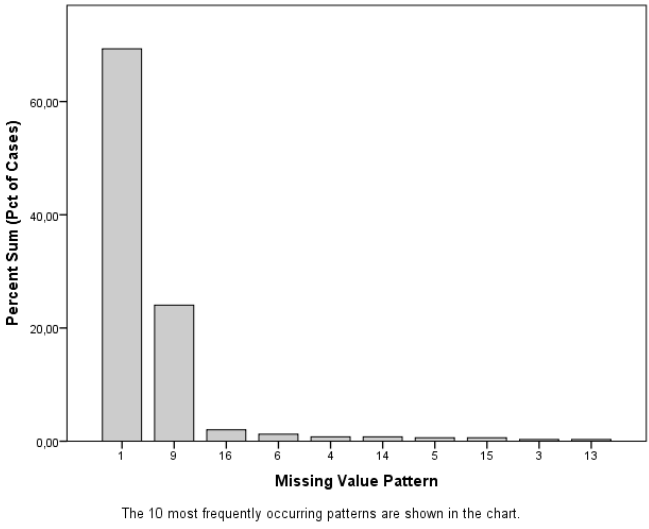
Slika 2. Pregled varijabli po nedostajućim podacima.

Potom sledi prikaz obrazaca nedostajućih podataka (Slika 3). U redovima ovog grafika su numerisani obrasci nedostajanja podataka, a u kolonama su varijable. Prvi obrazac, u prvom redu, obrazac je u kojem nema nedostajućih podataka. Dakle, ovaj red predstavlja grupu slučajeva na kom nema nedostajućih podataka. Drugi obrazac je u drugom redu i u našem primeru predstavlja grupu slučajeva kojima nedostaje samo materijalno stanje i tako svaki sledeći red predstavlja novi obrazac u kojem nedostaje sve veći broj slučajeva i varijabli. Poslednji, 16. red, tj. obrazac, jeste obrazac po kojem nedostaju podaci u vezi s obrazovanjem oba roditelja i školskim uspehom. Dakle, ovaj red predstavlja grupu slučajeva koji imaju nedostajuće podatke na ove tri varijable. Kolone su poređane u odnosu na procenat nedostajućih podataka po varijabli. Tako, varijabla razred nema nedostajuće podatke, dok najveći broj nedostajućih podataka ima varijabla uspeh. Ovaj prikaz služi vizuelnoj inspekciji mehanizma po kojem podaci nedostaju. Ukoliko se uoči neka pravilnost ili koncentracija u osenčenim poljima, onda se može konstatovati da postoji neki obrazac po kojem podaci nedostaju tj. podaci *ne* nedostaju po slučajnom obrascu. Ukoliko postoji (konstantan) porast ili smanjenje osenčenih površina (kao u našem primeru) to ukazuje da nedostajući podaci monotono nedostaju tj. prisutan je monotonicitet (npr. ako po jednom obrascu nedostaje jedan podatak, po drugom dva, po sledećem tri ili više, itd.). Monotonicitet je samo jedan od MNAR mehanizma po kojem podaci ne nedostaju po slučajnom redosledu. Ukoliko se tako nešto ne uočava i ukoliko osenčena polja predstavljaju „ostrvca" koja su nasumično raspoređena, može se zaključiti da nedostajući podaci ne pokazuju monotonicitet. Određenje prisutnosti monotoniciteta je potrebno zbog određenja metoda imputacije, mada se u SPSS-u nudi i automatski odabir metoda u skladu sa podacima. Na Slici 3 se može primetiti da postoji koncentracija osenčenih polja u donjem desnom uglu, što je uokvireno, te da je u podacima najverovatnije prisutan monotonicitet.



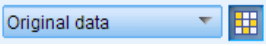
Slika 3. Obrasci po kojima podaci nedostaju.

I na kraju sledi grafički prikaz obrazaca nedostajućih podataka (Slika 4). Može se videti da je najčešći prvi obrazac po kojem, zapravo, nema nedostajućih podataka, pa potom obrazac 9 po kojem nedostaju materijalno stanje i školski uspeh, dok su ostali obrasci zastupljeni u relativno podjednakom procentu.



Slika 4. Procenat slučajeva s određenim obrascem po kojem podaci nedostaju.

Nakon uvida u strukturu nedostajućih podataka, možemo pristupiti zameni. Iz padajućeg menija odabiramo *Analyze/Multiple Imputation/Impute Missing Data Values*… i prebacimo varijable u okvir *Variables in Model*. Broj imputacija se može menjati, ali po preporukama zadržaćemo 5 imputacija. Matrica s zamenjenim nedostajućim vrednostima može biti sačuvana kao nova matrica ili u okviru neke postojeće matrice. Mi ćemo je sačuvati kao novu matricu i daćemo joj ime, na primer, *impute.matrix*. U sledećem jezičku *Method* možemo odabrati metod imputacije. U SPSS-u se nude dva metoda: 1. MCMC koji se primenjuje u slučaju da podaci nedostaju po slučajnom obrascu, i 2. metod koji se primenjuje u slučaju da postoji monotonicitet. U našem primeru ćemo zadržati automatsku procenu metoda, budući da će se u okviru ove opcije podaci skenirati u odnosu na monotonicitet i ukoliko se monotonicitet detektuje, primeniće se drugi metod. U okviru jezička *Constraints* možemo dobiti procenat nedostajućih podataka po varijabli i deskriptivne pokazatelje varijabli klikom na dugme *Scan Data*. U okviru ovog jezička se nude još neke opcije, ali one nisu predmet ove demonstracije. U sledećem jezičku *Output* možemo čekirati sve tri ponuđene kućice. Ukoliko čekiramo opciju za kreiranje istorije iteracija u novoj matrici, moramo dati ime toj novoj matrici, na primer, *iteration.history*.

U okviru ispisa, u tabeli *Imputation Models* može se videti koji tip modela je primenjen za koje varijable, u skladu sa definisanim nivoom merenja (logistička ili linearna regresija), kao i koji su se prediktori koristili za predikciju nedostajućih vrednosti u okviru svake od varijabli. U ovoj tabeli se takođe može videti broj nedostajućih i imputovanih vrednosti. Potom, za svaku varijablu je dat prikaz originalnih vrednosti unutar varijable i imputovane vrednosti u svakoj imputaciji, kao i konačni broj vrednosti nakon svake imputacije. Kada odemo na novu matricu sa zamenjenim vrednostima (*impute.matrix*), u desnom uglu možemo videti padajući meni 

. U njemu možemo videti na koji način su podaci zamenjeni u svakoj imputaciji od 5 sprovedenih. U matrici će ćelije sa zamenjenim vrednostima biti osenčene.

Nad zamenjenim podacima dobijenim u poslednjoj, 5. imputaciji, možemo raditi obradu podataka. Ukoliko se određena obrada može sprovesti nad novim, zamenjenim podacima, pored opcije za obradu će se nalaziti oznaka  . Na primer, nas je interesovalo da li postoje polne razlike u uspehu, pa smo sproveli neparametrijski *t*-test tj. *Mann-Whitney U* test budući da nam je uspeh definisan kao ordinalna varijabla. U izlazu se dobijaju rezultati za originalni set podataka i za setove podataka u svakoj imputaciji. Na osnovu Slike 5 može se videti da *p*-nivo značajnosti u originalom setu podataka ukazuje na to da nema značajnih razlika (p = .068), dok u matrici sa zamenjenim podacima u 5. imputaciji rezultat pokazuje da postoje značajne razlike (p = .008). Ovo je dobar primer kako zamena nedostajućih podataka može promeniti rezultat.

Test Statistics[a]

Imputation_ Imputation Number		Uspeh
0 Original data	Mann-Whitney U	7065,000
	Wilcoxon W	36468,000
	Z	-1,823
	Asymp. Sig. (2-tailed)	,068
1	Mann-Whitney U	10590,500
	Wilcoxon W	60993,500
	Z	-2,207
	Asymp. Sig. (2-tailed)	,027
2	Mann-Whitney U	10167,500
	Wilcoxon W	58995,500
	Z	-2,520
	Asymp. Sig. (2-tailed)	,012
3	Mann-Whitney U	10482,000
	Wilcoxon W	59310,000
	Z	-2,135
	Asymp. Sig. (2-tailed)	,033
4	Mann-Whitney U	10340,000
	Wilcoxon W	59168,000
	Z	-2,309
	Asymp. Sig. (2-tailed)	,021
5	Mann-Whitney U	9963,500
	Wilcoxon W	59104,500
	Z	-2,644
	Asymp. Sig. (2-tailed)	,008

a. Grouping Variable: Pol

Slika 5. Rezultati *U* testa u originalnim podacima i podacima iz svake sprovedene imputacije.

Reference

Allison, P. D. (2002). *Missing data*, *Sage University papers series on quantitative applications in the social sciences*, *series 07–136.* Thousand Oaks, CA: Sage.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5–37.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, *15*, 651–675.

Brown, R. L. (1994). Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods. *Structural Equation Modeling*, *1*, 287–316.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrated approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods*, *1*, 421–483.

Chen, G., & Astebro, T. (2003). How to deal with missing categorical data: Test of a simple Bayesian method. *Organizational Research Methods*, *6*, 309–327.

Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, *78*(1), 69–84.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *Springer Plus*, *2*(1), 1–17.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know. *Organizational Research Methods*, *6*(3), 282–308.

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science*, *8*, 361–378.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 87–114). New York: John Wiley & Sons.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*, 197–218.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213.

Huang, W., & Bentler, P. (2009). Estimating latent variable interactions with binary and ordinal data. *Multivariate Behavioral Research*, *44*, 852.

Ilić, I. (2012). *Ocenjivanje indeksa repa raspodele korišćenjem nekompletnih uzoraka* (Neobjavljena doktorska disertacija). Matematički fakultet, Univerzitet u Beogradu, Beograd.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Association*, *95*(1), 49–69.

Little, R. J. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*, 1227–1237.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Liu, J. S. (1994). Fraction of missing information and convergence rate for data augmentation. In J. Sall & A. Lehman (Eds.), *Computing Science and Statistics: Proceedings of the 26th Symposium on the Interface* (pp. 490–496). Interface Foundation of North America, Fairfax Station, VA.

Mundfrom, D. J., & Whitcomb, A. J. (1998). Imputing missing values: The effect on the accuracy of classification. *Multiple Linear Regression Viewpoints*, *25*(1), 13–19.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*, 328–362.

Peng, C. Y. J., & Zhu, J. (2008) Comparison of two approaches for handling missing covariates in logistic regression. *Educational Psychological Measurement*, *68*(1), 58–77.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, *7*, 353–383.

Rubin, D. B. (1987). *Multiple imputation for non response in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*(1), 3–15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545–571.

Switzer, F. S., III, Roth, P. L., & Switzer, D. M. (1998). Systematic data loss in HRM settings: A Monte Carlo analysis. *Journal of Management*, *24*, 763–779.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th edition). Boston, MA: Allyn and Bacon.

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistic in Medicine*, *18*, 681–694.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues*, *applied approaches and specific examples* (pp. 219–240). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mirjana Oblaković

Center for Applied Statistics, University of Novi Sad

Valentina Sokolovska

Department of Sociology, Faculty of Philosophy, University of Novi Sad

Bojana Dinić

Department of Psychology, Faculty of Philosophy, University of Novi Sad

TREATMENTS OF MISSING DATA

This paper presented a critical review of the most commonly used treatments of missing data: traditional, such as case deletion and single imputation (mean imputation, imputation by regression, hot deck imputation), and modern, such as multiple imputation methods and maximum likelihood methods (EM algorithm, FIML method). We described their advantages and disadvantages and recommendations regarding selection of treatment in relation to the missing data mechanism, the type and measurement level of variables, sample size, etc. Also, paper included an overview of treatment practices of categorical and numerical missing data in psychology articles published in top journals. It was concluded that the most commonly used treatment is the traditional listwise deletion, while multiple imputation is a slightly less used method. Thus, we provided an example of implementation of multiple imputation in SPSS.

Keywords: missing data analysis, multiple imputation, categorical variables, numerical variables