# STATISTIČKE I MATEMATIČKE METODE ZA REŠAVANJE PROBLEMA KLASTEROVANJA POŠTANSKIH PODATAKA KADA SU ONI NEPOTPUNI

Nataša Glišović[1,2], Tatjana Davidović[2], Nebojša Bojović[3], Nikola Knežević[3]
[1]Državni Univerzitet u Novom Pazaru, Departman za matematičke nauke
[2]Matematički Institut SANU
[3]Univerzitet u Beogradu, Saobraćajni fakultet
nglisovic@np.ac.rs, tanjad@mi.sanu.ac.rs,
nb.bojovic@sf.bg.ac.rs, n.knezevic@sf.bg.ac.rs

**Sadržaj:** *Kvalitet podataka je presudni faktor od kojeg zavisi uspešnost obrade podataka. Značajnu ulogu u kvalitetu podataka osim izvora imaju i postupci pretprocesiranja podataka. Podaci u izvornom obliku mogu biti nekompletni, atributi mogu imati nedostajuće vrednosti ili može postojati nedostatak atributa. Cilj ovog istraživanja je da pokažemo matematičke i statističke metode koje rešavaju probleme svrtstavanja podataka u grupe kao i rada sa nedostajućim podacima. Metode su testirane na bazi podataka Evropske komisije koja sadrži podatke vezane za poštanski saobraćaj.*
**Ključne reči:** *poštanski saobraćaj, nedostajući podaci, problem p-medijane, metaheuristike.*

## 1. Uvod

Poštanski saobraćaj i usluge predstvaljaju jedan od servisa od opšteg društvenog interesa. Praćenje i analiza učinka poštanskih kompanija imaju ključnu ulogu u podizanju nivoa kvaliteta poštanske usluge. Statistike i pokazatelji koji karakterišu poštanski saobraćaj su brojni i često je njihovo prikupljanje i formiranje baza podataka ograničeno njihovom dostupnošću i efikasnošću samog sistema. Primena modernih statističkih i matematičkih metoda u evaluaciji poslovanja poštanskih kompanija omogućuje sveobuhvatnu analizu koja uključuje veliki broj pokazatelja, kao i veliku količinu podataka.

Podaci u izvornom obliku mogu biti nekompletni, atributi mogu imati nedostajuće vrednosti, ili može postojati nedostatak atributa. Isto tako može se pojaviti nekonzistentnost unutar samih podataka kao posledica nedoslednosti u označavanju pojedinih kategorija ili grupa. Kada se u podacima naiđe na nedostajuću vrednost, tada se u procesu analize podataka koriste metode za predviđanje nedostajućih vrednosti npr. neuronske mreže, regresione metode, linearna interpolacija, Bajesove mreže, stabla odlučivanja i slično [10]. U ovom radu taj problem prevaziđen je korišćenjem rastojanja

koje se može primeniti i u slučaju nedostajućih vrednosti [1], tako da su vrednosti ostavljene u izvornom obliku onakvim kakve jesu, tj. nismo aproksimirane nedostajuće vrednosti u bazi.

Pored problema nedostajućih podataka, drugi aspekt obuhvata i heterogenost samih pokazatelja i podataka koji karakterišu poslovanje poštanskih kompanija. Klasterovanje predstvalja pristup kojim se uspešno može tretirati heterogenost podataka. Cilj ovog rada je analiza dostupnih baza podataka o poštanskom saobraćaju kojom se podaci grupišu u klastere čiji broj je dobijen statističkim pretprocesiranjem.

Rad je podeljen u nekoliko odeljaka. Metodologija predložena za rešavanje problema data je u narednom odeljku. Opis podataka kao i rezultati istraživanja dati su u Odeljku 3. Poslednji odeljak sadži zaključna razmatranja.

## 2. Metodologija

### 2. 1. Nedostajući podaci

Problem nedostajućih podataka sve češće se javlja pri analizi savremenih baza podataka. Podaci mogu nedostajati iz više razloga. Neki od njih su: podaci nisu raspoloživi, došlo je do grešaka u radu sa opremom, nekonzistentnosti sa drugim podacima, pa su zato izbrisani, nisu unešeni zbog nerazumevanja, nisu smatrani bitnim u trenutku unosa itd. Bitna je odluka šta raditi sa nedostajućim podacima. Neke od mogućnosti su [3]:

- Izbrisati elemente kod kojih se javljaju nedostajući podaci-što nije preporučljivo posebno kod klasifikacije, a naročito ako nedostajuće vrednosti varijaju od elementa do elementa, tj. nedostaju različiti elementi kod različitih objekata (vektora).
- Ručno popunjavanje nedostajućih vrednosti koje je zamorno i često neizvodljivo.
- Automatsko popunjavanje: nekom opštom konstantom, srednjom vrednosti elemenata za sve objekte (vektore) koji pripadaju istoj klasi.
- Najverovatnija vrednost-zaključak se donosi na osnovu Bajesove formule ili prema stablu odlučivanja.

Kako ni jedna od navedenih mogućnosti ne obezbeđuje zadovoljavajuću transformaciju polazne baze, u ovom istraživanju problem je rešen korišćenjem rastojanja koje se koristi u slučajevima kada podaci nedostaju, predloženom od strane Glišović i Rašković [1]. Ovo rastojanje zasnovano je na logičkim formulama i ne zahteva popunjavanje nedostajućih podataka niti brisanje nekih od atributa što je njegova osnovna prednost.

### 2.2. Pretprocesiranje

Normiranje podataka se često primenjuje kada je potrebno izbeći veliki uticaj pojedine promenljive koja gravitira ka visokim apsolutnim vrednostima kod reševanja problema klasterovanja.

Od metoda normiranja podataka koje se najviše koriste u obradi podataka su:

- Min-max normiranje
- Z-sklairanje
- Decimalno skaliranje

S obzirom na prirodu podataka koji se koriste u ovom istraživanju (postoje nedostajuće vrednosti, tj. nisu nam poznate maximalna i minimalna vrednost niza) opredelili smo se za Z skaliranje, kao metodu adekvatnu u ovom koraku obrade podataka.

$$y' = \frac{y - y_s}{\sigma_y}, \tag{1}$$

gde je *y'* nova (normirana vrednost), *y* izvorna vrednost atributa, $y_s$ srednja vrednost, a $\sigma_y$ standardna devijacija poznatih (postojećih) atributa. Srednja vrednost i standardna devijacija se računaju na uobičajeni način:

$$y_s = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_s)^2} \tag{3}$$

## 2. 3. Klasterovanje

Klasterovanje podrazumeva da se slični podaci (u odnosu na odgovarajuće atribute) grupišu zajedno u grupe koje nazivamo klasteri. Dok su elementi unutar klastera slični, klasteri se među sobom razlikuju. Klasterovanje je vid nenadgledanog učenja (engl. unsupervised learning) jer klasteri nisu određeni pre ispitivanja podataka.

Postoji nekoliko formulacija problema klasterovanja zavisno od funkcije cilja koja se optimizuje, a u ovom istraživalju je korišćena formulacija preko problema p-medijane [8].

Neka su $x_{ij}$ i $y_j$ binarne promenljive definisane na sledeći način:

$$x_{ij} = \begin{cases} 1, & \text{ako se objekat } i \text{ nalazi u klasteru } j, \\ 0, & \text{inače.} \end{cases}$$

$$y_j = \begin{cases} 1, & \text{ako objekat } j \text{ reprezentuje odgovarajući klaster,} \\ 0, & \text{inače.} \end{cases}$$

$$\min \sum_i \sum_j d_{ij} x_{ij} \tag{4}$$

t. d.

$$\sum_j x_{ij} = 1 \text{ za svako } i \tag{5}$$

$$x_{ij} \le y_j \text{ za svako } i,\ j \tag{6}$$

$$\sum_j y_j = p \tag{7}$$

$$x_{ij}, y_j \in \{0,1\} \tag{8}$$

## 2. 4. Osnovna metoda promenljivih okolina

Metoda promenljivih okolina je metaheuristika koja je predstavljena devedesetih godina prošlog veka [7][9] nakon čega je doživela mnogo promena i ekstenzija [5][6], kao i uspešnih primena [4].

Osnovna metoda promenljivih okolina (BVNS) je najrasprostranjenija varijanta metode promenljivih okolina jer obezbeđuje više preduslova za dobijanje kvalitetnijih konačnih rešenja. Kod BVNS metode osnovni koraci sadržani su u petlji u kojoj menjamo indeks okoline i, određujemo slučajno rešenje iz te okoline (korak razmrdavanja), izvršavamo proceduru lokalnog pretraživanja počev od tog slučajnog rešenja i proveravamo kvalitet dobijenog lokalnog minimuma u odnosu na trenutno najbolje rešenje. Ove korake ponavljamo dok ne bude zadovoljen neki od kriterijuma zaustavljanja. Uloga koraka razmrdavanja je da obezbedi diversifikaciju pretraživanja. Prilikom svakog odabira okoline početna rešenja generišemo na slučajan način kako bi obezbedili pretraživanje različitih regiona kod svakog sledećeg razmrdavanja u okolini i. Lokalnim pretraživanjem rešenja dobijenog razmrdavanjem intenzivira se pretraga rešenja u njegovoj okolini. Adekvatan balans između intenzifikacije i diversifikacije obezbeđuje se pravilnim izborom vrednosti $k_{max}$, osnovnog parametara BVNS metode.

Okoline koje se koriste u BVNS metodi razlikujemo po broju transformacija (rastojanju) ili po vrsti transformacija (metrici). Napominjemo da okoline za izbor slučajnog rešenja (razmrdavanje) i lokalno pretraživanje ne moraju biti istog tipa. Pseudokod BVNS metode dat je na slici 1.

*Inicijalizacija*. Izabrati početno rešenje $x \in X$ i definisati kriterijum zaustavljanja STOP=0
**Ponavljaj**
{
  $i = 1$
  **Ponavljaj**
  {
    Generisati slučajno rešenje $x'$ u $i$-toj okolini od $X$- Razmrdavanje(i).
    $x''$=Lokalno pretraživanje ($x'$)
    **Ako** je lokalni minimum bolji od trenutnog minimuma
      $x_{optimalno} = x''$
      $f(x_{optimalno}) = f(x'')$
      $i = 1$
    **Inače** preći u $i = i + 1$
    **Ako** je zadovoljen kriterijum zaustavljanja
      STOP=1.
  } dok nije $i = i_{max}$ ili STOP=1
} sve dok nije STOP=1

*Slika 1. Pseudokod BVNS metode*

BVNS metoda za kalsterovanje podataka vezanih za poštanski saobraćaj implementirana je po uzoru na rad Glišović N., Davidović T. i Rašković M. [2].

## 3. Rezultati istraživanja

Podaci koji su korišćeni u ovom istraživanju nalaze se u bazama Evropske komisije, i dostupne su sa adrese http://ec.europa.eu/eurostat/web/postal-services/data/database. Postoji šest baza koje za svaku od zemalja sadrže sledeće podatke:

- Ukupan obim poštanskih usluga u unutrašnjem i međunarodnom saobraćaju, Prihod od poštanskih usluga u unutrašnjem saobraćaju
- Cene poštanskih usluga u unutrašnjem saobraćaju
- Kvalitet i rokovi prenosa u unutrašnjem i međunarodnom saobraćaju
- Dostupnost poštanske mreže
- Broj zaposlenih u nacionalnim poštanskim operatorima
- Obim pismonosnih usluga u unutrašnjem i međunarodnom saobraćaju

Posmatrane baze podataka obuhvataju podatke za period 2004-2011 godina za 31 broj zemalja: Austrija, Belgija, Bugarska, Hrvatska, Kipar, Češka, Danska, Estonija, Finska, Bivša Jugoslovenska Republika Makedonija, Francuska, Nemačka, Grčka, Mađarska, Island, Irska, Italija, Latvija, Litvanija, Luksemburg, Malta, Holandija, Norveška, Poljska, Portugal, Rumunija, Slovačka, Slovenija, Španija, Švedska, Velika Britanija. Sve korišćene baze karakterišu se nedostajućim podacima. Važno je napomenuti da postoje zemlje za koje nisu navedeni podaci ni za jednu godinu. Te zemlje isključene su iz analize u odgovarajućoj bazi i podrazumevano je da su one svrstane u jedan zajednički klaster.

Nakon isključivanja zemalja kod kojih nisu postojali podaci ni za jednu godinu određen je procenat nedostajućih podataka. Karakteristike svake od baza sumirane su u tabeli 1.

*Tabela 1. Opis svake baze koja je analizirana iskazanje kroz broj zemalja koje se nalaze u bazi, broj Zemalja kod kojih nema podataka ni za jednu godinu (broj isključenih zemalja), kao i procenat nedostajućih podataka.*

| Naziv baze | Broj zemalja | Broj isključenih zemalja | Procenat nedostajućih podataka |
|---|---|---|---|
| Ukupan obim poštanskih usluga u unutrašnjem i međunarodnom saobraćaju | 31 | 8 | 30% |
| Prihod od poštanskih usluga u unutrašnjem saobraćaju | 30 | 0 | 14% |
| Cene poštanskih usluga u unutrašnjem saobraćaju | 31 | 0 | 7% |
| Kvalitet i rokovi prenosa u unutrašnjem i međunarodnom saobraćaju | 31 | 0 | 10% |
| Dostupnost poštanske mreže | 31 | 3 | 11% |
| Broj zaposlenih u nacionalnim poštanskim operatorima | 31 | 0 | 11% |
| Obim pismonosnih usluga u unutrašnjem i međunarodnom saobraćaju | 31 | 3 | 12% |

Prvo je izvršeno pretprocesiranje, zatim analiza kako podataka tako i njihove prirode formirane su grupe sličnih zemalja sa sličnim vrednostima atributa (određeni su intervali vrednosti u svakoj grupi). Zatim je primenjen BVNS algoritam za klasterovanje u okviru svake baze podataka na onoliko klastera koliko je određeno u pretprocesiranju podataka.

BVNS metoda implementirana je u C# programskom jeziku na računaru HP-15-d055, pod operativnim sistemom Windows 10 Pro. U pretprocesiranju za svaku od baza, na osnovu analiza stručnjaka, ocenjeno je da su razvrstavnja u tri grupe najadakvatnija. S obzirom na stohastičku prirodu metoda vršeno je 100 restartovanja. Najbolja rerešenja, kao i broj puta koliko su dostignuta, zajedno sa prosečnim vremenom potrebnim za nalaženje najboljih rešenja dati su u tabeli 2. Najbolje rešenje karakterisano je vrednošću funkcije cilja, tj. zbirom rastojanja unutar klastera. Metoda je uzvršavan 100 puta i pokazala je veliku stabilnost za svaku bazu.

*Tabela 2. Rezultati rada BVNS-a za svaku bazu podataka data po optimalnim rešenjima, uspešnosti i vremenu dolaska do optimalnih rešenja.*

| BVNS primenjen na bazama | Njabolje rešenje | Broj dostignutih najboljiih rešenja (uspešnost) | Vreme dolaska PROSEČNO do optimalnog rešenja (sekundama) |
|---|---|---|---|
| Ukupan obim poštanskih usluga u unutrašnjem i međunarodnom saobraćaju | 7954014.83 | 88 | 0.16 |
| Prihod od poštanskih usluga u unutrašnjem saobraćaju | 20188.09 | 100 | 0.3 |
| Cene poštanskih usluga u unutrašnjem saobraćaju | 2.40 | 99 | 0.45 |
| Kvalitet i rokovi prenosa u unutrašnjem i međunarodnom saobraćaju | 371.46 | 92 | 0.64 |
| Dostupnost poštanske mreže | 1163584.85 | 92 | 0.65 |
| Broj zaposlenih u nacionalnim poštanskim operatorima | 303556.29 | 100 | 0.14 |
| Obim pismonosnih usluga u unutrašnjem i međunarodnom saobraćaju | 14360498.72 | 95 | 0.62 |

Za svaku od ovih baza dajemo pregled klastera koji su dobijeni gde su različitim bojama predstavljeni različiti klasteri (Tabela 3.).

*Tabela 3. Klasteri dobijeni primenom metode BVNS..*

| Ukupan obim pošt. usluga u unutraš. i međun. saobraćaju. | Prihod od poštanskih usluga u unutraš. saobraćaju | Cene poštanskih usluga u unutraš. saobraćaju | Kvalitet i rokovi prenosa u unutraš.i međun. saobraćaju | Dostupnost poštanske mreže | Broj zaposlenih u NPO | Obim pismonos. usluga u unutraš. i međun. saobraćaju |
|---|---|---|---|---|---|---|
| Bulgaria | Belgium | Belgium | Belgium | Belgium | Belgium | Bulgaria |
| Czech R. | Bulgaria | Bulgaria | Bulgaria | Bulgaria | Bulgaria | Czech R. |
| Denmark | Czech R. | Czech R. | Czech R. | Czech R. | Czech R. | Denmark |
| Estonia | Denmark | Denmark | Denmark | Denmark | Denmark | Germany |
| Greece | Germany | Germany | Germany | Germany | Germany | Estonia |
| Spain | Estonia | Estonia | Estonia | Estonia | Estonia | Ireland |
| Croatia | Ireland | Ireland | Ireland | Ireland | Ireland | Greece |
| Italy | Greece | Greece | Greece | Greece | Greece | Spain |
| Cyprus | Spain | Spain | Spain | Spain | Spain | Croatia |
| Latvia | France | France | France | Croatia | France | Italy |
| Lithuania | Croatia | Croatia | Croatia | Cyprus | Croatia | Cyprus |
| Luxembourg | Italy | Italy | Italy | Latvia | Italy | Latvia |
| Hungary | Cyprus | Cyprus | Cyprus | Luxembourg | Cyprus | Lithuania |
| Austria | Latvia | Latvia | Latvia | Hungary | Latvia | Luxembourg |
| Poland | Lithuania | Lithuania | Lithuania | Malta | Lithuania | Hungary |
| Romania | Luxembourg | Luxembourg | Luxembourg | Netherlands | Luxembourg | Malta |
| Slovenia | Hungary | Hungary | Hungary | Austria | Hungary | Netherlands |
| Slovakia | Malta | Malta | Malta | Poland | Malta | Austria |
| Finland | Netherlands | Netherlands | Netherlands | Portugal | Netherlands | Poland |
| Sweden | Austria | Austria | Austria | Romania | Austria | Portugal |
| Iceland | Poland | Poland | Poland | Slovenia | Poland | Romania |
| Norway | Portugal | Portugal | Portugal | Slovakia | Portugal | Slovenia |
| FYRM | Romania | Romania | Romania | Finland | Romania | Slovakia |
| | Slovenia | Slovenia | Slovenia | Sweden | Slovenia | Finland |
| | Slovakia | Slovakia | Slovakia | UK | Slovakia | Sweden |
| | Finland | Finland | Finland | Iceland | Finland | Iceland |
| | Sweden | Sweden | Sweden | Norway | Sweden | Norway |
| | UK | UK | UK | FYRM | UK | FYRM |
| | Iceland | Iceland | Iceland | | Iceland | |
| | Norway | Norway | Norway | | Norway | |
| | | FYRM | FYRM | | FYRM | |

## 4. Zaključak

U radu je izvršena analiza baza podataka vezanih za poštanski saobraćaj u evropskim zemljama. Na osnovu te analize podaci su grupisani u odgovarajući broj klastera koji sadrže slične objekte. Za klasterovanje je korišćena osnovna metoda promenljivih okolina koja koristi rastojanje pogodno za objekte u bazama kod kojih nedostaju podaci.

Predložena metoda pokazala je veliku stabilnost i uspešnost u klasifikaciji podataka poštanskih servisa i usluga. Tako gurpisani podaci mogu se dalje koristiti u cilju što bolje analize zemalja čiji su se poštanski servisi i usluge razmatrani. Predloženi pristup može omogućiti primenu benčmarking i drugih alata u cilju analize poštanskih servisa kada su raspoloživi podaci nekompletni.

**Zahvalnica**

Ovaj rad je rezultat istraživanja na projektu III 044006 "Razvoj novih informaciono - komunikacionih tehnologija, korišćenjem naprednih matematičkih metoda, sa primenama u medicini, energetici, telekomunikacijama, e-upravi i zaštiti nacionalne baštine" i projektu TR 36022 "Upravljanje kritičnom infrastrukturom za održivi razvoj u poštanskom, komunikacionom i železničkom sektoru Republike Srbije" koje finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

**Literatura**

[1] Glišović, N., Rašković. M., Optimization for Classifying the Patients Using the Logic Measures for Missing Data, Scientific publications of the State University of Novi Pazar Ser. a: Appl. Math. Inform. and Mech. vol. 9, 1, 91-101, 2017.

[2] Glišović, N., Davidović, T., and Rašković, M., Klasterovanje kada podaci nedostaju korišćenjem metode promenljivih okolina, SYM-OP-IS, Zlatibor, 25-28. septembra, pp. 158-165, 2017.

[3] Graham, J. W., Missing Data: Analysis and Design. Springer Science and Business Media, New York, 2012.

[4] Hansen, P. and Mladenović, N., Variable neighborhood search. In Search methodologies (pp. 313-337). Springer US, 2014.

[5] Hansen, P., Mladenović, N. and Pérez, J. A. M., Variable neighborhood search: methods and applications. Annals of Operations Research, *175*(1), 367-407, 2010.

[6] Hansen, P., Mladenović, N., Brimberg, J. and Perez, J. A. M., Variable neighborhood search Handbook of Metaheuristics ser. International Series in Operations Research & Management Science, 146, 61-86, 2010.

[7] Mladenović, N., A Variable neighborhood algorithm – a new metaheuristic for combinatorial optimization, Abstracts of papers presented at Optimization Days, Montreal, p. 112, 1995.

[8] Mladenović, N., Brimberg, J., Hansen, P., Moreno-Perez JA, The p-median problem: a survey of metaheuristic approaches. European Journal of Operational Research 179:927–939, 2007.

[9] Mladenović, N., Hansen, P., Variable neighborhood search. Computers and Operations Research; 24(11); 1097–1100, 1997.

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

***Abstract**: Data quality is a crucial factor that depends on the success of data processing. A significant role in the quality of data other than sources has a process of data preprocessing. The data in the original form may be incomplete, the attributes may have missing values, or there may be a lack of attributes. The aim of this research is to show the mathematical and statistical methods that solve the problem of grouping into groups as well as working with missing data. The methods were tested on the European Commission database, the database statistics database.*

***Key words**: postal traffic, missing data, p-median problem, metaheuristics.*

**STATISTICAL AND MATHEMATICAL METHODS FOR SOLVING THE POSTAL DATA CLASSIFICATION PROBLEMS WHEN ARE MISSING DATA**

Nataša Glišović, Tatjana Davidović, Nebojša Bojović, Nikola Knežević