УНИВЕРЗИТЕТ У БЕОГРАДУ

ФАКУЛТЕТ ОРГАНИЗАЦИОНИХ НАУКА

Немања С. Миленковић

МЕТОДОЛОГИЈА ОТКРИВАЊА НЕСТАНДАРДНИХ ОПСЕРВАЦИЈА У *K*–ДИМЕНЗИОНОМ ПРОСТОРУ

докторска дисертација

Београд, 2019.

UNIVERSITY OF BELGRADE

FACULTY OF ORGANIZATIONAL SCIENCES

Nemanja S. Milenković

METHODOLOGY FOR OUTLIER DETECTION IN *K*–DIMENSIONAL SPACE

doctoral dissertation

Belgrade, 2019.

МЕНТОР:

др Зоран Радојичић, редовни професор
Факултет организационих наука, Универзитет у Београду

ЧЛАНОВИ КОМИСИЈЕ:

др Драган Вукмировић, редовни професор
Факултет организационих наука, Универзитет у Београду

др Милица Булајић, редовни професор
Факултет организационих наука, Универзитет у Београду

др Александар Ђоковић, ванредни професор
Факултет организационих наука, Универзитет у Београду

др Свјетлана Јанковић-Шоја, доцент
Пољопривредни факултет, Универзитет у Београду

Датум одбране рада: _____

МЕТОДОЛОГИЈА ОТКРИВАЊА НЕСТАНДАРДНИХ ОПСЕРВАЦИЈА У *K*–ДИМЕНЗИОНОМ ПРОСТОРУ

Сажетак:

Предмет истраживања ове докторске дисертације је формирање методологије за откривање мултиваријационих нестандардних опсервација кроз унапређење методе Ивановићевог одстојања.

Откривање нестандардних опсервација у *k*-димензионом простору је подједнако важно као и њихово откривање у једној димензији. Под појмом „нестандардна опсервација" се подразумева она опсервација која је на неки начин неконзистентна са преосталима из посматраног скупа. Откривање мултиваријационих нестандардних опсервација се најчешће спроводи коришћењем методе Махаланобисовог одстојања.

Ивановићево одстојање се користи у циљу мерења интензитета неке појаве, коришћењем већег броја изабраних индикатора. Унапређена метода Ивановићевог одстојања тестира значајност сваког од посматраних индикатора коришћењем одговарајуће *F* статистике. Кроз употребу дефинисаних процедура за елиминацију и/или селекцију индикатора, нова метода тежи формирању „оптималног" скупа индикатора, редукујући димензију посматраног комплексног проблема.

Метода секвенцијалног Ивановићевог одстојања узима у обзир дискриминациону моћ сваког од коришћених индикатора. У складу с тим, формира се јединствена вредност одстојања за сваку опсервацију из посматраног скупа. Резултати истраживања показали су да се ова метода може успешно користити за откривање мултиваријационих нестандардних опсервација.

Кључне речи:

нестандардне опсервације, *k*-димензиони простор, Ивановићево одстојање, мултиваријациона анализа, индикатори, елиминација, селекција.

Научна област:

Техничке науке

Ужа научна област:

Рачунарска статистика

УДК број:

_____

METHODOLOGY FOR OUTLIER DETECTION IN *K*–DIMENSIONAL SPACE

The subject of this doctoral dissertation is the development of the methodology for detecting multivariate outliers through the modification of the Ivanović (I-distance) distance method.

Detecting outliers in the *k*-dimensional space is as important as detecting them in a single dimension. The term outlier refers to the observation which is in some way inconsistent with the rest of the observations in a data set. Multivariate outliers are most commonly detected using the Mahalanobis distance method.

I-distance is used to measure the intensity of an occurrence, using a number of selected indicators. An improved method of I-distance tests the significance of each of the observed indicators using the appropriate F statistics. Through defined procedures for the elimination and /or selection of indicators, the new method seeks to form an optimal set of indicators, while reducing the dimension of the complex problem at hand.

The stepwise I-distance method takes into account the discriminatory power of each of the indicators used. Accordingly, a unique I-distance value is formed for each observation from the observed set. The research results show that this method can be used to detect multivariate outliers.

Keywords:

outliers, *k*-dimensional space, Ivanović distanse, multivariate analysis, indicators, elimination, selection.

Scientific Area:

Tehnical sciences

Specific Scientific Area:

Computational statistics

UDK Number:

_____

САДРЖАЈ

1. УВОД.....1
2. МУЛТИВАРИЈАЦИОНА АНАЛИЗА.....6
2.1 Анализа главних компонената.....11
2.2 Факторска анализа.....16
2.3 Вишеструка линеарна регресија.....23
2.4 Дискриминациона анализа.....32
3. НЕСТАНДАРДНЕ ОПСЕРВАЦИЈЕ.....39
3.1 Методе откривања универијационих нестандардних опсервација.....40
3.2 Методе откривања мултиваријационих нестандардних опсервација.....43
4. ИВАНОВИЋЕВО ОДСТОЈАЊЕ.....53
4.1 Дефиниција и особине И-одстојања.....55
4.2 Обично И-одстојање.....61
4.3 Квадратно И-одстојање.....61
4.4 Структурно И-одстојање.....62
4.5 Редоследна класификација и И-одстојање.....64
5. СЕКВЕНЦИЈАЛНО ИВАНОВИЋЕВО ОДСТОЈАЊЕ.....66
5.1 Процедура Ивановићевог одстојања за постепену елиминацију индикатора.....70
5.2 Процедура Ивановићевог одстојања за постепену селекцију индикатора.....77
5.3 Процедура Ивановићевог одстојања „корак по корак”.....82
6. ПРИМЕНА СЕКВЕНЦИЈАЛНОГ ИВАНОВИЋЕВОГ ОДСТОЈАЊА.....84
6.1 Откривање нестандардних опсервација – економска развијеност.....86
6.2 Откривање нестандардних опсервација – пољопривредна производња.....92
6.3 Откривање нестандардних опсервација – перформансе кошаркаша.....100
7. ЗАКЉУЧАК.....100
ЛИТЕРАТУРА.....111
ПРИЛОЗИ.....125

1. УВОД

Идентификација нестандардних опсервација у *k*-димензионом простору је подједнако важна као и њихова идентификација у једној димензији (Barnet & Lewis, 1994). Овај тип опсервација може довести до погрешног израчунавања параматара узорка (Liu et al., 2012), а самим тим и лоше процене параметара популације. Постоји велики број дефиниција нестандардних опсервација. Најчешће цитирана дефиниција је да је то „опсервација која толико одступа од осталих, да изазива сумњу да је генерисана неким другачијим механизмом" (Hawkins, 1980).

Методе откривања нестандардних опсервација имају широку примену у природним и друштвеним појавама. Њихова основна подела је на униваријационе (методе које проналазе екстремне вредности једне променљиве) и мултиваријационе (методе које проналазе нетипичне комбинације измерених вредности неколико променљивих). Још једна подела метода за откривање нестандардних опсервација је на параметарске и непараметарске (Han et al., 2011). Параметарске методе захтевају познату расподелу одабраних променљивих, па су због тога често непогодне за коришћење. Непараметарске методе се могу поделити у две подкласе. Прву чине методе које су засноване на техникама кластеровања, а другу методе које се базирају на мерама удаљености, односно одстојању.

Када су у питању методе засноване на одстојању, откривање мултиваријационих нестандардних опсервација се најчешће спроводи коришћењем Махаланобисовог одстојања. Махаланобисов метод израчунава одстојање опсервације од израчунате средине свих опсервација за све измерене променљиве. Велика вредност одстојања сугерише да посматрана опсервација значајно одступа од осталих у *k*-димензионом простору дефинисаном посматраним променљивима (Stevens, 1984).

У овој докторској дисертацији, за откривање нестандардних опсервација у *k*-димензионом простору, коришћена је метода Ивановићевог одстојања. Ивановић (1973) је ову методологију формирао у циљу рангирања земаља на

основу њиховог нивоа социо-економског развоја, измереног коришћењем неколико индикатора. Одабрани индикатори су сублимирани у једну вредност, која ће након тога представљати ранг земље.

Анализе спроведене коришћењем Ивановићевог одстојања су бројне. Коришћено је при рангирању најбољих светских универзитета (Jeremić et al., 2011a; Jovanović et al., 2012; Dobrota M. P. et al., 2016; Dobrota M. M. & Dobrota M. P., 2016), мерењу индекса људског развоја (Jeremić et al., 2011b), економске развијености земаља (Jednak et al., 2018; Milenković et al., 2016a; Milenković et al., 2016b) одрживог развоја (Radojičić et al., 2012), ефикасности образовања у основним школама (Milenković et al., 2013), испитивању ИКТ инфраструктуре (Jeremić et al., 2011c; Jeremić et al., 2011d; Dobrota M. P. et al., 2012; Dobrota M. P. et al., 2015) мерењу развијености здравствених система земаља (Al-Lagilli et al., 2011; Jeremić et al., 2011e; Jeremić et al., 2012; Seke et al., 2013), мерењу ефикасности банака (Bulajić et al., 2013) итд. Ивановићево одстојање је такође коришћено у мерењу социо-економске развијености земаља Блиског Истока и Северне Африке (Al Lagili 2013, Milenković et al., 2014).

За одређени вектор индикатора $X^T=(X_1,X_2,...,X_k)$ изабраних да репрезентују опсервације, Ивановићево одстојање између две опсервације се дефинише као:

$$D(r,s)=\sum_{i=1}^{k}\frac{\left|d_i\left(r,s\right)\right|}{\sigma_i}\prod_{j=1}^{i-1}\left(1-r_{ji.12\ldots j-1}\right)$$

где је $d_i(r,s)$ одстојање између вредности индикатора X_i опсервација e_r и e_s, тј. дискриминациони ефекат, σ_i стандардна девијација од X_i, а $r_{ji.12..j-1}$ је парцијални коефицијент корелације између X_i и X_j, ($j<i$). Рачунање вредности Ивановићевог одстојања је итеративно и врши се кроз неколико етапа (Ivanović, 1977). Прво се израчунава вредност дискриминационог ефекта за индикатор X_1 (најважнији индикатор, онај који пружа највећу количину информација о посматраној појави). Након тога се додаје вредност дискриминационог ефекта индикатора X_2 који није покривен индикатором X_1, па се процедура понавља за све индикаторе. (Al Lagili, 2013).

Учешће посматраних индикатора у Ивановићевом одстојању опада са њиховим рангом, па се у пракси често користи квадратно Ивановићево одстојање, које је дато као

$$D^2(r,s) = \sum_{i=1}^{k} \frac{d_i^2(r,s)}{\sigma_i^2} \prod_{j=1}^{i-1} (1 - r_{ji.12\ldots j-1}^2)$$

Углавном се у истраживањима дешава да између посматраних индикатора постоје негативни коефицијенти корелације и негативни парцијални коефицијенти корелације. То је још један од разлога за употребу квадратног Ивановићевог одстојања уместо обичног.

Предмет истраживања докторске дисертације биће анализа постојећих метода за откривање нестандардних опсервација у *k*-димензионом простору. Посебна пажња ће бити посвећена употреби Ивановићег одстојања као мултиваријационе методе и његове могућности за откривање мултиваријационих нестандардних опсервација.

Циљ истраживања је да се унапреди метода Ивановићевог одстојања кроз дефинисање процедура за постепену елиминацију индикатора, постепену селекцију индикатора, као и процедуре „корак по корак“. Поменуте процедуре ће бити коришћене у циљу редукције димензије проблема, као и за откривање мултиваријационих нестандардних опсервација.

На основу дефинисаног предмета и циља истраживања, формиране су следеће хипотезе:

Општа хипотеза

- Могуће је извршити идентификацију мултиваријационих нестандардних опсервација коришћењем методологије Ивановићевог одстојања кроз процедуре за постепену селекцију и/или елиминацију индикатора.

Посебне хипотезе

- Могуће је имплементирати процедуре, засноване на Ивановићевом одстојању, за постепену селекцију индикатора, постепену елиминацију индикатора, као и процедуру „корак по корак“,.
- Коришћењем Ивановићевог одстојања и дефинисаних процедура за постепену селекцију и/или елиминацију индикатора, могуће је идентификовати статистички значајне индикатора за посматрани вишедимензиони проблем.
- Идентификација нестандардних опсервација кроз Ивановићево одстојање коришћењем дефинисаних процедура за постепену селекцију и/или елиминацију индикатора може омогућити проналажење карактеристичних опсервација у *k*-димензионом простору, као и детаљнији увид у структуру података посматраног проблема.

У докторској дисертацији ће бити примењене следеће методе истраживања: методе дескриптивне анализе, методе индукције-дедукције, методе анализе и синтезе и методе компаративне анализе. Поред општих метода истраживања, користиће се и посебне статистичке методе које произилазе из формулисаног предмета и циља истраживања: методе мултиваријационе анализе, корелациона анализа, метода Ивановићевог одстојање, као и метода Махаланобисовог одстојања. Могућност примене предложене методологије за откривање мултиваријационих нестандардних опсервација ће бити приказана на реалним примерима и подацима, уз коришћење савремених статистичких пакета.

Наредно поглавље посвећено је методама мултиваријационе анализе које се најчешће користе у статистичким истраживањима. Детаљније су описане мултиваријационе методе које врше редукцију димензије проблема (анализа главних компонената и факторска анализа), као и методе које имају дефинисане процедуре за постепену селекцију и елиминацију променљивих из модела (вишеструка линеарна регресија и дискриминациона анализа).

У трећем поглављу објашњени су основни концепти универијационих и мултиваријационих нестандардних опсервација, као и методе за њихово

откривање. Посебна пажња је посвећена методи Махаланобисовог одстојања. У четвртом поглављу је дат детаљан опис методологије Ивановићевог одстојања. Пето поглавље чини тежиште ове докторске дисертације. У њему су формиране процедуре Ивановићевог одстојања за постепену елиминацију индикатора, постепену селекцију индикатора, као и процедура „корак по корак". Процедуре су детаљно објашњене кроз пример економске развијености земаља Европске уније. Шесто поглавље чине студије случаја, у којима су дефинисане процедуре коришћене у циљу откривања мултиваријационих нестандардних опсервација. Седмо поглавље се односи на закључна разматрања, научни допринос докторске дисертације и правце будућих истраживања.

2. МУЛТИВАРИЈАЦИОНА АНАЛИЗА

Развој информационо-комуникационих технологија је у последњих неколико десетина година омогућио прикупљање и обраду великих количина података. Статистички софтверски пакети који су се симултано развијали, омогућили су спровођење комплетних мултиваријационих статистичких анализа на великим базама података у року од само неколико секунди. Термин мултиваријационе анализе се користи да се представе напредне статистичке методе које се спроводе над великим бројем променљивих и опсервација. Мултиваријациона анализа је веома погодна, чак и неопходна за употребу, када се исте опсервације описују великим бројем променљивих које могу бити међусобно повезане (Radojičić, 2007). Променљиве које се укључују у мултиваријациону анализу могу бити нумеричке (квантитативне) и ненумеричке (квалитативне).

Мултиваријациона анализа се спроводи над сложеним скуповима података, односно када постоји велики број међусобно независних и зависних променљивих које су у корелацији, како би се обезбедиле што свеобухватније статистичке анализе (Jeremić, 2012). За мултиваријациону статистичку анализу, одговарајући скупови података се морају формирати од вредности које одговарају броју променљивих у односу на број ентитета. Они могу бити организовани као матрице података, корелационе матрице, матрице варијанси-коваријанси, матрице суме квадрата, као низ резидуала (Anderson, 1966). Може се рећи и да мултиваријациона анализа представља скуп статистичких метода које симултано анализирају вишедимензиона мерења добијена за сваку јединицу посматрања, односно опсервацију из скупа који посматрамо (Kovačić, 1994).

Претпоставимо да смо током мерења сакупили податке за n опсервација о њихових k својстава. Тако прикупљени подаци представљају основу мултиваријационе анализе, а приказујемо их кроз матрицу података. Сваки ред у њој представља једну опсервацију на којој су измерене вредности свих променљивих, док свака колона представља вредност једне променљиве за све опсервације из скупа (Kovačić, 1994). Тако дефинисана матрица података би имала изглед приказан у наредној табели:

	Пром. 1	Пром. 2	...	Пром. *j*	...	Пром. *k*
Опсервација 1	X_{11}	X_{12}	...	X_{1j}	...	X_{1k}
Опсервација 2	X_{21}	X_{22}	...	X_{2j}	...	X_{2k}
...	...	...		...		...
Опсервација *i*	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ik}
...	...	...		...		...
Опсервација *n*	X_{n1}	X_{n2}	...	X_{nj}	...	X_{nk}

где елемент X_{ij} представља вредност *j*-те променљиве за *i*-ту опсервацију. У матричној нотацији ову променљиву означавамо са X, односно $\left[X_{ij} \right]$ где важи да $i = 1, 2, ..., n;\ j = 1, 2, ..., k$

Избор одговарајуће мултиваријационе методе која ће се спровести над прикупљеном матрицом података зависи од неколико фактора: типова података, врсте проблема који је потребно испитати, карактеристика саме методе и дефинисаних циљева истраживања. Директно закључивање о међузависности неколико променљивих у *k*-димензионом простору са великим бројем опсервација је готово неизводљиво. Како би се то превазишло, користе се методе мултиваријационе статистичке анализе у циљу редуковања димензија посматраног проблема због лакше интерпретације резултата. Изузев овог задатка који је дескриптивне природе, мултиваријациона анализа такође испитује и степен међузависности променљивих, као и њихову статистичку значајност (Bulajić, 2002). Неке од метода мултиваријационе анализе су истраживачког карактера и зато се не користе за тестирање постављених хипотеза, већ за њихово генерисање.

Методе мултиваријационе анализе класификоване су према различитим класификационим критеријумима (Dobrota, M. M., 2018). На основу прве класификације, могуће их је поделити према томе да ли су методе усмерене ка испитивању међузависности променљивих (индикатора) или међузависности јединица посматрања (опсервација). Ако се испитује међузависност променљивих, у анализи се полази од коваријационо-дисперзионе или корелационе матрице. Када су предмет интересовања опсервације, дефинише се

мера одстојања између две опсервације, а затим се анализа започне од одговарајуће матрице одстојања (Đoković, 2013).

Према другој класификацији, мултиваријационе методе се деле у две групе: методе зависности и методе међузависности (Jeremić, 2012). Уколико се испитује зависност између два скупа променљивих, где један скуп представља зависне, а други независне променљиве, тада се говори о методама зависности. Уколико нема теоријског основа за поделу променљивих на два овако дефинисана подскупа (зависних и независних), тада се користе методе међузависности. Методе зависности покушавају да оцене или предвиде једну или више зависних променљивих на основу скупа независних променљивих. Методе међузависности се не користе за предвиђање, већ за редукцију скупа података и његово упрошћавање (Kovačić, 1994). На основу ове поделе, укратко ће бити објашњене неке од метода зависности и метода међузависности.

Методе зависности (Jeremić, 2012):

1. *Мултиваријациона (вишеструка) регресија* – Ово је најпознатија метода мултиваријационе анализе. Разликују се два случаја: анализа зависности једне променљиве од скупа неколико независних променљивих (универијациони регресиони модел) и анализа зависности више променљивих од скупа неколико независних променљивих (мултиваријациони регресиони модел). Код оба случаја задатак је оцењивање или предвиђање средње вредности (средњих вредности) зависне променљиве на бази познатих вредности независних променљивих.

2. *Каноничка корелациона анализа* – Ова мултиваријациона метода се може сматрати уопштеном вишеструком регресионом анализом. Она анализира линеарну зависност између скупа независних и скупа зависних променљивих. Приликом рачунања каноничке корелације, формирају се две линеарне комбинације, једна за скуп независних, а друга за скуп зависних променљивих. Коефицијенти ових линеарних комбинација

одређују се уз услов да коефицијент корелације између њих буде максималан.

3. *Дискриминациона анализа* – Ова анализа бави се проблемом раздвајања група и распоређивањем опсервација у унапред дефинисане групе. Дискриминациона анализа омогућава да се из скупа одабраних независних променљивих идентификује она која је највише допринела раздвајању група. Осим тога, омогућава и предвиђање вероватноће да ће свака од опсервација из посматраног скупа припасти једној од дефинисаних група.

4. *Мултиваријациона анализа варијансе (МАНОВА)* – Ова мултиваријациона метода се користи за испитивање утицаја различитих нивоа једне или више „експерименталних" променљивих на две или више зависних променљивих. Мултиваријациона анализа варијансе представља уопштење једнодимензионе анализе варијансе (АНОВА). Користи се када је могуће испитати утицај сваког од третмана дефинисаних фактора. Њен основни циљ је тестирање хипотезе везано за варијансу ефекта група две или више зависних променљивих.

5. *Логит анализа* – Ова анализа представља мултиваријациону анализу која се користи када је у регресионом моделу зависна променљива дихотомног типа (на пример, променљива *Пол* са модалитетима: мушко-женско). Такав модел се назива регресиони модел са квалитативном зависном променљивом. Зависну променљиву представља Логит функција (логаритам количника вероватноћа да ће дефинисана зависна променљива узети вредност првог или другог модалитета.

Методе међусобне зависности (Kovačić, 1994):

1. *Анализа главних компонената* – Ова мултиваријациона метода се користи за редукцију већег броја оригиналних променљивих на мањи број њихових линераних комбинација које називамо главним компонентама. Циљ методе је да се поменутим линеарним комбинацијама објасни највећи део

варијабилитета оригиналног скупа, смањујући притом димензију посматраног проблема. Анализа главних компонената је сконцентрисана на дијагоналне елементе коваријационо-дисперзионе матрице, односно на варијансе. Главне компоненте су конструисане тако да су међусобно некорелисане, што значи да свака од њих објашњава посебан део варијабилитета оригиналних променљивих. Већина истраживача сматра анализу главних компонената првим кораком у спровођењу факторске анализе.

2. *Факторска анализа* – Ова метода је слична методи главних компонената јер се такође користи за редукцију скупа оригиналних променљивих, али у циљу формирања мањег броја фактора. Поменути фактори се формирају груписањем оригиналних променљивих. За разлику анализе главних компонената, факторска анализа претпоставља формирање модела који дели укупан варијабилитет на заједнички и специфични. Заједнички варијабилитет је онај који свака од променљивих дели са осталима из посматраног скупа, а специфични онај који је посебан за сваку од променљивих. Факторска анализа је, за разлику од анализе главних компонената, сконцентрисана на вандијагоналне елементе коваријационо-дисперзионе матрице, односно на коваријансе.

3. *Анализа груписања* – Ова анализа се такође користи за редукцију оригиналног скупа података. За разлику од претходне две, сконцентрисана на груписање опсервација. Овом анализом се опсервације групишу на основу њихове међусобне сличности. Задатак анализе груписања је подела опсервација у мањи број група, тако да су елементи унутар група међусобно слични, а истовремено различити од елемената из других група. Методе анализе груписања се могу поделити на хијерархијске и нехијерархијске.

4. *Вишедимензионо пропорционално приказивање* – Ова мултиваријациона метода је као и претходна оријентисана на опсервације. Она користи меру сличности, односно разлике између опсервација у циљу њиховог

приказивања у простору. Изведена просторна репрезентација садржи геометријски распоред тачака на мапи, где се свака тачка односи на једну опсервацију. Уколико се за вишедимензионо пропорционално приказивање користи мера блискости добијена на основу квантитативних променљивих, ову методу називамо квантитативном. Ако смо за рачунање мера сличности користили квалитативне променљиве, тада користимо тзв. квалитативно вишедимензионо пропорционално приказивање.

5. *Логлинеарни модели* – Ови модели испитују међусобну зависност квалитативних променљивих које формирају вишедимензиону табелу контигенције. Уколико се једна од променљивих у табели контигенције може сматрати зависном, тада се на основу оцењених логлинеарних модела могу извести већ поменути логит модели.

Осим описаних мултиваријационих статистичких метода, у научним истраживања се користи још много сличних метода, од којих су неке имплементиране у разне статистичке пакете. У овој дисертацији, детаљније ће бити описане мултиваријационе методе које врше редукцију димензије проблема (анализа главних компонената и факторска анализа), као и методе које имају дефинисане алгоритме за постепену селекцију и елиминацију променљивих из модела (вишеструка линеарна регресија и дискриминациона анализа).

2.1 Анализа главних компонената

Анализа главних компонената се користи за редукцију већег броја оригиналних променљивих на мањи број њихових линеарних комбинација које називамо главним компонентама. Темеље ове анализе поставили су Пирсон (1901) је Хотелинг (1933), а временом је постала једна од најчешће коришћених мултиваријационих метода за редукцију података (Vidal et al., 2016). Њен основни циљ је да се формираним линеарним комбинацијама објасни највећи део варијабилитета оригиналног скупа променљивих, смањујући при том димензију посматраног проблема и олакшавајући интерпретацију резултата (Landau &

Everitt, 2004). Главне компоненте су конструисане тако да су међусобно некорелисане, што значи да свака од њих објашњава посебан део варијабилитета оригиналних променљивих.

Анализа главних компонената се спроводи над редундантним подацима, што значи над променљивима које су међусобно корелисане (јер се односе на исти вишедимензиони проблем). На поменутој вези између посматраних променљивих се темељи претпоставка да је могуће смањити димензију посматраног проблема кроз формирање мањег броја њихових линеарних комбинација тј. главних компонената (O'Rourke & Hatcher, 2013). Њихов број једнак је броју оригиналних променљивих, јер је за *p* посматраних непознатих могуће формирати највише *p* њихових линеарно независних комбинација. Међутим, само првих неколико главних компонената обухватају највећи део варијабилитета посматраног скупа података. У складу с тим, постоји више метода које предлажу начин за одабир броја главних компонената (Kovačić, 1994).

Ако претпоставимо да је X *p*-димензиони вектор са одговарајућом коваријационо-дисперзионом матрицом Σ, можемо формирати линеарну комбинацију $Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + ... + \alpha_{1p}X_p = \alpha_1^T X$ где су $\alpha_{11}, \alpha_{12}, ..., \alpha_{1p}$ коефицијенти које је потребно одредити како би се максимизирала варијанса од Y_1. Како је $Var(Y_1) = Var(\alpha_1^T X) = \alpha_1^T \Sigma \alpha_1$ и како се $\alpha_1^T \Sigma \alpha_1$ може произвољно повећавати множењем вектора α_1^T произвољним скаларом, уводи се ограничење да је вектор коефицијената јединичне дужине, тј. да је $\alpha_1^T \alpha_1 = 1$ (Kovačić, 1994).

Проблем максимизације варијансе главне компоненте $\alpha_1^T \Sigma \alpha_1$ уз ограничење $\alpha_1^T \alpha_1 = 1$ се решава коришћењем Лагранжових множитеља, максимизацијом Лагранжове функције $\alpha_1^T \Sigma \alpha_1 - \lambda(\alpha_1^T \alpha_1 - 1)$ где је λ Лагранжов множитељ (Bulajić, 2002). Диференцирањем функције по коефицијентима α_1 и изједначавањем добијеног израза са нулом, добијамо $\Sigma\alpha_1 - \lambda\alpha_1 = 0$ што се након извлачења вектора коефицијената α_1 после заграде своди на $(\Sigma - \lambda I)\alpha_1 = 0$. Како је α_1 ненула вектор, да би се добило нетривијално решење дате једначине, потребно је

да детерминанта $|\Sigma - \lambda I|$ буде једнака нули, што значи да λ мора бити један од карактеристичних корена коваријационо-дисперзионе матрице Σ (Jeremić, 2012).

Множењем претходно дате матричне једначине са α_1^T са леве стране добија се $\alpha_1^T \Sigma \alpha_1 - \lambda \alpha_1^T \alpha_1 = 0$, а како је $\alpha_1^T \alpha_1 = 1$ следи да је $\alpha_1^T \Sigma \alpha_1 = \lambda = Var(Y_1)$. С обзиром на то да је циљ да се максимизира варијабилитет, за λ_1 узећемо највећи карактеристични корен. На основу услова $(\Sigma - \lambda I)\alpha_1 = 0$ следи да је α_1 одговарајући карактеристични вектор придружен карактеристичном корену λ_1. Његовим нормирањем $\alpha_1^T \alpha_1 = 1$ добијамо тражени вектор α_1 (Kovačić, 1994).

Након тога, поступак се понавља како би се одредила следећа линеарна комбинација оригиналних променљивих, уз додатни услов да је коваријанса између прве и друге главне компоненте једнака нули. Ако су сви карактеристични корени матрице Σ међусобно различити $(\lambda_1 > \lambda_2 > ... > \lambda_p > 0)$ тада постоји p различитих линеарних комбинација оригиналних променљивих. Вектори коефицијената $\alpha_1, \alpha_2, ..., \alpha_p$ представљају карактеристичне векторе коваријационо-дисперзионе матрице Σ који су придружени коренима $\lambda_1, \lambda_2, ..., \lambda_p$ (Dobrota, M. P., 2013). Из свега наведеног, може се закључити да су особине главних компоненти следеће:

- $E(Y_j) = 0, \quad j = 1, 2, ..., p$
- $Var(Y_j) = \lambda_j$
- $Cov(Y_i, Y_j) = 0, \quad i \neq j$
- $Var(\lambda_1) \geq Var(\lambda_2) \geq ... \geq Var(\lambda_p) \geq 0$

Из последње особине се може приметити да се као прва и најважнија главна компонента изабере она линеарна комбинација која објашњава највећи део варијабилитета посматраног скупа оригиналних променљивих. Друга главна компонента обухвата максималан део преосталог дела варијабилитета итд. У случају када посматрамо p оригиналних променљивих на основу којих смо (коришћењем матрице Σ) израчунали p различитих карактеристичних корена, могуће је формирати p главних компонената (уз услов да су међусобно

независне). Међутим, у анализи се задржава само првих неколико главних компонената којима је објашњен највећи део варијабилитета оригиналних променљивих.

Прва главна компонента је корелисана са неколико (најчешће са већим бројем) оригиналних променљивих. Друга главна компонента ће бити у јакој корелацији са променљивима које нису испољиле значајну повезаност са првом главном компонентом (O'Rourke & Hatcher, 2013), што је последица особине независности линеарних комбинација. Свака од наредних компоненти се формира уз исте услове: покрива варијабилитет који није објашњен претходним компонентама и линеарно је независна у односу на њих.

Када би се у анализи задржале све формиране главне компоненте, варијабилитет оригиналног скупа података био би у потпуности објашњен. То произлази из особине да је траг коваријационо-дисперзионе матрице Σ (сума свих варијанси оригиналних променљивих) једнак трагу дијагоналне матрице Λ (чији су елементи израчунати карактеристични корени). Међутим, циљ анализе главних компонената је редукција и смањивање димензије посматраног проблема, што значи да ће број задржаних главних компонената бити дефинисан кроз количину варијабилитета оригиналног скупа променљивих који је потребно објаснити.

Релативни допринос сваке главне компоненте у објашњењу укупног варијабилитета може се израчунати кроз однос одговарајућег карактеристичног корена и генерализоване варијансе:

$$\frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}, \quad j = 1, 2, \ldots, p$$

Ако са A означимо матрицу чији су редови карактеристични вектори, онда је коваријационо-дисперзиона матрица $\Sigma = A^T \Lambda A$, односно у развијеном облику:

$$\Sigma = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_p \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \begin{bmatrix} \alpha_1^T \\ \alpha_2^T \\ \vdots \\ \alpha_p^T \end{bmatrix} = \lambda_1 \alpha_1 \alpha_1^T + \lambda_2 \alpha_2 \alpha_2^T + \dots + \lambda_p \alpha_p \alpha_p^T = \sum_{j=1}^{p} \lambda_j \alpha_j \alpha_j^T$$

тако да се допринос *j*-те главне компоненте коваријационо-дисперзионој матрици Σ може израчунати као $\lambda_j \alpha_j \alpha_j^T$ (Kovačić, 1994).

Иако је анализа главних компонената разумљива и лако применљива, поседује одређене мане. Један од највећих недостатака је што није у могућности да одвоји грешку мерења од заједничке варијансе (Pett et al., 2003). Још једна од мана је што се може десити да првих неколико главних компонената не обухвате највећи део варијабилитета. Ако не постоји велика разлика у вредностима израчунатих карактеристичних корена, адекватност њене примене је дискутабилна (Pedhazur & Schmelkin, 1991). Разни аутори скрећу пажњу на осетљивост методе главних компонената на присуство нестандардних опсервација (Naik, 2018; Candes et al., 2011; Lam & Choy, 2019). Још један од проблема у интерпретацији добијених резултата јавља се у случају када су оригиналне променљиве исказане у различитим мерним јединицама. Ако се деси да нека од променљивих има знатно већу варијансу од осталих, доминираће првом главном компонентом без обзира на корелациону структуру података (Radojičić, 2007). Тај проблем се може превазићи на два начина. Први је да се уместо коефицијената линеарне комбинације користе коефицијенти корелације између оригиналних променљивих и главних компонената, а други да се комплетна анализа базира на корелационој, а не коваријационо-дисперзионој матрици (Kovačić, 1994). С обзиром на то да се корелациона матрица може сматрати коваријационом матрицом стандардизованих променљивих, укупан варијабилитет мерен генерализованом ваијансом биће једнак *p,* где је *p* број оригиналних променљивих, димензија корелационе матрице и њен траг (Jeremić, 2012). Коефицијент корелације између *k*-те оригиналне променљиве и *j*-те главне компоненте једнак је $\alpha_{jk}\sqrt{\lambda_j}$.

Важно је скренути пажњу на количину варијабилитета оригиналних променљивих који је објашњен задржатим главним компонентама, јер он показује степен апроксимације варијансе сваке од променљивих појединачно (Dobrota, M. P., 2013). На основу израза ортогоналне декомпозиције коваријационо-дисперзионе матрице $\Sigma = A^T \Lambda A$ варијанса *k*-те променљиве се може израчунати као

$$\sigma_{kk}^2 = \sum_{j=1}^{p} \lambda_j \alpha_{jk}^2, \quad k = 1, 2, ..., p$$

То значи да је допринос сваке главне компоненте варијанси *k*-те променљиве једнак квадрату коефицијента корелације између посматране главне компоненте и те оригиналне променљиве. Допринос свих главних компонената *k*-тој променљивој се израчунава као сума квадрата свих елемената у *k*-том реду корелационе матрице $A\Lambda^{1/2}$ (Kovačić, 1994). Количник добијене суме и одговарајуће варијансе оригиналне променљиве представља пропорцију варијансе те променљиве која је објашњена задржаним главним компонентама (Jeremić, 2012). Ова пропорција се у анализи главних компонената, као и у факторској анализи, назива комуналитет оригиналне променљиве. Коришћењем корелационе уместо коваријационо-дисперзионе матрице, одмах се добија пропорција варијансе оригиналне променљиве објашњене задржаним главним компонентама, јер је стандардизацијом варијанса сведена на јединичну вредност.

2.2 Факторска анализа

Факторска анализа је метод мултиваријационе анализе који се користи за опис међусобне зависности великог броја оригиналних променљивих коришћењем мањег броја латентних променљивих које се називају фактори. Факторска анализа се, као и анализа главних компонената, користи за редукцију димензије проблема кроз посматрање новоформираних променљивих. Често се ове две мултиваријационе методе поистовећују. Неки истраживачи анализу главних компонената сматрају првом фазом спровођења факторске анализе (Kovačić, 1994), односно њеном поткатегоријом (Harris, 2001), или једноставно једном од метода факторске анализе (Pages, 2015).

У анализи главних компонената, линеарне комбинације се одређују уз услов да се обухвати максималан варијабилитет оригиналних променљивих коришћењем минималног броја главних компонената. У факторској анализи, фактори се дефинишу с циљем да се обухвати максимум међукорелације оригиналних променљивих (Reyment & Joreskog, 1996). Из овога се може закључити да је анализа главних компонената оријентисана на коваријационо-дисперзиону, а факторска анализа на корелациону матрицу.

Још једна од разлика између ове две методе је у томе што факторска анализа претпоставља постојање теоријског модела којим се успоставља релација између опсервација и издвојених фактора. Анализа главних компонената посматра укупан варијабилитет оригиналног скупа података, док факторска анализа разлаже тај варијабилитет на заједнички и специфични део (Kovačić, 1994). Заједнички део је онај који свака променљива дели са преосталима, а специфичан онај који је посебан и карактеристичан за сваку променљиву.

Факторска анализа развијена је крајем 19. и почетком 20. века кроз покушаје психолога да истраже феномен интелигенције. Њеним утемељивачем може се сматрати Чарлс Спирман, који је тестирао повезаност коефицијената корелације између различитих тестова и дошао до закључка да је те резултате могуће представити једноставним моделом (Spearman, 1904). Каснија истраживања проширила су првобитни модел факторске анализе, па се *p*-димензиони вектор X (са вектором средина μ и одговарајућом коваријационо-дисперзионом матрицом Σ) може изразити преко скупа од *m* новоформираних променљивих, које ћемо назвати заједничким факторима $(F_1, F_2, ..., F_m)$ и *p* специфичних фактора $(\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)$. Модел факторске анализе у развијеном облику се може записати као:

$$
\begin{aligned}
X_1 - \mu_1 &= \beta_{11}F_1 + \beta_{12}F_2 + ... + \beta_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= \beta_{21}F_1 + \beta_{22}F_2 + ... + \beta_{2m}F_m + \varepsilon_2 \\
&... \\
X_p - \mu_p &= \beta_{p1}F_1 + \beta_{p2}F_2 + ... + \beta_{pm}F_m + \varepsilon_p
\end{aligned}
$$

или у матричној нотацији

$$X_{p\times 1} - \mu_{p\times 1} = B_{p\times m} F_{m\times 1} + \varepsilon_{p\times 1}$$

где је

$$X - \mu = \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix}, \quad F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}, \quad B = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pm} \end{bmatrix}.$$

Елементи матрице B, тј. β_{ij} представљају факторска оптерећења *i*-те промењиве на *j*-ти фактор (Kovačić, 1994). Уз следећа ограничења модела факторске анализе:

1. $E(F) = 0, \quad Cov(F) = E(FF^T) = \Phi$ (где је $\Phi = I$ у ортогоналном моделу),

2. $E(\varepsilon) = 0, \quad Cov(\varepsilon) = E(\varepsilon\varepsilon^T) = \Psi = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix},$

3. $Cov(\varepsilon, F) = E(\varepsilon F^T) = 0$.

матрицу Σ можемо разложити на следећи начин (Jolliffe, 2002):

$$\Sigma = BB^T + \Psi$$

где је Σ коваријационо-дисперзиона матрица оригиналних промењливих, B је матрица факторских оптерећења а Ψ дијагонална матрица специфичних фактора (Bulajić, 2002). Када су фактори међусобно независни (у ортогоналном моделу факторске анализе), елементи матрице B су коефицијенти корелације између оригиналних промењливих и фактора. Варијансе оригиналних промењливих су у

том случају јединичне. На основу датог разлагања коваријационо-дисперзионе матрице, варијанса *i*-те промењиве се може израчунати као:

$$Var(X_i) = \sigma_{ii} = \beta_{i1}^2 + \beta_{i2}^2 + ... + \beta_{im}^2 + \psi_i = 1$$

односно

$$Var(X_i) = \sigma_{ii} = \sum_{i=1}^{m} \beta_{ij}^2 + \psi_i, \quad i = 1, 2, ..., p$$

Део варијансе који је објашњен заједничким факторима назива се *комуналитет* (Jackson, 1991) и израчунава као $h_i^2 = \sum_{i=1}^{m} \beta_{ij}^2$ тј. као сума квадрата *i*–тог реда у матрици факторских оптерећења. *Специфична варијанса* ($\psi_i = 1 - h_i^2$) је варијабилитет који није објашњен заједничким факторима, већ се приписује специфичном фактору. Укупна варијанса оригиналног скупа променљивих (генерализована варијанса) може се разложити на објашњени и необјашњени део:

$$tr(\Sigma) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_{ij}^2 + \sum_{i=1}^{p} \psi_i$$

Ако са h означимо укупан комуналитет, где је

$$h = \sum_{i=1}^{p} h_i^2 = \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_{ij}^2$$

Тада је генерализована варијанса једнака

$$tr(\Sigma) = h + tr(\Psi)$$

што значи да је једнака збиру две компоненте, укупног комуналитета и укупне специфичне варијансе (Kovačić, 1994).

Пропорција укупног комуналитета која се приписује *j*-том заједничком фактору може се израчунати као количник суме квадрата факторских оптерећења *j*-тог фактора (сума квадрата *j*-те колоне) и укупног комуналитета, односно

$$\frac{\sum_{i=1}^{p} \beta_{ij}^2}{h}$$

Проценат укупног варијабилитета који објашњава сваки од фактора представља један од важних елемената при доношењу одлуке о одређивању броја издвојених фактора (Dobrota M. P., 2013). Укупна варијанса објашњена сваким фактором представља сопствену вредност корелационе матрице за дати фактор (Jeremić, 2012). Сопствене вредности сваког фактора се могу израчунати и као сума квадрата факторских оптерећења по колонама матрице B. На основу њих се одређује кумулативни проценат објашњеног варијабилитета за сваки фактор посебно, при чему је важно напоменути да су издвојени фактори уређени по опадајућем редоследу у односу на количину објашњене варијансе.

Постоји неколико критеријума за одређивање броја фактора. Један од њих је свакако проценат варијабилитета који се објашњава издвојеним факторима. Како су фактори уређени у опадајући низ на основу сопствених вредности, задржава се број фактора који је потребан како би кумулативна пропорција објашњеног варијабилитета прешла унапред задату границу (Osborne & Banjanovic, 2016).

Међутим, најпознатији је критеријум јединичног корена, тј. Кајзеров критеријум (Kaiser, 1960; 1970). Кајзер предлаже да се задрже фактори чије су сопствене вредности веће од 1, односно они фактори који доприносе варијабилитету више од сваке променљиве појединачно. Многи истраживачи сматрају да овај критеријум није довољно прецизан (Velicer et al, 2000; Costello & Osborne, 2005). Ако је број променљивих већи од 50, издваја се сувише велики број фактора, а ако је мањи од 20, број издвојених фактора је превише мали (Kovačić, 1994). Из тог разлога се саветује комбиновање овог критеријума са још неким. Познат је и критеријум минималне сопствене вредности, који представља модификацију Кајзеровог критеријума. Критеријум дефинише да је потребно задржати оне факторе чија је сопствена вредност већа од просечног комуналитета (Osborne & Banjanovic, 2016).

За одређивање броја фактора може се користити и Кателов *scree* тест (Cattell, 1966) који се заснива на графичком представљању сопствених вредности фактора. Како су оне сортиране у опадајући низ (Thompson, 2004), специфична варијанса у одређеном тренутку преузима доминацију над заједничком варијансом. На графичком приказу изломљене линије идентификује се тачка прелома, након које она апроксимативно постаје права линија. Редни број фактора који припада тој тачки представља сугерисан број издвојених фактора (Kovačić, 1994). У циљу издвајања оптималног броја фактора, потребно је користити више од једног критеријума.

Матрица факторских оптерећења добијена након фазе екстракције, показује везу између оригиналних променљивих и издвојених фактора. Већина оригиналних променљивих је у значајној корелацији са неколико фактора, што отежава интерпретацију резултата анализе. Због тога се спроводи ротација фактора, која за циљ има генерисање „једноставне структуре", у којој би свака променљива имала висок степен корелације са тачно једним фактором. Овакву структуру није лако постићи, али је пожељно приближити јој се што је могуће више (Bulajić, 2002).

Ротација фактора не утиче на промене вредности комуналитета и специфичних варијанси, али се мења проценат варијабилитета који објашњава сваки од фактора (Jeremić, 2012). Различити типови ротације генеришу различита решења (Radojičić, 2007). Постоје две групе метода ротација, ортогоналне и неортогоналне.

Ортогонална ротација подразумева да се задржава прав угао између факторских оса што значи да фактори остају међусобно независни (Osborne & Banjanovic, 2016). Најчешће коришћене ортогоналне методе ротације су:

- *Varimax* метода, која тежи да максимизира варијансу унутар сваког фактора повећавањем вредности већих факторских оптерећења и смањивањем вредности оних која су мања. На овај начин се фактори упрошћавају кроз идентификацију скупа променљивих које имају високу корелацију с једним, а ниску са осталим факторима.

- *Quartimax* метода, која је сконцентрисана на оригиналне промењљиве и максимизира разлике факторских оптерећења у сваком реду посебно. Циљ је да свака промењљива има високу корелацију са тачно једним фактором.
- *Equamax* метода је комбинација претходне две методе, јер тежи ка јасном одвајању фактора посматрајући обе димензије истовремено – максимизацијом високих факторских оптерећења у свакој колони и издвајањем тачно једне јаке корелације у сваком реду факторске матрице.

За разлику од ортогоналне, неортогонална ротација дозвољава промену угла између факторских оса, чиме се нарушава претпоставка о независности фактора. Методе неортогоналне ротације које се најчешће користе су:

- *Promax* метода, која почиње спровођењем ортогоналне методе, а затим кроз итерације постепено дозвољава корелацију између фактора. Ова метода се спроводи када је број промењљивих и опсервација велики.
- *Direct oblimin* метода, којом се постиже максимизација сопствених вредности, али се може нарушити једноставна структура података.

Истраживачи чешће користе методе ортогоналне ротације, из разлога што се некорелисани фактори лакше интерпретирају. У случају независности фактора, ортогоналне и неортогоналне ротације често генеришу сличне резултате (Osborne & Banjanovic, 2016).

Након издвајања фактора и придруживања сваке оригиналне промењљиве тачно једном фактору, следи фаза интерпретације. Сваком фактору се додељују имена на основу скупа промењљивих које му припадају, водећи притом рачуна о висини и предзнаку факторских оптерећења. Већи значај се придаје оним промењљивима које имају највиши степен корелације са посматраним фактором (Kovačić, 1994).

2.3 Вишеструка линеарна регресија

Вишеструка линеарна регресија се користи за предвиђање вредности зависне променљиве на основу вредности две или више независних (предикторских) променљивих. Нека је Y зависна променљива и нека су $X_1, X_2, ..., X_p$ независне променљиве. Тада је линеарни регресиони модел дат једначином:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

при чему су $\beta_0, \beta_1, \beta_2, ... \beta_p$ непознати параметри које би требало оценити, а ε грешка мерења, тј. резидуали. Уколико имамо k зависних променљивих, модел можемо записати као

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i \qquad i = 1, ..., k$$

У матричном облику, модел вишеструке линеарне регресије се записује на следећи начин (Freund et al., 2006):

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{p1} & \cdots & x_{pk} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

односно

$$Y = X\beta + \varepsilon$$

при чему се $X\beta$ зове системска компонента модела, а ε случајна компонента модела. Претпоставке вишеструког линеарног регресионог модела су (Vuković & Bulajić, 2014):

1. *Линеарност.* Зависна променљива *Y* је функција облика

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon,$$

где је ε случајна променљива.

2. *Константна очекивана вредност.* Случајна променљива ε има очекивану вредност једнаку нули, тј.

$$E\{\varepsilon\} = 0\,.$$

3. *Хомоскедастичност.* За сваку вредност независних променљивих $X_1, X_2, ..., X_k$, варијаса ε је константна и једнака σ^2, тј.

$$Var\{\varepsilon\} = \sigma^2\,.$$

4. *Нормална расподела.* Претпоставља се да случајна променљива ε има Нормалну расподелу

$$\varepsilon : N\left(0; \sigma^2\right).$$

За сваку опсервацију добија се вредност зависне променљиве *Y* и вредности независних променљивих $X_1, X_2, ..., X_k$. Разлике између измерених вредности *Y* и њених очекиваних вредности су резидуали

$$e = Y - \left(\hat{\beta}_0 + \hat{\beta}_1 X_1 \ + \ldots + \ \hat{\beta}_k X_k\right)$$

За оцене регресионих коефицијената користи се метод најмањих квадрата (Montgomery et al., 2012) , док се непознати параметри оцењују тако да укупан збир квадрата буде минималан (Gordon, 2015). Збир квадрата разлике се може означити са:

$$\sum\left[Y - \left(\beta_0 + \beta_1 X_1 \ + \ldots + \ \beta_k X_k\right)\right]^2 = S\left(\beta_0, \beta_1, \ldots, \beta_k\right)$$

где је $S(\beta_0, \beta_1, \ldots, \beta_k)$ функција регресионих коефицијената а $\sum$ означава збир квадрата разлика за све елементе узорка. Изједначавањем парцијалних извода функције *S* по непознатим параметрима модела са нулом, добија се систем од *k*+1 линеарних једначина, сличан систему у простом линеарном регресионом моделу (Vuković & Bulajić, 2014). Овај систем се назива систем нормалних једначина и његово решење представљаће оцене регресионих коефицијената означене са $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

На основу израчунатих статистика регресионих коефицијената може се закључити колики је релативни утицај или важност сваке независне променљиве. У случају стандардизованог облика регресионог модела, регресиони коефицијенти указују на корелацију између независних и зависних променљивих. Вишеструка регресија такође показује колико је јака међузависност зависне променљиве са свим независним променљивима преко коефицијента корелације *R*. Коефицијент детерминације R^2 показује колики је проценат варијабилитета зависне променљиве објашњен варијабилитетом независних променљивих (Cohen et al., 2003).

У случају вишеструке регресије, усложњава се процес тестирања хипотеза, које се осим на цео модел, могу односити и на појединачне параметре. У тестирању хипотеза за појединачне регресионе параметре којима се проверава утицај сваке од независних променљивих посебно, користе се сличне статистике као у простом линеарном регресионом моделу. Оне имају Студентову расподелу са $n - k - 1$ степени слободе (Chatterjee & Simonoff, 2013). Квалитет целог модела тестира се применом *F* теста, где се у однос стављају варијабилитет обухваћен регресијом и резидуални варијабилитет (Vuković & Bulajić, 2014), јер и у вишеструком регресионом моделу такође важи да се укупан варијабилитет може разложити на објашњени (регресиони збир квадрата) и необјашњени (резидуални збир квадрата). За меру успешности објашњења варијабилитета зависне променљиве Y преко промена независних променљивих $X_1, X_2, \ldots, X_k$ користи се коефицијент детерминације R^2. Тестира се хипотеза:

$$H_0(\beta_1 = \beta_2 = \ldots = \beta_k = 0),$$

што у ствари значи да независне промењиве $X_1, X_2, ..., X_k$ не утичу на вредности зависне променљиве. За тестирање ове хипотезе користи се статистика

$$F = \frac{R^2}{1-R^2} \cdot \frac{(n-k-1)}{k}$$

која има Фишерову расподелу са k и $(n-k-1)$ степени слободе, где је n укупан број опсервација, а k број независних променљивих у регресионој једначини (Schroeder et al., 1986). Ако је добијена вредност статистике F из узорка већа од критичне вредности F_0, коју одређујемо из таблица функције Фишерове расподеле, хипотезу H_0 одбацујемо.

Приликом прављења регресионог модела, поставља се питање колико је независних променљивих потребно да би се добро предвидела зависна променљива. Способност предвиђања независних променљивих није могуће утврдити искључиво посматрањем коефицијената корелација између независних и зависне променљиве, већ се у ту сврху користе методе вишеструке линеарне регресије са одговарајућим техникама.

Постоје две основне групе метода селекције независних променљивих (Royston & Sauerbrei, 2008). Првој групи метода припадају оне које испитују свих 2^k могућих модела, па одабирају онај који је оптималан. Када је број независних променљивих велики, ове методе су непрактичне (Seber & Lee, 2002). Другу групу метода чине методе постепене регресије (енг. *stepwise regression*) у којима се одлучује које променљиве и којим редоследом се укључују или искључују из модела.

Постепена регресија

Ако су независне променљиве међусобно некорелисане, процена доприноса регресионом моделу сваке од њих се лако идентификује (Tabachnick & Fidell, 2013). Међутим, у пракси увек постоји одређени степен корелације између предиктора. Постепена регресија омогућава да се елиминишу променљиве које због дуплицитета у информацијама не доприносе (или у малој мери доприносе) тачности у предвиђању модела. На тај начин се променљиве секвенцијално укључују или искључују из регресионог модела. Са k независних променљивих, ове методе ће укључити процену највише $k + 1$ једначина, у поређењу са 2^k неопходних за испитивање свих могућих једначина (Chatterjee & Hadi, 2012).

Можемо разликовати три методе постепене регресије (Yan & Su, 2009): метода за постепену селекцију променљивих (енг. *forward selection*), метода за постепену елиминацију променљивих (енг. *backward elimination*) и модификација методе за постепену селекцију која се назива „корак по корак" (енг. *stepwise*). Све три методе се заснивају на постепеним корацима и математичким одлукама (Field, 2005).

1) Метода за постепену селекцију променљивих

Метода за постепену селекцију променљивих почиње једначином која не садржи предикторске променљиве, већ само константу β_0. Прва независна променљива укључена у једначину је она која има највиши коефицијент корелације са зависном променљивом Y (Sen & Srivastava, 1990). То је уједно и независна променљива са највећом вредношћу поменуте F статистике (Draper & Smith, 1998). Ако први предиктор значајно побољша могућност модела да предвиди зависну променљиву (када је регресиони коефицијент β_1 значајно различит од нуле), он се задржава у једначини и тражи се следећи предиктор. Независна променљива која улази у једначину као други предиктор је она која има највећи коефицијент корелације са зависном променљивом Y, након што је

зависна променљива подешена за ефекат првог предиктора, односно независне променљиве са највишим коефицијентом корелације са резидуалима из првог корака (Chatterjee & Hadi, 2012). Након тога се тестира значајност регресионог коефицијента другог предиктора β_2. Ако је β_2 статистички значајан, прелази се на тражење трећег предиктора на исти итеративан начин.

Поступак се прекида када више нема статистички значајаних регресионих коефицијената, или када су све независне променљиве укључене у регресиони модел. Како се полази од једначине која садржи само константу, а метода се завршава најкасније када је у регресионом моделу свих k независних променљивих, постоји могућност за формирање највише k+1 једначина.

2) Метода за постепену елиминацију променљивих

За разлику од претходне, метода за постепену елиминацију променљивих полази од регресионог модела који садржи све независне променљиве. У првом кораку, елиминише се независна променљива са најмањим доприносом смањења суме квадратне грешке (Chatterjee & Hadi, 2012). Другим речима, елиминише се променљива која има најмању значајност коефицијента корелације са зависном променљивом.

Ако су сви регресиони коефицијенти значајни, ниједна променљива неће бити елиминисана из модела. Под претпоставком да постоји једна или више променљивих које немају статистички значајне регресионе коефицијенте, елиминише се променљива са најмањом вредношћу одговарајућег теста. Једном елиминисана променљива се не може вратити у модел (Sutter & Kalivas, 1993). Након тога се формира једначина са преосталих k променљивих (међу којима је и константа β_0).

Поступак се прекида када су сви регресиони коефицијенти значајни или када су све независне променљиве елиминисане из модела. Метода за постепену елиминацију променљивих такође може генерисати највише k+1 једначина.

3) Метода „корак по корак”

Метода „корак по корак” је комбинација претходне две методе (Weisberg, 2005), при чему се у сваком кораку врши селекција независне променљиве која има статистички најзначајнију корелацију са зависном променљивом. Ако се деси да – у току спровођења методе – парцијална корелација неке од независних променљивих које су већ у регресионом моделу изгуби статистичку значајност (тј. вредност значајности пређе унапред дефинисани ниво), та променљива се елиминише из регресионог модела (Chatterjee & Hadi, 2012). Променљиве које се налазе у моделу у последњем кораку представљају најбољу комбинацију за предвиђање зависне променљиве (Yu et al., 2014).

Уколико се методе спроводе над некорелисаним или слабо корелисаним независним променљивима, резултати избора предиктора све три методе се углавном поклапају (Chatterjee & Hadi, 2012). Препорука за употребу наведених метода је да се превасходно користи метода за постепену елиминацију променљивих. Приликом њеног коришћења, све независне променљиве се укључују у регресиони модел у првом кораку и свака од њих је доступна за анализу, без обзира да ли ће бити укључена у коначном регресионом моделу.

У статистичком пакету СПСС се одређују границе нивоа значајности за селекцију (енг. *Probability of F for Entry*) и за елиминацију (енг. *Probability of F for Removal*) независних променљивих. Поменута два нивоа значајности не морају (и најчешће немају) исту вредност. Углавном је ниво значајности за селекцију предиктора ($p<0.05$) нижи од нивоа за елиминацију ($p>0.1$). На основу израчунате p вредности Фишерове статистике, метода „корак по корак” започиње укључивањем независне променљиве са најмањом p вредности. У следећем кораку се поступак рачунања Фишерове статистике и одговарајуће p вредности понавља за преосталих k–1 независних променљивих. Након тога се у регресиони модел укључује променљива са најмањом p вредности из тренутног скупа независних променљивих.

Селекција и елиминација независних променљивих се спроводи кроз операције „брисања”, које се односе на елиминацију редова и колона корелационе

матрице. Ове операције се користе како би се методом најмањих квадрата оцениле значајности регресионих коефицијената и одговарајућих статистика (Dempster, 1969). Ако је $\tilde{R}$ нова матрица која се добија елиминацијом *k*-тог реда и колоне корелационе матрице R. Елементи матрице $\tilde{R}$ су:

$$\tilde{r}_{kk} = \frac{1}{r_{kk}}$$

$$\tilde{r}_{ik} = \frac{r_{ik}}{r_{kk}}, \quad i \neq k$$

$$\tilde{r}_{kj} = -\frac{r_{kj}}{r_{kk}}, \quad j \neq k$$

и нека је

$$\tilde{r}_{ij} = \frac{r_{ij}r_{kk} - r_{ik}r_{kj}}{r_{kk}}, \quad i \neq k, j \neq k$$

Ако се операције елиминације примене на сваки ред подматрице R_{11} у матрици

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

где R_{11} садржи независне променљиве које се налазе у регресионом моделу у тренутном кораку, онда је

$$\tilde{R} = \begin{bmatrix} R_{11}^{-1} & -R_{11}^{-1}R_{12} \\ R_{21}R_{11}^{-1} & R_{22} - R_{21}R_{11}^{-1}R_{12} \end{bmatrix}$$

Последњи ред подматрице $R_{21}R_{11}^{-1}$ садржи стандардизоване регресионе коефицијенте, а подматрица $R_{22} - R_{21}R_{11}^{-1}R_{12}$ се може користити за израчунавање парцијалних коефицијената корелације независних променљивих које се не налазе у регресионом моделу.

Нека је r_{ij} елемент матрице $\tilde{R}$ којој припадају променљиве X_i и X_j. Променљиве се елиминишу или селектују једна по једна. Променљива X_k је квалификована за селекцију ако се тренутно не налази у моделу и ако важи $r_{kk} \geq t$, где је t ниво толеранције са референтном вредношћу 0.0001. Такође, за сваку променљиву X_j која се тренутно налази у регресионом моделу мора да важи

$$\left(r_{jj} - \frac{r_{jk} r_{ky}}{r_{kk}} \right) t \leq 1$$

Овај услов се поставља како селекција нове променљиве не би редуковала ниво толеранције за променљиве које се већ налазе у моделу.

Вредност статистике F-за-селекцију за променљиву X_k се израчунава као:

$$F - to - enter_k = \frac{(C - p^* - 1) V_k}{r_{yy} - V_k}$$

и има 1 и $C - p^* - 1$ степени слободе, где p^* означава број регресионих коефицијената који се тренутно налазе у моделу, док је

$$V_k = \frac{r_{yk} r_{ky}}{r_{kk}}$$

Вредност статистике F-за-елиминацију за променљиву X_k се израчунава као:

$$F - to - remove_k = \frac{(C - p^*) V_k}{r_{yy} - V_k}$$

и има 1 и $C - p^*$ степени слободе.

Методе постепене регресије нису увек најбоље решење када је потребно елиминисати неке од независних променљивих. Ако је однос између броја опсервација и броја независних променљивих мали, методе постепене регресије могу одабрати погрешне предикторе (Freedman, 1983). Често се дешава да статистички софтверски пакети користе погрешан број степени слободе, што доводи до повећања вероватноће добијања погрешне статистичке значајности регресионих коефицијената. (Thompson, 1995). Упркос томе, методе постепене регресије су најчешће коришћене методе за секвенцијалну елиминацију и селекцију независних променљивих (Smith, 2018).

2.4 Дискриминациона анализа

Дискриминациона анализа је метода мултиваријационе анализе која се бави раздвајањем различитих група и алокацијом опсервација у унапред дефинисане групе. (Kovačić, 1994). Разликују се две врсте дискриминационе анализе, дескриптивна и предиктивна (Habbema & Hermans, 1977). Када су групе унапред дефинисане, циљ дискриминационе анализе може бити да се утврде и објасне разлике између група (дескриптивна дискриминациона анализа), као и предвиђање којој од група ће посматране опсервације припасти (предиктивна дискриминациона анализа) на основу измерених вредности променљивих (Huberty, 1994). Разлика између ова два типа дискриминационе анализе је у томе што је у предиктивној анализи фокус на класификацији опсервација, док је циљ дескриптивне анализе да се опишу разлике између дефинисаних група (Stevens, 1996). Ови циљеви се често преклапају, па се технике анализе за раздвајање између група истовремено могу користити и за класификацију опсервација у такве предефинисане групе (Dobrota M. M., 2018).

Дискриминациона анализа има доста сличности са вишеструком регресијом. Главна разлика између ове две мултиваријационе методе је у томе што је у вишеструкој регресији зависна променљива континуалног типа, док је у дескриптивној анализи категоријска. Математичке методе које се користе у дискриминационој анализи су сличне онима у мултиваријационој анализи варијансе, са битном разликом у типу променљивих. У мултиваријационој

анализи варијансе, фактори су квалитативни, а зависна променљива квантитативна, док је у дискриминационој анализи зависна променљива квалитативна, а независне променљиве квантитативне (Kovačić, 1994). Заједничка карактеристика све три методе је да на основу независних променљивих предвиђају или описују понашање зависне променљиве.

Први корак у дискриминационој анализи је одређивање дескриптивних мера за сваку од предефинисаних група (Radojičić, 2001). У најпростијем случају (зависна променљива са две групе и једна независна променљива), одређује се гранична тачка класификације, помоћу које ће се дефинисати припадност опсервација формираним групама (Dobrota M. M., 2018). У општем случају, формирају се линеарне комбинације независних променљивих $X_1, X_2, ..., X_k$ на основу којих ће се дискриминација између унапред дефинисаних група извршити тако да вероватноћа погрешне класификације опсервације буде минимизирана (Kovačić, 1994). Линеарном комбинацијом независних променљивих за сваку опсервацију одређује се дискриминациони скор, који се након тога трансформише у апостериорну вероватноћу да опсервација потиче из једне од група.

Да би се постигао први циљ дискриминационе анализе (најбоље могуће раздвајање група), потребно је дефинисати критеријум дискриминације. Њега је дефинисао Фишер, формирањем линеарних дискриминационих функција (Fisher, 1936). У случају две групе, формира се једна дискриминациона функција. У случају више група, број дискриминационих функција је најчешће једнак броју категорија зависне променљиве (Huberty & Olejnik, 2006), при чему ће само првих неколико бити статистички значајне. Прва дискриминациона функција се одређује тако што се максимизира релативни однос варијација између и унутар група. Свака наредна функција се формира уз услов максимизације преосталог дела варијација између и унутар група, као и некорелисаности дискриминационих скорова одређених на основу претходних функција (Kovačić, 1994).

Технике дискриминационе анализе такође могу бити коришћене да би се одредило које променљиве доприносе класификацији. Фишерова дискриминациона линеарна функција је дефинисана као:

$$Z = a_1 X_1 + a_2 X_2 + \ldots + a_k X_k$$

где су $a_1, a_2, ..., a_k$ дискриминациони коефицијенти. Они се могу интерпретирати као и регресиони коефицијенти, што значи да показују допринос раздвајању група сваке променљиве појединачно. Стандардизовани дискриминациони коефицијенти се користе за процену значајности сваке од променљивих за класификацију опсервација. Апсолутна вредност коефицијента показује степен релативног доприноса независних променљивих, а његов предзнак показује смер утицаја (Kovačić, 1994). Како дискриминациони коефицијенти могу бити нестабилни, саветује се да се интерпретација дискриминационих функција базира на коефицијентима корелације оригиналних променљивих и дискриминационих функција. Ови коефицијенти се називају дискриминациона оптерећења или структурни коефицијенти корелације.

Поред утврђивања броја одабраних дискриминационих функција, могуће је и испитати које од независних променљивих имају јаку, а које слабу дискриминациону моћ. Такође, може се десити да две или више променљивих пружају исту дискриминациону информацију (Klecka, 1980). У том случају, довољно је задржати само ону променљиву која је најбољи дискриминатор. Селекција и елиминација променљивих се спроводи кроз методе постепене дискриминационе анализе (енг. *stepwise discriminant analysis*).

Постепена дискриминациона анализа

Постепена дискриминациона анализа се спроводи на основу истих концепата који се користе код постепене регресионе анализе, што значи да постоје три методе (Rencher & Christensen, 2012): метода за постепену селекцију променљивих (енг. forward selection), метода за постепену елиминацију променљивих (енг. backward elimination) и модификација методе за постепену селекцију која се назива „корак по корак” (енг. stepwise). Ове методе се најчешће користе у дескриптивној дискриминационој анализи (Whitaker, 1997).

За избор променљивих у дискриминационој анализи користе се следеће статистике:

$$\overline{x}_{ij} = \frac{\left(\sum_{k=1}^{m_j} f_{jk} x_{ijk} \right)}{n_j}$$

као аритметичка средина променљиве *i* за групу *j*, док

$$\overline{x}_{i\bullet} = \frac{\left(\sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} x_{ijk} \right)}{n}$$

представља аритметичку средину *i*-те променљиве (укупна аритметичка средина). Истом аналогијом се израчунава и варијанса:

$$S_{ij}^2 = \frac{\left(\sum_{k=1}^{m_j} f_{jk} x_{ijk}^2 - n_j \overline{x}_{ij}^2 \right)}{n_j - 1}$$

односно

$$S_{i\bullet}^2 = \frac{\left(\sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} x_{ijk}^2 - n \overline{x}_i^2 \right)}{n - 1}$$

где је са g обележен број група, p почетни број променљивих, q број изабраних променљивих, x_{ijk} вредност променљиве i за опсервацију k у групи j, f_{ij} тежински коефицијент за опсервацију k у групи j, m_j број опсервација у групи j, n_j сума тежинских коефицијената (обим узорка коригован за тежинске коефицијенте) у групи j, а n представља укупну суму тежина (обим узорка коригован за тежинске коефицијенте).

За израчунавање статистике *F*, базиране на проблемима анализе варијансе, потребни су нам и резултати подматрице сума квадрата унутар група и укупна сума квадрата. Елементи

$$\omega_{il} = \sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} x_{ijk} x_{ljk} - \sum_{j=1}^{g} \frac{\sum_{k=1}^{m_j} f_{jk} x_{ijk} \sum_{k=1}^{m_j} f_{jk} x_{ljk}}{n_j}$$

представљају суму квадрата унутар група и дефинишу матрицу *W*, а укупну суму квадрата (тј. матрицу *T*) чине елементи

$$t_{il} = \sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} x_{ijk} x_{ljk} - \frac{\sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} x_{ijk} \sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} x_{ljk}}{n}$$

Поред дефинисаних израза потребно је израчунати и коваријансе тј. матрицу коваријанси

$$C = \frac{W}{n-g} \qquad n > g$$

где се одговарајуће коваријансе рачунају као

$$c_{il}^{(j)} = \frac{\sum_{k=1}^{m_j} f_{jk} x_{ijk} x_{ljk} - \overline{x}_{ij} \overline{x}_{lj} n_j}{n_j - 1}$$

Укупна коваријациона матрица се израчунава изразом

$$T' = \frac{T}{n-1}.$$

Вредности Фишерове статистике F_i је

$$F_i = \frac{(t_{ii} - \omega_{ii})(n-g)}{\omega_{ii}(g-1)}$$

са g-1 и n-g степени слободе, а Вилксова ламбда се рачуна као

$$\Lambda_i = \frac{\omega_{ii}}{t_{ii}}$$

са 1, g-1 и n-g степени слободе.

У сваком кораку, за све променљиве се израчунавају вредности одговарајуће F статистике (Dixon, 1973). Ако је за променљиве у моделу вредност F статистике већа од границе F-за-елиминацију, променљива се искључује из модела. Када су у питању променљиве које нису у моделу, вредност њихове F статистике мора бити мања од вредности F-за-селекцију како би биле укључене у модел. У сваком кораку се може укључити или искључити највише једна променљива.

Ако се закључци доносе на основу вероватноћа, онда се од свих променљивих (које су квалификоване за укључивање у модел) укључује она са најмањом p вредношћу. Ако постоје променљиве које су квалификоване за искључивање из модела, искључује се она са највећом p вредношћу.

Током спровођења процедуре „корак по корак”, матрица W се на сваком кораку замењује новом матрицом W^* помоћу симетричних оператора које је дефинисао Демпстер (1969). Ако је првих k променљивих укључено у модел, W се може поделити као:

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

где је W_{ii} матрица димензија *qxq*. На исти начин се трансформише и матрица *T* и према Демпстеровом правилу формира нову матрицу *T**. Ниво толеранције за укључивање тј. искључивање променљивих из модела дефинише се следећим правилом:

$$TOL_i = \left\{ \begin{array}{ll} 0 & \textit{ако је } \omega_{ii} = 0 \\ \dfrac{\omega_{ii}^*}{\omega_{ii}} & \textit{променљива и није укључена, а } \omega_{ii} \neq 0 \\ \dfrac{-1}{\omega_{ii}^* \omega_{ii}} & \textit{променљива и јесте укључена, а } \omega_{ii} \neq 0 \end{array} \right\}$$

Ако је толеранција променљиве мања или једнака од задате границе толеранције (или се њеним укључивањем у анализу смањује толеранција неке друге променљиве у моделу) неће се рачунати одговарајуће вредности статистике *F*-за-селекцију и *F*-за-елиминацију. У супротном, за дефинисање граничних *F* статистика користе се следеће формуле:

F-за-елиминацију:

$$F_i = \frac{(\omega_{ii}^* - t_{ii}^*)(n-q-g+1)}{t_{ii}^*(g-1)}$$

са *g*-1 и *n-q-g*+1 степени слободе, као и

F-за-селекцију:

$$F_i = \frac{(t_{ii}^* - \omega_{ii}^*)(n-q-g)}{\omega_{ii}^*(g-1)}$$

са *g*-1 и *n-q-g* степени слободе.

3. НЕСТАНДАРДНЕ ОПСЕРВАЦИЈЕ

У прикупљеним подацима, често се јављају елементи посматрања који су на неки начин неконзистентни са већином преосталих из посматраног скупа. Такве опсервације се називају нестандардне (енг. *outliers*). Њихово постојање у подацима може изазвати проблеме при спровођењу статистичких анализа (Choi et al., 2018). Нестандардне опсервације могу довести до погрешног израчунавања параметара узорка (Liu et al., 2012), а самим тим и лоше процене параметара популације.

Један од првих корака које је потребно спровести у циљу добијања кохерентне анализе је откривање нестандардних опсервација. Може се десити да тако дефинисане опсервације буду само погрешно унети подаци (Liu & Zumbo, 2012). Због тога је важно идентификовати их пре спровођења анализа (Liu et al., 2004). Када се спроводе вишеструка мерења неке појаве, промене у начину мерења могу довести до појављивања нестандардних опсервација (Pena & Prieto, 2001).

Како би резултати били релевантни, велики број статистичких анализа претпоставља уклањање нестандардних опсервација из прикупљених података пре самог спровођења неке статистичке методе (Barnett & Lewis, 1994). Тачна дефиниција нестандардних опсервација често зависи од претпоставки везаних за структуру прикупљених података, као и од метода за њихово откривање. Може се рећи да се неке дефиниције сматрају довољно општим да би се могле применити на различите типове података и статистичких метода.

Једна од дефиниција претпоставља да је нестандардна опсервација она која толико одступа од осталих, да изазива сумњу како је генерисана неким другачијим механизмом (Hawkins, 1980). Барнет и Луис (1994) дефинишу нестандардну опсервацију као ону чије вредности посматраних променљивих значајно одступају од вредности за преостале опсервације из посматраног скупа. Још једна дефиниција нестандардне опсервације је да она представља опсервацију која је на неки начин неконзистентна са преосталим опсервацијама из посматраног скупа (Kovačić, 1994).

Методе откривања нестандардних опсервација имају широку примену у друштвеним и природним појавама, као што су откривање превара са кредитним картицама (Bhattacharyya et al., 2011), медицинска дијагностика (Podgorelec et al., 2005), упади у рачунарске мреже (Casas et al., 2012), детекција кривичних дела (Grubesic, 2006), предвиђање временске прогнозе (Lu et al., 2007), географски информациони системи (Bakon ey al., 2017), брзина возила у саобраћају (Saha et al., 2016), проток података у реалном времену (Bhushan et al., 2015) итд.

Поменуте методе могу се поделити на универијационе и мултиваријационе. У статистичким истраживањима чешће се користе мултиваријационе методе, јер је већина постављених проблема комплексна и вишедимензиона. Још једна фундаментална подела метода за откривање нестандардних опсервација је на параметарске и непараметарске (Han et al., 2011).

Параметарске методе за своју примену захтевају познату расподелу коришћених променљивих, или барем добро процењене непознате параметре расподеле. Ове методе означавају нестандардне опсервације као оне које одступају од претпостављеног модела расподеле. Оне су често непогодне за коришћење при анализи великог броја променљивих и опсервација, из разлога што је у том случају тешко испунити све предуслове за њихово коришћење.

Унутар класе непараметарских метода, могу се одредити две подкласе. Прву представљају методе откривања законитости у подацима, тј. методе засноване на одстојању. Ове методе се обично базирају на мерама удаљености и могу да раде са великим базама података. Друга подкласа непараметарских метода заснива се на техникама кластеровања, при чему нестандардне опсервације углавном формирају засебан кластер (Gan & Ng, 2017; Iqbal et al., 2017).

3.1 Методе откривања универијационих нестандардних опсервација

Откривање универијационих нестандардних опсервација подразумева коришћење једне променљиве како би се пронашле вредности које на неки начин одступају од највећег дела преосталих вредности. На пример, нестандардна

опсервација може бити било која вредност која одступа за више од три стандардне девијације од израчунате средње вредности (Jackson & Chen, 2004).

Код униваријационих нестандардних опсервација, некарактеристичне вредности заправо представљају екстремне вредности. Униваријационе методе могу бити ефикасне, али је проблем са коришћењем тих метода што су израчуната средња вредност и стандардна девијација изузетно осетљиве на присуство екстремних вредности (Bauder & Khoshgoftaar, 2017). Поставља се и питање расподеле којој подлеже дата променљива.

Методе које се најчешће користе за откривање униваријационих нестандардних опсервација су метод стандардне девијације, Z-скорови и боксплот метод. Ове методе су применљиве у случају када је расподела података симетрична и када је највећи број измерених вредности сконцентрисан око аритметичке средине (нпр. нормална расподела). Ако подаци подлежу нормалној расподели, лако се израчунава вероватноћа постојања екстремних вредности.

1) Метод стандардне девијације

Код метода стандардне девијације, оне вредности које су за више од две (или више од три) стандардне девијације удаљене од аритметичке средине, сматрају се екстремним вредностима. У случају када подаци не подлежу нормалној расподели, израчунавање вероватноће, односно процента нестандардних опсервација се ослања на Чебишевљеву теорему (Bain & Engelhardt, 1992). Иако је ова теорема применљива на било који тип расподеле, ограничена је тако да се концентрише само на онај део података који се сигурно налази у израчунатом интервалу. Ако подаци подлежу нормалној расподели, највећи део измерених вредности је сконцентрисан у околини просечне вредности. На удаљености једне, две и три стандардне девијације од аритметичке средине се налази приближно 68%, 95% и 99,7% измерених вредности, респективно. Као нестандардне опсервације се дефинишу оне вредности које су за више од две (или више од три) стандардне девијације удаљене од аритметичке средине.

2) *Z*-скорови

Још једна од често коришћених метода откривања универијационих нестандардних опсервација је метода *Z*-скорова. Најчешћа примена ове методе је у случају када подаци подлежу нормалној расподели. Измерено обележје се стандардизује преко формуле

$$Z_i = \frac{x_i - \overline{x}}{\sigma}$$

У случају када обележје подлеже нормалној расподели, нестандардним опсервацијама се могу прогласити оне вредности које имају апсолутну вредност већу од 3. Тада се ова метода може поистоветити са методом стандардне девијације. Највећа могућа вредност *Z*-скора зависи од обима узорка (Shiffler, 1988), а израчунава се као $(n-1)/\sqrt{n}$. Како *Z*-скор не може имати вредност већу од 3 у узорцима чији је обим мањи или једнак од 10, ова метода се не може сматрати адекватном за примену на малим узорцима.

Још једна њена мана је што израчуната стандардна девијација може бити превише велика због неколико или, чак, само једне измерене екстремне вредности. Ово може изазвати ефекат маскирања, односно прикривање једне нестандардне опсервације другом, која има израженије одступање.

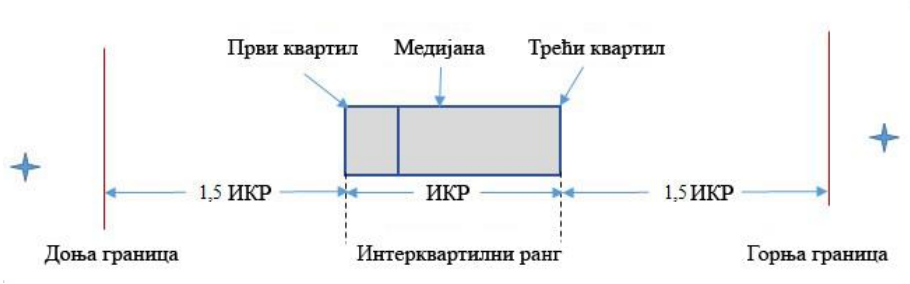
3) Боксплот метода

Најпознатија универијациона графичка метода за откривање нестандардних опсервација је боксплот метода (Tukey, 1977). Она је мање осетљива на екстремне вредности у односу на претходне методе које користе аритметичку средину и стандардну девијацију, јер користи квартиле, који су отпорнији на екстремне вредности. Ова метода дефинише доњу и горњу границу интервала унутар којег се опсервације не третирају као нестандардне.

Први корак је да се пронађе интерквартилни ранг (Dovoedo & Chakraborti, 2015). Он обухвата 50% посматраних опсервација које се налазе између првог и

трећег квартила. Након тога се за доњу границу интервала дефинише вредност која је мања од доњег квартила за једну и по дужину интерквартилног ранга. Горња граница поменутог интервала се формира на сличан начин (као вредност која је већа од горњег квартила за једну и по дужину интерквартилног ранга).

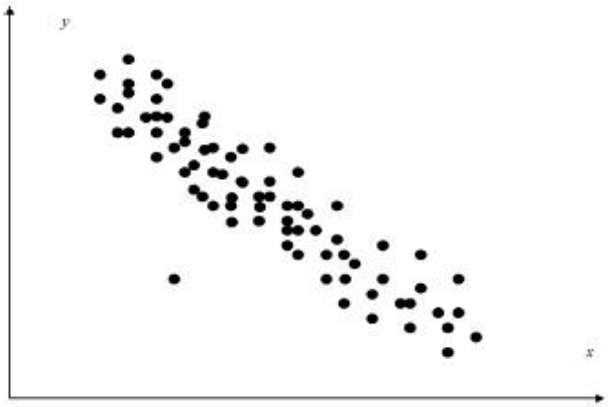
Све опсервације које припадају овако дефинисаном интервалу не третирају се као нестандардне. Преостале опсервације које имају вредности ниже од доње границе (или више од горње границе) сматрамо екстремним вредностима (слика 1).



Слика 1: Боксплот дијаграм

3.2 Методе откривања мултиваријационих нестандардних опсервација

Мултиваријационе нестандардне опсервације је тешко идентификовати посматрањем сваке променљиве појединачно. Њихово откривање је могуће једино спровођењем неке од мултиваријационих метода, испитивањем повезаности посматраних променљивих. Најпростији пример мултиваријационе нестандардне опсервације (у две димензије) се може видети на слици 2.



Слика 2: Мултиваријациона нестандардна опсервација

Опсервација која се налази у доњем левом углу очигледно одступа од осталих, иако ни за једну од посматраних променљивих појединачно нема екстремну вредност. Процес откривања мултиваријационих нестандардних опсервација је компликованији од њихове идентификације у једној димензији (Majewska, 2015), из разлога што постоји сумња да њих не генеришу исти механизми као остале опсервације. Са повећањем броја димензија, компликује се процес откривања нестандардних опсервација (Rocke & Woodruff, 1996). У три или више димензија, јако их је тешко или немогуће идентификовати графички (Hardin & Rocke, 2005), јер могу одступати у више различитих праваца (Leys et al., 2018). Нестандардне опсервације се могу класификовати у три широке категорије: глобалне, контекстуалне и колективне (Han et al., 2011).

Глобалне нестандардне опсервације се најлакше идентификују, јер оне значајно одступају од осталих по највећем броју променљивих које посматрамо. Самим тим, очигледно су издвојене из скупа свих посматраних опсервација. Контекстуалне нестандардне опсервације је теже идентификовати. Њихове измерене вредности променљивих не одступају значајно од осталих, али су условљене ситуацијом коју посматрамо. На пример, температура ваздуха од 28°C није екстремно висока, али ако је та температура измерена крајем октобра, онда је реч о контекстуалној нестандардној опсервацији. Колективне нестандардне опсервације се разликују од претходна два поменута типа јер се јављају у групама, па као скуп опсервација значајно одступају од осталих. Када се појединачно посматрају, ниједна од њих не мора бити нестандардна.

При идентификацији мултиваријационих нестандардних опсервација може доћи до две врсте грешака (Ro et al., 2015): препознавање нестандардних опсервација као стандардних и препознавање стандардних опсервација као нестандардних (Todeschini et al., 2013). Прва се јавља када једна нестандардна опсервација „маскира” другу која се налази близу ње у *k*-димензионом простору, па се друга опсервација у присуству прве сматра стандардном (Sajesh & Srinivasan, 2013). Када би се прва опсервација уклонила из скупа, друга би аутоматски била детектована као нестандардна. Овај ефекат се назива ефектом маскирања (енг. *masking effect*) и често настаје када постоји неколико груписаних

опсервација које повуку израчунату аритметичку средину и варијансу ка себи (Acuna & Rodriguez, 2004).

Друга грешка настаје када се нека опсервација идентификује као нестандардна само у присуству друге, израженије нестандардне опсервације (Thennadil et al., 2018). Елиминисањем друге поменуте опсервације, прва постаје стандардна. Овај ефекат се зове ефекат преплављивања (енг. *swamping effect*) и такође се појављује када неколико удаљенијих опсервација одвуку израчунату аритметичку средину и варијансу од највећег броја опсервација из посматраног скупа (Mavridis & Moustaki, 2008).

Постоји велики број мултиваријационих техника које се користе за идентификацију нестандардних опсервација (Cao et al., 2018). Методе које покушавају да превазиђу ефекат маскирања и ефекат преплављивања су засноване на мерама одстојања. Потенцијалне нестандардне опсервације имају високе вредности измерених одстојања (Filzmoser et al., 2005). У овој дисертацији, описане су методе Махаланобисовог одстојања, Куковог одстојања и метода утицајних вредности.

1) Махаланобисово одстојање

Како би се установиле разлике између дефинисаних група опсервација, потребно је измерити више променљивих односно индикатора по којима се поменуте групе разликују. Индијски математичар и статистичар Махаланобис је предложио меру одстојања на основу које би се могле измерити разлике, али и међусобне везе између људи различите расне и етничке припадности (Mahalanobis, 1930). Од тада, Махаланобисово одстојање као мултиваријациона мера заузима важну улогу у мултиваријационој анализи и проналази примену у различитим статистичким областима. Најчешће се користи као мера одстојања у вишеструкој линеарној регресији, односно као показатељ колико једна опсервација одступа од неког дефинисаног или израчунатог просека.

У анализи груписања, како би се нова опсервација класификовала у неки од већ утврђених кластера, потребно је на првом месту израчунати коваријационо-

дисперзионе матрице за сваки од кластера посебно, на основу опсервација које су већ унутар кластера. Након тога, израчунава се Махаланобисово одстојање нове опсервације од сваког кластера посебно, па се на основу добијених резултата поменута опсервација групише у кластер којем је најближа (тј. у кластер од којег најмање одступа). Махаланобисово одстојање опсервације $\vec{x} = (x_1, x_2, ..., x_n)^T$ од групе опсервација са вектором средина $\vec{\mu} = (\mu_1, \mu_2, ..., \mu_n)^T$ и коваријационо-дисперзионом матрицом Σ се дефинише као

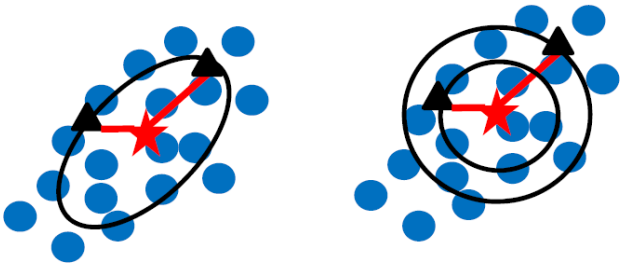
$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})}$$

У пракси се Махаланобисово одстојање често користи када имамо две популације или две групе опсервација, како би се измериле разлике између посматраних група (De Maesschalck et al., 2000). У том случају, израчунава се разлика између два *к*-димензиона вектора $\vec{\mu}_1$ и $\vec{\mu}_2$ и користи се квадратно Махаланобисово одстојање

$$D_M^2 = (\vec{\mu}_1 - \vec{\mu}_2)^T \Sigma^{-1} (\vec{\mu}_1 - \vec{\mu}_2)$$

где Σ^{-1} представља заједничку несингуларну инверзну коваријационо-дисперзиону матрицу. Када би посматране промењљиве биле некорелисане и стандардизоване, матрица Σ би била јединична матрица. У том случају, израчунавање квадратног Махаланобисовог одстојања свело би се на израчунавање квадратног Еуклидског одстојања између вектора $\vec{\mu}_1$ и $\vec{\mu}_2$. Коришћење инверзне коваријационо-дисперзионе матрице дозвољава употребу промењљивих (исказаних у различитим мерним јединицама) које не морају бити међусобно независне (Mahalanobis, 1936). Махаланобисово одстојање је осетљиво на присуство нестандардних опсервација, јер његова вредност зависи од израчунате средине и коваријационо-дисперзионе матрице (Gimenez et al., 2012). Високе вредности овог одстојања могу сугерисати да је у питању нестандардна опсервација (Kosinski, 1999).

Основна идеја Махаланобисовог одстојања је представљена на слици 3 (Brereton & Lloyd, 2016). Претпоставимо да елипса представља границу дводимензионог дијаграма растурања. Свака од тачака на елипси је, хипотетички, подједнако удаљена од центра елипсе. То значи да два црна троугла са слике имају различите вредности Еуклидског одстојања, а једнаке вредности Махаланобисовог одстојања.



Слика 3: Махаланобисово и Еуклидско одстојање

Махаланобисово одстојање има честу примену у откривању правилности у посматраним подацима при спровођењу дискриминационе анализе, када је на основу измерених вредности на узорку нову опсервацију потребно класификовати (доделити, идентификовати или препознати) у неку од две групе са што мањим степеном грешке у датим околностима. Један од таквих примера је спровођење дискриминационе анализе у циљу доношења одлуке о прихватању или одбијању захтева клијента за подизање кредита у банци.

Иако се Махаланобисово одстојање користи у готово свим постојећим статистичким софтверским пакетима, постоје одређене чињенице које је неопходно имати у виду. Претпоставка сваке од мултиваријационих метода која користи континуалне променљиве је да подаци подлежу вишедимензионој нормалној расподели. Неиспуњавање овог услова може значајно утицати на резултате анализе. Такође, због валидности резултата је пожељно да број опсервација буде већи од броја променљивих (Brereton & Lloyd, 2016). Махаланобисово одстојање као мера нема дефинисану ознаку у научној и стручној литератури. Многи аутори у својим радовима квадратно

Махаланобисово одстојање називају „генерализованим" одстојањем. Ово одстојање се често користи због следећих предности (Warren et al., 2011):

- Дефинише нумеричке и графичке границе за идентификацију нестандардних опсервација;
- Изузетно је флексибилно и омогућава коришћење независних избора вредности центроида и коваријационо-дисперзионих матрица;
- Адекватна је техника за умањење утицаја нестандардних опсервација;
- Може открити некарактеристичне законитости у мултиваријационим опсервацијама;
- Представља алтернативу регресионој анализи када не постоје конкретни параметри које је потребно предвидети.

Махаланобисов метод израчунава одстојање опсервације од израчунате средине свих опсервација за све измерене променљиве. Велика вредност одстојања сугерише да посматрана опсервација значајно одступа од осталих у *k*-димензионом простору дефинисаном посматраним променљивима (Stevens, 1984). Основна идеја Махаланобисовог одстојања је да се подаци трансформишу тако да буду стандардизовани и некорелисани. Иако се ово одстојање може израчунати на било ком скупу података, најтачнији резултати се добијају када постоји поклапање расподеле података са вишедимензионом нормалном расподелом (Bauder & Khoshgoftaar, 2017). Ако поменуто поклапање не постоји, израчунате просечне вредности неће бити добар показатељ централног дела података, а варијансе ће неадекватно представљати простирање података кроз *k*-димензиони простор, што може довести до погрешне идентификације нестандардних опсервација. Под претпоставком да подаци подлежу вишедимензионој нормалној расподели, може се доказати да израчунато Махаланобисово одстојање подлеже хи-квадрат расподели, ако је обим узорка довољно велики (Bogl et al., 2017).

Мана Махаланобисовог одстојања је његова осетљивост на ефекат маскирања и ефекат преплављивања (Hadi, 1992). Овај проблем се манифестује тако што се не подразумева да је опсервација са великом вредношћу одстојања сигурно нестандардна. Ефекат маскирања може смањити вредност израчунатог одстојања нестандардне опсервације. То се дешава у случају када мала група

нестандардних опсервација повуче аритметичку средину узорка и израчунат варијабилитет ка себи. С друге стране, ефекат преплављивања може повећати вредност одстојања оних опсервација које нису нестандардне. У том случају група нестандардних опсервација одвуче аритметичку средину узорка и израчунат варијабилитет од највећег броја опсервација које нису нестандардне (Penny & Jolliffe, 2001). Како бисмо утврдили да ли је вредност одстојања за неку опсервацију превише велика, а након тога и да ли је та опсервација нестандардна, вредност измереног одстојања се пореди са 99. перцентилом хи-квадрат расподеле која има *k* степени слободе, где *k* представља број променљивих (Oyeyemi et al., 2015). Ако је вредност Махаланобисовог одстојања већа, онда се опсервација проглашава нестандардном.

2) Куково одстојање

Још једна од метода која се често користи при идентификацији мултиваријационих нестандардних опсервација је Куково одстојање (Cook, 1977). Куково одстојање користи се за процену утицаја појединачних опсервација при спровођењу методе најмањих квадрата у вишеструкој линеарној регресији (Kannan & Manoj, 2015). Куково одстојање за *i*-ту опсервацију се израчунава тако што се та опсервација уклони из анализе, а затим се израчуна за колико су се промениле вредности оцењених регресионих параметара у моделу, преко следеће формуле

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\hat{\sigma}^2}$$

$\hat{Y}_j$ означава предвиђену вредност за опсервацију *j* из потпуног регресионог модела, $\hat{Y}_{j(i)}$ је предвиђена вредност за опсервацију *j* из регресионог модела из којег је уклоњена *i*-та опсервација, *p* је број подешених параметара, а $\hat{\sigma}^2$ средња квадратна грешка. Формула се може записати и као

$$D_i = \frac{e_i^2}{p\hat{\sigma}^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

где је e_i^2 *i*-ти резидуал, а h_{ii} утицајне вредности, које представљају дијагоналне елементе пројекционе матрице $H = X(X^TX)^{-1}X^T$. Куково одстојање се може израчунати и преко формуле

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{-i})^T (X^TX)(\hat{\beta} - \hat{\beta}^{-i})}{p\hat{\sigma}^2}$$

где је $\hat{\beta}$ матрица оцењених параметара регресионог модела, а $\hat{\beta}^{-i}$ матрица оцењених параметара регресионог модела из којег је уклоњена *i*-та опсервација.

Постоји неколико могућности интерпретације вредности Куковог одстојања како би се испитало да ли је нека опсервација нестандардна (Cook & Weisberg, 1982; Cook & Weisberg, 1999; Sarkar et al., 2011; Algur & Biradar, 2017):

- Свака опсервација са вредношћу Куковог одстојања која је више од три пута већа од просека, представља потенцијалну нестандардну опсервацију;
- Вредности Куковог одстојања које су веће од 4/*n*, где *n* представља обим узорка, могу одговарати нестандардним опсервацијама;
- Свака опсервација која има вредност Куковог одстојања већу од 1 утиче на оцењене вредности параметара регресионог модела, па би је требало додатно испитати;
- Још један од начина је коришћење одговарајуће F статистике у идентификацији нестандардних опсервација, где перцентил већи од 50 може бити високо утицајна тачка.

3) Утицајне вредности

Потенцијалне нестандардне опсервације се могу идентификовати израчунавањем одговарајућих утицајних вредности (енг. *leverage values*). Оне се,

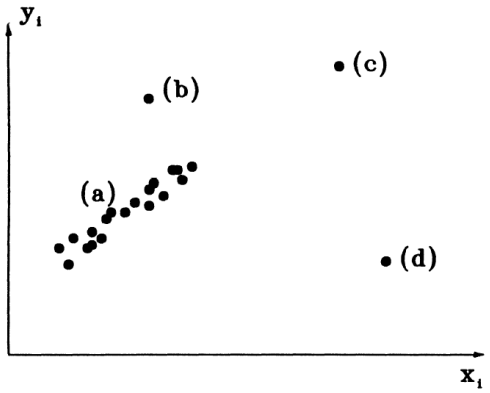
слично као и Куково одстојање, користе за процену утицаја сваке од опсервација у регресионом моделу. Опсервације које имају изузетно високе или ниске вредности променљивих могу се сматрати утицајним тачкама (Sarkar et al., 2011). Оне утичу на израчунавање просечне вредности и варијансе посматраних променљивих.

Утицајне вредности h_{ii} се израчунавају у пројекционој матрици $H = X(X^T X)^{-1} X^T$ као њени дијагонални елементи. Такође, утицајне вредности се могу израчунати коришћењем Махаланобисовог одстојања преко везе

$$h_{ii} = \frac{(D_{M_i})^2}{n-1} + \frac{1}{n}$$

Утицајне вредности је тешко препознати у вишедимензионом простору (Rousseeuw & van Zomeren, 1990). Оне су повезане искључиво са вредностима независних променљивих у регресионом моделу. Можемо их поделити на „добре" и „лоше". „Добре" утицајне вредности су оне које прате правац регресионе праве, тј. не одступају много од регресионе хиперравни у *k*-димензионом простору. „Лоше" утицајне вредности, за разлику од „добрих", негативно утичу на регресиони модел, повлачећи праву или хиперраван ка себи, што доводи до неповољних промена вредности регресионих коефицијената (Cook, 1998). И једне и друге узимају вредности са интервала од 0 до 1, где виша вредност показује већи утицај опсервације на формирање регресионог модела.

Све утицајне вредности које су веће од 3*p*/*n* (где је *p* број оцењених регресионих параметара укључујући и константу, а *n* број опсервација) требало би додатно испитати како би се одредило да ли је посматрана опсервација нестандардна (Kannan & Manoj, 2015). Визуелан приказ „добрих" и „лоших" утицајних вредности може се видети на слици 4.



Слика 4: Добре и лоше утицајне вредности

Тачка (a) је опсервација која се не разликује много од највећег броја опсервација из скупа, тачка (b) је пример нестандардне опсервације, тачка (c) представља добру утицајну вредност, док је тачка (d) пример лоше утицајне вредности.

4. ИВАНОВИЋЕВО ОДСТОЈАЊЕ

Када посматрамо неку комплексну појаву, морамо бити свесни да на њу утиче велики број индикатора, при чему сваки од њих пружа само један део информације о посматраној појави. Ако за пример узмемо социо-економску развијеност земаља, ндикатори као што су бруто национални доходак по глави становника, незапосленост, стопа писмености, број корисника интернета, број лекара на 1000 становника, трошкови за здравство и здравствени систем по глави становника су само неки од индикатора који пружају информације о посматраном феномену. Због тога је јако тешко одредити јединствен глобални индекс који би на један апсолутан начин исказао степен социо-економске развијености једне земље (Ivanović, 1973).

Парцијалне информације које пружају одабрани индикатори се разликују, тако да ће глобална информација о посматраном феномену бити у општем случају потпунија ако посматрамо велики број различитих индикатора. Такође, можемо закључити да неки садрже већу, а неки мању количину информација, па из тог разлога не можемо сваком индикатору придати исти значај (Ivanović, 1977). Ту се појављује проблем пондерисања изабраних индикатора, како би се избегло да они који пружају малу кочичину информација имају велики значај и обратно.

Веома је битно водити рачуна о варијабилитету сваког од посматраних индикатора. Одступање између две опсервације које постоји у односу на један индикатор је значајније уколико је његова варијанса у посматраном скупу елемената мања (Bulajić, 2002). Проблем рангирања индикатора по значају се из тог разлога још више компликује.

Индикатори које анализирамо при посматрању неког феномена су међусобно стохастички зависни. Због тога ће информација о посматраној појави коју пружа један индикатор бити у одређеној мери садржана и у укупној информацији коју пружају остали индикатори (Ivanović, 1956). Како бисмо избегли дуплицитете информација при рачунању глобалног индекса о посматраном феномену, морамо издвојити дискриминациони ефекат сваког индикатора посебно. Информација коју пружа један индикатор не би требало да

буде садржана у информацијама које пружају остали (Ivanović, 1977). Једино на тај начин сваки од њих може пружити јединствен допринос глобалном индексу. Са повећањем броја индикатора, повећава се и ризик да ће доћи до мешања праве и привидне зависности, а тиме и до одстрањивања извесне количине информација које заправо нису биле дуплициране (Ivanović, 1988).

Ако са $X = x_1, x_2, ..., x_k$ означимо скуп изабраних индикатора, а са $P = p_1, p_2, ..., p_n$ скуп опсервација на којима посматрамо одређену комплексну појаву, онда разлика $d_i(r,s) = x_{ir} - x_{is}$ дефинише дискриминациони ефекат обележја X_i (Bulajić, 2002) у уређеном пару опсервација (P_r, P_s). Дискриминациони ефекат скупа обележја X у уређеном пару опсервација (P_r, P_s) је вектор $d_x(r,s) = \langle d_1(r,s), ..., d_k(r,s) \rangle$, док матрица

$$d_x(P) = \begin{bmatrix} 0 & d_x(1,2) & \cdots & d_x(1,n) \\ -d_x(1,2) & 0 & \cdots & d_x(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ -d_x(1,n) & -d_x(2,n) & \cdots & 0 \end{bmatrix}$$

представља ефекат дискриминације од X у P (Jeremić, 2012).

Основни проблем који је потребно решити је како изабрати кључне индикаторе који ће носити велику количину информација о посматраној појави, при чему ће сваки индикатор пружати део информација које нису садржане у осталим индикаторима. Уз то, број индикатора би требало да буде ограничен, што значи да је потребно изабрати одређен број индикатора чија ће укупна информација садржати готово све потребне информације о посматраном феномену (Al Lagili, 2013).

4.1 Дефиниција и особине И-одстојања

Природа неког комплексног проблема не дозвољава да се конструише јединствен глобални индекс који би на један апсолутан начин исказивао степен јачине посматране појаве. Међутим, оно што бисмо могли одредити је релативан положај једне опсервације (из посматраног скупа) у односу на преостале (Ivanović & Fanchette, 1973). Тако долазимо до појма „одстојања" између две опсервације у односу на посматрану комплексну појаву. Осим већ поменутих услова, израчунато одстојање би требало да задовољава и услове које намећу својства одступања у једном математичком простору, као и услове који обезбеђују једнозначност решења (Al Lagili, 2013).

Ако са $D(r,s)$ означимо одстојање између опсервација P_r и P_s. сваку опсервацију можемо представити у виду тачке једног тополошког простора (Al Lagili, 2013). Да би тај простор био метричан, потребно је да одстојање задовољава следеће услове (Ivanović, 1977):

1. *Ненегативност.* Одстојање је ненегативан реалан број, тј.

$$D(r,s) \geq 0 \qquad \text{и} \qquad D(r,r) = 0.$$

2. *Комутативност.* Одстојање између P_r и P_s једнако је одстојању између P_s и P_r, тј.

$$D(r,s) = D(s,r)$$

3. *Триангуларност.* За било које три опсервације P_r, P_s и P_q, мора важити следећа релација

$$D(r,s) + D(s,q) \geq D(r,q)$$

4. *Услов хомогености.* Одстојање између две опсервације је хомогена функција разлика између њихових одговарајућих вредности индикатора. Због тога ће одстојање између две опсервације бити једнако нули ако и само ако важи да је $d_i(r,s) = 0 \quad \forall i = 1,...,k$

5. *Услов раста.* Одстојање је неопадајућа функција свих тих разлика.

6. *Услов варијабилитета* тј. разлике $d_i(r,s)$, $i \in \{1,\ldots,k\}$, требало би да буду тако пондерисане да је њихово учешће у одстојању $D(r,s)$ обрнуто сразмерно стандардној девијацији одговарајућих индикатора X_i, $i \in \{1,\ldots,k\}$. Разлике $d_i(r,s)$ појављиваће се зато у облику $\left|d_i\left(r,s\right)\right| / \sigma_i$ или $d_i^{\,2}(r,s) / \sigma_i^{\,2}$.

7. *Услов анулирања дуплицитета у информацији.* Стохастичка међузависност индикатора може довести до понављања истих информација Из тог разлога, одстојање $D(r,s)$ би требало конструисати тако да та понављања буду искључена, што значи да сваки индикатор пружа јединствен допринос у формирању одстојања.

8. *Услов асиметрије.* Како посматрани индикатори немају исти значај, потребно је одредити њихову ранг-листу на основу количине информација коју пружају о посматраној појави. Одстојање ће се затим конструисати тако да снижавању ранга једног индикатора одговара смањење његовог учешћа у одстојању и то за ону количину информације коју дају индикатори вишег ранга.

9. *Услов независности.* Ако су сви индикатори међусобно независни, дискриминациони ефекти биће дисјунктни, па самим тим неће доћи до понављања истих количина информације. У том случају, одстојање би требало да има облик

$$D(r,s) = \sum_{i=1}^{k} \frac{\left|d_i(r,s)\right|}{\sigma_i} \qquad \text{или} \qquad D^2(r,s) = \sum_{i=1}^{k} \frac{d_i^2(r,s)}{\sigma_i^2}$$

10. *Услов линеарне зависности.* Ако између свих индикатора постоји линеарна функционална зависност, дискриминациони ефекат сваког од индикатора биће садржан у ефекту претходног. У том случају, сви дискриминациони

ефекти би били садржани у ефекту индикатора који је први по значају, па се израз за одстојање своди на

$$D(r,s) = \frac{|d_1(r,s)|}{\sigma_1} \quad \text{или} \quad D^2(r,s) = \frac{d_1^2(r,s)}{\sigma_1^2}$$

11. *Услов независних група.* Ако је једна група од m индикатора независна од преосталих k-m индикатора, потребно је да постоји релација $D_k(r,s) = D_m(r,s) + D_{k-m}(r,s)$. У том случају, одстојање између опсервација P_r и P_s можемо да израчунамо независно, прво на основу првих m индикатора, а затим на основу преосталих k-m. Тражено одстојање, базирано на свих k индикатора, биће тада једнако збиру претходна два.

12. *Независност од почетка.* Увек можемо конструисати две фиктивне опсервације P_+ и P_- чије су одговарајуће вредности индикатора, x_i^+ и x_i^-, произвољно изабране, али тако да за сваку посматрану опсервацију и сваки изабрани индикатор важи релација

$$x_i^- \le x_{ir} \le x_i^+ \qquad i \in \{1,\ldots,k\}$$

Услови 11 и 12 обезбеђују једнозначност решења (Ivanović, 1977).

13. *Технички услов.* Ако је на основу k индикатора израчунато одстојање $D_k(r,s)$, између опсервација P_r и P_s и ако се накнадно дода још један ($k+1$)–ви индикатор, пожељно је да израз за ново одстојање $D_{k+1}(r,s)$ буде једнак збиру из претходног, већ израчунатог одстојања и једне додатне величине која одговара утицају новог индикатора X_{k+1}, што значи да би требало да важи једнакост:

$$D_{k+1} = D_k + E_{k+1}$$

где је E_{k+1} додатак који се односи на нови индикатор. За добијање вредности D_{k+1} довољно је тада израчунати E_{k+1} и томе додати већ познату вредност D_k.

Ако је изабрано k индикатора са следећим редоследом по значају информације коју пружају о посматраној појави: $X =< X_1, X_2, ..., X_k >$ и ако је $P = \{P_1, P_2, ..., P_n\}$ посматрани скуп елемената, располагаћемо следећом табелом:

Индикатор / Опсервација	X_1	X_2	...	X_k
P_1	x_{11}	x_{21}	...	x_{k1}
P_2	x_{12}	x_{22}	...	x_{k2}
...	...	...	...	...
P_n	x_{1n}	x_{2n}	...	x_{kn}

Израчунавање параметара индикатора X_i захтева познавање коефицијената пондерације основних елемента x_{ij}. За различите индикаторе, коефицијенти пондерације не морају бити исти (Bulajić, 2002). Ако са f_i^r означимо релативни коефицијент пондерације од x_{ir}, имаћемо табелу:

Индикатор / Опсервација	X_1	X_2	...	X_k
P_1	f_1^1	f_2^1	...	f_k^1
P_2	f_1^2	f_2^2	...	f_k^2
...	...	...	...	...
P_n	f_1^n	f_2^n	...	f_k^n

Аритметичка средина и варијанса индикатора X_i биће:

$$\bar{X}_i = \sum_{r=1}^{n} f_i^r x_{ir} \qquad \sigma_i^2 = \sum_{r=1}^{n} f_i^r x_{ir} - \bar{X}_i^2 \qquad i = 1,...,k$$

Израчунавање коваријансе w_{ij} захтева познавање дводимензионалних коефицијената пондерације f_{ij}^r у односу на индикаторе X_i и X_j. Како се у пракси ретко када располаже дводимензионалним распоредима $\left[f_{ij}^r \right]$, обично се користе апроксимативне оцене

$$(f_{ij}^r)^* = \frac{\sqrt{f_i^r f_j^r}}{F_{ij}}$$

где је

$$F_{ij} = F_{ji} = \sum_{r=1}^{n} \sqrt{f_i^r f_j^r} \qquad i = 1,...,k \qquad j = 1,...,k$$

Одговарајућа апроксимативна вредност коваријансе биће

$$w_{ij} = \frac{1}{F_{ij}} \sum_{r=1}^{n} \sqrt{f_i^r f_j^r} (x_{ir} - \bar{X}_i)(x_{jr} - \bar{X}_j)$$

а обичног коефицијента корелације

$$r_{ij} = \frac{w_{ij}}{\sigma_i \sigma_i} \qquad i = 1,...,k \qquad j = 1,...,k$$

Преко елемента корелационе матрице

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{12} & 1 & \cdots & r_{2k} \\ \vdots & \cdots & \ddots & \vdots \\ r_{1k} & r_{2k} & \cdots & 1 \end{bmatrix}$$

могу се израчунати парцијални коефицијенти корелације (Ivanović, 1977)

$$r_{ji.t} = \frac{r_{ij} - r_{jt} r_{it}}{\sqrt{(1 - r_{jt}^2)(1 - r_{it}^2)}}$$

за $i > j$ и $i = 1, \ldots, k$ $j = 1, \ldots, k$ $t \notin \{i, j\}$

Итеративним поступком могу се израчунати и следећи парцијални коефицијенти корелације

$$r_{ji.12\ldots j-1} = \frac{r_{ij12\ldots j-2} - r_{j-1,i.12\ldots j-2} r_{j-1,j.12\ldots j-2}}{\sqrt{(1 - r_{j-1,i.12\ldots j-2}^2)(1 - r_{j-1,j.12\ldots j-2}^2)}}$$

па се на тај начин формира матрица парцијалних корелација

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{12} & 1 & r_{23.1} & \cdots & r_{2k.1} \\ r_{13} & r_{23.1} & 1 & \cdots & r_{3k.12} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ r_{1k} & r_{2k.1} & r_{3k.12} & \cdots & 1 \end{bmatrix}$$

Према типу података и одстојања по појединачним индикаторима, разликују се три врсте И-одстојања (Đoković, 2013):

- Обично И-одстојање
- Квадратно И-одстојање
- Структурно И-одстојање

4.2 Обично И-одстојање

За изабрани скуп индикатора $X =< X_1, X_2, ..., X_k >$ рангираних према значајности информације коју пружају о посматраној комплексној појави, И-одстојање између опсервација P_r и P_s тога скупа дефинише се изразом

$$D(r,s) = \sum_{i=1}^{k} \frac{|d_i(r,s)|}{\sigma_i} \prod_{j=1}^{i-1} (1 - r_{ji.12...j-1})$$

где је $d_i(r,s)$ одстојање између вредности индикатора X_i за опсервације P_r и P_s, тј. $d_i(r,s) = x_{ir} - x_{is}, i \in \{1,...,k\}$, σ_i стандардна девијација од X_i , а $r_{ji.12...j-1}$ коефицијент парцијалне корелације између X_i и X_j, *(j < i).*

Може се приметити да је конструкција И-одстојања поступна. Почиње се са интеграцијом целокупног дискриминационог ефекта индикатора X_1, тј. оног који садржи највећу количину информације о посматраној појави. Након тога се додаје онај део дискриминационог ефекта другог (по рангу) индикатора који није био већ укључен у дискриминационом ефекту првог, па затим онај део дискриминационог ефекта трећег индикатора који није био већ укључен у дискриминационом ефекту прва два итд. Овако дефинисано И-одстојање задовољава свих 13 наведених услова (Ivanović, 1975).

4.3 Квадратно И-одстојање

Квадратно И-одстојање дато је формулом

$$D^2(r,s) = \sum_{i=1}^{k} \frac{d_i^2(r,s)}{\sigma_i^2} \prod_{j=1}^{i-1} (1 - r_{ji.12...j-1}^2)$$

Квадратно И-одстојање није једнако квадрату обичног И-одстојања. Како је $r^2 \leq |r|$, то је учешће мање значајних индикатора јаче у квадратном него у

обичном И-одстојању. Учешће посматраних индикатора у И-одстојању опада са њиховим рангом, па у пракси постоји тенденција да се користи квадратно И-одстојање, ако је велики број изабраних индикатора. Често се дешава и да између посматраних индикатора постоје негативни коефицијенти корелације и негативни парцијални коефицијенти корелације, што је још један од разлога за употребу квадратног уместо обичног И-одстојања.

4.4 Структурно И-одстојање

Појам сличности између две опсервације или две структуре може бити уско везан за класификацију по којој су те опсервације класификоване, односно за класификацију по којој су изведене посматране структуре (Ivanović, 1978). Конструкција структурног И-одстојања полази од особина обичног И-одстојања између две опсервације P_r и P_s:

$$D(r,s)=\sum_{i=1}^{k}\frac{\left|\bar{X}_{ri}-\bar{X}_{si}\right|}{\sigma_i}\prod_{j=1}^{i-1}(1-r_{ji.12\ldots j-1})$$

Уместо апсолутних вредности разлика аритметичких средина индикатора X_i опсервација P_r и P_s, користи се збир апсолутних вредности разлика одговарајућих фреквенција (Ivanović, 1977)

$$\left|\bar{X}_{ri}-\bar{X}_{si}\right|\approx\sum_{l=1}^{m}\left|f_{ril}-f_{sil}\right|$$

који представља одстојање између структура P_r и P_s за индикатор X_i у односу на класификацију K_i.

Ако су све разлике једнаке нули, израчунато одстојање ће такође бити једнако нули, што значи да ће обе структуре бити идентичне у односу на класификацију K_i (Radojičić, 2007). Ако је једна од разлика различита од нуле,

тада постоји барем још једна разлика такође различита од нуле, али супротног знака, јер је

$$\sum_{l=1}^{m} f_{ril} = \sum_{l=1}^{m} f_{sil} = 1$$

Уместо обичне стандардне девијације σ_i која се појављује у обичном И-одстојању, сада се појављује стандардна девијација структуре у односу на X_i (Ivanović, 1962), тј. варијанса структуре индикатора X_i која је дата детерминантом

$$S_i^2 = \begin{vmatrix} \sum_{r=1}^{n} f_{ri1}^2 & \sum_{r=1}^{n} f_{ri1} f_{ri2} & \cdots & \sum_{r=1}^{n} f_{ri1} f_{rim} \\ \sum_{r=1}^{n} f_{ri2} f_{ri1} & \sum_{r=1}^{n} f_{ri2}^2 & \cdots & \sum_{r=1}^{n} f_{ri2} f_{rim} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{r=1}^{n} f_{rim} f_{ri1} & \sum_{r=1}^{n} f_{rim} f_{ri2} & \cdots & \sum_{r=1}^{n} f_{rim}^2 \end{vmatrix}$$

Уместо парцијалног коефицијента корелације $r_{ji.12\ldots j-1}$ требало би користити одговарајући парцијални колективни коефицијент корелације $R_{ji.12\ldots j-1}$. Међутим, због комплексног израчунавања овог коефицијента, користи се апсолутна вредност обичног колективног коефицијента корелације R_{ij} (Ivanović, 1977), који је дефинисан са

$$R_{ij}^2 = \frac{Q_{ij}^2}{S_i^2 S_j^2}$$

Тако се долази до структурног И-одстојања између опсервација P_r и P_s за индикаторе $X_1, \ldots, X_k$ у односу на респективне класификације $K_1, \ldots, K_k$ које се исказују изразом

$$D(r,s)=\sum_{i=1}^{k}\sum_{l=1}^{m}\left|f_{ril}-f_{sil}\right|\frac{1}{S_i}\prod_{j=1}^{i-1}(1-\left|R_{ij}\right|)$$

Код структурног И-одстојања упоређују се структуре, а не вредности индикатора X_i (Bulajić, 2002). Због тога је логично да се пореде фреквенције оних класа које садрже средње вредности индикатора X_i, а затим поредити одговарајуће леве, односно десне фреквенције (Ivanović, 1977).

4.5 Редоследна класификација и И-одстојање

Метода И-одстојања омогућава да се формира ранг-листа посматраних опсервација, тако што се фиксира та једна опсервација која ће представљати реперну тачку. За њене вредности индикатора се најчешће дефинишу минималне вредности индикатора из посматраног скупа опсервација. Вредност обележја фиктивне опсервације P_- дефинисана је са

$$x_i^- = \min\{x_{ir}\}, 1\le r\le n \qquad i\in\{1,...,k\}$$

Овако дефинисаа опсервација P_- била би најслабији елемент у скупу P. И-одстојање између опсервације P_r и фиктивне најслабије опсервације P_- дефинише релативан степен јачине P_r (Radojičić, 2007). Образац И-одстојања се своди на

$$D_r^- = \sum_{i=1}^{k}\frac{x_{ir}-x_i^-}{\sigma_i}\prod_{j=1}^{i-1}\left(1-r_{ji.12...j-1}\right)$$

Ово одстојање се може одредити за сваку опсервацију из посматраног скупа. Након тога, све опсервације се могу рангирати на основу вредности тако добијених И-одстојања према степену јачине посматране појаве (Ivanović, 1977). Тада ће опсервације са вишим ранговима припадати групи најбољих, а са нижим групи најслабијих опсервација (Ivanović, 1981). Сличност између две опсервације биће већа уколико је разлика између њихових одстојања мања (Ivanović, 1972b).

За базну опсервацију може се узети и фиктивна најјача опсервација P_+ унутар посматраног скупа

$$x_i^+ = \max\left\{x_{ir}\right\}, i \le r \le n \qquad i \in \left\{1,...,k\right\}$$

Одговарајуће И-одстојање опсервације P_r тада би било

$$D_r^+ = \sum_{i=1}^{k} \frac{\left|x_i^+ - x_{ir}\right|}{\sigma_i} \prod_{j=1}^{i-1} \left(1 - r_{ji.12...j-1}\right)$$

Ако поново сортирамо све опсервације према величини њиховог И-одстојања од фиктивне опсервације P_+, добијени редослед биће инверзан претходном редоследу. То значи да ће резултати бити идентични, тако да је ирелевантно да ли се за фиктивну опсервацију поставља P_- или P_+. Ову једнозначност решења обезбеђује услов 12 (Ivanović, 1977).

5. СЕКВЕНЦИЈАЛНО ИВАНОВИЋЕВО ОДСТОЈАЊЕ

У свакој мултиваријационој статистичкој анализи користи се велики број променљивих. Методе зависности и међузависности покушавају да на основу великог броја оригиналних променљивих сазнају неку нову информацију о комплексном проблему који се посматра. Свака од оригиналних променљивих пружа јединствен допринос решавању мултидимензионог проблема и његовој интерпретацији. Анализа главних компонената и факторска анализа врше редукцију оригиналног скупа променљивих кроз формирање мањег броја линеарних комбинација или фактора. Вишеструка регресија и дискриминациона анализа пружају могућност испитивања значајности сваке од променљивих, а након тога њихову елиминацију или селекцију кроз дефинисане секвенцијалне процедуре.

Метода Ивановићевог одстојања, која је детаљно објашњена у претходном поглављу, има способност синтетизовања великог броја променљивих у само једну зависну променљиву, која мери интензитет посматраног комплексног проблема. Ивановић је ову методу креирао с циљем мерења социо-економске развијености земаља. Велики број индикатора је агрегиран у јединствену вредност, за сваку земљу посебно. Тај процес се врши постепено (Ivanović, 1975), кроз *k* корака (где *k* представља укупан број индикатора):

- израчунава се вредност дискриминационог ефекта првог индикатора X_1 (најважнији индикатор, онај који пружа највећу количину информација о посматраном проблему),
- израчунава се и додаје вредност дискриминационог ефекта другог индикатора X_2 који није покривен дискриминационим ефектом индикатора X_1,
- израчунава се и додаје вредност дискриминационог ефекта трећег индикатора X_3 који није покривен дискриминационим ефектима индикатора X_1 и X_2,
- овај поступак се понавља за све индикаторе.

Овако формирано Ивановићево одстојање задовољава поменутих 13 услова за дефинисање мере одстојања. Када је у питању редослед укључивања индикатора у процес израчунавања Ивановићевог одстојања, он мора одговарати количини информација коју пружа сваки од индикатора, од највеће до најмање. Идеја Ивановићевог одстојања је да се кроз итеративни поступак укључивања дискриминационог ефекта сваког од индикатора посебно, избегне дуплицитет информација, односно да сваки посматрани индикатор има јединствен допринос креирању поменутог глобалног индекса. На тај начин се пружа објективан приступ процесу рангирања.

До коначног редоследа укључивања индикатора се долази итеративним путем. Идеја оптималног редоследа почива на корелацијама између коришћених индикатора и израчунатог Ивановићевог одстојања. На крају сваке фазе, израчунају се коефицијенти корелације између испитиваних индикатора и добијеног одстојања (Ivanović, 1979). Након тога, индикатори се у наредну фазу укључују на основу јачине израчунате корелације, од најмање до највеће. Својим узајамним односима, индикатори коригују редослед из претходне фазе, тежећи дефинитивном, оптималном редоследу (Ivanović, 1982). Суштина методе је у формирању редоследа у којем су индикатори рангирани на основу свог значаја у односу на посматрани комплексни проблем. У циљу мерења интензитета неке појаве, потребно да сви индикатори буду истосмерно оријентисани са израчунатим Ивановићевим одстојањем. Ако је неки индикатор разносмерно оријентисан, потребно је извршити његову линеарну трансформацију, што се најчешће ради одабиром његовог комплементарног индикатора (Ivanović, 1969).

За k почетних индикатора постоји $k!$ могућих редоследа индикатора и k решења методе Ивановићевог одстојања. Јеремић (2012) је предложио почетно решење, верификовано кроз бројне научне радове, које се заснива на идеји да индикатор који најбоље корелира са осталим индикаторима буде први уврштен у образац за израчунавање Ивановићевог одстојања. Посебно се истакло решење које се добило применом унапређења почетног редоследа индикатора. То је решење у којем је сума корелација индикатора са вредношћу Ивановићевог одстојања највећа (Jeremić, 2012). Овако добијено решење је у складу са

принципима генерализоване варијансе и најбоље покрива варијабилитет почетног скупа индикатора. На овај начин, значајно је смањен број итерација методе.

Примене методе Ивановићевог одстојања су бројне. Коришћена је при рангирању најбољих светских универзитета (Jeremić et al., 2011a; Jovanović et al., 2012; Dobrota M. P. et al., 2016; Dobrota M. M. & Dobrota M. P., 2016), мерењу индекса људског развоја (Jeremić et al., 2011b), економске развијености земаља (Jednak et al., 2018; Milenković et al., 2016a; Milenković et al., 2016b) одрживог развоја (Radojičić et al., 2012), ефикасности образовања у основним школама (Milenković et al., 2013), испитивању ИКТ инфраструктуре (Jeremić et al., 2011c; Jeremić et al., 2011d; Dobrota M. P. et al., 2012; Dobrota M. P. et al., 2015) мерењу развијености здравствених система земаља (Al-Lagilli et al., 2011; Jeremić et al., 2011e; Jeremić et al., 2012; Seke et al., 2013), мерењу ефикасности банака (Bulajić et al., 2013) итд.

Ивановићево одстојање је такође коришћено у мерењу социо-економске развијености земаља Блиског Истока и Северне Африке (Milenković et al., 2014). У овом истраживању, коришћено је 19 индикатора развијености који су били подељени у четири категорије (економски, социјални, индикатори развијености здравствених система и ИКТ индикатори). Идеја је, осим мерења развијености, била такође и редукција димензије проблема кроз смањење броја индикатора. Циљ је био одређивање оптималног броја индикатора, јер је с једне стране важно ограничити њихов број, а с друге стране пружити што већу информацију о посматраном проблему (Ivanović, 1972a).

У поменутом истраживању, коришћени су коефицијенти корелације сваког од индикатора са израчунатим Ивановићевим одстојањем, као и њихове статистичке значајности. Пошло се од претпоставке да коефицијент корелације сваког од индикатора пружа информацију о томе колико је индикатор важан у испитивању посматраног комплексног проблема. Већа вредност коефицијента корелације са израчунатим Ивановићевим одстојањем означава бољу повезаност индикатора, па самим тим и његову већу дискриминациону моћ и количину информација о социо-економској развијености земаља. Основна идеја овог рада била је у елиминацији оних индикатора чији је коефицијент корелације са

израчунатим Ивановићевим одстојањем низак и није статистички значајан. Један од два разлога објашњава ову појаву: или посматрани индикатор није релевантан за испитивање посматраног проблема, или је његов дискриминациони ефекат већ обухваћен претходним индикаторима. У оба случаја, такав индикатор би требало искључити из анализе (Milenković et al., 2014).

У мултиваријационим статистичним методама које омогућавају секвенцијалну селекцију или елиминацију променљивих, тај процес се одвија итеративно, тако што се у сваком кораку у модел убацује и/или избацује једна променљива. Идеја ове дисертације је у постепеној селекцији и/или елиминацији променљивих на основу информације о количини варијабилитета који пружа сваки од посматраних индикатора. У те сврхе, коришћен је коефицијент детерминације R^2 који је добијен као квадрат коефицијента корелације сваког од индикатора са израчунатим Ивановићевим одстојањем. Одговарајућа статистика за тестирање значајности коефицијента детерминације је

$$F = \frac{R^2}{1-R^2} \cdot \frac{(n-k-1)}{k}$$

која има Фишерову расподелу са k и $n-k-1$ степени слободе (Chatterjee & Hadi, 2012). Након тога, израчуната је статистичка значајност свих добијених вредности F статистике коришћењем Фишерове функције густине и функције расподеле за задат број степени слободе. Методологија Ивановићевог одстојања је унапређена дефинисањем процедуре за постепену селекцију индикатора, постепену елиминацију индикатора, као и процедуре „корак по корак". Вредности које су коришћене за постепену селекцију и/или елиминацију индикатора заснивају се на статистичкој значајности израчунате F статистике за сваки од индикатора.

Границе статистичке значајности за селекцију и елиминацију индикатора дефинисане су у складу са нивоима значајности дате F статистике. Како би индикатор био укључен у анализу, потребно је да статистичка значајност одговарајуће F статистике буде мања од 0.05 (p<0.05). За искључивање индикатора из анализе, ова граница је нешто већа. Ако је статистичка значајност

одговарајуће F статистике већа од 0.1 (p>0.1), индикатор се искључује из даље анализе (IBM Corporation, 2013).

Како би се спровеле овако дефинисане процедуре, прикупљени су подаци о економској развијености за 28 земаља Европске уније. Коришћено је следећих седам индикатора: 1. Бруто Домаћи Производ (БДП) по глави становника (у доларима); 2. Домаћи кредити у приватном сектору (у процентима БДП-а); 3. Извоз добара и услуга (у процентима БДП-а); 4. Стране директне инвестиције, нето одлив у процентима БДП-а); 5. Раст БДП-а (годишњи проценат); 6. Инфлација, потрошачке цене (годишњи проценат); 7. Незапосленост (проценат укупног радно способног становништва).

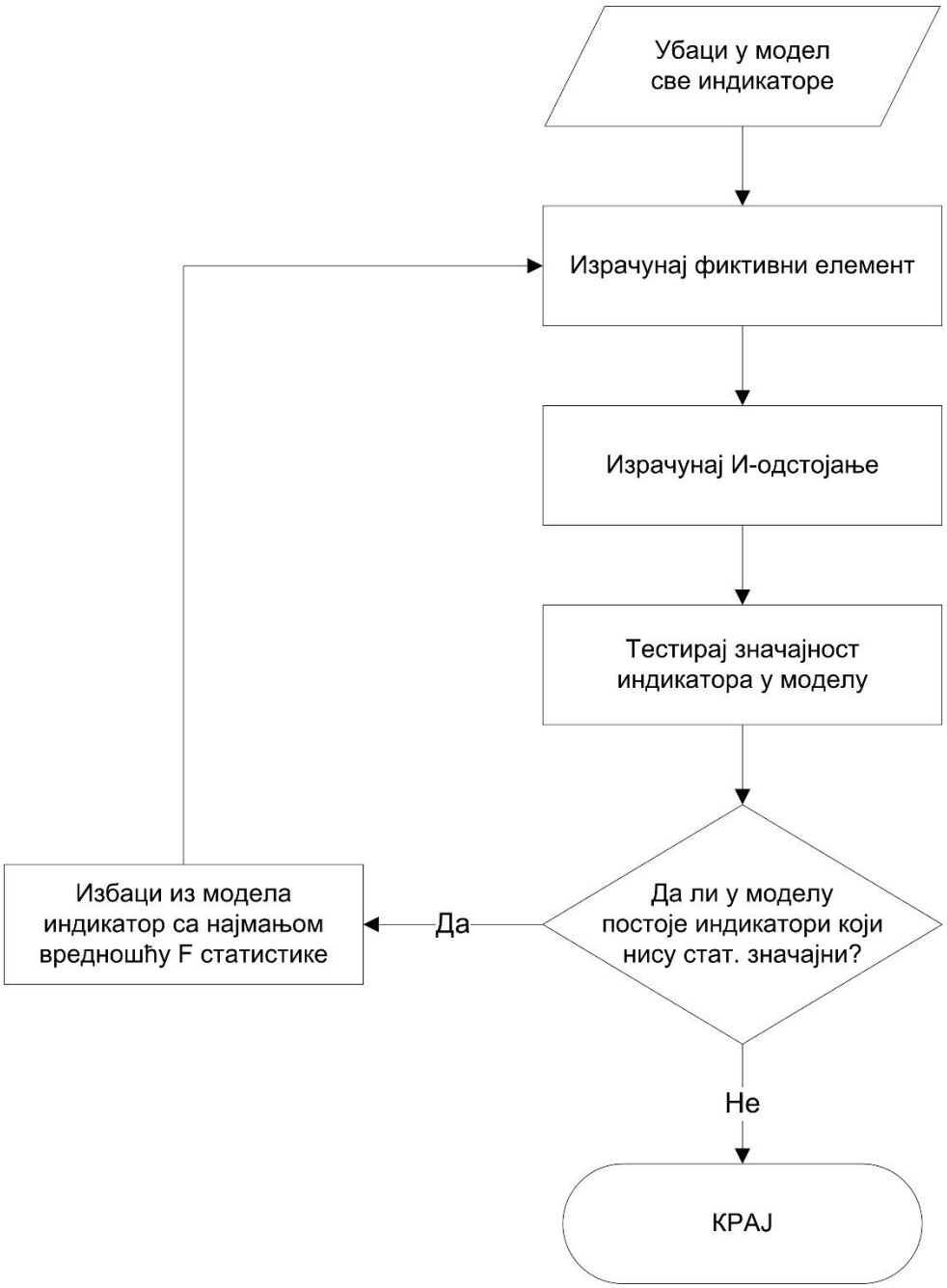
Ови индикатори су званично предложени од стране Светске банке као индикатори економског развоја земаља. Подаци за 2016. годину (последња година за коју су постојале вредности свих индикатора за све земље), преузети су са веб сајта Светске банке. Индикатори „Инфлација" и „Незапосленост" су линеарно трансформисани, како би били позитивно оријентисани са израчунатом вредношћу Ивановићевог одстојања.

Након спровођења сваке од три поменуте процедуре, дошло се до коначних резултата о значајности индикатора, а такође и економској развијености земаља Европске уније.

5.1 Процедура Ивановићевог одстојања за постепену елиминацију индикатора

Процедура за постепену елиминацију индикатора подразумева да се у првом кораку у анализу укључе сви индикатори. За мерење економске развијености земаља Европске уније у првом кораку користиће се поменутих седам индикатора. Циљ ове процедуре је редукција димензије проблема кроз постепено искључивање индикатора који буду идентификовани као статистички незначајни. У сваком кораку се искључује по један индикатор. Процедура се понавља све док у моделу не остану само они индикатори који имају статистички значајну вредност F статистике. Почетан редослед индикатора је дефинисан

сумом апсолутних корелација (Jeremić, 2012) и индикатори су укључени у процедуру почевши од оног са највећом сумом, до последњег индикатора са најмањом сумом, респективно. Алгоритам процедуре за постепену елиминацију дат је на следећој слици:



Слика 5: Алгоритам процедуре за постепену елиминацију индикатора

Редослед укључивања индикатора је представљен у табели 1.

Табела 1: Почетни редослед индикатора

РБ	*Индикатор*	*Сума корелација*
1.	БДП	3.539
2.	Инфлација	3.059
4.	Извоз добара и услуга	2.802
5.	Незапосленост	2.483
6.	Домаћи кредити у приватном сектору	2.450
7.	Стране директне инвестиције	2.350
8.	Раст БДП-а	2.312

Процедура је започета на основу добијеног почетног редоследа. Кроз две итерације израчунавања Ивановићевог одстојања, редослед коришћених индикатора се поновио, што значи да смо добили коначно решење у првом кораку процедуре за постепену елиминацију индикатора, које је представљено у табели 2.

Табела 2: Значајност индикатора у првом кораку – седам индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Извоз добара и услуга	0.800	0.640	5.079	0.00193
2.	БДП	0.747	0.558	3.607	0.01124
3.	Стране директне инвестиције	0.690	0.476	2.596	0.04443
4.	Инфлација	0.530	0.281	1.116	0.39147
5.	Незапосленост	0.178	0.032	0.093	0.99816
6.	Домаћи кредити у приватном сектору	0.142	0.020	0.059	0.99959
7.	Раст БДП-а	0.059	0.003	0.010	0.99999

Редослед земаља Европске уније на основу економске развијености коришћењем свих седам индикатора представљен је у табели 3.

Табела 3: Редослед земаља Европске уније – седам индикатора

РБ	*Земља*	*И-одстојање*
1	Луксембург	82.81
2	Италија	42.53
3	Мађарска	37.81
4	Кипар	35.85
5	Холандија	34.79
6	Данска	29.66
7	Словачка	25.39
8	Румунија	23.37
9	Велика Британија	20.70
10	Чешка	20.69
11	Бугарска	20.67
12	Пољска	20.56
13	Шведска	20.10
14	Литванија	20.02
15	Немачка	19.89
16	Хрватска	19.60
17	Малта	19.41
18	Финска	19.32
19	Летонија	19.25
20	Словенија	19.00
21	Грчка	18.79
22	Француска	18.42
23	Белгија	18.24
24	Шпанија	17.91
25	Естонија	17.32
26	Аустрија	16.94
27	Португал	15.34
28	Италија	14.19

На основу резултата из табеле 2 можемо видети да је први индикатор који ће бити елиминисан из даље анализе „Раст БДП-а", јер статистичка значајност израчунате *F* статистике има највишу вредност. Ова процедура се понавља све док у анализи не остану само они индикатори са статистички значајном вредношћу *F* статистике. Резултати значајности индикатора у другом кораку представљени су у наредној табели.

Табела 4: Значајност индикатора у другом кораку – шест индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	БДП	0.815	0.664	6.924	0.00037
2.	Извоз добара и услуга	0.757	0.573	4.698	0.00353
3.	Стране директне инвестиције	0.703	0.494	3.420	0.01631
4.	Инфлација	0.556	0.309	1.566	0.20621
5.	Домаћи кредити у приватном сектору	0.231	0.053	0.197	0.97395
6.	Незапосленост	0.166	0.028	0.099	0.99568

У другом кораку је искључен индикатор „Незапосленост". Након тога се прешло на трећи корак процедуре, а значајност индикатора се може видети у следећој табели.

Табела 5: Значајност индикатора у трећем кораку – пет индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	БДП	0.788	0.621	7.208	0.00039
2.	Извоз добара и услуга	0.715	0.511	4.602	0.00504
3.	Стране директне инвестиције	0.705	0.497	4.348	0.00666
4.	Инфлација	0.647	0.419	3.168	0.02648
5.	Домаћи кредити у приватном сектору	0.277	0.077	0.366	0.86652

Процедура за постепену елиминацију индикатора завршила се у четвртом кораку, у којем су преостала четири индикатора са значајном вредношћу *F* статистике: „БДП", „Извоз добара и услуга", „Стране директне инвестиције" и „Инфлација". Коначни резултати процедуре представљени су у табели 6.

Табела 6: Значајност индикатора у последњем кораку – четири индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	БДП	0.797	0.635	10.012	0.00000
2.	Извоз добара и услуга	0.767	0.588	8.216	0.00028
3.	Стране директне инвестиције	0.730	0.533	6.560	0.00112
4.	Инфлација	0.601	0.361	3.251	0.02977

Редослед земаља Европске уније на основу економске развијености коришћењем статистички значајних индикатора представљен је у табели 7.

Табела 7: Коначан редослед земаља Европске уније

РБ	*Земља*	*И-одстојање*
1	Луксембург	74.80
2	Ирска	34.12
3	Мађарска	27.52
4	Холандија	24.92
5	Кипар	17.47
6	Словачка	16.84
7	Грчка	16.57
8	Данска	16.53
9	Француска	13.19
10	Финска	12.75
11	Шпанија	12.45
12	Белгија	12.38
13	Хрватска	12.06
14	Немачка	11.43
15	Летонија	11.20
16	Пољска	10.97
17	Шведска	10.47
18	Словенија	10.30

19	Аустрија	10.16
20	Италија	10.13
21	Литванија	10.00
22	Чешка	9.94
23	Велика Британија	9.25
24	Естонија	8.70
25	Португал	8.46
26	Бугарска	7.62
27	Румунија	7.09
28	Малта	5.77

Резултати показују да је земља са највећим степеном економске развијености Луксембург. Разлог можемо пронаћи у чињеници да је ова земља имла највише вредности за три индикатора (БДП, Стране директне инвестиције и Извоз добара и услуга), а за индикатор „Инфлација" најмању вредност, па се може рећи да су добијени резултати очекивани.

Следећа земља на ранг листи је Ирска, која такође има изразито високе вредности за прва три најзначајнија индикатора. Најслабије развијене земље су Бугарска (најнижи БДП и највиши степен инфлације), Румунија (други најнижи БДП и други највиши степен инфлације) и Малта (најнижи проценат Страних директних инвестиција).

Израчунати су и коефицијенти корелације између коначних Ивановићевих одстојања у сваком од корака процедуре за постепену елиминацију индикатора, а резултати су представљени у табели 8.

Табела 8: Коефицијенти корелације између израчунатих И-одстојања

	И-одст. – 7 инд.	*И-одст. – 6 инд.*	*И-одст. – 5 инд.*	*И-одст. – 4 инд.*
И-одст. – 7 инд.	1.000**	0.983**	0.969**	0.965**
И-одст. – 6 инд.	0.983**	1.000**	0.987**	0.974**
И-одст. – 5 инд.	0.969**	0.987**	1.000**	0.978**
И-одст. – 4 инд.	0.965**	0.974**	0.978**	1.000**

Коришћењем процедуре Ивановићевог одстојања за елиминацију индикатора, редукована је димензија проблема, а број посматраних индикатора смањен је са седам на четири (смањање димензије проблема за 42.85%). Том приликом, изгубљена је количина информација од свега 6.88%, што се може израчунати коефицијентом детерминације између Ивановићевог одстојања израчунатог на почетних седам индикатора и коначна четири индикатора (R=0.965; p<0.01).

5.2 Процедура Ивановићевог одстојања за постепену селекцију индикатора

При коришћењу стандардних процедура за постепену селекцију индикатора, у првом кораку се у анализу укључи само један индикатор. Ивановићево одстојање се не може израчунати за мање од два индикатора (због парцијалних коефицијената корелације). Из тог разлога, почетан избор индикатора је дефинисан коришћењем суме апсолутних корелација избором два индикатора са највећом сумом. Алгоритам процедуре за постепену селекцију индикатора дат је на слици 6:

Слика 6: Алгоритам процедуре за постепену селекцију индикатора

Анализа је започета укључивањем индикатора „БДП” и „Инфлација”, на основу којих је израчунато Ивановићево одстојање. Решење првог корака процедуре за постепену селекцију индикатора представљено је у табели 9.

Табела 9: Значајност индикатора у првом кораку – два индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	БДП	0.808	0.653	23.509	0.00000
2.	Инфлација	0.735	0.540	14.687	0.00006
3.	Извоз добара и услуга	0.686	0.471	11.111	0.00035
4.	Стране директне инвестиције	0.529	0.280	4.857	0.01651
5.	Раст БДП-а	-0.269	0.072	0.973	0.39172
6.	Домаћи кредити у приватном сектору	0.235	0.055	0.730	0.49189
7.	Незапосленост	-0.114	0.013	0.164	0.84982

Резултати показују да су индикатори „Извоз добара и услуга” и „Стране директне инвестиције” статистички значајни. У следећем кораку процедуре, укључује се индикатор „Извоз добара и услуга”, јер статистичка значајност његове *F* статистике има мању вредност. Ова процедура се понавља све док постоји барем један индикатор чија је израчуната вредност *F* статистике значајна. Резултати другог корака представљени су у табели 10.

Табела 10: Значајност индикатора у другом кораку – три индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	БДП	0.819	0.671	16.298	0.00001
2.	Извоз добара и услуга	0.774	0.559	11.954	0.00006
3.	Инфлација	0.661	0.437	6.208	0.00283
4.	Стране директне инвестиције	0.502	0.252	2.695	0.06854
5.	Раст БДП-а	-0.230	0.053	0.448	0.72097
6.	Домаћи кредити у приватном сектору	0.169	0.028	0.234	0.87148
7.	Незапосленост	-0.027	0.001	0.006	0.99936

У другом кораку су идентификована три статистички значајна индикатора: Стране директне инвестиције, „БДП”, „Извоз добара и услуга” и „Инфлација”. Преостали индикатори нису статистички значајни (p>0.05), па се овај корак проглашава као последњи. Одговарајуће Ивановићево одстојање, које је израчунато на основу три статистички значајна индикатора, представљено је у табели 11.

Табела 11: Редослед земаља Европске уније – три индикатора

РБ	*Земља*	*И-одстојање*
1	Луксембург	50.87
2	Ирска	17.82
3	Грчка	10.61
4	Кипар	9.84
5	Словачка	9.42
6	Данска	8.71
7	Холандија	7.34
8	Француска	6.06
9	Малта	5.79
10	Хрватска	5.74
11	Аустрија	5.03
12	Шпанија	4.95
13	Летонија	4.46
14	Финска	4.23
15	Белгија	4.15
16	Немачка	4.09
17	Пољска	3.84
18	Шведска	3.64
19	Словенија	3.38
20	Италија	3.28
21	Мађарска	3.25
22	Литванија	2.88
23	Чешка	2.85

24	Велика Британија	2.01
25	Естонија	1.76
26	Португал	1.05
27	Бугарска	0.47
28	Румунија	0.15

У табели 12 могу се видети коефицијенти корелације између Ивановићевог одстојања израчунатог на оригиналном скупу података (осам индикатора), почетног одстојања у процедури за постепену селекцију (два индикатора) и коначног о одстојања у процедури за постепену селекцију (три индикатора).

Табела 12: Коефицијенти корелације између израчунатих И-одстојања

	И-одст. – 8 инд.	*И-одст. – 2 инд.*	*И-одст. – 3 инд.*
И-одст. – 8 инд.	1.000^{**}	0.885^{**}	0.909^{**}
И-одст. – 2 инд.	0.885^{**}	1.000^{**}	0.988^{**}
И-одст. – 3 инд.	0.909^{**}	0.988^{**}	1.000^{**}

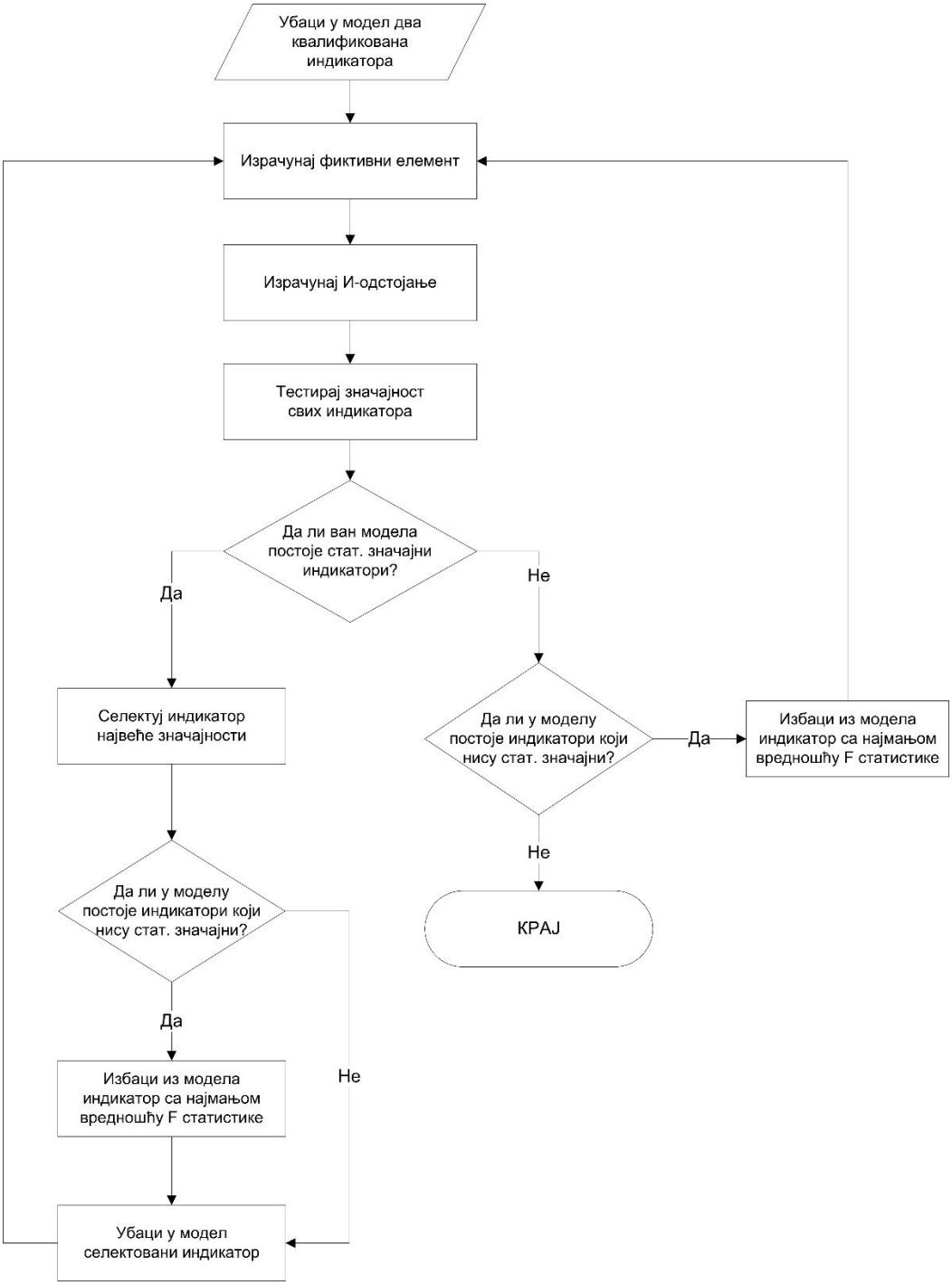
Под претпоставком је да наведени проблем у потпуности објашњен са почетних седам индикатора, први корак анализе која укључује само два индикатора обухватила би 78.32% варијабилитета, док коначни резултат процедуре за постепену селекцију индикатора објашњава 82.63%, с тим да су димензије проблема са 28,57% (два коришћена индикатора) доведене на 42,85% (три коришћена индикатора). У овој ситуацији, поређењем процедуре елиминације и процедуре селекције индикатора можемо закључити да боље резултате добијамо процедуром елиминације индикатора.

5.3 Процедура Ивановићевог одстојања „корак по корак”

Процедура Ивановићевог одстојања „корак по корак” је модификација процедуре за постепену селекцију индикатора. Може се рећи да заправо представља комбинацију претходне две процедуре. Њен почетак и ток је исти као и код процедуре за постепену селекцију, а разлика је у томе што се у сваком кораку тестира значајност и оних индикатора који се већ налазе у моделу. У случају да израчуната F статистика неког од индикатора изгуби статистичку значајност, такав индикатор је потребно искључити из модела. То се дешава када вредност израчунате F статистике пређе дефинисану границу за елиминацију ($p > 0.1$, као у процедури за постепену елиминацију индикатора). Граница за селекцију индикатора је иста као и код процедуре за постепену селекцију ($p < 0.05$).

У овом примеру, резултати процедуре „корак по корак” идентични су са резултатима који су добијени коришћењем методе за постепену селекцију индикатора, што се може видети у табелама 9, 10 и 11. У другом кораку, ниједан од индикатора који су у том тренутку били у моделу није изгубио статистичку значајност, па се резултати у потпуности поклапају, што се у општем случају не мора десити (Chatterjee & Hadi, 2012). Алгоритам процедуре „корак по корак“ дат је на слици 7.

При избору описаних процедура, саветује се коришћење процедуре за постепену елиминацију индикатора. Њена предност у односу на остале је у томе што се полази од скупа свих измерених индикатора, што значи да ће сваки од њих сигурно бити укључен у анализу. Осим тога, идеја Ивановићевог одстојања је смањивање димензије комплексног проблема, што се може учинити једино кроз постепену елиминацију која почиње од оригиналног скупа измерених индикатора.



Слика 6: Алгоритам процедуре за постепену селекцију индикатора

6. ПРИМЕНА СЕКВЕНЦИЈАЛНОГ ИВАНОВИЋЕВОГ ОДСТОЈАЊА

Ивановићево одстојање пружа могућност интегрисања великог броја оригиналних индикатора у само један, композитни индикатор. Вредности израчунатог одстојања се могу користити у циљу мерења интензитета посматране појаве, као и за рангирање елемената посматрања. У случају тестирања развијености земаља, Ивановићево одстојање се рачунало мерењем удаљености сваке земље од једне фиктивне, која је преузела минималне вредности индикатора из посматраног скупа. Притом, неопходно је да сви индикатори буду истосмерно оријентисани са израчунатим Ивановићевим одстојањем. На тај начин, добијен је редослед земаља у којем већа вредност одстојања означава већи степен развијености.

Као фиктивни елемент, могла је бити дефинисана земља са максималним вредностима свих посматраних индикатора. У том случају, најмања вредност Ивановићевог одстојања припадала би најразвијенијој земљи, док би ранг листа на основу развијености била формирана од најмање до највеће вредности Ивановићевог одстојања.

Једна од хипотеза ове дисертације је да се Ивановићево одстојање може користити у циљу откривања нестандардних опсервација. Под појмом нестандардне опсервације подразумевају се све оне јединице посматрања које на неки начин нису конзистентне са осталима, односно оне опсервације које одступају од осталих у више од једне димензије, у било ком смеру.

Откривање мултиваријационих нестандардних опсервација се најчешће спроводи коришћењем Махаланобисовог одстојања. Ако је обим узорка довољно велики, израчунато Махаланобисово одстојање подлеже хи-квадрат расподели (Bogl et al., 2017) са k степени слободе, где k представља број индикатора (Brereton, 2014). Коришћењем израчунатог Махаланобисовог одстојања и одговарајуће статистичке значајности за сваку од опсервација посебно, као нестандардне се идентификују оне опсервације чија је значајност мања од 0.001 (Tabachnick & Fidell, 2013). Махаланобисово одстојање се рачуна помоћу вектора

средњих вредности и коваријационо-дисперзионе матрице оригиналних променљивих. Као нестандардне се идентификују оне опсервације које имају високу вредност одстојања, однодно оне које значајно одступају од израчунатих средњих вредности индикатора.

Како би се Ивановићево одстојање могло користити у циљу откривања нестандардних опсервација, потребно је дефинисати фиктивни елемент који ће преузети *просечне* вредности измерених индикатора. Након тога, израчунава се одговарајуће Ивановићево одстојање. Опсервације које значајно одступају од просека у више од једне димензије, могу се идентификовати као потенцијалне нестандардне опсервације. Из тог разлога, одабрани индикатори не морају бити истосмерно оријентисани са израчунатим Ивановићевим одстојањем, јер ће у овом случају одстојање представљати степен одступања неке опсервације од измерених просечних вредности. Високе вредности израчунатог Ивановићевог одстојања сугеришу да опсервација може бити нестандардна.

Коришћење израчунатог Ивановићевог одстојања у циљу откривања нестандардних опсервација може бити примењено при директној редукцији димензије проблема, односно када се полази од скупа свих измерених индикатора. Из тог разлога, коришћена је метода за постепену елиминацију индикатора. Као што је већ поменуто, у истраживањима се најчешће користе методе за постепену елиминацију (Mantel, 1970), јер оне омогућавају да сви индикатори у првом кораку буду укључени у анализу, обезбеђујући на тај начин већу непристрасност од методе за постепену селекцију.

На одабраном скупу индикатора, израчуната су два одстојања, Махаланобисово и Ивановићево. Након тога, истим тестом је испитивана значајност поменутих одстојања. Граница значајности за идентификацију нестандардне опсервације је дефинисана у складу са правилима тестирања значајности Махаланобисовог одстојања (p<0.001). Израчунато Ивановићево одстојање је након тога коришћено у циљу откривања нестандардних опсервација.

6.1 Откривање нестандардних опсервација – економска развијеност

У овој студији случаја, откривање нестандардних опсервација спроведено је коришћењем седам индикатора економске развијености земаља Европске уније. У питању су исти подаци који су коришћени при редукцији димензије проблема кроз дефинисане процедуре секвенцијалног Ивановићевог одстојања. Као фиктивна опсервација, формиран је елемент са просечним вредностима поменутих индикатора. Резултати првог корака дати су у следећој табели:

Табела 13: Економска развијеност – први корак (седам индикатора)

Земља	*Махаланобисово одстојање*	*Значајност М-одст.*	*Ивановићево одстојање*	*Значајност И-одст.*
Луксембург	18.52	0.00981	32.04	0.00004
Малта	19.27	0.00739	11.76	0.10862
Грчка	15.06	0.03518	10.82	0.14688
Кипар	16.29	0.02258	9.51	0.21809
Румунија	7.71	0.35859	7.44	0.38432
Ирска	11.11	0.13367	7.31	0.39770
Мађарска	13.86	0.05368	5.98	0.54254
Бугарска	6.77	0.45332	5.54	0.59401
Шпанија	9.99	0.18916	4.14	0.76371
Велика Британија	5.43	0.60765	4.11	0.76738
Данска	7.19	0.40969	3.99	0.78041
Хрватска	4.46	0.72574	3.00	0.88461
Холандија	3.84	0.79791	2.77	0.90568
Словачка	5.17	0.63890	2.61	0.91822
Аустрија	4.76	0.68948	2.29	0.94210
Шведска	4.16	0.76141	2.08	0.95502
Француска	3.81	0.80121	2.05	0.95690
Белгија	5.04	0.65558	1.83	0.96862
Немачка	4.01	0.77903	1.82	0.96916
Италија	2.60	0.91923	1.68	0.97556

Португал	3.51	0.83404	1.67	0.97601
Литванија	1.52	0.98173	1.59	0.97911
Чешка	2.09	0.95468	1.35	0.98697
Естонија	3.38	0.84817	1.25	0.98976
Пољска	4.57	0.71193	1.19	0.99114
Летонија	2.08	0.95538	1.08	0.99342
Словенија	1.06	0.99378	0.99	0.99506
Финска	1.73	0.97312	0.75	0.99796

У првом кораку, на оригиналних седам индикатора, Махаланобисово одстојање није идентификовало ниједну нестандардну опсервацију. Ивановићево достојање је пронашло једну земљу која се значајно разликује од осталих, а то је Луксембург. Разлоге можемо пронаћи у вредностима индикатора „БДП", „Стране директне инвестиције", „Извоз добара и услуга" и „Инфлација", који су поменути у претходном поглављу. У наредној табели, приказане су вредности израчунатих одстојања, одговарајући рангови, као и разлике додељених рангова.

Табела 14: Вредности и рангови одстојања – први корак (седам индикатора)

Земља	*Махаланобисово одстојање*	*М-ранг*	Ивановићево одстојање	*И-ранг*	*Разлика рангова*
Луксембург	18.52	2	32.04	1	-1
Малта	19.27	1	11.76	2	1
Грчка	15.06	4	10.82	3	-1
Кипар	16.29	3	9.51	4	1
Румунија	7.71	8	7.44	5	-3
Ирска	11.11	6	7.31	6	0
Мађарска	13.86	5	5.98	7	2
Бугарска	6.77	10	5.54	8	-2
Шпанија	9.99	7	4.14	9	2
Велика Британија	5.43	11	4.11	10	-1
Данска	7.19	9	3.99	11	2

Хрватска	4.46	16	3.00	12	-4
Холандија	3.84	19	2.77	13	-6
Словачка	5.17	12	2.61	14	2
Аустрија	4.76	14	2.29	15	1
Шведска	4.16	17	2.08	16	-1
Француска	3.81	20	2.05	17	-3
Белгија	5.04	13	1.83	18	5
Немачка	4.01	18	1.82	19	1
Италија	2.60	23	1.68	20	-3
Португал	3.51	21	1.67	21	0
Литванија	1.52	27	1.59	22	-5
Чешка	2.09	24	1.35	23	-1
Естонија	3.38	22	1.25	24	2
Пољска	4.57	15	1.19	25	10
Летонија	2.08	25	1.08	26	1
Словенија	1.06	28	0.99	27	-1
Финска	1.73	26	0.75	28	2

У табели 14 се може приметити да су Махаланобисово и Ивановићево одстојање приказали сличне резултате. Обе методе су идентификовале Луксембург, Малту, Грчку и Кипар као земље са највећим вредностима одстојања (односно као потенцијалне нестандардне опсервације). Тестирана је зависност између Махаланобисовог и Ивановићевог одстојања. Израчунати коефицијент корелације износио је 0.927.

Након тога, примењена је процедура за постепену елиминацију, како би се дошло до статистички значајних индикатора за откривање мултиваријационих нестандардних опсервација. У првом кораку, индикатор који је елиминисан је „Незапосленост”, што се може видети у наредној табели:

Табела 15: Значајност индикатора – први корак (седам индикатора)

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Извоз добара и услуга	0.739	0.546	3.438	0.01402
2.	БДП	0.616	0.379	1.747	0.15468
3.	Инфлација	0.461	0.212	0.771	0.61808
4.	Стране директне инвестиције	0.314	0.098	0.312	0.93971
5.	Домаћи кредити у приватном сектору	0.180	0.032	0.096	0.99803
6.	Раст БДП-а	0.099	0.010	0.028	0.99996
7.	Незапосленост	0.051	0.003	0.007	1.00000

Анализа је настављена коришћењем преосталих шест индикатора. Елиминација индикатора по корацима је била следећа (међурезултати значајности индикатора су дати у прилогу):

- 1. корак – Незапосленост
- 2. корак – Раст БДП-а
- 3. корак – Домаћи кредити у приватном сектору
- 4. корак – Стране директне инвестиције
- 5. корак – Инфлација
- 6. корак – Коначни резултати

У шестом кораку, оба преостала индикатора, „Извоз добара и услуга” и „БДП”, идентификовани су као статистички значајни, па је тај корак проглашен за последњи. Значајност индикатора представљена је у табели 16.

Табела 16: Значајност индикатора – последњи корак (два индикатора)

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Извоз добара и услуга	0.785	0.616	20.071	0.00001
2.	БДП	0.774	0.599	18.678	0.00001

Вредности Махаланобисовог и Ивановићевог одстојања са одговарајућим значајностима су приказани у табели 17.

Табела 17: Економска развијеност – последњи корак (два индикатора)

Земља	*Махаланобисово одстојање*	*Значајност М-одст.*	*Ивановићево одстојање*	*Значајност И-одст.*
Луксембург	17.37	0.00017	22.73	0.00001
Ирска	3.66	0.16014	3.95	0.13850
Малта	5.09	0.07838	2.83	0.24274
Грчка	0.88	0.64297	1.17	0.55674
Румунија	0.94	0.62549	0.98	0.61135
Велика Британија	2.01	0.36611	0.96	0.61892
Бугарска	2.14	0.34224	0.87	0.64697
Италија	1.32	0.51570	0.87	0.64869
Француска	1.66	0.43680	0.86	0.65164
Шпанија	0.94	0.62632	0.75	0.68880
Хрватска	0.89	0.64161	0.73	0.69384
Финска	1.48	0.47680	0.64	0.72685
Мађарска	2.24	0.32575	0.63	0.73008
Португал	0.47	0.79047	0.62	0.73444
Словачка	1.91	0.38554	0.61	0.73856
Шведска	1.76	0.41459	0.53	0.76862
Немачка	1.63	0.44229	0.48	0.78745
Пољска	0.51	0.77437	0.45	0.79937
Летонија	0.84	0.65621	0.44	0.80307
Аустрија	1.49	0.47420	0.41	0.81402
Холандија	0.45	0.79834	0.40	0.81821
Данска	1.16	0.55920	0.32	0.85161
Естонија	0.99	0.60901	0.27	0.87219
Белгија	0.20	0.90569	0.25	0.88316
Литванија	0.76	0.68283	0.22	0.89619
Словенија	0.55	0.75830	0.15	0.92621
Чешка	0.43	0.80774	0.13	0.93718
Кипар	0.20	0.90262	0.10	0.95181

У последњем кораку, обе методе су идентификовале само једну нестандардну опсервацију, а то је Луксембург. Осим Луксембурга, као земље са већим вредностима одстојања су се истакле Ирска и Малта. Међутим, значајност израчунатих одстојања је показала да нису у питању нестандардне опсервације. У табели 18, осим израчунатих одстојања, могу се видети и одговарајући рангови, као и разлике додељених рангова.

Табела 18: Вредности и рангови одстојања – последњи корак (два индикатора)

Земља	*Махаланобисово одстојање*	*М-ранг*	Ивановићево одстојање	*И-ранг*	*Разлика рангова*
Луксембург	17.37	1	22.73	1	0
Ирска	3.66	3	3.95	2	-1
Малта	5.09	2	2.83	3	1
Грчка	0.88	19	1.17	4	-15
Румунија	0.94	16	0.98	5	-11
Велика Британија	2.01	6	0.96	6	0
Бугарска	2.14	5	0.87	7	2
Италија	1.32	13	0.87	8	-5
Француска	1.66	9	0.86	9	0
Шпанија	0.94	17	0.75	10	-7
Хрватска	0.89	18	0.73	11	-7
Финска	1.48	12	0.64	12	0
Мађарска	2.24	4	0.63	13	9
Португал	0.47	24	0.62	14	-10
Словачка	1.91	7	0.61	15	8
Шведска	1.76	8	0.53	16	8
Немачка	1.63	10	0.48	17	7
Пољска	0.51	23	0.45	18	-5
Летонија	0.84	20	0.44	19	-1
Аустрија	1.49	11	0.41	20	9
Холандија	0.45	25	0.40	21	-4
Данска	1.16	14	0.32	22	8

Естонија	0.99	15	0.27	23	8
Белгија	0.20	28	0.25	24	-4
Литванија	0.76	21	0.22	25	4
Словенија	0.55	22	0.15	26	4
Чешка	0.43	26	0.13	27	1
Кипар	0.20	27	0.10	28	1

Коефицијент корелације између коначних вредности Махаланобисовог и Ивановићевог одстојања износио је 0.69 (p<0.01). Резултати првог и последњег корака дефинитивно потврђују чињеницу да је методологију Ивановићевог одстојања могуће користити у циљу откривања мултиваријационих нестандардних опсервација. На оригиналном скупу индикатора, Ивановићево одстојање је показало већу моћ при идентификацији мултиваријационих нестандардних опсервација. Разлог можемо пронаћи у томе што Махананобисово одстојање не задовољава услов раста, услов анулирања дуплицитета у информацијама, услов асиметрије и услов независности од почетка (Ivanović, 1963).

7.2 Откривање нестандардних опсервација – пољопривредна производња

Пољопривредна статистика је важно средство за праћење и управљање тржиштем пољопривредних производа. Раст пољопривредне производње не само да развија економију и смањује незапосленост, већ такође помаже снижавању цена главних потрошачких производа и смањивању зависности земље од увоза.

Анкетари прикупљају податке о пољопривредној производњи непосредно од земљорадника (INSTAT, 2016). Највећа количина података се прикупља кроз Анкету о пољопривредној производњи (RZS, 2014) и Анкету о структури пољопривредних газдинстава (RZS, 2018). У овој студији случаја, откривање нестандардних опсервација спроведено је коришћењем пет индикатора пољопривредне производње дистриката у Албанији, која је имплементирала моделе анкета из Србије. Подаци су добијени агрегацијом микроподатака са пописа пољопривреде из 2016. године за сваки дистрикт посебно.

Почетни скуп чини пет најважнијих индикатора пољопривредне производње, а то су: 1. Искоришћена пољопривредна површина (у m^2); 2. Укупна обрадива површина (у m^2); 3. Укупна жетва (у m^2); 4. Производња ндустријских биљака (у m^2); 5. Производња сена (у m^2). Резултати првог корака дати су у наредној табели:

Табела 19: Пољопривредна производња – први корак (пет индикатора)

Дистрикт	*Махаланобисово одстојање*	*Значајност М-одст.*	*Ивановићево одстојање*	*Значајност И-одст.*
Маласи и Маде	28.83	0.00003	25.47	0.00011
Корча	30.87	0.00001	24.55	0.00017
Лушње	17.83	0.00317	18.83	0.00207
Ђирокастра	21.17	0.00075	12.46	0.02903
Фиер	9.27	0.09866	5.00	0.41592
Саранда	17.52	0.00362	3.32	0.65033
Валона	11.23	0.04703	3.03	0.69579
Каваја	2.07	0.83914	0.82	0.97593
Хас	0.77	0.97873	0.81	0.97665
Пука	0.75	0.98006	0.80	0.97683
Мирдита	0.74	0.98054	0.79	0.97764
Мат	0.66	0.98493	0.73	0.98135
Тропоја	0.80	0.97677	0.70	0.98288
Пекин	0.55	0.99027	0.67	0.98462
Кукеш	0.90	0.97031	0.64	0.98586
Булкиза	0.63	0.98636	0.64	0.98610
Либражд	0.54	0.99058	0.64	0.98619
Кучова	0.51	0.99184	0.64	0.98630
Пограделц	0.55	0.99015	0.61	0.98748
Малакастра	0.75	0.98029	0.57	0.98933
Делвина	6.87	0.23043	0.54	0.99065
Скрапар	0.48	0.99280	0.49	0.99249
Курбин	0.30	0.99765	0.41	0.99493

Колоња	0.58	0.98889	0.36	0.99627
Грамш	0.28	0.99797	0.36	0.99643
Девол	0.51	0.99158	0.35	0.99658
Пармет	1.38	0.92622	0.30	0.99764
Дибар	0.59	0.98837	0.30	0.99767
Круја	0.32	0.99719	0.29	0.99776
Тирана	5.89	0.31713	0.25	0.99841
Тепелена	0.59	0.98848	0.25	0.99850
Драч	5.49	0.35941	0.23	0.99879
Скадар	2.23	0.81677	0.18	0.99930
Лежје	1.00	0.96269	0.14	0.99965
Берат	0.59	0.98832	0.10	0.99984
Елбасан	0.93	0.96789	0.10	0.99985

У првом кораку, Махаланобисово одстојање је отктило три нестандардне опсервације: Маласи и Маде, Корчу и Ћирокастру, док је Ивановићево одстојање само прве две опсервације идентификовало као нестандардне. У табели 20 представљене су вредности израчунатих одстојања, одговарајући рангови, као и разлике додељених рангова.

Табела 20: Вредности и рангови одстојања – први корак (пет индикатора)

Дистрикт	*Махаланобисово одстојање*	*М-ранг*	*Ивановићево одстојање*	*И-ранг*	*Разлика рангова*
Маласи и Маде	28.83	2	25.47	1	-1
Корча	30.87	1	24.55	2	1
Лушње	17.83	4	18.83	3	-1
Ћирокастра	21.17	3	12.46	4	1
Фиер	9.27	7	5.00	5	-2
Саранда	17.52	5	3.32	6	1
Валона	11.23	6	3.03	7	1
Каваја	2.07	12	0.82	8	-4

Хас	0.77	18	0.81	9	-9
Пука	0.75	19	0.80	10	-9
Мирдита	0.74	21	0.79	11	-10
Мат	0.66	22	0.73	12	-10
Тропоја	0.80	17	0.70	13	-4
Пекин	0.55	29	0.67	14	-15
Кукеш	0.90	16	0.64	15	-1
Булкиза	0.63	23	0.64	16	-7
Либражд	0.54	30	0.64	17	-13
Кучова	0.51	32	0.64	18	-14
Пограделц	0.55	28	0.61	19	-9
Малакастра	0.75	20	0.57	20	0
Делвина	6.87	8	0.54	21	13
Скрапар	0.48	33	0.49	22	-11
Курбин	0.30	35	0.41	23	-12
Колоња	0.58	27	0.36	24	-3
Грамш	0.28	36	0.36	25	-11
Девол	0.51	31	0.35	26	-5
Пармет	1.38	13	0.30	27	14
Дибар	0.59	25	0.30	28	3
Круја	0.32	34	0.29	29	-5
Тирана	5.89	9	0.25	30	21
Тепелена	0.59	26	0.25	31	5
Драч	5.49	10	0.23	32	22
Скадар	2.23	11	0.18	33	22
Лежје	1.00	14	0.14	34	20
Берат	0.59	24	0.10	35	11
Елбасан	0.93	15	0.10	36	21

Из резултата у претходној табели, може се приметити да су обе методе израчунале високе вредности одстојања за првих седам дистриката са ранг листе. Коефицијент корелације између израчунатих одстојања износио је 0.441 (p<0.01).

Након добијених резултата, израчунате су статистичке значајности посматраних индикатора, а резултати су представљени у табели 21.

Табела 21: Значајност индикатора – први кораку (пет индикатора)

РБ	*Индикатор*	*R*	R^2	*F*	*P*
1.	Производња ндустријских биљака	0.738	0.545	7.177	0.00016
2.	Укупна обрадива површина	0.625	0.391	3.846	0.00823
3.	Производња сена	0.615	0.378	3.650	0.01068
4.	Укупна жетва	0.505	0.255	2.054	0.09932
5.	Искоришћена пољопривредна површина	0.307	0.094	0.624	0.68238

У првом кораку, „Искоришћена пољопривредна површина" је идентификована као индикатор који није статистички значајан, па је искључен из даље анализе. Елиминација индикатора по корацима је била следећа (међурезултати значајности индикатора су дати у прилогу):

- 1. корак – Искоришћена пољопривредна површина
- 2. корак – Укупна жетва
- 3. корак – Коначни резултати

Резултати значајности индикатора у последњем кораку приказани су у табели 22:

Табела 22: Значајност индикатора – последњи корак (три индикатора)

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Производња сена	0.715	0.511	11.157	0.00004
2.	Производња индустријских биљака	0.676	0.457	8.976	0.00018
3.	Обрадива површина	0.516	0.266	3.871	0.01810

Како су у трећем кораку сви индикатори идентификовани као статистички значајни, добијени резултати су проглашени за коначне. Израчунато Махаланобисово и Ивановићево одстојање, као и одговарајуће значајности, приказане су у наредној табели.

Табела 23: Пољопривредна производња – последњи корак (три индикатора)

Дистрикт	*Махаланобисово одстојање*	*Значајност М-одст.*	*Ивановићево одстојање*	*Значајност И-одст.*
Корча	28.48	0.00000	30.24	0.00000
Маласи и Маде	27.38	0.00000	25.24	0.00001
Лушње	13.33	0.00398	15.44	0.00147
Ђирокастра	15.17	0.00168	6.73	0.08118
Фиер	3.45	0.32765	3.78	0.28605
Валона	5.02	0.17057	2.06	0.56021
Каваја	1.28	0.73459	1.01	0.79817
Саранда	2.42	0.49011	0.95	0.81233
Хас	0.51	0.91604	0.58	0.90029
Пука	0.51	0.91610	0.58	0.90073
Мирдита	0.50	0.91804	0.57	0.90276
Мат	0.46	0.92790	0.54	0.91025
Пекин	0.43	0.93468	0.49	0.92008
Кучова	0.41	0.93899	0.48	0.92416
Тропоја	0.49	0.92175	0.47	0.92594
Либражд	0.41	0.93790	0.46	0.92671
Булкиза	0.42	0.93606	0.45	0.93034
Кукеш	0.41	0.93897	0.44	0.93114
Пограделец	0.39	0.94194	0.44	0.93158
Малакастра	0.36	0.94856	0.40	0.94077
Скрапар	0.30	0.95928	0.36	0.94795
Колоња	0.22	0.97440	0.30	0.96029
Курбин	0.29	0.96127	0.29	0.96143
Грамш	0.27	0.96545	0.26	0.96659
Круја	0.18	0.98018	0.24	0.97035
Пармет	0.17	0.98152	0.24	0.97113
Делвина	0.32	0.95712	0.23	0.97239
Девол	0.26	0.96652	0.23	0.97303
Дибар	0.34	0.95330	0.22	0.97397

Тепелена	0.15	0.98562	0.20	0.97750
Скадар	0.16	0.98447	0.16	0.98359
Тирана	0.18	0.98022	0.14	0.98643
Лежје	0.09	0.99337	0.10	0.99150
Елбасан	0.09	0.99259	0.08	0.99380
Драч	0.10	0.99209	0.08	0.99401
Берат	0.06	0.99656	0.06	0.99629

Коначни резултати показују да су обе методе идентификовале само две нестандардне опсервације, Корчу и Маласи и Маде. Област Корче је позната по производњи сена; 40% укупне количине у целој земљи се производи у овом дистрикту. Такође, област Корче има трећу највећу укупну обрадиву површину. Када је у питању дистрикт Маласи и Маде, он има највећу производњу индустријских биљака (преко 25%) у целој земљи, а једну од најслабијих производња сена. У табели 24 су дате вредности израчунатих одстојања, одговарајући рангови, као и разлике додељених рангова.

Табела 24: Вредности и рангови одстојања – последњи корак (три индикатора)

Дистрикт	*Махаланобисово одстојање*	*М-ранг*	*Ивановићево одстојање*	*И-ранг*	*Разлика рангова*
Корча	28.48	1	30.24	1	0
Маласи и Маде	27.38	2	25.24	2	0
Лушње	13.33	4	15.44	3	-1
Ђирокастра	15.17	3	6.73	4	1
Фиер	3.45	6	3.78	5	-1
Валона	5.02	5	2.06	6	1
Каваја	1.28	8	1.01	7	-1
Саранда	2.42	7	0.95	8	1
Хас	0.51	9	0.58	9	0
Пука	0.51	10	0.58	10	0
Мирдита	0.50	11	0.57	11	0

Мат	0.46	13	0.54	12	-1
Пекин	0.43	14	0.49	13	-1
Кучова	0.41	18	0.48	14	-4
Тропоја	0.49	12	0.47	15	3
Либражд	0.41	16	0.46	16	0
Булкиза	0.42	15	0.45	17	2
Кукеш	0.41	17	0.44	18	1
Пограделц	0.39	19	0.44	19	0
Малакастра	0.36	20	0.40	20	0
Скрапар	0.30	23	0.36	21	-2
Колоња	0.22	27	0.30	22	-5
Курбин	0.29	24	0.29	23	-1
Грамш	0.27	25	0.26	24	-1
Круја	0.18	28	0.24	25	-3
Пармет	0.17	30	0.24	26	-4
Делвина	0.32	22	0.23	27	5
Девол	0.26	26	0.23	28	2
Дибар	0.34	21	0.22	29	8
Тепелена	0.15	32	0.20	30	-2
Скадар	0.16	31	0.16	31	0
Тирана	0.18	29	0.14	32	3
Лежје	0.09	35	0.10	33	-2
Елбасан	0.09	34	0.08	34	0
Драч	0.10	33	0.08	35	2
Берат	0.06	36	0.06	36	0

Разлике између рангова показују да постоји велики степен поклапања у коришћеним методама одстојања. Такође, може се закључити да су осим поменута два дистрикта, Лушње и Ђирокастра имали релативно високе вредности израчунатих одстојања. Коефицијент корелације између Махаланобисовог и Ивановићевог одстојања на основу три индикатора износио је 0.973.

7.3 Откривање нестандардних опсервација – перформансе кошаркаша

НБА лига је најконкурентнија лига у свету кошарке. Велике количине прикупљених података са сваке утакмице захтевају мултидисциплинарни приступ при њиховој обради (Dehesa et al., 2019). У циљу предвиђања перформанси играча и тимова, најчешће се користи вишеструка регресиона анализа (Yang, 2015). Такође, са великим успехом се предвиђају редоследи тимова у регуларном делу сезоне (Hill, 2018), где свака екипа одигра 82 утакмице. Овакав интензитет такмичења доводи до повећања вероватноће повреда играча (Mack et al., 2018).

Стратегије тимова су засновне на статистичким параметрима и индикаторима перформанси, о чему најбоље сведочи чињеница да сваки тим у саставу стручног штаба има најмање два статистичара. Најмеродавније податке обезбеђују записничари запослени од стране НБА лиге, који на свакој утакмици прате велики број индикатора. На основу њихових вредности, формирају се модели за рангирање играча, који помажу да се утврде кључни индикатори перформанси играча (Mertz et al., 2016). Такође, посебна пажња се посвећује индикаторима перформанси тимова, који обезбеђују њихова међусобна поређења, али и анализу успешности тренутне сезоне у односу на претходне.

У трећем примеру у овој дисертацији, откривање нестандардних опсервација спроведено је над 30 најбољих плејмејкера НБА лиге. Из сваког тима је одабран најбољи плејмејкер и мерене су вредности девет индикатора: 1. Просечан број скокова; 2. Просечан број изгубљених лопти; 3. Просечан број фаулова; 4. Просечан број блокада; 5. Просечан број асистенција; 6. Просечан број украдених лопти; 7. Просечан број погођених тројки; 8. Просечан број поена; 9. Просечан шут за три поена. Вредности Махаланобисовог и Ивановићевог одстојања, са одговарајућим значајностима, дати су у табели 25:

Табела 25: Перформансе кошаркаша – први корак (девет индикатора)

Играч	*Махаланобисово одстојање*	*Значајност М-одст.*	*Ивановићево одстојање*	*Значајност И-одст.*
Бен Симонс	23.79	0.00465	29.46	0.00054
Расел Вестбрук	18.76	0.02732	17.92	0.03608
Џејмс Харден	13.94	0.12457	12.33	0.19555
Џорџ Хил	8.25	0.50876	6.95	0.64281
Пети Милс	5.97	0.74276	6.51	0.68832
Д. Ј. Аугустин	5.05	0.83004	5.96	0.74388
Крис Дан	10.62	0.30233	5.43	0.79544
Стеф Кари	11.60	0.23668	5.41	0.79712
Џон Вол	11.60	0.23701	4.50	0.87569
Лу Вилијамс	8.13	0.5215	4.29	0.89104
Колин Секстон	9.61	0.38291	3.96	0.91421
Дарен Колисон	7.73	0.56149	3.18	0.95670
Лука Дончић	10.67	0.29869	2.98	0.96511
Реџи Џексон	7.64	0.57126	2.92	0.96743
Треј Јанг	11.42	0.2482	2.48	0.98143
Џејру Холидеј	6.63	0.67562	2.32	0.98541
Џеф Тег	7.35	0.60052	2.00	0.99153
Крис Пол	11.64	0.23429	1.86	0.99355
Горан Драгић	3.77	0.92613	1.85	0.99361
Кемба Вокер	4.47	0.87766	1.62	0.99620
ДиАнђело Расел	4.60	0.86805	1.48	0.99733
Кајл Лаури	9.32	0.40835	1.46	0.99743
ДеАрон Фокс	5.45	0.79378	1.44	0.99756
Кајри Ирвинг	4.60	0.86768	1.39	0.99791
Донован Мичел	6.86	0.65197	1.36	0.99805
Денис Смит	7.84	0.54989	1.14	0.99904
Мајк Конли	6.78	0.65996	1.07	0.99925
Дејмијан Лилард	4.70	0.85961	1.01	0.99941
Лонзо Бол	8.36	0.49864	0.97	0.99950
Џамал Мареј	3.86	0.92028	0.83	0.99974

Рангови израчунатих вредности одстојања рангова приказани су у табели 26.

Табела 26: Вредности и рангови одстојања – први корак (девет индикатора)

Играч	*Махаланобисово одстојање*	*М-ранг*	*Ивановићево одстојање*	*И-ранг*	*Разлика рангова*
Бен Симонс	23.79	1	29.46	1	0
Расел Вестбрук	18.76	2	17.92	2	0
Џејмс Харден	13.94	3	12.33	3	0
Џорџ Хил	8.25	13	6.95	4	-9
Пети Милс	5.97	22	6.51	5	-17
Д. Ј. Аугустин	5.05	24	5.96	6	-18
Крис Дан	10.62	9	5.43	7	-2
Стеф Кари	11.60	5	5.41	8	3
Џон Вол	11.60	6	4.50	9	3
Лу Вилијамс	8.13	14	4.29	10	-4
Колин Секстон	9.61	10	3.96	11	1
Дарен Колисон	7.73	16	3.18	12	-4
Лука Дончић	10.67	8	2.98	13	5
Реџи Џексон	7.64	17	2.92	14	-3
Треј Јанг	11.42	7	2.48	15	8
Џејру Холидеј	6.63	21	2.32	16	-5
Џеф Тег	7.35	18	2.00	17	-1
Крис Пол	11.64	4	1.86	18	14
Горан Драгић	3.77	30	1.85	19	-11
Кемба Вокер	4.47	28	1.62	20	-8
ДиАнђело Расел	4.60	27	1.48	21	-6
Кајл Лаури	9.32	11	1.46	22	11
ДеАрон Фокс	5.45	23	1.44	23	0
Кајри Ирвинг	4.60	26	1.39	24	-2
Донован Мичел	6.86	19	1.36	25	6
Денис Смит	7.84	15	1.14	26	11
Мајк Конли	6.78	20	1.07	27	7
Дејмијан Лилард	4.70	25	1.01	28	3
Лонзо Бол	8.36	12	0.97	29	17
Џамал Мареј	3.86	29	0.83	30	1

У првом кораку, коришћењем почетних девет индикатора Махаланобисово одстојање није идентификовало ниједну нестандардну опсервацију. Ивановићево одстојање је идентификовало Бена Симонса као плејмејкера који се издваја од осталих играча. Бен Симонс је играч који никада не покушава са шутевима за 3 поена, али има сјајан учинак у преосталим индикаторима (8.8 скокова, 7.7 асистенција и 0.8 блокада по мечу). Тестирана је и значајност посматраних 9 индикатора и добијени су следећи резултати:

Табела 27: Значајност индикатора у првом кораку (девет индикатора)

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Просечан шут за три поена	0.744	0.554	2.755	0.02833
2.	Просечан број скокова	0.641	0.411	1.550	0.19796
3.	Просечан број изгубљених лопти	0.379	0.144	0.373	0.93484
4.	Просечан број фаулова	0.332	0.110	0.275	0.97423
5.	Просечан број блокада	0.330	0.109	0.272	0.97534
6.	Просечан број асистенција	0.250	0.063	0.148	0.99710
7.	Просечан број украдених лопти	0.212	0.045	0.105	0.99925
8.	Просечан број погођених тројки	0.205	0.042	0.097	0.99943
9.	Просечан број поена	0.105	0.011	0.025	1.00000

У првом кораку, као индикатор са најмањом статистичком значајношћу идентификован је „Просечан број поена”, па је искључен из даље анализе. Елиминација индикатора по корацима је била следећа (међурезултати значајности индикатора су дати у прилогу):

- 1. корак – Просечан број поена
- 2. корак – Просечан број украдених лопти
- 3. корак – Просечан број погођених тројки
- 4. корак – Просечан број асистенција
- 5. корак – Просечан број фаулова
- 6. корак – Просечан број изгубљених лопти
- 7. корак – Коначни резултати

Седми корак је уједно и последњи, јер су сва три преостала индикатора идентификована као статистички значајна. Коначни резултати значајности преосталих индикатора дати су у наредној табели:

Табела 28: Значајност индикатора у последњем кораку (три индикатора)

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Просечан шут за три поена	0.839	0.873	59.575	0.00000
2.	Просечан број скокова	0.649	0.573	11.630	0.00005
3.	Просечан број блокада	0.367	0.373	5.156	0.00625

Коначни резултати израчунатих одстојања са одговарајућим статистичким значајностима приказани су у табели 27. У последњем кораку, обе методе су издвојиле само једну нестандардну опсервацију. То је био исти кошаркаш којег је идентификовала метода Ивановићевог одстојања у првом кораку, коришћењем оригиналног скупа индикатора перформанси.

Табела 29: Перформансе кошаркаша – последњи корак (три индикатора)

Играч	*Махаланобисово одстојање*	*Значајност М-одст.*	*Ивановићево одстојање*	*Значајност И-одст.*
Бен Симонс	22.14	0.00006	27.88	0.00000
Расел Вестбрук	13.86	0.00310	9.22	0.02650
Џон Вол	8.48	0.03713	3.93	0.26907
Д. Ј. Аугустин	2.52	0.47082	3.31	0.34592
Џејру Холидеј	4.13	0.24778	2.44	0.48674
Џејмс Харден	4.00	0.26162	2.36	0.50202
Лука Дончић	4.94	0.17654	2.30	0.51169
Пети Милс	1.65	0.64880	2.17	0.53871
Дарен Колисон	1.39	0.70804	1.84	0.60572
Колин Секстон	1.35	0.71624	1.84	0.60645
Џорџ Хил	3.26	0.35347	1.72	0.63357
Стеф Кари	3.26	0.35379	1.67	0.64260
Реџи Џексон	1.59	0.66237	1.55	0.67102

Лу Вилијамс	1.51	0.67991	1.26	0.73899
Горан Драгић	1.71	0.63445	1.17	0.76014
ДеАрон Фокс	2.25	0.52229	0.84	0.84097
Кајри Ирвинг	1.84	0.60688	0.83	0.84133
Џеф Тег	1.78	0.62033	0.71	0.87192
Денис Смит	1.39	0.70810	0.53	0.91209
Трej Јанг	1.09	0.78049	0.52	0.91550
ДиАнђело Расел	0.52	0.91382	0.47	0.92450
Мајк Конли	0.24	0.97017	0.29	0.96189
Кајл Лаури	0.42	0.93603	0.25	0.96931
Крис Дан	0.64	0.88608	0.24	0.97066
Лонзо Бол	0.23	0.97303	0.23	0.97319
Дејмијан Лилард	0.24	0.97047	0.12	0.98982
Џамал Марej	0.18	0.98108	0.10	0.99214
Крис Пол	0.19	0.97917	0.09	0.99326
Донован Мичел	0.14	0.98636	0.07	0.99488
Кемба Вокер	0.07	0.99514	0.03	0.99859

У табели 30 приказани су одговарајући рангови, као и њихова разлика. Резултати показују потпуно поклапање за пет од првих шест рангираних играча. Плејмејкер који се на основу вредности израчунатих одстојања издвојио од преосталих је Расел Вестброк. У питању је играч који просечно бележи 11.1 скокова по утакмици, што свакако није карактеристично за позицију на којој игра.

Табела 30: Вредности и рангови одстојања – последњи корак (три индикатора)

Играч	*Махаланобисово одстојање*	*М-ранг*	*Ивановићево одстојање*	*И-ранг*	*Разлика рангова*
Бен Симонс	22.14	1	27.88	1	0
Расел Вестброк	13.86	2	9.22	2	0
Џон Вол	8.48	3	3.93	3	0
Д. Ј. Аугустин	2.52	9	3.31	4	-5
Џејру Холидеј	4.13	5	2.44	5	0

Џејмс Харден	4.00	6	2.36	6	0
Лука Дончић	4.94	4	2.30	7	3
Пети Милс	1.65	14	2.17	8	-6
Дарен Колисон	1.39	17	1.84	9	-8
Колин Секстон	1.35	19	1.84	10	-9
Џорџ Хил	3.26	7	1.72	11	4
Стеф Кари	3.26	8	1.67	12	4
Реџи Џексон	1.59	15	1.55	13	-2
Лу Вилијамс	1.51	16	1.26	14	-2
Горан Драгић	1.71	13	1.17	15	2
ДеАрон Фокс	2.25	10	0.84	16	6
Кајри Ирвинг	1.84	11	0.83	17	6
Џеф Тег	1.78	12	0.71	18	6
Денис Смит	1.39	18	0.53	19	1
Трej Јанг	1.09	20	0.52	20	0
ДиАнђело Расел	0.52	22	0.47	21	-1
Мајк Конли	0.24	24	0.29	22	-2
Кајл Лаури	0.42	23	0.25	23	0
Крис Дан	0.64	21	0.24	24	3
Лонзо Бол	0.23	26	0.23	25	-1
Дејмијан Лилард	0.24	25	0.12	26	1
Џамал Марej	0.18	28	0.10	27	-1
Крис Пол	0.19	27	0.09	28	1
Донован Мичел	0.14	29	0.07	29	0
Кемба Вокер	0.07	30	0.03	30	0

Након ових резултата, тестирана је зависност између Махаланобисовог и Ивановићевог одстојања. Коефицијент корелације износио је 0.914. То значи да је још једном показано како ове две методе приказују сличне резултате.

7. ЗАКЉУЧАК

Методологија Ивановићевог одстојања се заснива на претпостваци да је могуће измерити величину неке појаве кроз формирање јединственог синтетизованог индикатора. У досадашњим истраживањима, Ивановићево одстојање коришћено је у сврху мерења интензитета неке комплексне појаве, при чему је било неопходно да сви посматрани индикатори буду истосмерно оријентисани са израчунатим одстојањем. Као референтни елемент, коришћена је фиктивна опсервација која је преузела минималне вредности свих посматраних индикатора.

У овој дисертацији, методологија Ивановићевог одстојања је унапређена кроз дефинисање секвенцијалних процедура за постепену елиминацију индикатора, постепену селекцију индикатора, као и процедуре "корак по корак", која заправо представља модификацију дефинисане процедуре селекције. Свака од процедура приказана је кроз одговарајуће алгоритме. У циљу тестирања значајности оригиналног скупа индикатора, коришћена је одговарајућа F статистика. Као критеријум за доношење одлуке о елиминацији и/или селекцији посматраних индикатора, коришћен је ниво значајности поменуте статистике. Границе нивоа значајности дефинисане су у складу са одговарајућим методама у вишеструкој линеарној регресији и дискриминационој анализи.

Процедуре секвенцијалног Ивановићевог одстојања су детаљно објашњене, а након тога приказане на примеру економске развијености земаља Европске уније. На почетку анализе, одабрано је седам индикатора економске развијености, препоручених од стране Светске Банке. У процедури за постепену елиминацију, број посматраних индикатора се са оригиналних седам свео на четири индикатора који су идентификовани као статистички значајни. Уз редукцију димензије проблема од 42.85%, изгубљено је само 6.88% информација. Процедура за постепену селекцију индикатора је у последњем кораку идентификовала три статистички значајна индикатора. Редукцијом димензије проблема од 57.15% изгубљено је 17.37% информација. У процедури "корак по корак", коначни резултати су се поклапали са резултатима постепене селекције

индикатора, јер ниједан од индикатора за време спровођења процедуре није изгубио статистичку значајност.

Главна хипотеза ове дисертације је била да је, кроз дефинисане процедуре секвенцијалног Ивановићевог одстојања, могуће извршити идентификацију мултиваријационих нестандардних опсервација. Као референтни елемент, при формирању одстојања је коришћена фиктивна опсервација која је преузела просечне вредности свих посматраних индикатора. Из тог разлога, није било неопходно извршити линеарну трансформацију индикатора који нису истосмерно оријентисани са израчунатим Ивановићевим одстојањем.

У првој студији случаја, коришћени су већ поменути индикатори развијености земаља Европске уније. Осим Ивановићевог одстојања, коришћењем истих индикатора је израчунато и Махаланобисово одстојање, која је најчешће коришћена метода за идентификацију мултиваријационих нестандардних опсервација. На оригиналном скупу индикатора, Махаланобисово одстојање није идентификовало ниједну нестандардну опсервацију, док је Ивановићево одстојање идентификовало Луксембург, као земљу чије вредности посматраних индикатора значајно одступају од већине преосталих земаља из посматраног скупа. Коефицијент корелације између израчунатих одстојања на почетном скупу индикатора био је 0.927. Кроз процедуру за постепену елиминацију, задржана су само два индикатора која су идентификована као статистички значајна. У последњем кораку, оба одстојања су идентификовала само Луксембург као нестандардну опсервацију. Корелација између одстојања је износила 0.690.

Друга студија случаја односила се на пољопривредну статистику. Агрегацијом микропододатака са пописа пољопривреде, добијено је пет индикатора. На почетном скупу, Махаланобисово одстојање је открило три нестандардне опсервација, а Ивановићево одстојање две, уз међусобни коефицијент корелације 0.441. Редукцијом димензије проблема, почетни број индикатора је смањен на три најзначајнија. На основу статистички значајних индикатора, оба одстојања су идентификовала иста два дистрикта као јединице посматрања које се разликују од преосталих. Корелација између израчунатих одстојања износила је 0.973.

У трећој студији случаја, испитиване су перформансе кошаркаша НБА лиге који играју на позицији плејмејкера. На почетном скупу од девет индикатора, Махаланобисово одстојање није идентификовало ниједну нестандардну опсервацију, док је Ивановићево одстојање идентификовало једног кошаркаша. Корелација између одстојања је била -0.573. Коришћењем процедуре за постепену елиминацију, издвојена су три статистички значајна индикатора. Коначни резултати за оба одстојања су идентификовала истог кошаркаша као играча чије перформансе значајно одступају од осталих. То је био Бен Симонс, исти кошаркаш који је у првом кораку идентификован коришћењем Ивановићевог одстојања. Коефициент корелације између коначних вредности Махаланобисовог и Ивановићевог одстојања износио је 0.914.

Као главни научни доприноси ове докторске дисертације се могу навести унапређење методологије Ивановићевог одстојања кроз дефинисане процедуре за секвенцијалну селекцију и/или елиминацију индикатора, као и коришћење поменутих процедура у циљу откривања мултиваријационих нестандардних опсервација.

Међу осталим доприносима се могу издвојити:

- Евалуација постојећих метода за откривање универијационих и мултиваријационих нестандардних опсервација;
- Детаљан приказ мултиваријационих статистичких метода које се користе у циљу редукције димензије проблема, као и метода које;
- Детаљан приказ мултиваријационих статистичких метода које имају дефинисане процедуре за постепену елиминацију и селекцију променљивих;
- Потврђивање постављених хипотеза кроз резултате добијене употребом методологије секвенцијалног Ивановићевог одстојања.

Као правци будућих истраживања могу се дефинисати:

- Аутоматизација дефинисаних процедура секвенцијалног Ивановићевог одстојања;

- Формирање процедуре за постепену елиминацију идентификованих мултиваријационих нестандардних опсервација;
- Дефинисање одговарајуће статистике за тестирање значајности Ивановићевог одстојања, измереног од просечног фиктивног елемента.

ЛИТЕРАТУРА

1. Acuna, E., & Rodriguez, C. A. (2004). Meta Analysis Study of Outlier Detection Methods in Classification. *Proceedings IPSI 2004,* Venice, Italy.
2. Algur, S. P., & Biradar, J. G. (2017). Cooks Distance and Mahanabolis Distance Outlier Detection Methods to identify Review Spam, *International Journal of Engineering and Computer Science,6*(6), 21638-21649.
3. Al-Lagilli, S., Jeremic, V., Seke, K., Jeremic, D., & Radojicic, Z. (2011). Evaluating the health of nations: a Libyan perspective. *Libyan Journal of Medicine, 6*, 1-2.
4. Al Lagili S. A. M. (2013). *Model ocenjivanja indikatora poređenja performansi razvoja zemalja arapskog regiona, Doktorska disertacija*, Beograd, Fakultet organizacionih nauka, Univerzitet u Beogradu
5. Anderson, T. W. (1966). *An Introduction to Multivariate Statistical Analysis (7th edition).* London: John Wiley and Sons, Inc.
6. Bain, L., & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics (2nd edition).* Duxbury Classic series.
7. Bakon, M., Oliveira, I., Perissin, D., Sousa, J. J., & Papco, J. (2017). A Data Mining Approach for Multivariate Outlier Detection in Post-Processing of Multi-Temporal InSAR Results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(6), 2791-2798.
8. Barnett, V., & Lewis, T. (1994). *Outliers in statistical data (3rd edition).* New York: John Wiley & Sons, Inc.
9. Bauder, A. R., & Khoshgoftaar, M. T. (2017). Multivariate outlier detection in medicare claims payments applying probabilistic programming methods. *Health Services and Outcomes Research Methodology*, *17*, 256–289.
10. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602-613.

11. Bogl, M., Filzmoser, P., Gschwandtner, T., Lammarsch, T., Leite, R. A., Miksch, S., & Rind, A. (2017). Cycle Plot Revisited: Multivariate Outlier Detection Using a Distance-Based Abstraction. *Computer Graphics Forum, 36*(3), 227-238.
12. Brereton, R. G. (2014). The chi squared and multinormal distributions. *Journal of Chemometrics, 29*, 9-12.
13. Brereton, R. G., & Lloyd, G. R. (2016). Re-evaluating the role of the Mahalanobis distance measure. *Journal of Chemometrics, 30*, 134–143.
14. Bhushan, A., Sharker, M. H., & Karimi, H. A. (2015). Incremental principal component analysis based outlier detection methods for spatiotemporal data streams. *Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2*(4), International Workshop on Spatiotemporal Computing, Fairfax, Virginia, USA.
15. Bulajić, M. (2002). *Geodemografski model tržišnog prostora Srbije. Doktorska disertacija,* Beograd: Fakultet organizacionih nauka, Univerzitet u Beogradu.
16. Bulajić, M., Jeremić, V., Knežević, S., & Žarkić-Joksimović, N. (2013). A Statistical Approach to Evaluating Efficiency of Banks. *Economic Research-Ekonomska Istraživanja, 26*(4), 91–100.
17. Candes, E.J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM, 58*(3), article 11.
18. Cao, N., Lin, Y. R., Gotz, D., & Du, F. (2018). Z-Glyph: Visualizing outliers in multivariate data. *Information Visualization, 17*(1), 22–40.
19. Casas, P., Mazel, J., & Owezarski, P. (2012) Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge. *Computer Communications, 35*(7), 772-783.
20. Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.
21. Chatterjee, S., & Hadi, A. S. (2012*). Regression Analysis by Example (5th edition).* New Jersey: John Wiley & Sons, Inc.
22. Chatterjee, S., & Simonoff, J. S. (2013). *Handbook of Regression Analysis.* New Jersey: John Wiley & Sons, Inc.

23. Choi, Y., Park, C. G., & Lee, K. E. (2018). Evaluation of outlier detection methods for multiple linear regression model. *Journal of the Korean Data & Information Science Society, 29*(6), 1663-1677.

24. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd edition).* New Jersey: Lawrence Erlbaum Associates, Inc.

25. Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 19*(1), 15–18.

26. Cook, R. D. (1998). *Regression Graphic: Ideas for Studying Regression through Graphics*. New York: John Wiley & Sons.

27. Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.

28. Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics*. New York: John Wiley & Sons, Inc.

29. Costello, A. B., & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, & Evaluation, 10*(7), 1-9.

30. De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems, 50*(1), 1–18.

31. Dehesa, R., Vaquera, A., Goncalves, B., Mateus, N., Gomez-Ruano, M. A., & Sampaio, J. (2019). Key game indicators in NBA players' performance profiles. *Kinesiology, 51,* 92-101.

32. Dempster, A. P. (1969). *Elements of Continous Multivariate Analysis*. Reading: Addison-Wesley.

33. Dixon, W. J. (1973). *BMD Biomedical computer programs.* Los Angeles: University of California Press.

34. Dobrota, M. M. (2018). *Statistički pristup definisanju zone osetljivosti u metodama daljinskog uzorkovanja. Doktorska disertacija,* Beograd: Fakultet organizacionih nauka, Univerzitet u Beogradu.

35. Dobrota, M. M., & Dobrota, M. P. (2016). ARWU ranking uncertainty and sensitivity: What if the award factor was Excluded? *Journal of the Association for Information Science and Technology*, *67*(2), 480–482.

36. Dobrota, M. P. (2013). *Statistički pristup formiranju kompozitnih indikatora zasnovan na Ivanovićevom odstojanju. Doktorska disertacija,* Beograd: Fakultet organizacionih nauka, Univerzitet u Beogradu.

37. Dobrota, M. P., Jeremić, V. & Marković, A. (2012). A new perspective on the ICT Development Index. *Information Development, 28*(4), 81-85.

38. Dobrota, M. P., Martić, M., Bulajić, M., & Jeremić, V. (2015). Two-phased composite I-distance indicator approach for evaluation of countries' information development. *Telecommunications Policy*, 39(5), 406–420.

39. Dobrota, M. P., Bulajić, M., Bornmann, L., & Jeremić, V. (2016). A new approach to the QS university ranking using the composite I-distance indicator: Uncertainty and sensitivity analyses. *Journal of the Association for Information Science and Technology*, *67*(1), 200-211.

40. Dovoedo, Y. H., & Chakraborti, S. (2015). Boxplot-Based Outlier Detection for the Location-Scale Family. *Communications in statistics-simulation and computation, 44*(6), 1492-1513.

41. Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis (3rd edition).* New Jersey: John Wiley & Sons.

42. Đoković, A. (2013). *Strukturna korelaciona analiza u interpretaciji vektorskih koeficijenata korelacije. Doktorska disertacija,* Beograd: Fakultet organizacionih nauka, Univerzitet u Beogradu.

43. Field, A. (2005). Discovering statistics using SPSS *(2nd edition).* Thousand Oaks, CA: Sage Publications, Inc.

44. Filzmoser, P., Garrett, G. R., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences, 31*, 579–587.

45. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics, 7*(2), 179-188.

46. Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician 37*(2), 152–155.
47. Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression analysis: statistical modeling of a response variable (2nd edition).* Burlington, MA: Elsevier.
48. Gan, G., & Ng, M. K. (2017). *k* -means clustering with outlier removal. *Pattern Recognition Letters, 90,* 8–14.
49. Giménez, E., Crespi, M., Garridoa, M. S., & Gil, A. J. (2012). Multivariate outlier detection based on robust computation of Mahalanobis distances. Application to positioning assisted by RTK GNSS Networks, *International Journal of Applied Earth Observation and Geoinformation, 16,* 94–100.
50. Gordon, R. (2015). *Regression Analysis for the Social Sciences (2nd edition).* New York: Taylor & Francis.
51. Grubesic, T. H. (2006). On the application of fuzzy clustering for crime hot spot detection. *Journal of Quantitative Criminology, 22*(1), 77-105.
52. Habbema, J. D. F., & Hermans, J. (1977). Selection of Variables in Discriminant Analysis by F-statistic and Error Rate. *Technometrics, 19*(4), 487-493.
53. Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statisttical Society B, 54,* 761-771.
54. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques (3rd edition).* Morgan Kaufmann Publishers. Waltham, MA: Morgan Kaufmann Publishers.
55. Hardin, J., & Rocke, D. M. (2005). The Distribution of Robust Distances. *Journal of Computational and Graphical Statistics, 14*(4), 928-946.
56. Harris, J. R. (2001). *A Primer of Multivariate Statistics (3rd edition).* New Jersey: Lawrence Erlbaum Associates.
57. Hawkins, D. (1980). *Identification of Outliers.* London: Chapman and Hall.
58. Hill, B. (2018). Shadow and Spillover Effects of Competition in NBA Playoffs. *Journal of Sports Economics, 19*(8), 1067-1092.
59. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417-441.

60. Huberty, C. J. (1994). *Applied discriminant analysis.* New York: John Wiley and Sons, Inc.

61. Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and Discriminant Analysis (2nd edition).* New Jersey: John Wiley & Sons.

62. IBM Corporation (2013). *IBM SPSS Statistics 22 Algorithms*.

63. INSTAT (2016). *Agriculture and Livestock Statistics*. Institute of Statistics. Tirana Albania.

64. Iqbal, M. Z., Riaz, M., & Nasir, W. (2017). Multivariate outlier detection: a comparison among two clustering techniques. *Pakistan Journal of Agricultural Sciences, 54*(1), 227-231.

65. Ivanović, B. (1956). Projekat za klasifikaciju srezova FNRJ prema stepenu ekonomske razvijenosti. *Statistička revija, 4,* 287-297.

66. Ivanović, B. (1962). Struktura I-odstojanja između statističkih skupova. *Statistička revija, 1*, 1-14.

67. Ivanović, B. (1963). *Diskriminaciona analiza – sa primenom u ekonomskim istraživanjima.* Beograd: Institut za ekonomska istraživanja.

68. Ivanović, B. (1969). Neke dopune u vezi sa određivanjem I-odstojanja. *Statistička revija, 4*, 297-300.

69. Ivanović, B. (1972a). Klasifikacija i izbor statističkih obeležja. *Statistička revija, 1-2*, 63-74.

70. Ivanović, B. (1972b). Klasifikacija skupa objekata prema stepenu sličnosti sa primenom u razradi tipologije zemalja prema njihovom socio-ekonomskom profilu. *Statistička revija, 1-2*, 115-127.

71. Ivanović, B. (1973). *A method of establishing a list of development indicators.* Paris: United Nations Educational, Scientific and Cultural Organization.

72. Ivanović, B., & Fanchette, S. (1973). *Grouping and ranking of 30 countries of Sub-Saharan Africa, two distancebased methods compared.* Paris: United Nations Educational, Scientific and Cultural Organization.

73. Ivanović, B. (1975). Socio-ekonomski nivoi i profil razvijenih zemalja i zemalja u razvoju u 1970. godini. *Statistička revija, 3-4*, 183-198.

74. Ivanović, B. (1977). *Teorija klasifikacije.* Beograd: Institut za ekonomiku industrije.

75. Ivanović, B. (1978). Mere sličnosti vezane za jednu klasifikaciju. *Statistička revija, 1-2,* 48-62.

76. Ivanović, B. (1979). Određivanje značajnosti kriterijuma razvijenosti pomoću metode I-odstojanja. *Statistička revija, 3-4,* 276-280.

77. Ivanović, B. (1981). Problemi statističkih selekcija kod određivanja grupe najslabijih elemenata jednog skupa. *Statistička revija, 1-2,* 47-59.

78. Ivanović, B. (1982). Primena I-korelacije u metodologiji I-odstojanja i nov način određivanja redosleda indikatora prema stepenu značajnosti. *Statistička revija, 3-4*, 189-203.

79. Ivanović, B. (1988). Grupisanje obeležja preko metoda automatske klasifikacije. *Statistička revija, 1-2,* 11-20.

80. Jackson, D. A., & Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. *Environmetrics, 15*, 129–139.

81. Jackson, J. E. (1991). *A user's guide to principal components*. New York: John Wiley & Sons, Inc.

82. Jednak, S., Kragulj, D., & Bulajić, M. (2018). A comparative analysis of development in Southeast European countries. *Technological and Economic Development of Economy, 24*(1), 253–270.

83. Jeremić, V., Bulajić, M., Martić, M., & Radojičić, Z. (2011a). A fresh approach to evaluating the academic ranking of world universities. *Scientometrics, 87*(3), 587-596.

84. Jeremić, V., Išljamović, S., Petrović, N., Radojičić, Z., Marković, A., & Bulajić, M. (2011b). Human development index and sustainability: What's the correlation? *Metalurgia International, 16*(7), 63-67.

85. Jeremić, V., Vukmirović, D., Radojičić, Z., & Đoković, A. (2011c). Towards a framework for evaluating ICT infrastructure of countries: a Serbian perspective. *Metalurgia International, 16*(9), 15-18.

86. Jeremić, V., Marković, A., & Radojičić, Z. (2011d). ICT as crucial component of socio-economic development. *Management, 16*(60), 5-9.

87. Jeremić, V., Seke, K., Radojičić, Z., Jeremić, D., Marković, A., Slović, D., & Aleksić, A. (2011e). Measuring health of countries: a novel approach. *HealthMED, 5*(6), 1762-1766.

88. Jeremić, V. (2012). *Statistički model efikasnosti zasnovan na Ivanovićevom odstojanju. Doktorska disertacija,* Beograd: Fakultet organizacionih nauka, Univerzitet u Beogradu.

89. Jeremić, V., Bulajić, M., Martić, M., Marković, A., Savić, G., Jeremić, D., & Radojičić, Z. (2012). An Evaluation of European Countries Health Systems through Distance Based Analysis. *Hippokratia, 16*(2), 170-174.

90. Jolliffe, I. F. (2002). *Principal Component Analysis (2nd edition).* New York: Springer.

91. Jovanović, M., Jeremić, V., Savić, G., Bulajić, M., & Martić, M. (2012). How does the normalization of data affects the ARWU ranking? *Scientometrics, 93*(2), 319-327.

92. Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement, 20(1)*, 141-151.

93. Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401-415.

94. Kannan, K. S., & Manoj, K. (2015). Outlier Detection in Multivariate Data. *Applied Mathematical Sciences, 47*(9), 2317 – 2324.

95. Klecka, W. R. (1980). *Discriminant Analysis.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-019. Iowa, US: Sage.

96. Kosinski, S. A. (1999). A procedure for the detection of multivariate outliers. *Computational Statistics & Data Analysis,* 29, 145-161.

97. Kovačić, Z. (1994). *Multivarijaciona analiza.* Beograd: Ekonomski fakultet.

98. Lam, B. S. Y., & Choy, S. K. (2019). A Trimmed Clustering-Based l_1-Principal Component Analysis Model for Image Classification and Clustering Problems with Outliers. *Applied Science, 9*(8), 1562.

99. Landau, S., & Everitt, B. S. (2004). *A Handbook of Statistical Analyses using SPSS*. Boca Raton: Chapman & Hall/CRC Press LLC.

100. Leys, C., Klein, O., Dominicy, Z., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology, 74,* 150–156.

101. Liu H., Shah S., & Jiang W. (2004). On-line outlier detection and data cleaning. *Computers and Chemical Engineering, 28,* 1635–1647.

102. Liu, Y., Zumbo, B. D., & Wu, A. D. (2012). A Demonstration of the Impact of Outliers on the Decisions About the Number of Factors in Exploratory Factor Analysis. *Educational and Psychological Measurement, 72*(2), 181–199.

103. Liu, Y., & Zumbo, B. D. (2012). Impact of Outliers Arising from Unintended and Unknowingly Included Subpopulations on the Decisions About the Number of Factors in Exploratory Factor Analysis. *Educational and Psychological Measurement, 72*(3), 388–414.

104. Lu, C. T., Kou, Y., Zhao, J., & Chen, L. (2007). Detecting and tracking regional outliers in meteorological data. *Information Sciences, 177*(7), 1609-1632.

105. Mack, C. D., DiFiori, J. P., Meisel, P. L., & Dreyer, N. A. (2018). A second look at NBA game schedules: Response to Teramoto et al. *Journal of Science and Medicine in Sport, 21,* 228-229.

106. Mahalanobis, P. C. (1930). On tests and measures of groups divergence. *Journal of Asiatic Sociology of Bengal, 26,* 541–588.

107. Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India, 2,* 49-55.

108. Majewska, J. (2015). Identification of multivariate outliers – problems and challenges of visualization methods. *Economic Studies, 247,* 69-83.

109. Mantel, N. (1970). Why step-down procedures in variable selecton. *Technometrics,* 12, 621–625.

110. Mavridis, D., & Moustaki, I. (2008). Detecting Outliers in Factor Analysis Using the Forward Search Algorithm. *Multivariate Behavioral Research, 43*(3), 453-475.

111. Mertz, J., Hoover, L. D., Burke, J. M., Bellar, D., Jones, M. L., Leitzelar, B., & Judge, L. W. (2016). Ranking the Greatest NBA Players: A Sport Metrics Analysis. *International Journal of Performance Analysis in Sport, 16*(3), 737-759.

112. Milenković, N., Đoković, A., Milenković, J., Milanović, N. & Vukmirović, D. (2013). *Measuring effectiveness of elementary school education in Serbia - a multivariate statistical approach,* 32nd International Conference on Organizational Science Development, Portorož, Slovenia.

113. Milenković, N., Vukmirović, J., Bulajić, M., & Radojičić, Z. (2014). A multivariate approach in measuring socio-economic development of MENA countries. *Economic Modelling, 38,* 604-608.

114. Milenković, N., Đoković, A., & Vukmirović, A. (2016a). *Europe 2020 Strategy – A multivariate approach,* XV International Symposium SymOrg 2016: Reshaping the future through sustainable business development and enterpreneurship, 228-233.

115. Milenković, N., Đoković, A., Totić, S., & Radojičić, M. (2016b). *Bruto Drustveni Proizvod i Bruto Nacionalni Dohodak kao indikatori ekonomskog razvoja,* SYM-OP-IS 2014, Divčibare, 14-18.

116. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis (5th edition).* Chichester: John Wiley & Sons, Ltd.

117. Naik, G. R. (2018). *Advances in Principal Component Analysis - Research and Development.* Singapore: Springer.

118. O'Rourke, N., & Hatcher, L. (2013). *A Step-by-Step Approach to Using SAS® for Factor Analysis and Structural Equation Modeling (2nd edition).* North Carolina: SAS Institute Inc.

119. Osborne, J. W., & Banjanovic, E. S. (2016). *Exploratory Factor Analysis with SAS®*. Cary, North Carolina: SAS Institute Inc.

120. Oyeyemi, G. M., Bukoye, A., & Akeyede, I. (2015). Comparison of Outlier Detection Procedures in Multiple Linear Regressions. *American Journal of Mathematics and Statistics, 5*(1), 37-41.

121. Pages, J. (2015). *Multiple Factor Analysis by Example Using R.* Boca Raton: Taylor & Francis Group, LLC.

122. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 2*¸ 559-572.

123. Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* New Jersey: Lawrence Erlbaum Associates, Inc.

124. Pena, D., & Prieto, J. F. (2001). Multivariate Outlier Detection and Robust Covariance Matrix Estimation. *Technometrics, 43*(3), 286-310.

125. Penny K. I., & Jolliffe I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. *The Statistician, 50*(3), 295-308.

126. Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making Sense of Factor Analysis – The Use of Factor Analysis for Instrument Development in Health Care Research.* Sage Publications, Inc.

127. Podgorelec, V., Heričko, M., & Rozman, I. (2005). *Improving mining of medical data by outliers prediction.* IEEE Symposium on Computer-Based Medical Systems. 91-96.

128. Radojičić, Z. (2001). *Statističko merenje intenziteta pojava. Magistarski rad,* Beograd: Fakultet organizacionih nauka, Univerzitet u Beogradu.

129. Radojičić, Z. (2007). *Statistički model ocenjivanja na subjektivno procenjenim karakteristikama. Doktorska disertacija,* Beograd: Fakultet organizacionih nauka, Univerzitet u Beogradu.

130. Radojičić, Z., Išljamović, S., Petrović, N., & Jeremić, V. (2012). A novel approach to evaluating sustainable development. *Problemy Ekorozwoyu, 7,* 81–85.

131. Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis (3rd edition).* New York: John Wiley & Sons, Inc.

132. Reyment, R. A., & Joreskog, K. G. (1996). *Applied Factor Analysis in the Natural Sciences.* Cambridge: Cambridge Universitty Press.

133. Ro, K., Zou, C., & Wang, Z. (2015). Outlier detection for high-dimensional data. *Biometrika, 102*(3), 589–599.

134. Rocke, M. D., & Woodruff, D. L. (1996). Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association, 435*(91), 1047-1061.

135. Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association, 85*, 633-639.

136. Royston, P., & Sauerbrei, W. (2008). *Multivariable Model-Building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. New Jersey: John Wiley & Sons, Inc.

137. RZS (2014). *Anketa o poljoprivrednoj proizvodnji*, Republički zavod za statistiku Srbije. Beograd, Srbija.

138. RZS (2018). *Anketa o strukturi poljoprivrednih gazdinstava*, Republički zavod za statistiku Srbije. Beograd, Srbija.

139. Saha, P., Roy, N., Mukherjee, D., & Sarkar, A. K. (2016). Application of Principal Component Analysis for Outlier Detection in Heterogeneous Traffic Data. *Procedia Computer Science, 83*, 107–114.

140. Sajesh, T. A., & Srinivasan, M. R. (2013). An Overview of Multiple Outliers in Multidimensional Data. *Sri Lankan Journal of Applied Statistics, 14*(2), 87–120.

141. Sarkar, S. K., Midi, H., & Rana, S. (2011). Detection of outliers and influential observations in binary logistic regression: an empirical study. *Journal of Applied Sciences, 11*(1), 26-35.

142. Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis*: *An introductory guide.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-057. Newbury Park, CA: Sage.

143. Seber, G. A. F., & Lee, A. J. (2002). *Linear Regression Analysis (2nd edition).* New Jersey: John Wiley & Sons, Inc.

144. Seke, K., Petrović, N., Jeremić, V., Vukmirović, J., Kilibarda, B., & Martić, M. (2013). Sustainable development and public health: rating European countries. *BMC Public Health, 13*, 1-7.

145. Sen, A., & Srivastava, M. (1990). *Regression analysis: Theory, methods, and applications*. New York: Springer.

146. Shiffler, R. E. (1988). Maximum Z-Score and outliers. *The American Statistician, 42*(1), 79-80.

147. Smith, G. (2018). Step away from stepwise. *Journal of Big Data, 5*(32), 1-12.

148. Spearman, C. (1904). “General inteligence” objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

149. Stevens, J. P. (1984). *Outliers and influential data points in regression analysis.* Psychological Bulletin, 95(2), 334-344.

150. Stevens, J. P. (1996). *Applied multivariate statistics for the social sciences (3rd edition).* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

151. Sutter, J. M., & Kalivas, J. H. (1993). Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchemical Journal, 47,* 60-66.

152. Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (5th edition).* Boston: Pearson Education, Inc.

153. Thennadil, N. S., Dewar, M., Herdsman, C., Nordon, A., & Becker, E. (2018). Automated weighted outlier detection technique for multivariate data. *Control Engineering Practice, 70,* 40–49.

154. Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. *Education and Psychological Measurement, 55*(4), 525-534.

155. Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis - Understanding Concepts and Applications*. Washington: American Psychological Association.

156. Todeschini, R., Ballabio, D., Consonni, V., Sahigara, F., & Filzmoser, P. (2013). Locally centred Mahalanobis distance: A new distance measure with salient features towards outlier detection. *Analytica Chimica Acta, 787,* 1– 9.

157. Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesely.

158. Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). *Construct Explication through Factor or Component Analysis: A Review and Evaluation of Alternative Procedures for Determining the Number of Factors or Components.* In R. D. Goffin & E. Helmes (Eds.), *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy,* 41-71. Boston: Kluwer Academic Publishers.

159. Vidal, R., Ma, Y., & Sastry, S. S. (2016). *Generalized Principal Component Analysis*. New York: Springer.

160. Vuković, N. & Bulajić, M. (2014). *Osnove statistike.* Beograd: Fakultet organizacionih nauka.

161. Warren, R., Smith, R. E., & Cybenko, A. K. (2011). Use of Mahalanobis distance for detecting outliers and outlier clusters in markedly non-normal data: a vehicular traffic example. *Air Forse Research Laboratory*, United Stater Air Force – Intern Report.

162. Weisberg, S. (2005). *Applied linear regression (3rd edition).* New Jersey: John Wiley & Sons, Inc.

163. Withaker, J. S. (1997). *Use of Stepwise Methodology in Discriminant Analysis.* Annual Meeting of the Southwest Educational Research Association, Austin, TX.

164. Yan, X., & Su, X. (2009). *Linear Regression Analysis: Theory and Computing.* Singapore: World Scientific Publishing Co. Pte. Ltd.

165. Yang, Y. S. (2015). Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics. Doctoral dissertation, PhD thesis, UC: Berkley.

166. Yu, T., Yu, G., Li, PY., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics 101*(2), 1233-1252.

ПРИЛОЗИ

Табела А1: Индикатори економске развијености земаља Европске уније

Земља	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Аустрија	84.1800	52.262	-7.325	0.300	50,521.48	1.089	6.014
Белгија	64.487	82.938	6.143	0.900	46,416.85	1.584	7.830
Бугарска	52.766	63.979	1.589	4.673	19,242.62	2.246	7.575
Хрватска	60.509	49.013	-0.556	3.888	23,710.41	-0.084	13.103
Кипар	226.125	64.710	5.758	2.561	32,707.87	-0.684	12.950
Чешка	51.312	79.541	1.607	2.396	34,749.21	1.236	3.951
Данска	169.935	53.584	5.723	1.171	49,029.02	-0.046	6.181
Естонија	71.935	78.981	0.812	2.034	29,743.34	1.593	6.762
Финска	94.760	36.033	7.665	1.842	43,378.15	0.791	8.818
Француска	97.617	30.155	2.612	0.784	41,357.99	0.181	10.057
Немачка	77.225	46.119	2.623	1.124	48,943.10	1.329	4.122
Грчка	108.782	30.459	-0.818	0.171	26,764.96	-0.956	23.539
Мађарска	34.123	89.537	53.165	2.515	26,700.76	0.959	5.115
Ирска	47.450	121.582	33.524	3.961	71,388.76	0.021	8.366
Италија	85.692	29.802	0.820	1.030	38,380.17	0.848	11.688
Летонија	67.640	60.042	0.888	3.147	25,586.29	0.274	9.643
Литванија	42.934	74.451	1.898	3.654	29,862.32	0.951	7.862
Луксембург	102.387	221.268	53.969	0.885	102,389.44	-1.309	6.291
Малта	84.506	136.109	-59.357	2.850	37,363.33	1.628	4.900
Холандија	115.734	82.449	36.689	1.667	50,538.61	0.584	5.800
Пољска	54.560	52.260	2.370	2.909	27,383.26	0.416	6.161
Португал	111.840	40.129	2.842	1.940	30,658.63	1.518	11.066
Румунија	28.176	41.335	0.669	5.423	23,050.01	2.052	5.901
Словачка	56.999	94.625	4.585	3.192	30,460.38	-0.448	9.670
Словенија	46.671	77.654	1.062	3.073	32,723.07	0.896	8.000
Шпанија	111.340	32.947	4.079	3.187	36,305.22	0.283	19.635
Шведска	128.822	44.273	0.349	1.946	48,904.55	1.593	6.990
Велика Британија	134.256	28.255	1.977	1.211	42,656.22	1.974	4.813

1 - Домаћи кредити у приватном сектору (у процентима БДП-а), 2 - Извоз добара и услуга (у процентима БДП-а), 3 - Стране директне инвестиције, нето одлив (у процентима БДП-а), 4 - Раст БДП-а (годишњи проценат), 5 - БДП по глави становника (у доларима), 6 – Инфлација, потрошачке цене (годишњи проценат), 7 - Незапосленост (проценат укупног радно способног становништва).

Табела А2: Значајност индикатора у другом кораку – 6 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Извоз добара и услуга	0.792	0.627	5.890	0.00099
2.	БДП	0.652	0.425	2.588	0.04896
3.	Инфлација	0.408	0.166	0.699	0.65345
4.	Стране директне инвестиције	0.333	0.111	0.437	0.84614
5.	Домаћи кредити у приватном сектору	0.153	0.023	0.084	0.99727
6.	Раст БДП-а	0.061	0.004	0.013	0.99999

Табела А3: Значајност индикатора у трећем кораку – 5 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Извоз добара и услуга	0.818	0.669	8.898	0.00010
2.	БДП	0.674	0.454	3.663	0.01460
3.	Инфлација	0.431	0.186	1.004	0.43870
4.	Стране директне инвестиције	0.342	0.117	0.583	0.71285
5.	Домаћи кредити у приватном сектору	0.191	0.036	0.167	0.97223

Табела А4: Значајност индикатора у четвртом кораку – 4 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Извоз добара и услуга	0.836	0.699	13.346	0.00001
2.	БДП	0.699	0.489	5.494	0.00296
3.	Инфлација	0.376	0.141	0.947	0.45499
4.	Стране директне инвестиције	0.333	0.111	0.717	0.58887

Табела А5: Значајност индикатора у петом кораку – 3 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Извоз добара и услуга	0.757	0.573	10.738	0.00012
2.	БДП	0.724	0.524	8.813	0.00041
3.	Инфлација	0.436	0.190	1.878	0.16032

Табела А6: Индикатори развијености пољопривреде

Дистрикт	*1*	*2*	*3*	*4*	*5*
Берат	6,923,070.00	68,852.00	83,420.00	1,018,950.00	4,279,420.00
Булкиза	907,410.00	0.00	25,500.00	203,150.00	645,850.00
Делвина	64,006,630.00	0.00	0.00	1,651,000.00	6,187,050.00
Девол	2,625,370.00	0.00	80,900.00	610,800.00	2,162,880.00
Дибар	8,889,312.00	0.00	114,450.00	1,019,725.00	2,181,840.00
Драч	14,481,453.00	26,700.00	130,870.00	3,788,300.00	7,395,765.00
Елбасан	9,407,185.00	275,100.00	50,350.15	2,417,880.15	6,007,090.00
Фиер	33,024,524.00	5,010.00	445,700.00	7,184,535.00	20,020,098.00
Грамш	6,023,530.00	161,400.00	26,800.00	703,760.00	1,773,720.00
Ђирокастра	159,603,901.00	95,000.00	1,010.00	4,990,600.00	31,964,282.00
Хас	182,200.00	0.00	0.00	5,500.00	47,000.00
Каваја	7,037,341.00	0.00	349,700.00	2,657,500.00	6,279,884.00
Колоња	4,555,795.00	0.00	14,325.00	732,675.00	2,712,220.00
Корча	25,418,209.00	80,200.00	1,447,345.00	5,646,335.00	22,317,610.00
Круја	3,221,525.00	0.00	31,450.00	1,247,900.00	2,938,860.00
Кучова	2,753,658.00	200.00	9,200.00	291,730.00	770,880.00
Кукеш	1,286,890.00	9,500.00	20,525.00	54,050.00	713,000.00
Курбин	2,067,690.00	32,000.00	46,090.00	764,590.00	1,536,190.00
Лежје	5,564,551.00	700	78,300.00	2,191,420.00	4,555,145.00
Либражд	1,363,090.00	0	15,880.00	284,935.00	711,255.00
Лушње	38,924,355.00	1,749,086.00	430,370.00	8,671,780.00	31,136,895.00
Маласи и Маде	4,764,584.00	3,040,889.00	14,060.00	159,640.00	3,704,079.00
Малакастра	1,818,558.00	0.00	27,420.00	218,190.00	1,093,439.00
Мат	807,800.00	750.00	400.00	148,500.00	418,500.00
Мирдита	203,630.00	0.00	1,440.00	36,540.00	102,470.00
Пекин	1,011,450.00	2,950.00	7,565.00	273,315.00	612,250.00
Пармет	3,783,819.00	0.00	20,900.00	679,670.00	3,456,039.00
Пограделц	1,058,319.00	0.00	19,140.00	314,820.00	850,927.00
Пука	100,950.00	0.00	650.00	21,950.00	46,615.00
Саранда	133,482,180.00	0.00	15,000.00	2,137,100.00	14,907,240.00
Скрапар	1,827,160.00	1,500.00	23,450.00	525,780.00	1,561,190.00
Скадар	6,268,110.00	40,000.00	16,080.00	2,630,010.00	5,095,740.00
Тепелена	12,303,093.00	20,650.00	30,930.00	630,958.00	3,391,886.00
Тирана	7,508,213.00	300.00	34,700.00	3,587,910.00	6,127,331.00
Тропоја	986,550.00	0.00	35,950.00	8,000.00	251,950.00
Валона	39,084,040.00	3,600.00	10,400.00	4,693,000.00	19,857,100.00

1 - Искоришћена пољопривредна површина, 2 – Производња ндустријских биљака, 3 – Производња сена, 4 - Укупна жетва, 5 - Укупна обрадива површина

Табела А7: Значајност индикатора у другом кораку – 4 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Производња сена	0.721	0.520	8.390	0.00010
2.	Производња индустријсих биљака	0.677	0.458	6.558	0.00060
3.	Обрадива површина	0.538	0.289	3.157	0.02748
4.	Укупна жетва	0.464	0.215	2.126	0.10119

Табела А8: Индикатори перформанси кошаркаша

Играч	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Бен Симонс	16.9	0.0	0.0	8.8	7.7	3.5	1.4	0.8	2.6
Крис Пол	15.6	2.2	35.8	4.6	8.2	2.6	2.0	0.3	2.5
Колин Секстон	16.7	1.5	40.2	2.9	3.0	2.3	0.5	0.1	2.3
Д. Ј. Аугустин	11.7	1.6	42.1	2.5	5.3	1.6	0.6	0.0	1.4
ДиАнђело Расел	21.1	2.9	36.9	3.9	7.0	3.1	1.2	0.2	1.7
Дејмијан Лилард	25.8	3.0	36.9	4.6	6.9	2.7	1.1	0.4	1.9
Дарен Колисон	11.2	1.0	40.7	3.1	6.0	1.6	1.4	0.1	1.8
ДеАрон Фокс	17.3	1.1	37.1	3.8	7.3	2.8	1.6	0.6	2.5
Денис Смит	13.6	1.3	32.2	2.9	4.8	2.9	1.3	0.4	2.4
Донован Мичел	23.8	2.4	36.2	4.1	4.2	2.8	1.4	0.4	2.7
Џорџ Хил	7.6	0.8	31.4	2.5	2.3	0.9	0.9	0.1	1.7
Горан Драгић	13.7	1.6	34.8	3.1	4.8	2.0	0.8	0.1	2.3
Џамал Мареј	18.2	2.0	36.7	4.2	4.8	2.1	0.9	0.4	2.0
Џејмс Харден	36.1	4.8	36.8	6.6	7.5	5.0	2.0	0.7	3.1
Џеф Тег	12.1	0.8	33.3	2.5	8.2	2.3	1.0	0.4	2.1
Џон Вол	20.7	1.6	30.2	3.6	8.7	3.8	1.5	0.9	2.2
Џејру Холидеј	21.2	1.8	32.5	5.0	7.7	3.1	1.6	0.8	2.2
Кемба Вокер	25.6	3.2	35.6	4.4	5.9	2.6	1.2	0.4	1.6
Крис Дан	11.3	0.7	35.4	4.1	6.0	2.3	1.5	0.5	3.6
Кајл Лаури	14.2	2.4	34.7	4.8	8.7	2.8	1.4	0.5	2.6
Кајри Ирвинг	23.8	2.6	40.1	5.0	6.9	2.6	1.5	0.5	2.5
Лонзо Бол	9.9	1.6	32.9	5.3	5.4	2.2	1.5	0.4	2.4
Лу Вилијамс	20.0	1.4	36.1	3.0	5.7	2.4	0.8	0.1	1.1
Лука Дончић	21.2	2.3	32.7	7.8	6.0	3.4	1.1	0.3	1.9
Мајк Конли	21.1	2.2	36.4	3.4	6.4	1.9	1.3	0.3	1.8
Пети Милс	9.9	1.9	39.4	2.2	3.0	1.1	0.6	0.1	1.6
Реџи Џексон	15.4	2.1	36.9	2.6	4.2	1.8	0.7	0.1	2.5
Расел Вестбрук	22.9	1.6	29.0	11.1	10.7	4.5	1.9	0.5	3.4
Стеф Кари	27.3	5.1	43.7	5.3	5.2	2.8	1.3	0.4	2.4
Трај Јанг	19.1	1.9	32.4	3.7	8.1	3.8	0.9	0.2	1.7

1 - Просечан број поена, 2 - Просечан број погођених тројки, 3 - Просечан шут за три поена, 4 - Просечан број скокова, 5 - Просечан број асистенција, 6 - Просечан број изгубљених лопти, 7 - Просечан број украдених лопти, 8 - Просечан број блокада, 9 - Просечан број фаулова.

Табела А9: Значајност индикатора у другом кораку – 8 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Просечан шут за три поена	0.764	0.584	3.680	0.00784
2.	Просечан број скокова	0.636	0.404	1.783	0.13752
3.	Просечан број изгубљених лопти	0.371	0.138	0.419	0.89661
4.	Просечан број блокада	0.334	0.112	0.330	0.94497
5.	Просечан број фаулова	0.318	0.101	0.295	0.95962
6.	Просечан број асистенција	0.254	0.065	0.181	0.99106
7.	Просечан број погођених тројки	0.229	0.052	0.145	0.99571
8.	Просечан број украдених лопти	0.199	0.040	0.108	0.99846

Табела А10: Значајност индикатора у трећем кораку – 7 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Просечан шут за три поена	0.782	0.612	4.947	0.00177
2.	Просечан број скокова	0.638	0.407	2.157	0.07956
3.	Просечан број изгубљених лопти	0.369	0.136	0.495	0.82772
4.	Просечан број блокада	0.349	0.122	0.436	0.86898
5.	Просечан број фаулова	0.310	0.096	0.334	0.92967
6.	Просечан број асистенција	0.257	0.066	0.222	0.97592
7.	Просечан број погођених тројки	0.237	0.056	0.187	0.98523

Табела А11: Значајност индикатора у четвртом кораку – 6 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Просечан шут за три поена	0.809	0.654	7.261	0.00019
2.	Просечан број скокова	0.628	0.394	2.496	0.05228
3.	Просечан број изгубљених лопти	0.340	0.116	0.501	0.80102
4.	Просечан број блокада	0.325	0.106	0.453	0.83558
5.	Просечан број фаулова	0.287	0.082	0.344	0.90607
6.	Просечан број асистенција	0.264	0.070	0.287	0.93703

Табела А12: Значајност индикатора у петом кораку – 5 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Просечан шут за три поена	0.823	0.677	10.076	0.00003
2.	Просечан број скокова	0.627	0.393	3.109	0.02649
3.	Просечан број изгубљених лопти	0.349	0.122	0.666	0.65301
4.	Просечан број блокада	0.345	0.119	0.649	0.66532
5.	Просечан број фаулова	0.274	0.075	0.390	0.85099

Табела А13: Значајност индикатора у шестом кораку – 4 индикатора

РБ	*Индикатор*	*R*	R^2	*F*	*p*
1.	Просечан шут за три поена	0.858	0.736	17.439	0.00000
2.	Просечан број скокова	0.591	0.349	3.355	0.02495
3.	Просечан број блокада	0.371	0.138	0.998	0.42732
4.	Просечан број изгубљених лопти	0.324	0.105	0.733	0.57804

БИОГРАФИЈА

Немања Миленковић је рођен 25.6.1986. године у Београду. Основну школу "Бранислав Нушић" и Дванаесту београдску гимназију "Димитрије Туцовић" је завршио у Београду. Факултет организационих наука уписује 2005. године на одсеку Информациони системи и технологије. Дипломирао је у јуну 2010. године са оценом 10 на дипломском раду са називом *"Основне поставке научног рада професора доктора Бранислава Ивановића"*. Школске 2010/2011 је уписао дипломске академске студије – Мастер на Факултету организационих наука, смер Операциона истраживања и рачунарска статистика, студијска група – Рачунарска статистика. Завршни мастер рад на тему *"Теорија и пракса Ивановићевог одстојања"*, одбранио је у јануару 2012. године.

Докторске студије уписао је школске 2014/2015 године на Факултету организационих наука на студијском програму Информациони системи и квантитативни менаџмент – изборно подручје Операциона истраживања. Положио је све испите са просечном оценом 10. У октобру 2018. године одбранио је приступни рад за израду дисертације под називом „Методологија откривања нестандардних опсервација у *k*-димензионом простору".

У радни однос на Факултету организационих наука као стручни сарадник ступио је новембру 2011. године. У априлу 2013. године изабран је у звање сарадника у настави за ужу научну област Рачунарска статистика. У априлу 2015. године изабран је у звање асистента за ужу научну област Рачунарска статистика.

Од маја до септембра 2018. године учествовао је на пројекту Европске уније "Support for the Improvement of Statistical Information System – Albania (EuropeAid/136334/IH/SER/AL)" као експерт за пројектовање и имплементацију статистичког регистра пољопривредних газдинстава.

На Thomson Reuters листи, његов идентификациони истраживачки број је O-2082-2013.

Објавио је значајан број научних и стручних радова од којих се посебно истичу:

- **Milenkovic, N.,** Vukmirovic, J., Bulajic, M., & Radojicic, Z.: *A multivariate approach in measuring socio-economic development in MENA countries.* Economic Modelling, Vol 38, 2014, pp. 604-608, (ISSN: 0264-9993), (IF 2012-0.557), (5-Year Impact Factor: 0.699), **M23.** doi: 10.1016/j.econmod.2014.02.011
- Radaković, J. A., Petrović, N., **Milenković, N.,** Stanojević, K., & Đoković, A.: *Improving Students' Higher Environmental and Climate Change Knowledge: A Case Study,* Polish Journal of Environmental Studies, Vol 26, Issue 6, 2017, pp. 2711-2719, (ISSN: 1230-1485), (IF 2016-0.793), (5-Year Impact Factor: 0.961), **M23.** doi: 10.15244/pjoes/69645
- **Milenković, N.,** Đoković, A., & Vukmirović, A.*: Europe 2020 Strategy – A multivariate approach,* XV International Symposium SymOrg 2016: Reshaping the future through sustainable business development and enterpreneurship, Zlatibor, 2016, pp. 228-233, (ISBN: 978-86-7680-326-2), **M33.**
- Ćirović, M., Delibašić, B., Petrović, N., Makajić-Nikolić, D. & **Milenković, N.**: *A ski Slopes Injury Risk Evaluation based on FMEA method,* 33th International Conference on Organizational Science Development, Portorož, Slovenia 2014, **M34.**

Рад „*A Ski Slopes Injury Risk Evaluation based on FMEA method*" проглашен је за најбољи рад на конференцији „33th International Conference on Organizational Science Development", Portorož, Slovenia 2014.

Изјава о ауторству

Име и презиме аутора: Немања Миленковић

Број индекса: 2014/5001

Изјављујем

да је докторска дисертација под насловом

Методологија откривања нестандардних опсервација у *k*–димензионом простору

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио ауторска права и користио интелектуалну својину других лица.

Потпис аутора

У Београду, _____

_____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора: Немања Миленковић

Број индекса: 2014/5001

Студијски програм: Информациони системи и квантитативни менаџмент

Наслов рада: Методологија откривања нестандардних опсервација у *k*-димензионом простору

Ментор: др Зоран Радојичић

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, _____

_____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Методологија откривања нестандардних опсервација у *k*-димензионом простору

која је моје ауторско дело.

Дисертацију са свим прилозима предао сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио.

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

Потпис аутора

У Београду, _____

_____