УНИВЕРЗИТЕТ У КРАГУЈЕВЦУ
ЕКОНОМСКИ ФАКУЛТЕТ

Милан Стаменковић

МУЛТИВАРИЈАЦИОНО СТАТИСТИЧКО МОДЕЛИРАЊЕ У ФУНКЦИЈИ МЕРЕЊА СТЕПЕНА ЕКОНОМСКЕ РАЗВИЈЕНОСТИ ТЕРИТОРИЈАЛНИХ ЈЕДИНИЦА

Докторска дисертација

Крагујвац, 2019. године

ИДЕНТИФИКАЦИОНА СТРАНИЦА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

I Аутор
Име и презиме: Милан Стаменковић
Датум и место рођења: 18.09.1983. године у Крагујевцу
Садашње запослење: Финансијско-рачуноводствени аналитичар
II Докторска дисертација
Наслов: Мултиваријационо статистичко моделирање у функцији мерења степена економске развијености територијалних јединица
Број страница: 254
Број табела: 67; број слика: 50
Број библиографских јединица: 209+1
Установа и место где је рад израђен: Економски факултет Универзитета у Крагујевцу
Научна област (УДК): 519.2:330.45(043.3)
Ментор: Проф. др Мирко Савић, Економски факултет, Универзитет у Новом Саду
III Оцена и одбрана
Датум пријаве теме: 06.06.2017. године
Број одлуке и датум прихватања теме докторске/уметничке дисертације: IV-02-938/8, од 11.10.2017. године
Комисија за оцену научне заснованости теме и испуњености услова кандидата: 1. Др Петар Веселиновић, редовни професор Економског факултета Универзитета у Крагујевцу, ужа научна област Општа економија и преивредни развој 2. .Др Весна Јанковић-Милић, ванредни професор Економског факултета Универзитета у Нишу, ужа научна област Економска статистика, примена математичких и статистичких метода у економским истраживањима 3. Др Предраг Мимовић, редовни професор Економског факултета Универзитета у Крагујевцу, ужа научна област Статистика и информатика
Комисија за оцену и одбрану докторске/уметничке дисертације:
Датум одбране дисертације:

С А Д Р Ж А Ј

Апстракт

У докторској дисертацији су разматрана суштинска теоријска одређења одабраних мултиваријационих статистичких метода међузависности и зависности и сагледани њихови апликативни потенцијали за моделирање комплексних, мултидимензионих економских феномена од интереса, чиме је истовремено опредељен предмет истраживања. Афирмацију примене мултиваријационих статистичких метода у домену економских истраживања одражава основни циљ дисертације, који подразумева креирање иновативног концептуално-методолошког оквира, заснованог на статистички валидној имплементацији како појединачних метода мултиваријационе анализе, тако и њихове комбинације на одређеном броју релевантних показатеља, у функцији мерења достигнутог степена економске развијености и, сходно томе, класификације територијалних јединица локалне самоуправе у Републици Србији.

У том смислу, у оквиру сваке методе детаљно су анализирани циљеви, типови и поступак спровођења, уз јасно разграничење истраживачких околности под којима се њихова примена сматра прикладном и статистички оправданом. На темељима важности адекватне припреме података за спровођење било које анализе података у контексту обезбеђивања научне заснованости добијених резултата и изведених закључака, посебна пажња је посвећена претпроцесирању мултиваријационих опсервација и елаборирању значаја испуњености статистичких претпоставки из перспективе валидне примене конкретне методе. Расветљавање комплексног и значајног питања валидације квалитета резултата мултиваријационог моделирања и, с тим у вези, проналажења „статистичких“ аргумента за избор оптималног решења, извршено је анализом бројних критеријума и метода за евалуацију резултата.

У емпиријском делу дисертације представљена су два оригинална концептуално-методолошка оквира анализе мултиваријационих података, и то: први, заснован на интегрисаној примени факторске анализе, анализе груписања и мултиваријационе анализе варијансе у функцији развоја мултиваријационог модела (форма композитног показатеља) за мерење степена економске развијености и класификацију јединица локалне самоуправе у Републици Србији, и, други, заснован на примени дискриминационе анализе у функцији развоја класификационог модела за разврставање анализираних територијалних јединица у једну од, према вредностима претходно утврђеног композитног показатеља степена економске развијености, емпиријски идентификованих група. Резултати истраживања указују на велики потенцијал комбиноване имплементације мултиваријационих статистичких метода у конципирању иновативних методолошких решења за анализу и разумевање економских феномена.

***Кључне речи**: мултиваријациона статистичка анализа, факторска анализа, анализа груписања, мултиваријациона анализа варијансе, дискриминациона анализа, композитни показатељ, степен економске развијености, регионални диспаритети*

Abstract

In this doctoral dissertation, the essential theoretical determinations of selected multivariate statistical methods of interdependence and dependence were examined, as well as their application potentials for modeling complex, multidimensional economic phenomena of interest were considered, which simultaneously defined the research subject. The affirmation of the application of multivariate statistical methods in the domain of economic research reflects the primary objective of the dissertation, which implies the development of an innovative conceptual-methodological framework, based on statistically valid implementation of individual methods of multivariate analysis, as well as their combinations on a number of relevant indicators in function of measuring the achieved degree of economic development and, accordingly, the classification of local self-government territorial units in the Republic of Serbia.

In this sense, within each method, the objectives, types and implementation procedures have been thoroughly analyzed, with a clear distinction of research circumstances under which their application is considered appropriate and statistically justified. On the basis of the importance of adequate data preparation for implementation of any data analysis, in terms of ensuring the scientific basis of the obtained results and conclusions drawn, special attention has been devoted to preprocessing of multivariate observations and elaboration of importance of fulfilling statistical assumptions from the perspective of valid application of particular method. The clarification of complex and significant question of validating the quality of multivariate modeling results and, in this regard, finding "statistical" arguments for choosing the optimal solution, was done by analyzing a number of different criteria and methods for evaluation of results.

Within the empirical part of the dissertation, the following two original conceptual-methodological frameworks of multivariate data analysis were presented: first, based on the integrated implementation of factor analysis, cluster analysis and multivariate analysis of variance in function of developing a specific multivariate model (in form of a composite indicator) for measuring the degree of economic development and classification of local self-government units in the Republic of Serbia, and, second, based on the implementation of discriminant analysis in function of developing a classification model for allocation of analyzed territorial units into one of the empirically identified groups, according to the values of the previously proposed composite indicator of degree of economic development. The results of the conducted research indicate the great potential of the combined implementation of multivariate statistical methods in conceiving innovative methodological solutions for the analysis and understanding of economic phenomena.

***Key words**: multivariate statistical analysis, factor analysis, cluster analysis, multivariate analysis of variance, discriminant analysis, composite indicator, degree of economic development, regional disparities*

СПИСАК СЛИКА

СПИСАК ТАБЕЛА

У В О Д

Један од најважнијих, али и најкомплекснијих друштвено-економских проблема са којим се креатори развојних политика и представници државних институција данас суочавају односи се на неравномерности у развоју између дефинисаних административно-територијалних целина, односно јединица различитог нивоа територијалног обухвата у саставу једне конкретне државе (*Rovan & Sambt*, 2003, стр. 265; *Maletić & Bucalo-Jelić*, 2016, стр. 13). Присуство регионалних разлика у погледу достигнутог степена развијености, посматрано из угла економске, друштвене (социјалне), еколошке или, пак, неке друге димензије развоја, својствено је како развијеним тако и, мада у већем интензитету, земљама у развоју (*Mohiuddin & Hashia*, 2012, стр. 86; *Miljačić & Paunović*, 2011, стр. 380). Полазећи од тога да је економски развој региона основа за реализацију националних економских циљева (*Jakopin*, 2015, стр. 109), остваривање интензивног раста и одрживог економског развоја земље и њене привреде нужно подразумева уважавање концепта регионалне једнакости, односно предузимање активности усмерених на уравнотежење и / или повећање степена развијености свих њених региона, а тиме и благостања свих њених становника (*Rovan & Sambt*, 2003, стр. 265; *Istrate & Horea-Serban*, 2016, стр. 209). У том контексту, потпуно је оправдана констатација да стварање услова за успостављање равномерног регионалног развоја представља приоретни задатак сваке државе и кључни корак у настојањима да се обезбеди успешна интеграција националне економије у глобалне економске токове (РЗР, 2009, стр. 88; *Krstić & Vukadinović*, 2011, стр. 554). Консеквентно, разматрање различитих аспеката и фактора равномерног регионалног развоја све више заокупља и истраживачку пажњу представника научне и стручне заједнице, о чему сведочи велики број публикованих радова и спроведених емпиријских студија.

Сагласно мишљењу бројних аутора (попут, *Abdollahzade & Sharifzadeh*, 2012, стр. 9; *Lukić & Anđelković-Stoilković*, 2017, стр. 66; *Puljiz & Maleković*, 2007, стр. 1), према којима се транзиционе земље и земље у развоју издвајају као посебно осетљиве на интензивирање проблема регионалних разлика, и Републику Србију карактеришу веома изражени међурегионални и унутаррегионални диспаритети и асиметричности у погледу степена развијености, са тенденцијом њиховог континуираног повећања (Влада РС, 2007а, стр. 1; *Winkler*, 2012, стр. 82; *Vukmirović*, 2013, стр. 39). Након вишедеценијског занемаривања питања регионализације у економској политици, а делимично и научним круговима (*Molnar*, 2013, стр. 66), значајан корак у признавању важности и неопходности обезбеђивања равномерног и одрживог регионалног развоја у Републици Србији, а тиме и ралізацији уставних надлежности државе из 2006. године, учињен је 2007. и 2009. године, доношењем *Стратегије регионалног развоја Републике Србије* (Влада РС, 2007а) и *Закона о регионалном развоју* (Влада РС, 2009а). Тиме је постављен адекватан институционални оквир за примену принципа и механизама управљања регионалним развојем на основу савремених (европских) концепција.

Објективно мерење достигнутог степена развијености и категоризација територијалних јединица (региона, субрегиона (управних округа) и / или јединица локалних самоуправа), уз сагледавање различитих аспеката (димензија) њихових

развојних специфичности (потенцијала и ограничења), представља важан извор информација које имају кључну улогу у поступку стратешког планирања равномерног регионалног развоја и ефикасног / ефективног спровођења политике регионалног развоја. У литератури се као главне и најчешће коришћене, за потребе квантификовања размера присутних регионалних диспаритета, углавном издвајају следеће развојне димензије: економска, друштвена (или социјална), еколошка, инфраструктурна и демографска (и / или образовна). При томе, свака од наведених димензија може бити посматрана као засебна мултидимензиона латентна променљива, чије се „мерење“ по правилу спроводи индиректно, на бази симултане анализе вредности и веза између више, из угла конкретне димензије, репрезентативних, директно мерљивих, нумеричких показатеља (Влада РС, 2007а, стр. 86; *Polednikova*, 2014, стр. 498). Не умањујући значај других димензија, при оцени нивоа регионалне развијености, неопходно је, ипак, апострофирати доминантну улогу економске димензије и њених показатеља (*Rovan & Sambt*, 2003, стр. 265; *Bojović*, 2010, стр. 10), путем којих се најбоље илуструју размере регионалних неравномерности. У начелу, одлука о избору једне или више развојних димензија и њихових кореспондентних показатеља условљена је дефинисаним циљевима истраживања, доступношћу података, а нарочито административно-територијалним нивоом јединица посматрања. Такође, полазећи од правилности да се регионалне асиметрије повећавају што је ниво посматрања нижи, потребно је истаћи да регионална истраживања спроведена на нижим нивоима територијалне агрегације обезбеђују (нај)бољи увид у величину асиметрије и регионалне (не)развијености (НАРР, 2012, стр. 23; *Mohiuddin & Hashia*, 2012, стр. 87).

Сходно апострофираној мултидимензионалности концепта регионалне развијености и појединачних развојних димензија, квантификовање достигнутог нивоа развијености територијалних јединица, са концептуално–методолошког становишта, представља захтеван и „тежак“ задатак (*Meyer et al.*, 2016, стр. 101). Његова реализација у условима вишеструке мултидимензионалности, условила је померање аналитичког оквира од једнодимензионог праћења вредности великог броја појединачних показатеља различитих развојних димензија ка развоју и примени разноврсних методолошких поступака заснованих на експлоатацији апликативних потенцијала метода мултиваријационе статистичке анализе (*Polednikova*, 2014, стр. 499; *Melecky*, 2012, стр. 979; *Meyer et al.*, 2016, стр. 101). Употребљени појединачно или у комбинацији, наведени статистички методи омогућавају мерење степена развијености конкретних територијалних јединица (углавном кроз развој одговарајућих композитних показатеља), њихову класификацију у интерно-хомогене / екстерно-хетерогене групе (сходно расположивим развојним потенцијалима и ограничењима) и / или предвиђање њихове припадности конкретној развојној категорији, а на бази вредности анализираних показатеља.

Генерално, последњих деценија бележи се широка примена метода мултиваријационе статистичке анализе скоро у свим научним областима (укључујући и проблематику у разматрању), и то не само као последица развоја компјутерске технике и софтверских производа, већ и потребе за симултаним анализирањем међузависности више од две променљиве и превазилажењем оних ограничења која су својствена једнодимензионом и дводимензионом приступу у анализи података. У складу са тим, веома је важно истаћи следеће три констатације:

(а) методе мултиваријационе статистичке анализе поседују изузетан апликативни потенцијал у истраживању структуре и моделирању релација унутар мултидимензионих економских феномена, а тиме и у домену истраживања регионалних неравномерности;

(б) мерење степена економске развијености територијалних јединица засновано на развоју и употреби композитног показатеља обезбеђује бољи и потпунији увид у размере присутних диспаритета у односу на алтернативни, униваријациони приступ појединачног посматрања већег броја засебних показатеља у контексту проблематке у разматрању;

(в) недовољна транспарентност, односно непотпуна методолошка спецификација спроведених аналитичких поступака, одсуство или некомплетност провере испуњености статистичких претпоставки и наглашена субјективност при евалуацији резултата представљају углавном присутне карактеристике не малог удела мултиваријационих истраживања спроведених у домену разматране проблематике.

Имајући у виду наведено, **предмет истраживања** у докторској дисертацији под насловом „Мултиваријационо статистичко моделирање у функцији мерења степена економске развијености територијалних јединица“ је анализа, у теоријском и апликативном смислу верификованих, метода мултиваријационе статистичке анализе и њихова шира афирмација у домену економских истраживања кроз испитивање и презентовање могућности њихове, статистички валидне и оправдане, појединачне и комбиноване, примене у моделирању релација између релевантних показатеља степена економске развијености јединица локалне самоуправе, односно територијалних јединица на нивоу општина.

Сходно дефинисаном предмету истраживања формулисан је и **основни циљ истраживања** који гласи: креирање концептуално-методолошког оквира заснованог на статистички валидној примени како појединачних метода мултиваријационе анализе тако и њихове комбинације на одређеном броју релевантних показатеља, у функцији обезбеђивања иновативног приступа мерењу достигнутог степена економске развијености и, сходно томе, класификацији територијалних јединица нивоа општина у Републици Србији.

Полазећи од дефинисаног основног циља и аналитичког карактера истраживања у оквиру докторске дисертације, прецизиран је и **сет посебних циљева истраживања**, који обухватају следеће:

- систематизовање концептуално–методолошких одређења одабраних метода мултиваријационе статистичке анализе података и анализирање њихових апликативних могућности у домену истраживања мултидимензионих економских феномена;
- креирање мултидимензионог статистичког модела, у форми одговарајућег композитног показатеља (Индекс Економске Развијености – ИЕР) за мерење достигнутог степена економске развијености територијалних јединица у саставу државе;
- креирање мултидимензионог класификационог модела који ће омогућити разврставање посматраних административно-територијалних јединица у одговарајуће интерно хомогене и екстерно хетерогене групе, на бази конкретних вредности одабраних појединачних показатеља степена економске развијености; и
- разматрање и истицање значаја темељне и адекватне провере испуњености претпоставки на којима се заснива статистички валидна и оправдана примена метода

мултиваријационе анализе података, као незаобилазног (мада углавном занемариваног) корака у обезбеђивању научне заснованости добијених резултата истраживања и формулисаних закључака.

Сходно опредељеном предмету и дефинисаним циљевима истраживања, формулисане су једна општа и, из ње изведене, три посебне хипотезе.

Општа хипотеза гласи: Статистички валидна примена одговарајуће комбинације метода мултиваријационе анализе одликује се високим практичним значајем и апликативним могућностима у домену моделирања, анализе и праћења мултидимензионих економских појава.

Посебне хипотезе су:

Хипотеза 1: Примена комбинације одабраних метода мултиваријационе анализе омогућава развој статистичког модела (у форми одговарајућег композитног индекса) који може допринети успешном и прецизном мерењу достигнутог степена економске развијености посматраних територијалних јединица у саставу државе.

Хипотеза 2: Примена комбинације одабраних метода мултиваријационе анализе омогућава развој одговарајућег класификационог статистичког модела који се може користити за прецизно и објективно разврставање посматраних територијалних јединица у одговарајуће интерно хомогене и екстерно хетерогене групе према достигнутом степену економске развијености.

Хипотеза 3: Методолошки валидна провера статистичких претпоставки у претпроцесирању података и употреба критеријума оптималности у екстракцији и евалуацији добијених резултата примењених мултиваријационих метода доприносе побољшању укупног квалитета креираних модела, а самим тим и повећању поузданости и објективности формулисаних закључака у вези са разматраним економским појавама.

За потребе реализације дефинисаних циљева и аргументације и евалуације постављених истраживачких хипотеза коришћена је квантитативна и квалитативна методологија истраживања карактеристична за област друштвених наука.

Рашчлањавање дефинисаног предмета дисертације на његове саставне делове, у циљу проучавања и стицања сазнања о сваком делу посебно, али и релацијама које постоје између њих, спроведено је коришћењем метода анализе. На тај начин је обезбеђено дефинисање и објашњење кључних појмова и концепата мултиваријационе анализе података (метод дескрипције), као и дубље схватање карактеристика и иманентних претпоставки мултиваријационих метода, са посебним осврtом на сагледавање њихових апликативних могућности у домену истраживања регионалне економске развијености територијалних јединица. Такође, коришћен је и метод синтезе, као посебно релевантан за повезивање постојећих теоријско-методолошких сазнања у функцији идејног осмишљавања специфичног комбиновања мултиваријационих метода у домену сагледавања регионалних неравномерности, али и за сумаризацију резултата и формулисање закључака дисертације. У теоријском и емпиријском делу истраживања, коришћени су и елементи индуктивног и дедуктивног закључивања.

Метод компарације је примењен за потребе критичког упоређивања истраживачких искустава и методолошких приступа из домена мерења степена регионалне развијености који су засновани на мултиваријационим методама, а у циљу идентификовања сличности и разлика, као и могућности за побољшање постојећих приступа и / или креирање

иновативног методолошког оквира. Поред наведеног, у емпиријском делу дисертације, исти научни метод је коришћен за сагледавање предности и ограничења резултата примене појединих метода у оквиру реализованог иновативног мултиваријационог статистичког моделирања.

У емпиријском делу истраживања је доминантна примена статистичког научног метода, која неизоставно инкорпорира елементе хипотетичко-дедуктивног метода и метода моделовања. Презентовање података и добијених (међу)резултата спроведене статистичке анализе извршено је коришћењем табеларно-графичког метода (метод визуелизације). За потребе реализације истраживања, коришћени су секундарни подаци изабраних показатеља економске развијености за сваку од територијалних јединица нивоа општина у Републици Србији, прикупљени из различитих, релевантних, званичних и периодичних публикација и електронских база података надлежних државних институција (*DeskResearch* метод). Статистичка анализа прикупљених података, неопходна израчунавања и визуелизација података су спроведени уз подршку следећих софтверских алата: *IBM SPSS Statistics* 20.0, *EduStat* 4.05, *Microsoft Office Excel*, *QI Macros for Excel* и *SYSTAT* 13.1.

Предмет и циљеви истраживања определили су структуру докторске дисертације, која је, поред увода и закључка, организована кроз теоријски и емпиријски део, и обухвата пет тематски заокружених и логички повезаних целина (Поглавља).

Прво поглавље под називом ***Концепцијски оквир и апликативни значај мултиваријационе статистичке анализе***, посвећено је разматрању концепцијских одређења мултиваријационе статистичке анализе података у циљу истицања њених специфичности у односу на једнодимензиони и дводимензиони приступ, као и представљању њеног значаја у разумевању и анализи сложених, мултидимензионих економских феномена. У наставку, истраживачка пажња је усмерена на класификацију широког спектра мултиваријационих метода. Будући да успешно спровођење мултиваријационе анализе, у контексту решавања конкретне проблемске ситуације, представља изразито комплексан концептуално–методолошки подухват, представљене су кључне фазе процеса креирања статистички валидног модела, као коначног исхода анализе. На темељима важности адекватне припреме података за спровођење било које анализе података у контексту обезбеђивања научне заснованости добијених резултата и изведених закључака, у наставку овог Поглавља, посебна пажња је посвећена кључним методолошким аспектима и специфичностима поступка претпроцесирања података мултидимензионе природе. Такође, сагледан је значај како метода визуелизације у приказивању мултиваријационих статистичких података и коначних резултата анализе, тако и развоја и примене софтверских алата за спровођење мултиваријационе анализе података, уз кратак осврт на етичке аспекте у вези са начином њиховог коришћења.

У оквиру наредна два поглавља пажња је фокусирана на разматрање концептуално-методолошких одређења изабране групе мултиваријационих метода. Избор метода извршен је са становишта њиховог доприноса реализацији дефинисаних циљева ове докторске дисертације, а у складу са класификацијом метода мултиваријационе анализе на методе међузависности и методе зависности.

Друго поглавље под називом ***Методолошка одређења одабраних мултиваријационих метода међузависности***, посвећено је детаљном приказу

методолошких аспеката примене факторске анализе и анализе груписања. При томе, најпре су представљене основне карактеристике, циљеви, типови, поступак спровођења и статистичке претпоставке факторске анализе. Даљим истраживањем у овом делу дисертације, посебан акценат је стављен на дефинисање једнофакторског модела факторске анализе и етапе у поступку његовог статистичког оцењивања. Са становишта тока анализе и квалитета изведених закључака, један од критичних корака у процесу креирања факторског модела односи се на доношење одлуке о „оптималном“ броју заједничких фактора који ће бити задржани у изведеном моделу. С тим у вези, елаборирана су суштинска одређења различитих критеријума за селекцију „оптималног“ броја фактора, уз кратак осврт на улогу и значај метода за ротацију фактора. Даљим излагањем указано је на широке апликативне могућности факторске анализе, уз апострофирање њене улоге у креирању композитних показатеља. Након факторске анализе усмерене на проналажење латентне структуре која описује однос између променљивих, у наставку истраживања презентована је анализа груписања, као мултиваријациони статистички метод намењен идентификовању „природног“ груписања анализом обухваћених јединица посматрања. Будући да је концепт блискости фундаментални концепт у поимању суштинских одређења анализе груписања, након дефинисања њених основних циљева и приказа поступка спровођења, пажња је фокусирана на дефинисање различитих мера блискости (сличности / различитости) између објеката. Сходно методолошкој варијететности анализе груписања, такође, сагледане су основне карактеристике метода хијерархијске и нехијерархијске процедуре груписања. На крају овог дела, детаљно је елаборирано питање евалуације квалитета резултата груписања, уз апострофирање важности статистички заснованих критеријума за избор оптималног решења, посматрано из угла броја и структуре формираних група.

Треће поглавље под називом ***Методолошка одређења одабраних мултиваријационих метода зависности***, посвећено је детаљној анализи методолошких аспеката примене мултиваријационе анализе варијансе (*MANOVA*) и дискриминационе анализе, као најчешће коришћених мултиваријационих метода усмерених на испитивање зависности између два раздвојена скупа (зависних и независних) променљивих. У оквиру сваке методе појединачно најпре су дефинисани циљеви, типови и поступак спровођења, уз разграничење истраживачких околности под којима се њихова примена сматра прикладном и статистички оправданом. Посебна пажња посвећена је претпроцесирању мултиваријационих опсервација и, консеквентно, елаборирању значаја испуњености статистичких претпоставки из угла њихове валидне примене ових метода. Даља разматрања у контексту *MANOVA* анализе (са фокусом на једнофакторски *MANOVA* модел) усмерена су на објашњење поступка статистичког закључивања у мултиваријационом контексту путем одговарајућих параметарских тестова за тестирање статистичке значајности разлика између вектора средина две, односно три или више мутиваријационих популација. Навођењем извесних смерницама везаних за примену и интерпретацију резултата *MANOVA* методе, уз апострофирање улоге накнадне анализе у форми вишеструке компарације, комплетиран је приказ општих и специфичних концептуално–методолошка својства ове методе. С друге стране, кроз даља разматрања у контексту дискриминационе анализе, у циљу расветљавања њених комплексних специфичности, најпре је, детаљно елаборирана централна статистичка активност која се односи на

креирање оцењеног дискриминационог модела (то јест, формирање једне или више одговарајућих линеарних дискриминационих варијата), а затим исцрпно анализирана његова статистичка и практична значајност. Сагласно са дефинисаним циљевима и извршеном поделом на дескрипривни и предиктивни аспект квантитативних активности у оквиру дискриминационе анализе, излагање у наставку посвећено је предиктивним аспектима дискриминационе анализе, као и статистичкој евалуацији предиктивне прецизности изведених класификационих правила.

Четврто поглавље, насловљено ***Регионалне неравномерности као изазов за примену мултиваријационе анализе података***, посвећено је сагледавању различитих аспеката концепта регионализације, који по својој природи, представља мултидимензиони феномен, а самим тим, у истраживачком смислу, атрактивно подручје за примену, у претходним поглављима елаборираних, мултиваријационих статистичких метода. Након сажетог осврта на значај регионализације и (не)равномерног регионалног развоја у контексту развоја националне економије, апострофирана је чињеница да дефинисање адекватног институционалног оквира на нивоу конкретне земље представља основу за отклањање или ублажавања развојних неравномерности. У том смислу, у наставку је указано на размере регионалних диспаритета и институционалне темеље изградње актуелног система спровођења регионалне политике у Републици Србији. У овом делу дисертације је, такође, представљена статистичка регионализација територијалног простора Републике Србије на територијалне јединице различитог нивоа посматрања, заснована на критеријумима *NUTS* класификационе методологије. Даља разматрања су усмерена на сагледавање значаја мерења степена развијености територијалних јединица, као основе за алокацију подстицајних средстава и других инструмената подршке од стране надлежних институција. Осим тога, наведене су различите димензија концепта регионалне развијености. При томе је апострофирана кључна улога економске димензије и показатеља економске развијености у оцени степена развијености територијалних јединица. Полазећи од мултидимензионе природе концепта регионалне развијености и појединачних развојних димензија, указано је на неопходност мултиваријационог приступа у сагледавању регионалних економских диспаритета и мерењу степена економске развијености. Детаљним прегледом релевантне литературе и спроведених емпиријских студија, које се односе на различите могућности и аспекте примене широког спектра метода мултиваријационе анализе у контексту сагледавања регионалних и унутар-регионалних неједнакости територијалних јединица различитог нивоа у различитим државама, уз критички осврт на представљена истраживања, комплетирано је излагање у оквиру овог поглавља.

Пето поглавље под називом ***Истраживање могућности и илустрација примене мултиваријационих метода у сагледавању степена економске развијености општина у Републици Србији*** је посвећено оригиналном емпиријском истраживању и односи се на конкретну примену комбинације одабраних метода мултиваријационе анализе у циљу развоја одговарајућих статистичких модела за мерење степена економске развијености јединица локалне самоуправе (градова / општина) у Србији и њихову класификацију на бази одабраних показатеља разматраног мултидимензионог економског феномена. Сходно томе, најпре је представљен дизајн емпиријског истраживања са фокусом на опис селектованих показатеља економске развијености, дефинисање просторно-временског

обухвата и коришћене изворе података. Даљим излагањем детаљно је презентован поступак развоја иновативног мултиваријационог статистичког модела, у форми одговарајућег композитног показатеља (ИЕР – индекс економске развијености) за мерење степена економске развијености јединица локалне самоуправе (ЈЛС) у Републици Србији. На бази резултирајућих ИЕР вредности извршено рангирање и класификација ЈЛС-а у контексту разматраног мултидимензионог феномена. Дефинисањем истраживачког проблема, представљањем методолошког оквира и провером статистичких претпоставки за валидну примену коришћене комбинације метода у овом сегменту емпиријског истраживања, обезбеђена је адекватна основа за развој композитног показатеља ИЕР. Посебна пажња посвећена је оцени валидности формираног мултиваријационог модела, након чега следи интерпретација резултата из угла конкретног подручја примене. Користећи резултате претходно предложене ИЕР класификације ЈЛС-а, у наставку је креиран мултиваријациони класификациони модел, заснован на дискриминационој анализи, за разврставање ЈЛС-а у Републици Србији у једну од претходно емпиријски идентификованих интерно хомогених и екстерно хетерогених група према степену економске развијености. На крају, сагледана су ограничења спроведеног истраживања и прецизирани могући правци будућег научно-истраживачког рада.

Логично, у Закључку су представљени сумарни резултати спроведеног истраживања и истакнути теоријско–практични доприноси дисертације, након чега следи попис коришћене иностране и домаће литературе.

Теоријски део

КОНЦЕПЦИЈСКИ ОКВИР И АПЛИКАТИВНИ ЗНАЧАЈ МУЛТИВАРИЈАЦИОНЕ СТАТИСТИЧКЕ АНАЛИЗЕ ПОДАТАКА

1.1. Мултиваријациона анализа података – кључне одреднице и значај

Подаци представљају неизоставни производ реализације сваке активности и сваког процеса у свим сегментима живота и рада, тако да функционисање савременог света, практично, није могуће замислити без података. Стога, разумевање својстава података и процеса који их стварају, као и идентификовање потенцијално корисних информација садржаних у подацима, кроз научно засновану (статистичку) анализу, чине полазну основу у објашњењу интересних (разматраних) феномена и решавању конкретних проблема. Заправо, анализа података омогућава конверзију података у информације и знање о структури, релацијама и тенденцијама које карактеришу посматране појаве обухватањем, мерењем и процесирањем њихових различитих карактеристика (променљивих) уз примену различитих методолошких поступака. Другим речима, статистичка анализа података се односи на испитивање података за потребе истраживања карактеристика јединица посматрања на којима се посматране појаве испољавају и сумирање резултата испитивања у форми релевантних законитости интерпретабилних у анализираном контексту.

Сходно броју анализираних променљивих разликују се три приступа у статистичкој анализи: једнодимензиони (униваријациони), дводимензиони (биваријациони) и вишедимензиони (мултиваријациони), односно приступи засновани на анализи варијација и/или међусобних реалација једне, две и више од две променљиве, респективно. За потребе реализације циљева конципираних фундаменталних и примењених истраживања, сам ток анализе веома често захтева кретање од униваријационог и биваријационог ка мултиваријационом приступу, чиме се обезбеђује целовито испитивање и, из различитих перспектива, сагледавање комплексне природе феномена (природних и друштвених) који су предмет проучавања.

Развој и примена статистичких метода и процедура за симултану анализу различитих карактеристика вишедимензионих појава у литератури су обухваћени синтагмом мултиваријациона статистичка анализа, при чему термин „мултиваријациона“ указује не само на присуство више променљивих у анализи, већ и на постојање односа повезаности и зависности између њих (*Đorđević i drugi*, 2011, стр.89). Мада се мултиваријациона анализа дефинише на различите начине, готово у свакој дефиниција иста је опредељена као скуп метода (*Kovačić*, 1994, стр. 1) или грана статистике (*Kramer*, 1978, стр. 848). Сходно томе, у наставку текста, суштинска одређења концепта мултиваријационе анализе се сублимирају кроз следећу дефиницију: мултиваријациона анализа представља скуп метода које омогућавају симултану анализу вишедимензионих мерења добијених за сваку јединицу посматрања једног или више узорака. Другим речима, мултиваријациона анализа је статистичка анализа међусобних релација (више од једне истовремено) између варијабли (више од две) унутар скупа анализом обухваћених варијабли.

Полазна основа у мултиваријационој анализи су мултиваријациони подаци, који се добијају мерењем p променљивих за n јединица посматрања. Табела 1.1.1 репрезентује мултиваријациони профил јединица посматрања у форми ($n \times p$) матрице, $\mathbf{X} = [x_{ij}]$. Редови матрице означавају сваку од n јединица посматрања (i=1, 2,..., l,..., n), а колоне анализом

обухваћене промењиве X_j (j=1, 2,..., m,..., p). Елемент матрице x_{ij} представља вредност j-те промењиве измерене на i-тој јединици посматрања.

При томе, измерене вредности, односно скупови мултидимензионалних опсервација могу се уклопити у једну од следећих основних категорија мултидимензионалних података или, пак, представљати њихову комбинацију (*Rencher*, 2002, стр. 4): (а) подаци једног узорка са више промењивих измерених на свакој јединици посматрања, (б) подаци једног узорка са два скупа промењивих измерених на свакој јединици посматрања, (в) подаци два узорка са више промењивих измерених на свакој јединици посматрања, и (г) подаци три или више узорака са више промењивих измерених на свакој јединици посматрања. За процесирање ових података користе се различити методолошки поступци мултиваријационе статистичке анализе који су примерени специфичним својствима сваке наведене категорије, као и формулисаним циљевима истраживања.

Табела 1.1.1. *Матрица мултиваријационих података*

Јединице посматрања	Промењиве					
	X_1	X_2	…	X_m	…	X_p
1	x_{11}	x_{12}	…	x_{1m}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2m}	…	x_{2p}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
l	x_{l1}	x_{l2}	…	x_{lm}	…	x_{lp}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{nm}	…	x_{np}

Извор: Ауторов приказ

Иако су идеје за многе методе мултиваријационе анализе настале у првим деценијма прошлог века, до њихове афирмације, интензивне употребе и апликативне дисперзије ка готово свим областима истраживања долази са појавом и развојем компјутерске технологије. Заправо, технолошки напредак и решења, пре свега, у домену складиштења података и развоја софтверских пакета за њихову анализу, омогућили су једноставну примену рачунски интензивних метода мултиваријационе анализе. Осим наведеног, разлог повећаног значаја и интересовања за примену и развој ових метода налази се и у чињеници да је мултидимензионалност инхерентно својство већина појава у савременом свету тако да њихово разумевање захтева анализу мноштва промењивих између којих, по правилу, постоје врло сложене релације. У том контексту, будући да су готово сви процеси, феномени и проблеми у економији, пословној економији и менаџменту по својој природи мултидимензионални, истраживања у функцији објашњења повезаности између економских појава представљају изузетно погодно тле за примену метода мултиваријационе анализе.

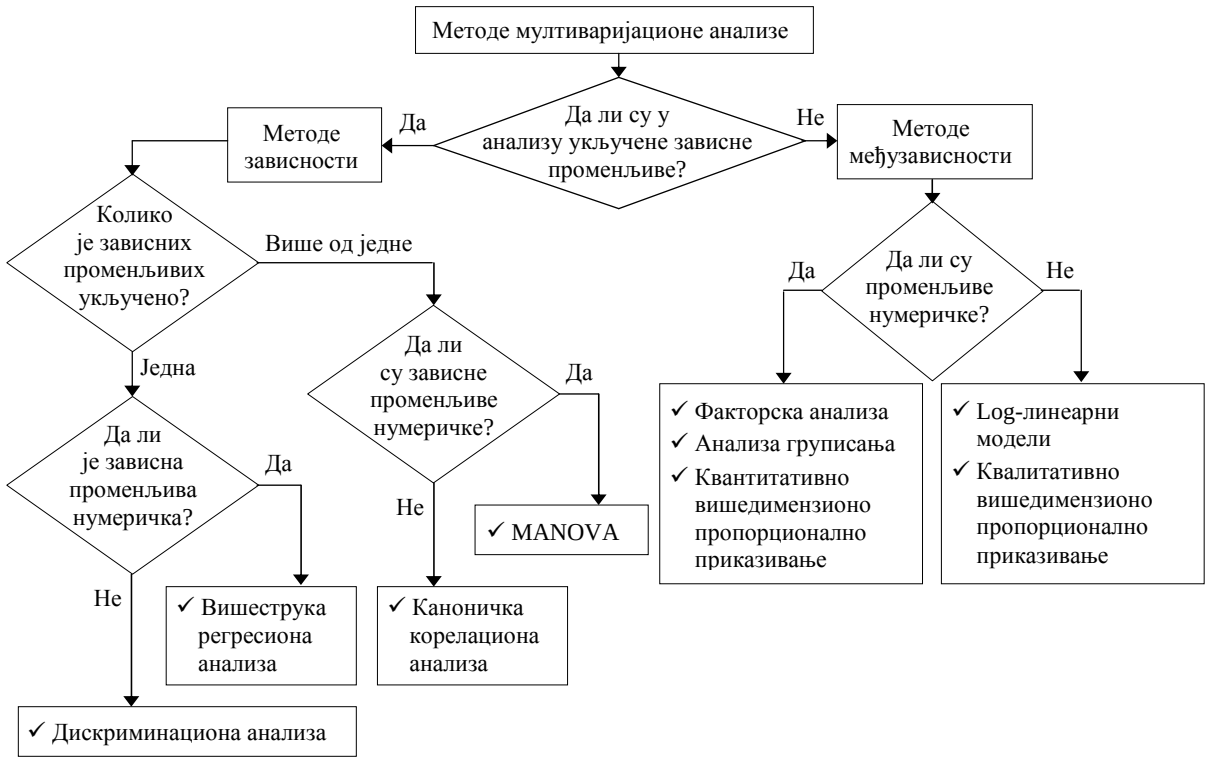
1.2. Класификација метода мултиваријационе анализе података

Мултиваријациона анализа обухвата широк спектар метода, чија класификација може бити извршена на основу различитих критеријума. У релевантној литератури, најчешће се, сходно одговору на питање: Да ли примена метода омогућава испитивање зависности између два подскупа варијабли, где један подскуп представља зависне, а други

независне променљиве?, прави разлика између метода зависности (енгл. *dependence methods*) и метода међузависности (енгл. *interdependence methods*).

Као што сам назив сугерише, суштина метода зависности је истраживање веза између подскупова променљивих, при чему је аналитички циљ усмерен на објашњење или предвиђање вредности зависних променљивих на основу подскупа независних променљивих. Сходно томе, ова група метода припада категорији предиктивних метода (*Kovačić*, 1994, стр. 3). С друге стране, методи међузависности омогућавају утврђивање односа између променљивих, при чему *a priori*, концептулано и теоријски, нема основа за формирање подскупова зависних и независних променљивих. Заправо, на почетку анализе све променљиве имају подједнаку важност и припадају истој групи. Примена метода међузависности омогућава да се установи како и зашто су анализиране променљиве међусобно повезане (*Sharma*, 1996, стр. 4), односно да се кроз симултану анализу свих променљивих у посматраном скупу и редукцију података објасне комплексна унутрашња структура и односи између података (*Kovačić*, 1994, стр. 3; *Đorđević i drugi*, 2011, стр. 90). У суштини, ова група метода припада категорији дескриптивних метода, мада се неке од њих врло успешно користе и за реализацију предиктивних задатака.

За даљу класификацију метода, а у циљу дубљег разумевања њихове природе, често се истичу додатна два критеријума која се односе на број (зависних) променљивих и поделу променљивих на квантитативне и категоријске (сходно типу мерне скале), при чему се при класификацији метода међузависности користи само други критеријум. На Слици 1.2.1 представљени су наведени критеријуми поделе и за сваку класификациону групу издвојени најчешће коришћени методи. Генерално, разумевање и познавање својстава различитих типова података и мерних скала представља веома битан фактор који утиче на избор одговарајуће методе мултиваријационе анализе у датој ситуацији.



Слика 1.2.1. *Класификација метода мултиваријационе анализе*

Извор: Ауторов визуелни приказ, прилагођено према *Kovačić* (1994, стр. 6)

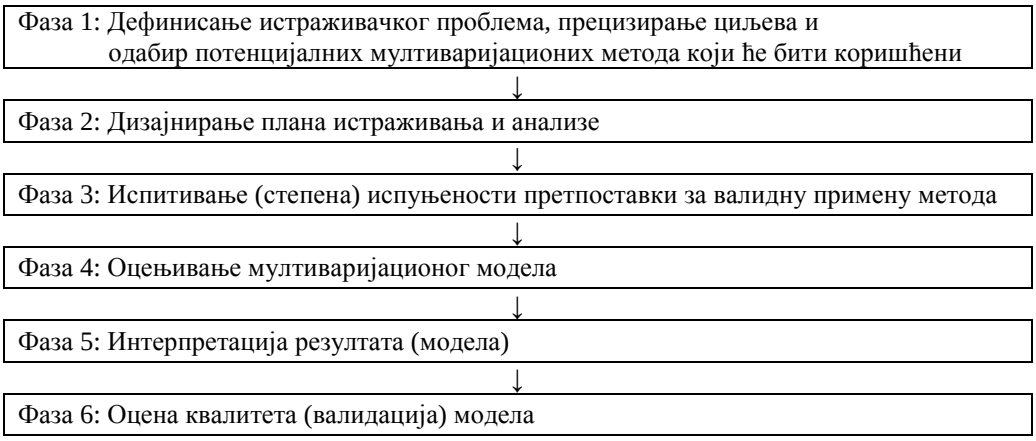
Чињеницу да је реч о богатој методолошкој апаратури за испитивање мултидимензионих феномена потврђују и бројне варијанте (алгоритми) постојећих фундаменталних метода. На пример, анализа груписања се успешно може применити за формирање група јединица посматрања чија су својства исказана не само путем квантитативних, већ и квалитативних променљивих, укључујући и њихову комбинацију. Поред наведених мултиваријационих метода анализе, у научним истраживањима користи се и низ других који се могу означити као методи који припадају овој групи с обзиром на њихова општа својства да се баве симултаном анализом више променљивих између којих постоје односи повезаности, попут, кореспондентне, *CHAID* и анализе здружених ефеката (енгл. *Conjoint Analysis*).

Имајући у виду презентовану методолошку разуђеност мултиваријационе анализе, може се констатовати да избор одговарајуће методе или њихове комбинације у конкретној проблемској ситуацији представља сложен задатак. У суштини, коначан избор детерминишу (поред већ истакнутих типова података) карактеристике дефинисаног проблема, циљеви истраживања (анализе), карактеристике метода, али и распон знања истраживача о својствима метода и поседовање аналитичких вештина у имплементацији истих. Будући да је кључна претпоставка квалитетне анализе података усаглашеност карактеристика примењених метода и конкретног проблемског контекста који је предмет разматрања, сходно истраживањима у емпиријском делу дисертације, у наредним Поглављима представљена су кључна концептуално-методолошка одређења одабраних метода мултиваријационе анализе, и то: факторске анализе, анализе груписања, мултиваријационе анализе варијансе и дискриминационе анализе.

1.3. Процесни приступ у креирању мултиваријационог статистичког модела

Успешно спровођење мултиваријационе анализе података, у контексту решавања конкретне проблемске ситуације, представља изузетно захтеван подухвт који превазилази проблем избора метода и пуке апликације „*user-friendly point-and-click*“ софтверских платформи. У складу са тим, мултидимензионално моделирање и, консеквентно, креирање статистички валидних модела, као исхода мултиваријационе анализе података, нужно подразумева дефинисање јасног процедуралног (општег) оквира у форми процесног модела за идентификовање законитости из података. У основи, сваки процесни модел садржи низ корака (фаза), којима су обухваћене серије процесних активности, са инкорпорираним повратним спрегама и итерацијама. Заправо, примена процесног приступа омогућава интеграцију свих активности релевантних са становишта разумевања конкретног истраживачког проблема и развоја квалитетног модела, као поједностављеног приказа реалности, који може имати одговарајућу употребну вредност за корисника.

Један од универзалних вишефазних поступака за мултиваријационо моделирање, према идеји изложеној од стране *Hair et al*. (2014, стр. 23-24), обухвата шест фаза (корака), које су представљене путем дијаграма тока на Слици 1.3.1. Наиме, овим процесним моделом развоја мултиваријационог модела дефинисан је општи оквир за развој, интерпретацију и верификацију резултата примене било којег мултиваријационог метода. У наставку текста следи кратак осврт на активности у оквиру сваке фазе појединачно.



Слика 1.3.1. *Процес креирања мултиваријационог модела*
Извор: Ауторова визуелна интерпретација, прилагођено према *Hair et al*. (2014, стр. 23-24)

Прва фаза обухвата дефинисање истраживачког проблема и циљева у конкретном истраживачком подручју, а затим и њихово превођење у статистичке (методолошке) проблеме и задатке. У зависности од тога да ли су прецизирани циљеви усмерени на испитивање релација зависности између променљивих или утврђивање структуралних односа и законитости у погледу међусобно сличног понашања променљивих, истраживач приступа избору потенцијалних метода мултиваријационе анализе које ће применити. Друга фаза обухвата одређивање променљивих на основу искуства истраживача и/или теоријских претпоставки, као и питања која се односе на избор узорка (од дефинисања циљне популације, избора метода узорковања, одређивања величине узорка, прецизирања просторног и временског обухвата података, до прикупљање података о елементима узорка). Саставна компоненета садржаја ове фазе су и активности које се односе на прелиминарну припрему података за анализу (попут, валидације, кодирања, пречишћавање и трансформације података), укључујући и прелиминарни план анализе података који се током истраживачког процеса може мењати. У трећој фази пажња се фокусира на проверу испуњености претпоставки које су повезане са валидном применом одабраног мултиваријационог метода, при чему су неке од њих општег карактера и важе за већину мултиваријационих метода, док су неке специфичне и везују се само за одређени метод. У четвртој фази приступе се процесирању података и одређивању, кореспондентно изабраном методу, оцењеног модела кроз оцењивање функције, коефицијената, пондера, скорова, фактора, група и слично. Такође, у овој фази се врши мерење, тестирање и утврђивање (сходно дефинисаним критеријумима) статистичке и практичне значајности резултирајућег модела. Четврта фаза обухвата интерпретацију модела у смислу оцењивања релативног значаја независних променљивих, интерпретације издвојених фактора или стандардизованих и нестандардизованих коефицијената, интерпретације карактеристичних профила формираних група, издвајање важних информација о предиктивној моћи модела итд. Ова фаза може условити потребу за понављањем активности спецификације променљивих и, последично, оцењивања – дефинисања новог модела, како би се на основу података из узорка формулисали валидни закључци о мултиваријационим релацијама у популацији. Заправо, у шестој фази оцењује се квалитет модела са аспекта поузданости и стабилности добијених резултата. Реч је о процени степена генерализације резултата и сагледавању потенцијалног утицаја

појединачних опсервација на опште резултате, јер се може десити да утврђене релације важе за анализиране податке, али не и у другим ситуацијама (*Đorđević i drugi*, 2011, стр. 163).

Поред универзалности (односно, апликативне неутралности), подразумевано својство представљеног поступка креирања мултиваријационог модела се односи на његову итеративну природу са постојањем вишеструких повратних веза између било које две фазе. При томе свака фаза садржи серију одлука којима је детерминисан правац даље анализе у наредним фазама. Међутим, које активности и у којем обиму чине садржај сваке фазе зависи од комплексности посматраног проблема, општег нивоа знања о проблему, вештина и способности истраживача. У том контексту, веома је важно истаћи да, универзални оквир процесног модела не угрожава креативност истраживача у домену начина експлоатације апликативних могућности појединачних метода мултиваријационе анализе или њихове комбинације. Логично, добро обучен и статистички едукован истраживач примениће сложеније методе анализе података, што имплицира богатији садржај активности у свакој фази.

1.4. Методолошки аспекти претпроцесирања података

Претпроцесирање расположивог узорка је фундаментална активност у сваком процесу анализе подтака, будући да омогућава не само разумевање саме природе података, већ обезбеђује и неопходне предуслове за ефикасну примену одговарајућих методолошких поступака у наредним фазама процеса анализе. На значај претпроцесирања указује и чињеница да је реч о најзахтевнијој фази у процесу креирања статистичког модела како са становишта времена, тако и у погледу ангажовања непоходних ресурса (људских, финансијских, информатичких и слично). У суштини, претпроцесирањем се, кроз припремање података за анализу и прелиминарну анализу, решава питање лоших (неквалитетних) података. Коначни резултат претпроцесирања мултиваријационих података је генерисање улазног скупа квалитетних података за креирање одговарајућег мултиваријационог модела. Сходно томе, квалитет података директно опредељује резултат мултиваријационе анализе, али могућност генерализације изведених закључака.

Пре него што се приступи одређивању оцењеног мултиваријационог модела, након прикупљања података и реализације припремних активности (у смислу кодирања, интеграције и провере валидности, потпуности и конзистентности података), неопходно је идентификовати кључне карактеристике улазних података како би се сагледале могућности за апликацију погодног методолошког поступка у оквиру конкретног истраживачког проблема. Заправо, валидна примена мултиваријационе анализе подразумева детаљно прелиминарно испитивање података и проверу одговарајућих статистичких претпоставки како на нивоу појединачних варијабли (универијациони ниво), тако и на нивоу варијате као комбинације појединачних варијабли обухваћених анализом (мултиваријациони ниво), уз коришћење адекватних графичких и рачунских метода.

Примена било којег метода мултиваријационе анализе захтева проверу испуњености серије претпоставки, при чему неке од њих имају својство заједничких претпоставки за већину метода, а неке специфичних претпоставки које се односе само на конкретни метод. При томе, важно је истаћи да се извесне статистичке претпоставке које су означене као

заједничке, ипак, разликују не само са аспекта значаја сваке од њих за конкретну анализу (метод), већ и са аспекта специфичности везаних за поступак испитивања да ли и у којој мери расположиви узорак једнодимензионих и мултиваријационих опсервација задовољава основне (заједничке) претпоставке на којима се заснива валидна имплементација изабране аналитичке процедуре.

Када је реч о примени параметарских метода мултиваријационе анализе, као доминантна из угла њене важности за квалитет финалних резултата моделирања, издваја се претпоставка о мултиваријационој нормалности (заједничког) распореда променљивих, и са њом повезана претпоставка о униваријационој нормалности распореда (појединачних) променљивих. При томе, униваријациона нормалност променљивих је неопходан, али не и довољан услов за постизање мултиваријационе нормалности. Заправо, испуњена униваријациона нормалност не подразумева, по аутоматизму, и испуњеност мултиваријационе нормалности, док обрнуто важи. Због истакнутог значаја претпоставке о нормалности распореда, у наставку текста се наводе изабрани, а у оквиру емпиријског дела дисертације примењени, методолошки поступци за проверу испуњености ове претпоставке. Наиме, испитивање једнодимензионе нормалности распореда променљивих, између осталог, може се спровести применом: *Anderson–Darling*-овог теста, *Shapiro–Wilk*-овог теста, *Z* теста симетричности и заобљености распореда ($Z_{skewness}$ и $Z_{kurtosis}$, респективно), хистограма фреквенција са нормалном кривом и дијаграма нормалне вероватноће. Испитивање, пак, мултиваријационе нормалности може укључити примену *Mardia* тестова симетричности и заобљености мултиваријационог распореда и / или *Henze-Zirkler* теста мултиваријационе нормалности распореда.

У непосредној вези са претпоставком о нормалности распореда је питање идентификовања једнодимензионих и мултидимензионих нестандардних опсервација, које су најчешће узрок нарушености ове претпоставке. Сходно значају и величини утицаја нестандардних опсервација на ток и квалитет резултата анализе података, развијене су бројне методе за њихово идентификовање. Поред већ наведених графичких приказа за испитивање униваријационе нормалности, графички прикази у форми *box-plot* дијаграма и одговарајуће дескриптивне статистичке мере централне тенденције, дисперзије и облика распореда представљају ефикасне начине за откривање присуства једнодимензионих нестандардних опсервација. Присуство мултиваријационих нестандардних опсервација се, такође, ефикасно идентификује путем одговарајућих графичких приказа и мера, попут Хи-квадрат *Q-Q* дијаграма и *Mahalanobis*-ове мере одстојања.

Генерално, у ситуацијама када је озбиљно нарушена нормалност распореда, као и било која друга претпоставка на којој се базира валидна примена одређеног метода, истраживач у конкретној ситуацији може поступити на један од следећих начина (*Sakia*, 1992, стр. 169): (а) да игнорише нарушеност претпоставки и настави са анализом као да су исте испуњене, (б) да установи која је претпоставка испуњена и, сходно томе, донесе одлуку о примени новог метода који узима у обзир управо ту претпоставку, (в) да креира нови модел који садржи кључне карактеристике оригиналног модела и испуњава све неопходне претпоставке, на пример, спровођењем одговарајућих трансформационих процеса на подацима или доношењем одговарајућих одлука о третирању нестандардних опсервација, и (д) да користи непараметарске процедуре које је оправдано применити чак и када су бројне претпоставке нарушене. Не улазећи у дубљу дискусију о разлозима „за и

против“ по свакој предложеној опцији (осим констатације да је прва опција не најмање пожељна, већ недопустива као изабрано решење), овом приликом, истиче се моћ трансформационих процеса у обезбеђивању испуњености претпоставке о нормалности распореда и ублажавању утицаја нестандардних опсервација. Заправо, трансформација података обезбеђује модификовање оригиналних променљивих како у циљу кориговања нарушености претпоставки које стоје у основи валидне примене конкретних мултиваријационих метода, тако и обезбеђења упоредивости променљивих исказаних различитим мерним јединицама.

Иако статистичка теорија и пракса обилују методолошким поступцима за трансформацију, дубља дискусија у том правцу је изван оквира ове дисертације. Услед наведеног, а сагласно потребама емиријског истраживања, следи кратак осврт на *Box–Cox*-ову трансформацију и нормализацију, као два вида трансформације који се разликују у погледу сврхе употребе у процесу моделирања.

Box–Cox-ова трансформација оригиналних вредности променљивих, усмерена на отклањање проблема за нарушеност претпоставке о нормалности распореда и ублажавање утицаја идентификованих нестандардних опсервација, дефинише се путем следећег израза (*Osborne*, 2010):

$$T(X_{ij}) = \frac{X_{ij}^{\lambda_j} - 1}{\lambda_j}, \text{ за } \lambda_j \neq 0,\ i = 1, 2,\ldots, n \text{ и } j = 1, 2,\ldots, p, \qquad (1.4.1)$$

где је: p – укупан број променљивих;

n – укупан број јединица посматрања;

$T(X_{ij})$ – трансформисана вредност i-те јединице посматрања за j-ту променљиву;

X_{ij} – оригинална вредност i-те јединице посматрања за j-ту променљиву;

λ_j – оптимална вредност трансформационог параметра за j-ту променљиву.

Нормализација је усмерена на елиминисање разлика у погледу мерних јединица и свођење вредности променљивих на упоредну основу. Реч је о виду трансформације који се односи на скалирање променљивих на начин да њихове вредности падају унутар специфичног интервала са малим размаком варијација, попут интервала [−1, +1], [0, 1] или [1, 10], независно од тога што оригиналне податке може да карактерише велика разлика између минималне и максималне вредности. Нормализација вредности променљивих, применом методе *min-max* трансформације (*OECD*, 2008; *Perišić & Wagner*, 2015), уз кориговање опсега вредности од 1 до 10, спроводи се путем следећег израза:

$$\text{позитивно кодирање } \rightarrow X'_{ij} = 9 \times \frac{X_{ij} - X_j^{\min}}{X_j^{\max} - X_j^{\min}} + 1, \text{ за } i = 1, 2,\ldots, n \text{ и } j = 1, 2,\ldots, p, \qquad (1.4.2)$$

$$\text{инверзно кодирање } \rightarrow X'_{ij} = -9 \times \frac{X_{ij} - X_j^{\min}}{X_j^{\max} - X_j^{\min}} + 10, \text{ за } i = 1, 2,\ldots, n \text{ и } j = 1, 2,\ldots, p\ , \qquad (1.4.3)$$

где је: $X_{ij}^{'}$ – нормализована вредност i-те јединице посматрања j-те променљиве;

X_{ij} – оригинална i-та вредност j-те променљиве;

X_j^{min} – минимална оригинална вредност j-те променљиве;

X_j^{max} – максимална оригинална вредност j-те променљиве.

Имајући у виду наведено, важно је нагласити да, резултати оног дела претпроцесних активности којима се испитује испуњеност претпоставки на којима се заснива валидна имплементација мултиваријационих метода примарно детерминишу употребну вредност изведених модела.

1.5. Графичко приказивање мултиваријационих података

Поред нумеричких метода, други фундаментални стуб на којем се заснива истраживање и анализа мултиваријационих података односи се на визуелизацију података, односно примену графичких метода. Визуелизација, кроз приказивање података применом различитих геометријских форми и фигура и у различитим димензијама, омогућава издвајање релевантних информација садржаних у подацима, односно приказ динамике (трендова) појава, утврђивање нивоа (стања) и структуре појава, праћење промена структуре у времену и простору, упоређивање променљивих и откривање релација у подацима, идентификовање нестандардних опсервација и, генерално, разумевање, објашњење и интерпретацију анализираних феномена.

Графичка анализа је често полазна тачка у анализи података, будући да људски визуелни систем поседује изузетне способности у погледу уочавања правилности у подацима који су представљени путем адекватних, добро дизајнираних, графичких ентитета (геометријских облика, ознака на географским картама и визуелних метафора). Међутим, поред приказа сирових и претпроцесираних података, графички методи имају значајну улогу у представљању међурезултата појединих фаза процеса креирања мултиваријационог модела, као и финалних резултата анализе (Милановић, 2018, стр. 140). Заправо, постоји „неограничени“ број графичких метода за приказивање података, од једноставних у форми *1D*, *2D* и *3D* графикона до изразито специфичних и сложених, развијених на темељу развоја и достигнућа у домену информационо-комуникационих технологија. Методи визуелизације мултиваријационих података могу се, генерално, класификовати у четири широке категорије (*Chan*, 2006, стр. 8): геометријски прикази, прикази помоћу слика, хијерархијски прикази и *Pixel* оријентисани прикази.

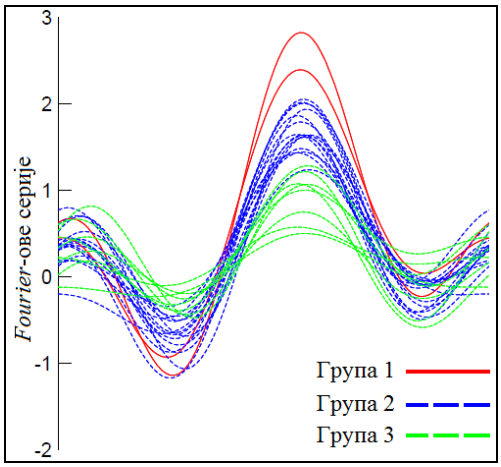
С обзиром на бројност визуелних формата података, у процесу анализе података засноване на визуелизацији, централно питање се односи на избор метода који ће се у конкретној ситуацији применити. У настојању да се избегну погрешна решења, неопходно је узети у обзир релевантне карактеристике података, које се тичу њиховог типа, количине и димензионалности (*Sachinopoulou*, 2001, стр. 28), уз неизоставно разматрање циљева истраживања које треба остварити.

Иако су у даљем тексту дисертације, нарочито у емпиријском делу истраживања, примењени различити графички методи, у наставку се, ради стицања потпуног увида у разноврсност и интересантност визуелизације мултиваријационих података, укратко представљају (и на Сликама 1.5.1 и 1.5.2 илуструју) два репрезентативна, веома често коришћена, приказа: *Andrews*-ове криве (припада категорији геометријских приказа) и *Chernoff*-ова лица (припада категорији приказа помоћу слика).

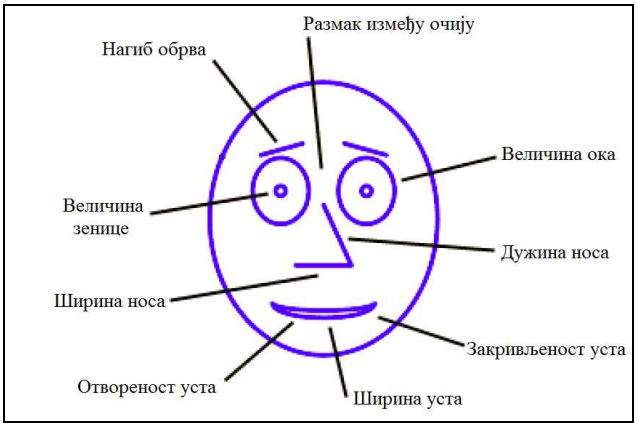
Аутор *Andrews* је у свом раду „*Plots of High-Dimensional Data*“, публикованом у часопису *Biometrics*, предложио процедуру конструкције графичког приказа који је познат под називом *Andrews*-ове криве (*Andrews*, 1972). Основна идеја овог приказа је да се свака

јединица посматрања са *p*-променљивих (односно, свака мултидимензиона опсервација) представи путем криве (линије) профила која је одређена коришћењем *Fourier*-ове трансформационе функције. *Andrews*-ов приказ је, у ствари, скуп *Fourier*-ових серија, при чему свака серија репрезентује један мултидимензиони податак. Као метод визуелизације, *Andrews*-ов приказ омогућава, након извршене стандардизације података, идентификовање скривених структура у релативно малом скупу података – скупу података са мање од хиљаду опсервација (*Moustafa*, 2011, стр. 374). Стога, кључне примене овог приказа односе се на (визуелно) груписање јединица посматрања у релативно хомогене групе и откривање присуства нестандардних опсервација у анализираном скупу података.

Статистичар *Chernoff* је у свом раду „*The Use of Faces to Represent Points in k-Dimensional Space Graphically*“, публикованом у часопису *Journal of the American Statistical Association*, предложио процедуру конструкције графичког приказа који је познат под називом *Chernoff*-ова лица (*Chernoff*, 1973). Реч је о приказу који се базира на коришћењу елемената и карактеристика људског лица (облик лица, дужина носа, положај обрва, величина очију, величина зеница, ширина усана итд.) за изражавање различитих вредности променљивих. Наиме, поједине варијабле додељују се различитим карактеристикама лица (на пример, промељива X_5 → ширина носа), након чега се компјутерски генерише израз лица за сваку јединицу посматрања и кореспондентну комбинацију вредности *p*-променљивих. На основу визуелне сличности *Chernoff*-ових лица, путем којих се репрезентује по једна мултидимензиона опсервација, без формализованих поступака груписања, могуће је издвојити релативно хомогене групе, уочити разлике између формираних група и открити нестандардне опсервације. Заправо, овај приказ побољшава способност истраживача да детектује и разуме важне феномене и служи као визуелни „уређај“ који олакшава меморисање кључних закључака (*Raciborski*, 2009, стр. 374). Погодан за визуелизацију мултидимензионих опсервација у случајевима када је $p \leq 18$, тако да се његове визуелне предности у идентификовању законитости смањују са повећањем броја јединица посматрања и / или броја променљивих. У циљу ублажавања наведеног недостатка, истраживач може у ситуацијама када је $p > 18$ приступити генерисању два лица по свакој мултидимензионој опсервацији (*Chernoff*, 2011, стр. 244).



Слика 1.5.1. *Andrews-ове криве*
Извор: Ауторов визуелни приказ



Слика 1.5.2. *Chernoff-ово лице*
Извор: Прилагођено према *Spinelli & Zhou* (2004, стр. 1)

Имајући у виду претходна разматрања, јасно је да графички методи представљају веома моћно средство за разликовање, класификацију, рангирање, компарацију и повезивање података у функцији идентификовања законитости из података. Мада су у статистици графикони иницијално коришћени, пре свега, у функцији илустрације табеларно представљених статистичких података и спровођења једноставних анализа, визуелизација података је данас посебно истраживачко поље које се интензивно развија и активно примењује за анализу и разумевање веома комплексних појава у различитим дисциплинама и областима.

1.6. Статистички програмски пакети за мултиваријациону анализу података

Развој информационо-комуникационих технологија иницирао је дизајнирање великог броја софтверских пакета за анализу података, међу којима посебно место припада онима који су ближе одређени атрибутом „статистички“, тако да је апликација статистичких софтвера постала круцијални сегмент сваке фазе процеса анализе података. За већину статистичких софтверских пакета је карактеристично да, поред програмских модула за примену фундаменталних статистичких метода униваријационе и биваријационе анализе, садрже широку лепезу нумеричких и графичких метода за анализу мултиваријационих опсервација. Заправо, (р)еволуција у домену компјутерске технологије је знатно убрзала и олакшала примену многих метода, али истовремено и проширила апликативни потенција мултиваријационе анализе у многим научним и практичним истраживањима кроз унапређење постојећих и развој нових алгоритамских решења. Осим тога, многе мултиваријационе методе практично није могуће имплементирати без адекватне софтверске подршке.

Као најчешће коришћени софтверски пакети комерцијалног типа, у оквиру којих су поред стандардних статистичких модула интегрисани и програмски модули за низ мултиваријационих метода, издвајају се: *IBM SPSS*, *SAS* и *STATA*, као и апликације засноване на програмима (који су уједно и програмерски језици) *MatLab* и *R*. Сваки од наведених софтверских алата поседује одређене карактеристике које га, између осталог, јасно диференцирају од других у погледу уграђених расположивих опција (процедура и функција) за примену метода мултиваријационе анализе. Чињеница да се број доступних алата непрекидно повећава, како због креирања модификованих верзија постојећих решења, тако и због појаве потпуно нових решења, јасно указује на растући (могло би се рећи чак и старатегијски) значај проблема избора доброг софтверског алата са становишта конкретног истраживачког подручја, институције и крајњег корисника - аналитичара. Међутим, питање евалуације статистичких софтвера кроз идентификовање њихових предности и недостатака у циљу усаглашавање коначног избора са аналитичким потребама је изван оквира ове дисертације.

Упркос значају коју софтверска подршка има при реализацији фаза процеса мултиваријационе анализе податка, ипак, доминанатна улога припада истраживачким знањима и вештинама у спровођењу исте. Стога одговорност за квалитет резултата мултиваријационе статистичке анализе сносе непосредни аналитичари који морају познавати ограничења и претпоставке разних метода како би у датој ситуацији извршили оптималан избор метода (и/или комбинације метода) и, последично, коректно

протумачили резултате. Заправо, сваки истраживач мора непристрасно и професионално приступити анализи података, како би се ризик неправилне примене метода и погрешних закључака свео на минимум. Услед наведеног значаја познавања ограничења и обезбеђења испуњености претпоставки за валидну примену метода, независно од софтверске имплементације, у наредним Поглављима су приказана кључна методолошка одређења изабраних метода мултиваријационе анализе на класичан начин коришћењем концепта матричне нотације података.

МЕТОДОЛОШКА ОДРЕЂЕЊА ОДАБРАНИХ МУЛТИВАРИЈАЦИОНИХ МЕТОДА МЕЂУЗАВИСНОСТИ

2.1. ФАКТОРСКА АНАЛИЗА

У многим областима и подручјима научног деловања и истраживања, нарочито у домену друштвених наука, често није могуће извршити директно мерење специфичних феномена или концепата, који су од примарног интереса и значаја за истраживача (*Everitt*, 2010, стр. 211). Заправо, реч је о концептима који се сматрају корисним у теоријском и практичном смислу, али који се не испољавају на начин својствен, у статистичкој анализи доминантно присутним, опажљивим (односно, опсервабилним) карактеристикама јединица посматрања чије је модалитете могуће директно „измерити" (*Barthlomew et al.*, 2008, стр. 176). Генерално, такви немерљиви (односно, неопсервабилни или неопажљиви) концепти представљају латентне променљиве (енгл. *latent variable*) (*Bartholomew*, 2011, стр. 502)[1], а њихово „мерење" и анализирање спроводи се индиректно, прикупљањем и испитивањем информација везаних за скуп конкретних опсервабилних променљивих (енгл. *indicator* / *manifest variables*), које се, на бази реалних и оправданих аргумената, могу сматрати њиховим адекватним показатељима (*Everitt*, 2010, стр. 211; *Barthlomew et al.*, 2008, стр. 176). Овакво индиректно испитивање заснива се на претпоставци да су анализом обухваћене (опсервабилне) индикатор-променљиве под утицајем латентне(их) променљиве(их), при чему наведена зависност индукује корелацију између њих. У том смислу, присуство одређеног степена квантитативног слагања варијација вредности две индикатор-променљиве може се сматрати доказом постојања заједничког извора утицаја, односно зависности њихових вредности од хипотетичке латентне променљиве (*Barthlomew et al.*, 2008, стр. 177). Метод мултиваријационе статистичке анализе који се најчешће користи за откривање и описивање међусобне зависности одговарајућег броја (две или више) опсервабилних индикатор-променљивих и (најмање једне) претпостављене латентне (неопсервабилне) случајне променљиве, углавном означене термином заједнички фактор (енгл. *common factor*), назива се факторска анализа (енгл. *Factor Analysis*) (*Kovačić*, 1994, стр. 215; *Đorđević i drugi*, 2011, стр. 139; *Everitt*, 2010, стр. 211).

Историјски посматрано, основна идеја факторске анализе (акроним, ФА), формално изложена у првим годинама XX века, произашла је превасходно из напора психолога тог доба да боље разумеју и истраже комплексан феномен људске „интелигенције" (*Kovačić*, 1994, стр. 215). У том контексту, највеће заслуге за утемељење и рани развој ФА-е припадају психологу *Charles*-у *Spearman*-у (*Bartholomew*, 2011, стр. 502; *Manly & Navarro Alberto*, 2017, стр. 121), захваљујући иновативној статистичкој процедури, представљеној у чланку под насловом „*'Општа интелигенција', објективно одређена и мерена*" (*Spearman*, 1904), коју је развио за потребе креирања првог квантитативног модела структуре когнитивних способности (*Vasić*, 2014, стр. 57), односно моделирање феномена људске интелигенције. Наиме, у изучавању каузалне релације између опште интелигенције (као латентне променљиве) и резултата различитих когнитивних тестова (односно, скупа опсервабилних индикатор–променљивих) којима су испитаници били подвргнути, професор *Spearman* је пошао од претпоставке да су резултати извршених тестова међусобно корелисани и да очекивана квантитативна слагања варијација парова

[1] Примери таквих неопсервабилних променљивих могу бити ниво опште интелигенције, политички став, или социо-економски статус испитаника (*Barthlomew et al.*, 2008, стр. 175), или пак, степен економске развијености разматраних територијалних јединица, одговарајућег административног нивоа (нпр. општине, окрузи, региони, државе и сл.).

резултата појединачних тестова могу у значајној мери бити објашњена „деловањем“ једног заједничког фактора, названог ниво опште интелигенције (*Timm*, 2002, стр. 496; *Sharma*, 1996, стр. 90). Резултати спроведеног психометријског истраживања, послужили су *Spearman*-у као основа за формулисање такозване двофакторске теорије когнитивних способности и интелигенције, према којој се резултат сваког теста може схватити и исказати као последица усаглашеног деловања следећа два фактора који условљавају вредност истог, и то (*Kovačić*, 1994, стр. 216; *Timm*, 2002, стр. 496; *Manly & Navarro Alberto*, 2017, стр. 121): ▸ први фактор, назван *ниво опште интелигенције* или *општа способност*, који је заједнички свим тестовима (енгл. *common factor*), и ▸ други, такозвана *јединствена способност* или *посебан таленат*, специфичан за сваки тест (односно, дисциплину којој припада) појединачно (енгл. *unique factor*). Захваљујући истраживањима која су уследила, превасходно од стране психолога *Louis*-а *Thurstone*-а (1931), извршена је генерализација *Spearman*-овог једнофакторског модела и мултиваријационе статистичке процедуре коришћене за његово извођење, чиме је омогућено издвајање и анализирање међузависности више од једног заједничког фактора и скупа опсервабилних индикатор-променљивих (*Manly & Navarro Alberto*, 2017, стр. 121; *Boslaugh & Watters*, 2008, стр. 299; *Timm*, 2002, стр. 496).

Због њене иницијалне повезаности са истраживањем неопсервабилних феномена из домена психологије, али и специфичне терминологије коришћене при интерпретирању резултата спроведених студија, до средине XX века, практична имплементација и развој факторске анализе, иако у суштини статистичке методе, претежно су били под ингеренцијом истраживача заинтересованих за психометрију (*Johnson & Wichern*, 2007, стр. 481). Третирање факторске анализе у највећој мери као „привилегије“ психолога у деценијама након њеног увођења (*Bartholomew*, 2011, стр. 502), утицало је на одлагање адекватног разматрања статистичких питања и проблема у вези са ФА у дужем временском периоду, током којег је овај статистички метод, како истичу *Lawley & Maxwell* (1962, стр. 209), представљао „црну овцу“ статистичке теорије. У другој половини XX века, појавом и рапидним развојем рачунара, створени су услови за ефикасно „превазилажење“ изразите рачунске комплексности ФА-е, као једног од кључних ограничавајућих фактора њене интензивније примене и развоја у претходном периоду. Могућност ефикасног спровођења компликованих прорачуна модерне факторске анализе коришћењем *user-friendly point-and-click* софтверских платформи допринела је широкој афирмацији њених апликативних могућности и, консеквентно, њеном позиционирању у академској и научној заједници као једне од најчешће коришћених мултиваријационих статистичких процедура при реализацији истраживања у бројним научним дисциплинама (*Lawley & Maxwell*, 1962, стр. 209; *Thompson*, 2004, стр. 4; *Brown*, 2015, стр. 10).

2.1.1. Циљеви и поступак спровођења факторске анализе

Генерално, ФА представља мултиваријациони статистички метод међузависности[2] који омогућава анализирање и боље разумевање комплексне структуре релација (веза или

[2] За разлику од мултиваријационих метода зависности, заснованих на експлицитној подели анализираних променљивих на зависне и независне, при спровођењу ФА-е врши се истовремено разматрање и равноправно третирање свих индикатор-променљивих (*Đorđević i drugi*, 2011, стр. 139), чије су вредности утврђене коришћењем нумеричке (интервалне / рацио) мерне скале (*Barthlomew et al.*, 2008, стр. 177).

корелација) које постоје између неколико (две или више) нумеричких индикатор-променљивих путем идентификовања мањег броја (једне или више) латентних променљивих (такозваних, заједничких фактора или фундаменталних димензија) за које се претпоставља да су у основи међусобне повезаности скупа оригиналних променљивих (*Thompson*, 2004, стр. 10; *Barthlomew et al.*, 2008, стр. 177; *Đorđević i drugi*, 2011, стр. 146; *Manly & Navarro Alberto*, 2017, стр. 121).

Другим речима, примарни циљ и сврха факторске анализе јесте ефикасна (уз минимални губитак) сумаризација информација садржаних у структури коваријационе (или корелационе) матрице релативно комплексног скупа опсервабилних променљивих путем идентификовања одређеног (унапред непознатог, али у поређењу са бројем индикатор-променљивих мањег) броја нових, неопсервабилних варијата, односно заједничких фактора који се сматрају „одговорним“ за корелацију између група анализираних индикатор-променљивих. (*Lawley & Maxwell*, 1962, стр. 210; *Timm*, 2002, стр. 496; *Barthlomew et al.*, 2008, стр. 177; *Hair et al.*, 2010, стр. 95). Сходно дефинисаном примарном циљу, може се констатовати да ФА представља моћан статистички алат за редукцију димензионалности конкретно разматраног истраживачког проблема, услед чега се често практикује као прелиминарни корак у имплементацији аналитичких метода осетљивих на број улазних променљивих (*Tuffery*, 2011, стр. 175).

У начелу, имплементација факторске анализе, у контексту реализације формулисаног циља, подразумева извођење одговарајућег статистичког модела, такозваног модела заједничког(их) фактора (енгл. *common factor*(*s*) *model*), којим се обезбеђује спона између опсервабилних индикатор-променљивих и неопсервабилне(их), латентне(их) променљиве(их) (*Kovačić*, 1994, стр. 215; *Brown*, 2015, стр. 11; *Everitt*, 2010, стр. 212; *Bartholomew*, 2011, стр. 502). По својој статистичкој природи, модел ФА представља модификовану верзију класичног линеарног регресионог модела [3] и заснива се на претпоставци да су присутна квантитативна слагања између варијација p индикатор-променљивих (у ознаци: X_1, X_2,..., X_p), најчешће приказана у форми одговарајуће ($p \times p$) корелационе матрице, заправо резултат њихове повезаности са једном или неколико латентних променљивих (*Sharma*, 1996, стр. 91; *Everitt*, 2010, стр. 211).

У зависности од броја издвојених заједничких фактора (у ознаци: F_1, F_2,…, F_m) неопходних за разумевање структуре полазне корелационе матрице, односно објашњење међусобне корелисаности p индикатор-променљивих, разликују се једнофакторски модели (енгл. *single-common factor model*), као најједноставнији облик модела ФА (израз 2.1.1.(А)) и вишеструки или *m*-факторски модели (енгл. *multiple-common factors model*), засновани на постојању два или више заједничких фактора који се сматрају одговорним за квантитативну повезаност p индикатор-променљивих (израз 2.1.1.(Б)) [4] (*Sharma*, 1996, стр. 93; 96; *Barthlomew et al.*, 2008, стр. 175). Наведени типови модела ФА (респективно) могу се, у општем облику, приказати на следећи начин, у форми одговарајућих система регресионих једначина (*Brown*, 2015, стр. 17; *Everitt*, 2010, стр. 212):

[3] *Barthlomew et al.* (2008, стр. 178) дефинише модел факторске анализе као скуп или систем простих или вишеструких регресионих модела у којима је скупина објашњавајућих променљивих, латентне природе, то јест неопсервабилна.

[4] Сходно методолошким потребама спроведеног емпиријског истраживања, презентираног у оквиру последњег поглавља дисертације, у наставку је пажња усмерена доминантно на елаборирање одлика једнофакторског модела ФА.

(А) (Б)

$$\left.\begin{aligned} X_1 &= \beta_1 F_1 + \varepsilon_1 \\ X_2 &= \beta_2 F_1 + \varepsilon_2 \\ &\vdots \\ X_p &= \beta_p F_1 + \varepsilon_p \end{aligned}\right\} X_j = \beta_j F_1 + \varepsilon_j \ (\text{за } j = 1, 2, ..., p) \quad \text{и} \quad \left.\begin{aligned} X_1 &= \beta_{11} F_1 + \beta_{12} F_2 + ... + \beta_{1m} F_m + \varepsilon_1 \\ X_2 &= \beta_{21} F_1 + \beta_{22} F_2 + ... + \beta_{2m} F_m + \varepsilon_2 \\ &\vdots \\ X_p &= \beta_{p1} F_1 + \beta_{p2} F_2 + ... + \beta_{pm} F_m + \varepsilon_p \end{aligned}\right\} X_j = \beta_{j1} F_1 + ... + \beta_{jm} F_m + \varepsilon_j \ (\text{за } j = 1, 2, ..., p \text{ и } m \geq 2) \quad (2.1.1)$$

где су:

X_j – j-та индикатор-променљива (за j = 1, 2,..., p);

F_f – f-ти заједнички (латентни) фактор (за f = 1, 2,..., m);

β_{jf} – факторско оптерећење j-те индикатор-променљиве и f-тог заједничког фактора;

ε_j – j-та случајна грешка (стохастички члан или резидуална компонента факторског модела), која се, сходно терминологији коришћеној у ФА, често назива и j-ти специфични фактор (енгл. *specific* / *unique factor*) будући да је својствен свакој појединачној променљивој X_j.

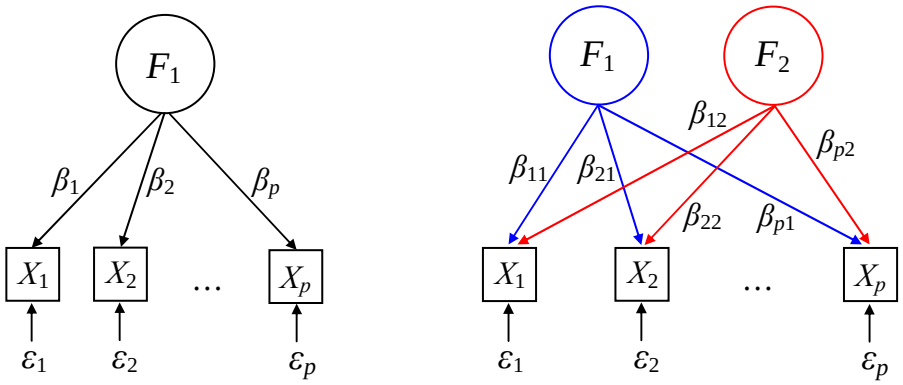
Основне и углавном навођене претпоставке на којима почива оптимално оцењивање непознатих параметара модела факторске анализе могу бити формулисане на следећи начин (*Sharma*, 1996, стр. 92; *Barthlomew et al.*, 2008, стр. 181; *Everitt*, 2010, стр. 213; *Manly & Navarro Alberto*, 2017, стр. 123; *Brown*, 2015, стр. 18; *Lawley & Maxwell*, 1962, стр. 211):

- Специфични фактори ε_j (за j = 1, 2,..., p) следе нормалан распоред са аритметичком средином 0 ($E(\varepsilon_j) = 0$) и варијансом $Var(\varepsilon_j)$, симболички, ε_j : N(0, $Var(\varepsilon_j)$), при чему између истих не постоји линеарна корелациона веза;
- Издвојени заједнички фактори (F_1, F_2,..., F_m) нису међусобно корелисани;
- Између заједничких фактора (F_1, F_2,..., F_m) и специфичних фактора (ε_1 ε_2,..., ε_p) не постоји линеарна корелација;
- Полазећи од чињенице да заједнички фактори представљају немерљиве променљиве, њихова мерна скала може бити арбитрарно подешена. Углавном се претпоставља да су исказани у стандардизованој форми са аритметичком средином 0 и варијансом 1;
- Модел ФА је у потпуности стандардизован, односно оригиналне индикатор-променљиве трансформисане су у стандардизоване нормално распоређене случајне променљиве са аритметичком средином 0 и варијансом 1.

Представљеним моделима факторске анализе се свака индикатор-променљива појединачно (X_1, X_2,...,X_p) исказује као линеарна комбинација (функција) једног или више заједничких фактора (F_1 или F_1, F_2,…, F_m) и једног специфичног фактора (ε_j) који одражава степен независности конкретно посматране променљиве у односу на све остале (*Kovačić*, 1994, стр. 4; *Sharma*, 1996, стр. 91; *Brown*, 2015, стр. 11), при чему је, уколико су оригиналне променљиве барем умерено корелисане, димензионалност изведеног система генерално мања од оригиналне димензионалности проблема, то јест $m < p$ (*Rencher*, 2002, стр. 408). Вредности факторских оптерећања (енгл. *factor loadings*), у ознаци β_{jf} (за j = 1, 2,…, p и f = 1, 2,…, m), показују смер и степен линеарног квантитативног слагања између варијација j-те променљиве и f-тог издвојеног заједничког фактора (*Sharma*, 1996, стр. 92; *Đorđević i drugi*, 2011, стр. 143) и, сходно томе, крећу се у интервалу $-1 \leq \beta_{jf} \leq +1$.

Консеквентно, веће вредности β_{jf} указују на већи степен повезаности конкретне индикатор-променљиве и посматраног заједничког фактора и обратно.

Посматрано из угла факторских модела заснованих на издвајању само једног заједничког фактора, уколико факторско оптерећење на нивоу било које оригиналне променљиве износи нула, тада се може констатовати да између те индикатор-променљиве и преосталих (p–1) променљивих, као одговарајућих мера идентификоване латентне променљиве, не постоји линеарна корелација (*Sharma*, 1996, стр. 92). Разлог наведеној тврдњи, као што је претходно истакнуто, садржан је у претпоставци модела према којој издвојени заједнички фактор заправо представља јединог одговорног „кривца“ за корелације које постоје између анализираних променљивих. Другим речима, корелисаност индикатор-променљивих сугерише да исте „деле“ (најмање) једну заједничку особину, која се ФА моделом може екстраховати у форми заједничког фактора, док вредности факторских оптерећења показују степен заступљености и утицаја наведене неопсервабилне особине на кретање вредности појединачних индикатора X_j. Елаборирane релације присутне између скупа индикатор-променљивих и једног или више заједничких фактора, на примеру једнофакторског и двофакторског модела ФА, могу се графички представити као на Слици 2.1.1.



Слика 2.1.1. Графички приказ релација између елемената на нивоу једнофакторског (лево) и двофакторског (десно) хипотетичког модела факторске анализе

Полазећи од претходно дефинисаног једнофакторског модела ФА (израз 2.1.1. и Слика 2.1.1.) укупан варијабилитет сваке појединачне индикатор-променљиве X_j, у ознаци, $Var(X_j)$, може се представити путем следећег израза (*Manly & Navarro Alberto*, 2017, стр. 123)[5]:

$$Var(X_j) = 1 = Var(\beta_j F + \varepsilon_j) = \beta_j^2 Var(F) + Var(\varepsilon_j) = \beta_j^2 + Var(\varepsilon_j), \quad \text{за } j = 1, 2, \ldots, p. \qquad (2.1.2)$$

Прецизније, полазећи од (p×p) корелационе матрице као улазне структуре, модел ФА имплицира раздвајање варијансе сваке појединачне индикатор-променљиве X_j на два дела и то (*Kovačić*, 1994, стр. 217; *Sharma*, 1996, стр. 92; *Everitt*, 2010, стр. 213; *Đorđević i drugi*, 2011, стр. 139; *Brown*, 2015, стр. 11; *Manly & Navarro Alberto*, 2017, стр. 123): ▸ први део, у ознаци h_j^2, који се назива заједничка варијанса (енгл. *common variance*) или комуналитет

[5] Генерализацијом израза (2.1.2), укупан варијабилитет појединачних индикатора у контексту примене вишефакторског модела ФА може се представити на следећи начин (*Everitt*, 2010, стр. 213; *Manly & Navarro Alberto*, 2017, стр. 123):

$$Var(X_j) = 1 = \sum_{f=1}^{m} \beta_{jf}^2 + Var(\varepsilon_j), \quad \text{за } j = 1, 2, \ldots, p. \qquad (2.1.3)$$

(енгл. *communality*) променљиве X_j, и ▸ други део, у ознаци $Var(\varepsilon_j)$ или ψ_j, под називом специфична или својствена варијанса (енгл. *specific / unique variance*) променљиве X_j. Вредност комуналитета конкретно посматране променљиве X_j репрезентује удео варијансе коју дата променљива „дели“ са осталим индикаторима путем заједничког фактора F, односно удео укупног варијабилитета променљиве X_j који је објашњен заједничким фактором. Будући да се, у случају једнофакторског модела ФА, утврђује се као квадрат вредности кореспондентног факторског оптерећења (β_j) који стоји уз издвојени фактор F,[6] вредност комуналитета сваке појединачне променљиве X_j креће се у интервалу $0 \leq h_j^2 = \beta_j^2 \leq 1$, при чему већа вредност имплицира да посматрана индикатор-променљива представља бољу и поузданију меру издвојеног заједничког фактора и обратно (*Sharma*, 1996, стр. 92). С друге стране, специфична варијанса променљиве X_j репрезентује удео укупног варијабилитета конкретне променљиве који је својствен само њој, односно који не „дели“ (који није заједнички) ни са једном другом индикатор-променљивом, будући да представља резултат деловања специфичног фактора ε_j. Полазећи од претпоставке модела ФА према којој је $Var(X_j) = 1$, удео укупне варијансе променљиве X_j који није корелисан и условљен заједничким фактором утврђује се као разлика укупног варијабилитета и комуналитета дате променљиве, односно (*Manly & Navarro Alberto*, 2017, стр. 123): $Var(\varepsilon_j) = \psi_j = Var(X_j) - \beta_j^2 = 1 - \beta_j^2$.

На основу претходних разматрања, евидентно је да се формираним моделом ФА целокупна коваријанса (или корелација) између p индикатор-променљивих објашњава заједничким фактором(има), а необјашњени део укупног варијабилитета појединачних променљивих придружује случајној грешци (*Kovačić*, 1994, стр. 4). Прихватљивост модела ФА у највећој мери је детерминисана степеном у којем оцене непознатих параметара факторског решења, конкретно факторских оптерећења, добро описују и објашњавају присутне релације између улазних индикатор-променљивих.[7] У том смислу, основна идеја и циљ факторске анализе, као мултиваријационе статистичке процедуре засноване на једном узорку (*Rencher*, 2002, стр. 409), огледа се у оцењивању модела ФА и идентификовању / издвајању најмањег броја (генерално, $m < p$) интерпретабилних заједничких фактора неопходних за адекватно објашњавање присутне корелације између улазних индикатор-променљивих (*Bartholomew*, 2011, стр. 502; *Sharma*, 1996, стр. 96; *Pituch & Stevens*, 2016, стр. 339).

Полазећи од изнетих спецификација модела и циљева ФА, у литератури се истичу разлике између следећа два основна и одвојена типа разматране мултиваријационе процедуре

---

[6] Посматрано из угла m-факторског модела ФА, комуналитет појединачних индикатор-променљивих утврђује се као збир квадрата кореспондентних факторских оптерећења на нивоу променљиве X_j и сваког од m издвојених заједничких фактора, симболички (*Everitt*, 2010, стр. 213; *Manly & Navarro Alberto*, 2017, стр. 123): $h_j^2 = \sum_{f=1}^{m} (\beta_{jf})^2$, за $j = 1, 2,..., p$. Консеквентно, специфични део укупног варијабилитета појединачних индикатора утврђује се путем следећег израза: $Var(\varepsilon_j) = \psi_j = Var(X_j) - h_j^2 = 1 - \sum_{f=1}^{m} (\beta_{jf})^2$, за $j = 1, 2,\ldots, p$.

[7] Наиме, полазећи од теоријских претпоставки на којима се заснива модел факторске анализе, може се показати да је корелација између две индикатор-променљиве (на пример, X_1 и X_2) заправо кореспондентна следећем изразу (*Everitt*, 2010, стр. 213; *Manly & Navarro Alberto*, 2017, стр. 123): $r_{X_1X_2} = \beta_{11} * \beta_{21} + \beta_{12} * \beta_{22} + ... + \beta_{1m} * \beta_{2m} = \Sigma\beta_{1f} * \beta_{2f}$ (за $f = 1, 2,\ldots, m$), односно, у случају једнофакторског модела ФА (*Brown*, 2015, стр. 18; *Barthlomew et al.*, 2008, стр. 180): $r_{X_1X_2} = \beta_1 * \beta_2$. Сходно наведеном, што је одступање, евидентирано поређењем вредности коефицијента просте линеарне корелације индикатор-променљивих и (збира) производа кореспондентних оцена факторских оптерећења, мање, компариране вредности су међусобно блиске, а модел се сматра бољим и квалитетнијим, односно генерисана матрица факторских оптерећења боље описује корелацију између оригиналних променљивих, и обратно (*Brown*, 2015, стр. 18).

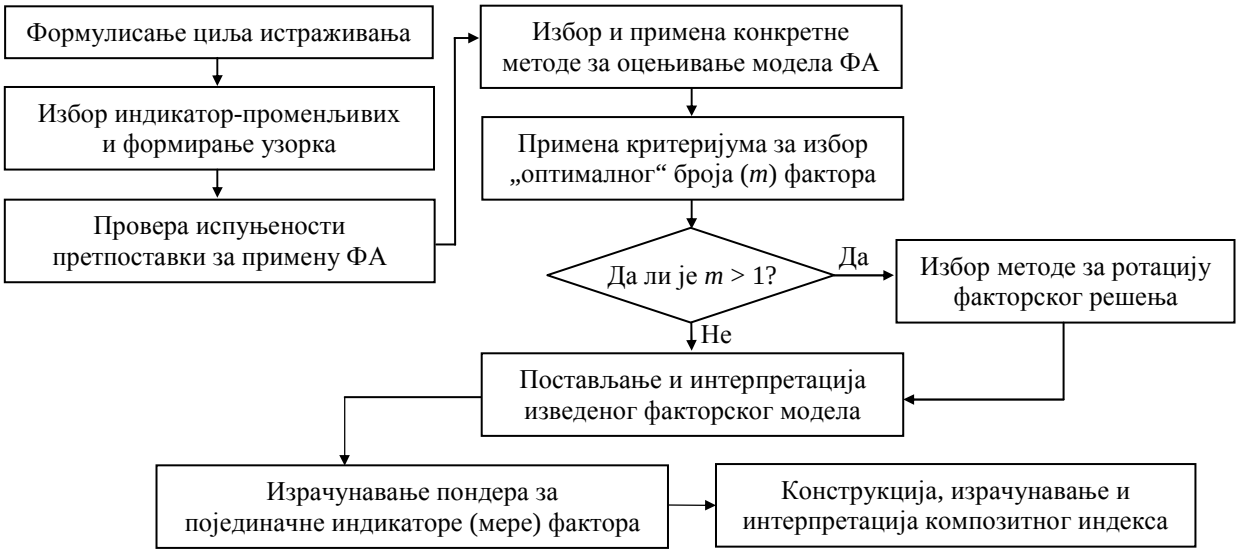
засноване на моделу заједничког(их) фактора и то: ▸ експлоративна факторска анализа (енгл. *Exploratory Factor Analysis, EFA*) и ▸ конфирматорна факторска анализа (енгл. *Confirmatory Factor Analysis, CFA*). Иако, у начелу, и *EFA* и *CFA* имају за циљ извођење одређеног факторског модела којим се описују присутне релације унутар скупа оригиналних индикатор-променљивих коришћењем мањег броја латентних променљивих, наведени типови ФА разликују се суштински са аспекта и у погледу броја и природе *a priori* спецификација и ограничења везаних за модел ФА (*Brown*, 2015, стр. 11). Наиме, у случају тзв. истраживачке примене ФА[8], аналитичар не мора да располаже никаквим конкретним претходним сазнањима и, на њима заснованим, очекивањима у погледу (*Sharma*, 1996, стр. 128): ▸ броја заједничких фактора који конституишу разматрани латентни феномен; ▸ броја индикатор-променљивих „груписаних" око сваког од издвојених (једног или више) фактора; или пак ▸ прецизне (појединачне) идентификације индикатора који репрезентују, односно представљају поуздану меру сваког, појединачно посматраног, заједничког фактора. Другим речима, *EFA* представља дескриптивну статистичку процедуру која се примењује за потребе истраживања извора корелација између мерљивих индикатора, без претходног специфицирања било каквих теоријских претпоставки у погледу структуре факторског модела, односно броја заједничких фактора и / или релација између заједничких фактора и индикатор-променљивих (*Kovačić*, 1994, стр. 217; *Đorđević i drugi*, 2011, стр. 140; *Thompson*, 2004, стр. 5; *Everitt*, 2010, стр. 211). Будући да се расположиви узорак мултиваријационих података користи за идентификовање (унапред непознате) структуре факторског модела, *EFA* се заправо третира као поступак вођен подацима (*Timm*, 2002, стр. 496; *Brown*, 2015, стр. 11). С друге стране, *CFA* представља инференцијалну статистичку процедуру која се користи за потребе емпиријске провере (то јест, негирања или потврде) одрживости и валидности *a priori* формулисане прецизне теоријске спецификације хипотетичке структуре одговарајућег факторског модела (*Sharma*, 1996, стр. 128; *Timm*, 2002, стр. 496; *Đorđević i drugi*, 2011, стр. 140). Заправо, *CFA* омогућава директну евалуацију прилагођености унапред дефинисаног факторског решења расположивим емпиријским подацима (*Thompson*, 2004, стр. 6), односно сагледавање да ли и у којој мери тестирана факторска структура добро одражава и репрезентује релације присутне унутар узорачке корелационе (или коваријационе) матрице опсервабилних индикатор-променљивих (*Brown*, 2015, стр. 11; *Everitt*, 2010, стр. 211).

Полазећи од карактеристика представљене две варијанте ФА, а у складу са методолошким потребама емпиријског истраживања реализованог у последњем поглављу дисертације, у наставку излагања модел заједничког фактора разматран је искључиво у контексту *EFA*-е. Такође, имајући у виду чињеницу да комплетан и потпун опис методолошке процедуре у основи *EFA*-е превазилази оквире ове докторске дисертације[9], на Слици 2.1.2. приказан је дијаграм тока поступка спровођења експлоративне ФА у функцији конструкције композитног индекса намењеног квантитативној оцени вредности латентне променљиве

[8] Према мишљењу *Thompson*-а (2004, стр. 5), експлоративна примена ФА превасходно обухвата типове факторске анализе попут оних оригинално предложених од стране *Spearman*-а (1904). С друге стране, методи конфирматорне ФА развијени су знатно касније, крајем 60-их и почетком 70-их година XX века (видети детаљније нпр. *Joreskog*, 1969).

[9] Генерално, имплементација *EFA* подразумева и обухвата читав низ комплексних одлука које је неопходно донети, односно избора које је потребно учинити, у циљу адекватног усмеравања и ефикасног управљања поступком креирања одговарајућег модела ФА (*Thompson*, 2004, стр. 27).

од интереса, односно одређивања висине тежинских коефицијената (пондера) који стоје уз појединачне индикаторе у саставу креиране мултиваријационе линеарне функције.



Слика 2.1.2. Шематски приказ поступка спровођења *EFA*-е у функцији креирања композитне мере латентне променљиве од интереса

Излагањем у оквиру Одељка 2.1.2, обухваћена је дискусија о кључним статистичким претпоставкама на којима се заснива валидна експлоатација *EFA*-е, као параметарског мултиваријационог статистичког метода. Сажети прикази кључних одређења најчешће коришћених метода за оцену факторског модела, односно екстракцију заједничког(их) фактора, као и процедура и критеријума за избор „оптималног“ броја заједничких фактора који ће бити издвојени, представљају садржај Одељка 2.1.3 и Одељка 2.1.4, респективно. Методолошко објашњење поступка конструкције композитног индикатора на бази резултата факторске анализе презентирано је у оквиру последњег одељка овог поглавља.

2.1.2. Статистичке претпоставке за примену факторске анализе

У циљу обезбезбеђивања статистички валидне примене експлоративне факторске анализе и, консеквентно, научне заснованости и генерализације резултирајућих закључака у контексту анализиране проблематике, неопходно је проверити и осигурати одговарајући степен испуњености следећих, у литератури најчешће апострофираних, претпоставки на нивоу узорка мултиваријационих опсервација:

- ✓ Униваријациона нормалност распореда (појединачних) индикатор-променљивих;
- ✓ Мултиваријациона нормалност (заједничког) распореда индикатор-променљивих[10];
- ✓ Одсуство униваријационих и мултиваријационих нетипичних опсервација;
- ✓ Постојање статистички значајне линеарне везе између парова променљивих;
- ✓ Одсуство мултиколинеарности и сингуларности између индикатор-променљивих; и
- ✓ Обезбеђивање „адекватне“ величине узорка мултиваријационих података.

[10] *Bartholomew* (2008, стр. 175) истиче да примена ФА, као мултиваријационе технике засноване на моделу (енгл. *model-based technique*), нужно подразумева испуњеност претпоставке која се тиче заједничког распореда анализираних променљивих на нивоу популације. Значај и важност претпоставке о мултиваријационој (и униваријационој) нормалности у контексту имплементације ФА, такође истичу и елаборирају и следећи аутори: *Yong & Pearce* (2013, стр. 80), *Zygmont & Smith* (2014, стр. 41), *Kasper & Ünlü* (2013, стр. 4).

Будући да корелациона матрица представља уобичајену полазну тачку у примени ФА (*Bartholomew*, 2008, стр. 183), централно место у наведеним предусловима управо припада процедури испитивања да ли између анализираних индикатор-променљивих постоји статистички значајна линеарна корелација и, ако постоји, каквог је степена (*Watkins*, 2018, стр. 226). Наведена констатација произилази из чињенице да (статистички значајна) линеарна корелисаност непрекидних нумеричких променљивих индиректно имплицира могућност њиховог „груписања“ (применом ФА) у једну или више релативно хомогених група индикатора, тако да променљиве унутар сваке групе представљају доминантне мере (показатеље) исте неопсервабилне димензије, односно заједничког фактора (*Sharma*, 1996, стр. 116). Логично посматрано, ако анализиране индикатор-променљиве нису у довољној мери међусобно корелисане, мало је вероватно да исте „деле“ заједнички фактор, чиме се озбиљно доводи у питање оправданост примене ФА, а тиме и корисност добијених резултата (*Bartholomew*, 2008, стр. 183; *Everitt*, 2010, стр. 229; *Pituch & Stevens*, 2016, стр. 341; *Hair et al.*, 2010, стр. 102; *Watkins*, 2018, стр. 226; *Johnson & Wichern*, 2007, стр. 488).
Први корак у поступку сагледавања и испитивања степена адекватности и погодности расположиве ($p \times p$) корелационе матрице за оправдану примену ФА, по правилу представља визуелна (сумарна) евалуација израчунатих, на нивоу узорка, вредности *Pearson*-ових коефицијената просте линеарне корелације (у ознаци, r) и интерпретација добијених резултата спроведеног тестирања хипотезе о њиховој појединачној статистичкој значајности на нивоу популације (*Sharma*, 1996, стр. 116; *Bartholomew*, 2008, стр. 183). У том контексту, као својеврсна подршка у превазилажењу формулисане методолошке дилеме, у литератури се најчешће наводе следеће препоруке формулисане на бази искуства, односно искуствена правила и неформални критеријуми:

- Највећи број вредности коефицијента просте линеарне корелације треба да буде изнад ±0,30, односно симболички: $r \geq | \pm 0{,}30 |$ (*Hair et al.*, 2010, стр. 102; *Yong & Pearce*, 2013, стр. 81; *Zygmont & Smith*, 2014, стр. 44; *Watkins*, 2018, стр. 226).

- При тестирању хипотезе о статистичкој значајности оцене коефицијента просте линеарне корелације, алтернативна хипотеза, којом се претпоставља да између датог пара индикатор-променљивих постоји линеарно квантитативно слагање на нивоу популације из које је узорак извучен (симболички: H_1: $\rho \neq 0$), мора бити усвојена уз одговарајући ризик грешке α. Заправо, променљиве које се не одликују статистички значајним линеарним корелацијама са преосталим индикаторима не би требало да буду обухваћене анализом усмереном на екстракцију заједничког(их) фактора (*Varmuza & Filzmoser*, 2009, стр. 81; *Hair et al.*, 2010, стр. 128).[11]

- Будући да „екстремно“ високе вредности коефицијената линеарне корелације ($r > | \pm 0{,}80|$ или $| \pm 0{,}90 |$) указују на могуће присуство мултиколинеарности или сингуларности у подацима, неопходно је размотрити могућност елиминисања из анализе индикатора који се сматра(ју) узроком наведених методолошких ограничења (*Yong & Pearce*, 2013, стр. 88; *Zygmont & Smith*, 2014, стр. 44).

[11] У контексту изнете искуствене сугестије, уколико између парова анализом обухваћених индикатор-променљивих доминира одсуство статистички значајне линеарне корелације *Watkins* (2018, стр. 224) препоручује употребу показатеља криволинијског (нелинеарног) квантитативног слагања уместо *Pearson*-овог коефицијента линеарне корелације.

Такође, посматрано из угла осталих претпоставки за валидну примену ФА, важно је истаћи да се (евентуална) нарушеност статистичке претпоставке о нормалности мултиваријационог распореда анализираних индикатор-променљивих негативно одражава на вредности *Pearson*-овог коефицијента линеарне корелације (то јест, умањује степен емпиријски утврђеног линеарног квантитативног слагања), а тиме, консеквентно, утиче и на резултате спроведене експлоративне факторске анализе (*Hair et al.*, 2010, стр. 102; *Zygmont & Smith*, 2014, стр. 41; *Watkins*, 2018, стр. 223).

У том смислу, *Watkins* (2018, стр. 224) истиче да у ситуацијама када претпоставка о нормалности није задовољена, *Pearson*-ов коефицијент линеарне корелације *r* не представља најпогоднији улазни елемент за спровођење *EFA* и, сходно томе, препоручује коришћење неких других, робустнијих (непараметарских) показатеља корелације између променљивих. Поред наведеног, присуство униваријационих и / или мултиваријационих нетипчних опсервација, поред њиховог неспорног утицаја на дистрибуциона својства анализираних нумеричких променљивих, такође, према мишљењу *Zygmont*-а *& Smith*-а (2014, стр. 42), може потенцијално нарушити прецизност оцена коефицијената линеарне корелације, а тиме и, као што је истакнуто, оцена непознатих параметара модела ФА.

Други корак у оцени степена међусобне повезаности индикатор-променљивих и могућности сумаризације информација садржаних у корелационој матрици иницијалних индикатора у виду заједничког(их) фактора, подразумева примену и интерпретацију формалних и објективнијих статистичких процедура попут *Kaiser-Meyer-Olkin*-ове мере адекватности узорка (енгл. *Kaiser-Meyer-Olkin measure of sampling adequacy*, *KMO-MSA*) и *Bartlett*-ов тест сферичности (енгл. *Bartlett's sphericity test*)[12], које представљају предмет анализе у наставку.

- Заснована на поређењу утврђених коефицијената просте линеарне и парцијалне корелације, *KMO* мера адекватности узорка представља статистички показатељ намењен сагледавању степена у којем су присутне корелације на нивоу расположивог узорка заправо функција варијансе заједничке свим индикаторима, пре него варијансе коју деле и која је својствена конкретним, појединачним паровима индикатор-променљивих (*Watkins*, 2018, стр. 226; *Zygmont & Smith*, 2014, стр. 44). Као специфични показатељ хомогености *p* индикатора (*Sharma*, 1996, стр. 116), *KMO* мера адекватности узорка на нивоу комплетне корелационе матрице (енгл. *Overall KMO-MSA*) утврђује се путем следећег израза (*Rencher*, 2002, стр. 445):

$$Overall\ KMO\text{-}MSA = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} q_{ij}^2}, \text{ за } i,j = 1, 2, \ldots, p \text{ и } i \neq j, \qquad (2.1.4)$$

где су r_{ij} вредности коефицијента просте линеарне корелације (то јест, елементи узорачке корелационе матрице $\mathbf{R}=[r_{ij}]$), док симболи q_{ij} означавају вредности коефицијента парцијалне корелације (то јест, елементи матрице $\mathbf{Q} = \mathbf{DR}^{-1}\mathbf{D} = [q_{ij}]$, при чему је $\mathbf{D} = [(\text{diag}\mathbf{R}^{-1})^{1/2}]^{-1}$). Вредности овог показатеља крећу се у интервалу од 0,00 до 1,00, при чему

[12] Предности наведених статистичких процедура нарочито долазе до изражаја у ситуацијама када је анализом обухваћен прилично велики број индикатор-променљивих, будући да је у таквим околностима реализација (генерално субјективне) визуелне „инспекције" корелационе матрице изразито отежана и комплексна (*Sharma*, 1996, стр. 116).

важи релација: када $\mathbf{R}\rightarrow\mathbf{I}$,[13] тада *Overall KMO-MSA* → 0, и обратно (*Timm*, 2002, стр. 507) [14]. Прецизније, за потребе интерпретације израчунатих вредности разматране мере у контексту формулисања закључка о степену адекватности узорка (односно, присутне „јачине“ корелација између индикатора) из угла оправдане имплементације ФА, у литератури се наводи следећа скала вредности са припадајућим тумачењима, предложена од стране *Kaiser*-а *& Rice*-а (1974, цитирано у: *Sharma*, 1996, стр. 116; *Timm*, 2002, стр. 507; *Watkins*, 2018, стр. 226): ▸ $MSA \geq 0{,}90$ (надпросечно); ▸ $0{,}90 > MSA \geq 0{,}80$ (изнад просека); ▸ $0{,}80 > MSA \geq 0{,}70$ (просечно); ▸ $0{,}70 > MSA \geq 0{,}60$ (испод просека); ▸ $0{,}60 > MSA \geq 0{,}50$ (знатно испод просека); ▸ $0{,}50 > MSA$ (неприхватљиво). Иако (углавном субјективно) дефинисана гранична вредност *Overall KMO* мере при доношењу позитивне одлуке о погодности анализираног узорка варира од истраживача до истраживача[15], у литератури је начелно присутна сагласност око става да вредност MSA=0,50 представља општи минимум који мора бити задовољен како би се примена ФА третирала као оправдана и ефективна (*Hair et al.*, 2010, стр. 103; *Zygmont & Smith*, 2014, стр. 44; *Yong & Pearce*, 2013, стр. 88; *Watkins*, 2018, стр. 227).

Поред наведеног, вредности мере адекватности узорковања могуће је утврдити и на нивоу сваке индикатор-променљиве појединачно (енгл. *variable-specific measures of intercorrelation*), а у циљу идентификовања променљиве(их) која(е) није(су) у довољној мери корелисана(е) са преосталим индикаторима, коришћењем следећег израза:

$$MSA_j = \frac{\sum_{j\neq i} r_{ij}^2}{\sum_{j\neq i} r_{ij}^2 + \sum_{j\neq i} q_{ij}^2}\text{, за } i,j = 1, 2,\ldots, p, \text{ и } i \neq j, \text{ при чему } r_{ij} \in \mathbf{R} \text{ и } q_{ij} \in \mathbf{Q}. \qquad (2.1.5)$$

Интерпретација појединачних MSA_j вредности врши се на идентичан начин као и вредности претходно разматране *Overall KMO-MSA* мере. Значај показатеља MSA_j нарочито долази до изражаја у ситуацији када је добијена вредност *Overall KMO* мере адекватности комплетног узорка испод постављеног минимално прихватљивог нивоа. Наиме, променљиве X_j које се одликују вредностима $MSA_j < 0{,}50$ представљају кандидате за искључивање из даље анализе чиме се, у крајњој мери, може утицати на повећање вредности *Overall KMO* мере, иницијално утврђене испод минимално прихватљивог нивоа (*Sharma*, 1996, стр. 116; *Hair et al.*, 2010, стр. 103).

- *Bartlett*-ов тест сферичности представља формалну статистичку процедуру која се користи за тестирање валидности нулте хипотезе, чији садржај може бити формулисан на један од следећих (алтернативних) начина (*Sharma*, 1996, стр. 123; *Pituch & Stevens*, 2016, стр. 341; *Zygmont & Smith*, 2014, стр. 44; *Thompson*, 2004, стр. 31):

H_0: Узорачка корелациона матрица потиче из популације у којој анализиране индикатор-променљиве нису међусобно (статистички значајно) корелисане; или

[13] Симбол **I** означава јединичну или матрицу идентитета, односно квадратну матрицу у којој су елементи на главној дијагонали јединице а остали (вандијагонални) елементи нуле. Детерминанта матрице **I** (у ознаци | **I** |) увек је једнака 1.

[14] Као допуна наведеној правилности, *Hair et al.* (2010, стр. 103) наводе и следеће „факторе“ који могу условити (у већој или мањој мери) повећање вредности разматране мере адекватности узорка, и то: (1) повећање величине узорка, (2) повећање просечне вредности коефицијената линеарне корелације, (3) повећање броја анализом обухваћених индикатор-променљивих, и (4) смањење броја присутних заједничких фактора у основи анализиране корелационе матрице.

[15] На пример, *Timm* (2002, стр. 507) и *Rencher* (2002, стр. 445) апострофирају коришћење граничне вредности на нивоу 0,80, за разлику од *Watkins*-а (2018, стр. 226) који као пожељну наводи вредност *Overall KMO* мере која је једнака или већа од 0,70, док се према мишљењу *Sharma*-е (1996, стр. 116) вредности наведене мере изнад 0,60 могу сматрати прихватљивим.

H_0: Корелациона матрица анализираних p индикатор-променљивих на нивоу популације (у ознаци, $\mathbf{R}=[\rho_{ij}]$, за $i,j = 1,2,...,p$) се не разликује (не одступа) од јединичне матрице **I**, симболички: H_0: $|\mathbf{R}| = |\mathbf{I}| = 1$.

Статистика *Bartlett*-овог теста сферичности апроксимативно следи χ^2 распоред са $[p\times(p-1)/2]$ степени слободе и израчунава се коришћењем следеће формуле (*Bartlett*, 1954, стр. 297):

$$\chi^2 = -(n-1-\frac{2p+5}{6})\times\log_e|\mathbf{R}| = \frac{(11+2p-6n)}{6}\times\log_e|\mathbf{R}|, \qquad (2.1.6)$$

где $\log_e|\mathbf{R}|$ представља природни логаритам детерминанте узорачке корелационе матрице, p број индикатор-променљивих X_j, а n величину узорка. Уколико је резултирајућа p-вредност мања од *a priori* дефинисаног нивоа значајности теста α, H_0 се одбацује и прихвата алтернативна хипотеза (симболички, H_1: $|\mathbf{R}| \neq |\mathbf{I}| \neq 1$.) која сугерише да се корелациона матрица на нивоу популације из које је изорак извучен статистички значајно разликује од јединичне матрице **I**, а тиме и, консеквентно, да се употреба ФА коришћењем расположиве корелационе матрице може сматрати оправданом. У супротном, може се закључити, уз одговарајући ризик грешке α, да су ненулте вредности коефицијената линеарне корелације у узорачкој корелационој матрици заправо резултат узорачке грешке.

Коначно, полазећи од чињенице да величина узорка у значајној мери може утицати на прецизност примењених статистичких поступака за оцењивање непознатих вредности параметара модела ФА, као и других формалних процедура статистичког закључивања (*Thompson*, 2004, стр. 24), неопходно је адекватну пажњу посветити разматрању питања која се тичу обезбеђивања неопходне величине узорка, n. У литератури су предложена бројна и различита искуствена правила заснована на исказивању „пожељне“ величина n у апсолутном укупном износу или пак као функције броја опсервација и анализираних индикатор-променљивих (односно, у форми односа n / p). Сходно наведеном, довољном и пожељном величином узорка за спровођење ФА, генерално, сматра се величина $n \geq 100$ јединица посматрања, док се минимално прихватљива величина најчешће дефинише на нивоу $n = 50$ опсервација (*Hair et al.*, 2010, стр. 101; *Zygmont & Smith*, 2014, стр. 41). Узорци којима је обухваћено 200 и више јединица посматрања сматрају се великим узорцима (*Zygmont & Smith*, 2014, стр. 41). Посматрано из угла броја анализираних индикатора, препоруке су углавном усмерене у правцу обезбеђивања величине узорка којом ће бити задовољен критеријум представљен рационом n / $p \approx 5 : 1$, односно 5 опсервација по свакој променљивој X_j. Наведени количник представља минимално прихватљив однос броја јединица посматрања и броја анализираних променљивих (*Thompson*, 2004, стр. 24; *Hair et al.*, 2010, стр. 101; *Yong & Pearce*, 2013, стр. 80; *Zygmont & Smith*, 2014, стр. 41; *Pituch & Stevens*, 2016, стр. 347). Најчешће прихватљива величина узорка подразумева постизање односа n / p на нивоу 10 : 1 (*Zygmont & Smith*, 2014, стр. 41; *Hair et al.*, 2010, стр. 101; *Yong & Pearce*, 2013, стр. 80), при чему поједини истраживачи препоручују и однос 20 : 1 (*Thompson*, 2004, стр. 24; *Hair et al.*, 2010, стр. 101; *Pituch & Stevens*, 2016, стр. 347) па чак и 30 : 1 (*Yong & Pearce*, 2013, стр. 80) у корист броја јединица посматрања обухваћених анализом.

Уколико расположиви узорак мултиваријационих опсервација и на њему заснована матрица коефицијената просте линеарне корелације **R** задовољавају, у одговарајућем

степену, елабориране статистичке претпоставке, у наставку анализе, сходно поступку представљеном на Слици 2.1.2, врши се избор и примена одговарајуће методе за оцењивање непознатих параметара модела факторске анализе, дефинисаног изразом 2.1.1.

2.1.3. Методе за оцењивање модела факторске анализе

У релевантној литератури постоји читав низ предложених метода намењених за оцењивање непознатих параметара модела факторске анализе, при чему се као најчешће коришћени издвајају следећи[16] (*Kovačić*, 1994, стр. 224; *Rencher*, 2002, стр. 415; *Johnson & Wichern*, 2007, стр. 488): метод главних компонената (енгл. *principal component method*, *PCm*), метод главних фактора (енгл. *principal factor method*, *PFm*) и метод највеће веродостојности (енгл. *maximum likelihood method*, *MLm*). Независно од изабраног метода, поступак издвајања заједничких фактора и креирање модела ФА захтева итеративно спровођење комплексних израчунавања, које се, због ефикасности и прецизности, углавном заснива на употреби рачунара, односно апликативних могућности расположивих комерцијалних статистичких програмских пакета (*Johnson & Wichern*, 2007, стр. 488). Међутим, у циљу адекватног разумевања статистичке логике у основи ФА, а тиме и смислене интерпретације резултирајућег модела, корисним (или боље речено, нужним) се сматра поседовање знања у погледу кључних концептуално-методолошких одређења и рачунских специфичности имплементације конкретно изабране процедуре оцењивања. У том смислу, полазећи од чињенице да комплетан опис наведених метода превазилази оквир ове дисертације, а у складу са методолошким потребама реализованог емпиријског истраживања, даљим излагањем елабориран је, у квантитативној форми, поступак примене методе главних компонената у контексту оцене факторских оптерећења (то јест, комуналитета и специфичних варијанси) изведеног модела ФА.[17] [18]

Генерално, математички приступ решавању проблема одређивања оцењених вредности факторских оптерећења модела ФА (израз 2.1.1), у контексту примене методе главних компонената, започиње проналажењем одговарајућих карактеристичних вредности (енгл. *eigenvalues*) λ_f корелационе матрице $\mathbf{R}$[19] за које је задовољена наредна једнакост:

$$\left|\mathbf{R}-\lambda_f\mathbf{I}\right| = \left|\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} - \lambda_f \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}\right| = \left|\begin{bmatrix} r_{11}-\lambda_f & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22}-\lambda_f & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp}-\lambda_f \end{bmatrix}\right| = 0, \quad (2.1.7)$$

---

[16] У литератури се наводе и следеће математичко-статистичке процедуре за екстракцију заједничких фактора присутних у структури анализираног *p*-димензионог узорка опсервација, попут: метод пондерисаних најмањих квадрата (енгл. *weighted least squares method*), метод непондерисаних најмањих квадрата (енгл. *unweighted least squares method*), метод каноничке ФА (енгл. *canonical factor analysis method*), метод центроида (енгл. *centroid method*), итд.

[17] Детаљна дискусија о методи главних фактора садржана је, на пример, у: *Johnson & Wichern* (2007, стр. 494–495), *Rencher* (2002, стр. 421–425), *Kovačić* (1994, стр. 225–227), *Sharma* (1996, стр. 107–108).

[18] Користан преглед кључних карактеристика методе највеће веродостојности обезбеђен је од стране бројних аутора, попут: *Johnson & Wichern* (2007, стр. 495–497), *Rencher* (2002, стр. 425–426), *Lawley & Maxwell* (1962, стр. 215–217), *Kovačić* (1994, стр. 231–233).

[19] Поступак оцењивања факторских оптерећења заснован на примени методе главних компонената могуће је, на идентичан начин, спровести и коришћењем узорачке коваријационе матрице, у ознаци **S**, једноставном заменом матрице **R** матрицом **S** у структури презентираних израза (*Johnson & Wichern*, 2007, стр. 489). У контексту изнете констатације, важно је истаћи да се у пракси, генерално, при оцени модела ФА чешће употребљава корелациона наспрам коваријационе матрице (*Rencher*, 2002, стр. 418).

при чему симболи $\{r_{ij}\}$ за $i, j = 1, 2,..., p$, означавају елементе симетричне ($p \times p$) узорачке корелационе матрице **R**, а λ_f вредност f-тог (за $f = 1, 2,..., m,..., p$) карактеристичног корена матрице **R**, и p – број индикатор-промењљивих. Детерминанта „новоформиране“ матрице $[\mathbf{R}-\lambda_f\mathbf{I}]$, у ознаци $|\mathbf{R}-\lambda_f\mathbf{I}|$ је заправо карактеристични полином реда p по λ_f, који се може приказати следећим изразом:

$$\left|\mathbf{R}-\lambda_f\mathbf{I}\right| = (-1)^p\left[\lambda_f^p + c_1\lambda_f^{p-1} + c_2\lambda_f^{p-2} + \ldots + c_p\lambda_f + c_{p+1}\right] = 0. \qquad (2.1.7a)$$

Решавањем представљене карактеристичне једначине p-тог реда (енгл. *characteristic equation*) могуће је одредити (максимално) p позитивних карактеристичних вредности λ_f за које је дата једнакост задовољена, при чему важи: $\lambda_1 \geq ... \geq \lambda_m \geq ... \geq \lambda_p \geq 0$. (*Johnson & Wichern*, 2007, стр. 488). Након одређивања карактеристичних вредности следи израчунавање кореспондентних, односно појединачним λ_f вредностима придружених, ($p \times 1$) карактеристичних вектора (енгл. *eigenvectors*) $\mathbf{a}_f^*$ матрице **R**, решавањем одговарајућег хомогеног система једначина на начин којим се обезбеђује ненарушавање једнакости представљене следећим изразом:

$$\underset{(p\times p)}{[\mathbf{R}-\lambda_f\mathbf{I}]}\,\underset{(p\times 1)}{\mathbf{a}_f^*} = \begin{bmatrix} r_{11}-\lambda_f & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22}-\lambda_f & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp}-\lambda_f \end{bmatrix} \times \begin{bmatrix} a_{1f}^* \\ a_{2f}^* \\ \vdots \\ a_{pf}^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \qquad (2.1.8)$$

Нормирањем карактеристичних вектора $\mathbf{a}_f^*$ за сваки од (иницијално идентификованих) p појединачно посматраних карактеристичних корена λ_f, утврђују се елементи ($p \times 1$) вектора $\mathbf{a}_f' = [a_{1f}, a_{2f},\ldots, a_{pf},]$ као основе за израчунавање оцењених вредности факторских оптерећења траженог факторског модела.

Прецизније, полазећи од парова појединачних карактеристичних вредности и придружених нормализованих карактеристичних вектора (енгл. *eigenvalue – eigenvector pairs*) матрице **R**, у ознаци (λ_f, $\mathbf{a}_f$), односно (λ_1, $\mathbf{a}_1$), (λ_2, $\mathbf{a}_2$),..., (λ_m, $\mathbf{a}_m$),..., (λ_p, $\mathbf{a}_p$), матрица оцењених вредности факторских оптерећења, у ознаци $\mathbf{B}=\{b_{jf}\}$, има следећи изглед (*Kovačić*, 1994, стр. 224; *Rencher*, 2002, стр. 417; *Johnson & Wichern*, 2007, стр. 490):

$$\mathbf{B} = \left[\sqrt{\lambda_1}\mathbf{a}_1 \vdots \ldots \vdots \sqrt{\lambda_m}\mathbf{a}_m \vdots \ldots \vdots \sqrt{\lambda_p}\mathbf{a}_p\right] = \left[\underbrace{\sqrt{\lambda_1}\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}}_{\substack{\text{факторска оптерећења} \\ \text{за први фактор } F_1}} \vdots \ldots \vdots \underbrace{\sqrt{\lambda_m}\begin{bmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{pm} \end{bmatrix}}_{\substack{\text{факторска оптерећења} \\ \text{за } m\text{-ти фактор } F_m}} \vdots \ldots \vdots \underbrace{\sqrt{\lambda_p}\begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix}}_{\substack{\text{факторска оптерећења} \\ \text{за } p\text{-ти фактор } F_p}}\right]. \qquad (2.1.9)$$

Наиме, применом методе главних компонената, вредности оцена факторских оптерећења представљају резултат скалирања појединачних (нормализованих) карактеристичних вектора ($\mathbf{a}_f$) квадратним кореном кореспондентних карактеристичних вредности (λ_f) матрице **R** (*Johnson & Wichern*, 2007, стр. 493). Резултирајућим, на основу изложеног поступка оцењеним, иницијалним факторским моделом (израз 2.1.10) у потпуности се

објашњава корелациона структура (представљена матрицом **R**) између анализираних индикатора, будући да је број обухваћених фактора F_f еквивалентан броју индикатор-променљивих X_j, симболички $f \equiv j = 1, 2,..., p$. Такође, изведени модел не укључује специфичне факторе индивидуалних индикатора (e_j) пошто, сходно претходној тврдњи, оцењене вредности, њима условљених, специфичних варијанси износе нула за сваку j променљиву (*Kovačić*, 1994, стр. 224; *Johnson & Wichern*, 2007, стр. 488).

$$\left.\begin{aligned} X_1 &= b_{11}F_1 + b_{12}F_2 + ... + b_{1m}F_m + ... + b_{1p}F_p \\ X_2 &= b_{21}F_1 + b_{22}F_2 + ... + b_{2m}F_m + ... + b_{2p}F_p \\ &\vdots \\ X_p &= b_{p1}F_1 + b_{p2}F_2 + ... + b_{pm}F_m + ... + b_{pp}F_p \end{aligned}\right\} X_j = \sum_{f=1}^{p} b_{jf}F_f \ , \quad \text{за } j = 1, 2,..., p \text{ и } e_j = 0. \qquad (2.1.10)$$

С обзиром да приказани факторски модел (израз 2.1.10), сачињен од p фактора, не обезбеђује редукцију оригиналне димензионалности конкретно разматраног истраживачког проблема, његова употребна вредност може се сматрати „у најмању руку“ дискутабилном. Логично, у контексту формулисаног примарног циља ФА, неопходно је извршити редефинисање иницијално развијеног факторског модела избором „оптималног“ броја заједничких фактора, m, тако да је $1 \leq m < p$.[20] Будући да, за разлику од p-факторског модела, не обезбеђује објашњење укупног већ највећег (односно, довољно великог) удела укупног (заједничког) варијабилитета скупа анализираних индикатора [21], оцењени (редуковани) m-факторски модел, представљен изразом 2.1.11, својом структуром, поред заједничких фактора F_f (за $f = 1, 2,..., m$) обухвата и специфичне факторе појединачних променљивих, e_j (за $j = 1, 2,..., p$).

$$\left.\begin{aligned} X_1 &= b_{11}F_1 + b_{12}F_2 + ... + b_{1m}F_m + e_1 \\ X_2 &= b_{21}F_1 + b_{22}F_2 + ... + b_{2m}F_m + e_2 \\ &\vdots \\ X_p &= b_{p1}F_1 + b_{p2}F_2 + ... + b_{pm}F_m + e_p \end{aligned}\right\} X_j = \sum_{f=1}^{m} b_{jf}F_f + e_j \ , \quad \text{за } j = 1, 2,..., p. \qquad (2.1.11)$$

Посматрано у контексту презентираног m-факторског модела, а у складу са методолошким напоменама представљеним у Одељку 2.1.1, оцењене вредности комуналитета појединачних индикатора утврђују се као сума квадрата вредности факторских оптерећења садржаних у кореспондентном j-том реду редуковане (са ($p \times p$) на ($p \times m$)) матрице **B**, представљене изразом (2.1.9), односно коришћењем следеће формуле (*Kovačić*, 1994, стр. 225; *Johnson & Wichern*, 2007, стр. 490; *Rencher*, 2002, стр. 418):

$$\hat{h}_j^2 = \sum_{f=1}^{m} b_{jf}^2 = b_{j1}^2 + b_{j2}^2 + ... + b_{jm}^2 \ , \quad \text{за } j = 1, 2,..., p. \qquad (2.1.12)$$

Такође, полазећи од претпоставке модела ФА према којој је, као последица извршене Z-стандардизације, укупан варијабилитет сваке променљиве X_j једнак јединици, специфична

[20] Најчешће коришћени приступи и критеријуми намењени свођењу крајње вредности индекса f (употребљеног за означавање издвојених заједничких фактора, F_f) са p на, произвољно означену вредност, m, у ситуацијама када „оптимални“ број заједничких фактора није *a priori* специфициран, представљају предмет разматрања у Одељку 2.1.4.

[21] Кумулативна пропорција укупне (стандардизоване) узорачке варијансе објашњена моделом ФА, заснованом на издвајању m заједничких фактора F_f, утврђује се као (*Johnson & Wichern*, 2007, стр. 492; *Rencher*, 2002, стр. 419): $(\sum_{f=1}^{m} \lambda_f) / p$.

варијанса, $\hat{\psi}_j$, то јест, удео укупне варијансе променљиве X_j који није корелисан и условљен заједничким факторима већ представља резултат дејства специфичног фактора e_j, оцењује се израчунавањем разлике између вредности укупног стандардизованог варијабилитета и оцењене вредности комуналитета на нивоу конкретне променљиве, односно симболички (*Kovačić*, 1994, стр. 225; *Johnson & Wichern*, 2007, стр. 490; *Rencher*, 2002, стр. 418):

$$\hat{\psi}_j = s_j^2 - \hat{h}_j^2 = 1 - \sum_{f=1}^{m} b_{jf}^2 = 1 - (b_{j1}^2 + b_{j2}^2 + ... + b_{jm}^2), \quad \text{за } j = 1, 2, ..., p. \tag{2.1.13}$$

Детаљно и корисно објашњење поступка и статистичке логике на којој се базира оцена степена прилагођености развијеног *m*-факторског модела мултиваријационим опсервацијама и релацијама између њих, односно евалуација квалитета апроксимације анализиране корелационе структуре, може се видети, на пример, у: *Johnson & Wichern* (2007, стр. 490-493), *Barthlomew et al.* (2008, стр. 186-187), *Rencher* (2002, стр. 419). На крају треба истаћи да поред мање изражене рачунске комплексности којом се одликује (*Đorđević i drugi*, 2011, стр. 151), додатни аргумент за избор методе главних компонената при оцењивању факторског модела садржан је и емпиријски верификованом мишљењу, према којем, у условима када су претпоставке везане за оправданост спровођења ФА задовољене, примена, на почетку одељка наведених, најчешће коришћених метода углавном резултира прилично сличним и конзистентним исходима (*Barthlomew et al.*, 2008, стр. 183; *Johnson & Wichern*, 2007, стр. 488).

2.1.4. Критеријуми за избор „оптималног“ броја фактора и интерпретација модела

Као што је већ истакнуто, у околностима када тачан број фактора није унапред прецизиран, након оцењивања факторских оптерећења, а у циљу обезбеђивања адекватне редукције димензионалности оригиналног истраживачког проблема која резултира довољно добром апроксимацијом анализиране коваријационе (**S**) или корелационе матрице (**R**), неопходно је донети одлуку о „оптималном“ броју (*m*) заједничких фактора[22] који ће бити задржани унутар модела ФА за потребе интерпретације и даље анализе. Будући да у великој мери опредељује квалитет изведених закључака, према мишљењу бројних аутора, доношење поменуте одлуке представља критични корак[23] у процесу креирања факторског модела којем треба посветити посебну пажњу (*Kovačić*, 1994, стр. 233; *Thompson*, 2004, стр. 31; *Everitt*, 2010, стр. 217; *Pituch & Stevens*, 2016, стр. 342; *Wilson & Cooper*, 2008, стр. 867). Сходно наведеном, у литератури су предложени бројни, како *ad-hoc* тако и формални, статистички критеријуми и поступци који могу бити коришћени

[22] Термин „оптимални“ заправо означава минимално потребан (или довољан) број заједничких фактора за задржавање (*Tuffery*, 2011, стр. 189). Другим речима, број издвојених заједничких фактора треба да буде, генерално, што је могуће мањи, али ипак довољно велики како би се обезбедила добра апроксимација оригиналних мултиваријационих опсервација и релација између њих, односно како би се формираним моделом осигурала квалитетна апроксимација реалних веза присутних између анализираних променљивих уз минимални губитак информација (*Barthlomew et al.*, 2008, стр. 124).

[23] Избор превеликог, или премалог броја фактора при специфицирању финалне верзије модела ФА може имати негативне, мање или више озбиљне, последице на резултате извршене анализе (*Wilson & Cooper*, 2008, стр. 866; *Đorđević i drugi*, 2011, стр. 153). Прецизније, уколико је изабрани број фактора сувише мали, тада ће значајни заједнички фактори бити изостављени из анализе, а исправност и репрезентативност откривене структуре нарушена, док ће, у супротном случају, избор сувише великог броја фактора у значајној мери отежати поступак интерпретације (*Kovačić*, 1994, стр. 233; *Đorđević i drugi*, 2011, стр. 153).

за доношење одлуке у погледу минимално потребног броја фактора који ће бити екстраховани у оквиру развијеног модела ФА. Међу расположивим приступима, генерално, најширом практичном применом одликују се следећи критеријуми[24]:

- *Kaiser-Guttman*-ово *правило* или *критеријум карактеристичне вредности* (енгл. *Kaiser-Guttman rule* / *latent root criterion*)

Према *Kaiser–Guttman*-овом правилу, у коначном моделу се задржава онолико заједничких фактора (у ознаци, *m*) колико има иницијално утврђених карактеристичних корена (у ознаци, λ_f) узорачке корелационе или коваријационе матрице чије су појединачне вредности веће од аритметичке средине свих *p* карактеристичних вредности, утврђене изразом $\bar{\lambda} = \Sigma\lambda_f / p$ (за $f = 1, 2,..., p$), где *p* означава број индикатор-променљивих (*Kovačić*, 1994, стр. 234; *Rencher*, 2002, стр. 427). У случају када се оцењивање параметара модела ФА заснива на употреби корелационе матрице (**R**), просек ($\bar{\lambda}$) карактеристичних вредности (λ_f) увек износи 1,[25] због чега се разматрани критеријум често у литератури назива и правило „карактеристична вредност > 1,0“ (*Kovačić*, 1994, стр. 207; *Johnson & Wichern*, 2007, стр. 491; *Barthlomew et al.*, 2008, стр. 124; *Pituch & Stevens*, 2016, стр. 342; *Brown*, 2015, стр. 23; *Thompson*, 2004, стр. 32).

Логика на којој се заснива примена овог критеријума директно је повезана са разумевањем улоге и значења карактеристичних вредности (λ_f) конкретно разматране матрице (**S** или **R**), у контексту сагледавања перформанси изведеног модела. У том смислу, дефинишући укупну узорачку варијансу као збир варијанси појединачних индикатор-променљивих,[26] појединачне карактеристичне вредности заправо репрезентују део укупне варијансе индикатора на нивоу узорка (у ознаци, Σs_{jj}^2, за $j = 1, 2,..., p$) који је објашњен или обухваћен сваким, појединачно посматраним, фактором F_f (*Brown*, 2015, стр. 22; *Tuffery*, 2011, стр. 189; *Barthlomew et al.*, 2008, стр. 124), односно симболички (*Johnson & Wichern*, 2007, стр. 491; *Brown*, 2015, стр. 22):

$$\lambda_f = (\sqrt{\lambda_f}\,\mathbf{a}_f)'(\sqrt{\lambda_f}\,\mathbf{a}_f) = \hat{h}_{1f}^2 + \hat{h}_{2f}^2 + ... + \hat{h}_{pf}^2 = b_{1f}^2 + b_{2f}^2 + ... + b_{pf}^2 \text{, за } j=1,2,...,p \text{ и } f=1,...,m,...,p, \quad (2.1.14)$$

где $\hat{h}_{1f}^2, \hat{h}_{2f}^2, ..., \hat{h}_{pf}^2$ означавају оцењене вредности комуналитета на нивоу сваког појединачног индикатора и конкретно посматраног *f*-тог заједничког фактора, а $b_{1f}^2, b_{2f}^2, ..., b_{pf}^2$ вредности кореспондентних факторских оптерећења у оцењеном моделу.

---

[24] Поред елабориранх мање формалних критеријума, у литератури се такође наводи (а у пракси неретко користи) и приступ заснован на спровођењу формалне процедуре тестирања статистичке хипотезе о „оптималном“ броју заједничких фактора (за детаљно објашњење видети: *Kovačić*, 1994, стр. 234–235; *Rencher*, 2002, стр. 428; *Johnson & Wichern*, 2007, стр. 501–503; *Pituch & Stevens*, 2016, стр. 343; *Lawley & Maxwell*, 1962, стр. 218–219; *Everitt*, 2010, стр. 218), као и приступ познат под називом „паралелна анализа“ (објашњење методолошких карактеристика видети у: *Thompson*, 2004, стр. 34–36; *Pituch & Stevens*, 2016, стр. 343; *Brown*, 2015, стр. 24–25; *Wilson & Cooper*, 2008, стр. 867).

[25] Наведена констатација произилази из чињенице да је збир карактеристичних вредности ($\Sigma\lambda_f$) корелационе матрице (**R**), као бројилац израза за аритметичку средину ($\bar{\lambda}$), једнак трагу матрице **R**, симболички $tr(\mathbf{R}) = r_{11}+r_{22}+...+r_{pp} \rightarrow tr(\mathbf{R}) = p$, будући да је $r_{11} = r_{22} = ... = r_{pp} = 1$. Аналогно, у случају када се креирање модела ФА заснива на коришћењу коваријационе матрице (**S**), збир утврђених карактеристичних вредности одговара трагу матрице **S**, односно $tr(\mathbf{S}) = s_{11}^2 + s_{22}^2 + ... + s_{pp}^2$, а просек ($\bar{\lambda}$) ће износити такође 1, под условом да је анализа спроведена на стандардизованим вредностима променљивих, односно уколико је варијанса анализираних појединачних индикатора једнака јединици ($s_{11}^2 + s_{22}^2 + ... + s_{pp}^2 = 1$).

[26] Уколико су оригинални индикатори стандардизовани, укупна варијанса на нивоу узорка једнака је броју индикатора, *p*, као и збир карактеристичних вредности ($\Sigma\lambda_f$) матрице **R** или **S** (*Barthlomew et al.*, 2008, стр. 124). Консеквентно, „допринос“ сваке појединачне индикатор-променљиве поменутом збиру једнак је 1 (*Đorđević i drugi*, 2011, стр. 151).

Консеквентно, пропорција укупне узорачке варијансе објашњена *f*-тим фактором може се дефинисати на следећи начин (*Thompson*, 2004, стр. 21; *Johnson & Wichern*, 2007, стр. 491; *Barthlomew et al.*, 2008, стр. 124):

$$\lambda_f / \sum_{f=1}^{p} \lambda_f = \lambda_f / p \,. \qquad (2.1.15)$$

Полазећи од наведеног, логика у основи примене *Kaiser–Guttman*-овог правила при избору броја заједничких фактора, *m*, садржана је у поређењу „величине“ дела укупне узорачке варијансе који је објашњен (у ознаци, λ_f) сваким од појединачних фактора F_f (као кандидата за задржавање у моделу) и стандардизованог износа варијансе индивидуалних индикатор-променљивих (односно, просек карактеристичних вредности, $\bar{\lambda}$=1). Прецизније, према овом критеријуму, потребно је да део укупне узорачке варијансе који је обухваћен и објашњен конкретним заједничким фактором (у ознаци, λ_f) буде већи од стандардизоване вредности варијансе на нивоу појединачних индиктора (односно, s_{jj}^2=1) да би посматрани заједнички фактор био задржан у коначном моделу (*Brown*, 2015, стр. 23; *Barthlomew et al.*, 2008, стр. 124). Другим речима, да би „заслужио“ (или боље речено „задржао“) своје место у моделу, посматрани фактор треба да „објасни“ варијансу која одговара најмање, у просеку, једном индикатору (*Hair et al.*, 2010, стр. 108). Сходно наведеном, *Kaiser–Guttman*-ово правило може бити формулисано и на следећи начин (*Đorđević i drugi*, 2011, стр. 151; *Brown*, 2015, стр. 23; *Pituch & Stevens*, 2016, стр. 342): заједнички фактори који се одликују карактеристичном вредношћу (λ_f) већом од јединице требају бити задржани у коначном моделу ФА, и обратно. Детаљни и корисни сумарни прегледи резултата истраживања спроведених у циљу испитивања прецизности разматраног критеријума могу се видети у: *Pituch & Stevens* (2016, стр. 342) и *Kovačić* (1994, стр. 234).

- *Критеријум заснован на објашњеном уделу укупне узорачке варијансе*
  (енгл. *Proportion of total sample variance explained*)

За разлику од претходног критеријума, заснованог на разматрању појединачних доприноса идентификованих фактора у објашњавању укупне узорачке варијансе, избор адекватног броја заједничких фактора (*m*), према овом (другом) приступу, условљен је обезбеђивањем, или боље речено, достизањем одговарајућег *a priori*, по правилу, субјективно специфицираног нивоа кумулативне пропорције објашњеног дела укупне варијансе узорка (*Kovačić*, 1994, стр. 207; *Rencher*, 2002, стр. 427; *Đorđević i drugi*, 2011, стр. 152; *Pituch & Stevens*, 2016, стр. 343), дефинисаног изразом (2.1.16). Наиме, полазећи од израза (2.1.15), удео укупне варијансе узорка који је објашњен путем првих неколико (*m*) издвојених, од укупно идентификованих *p*, заједничких фактора, утврђује се путем следећег израза (*Rencher*, 2002, стр. 427; *Thompson*, 2004, стр. 21; *Barthlomew et al.*, 2008, стр. 124):

$$\sum_{f=1}^{m} \lambda_f / \sum_{f=1}^{p} \lambda_f = \sum_{f=1}^{m} \lambda_f / p \text{, где је } m < (\text{или} \ll) \, p. \qquad (2.1.16)$$

Након дефинисања циљне (или граничне) вредности датог израза, приступа се постепеном издвајању појединачних заједничких фактора, рангираних у опадајућем низу сходно вредностима припадајућих карактеристичних корена ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m \geq ... \geq \lambda_p$), све док се редукованим (коначним) моделом не обезбеди испуњење очекивања у погледу постигнуте

пропорције објашњеног дела укупне варијансе. Гранична вредност критеријума утврђује се произвољно, на различитим нивоима, у зависности од конкретног истраживачког подручја и проблема, због чега се овај критеријум одликује високим степеном арбитрарности (*Kovačić*, 1994, стр. 207). У прилог наведеном, у литератури се најчешће сугерише *a pripri* дефинисање граничне вредности на нивоу од 80% ± (5–10%) (*Barthlomew et al.*, 2008, стр. 124; *Pituch & Stevens*, 2016, стр. 343; *Kovačić*, 1994, стр. 207; *Rencher*, 2002, стр. 427), уз напомене појединих аутора да се и модели са објашњеним уделом укупне варијансе на нивоу од 70% (*Đorđević i drugi*, 2011, стр. 152), 60% (*Hair et al.*, 2010, стр. 108), 50%, па чак и мање (*Pituch & Stevens*, 2016, стр. 343) могу сматрати задовољавајућим и прихватљивим у појединим истраживањима, нарочито у домену друштвених наука.

- *Критеријум заснован на Cattell*-овом *дијаграму „превоја“* или *„прелома“* (енгл. *Cattell's scree*[27] *test* / *plot*)

Овај приступ избору адекватног броја фактора заснива се на формирању и интерпретирању графичког приказа вредности карактеристичних корена λ_f (за $f = 1, 2,..., p$) матрице **S** или **R** (лоцираних на вертикалној оси) према њиховом редном броју, односно броју фактора сходно редоследу њихове екстракције (*Kovačić*, 1994, стр. 207; *Barthlomew et al.*, 2008, стр. 124; *Đorđević i drugi*, 2011, стр. 152; *Pituch & Stevens*, 2016, стр. 342). Поменути графички приказ резултира кривом која се, по правилу, у почетку одликује стрмим опадајућим нагибом (илуструјући нагли пад вредности карактеристичних корена), након чега постепено поприма приближно (апроксимативно) хоризонтални облик (као последица уравнотежења последњих вредности λ_f) са знатно блажим нагибом. Тачка у којој долази до драстичне промене нагиба конструисане праве линије, односно након које опадање вредности карактеристичних корена постаје знатно спорије и мање изражено, назива се тачка прелома или „лакат“ (*Barthlomew et al.*, 2008, стр. 124; *Kovačić*, 1994, стр. 208). Идеја на којој се заснива овај графички приступ огледа се у идентификовању тачке прелома и примени правила које гласи: број издвојених заједничких фактора једнак је броју карактеристичних корена λ_f чије вредности претходе идентификованој тачки прелома и хоризонаталној линији, односно које су позициониране на „стрмом“ делу дијаграма (*Kovačić*, 1994, стр. 208; *Costello & Osborne*, 2005, стр. 3; *Rencher*, 2002, стр. 428; *Pituch & Stevens*, 2016, стр. 342). Логично, у ситуацијама када није могуће јасно уочити тачку прелома, односно када је на дијаграму присутно више од једног „лакта“, употребна вредност овог критеријума сматра се малом, будући да је одлука о броју фактора у значајној мери под утицајем субјективне интерпретације истраживача (*Brown*, 2015, стр. 24; *Kovačić*, 1994, стр. 208).

Генерално, важно је такође истаћи да се одлука о броју заједничких фактора који ће бити задржани у коначном моделу ретко доноси коришћењем само једног од расположивих критеријума. Уместо тога, углавном се практикује и препоручује употреба различитих приступа, иако ће (вероватно) исти резултирати, у већој или мањој мери, различитим

[27] Назив овог критеријума *Cattell* је формулисао на основу аналогије са значењем геолошког појма, који на енглеском гласи *scree*, а односи се на камените остатке (наслаге одломљених стена и камења) у подножју стеновите стрме литице, односно планине (*Cattell*, 1983, стр. 16). Наиме, стрме литице планине *Cattell* је поистоветио са сигнификантним заједничким факторима које треба задржати, а тривијалне факторе са наслагама шљунка у подножју литице (*Thompson*, 2004, стр. 32).

проценама (*Đorđević i drugi*, 2011, стр. 153; *Kovačić*, 1994, стр. 234; *Pituch & Stevens*, 2016, стр. 344). У контексту наведеног, *Rencher* (2002, стр. 429) истиче да уколико су подаци добро прилагођени моделу факторске анализе, претходно елаборирана три критеријума ће готово увек резултирати усаглашеним вредностима броја фактора, m. Међутим, уколико такав исход није остварен, коначна одлука о броју издвојених фактора условљена је проценом могућности смислене и практично корисне интерпретације изведеног модела (*Barthlomew et al.*, 2008, стр. 124; *Kovačić*, 1994, стр. 234). Другим речима, знање и искуство истраживача из конкретног подручја којем разматрани проблем припада игра важну улогу при доношењу одлуке о задржавању појединих фактора и њиховом броју (*Pituch & Stevens*, 2016, стр. 343). Користан сумарни приказ искуствених правила и препорука везаних за примену различитих приступа при избору броја заједничких фактора представљен је од стране *Pituch*-а *& Stevens*-а (2016, стр. 344) и *Kovačića* (1994, стр. 209).

Након примене критеријума за избор „оптималног“ броја заједничких фактора, уколико је редукованим факторским моделом обухваћено два или више фактора (односно, ако је $m \geq 2$) неопходно је извршити њихову ротацију путем неког од расположивих, бројних метода за спровођење исте. Наиме, оправдање за спровођење ротације добијеног факторског решења огледа се у чињеници да иницијалне вредности оцена факторских оптерећења (дефинисане изразом (2.1.9)) углавном не омогућавају извођење недвосмислених закључака везаних за идентификовање конкретних индикатора X_j који доминантно опредељују први, односно други, или m-ти заједнички фактор F_f (за $f = 1, 2,..., m$) (*Kovačić*, 1994, стр. 238; *Đorđević i drugi*, 2011, стр. 154). У том смислу, поступком ротације фактора врши се поједностављење иницијалне, редуковане (са $(p \times p)$ на $(p \times m)$), матрице факторских оптерећења b_{jf} (за $j = 1, 2,..., p$ и $f = 1, 2,..., m$) у циљу добијања нове (ротиране) матрице факторских оптерећења којом се омогућава боље разумевање природе и структуре издвојених фактора, олакшавајући у значајној мери поступак интерпретације добијеног факторског решења (*Đorđević i drugi*, 2011, стр. 157; *Kovačić*, 1994, стр. 239; *Barthlomew et al.*, 2008, стр. 189). Логично, у случају када је добијено решење факторске анализе једнофакторског типа (то јест, када је $m = 1$), не постоји потреба за спровођењем метода ротације (*Brown*, 2015, стр. 27). Сходно наведеном, у ситуацијама када се факторска анализа користи за потребе разумевања присутних релација између коришћених индикатор-променљивих и обезбеђивања одговарајуће методолошке основе за креирање специфичног композитног показатеља намењеног мерењу једног конретног неопсервабилног феномена од интереса (кореспондентног издвојеном заједничком фактору), као што је то случај са емпиријским истраживањем спроведеним у оквиру ове дисертације, поменути једнофакторски модел ФА представља пожељно решење у примени ФА. Наиме, у наведеном контексту, издвајање већег броја фактора заправо представља индиректни сигнал недовољно прецизног одабира индикатора на којима се заснива конструкција осмишљеног композитног показатеља. Консеквентно, детаљније разматрање методолошких одређења поступка ротације факторског решења није обухваћено излагањем у наставку текста.[28]

[28] Користан и детаљан преглед у литератури предложених метода ортогоналне (правоугаоне) и/или неортогоналне (косе) ротације модела факторске анализе, представљен је од стране, на пример, следећих аутора: *Kovačić* (1994, стр. 237–246), *Rencher*, (2002, стр. 430–437), *Johnson & Wichern* (2007, стр. 504–513), *Barthlomew et al.* (2008, стр. 188–192), *Hair et al.* (2010, стр. 112–115), *Đorđević i drugi* (2011, стр. 154–158), *Brown* (2015, стр. 27–31), *Pituch & Stevens* (2016, стр. 344-345).

Коначно, за потребе интерпретације изведеног факторског модела, неопходно је размотрити оцењене вредности факторских оптерећења и кореспондентних комуналитета из угла њихове, како практичне, тако и статистичке значајности. У општем случају, полазећи од присутних релација између наведених елемената модела ФА (представљених изразом (2.1.12)), што је већа (апсолутна) вредност конкретног факторског оптерећења, b_{jf}, иста се сматра важнијом из угла интерпретације, будући да имплицира присуство сразмерно већег удела објашњеног дела укупне варијансе посматраног индикатора по основу датог фактора. Прецизније, у одсуству формалних статистичких поступака, у литератури су предложена следећа, широко коришћена, искуствена правила и критеријуми за утврђивање практичне сигнификантности факторских оптерећења, која гласе (*Hair et al.*, 2010, стр. 116; *Đorđević i drugi*, 2011, стр. 159):[29] (1) вредности факторских оптерећења у интервалу $0{,}30 < |b_{jf}| \leq 0{,}40$ задовољавају минимално прихватљив ниво практичне значајности; (2) факторска оптерећења чије се вредности крећу у опсегу $0{,}40 < |b_{jf}| \leq 0{,}50$ сматрају се важним за интерпретацију; (3) вредности $b_{jf} > \pm 0{,}50$ сматрају се практично сигнификатним; (4) факторска оптерећења изнад ± 0,70 сматрају се показатељима добро дефинисане факторске структуре и представљају жељени циљ сваког истраживача који примењује ФА. С друге стране, користан сумарни преглед практичних смерница предложених у контексту извођења закључака о статистичкој сигнификантности разматраних факторских оптерећења, у односу на величину анализираног узорка, може се видети код *Hair*-а *et al.* (2010, стр. 118). Посматрано из угла висине оцењених вредности појединачних комуналитета, као минимално прихватљив, углавном се у литератури наводи ниво од 0,50, односно 50% објашњеног укупног варијабилитета (*Hair et al.*, 2010, стр. 118; *Thompson*, 2004, стр. 20) на нивоу појединачних индикатора.

2.1.5. Употреба модела ФА у функцији развоја композитног индикатора

Генерално, оцењеним моделом ФА, као што је раније апострофирано, свака, појединачно посматрана, индикатор-променљива (X_j) представљена је као линеарна (адитивна) функција производа издвојених заједничких фактора (F_f) и кореспондентних оцењених вредности факторских оптерећења b_{jf}, с једне стране, и оцене специфичног фактора e_j, с друге стране, односно симболички: $X_j = (b_{j1}F_1 + b_{j2}F_2 + ... + b_{jm}F_m) + e_j$, за $j = 1, 2, ..., p$ и $f = 1,2,..., m$. Уколико је применом, у претходном одељку елаборираних, критеријума, као доминантно, издвојено присуство само једног заједничког фактора, модел ФА добија једноставнију форму, која гласи: $X_j = b_{j1}F_1 + e_j$, за $j = 1, 2, ..., p$.

Посматрано из угла представљеног једнофакторског модела, издвојени фактор је заправо неопсервабилна променљива (чије мерење није могуће извршити директно) и доминантни узрочник квантитативног слагања између анализираних индикатора. У том контексту, појединачни индикатори, односно анализиране опсервабилне (мерљиве) нумеричке променљиве X_j, могу се третирати као „помоћна средства“ неопходна за индиректно „мерење вредности“ фактора, као сложеног, мултидимензионог феномена од интереса за истраживача. Другим речима, полазећи од оцењеног једнофакторског модела ФА,

[29] Наведене искуствене препоруке су примењиве у условима када је анализираним узорком обухваћено 100 и више јединица посматрања и у ситуацији када је нагласак на практичној, а не статистичкој значајности (*Hair et al.*, 2010, стр. 116)

издвојени заједнички фактор могуће је исказати путем линеарне комбинације разматраних p индикатор-променљивих (X_j) и њима припадајућих тежинских коефицијената (у ознаци, w_j), односно у виду наменски креираног композитног показатеља (или индекса), чија општа математичка форма има следећи изглед (*Stamenković & Savić*, 2017, стр. 108; *Иглић*, 2014, стр. 104; 173):

$$f_{(i)} = \sum_{j=1}^{p} w_j x'_{ij} = w_1 x'_{11} + w_2 x'_{12} + \ldots + w_p x'_{1p}, \text{ за } j = 1, 2, \ldots, p \text{ и } i = 1, 2, \ldots, n, \qquad (2.1.17)$$

где коришћени симболи означавају:

x_{ij}' – нормализована вредност j-те променљиве на нивоу i-те јединице посматрања;

$f_{(i)}$ – „оцењена вредност“ издвојеног заједничког фактора на нивоу i-те опсервације;

w_j – релативна тежина (пондер) j-те променљиве у контексту заједничког фактора.

Прецизније, представљеним композитним показатељем врши се агрегирање пондерисаних вредности појединачних индикатора који репрезентују различите аспекте латентне циљне димензије, а у циљу утврђивања „оцењених вредности“ издвојеног заједничког фактора за сваку i-ту опсервацију, чиме се обезбеђује сумаризација и свођење вишедимензионог оригиналног истраживачког проблема на специфичну и, за разумевање и интерпретацију знатно једноставнију, униваријациону форму, која, по правилу, „вреди“ (нешто) више од простог збира саставних делова.[30] Одређивање тежинских коефицијената, w_j, додељених појединачним индикаторима у изразу (2.1.17) заснива се на анализи структуре удела укупне узорачке варијансе који је објашњен издвојеним (једно)факторским решењем.
Наиме, полазећи од релација дефинисаних изразима (2.1.14) и (2.1.15), у случају када је моделом ФА обухваћен само један заједнички фактор (односно, $m = 1$), релативни значај појединачних индикатора еквивалентан је релативном учешћу појединачних вредности оцена комуналитета (у ознаци $\hat{h}_j^2$) у укупном објашњеном уделу узорачке варијансе, односно вредности карактеристичне вредности (λ_1) која одговара издвојеном фактору F_1, или симболички (*Stamenković & Savić*, 2017, стр. 108; *Nardo et al.*, 2005, стр. 57; *Fernando et al.*, 2012, стр. 330):

$$w_j = \frac{\hat{h}_j^2}{\sum_{j=1}^{p} \hat{h}_j^2} = \frac{\hat{h}_j^2}{\lambda_1} = \frac{(b_j)^2}{\lambda_1}, \qquad (2.1.18)$$

при чему је (*Hudrlikova*, 2013, стр. 464): $0 \leq w_j \leq 1$ и $\Sigma w_j = 1$, за $j = 1, 2, \ldots, p$. Утврђени пондери заправо представљају релативно учешће квадрираних вредности оцена факторских оптерећења (у ознаци b_j) појединачних (нормализованих) индикатора X_j, у делу укупне узорачке варијансе који је објашњен издвојеним (једним) фактором у оквиру модела ФА. Сходно евидентној повезаности са оцењеним вредностима параметара

[30] У методолошком смислу, конструкција композитног показатеља представља комплексан процес сачињен од следећих (кључних) корака (*Nardo et al.*, 2005, стр. 9-14; *OECD*, 2008, стр. 20-21): (1) Одабир улазних индикатор-променљивих; (2) Решавање питања недостајућих вредности; (3) Нормализација оригиналних индикатор-променљивих; (4) Дефинисање модела пондерисања; (5) Избор процедуре агрегирања; (6) Анализа осетљивости и робустности креираног композитног индекса; (7) Презентација и интерпретација вредности развијеног композитног показатеља. Детаљан преглед и објашњење концептуално-методолошких одређења појединачних корака укључених у процес развоја композитног показатеља као и дискусија везана за уобичајене дилеме са којима се истраживач сусреће у датим настојањима, изложени су од стране следећих аутора: *Salzman* (2003), *Nardo et al.* (2005), *OECD* (2008), *Foa & Tanner* (2012), *Hudrlikova* (2013).

изведеног модела ФА, за представљени поступак, који се користи у одређивању пондера додељених појединачним индикаторима у структури креираног композитног показатеља, каже се да је заснован на статистичком моделу (*Nardo et al.*, 2005, стр. 12). Израчунате (појединачне) вредности формираног композитног показатеља, f_i, представљају погодну основу за рангирање и иницијалну класификацију разматраних јединица посматрања, у контексту конкретног (неопсервабилног) мултидимензионог феномена од интереса за истраживача. Формирана нова (композитна) променљива може бити употребљена као улазна компонента у поступку примене неких других метода мултиваријационе статистичке анализе, усмерених ка евалуацији квалитета формираног композитног индикатора и оцену извршене иницијалне класификације јединице посматрања.

2.2. АНАЛИЗА ГРУПИСАЊА

Полазећи од чињенице да класификација и груписање објеката или феномена од интереса, на бази њихове међусобне сличности, представља једну од основних, исконских концептуалних активности на којој почива функционисање људског ума (*Vercellis*, 2009, стр. 293; *Everitt et al.*, 2010, стр. 239), не изненађује широка популарност и разноврсност употребе анализе груписања (енгл. *cluster analysis*) при реализацији истраживања у домену бројних (готово свих) научних области и дисциплина данас.[31] Иако први, у научном контексту неформални, облици примене потичу из „далеке“ прошлости,[32] формална афирмација анализе груписања у научним оквирима уследила је, како *Bailey* (1975, стр. 59) наводи, током 30-их година XX века (примарно) захваљујући истраживачком раду *Driver*-а & *Kroeber*-а (1932) у области антропологије, односно *Zubin*-а (1938) и *Tryon*-а (1939)[33] у оквиру психологије. Изразита рачунска комплексност успорила је даљи развој и ограничила ширу примену метода анализе груписања све до касних 1950-их, односно до појаве првих рачунара (*Bailey*, 1975, стр. 59), који су процес груписања учинили рационалним (са аспекта „утрошеног“ времена), али и рачунски изводљивим, узимајући у обзир обимност и захтевност неопходних израчунавања[34] (*Aldenderfer & Blashfield*, 1984, стр. 8). Током 1960-их и 1970-их година XX века, анализа груписања постаје „главна тема“ у академским круговима и научно-истраживачкој заједници на глобалном нивоу, а један од највећих подстицаја рапидног развоја ове групе метода (у теоријском и апликативном смислу), поред убрзане експанзије и унапређења компјутерске технологије, како истичу *Aldenderfer & Blashfield* (1984, стр. 7) и *Bock* (2008, стр. 1), представљала је монографија под насловом „*Принципи нумеричке таксономије*“ (енгл.

[31] У релевантној литератури, у зависности од конкретног подручја примене, употребљавају се различити термини за описивање нумеричких метода који се могу сврстати под окриљем анализе груписања. Генерално, термин *кластер анализа* (енгл. *cluster analysis*) представља назив који се најчешће користи за означавање ове групе метода. Међутим, у литератури из биологије и зоологије веома често се употребљава термин *нумеричка таксономија* (енгл. *numerical taxonomy*) (*Tuffery*, 2011, стр. 236). У медицини и психологији углавном се користи термин *носологија* (енгл. *nosology*) (*Tuffery*, 2011, стр. 236), односно *Q-анализа* (енгл. *Q-analysis*) (*Everitt et al.*, 2011, стр. 5) респективно, док се у маркетинг истраживањима говори о *сегментацији* или *типологији* (енгл. *segmentation* / *typology*) (*Halkidi et al.*, 2001, стр. 107; *Tuffery*, 2011, стр. 236; *Everitt et al.*, 2011, стр. 5). Истраживачи у домену *data mining*-а, вештачке интелигенције и машинског учења пак користе назив *ненадгледано препознавање образаца* (енгл. *unsupervised pattern recognition*) (*Tuffery*, 2011, стр. 236). Без обзира на одомаћеност „кластер анализе“ у домаћој литератури, будући да исти, према мишљењу *Kovačić*-а (1994, стр. 255) није у духу српског језика, у наставку излагања коришћен је лингвистички примеренији назив, *анализа груписања*.

[32] Ослањајући се на сопствено и мишљење бројних других истраживача, *Andritsos* (2002, стр. 1) следећи пример истиче као први познати (документовани) случај (релативно једноставне, али практичне) примене аналитичког поступка који се може сматрати претечом анализе груписања: „Током епидемије колере у Лондону 1854. године, *John Snow* је користио посебну мапу за евидентирање локација на којима су евидентирани случајеви обољевања од ове опаке болести. Анализирањем креиране мапе, запажена је блиска повезаност концентрације забележених случајева болести и једног бунара који се налазио у централној улици Лондона. Уклањање пумпе за воду са „сумњивог“ бунара довело је до окончања епидемије.“

[33] Прва званична употреба термина *cluster analysis* у писаној форми, како наводи *Murtagh* (2016, стр. 21), везује се управо за рад психолога *Robert*-а *Tryon*-а и његову публикацију под насловом „*Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*“ из 1939. године, у којој је изложио детаљно објашњење комплексних квантитативних процедура за организовање (груписање) објеката на бази уочене сличности између њих.

[34] О размерама рачунске комплексности и интензивности процеса груписања без „помоћи“ рачунара, најбоље сведочи следећи пример, формулисан од стране *Aldenderfer*-а *& Blashfield*-а (1984, стр. 8): Полазни корак у поступку груписања 200 јединица посматрања подразумева формирање адекватне матрице сличности сачињене од 19.900 јединствених вредности, као основе за даљи ток анализе и примену једне од расположивих метода груписања. Ручно претраживање и рад са матрицама такве величине, према мишљењу истих аутора, представља мукотрпан и дуготрајан подухват у чију реализацију је само неколицина истраживача (или пак, њихових „несрећних“ асистената) спремно (и вољно) да се упусти.

Principles of numerical taxonomy), аутора (двојице биолога) *R. Sokal*-а & *P. Sneath*-а, која је публикована 1963. године. Од тог периода, заинтересованост истраживача (из бројних и различитих научних области) за испитивањем апликативних потенцијала, али и могућности за теоријско-методолошка унапређења и даљи развој анализе груписања, бележи стални раст.

2.2.1. Циљеви и поступак спровођења анализе груписања

Анализа груписања представља генерички термин који се користи за описивање широког спектра (фамилије) мултиваријационих статистичких процедура посебно креираних за откривање „природне“ структуре, односно класификације разматраних јединица посматрања унутар комплексних и (углавном) хетерогених скупова / узорака података (*Gore Jr.*, 2000, стр. 298). Заснован на уважавању суштинских одређења концепта блискости појединачних опсервација, заједнички примарни циљ ових процедура је подела конкретног скупа / узорка неструктурираних мултиваријационих опсервација на одређени (по правилу „мали“) број, међусобно искључивих и, у што је могуће већој мери, интерно-хомогених и екстерно-хетерогених, смислених група јединица посматрања (енгл. *clusters*). Наиме, исход примене било које процедуре груписања јесте одговарајућа класификациона структура анализираних опсервација, сачињена од одређеног броја група, при чему се за опсервације у саставу исте групе претпоставља да поседују „велику“ међусобну сличност, али уједно и „малу“ сличност са објектима изван те конкретне групе, у контексту вредности симултано разматраних свих променљивих обухваћених анализом (*Bijnen*, 1973, стр. 2). Такође, полазећи од констатације *Kaufman*-а *& Rousseeuw*-а (2005, стр. 1) да анализа груписања представља уметност проналажења (откривања) група у расположивим подацима, практични значај реализације дефинисаног циља анализе груписања заснива се на претпоставци да је конкретна структура „природних“ група[35] заиста и инхерентна, то јест присутна у посматраном скупу / узорку мултиваријационих опсервација (*Varmuza & Filzmoser*, 2009, стр. 265).[36] С обзиром на чињеницу да егзактне информације како у погледу броја тако и структуре група нису *a priori* расположиве, анализа груписања припада групи метода ненадгледаног учења, за разлику од класичних метода класификације, усмерених на извођење правила за алокацију опсервација у познати број претходно специфицираних група, или класа (*Rencher*, 2002, стр. 451; *Johnson & Wichern*, 2007, стр. 671; *Martinez & Martinez*, 2007, стр. 431; *Everitt et al.*, 2011, стр. 7). Такође, анализа груписања представља непараметарски мултиваријациони метод, будући да валидна апликација статистичких процедура груписања није условљена ригидним претпоставкама у погледу дистрибуције података (*Sharma*, 1996, стр. 197).

Полазну основу у анализи груписања представљају подаци уређени у форми ($n \times p$) матрице $\mathbf{X} = [x_{ij}]$, сачињене од n редова, који означавају јединице посматрања (објекте, или опсервације), и p колона, које репрезентују анализом обухваћене променљиве, или

[35] Будући да ће операционализација појединачних процедура груписања увек резултирати генерисањем специфичног броја група јединица посматрања без обзира да ли или не групе заиста и постоје у анализираном скупу / узорку, примена анализе груписања у условима изразите хомогености расположивих мултиваријационих опсервација може се схватити пре као „наметање непостојеће структуре груписања“ (енгл. *structure-imposing*) него „проналажење скривене структуре груписања у подацима“ (енгл. *structure-seeking*) (*Aldenderfer & Blashfield*, 1984, стр. 16).

[36] Употреба термина „природне групе“ подразумева схватање и третирање издвојених група као појединачних области у p-димензионом простору које се одликују великом густином тачака (као графичког приказа ($p \times 1$) вектора опсервација), а које су уједно јасно и очигледно раздвојене од других области (*Kovačić*, 1994, стр. 256).

карактеристике (Табела 2.2.1.). Наиме, елементи у j-тој колони представљају вредности промењиве X_j (за j=1, 2,…, p) измерене (или утврђене) за сваку од n појединачних јединица посматрања, док елементи i-тог реда кореспондирају вредностима сваке од p промењивих (у ознаци: x_{i1}, x_{i2},..., x_{ip}) утврђеним на нивоу i-тог објекта (за i=1,2,...,n). Мултиваријациони профил i-те опсервације, у контексту анализираних промењивих, означен је симболом $\mathbf{x}_i' = [x_{i1}, x_{i2},..., x_{ip}]$.

Табела 2.2.1. *Матрица мултиваријационих података у анализи груписања*

Јединице посматрања	Промењиве (X_j)				Векторска знака опсервација
	X_1	X_2	…	X_p	
1	x_{11}	x_{12}	…	x_{1p}	$\mathbf{x}_1' = [x_{11}, x_{12},\ldots, x_{1p}]$
2	x_{21}	x_{22}	…	x_{2p}	$\mathbf{x}_2' = [x_{21}, x_{22},\ldots, x_{2p}]$
⋮	⋮	⋮	…	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{np}	$\mathbf{x}_n' = [x_{n1}, x_{n2},\ldots, x_{np}]$

Извор: Ауторов приказ

Посматрано у контексту матрице оригиналних података $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$, циљ анализе груписања заправо представља проналажење класификационе структуре којом се остварује распоређивање n редова матрице $\mathbf{X}$ (то јест, n појединачних (p×1) вектора опсервација, $\mathbf{x}_i$) у g међусобно искључивих (дистинктивних) група, у ознаци C_k (за k = 1, 2,…, g), при чему је, по конвенцији, $g \ll n$, а $\forall C_k \subset \mathbf{X}$.

У начелу, као аналитички процес вођен подацима, анализа груписања представља корисно статистичко средство (алат) експлоративне, али и конфирматорне анализе података (*Gore, Jr.*, 2000, стр. 300). *Aldenderfer & Blashfield* (1984, стр. 9) објашњавају претходну констатацију, наводећи да се анализа груписања може користити не само за потребе откривања непознате структуре објеката унутар разматраног скупа / узорка, већ и за проверу одрживости постојећих претпоставки формулисаних у погледу присуства одређеног броја и структуре група у конкретном скупу / узорку података, а на основу резултата (претходно извршене) примене неких других метода анализе. Категоризација метода, предложених у литератури за реализацију експлоративних и конфирматорних циљева анализе груписања, може се извршити на следеће две, широко дефинисане, групе, односно фамилије метода, и то (*Murtagh & Heck*, 1987, стр. 56; *Rencher*, 2002, стр. 452; *Timm*, 2002, стр. 522; *Hardle & Simar*, 2003, стр. 308; *Shmueli et al.*, 2005, стр. 208; *Bartholomew et al.*, 2008, стр. 19): ▸ методи хијерархијског груписања (енгл. *hierarchical clustering methods*) и ▸ методи нехијерархијског груписања (енгл. *non-hierarchical / partitional clustering methods*). Генерално, избор и примена, како адекватне мере за квантитативно оцењивање степена блискости појединачних опсервација, тако и конкретне процедуре груписања, по правилу се истичу као фундаментални кораци у процесу идентификовања „природне“ структуре објеката, заснованом на симултаном разматрању њихових мултиваријационих профила. Дијаграм тока поступка спровођења анализе груписања представљен је на Слици 2.2.1.

Сходно методолошком приступу који је коришћен за потребе реализације дела емпиријског истраживања у дисертацији, у наставку су елаборирана кључна концептуално–методолошка одређења појединачних корака у спровођењу анализе груписања, са посебним осврtом на хијерархијске процедуре креирања модела груписања.

Слика 2.2.1. Шематски приказ поступка спровођења (хијерархијске) анализе груписања

Извор: Ауторов приказ

2.2.2. Кључна одређења концепта блискости између јединица посматрања

Кључни (централни) корак у реализацији дефинисаних циљева анализе груписања обухвата дефинисање и/или избор објективних статистичких критеријума, односно одговарајућих мера за квантитативно оцењивање степена блискости између свих парова расположивих јединица посматрања у контексту анализираних карактеристика. У домену анализе груписања, а у зависности од коришћеног критеријума за „мерење“, блискост (енг. *proximity*), као општи термин, може се тумачити из угла различитости (енг. *dissimilarity*) или пак сличности (енг. *similarity*) утврђене између јединица посматрања чије се груписање спроводи (*Kaufman & Rousseeuw*, 2005, стр. 4; *Everitt, et al.*, 2011, стр. 43; *Timm*, 2002, стр. 515; 516). У том смислу, коришћена мера блискости (енг. *proximity measure*) сматра се мером сличности (енг. *similarity measure*) уколико се њоме квантификује степен међусобне сличности, односно „близина“ посматране две опсервације, са аспекта вредности променљивих које их карактеришу, уз нотацију да веће вредности мере сличности указују да посматране опсервације поседују у већој мери сличне карактеристике, односно заједничка својства и обратно. *Kovačić* (1994, стр. 259), *Timm* (2002, стр. 519), *Kaufman & Rousseeuw* (2005, стр. 20) и *Nedović* (2016, стр. 9) наводе следећа математичка својства која углавном поседују мере сличности, у ознаци $sim(\mathbf{x}_q, \mathbf{x}_h)$ или $s_{(q,h)}$, дефинисане између било које две i-те опсервације, конкретно $\mathbf{x}_q$ и $\mathbf{x}_h$ (при чему $\mathbf{x}_q$, $\mathbf{x}_h \in \mathbf{X}$), која гласе:

1. Нормираност и ненегативност: $\forall \mathbf{x}_q, \mathbf{x}_h \in \mathbf{X}, \;\; 0 \le sim(\mathbf{x}_q, \mathbf{x}_h) \le 1$;
2. Симетричност (комутативност): $\forall \mathbf{x}_q, \mathbf{x}_h \in \mathbf{X}, \;\; sim(\mathbf{x}_q, \mathbf{x}_h) = sim(\mathbf{x}_h, \mathbf{x}_q)$;
3. Рефлексивност: $\forall \mathbf{x}_q \in \mathbf{X}, \;\; sim(\mathbf{x}_q, \mathbf{x}_q) = 1$;

4. Неразликовање идентичних: $\forall \mathbf{x}_q, \mathbf{x}_h \in \mathbf{X},\ sim(\mathbf{x}_q, \mathbf{x}_h) = sim(\mathbf{x}_q, \mathbf{x}_q) = 1 \Leftrightarrow \mathbf{x}_q = \mathbf{x}_h$;
5. Разликовање неидентичних: $\forall \mathbf{x}_q, \mathbf{x}_h \in \mathbf{X},\ [sim(\mathbf{x}_q, \mathbf{x}_h) \neq 1] < [sim(\mathbf{x}_q, \mathbf{x}_q) = 1] \Leftrightarrow \mathbf{x}_q \neq \mathbf{x}_h$.

С друге стране, уколико квантификовање степена блискости између посматраних опсервација подразумева израчунавање разлике односно „удаљености“ између њих у контексту вредности анализираних променљивих, тада се коришћена мера блискости сматра мером различитости (енг. *dissimilarity measure*) (*Kovačić*, 1994, стр. 259). Насупрот концепту сличности и кореспондентним мерама, већа „удаљеност“ јединица посматрања, манифестована високим вредностима мере различитости, имплицира присуство мањег степена блискости, док мале (позитивне) вредности и вредности једнаке нули указују на постојање изражене блискости, односно идентичност посматраних опсервација, респективно (*Aldenderfer & Blashfield*, 1984, стр. 25; *Murtagh & Heck*, 1987, стр. 3). *Izenman* (2008, стр. 412), *Timm* (2002, стр. 516) и *Nedović* (2016, стр. 1–4) као кључна математичка својства која поседује свака мера различитости, у ознаци $dis(\mathbf{x}_q, \mathbf{x}_h)$ или $d_{(q,h)}$, дефинисана за произвољно одабрану q-ту и h-ту опсервацију (при чему: $q, h \in i = 1, 2, \ldots, n$) издвајају следећа:

1. Ненегативност: $\forall \mathbf{x}_q, \mathbf{x}_h \in \mathbf{X},\ dis(\mathbf{x}_q, \mathbf{x}_h) \geq 0$;
2. Симетричност (комутативност): $\forall \mathbf{x}_q, \mathbf{x}_h \in \mathbf{X},\ dis(\mathbf{x}_q, \mathbf{x}_h) = dis(\mathbf{x}_h, \mathbf{x}_q)$;
3. Рефлексивност: $\forall \mathbf{x}_q \in \mathbf{X},\ dis(\mathbf{x}_q, \mathbf{x}_q) = 0$;
4. Неразликовање идентичних: $\forall \mathbf{x}_q, \mathbf{x}_h \in \mathbf{X},\ dis(\mathbf{x}_q, \mathbf{x}_h) = dis(\mathbf{x}_q, \mathbf{x}_q) = 0 \Leftrightarrow \mathbf{x}_q = \mathbf{x}_h$;
5. Разликовање неидентичних: $\forall \mathbf{x}_q, \mathbf{x}_h \in \mathbf{X},\ [dis(\mathbf{x}_q, \mathbf{x}_h) > 0] > [dis(\mathbf{x}_q, \mathbf{x}_q) = 0] \Leftrightarrow \mathbf{x}_q \neq \mathbf{x}_h$

Мере различитости које задовољавају наведене аксиоме називају се полу-метрике (енг. *semi-metrics*) (*Timm*, 2002, стр. 517; *Dinčić*, 2016, стр. 35). У контексту ефикасне имплементације анализе груписања, неопходно је указати и на посебну групу мера, као дела ширег концепта различитости, сачињену од такозваних мера одстојања односно удаљености, или једноставно метрика (енг. *metrics*) (*Dalbelo-Bašić*, 2011, стр. 397). За разлику од мера различитости у општем смислу које се дефинишу као полу-метрике, мере одстојања (енг. *distance measures*) поред наведених услова задовољавају и услов неједнакости троугла (енг. *triangle inequality*) према којем удаљеност између било ког пара опсервација ($\mathbf{x}_q$, $\mathbf{x}_h$) у p-димензионом простору не може бити већа од збира одстојања забележених између датих опсервација и било које произвољне, додатне опсервације $\mathbf{x}_l$, односно: $\forall\ \mathbf{x}_q, \mathbf{x}_h, \mathbf{x}_l \in \mathbf{X}$, $d(\mathbf{x}_q, \mathbf{x}_h) \leq d(\mathbf{x}_q, \mathbf{x}_l) + d(\mathbf{x}_l, \mathbf{x}_h)$. Констатацијом да квантитативном оценом блискости доминира концепт метрика, *Aldenderfer & Blashfield* (1984, стр. 18) истичу значај и наглашавају улогу мера одстојања у анализи груписања. Наиме, у случају када су све променљиве, обухваћене анализом груписања, непрекидне нумеричке природе, по правилу, блискост између разматраних опсервација се углавном квантификује применом одговарајућих мера одстојања (*Everitt, et al.*, 2011, стр. 49), будући да концепт метрике омогућава репрезентовање опсервација у форми тачака у p-димензионом систему (*Sharma*, 1996, стр. 42; *Aldenderfer & Blashfield*, 1984, стр. 18) и, сходно томе, прецизно и објективно мерење блискости (различитости) која кореспондира њиховој међусобној удаљености у координатном простору чија је димензионалност одређена бројем улазних променљивих. У том смислу, *Johnson & Wichern* (2007, стр. 674) препоручују, увек када је

то природом података омогућено, коришћење мера различитости, које задовољавају услове својствене мерама одстојања, за потребе груписања јединица посматрања. Међутим, без обзира на њихов неоспорни и очигледни значај, *Aldenderfer & Blashfield* (1984, стр. 19) напомињу да мере одстојања ни на који начин не представљају једино средство за репрезентовање блискости између опсервација, нарочито у случају када анализиране променљиве нису нумеричког типа, иако бројни математичари аргументовано оспоравају рутинску употребу мера сличности / различитости које не задовољавају наведена својства за квалификовање одређене мере као метрике. Став поменутих аутора, подржавају и *Berry & Linoff* (2004, стр. 361) и *Johnson & Wichern* (2007, стр. 674) образлажући наведено флексибилном природом већине процедура за груписање, која омогућава и дозвољава њихову успешну имплементацију и у условима када су поједини од наведених (формално дефинисаних) аксиома метрике (на пример, неједнакост троугла) релаксирани (ублажени). У том контексту, *Timm* (2002, стр. 517) напомиње да услов неједнакости троугла, као неопходан за дефинисање метрике и геометријску интерпретацију различитости између опсервација, представља довољан али не и неопходан предуслов за дефинисање и примену мере блискости, односно различитости. Наведену тврдњу употпуњују *Kaufman & Rousseeuw* (2005, стр. 16) и *Berry & Linoff* (2004, стр. 361) који истичу симетричност и рефлексивност као најважније услове за које се од мера различитости, коришћених у контексту анализе груписања, очекује и претпоставља да задовољавају, иако постоје и методи груписања који не захтевају испуњеност нити једне од њих (*Kaufman & Rousseeuw*, 2005, стр. 16).

Сублимирањем изнетих разматрања, евидентно је да се блискост између две опсервације може тумачити, у зависности од примењене мере за квантификовање исте, као сличност или различитост, односно одстојање, које означава нешто рестриктивнију различитост (*Murtagh & Heck*, 1987, стр. 3). Будући да и одстојање и различитост „мере“ идентичне опсервације нулом,[37] односно великим (позитивним) вредностима уколико се блискост између опсервација смањује, насупрот мерама сличности које високим вредностима унутар детерминисаног (пожељно и нормираног) опсега њихових вредности сугеришу присуство изражене блискости, оправдано је тврдити да су концепт сличности, на једној страни, и концепти различитости и одстојања, на другој страни, у међусобно инверзној, али комплементарној релацији (*Varmuza & Filzmoser*, 2009, стр. 58; *Bartholomew, et al.*, 2008, стр. 19). Блискост посматране две опсервације се сматра израженом или високом уколико је одстојање, односно различитост, између њих мало(а), односно сличност велика, док већа удаљеност имплицира да су посматране јединице посматрања мање сличне и обратно (*Hardle & Simar*, 2003, стр. 303; *Varmuza & Filzmoser*, 2009, стр. 58; *Everitt et al.*, 2011, стр. 43).

Систематско презентовање, на бази елемената оригиналне ($n \times p$) матрице мултиваријационих опсервација (Табела 2.2.1), утврђених вредности изабране, конкретне од многих, мере блискости између свих парова n јединица посматрања врши се њиховим организовањем у форми симетричне квадратне ($n \times n$), у општем смислу такозване, матрице блискости (енг. *proximity matrix*), односно матрице сличности (енг. *similarity matrix*), у ознаци $\mathbf{S}=[s_{(q,h)}]$ (за $q, h = 1,2,\ldots, n$), или матрице различитости / одстојања (енг.

[37] Јединице посматрања се сматрају идентичним уколико се обе описане, односно представљене карактеристикама које се одликују истим вредностима или модалитетима (*Aldenderfer & Blashfield*, 1984, стр. 25).

dissimilarity / *distance matrix*), у ознаци $\mathbf{D}=[d_{(q,h)}]$, (за q, h=1,2,…,n), у зависности од тога да ли припадајући елементи матрице означавају вредности мере сличности или мере различитости, односно мере одстојања, респективно (израз (2.2.1) [38]).

$$\mathbf{S}=\left[s_{(q,h)}\right]=\begin{bmatrix} s_{11}=1 & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22}=1 & \dots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{21} & \dots & s_{nn}=1 \end{bmatrix},\quad \mathbf{D}=\left[d_{(q,h)}\right]=\begin{bmatrix} d_{11}=0 & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22}=0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{21} & \dots & d_{nn}=0 \end{bmatrix}. \quad (2.2.1)$$

Формиране матрице блискости представљају полазну основу у спровођењу бројних процедура (алгоритама) груписања, превасходно у оквиру хијерархијских метода груписања, будући да њихови елементи представљају квантитативну меру сличности или различитости, односно одстојања између посматраних парова опсервација (*Everitt*, 2010, стр. 242). Важно је такође истаћи да се концепти различитости и одстојања заједно са кореспондентним мерама и матрицама, генерално, сматрају погоднијом алтернативом при спровођењу анализе груписања, у односу на употребу матрице сличности, будући да су најчешће коришћене процедуре груписања углавном дизајниране за анализирање различитости и рад са израчунатим одстојањима између јединица посматрања (*Dunham*, 2000, стр. 99; *Bartholomew et al.*, 2008, стр. 38; *Vercellis*, 2009, стр. 294). Сходно наведеном, у околностима када је квантификовање степена блискости између расположивих опсервација извршено коришћењем одређене мере сличности, пре имплементације адекватних алгоритама груписања, углавном се практикује и препоручује конверзија иницијално формиране матрице сличности у одговарајућу матрицу различитости (*Kovačić*, 1994, стр. 259; *Kaufman & Rousseeuw*, 2005, стр. 21). Основа за спровођење наведене трансформације садржана је у претходно елаборираној констатацији да је сличност, у интуитивном смислу, појам комплементаран појму различитости. Као једну од најчешће коришћених формула за трансформацију мере сличности, $s_{(q,h)}$, односно матрице $\mathbf{S}=[s_{(q,h)}]$ у меру различитости, $d_{(q,h)}$, односно матрицу $\mathbf{D}=[d_{(q,h)}]$, *Timm* (2002, стр. 519), *Kaufman & Rousseeuw* (2005, стр. 21) и *Everitt et al.* (2011, стр. 46) наводе[39]: $[d_{(q,h)}]=1-[s_{(q,h)}]$. Својства новоформиране мере различитости не морају нужно одговарати својствима метрике, односно мере одстојања (*Timm*, 2002, стр. 519), будући да су иста условљена одликама (претходно) коришћене мере сличности. Полазећи од изнетих разматрања, а имајући у виду и констатацију, на коју *Sharma* (1996, стр. xxx), *Everitt et al.* (2011, стр. 69) и *Manly & Navarro Alberto* (2017, стр. 97) указују, да примена различитих мера сличности или различитости на истом узорку расположивих мултиваријационих опсервација може, и често хоће, резултирати различитим бројем и / или конфигурацијом група (односно

[38] Симетричност представљених матрица блискости произилази из чињенице да оцењене вредности коришћених мера сличности или различитости / одстојања, сходно својству симетричности којим се одликују ове мере, не зависе од редоследа опсервација. У том смислу, бројни аутори (попут: *Kaufman & Rousseeuw*, 2005, стр. 16; *Bartholomew et al.*, 2008, стр. 19; *Vercellis*, 2009, стр. 296) истичу да је углавном довољно представити само вредности садржане унутар једне половине матрице, односно „троугла“ изнад или испод главне дијагонале, прецизније $[n\times(n-1)]/2$ вредности. Такође, будући да немају никакву конкретну улогу у процесу груписања, елементи садржани на главној дијагонали могу бити представљени и празним пољима (*Bartholomew et al.*, 2008, стр. 19).

[39] Примена представљеног израза за конверзију $s_{(q,h)} \rightarrow d_{(q,h)}$, подразумева да је разматрана мера сличности ненегативно (позитивно) дефинитна (енг. *non-negative definite*) и ограничена одозго са јединицом, односно да задовољава 1. услов – нормираност и ненегативност (*Timm*, 2002, стр. 519; *Johnson & Wichern*, 2007, стр. 677; *Everitt et al.*, 2011; стр. 46; *Nedović*, 2016, стр. 9; *Dinčić*, 2016, стр. 24).

решењима анализе груписања), сасвим је јасно због чега се доношење одлуке о избору конкретног критеријума за мерење блискости сматра, према мишљењу већине аутора (попут, на пример: *Hardle & Simar*, 2003, стр. 302; *Murtagh & Heck*, 1987, стр. 2–3), кључним питањем у контексту обезбеђивања ефикасне имплементације анализе груписања. Комплексност питања избора одговарајуће мере блискости произилази из изразите бројности[40], у литератури предложених, различитих начина за мерење сличности и / или различитости (одстојања) између парова опсервација и, сходно томе, неусаглашености ставова, упркос бројним компаративним студијама, у погледу најбоље мере која би била примењена у конкретним истраживачким околностима. Заправо, не постоји мера блискости која се може сматрати „оптималном“ или „савршеном“ за примену, како у апсолутном смислу, односно у свим истраживачким ситуацијама, тако и на нивоу специфичних, конкретних околности (*Manly & Navarro Alberto*, 2017, стр. 97; *Everitt et al.*, 2011, стр. 69).

Сходно наведеном, доминира мишљење да доношење одлуке о избору адекватне мере блискости у конкретном истраживању углавном представља активност која је под снажним утицајем субјективности, односно интуиције истраживача (*Johnson & Wichern*, 2007, стр. 673; *Izenman*, 2008, стр. 412; *Everitt et al.*, 2011, стр. 69; *Manly & Navarro Alberto*, 2017, стр. 83). Главне препоруке, које се наводе у литератури за потребе „усмеравања“ ове субјективности, сугеришу да избор у конкретном случају треба извршити у зависности од природе структуре анализираних варијабли, која може бити категоријска, нумеричка или мешовита, односно мерне скале (номинална, ординална, интервална, скала односа, или комбинација наведених) извршених мерења њихових вредности (*Timm*, 2002, стр. 515; *Hardle & Simar*, 2003, стр. 303; *Johnson & Wichern*, 2007; стр. 673; *Vercellis*, 2009, стр. 296; *Everitt et al.*, 2011, стр. 68; *Manly & Navarro Alberto*, 2017, стр. 97), подручја примене и циљева истраживања (*Johnson & Wichern*, 2007; стр. 673; *Shmueli, et al.*, 2005, стр. 211; *Timm*, 2002, стр. 515; *Manly & Navarro Alberto*, 2017, стр. 97), као и изабраног метода и / или процедуре груписања (*Kovačić*, 1994, стр. 273; *Everitt et al.*, 2011, стр. 68). Генерално, а у складу са најчешће коришћеном класификацијом мера блискости према природи анализираних варијабли, уколико се при груписању опсервација користе дихотомне (бинарне) или мултихотомне категоријске променљиве, најчешће разматране мере блискости, које се типично сматрају погодним у таквим околностима, јесу мере сличности (енг. *similarity* / *affinity* / *association measures*)[41] насупрот мерама различитости / одстојања које представљају адекватан избор када су све анализиране променљиве непрекидног нумеричког типа, мерене на интервалној или скали односа. У случају када су мултиваријационе опсервације описане модалитетима категоријских и вредностима нумеричких променљивих истовремено (када је структура променљивих мешовите

[40] У покушају да ближе представе размере (бројност) али и варијететност достигнућа у домену мерења блискости, *Berry & Linoff* (2004, стр. 360) наводе да постоји на десетине, ако не и стотине предложених и објављених мера сличности / различитости, док *Everitt et al.* (2011, стр. 68), користећи нешто „прецизирају“ формулацију, истичу да је у литератури дефинисан готово неограничен број различитих коефицијената сличности и / или различитости.

[41] Будући да дискусија о мерама сличности намењеним квантификовању степена блискости између јединица посматрања описаних категоријским (бинарним-симетричним или бинарним-асиметричним) карактеристикама превазилази предмет и оквире ове дисертације, детаљан преглед мера овог типа може се видети, на пример, у: *Aldenderfer & Blashfield* (1984, стр. 28–31), *Kovačić* (1994, стр. 266–270), *Timm* (2002, стр. 518–522), *Kaufman & Rousseeuw* (2005, стр. 20–30), *Johnson & Wichern* (2007, стр. 674–677), *Everitt et al.* (2011, стр. 46–48). Посебно се издваја компаративни приказ 76 мера сличности и одстојања у домену анализе бинарних категоријских променљивих, презентиран од стране *Choi, Cha & Tappert* (2010).

природе) најчешће се користи *Gower*-ова мера сличности[42]. Сходно наведеном, у наставку излагања извршен је детаљан преглед најчешће коришћених мера блискости (конкретно мера одстојања) погодних за спровођење анализе груписања у условима када су јединице посматрања описане искључиво вредностима непрекидних нумеричких променљивих, у складу са методолошким потребама спроведеног емпиријског истраживања у оквиру последњег поглавља дисертације.

2.2.3. Мере блискости објеката засноване на непрекидним нумеричким променљивим

У случају када су вредности свих анализираних променљивих измерене путем интервалне или скале односа, односно када су мултиваријационе опсервације описане вредностима p променљивих непрекидне нумеричке природе, блискост између расположивих n јединица посматрања се по правилу квантификује коришћењем неке од бројних мера различитости. При томе, доминантна улога у домену избора, захваљујући њиховом интуитивном значењу, припада метричким мерама различитости, односно мерама одстојања које углавном представљају специфичне форме посебне класе функција одстојања познатих под називом *Minkowski* метрике (енгл. *Minkowski distance metrics*), дефинисане у општем смислу следећим изразом (*Aldenderfer & Blashfield*, 1984, стр. 25; *Murtagh & Heck*, 1987, стр. 4):

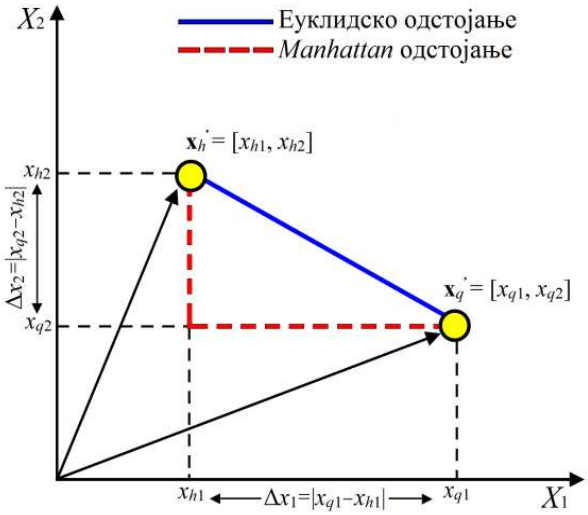
$$d_{Minkowski}(\mathbf{x}_h, \mathbf{x}_q) = \left[\sum_{j=1}^{p} \left|x_{hj} - x_{qj}\right|^{\psi}\right]^{\frac{1}{\psi}} = \sqrt[\psi]{\left(\left|x_{h1} - x_{q1}\right|^{\psi}\right) + \left(\left|x_{h2} - x_{q2}\right|^{\psi}\right) + \ldots + \left(\left|x_{hp} - x_{qp}\right|^{\psi}\right)}, \qquad (2.2.2)$$

где је ψ било који реалан број такав да је $\psi \geq 1$, односно $\psi = 1, 2, \ldots, \infty$. Сходно датом изразу, *Minkowski* одстојање између произвољно одабраних опсервација $\mathbf{x}_h' = [x_{h1}, x_{h2}, \ldots, x_{hp}]$ и $\mathbf{x}_q' = [x_{q1}, x_{q2}, \ldots, x_{qp}]$, у ознаци $d_{Minkowski}(\mathbf{x}_h, \mathbf{x}_q)$, заснива се, генерално, на израчунавању вредности корена степена ψ из збира апсолутних разлика појединачних вредности p нумеричких променљивих на нивоу посматраног пара опсервација подигнутих на степен ψ. Представљена мера назива се и L_ψ метрика (норма). Додељивањем различитих позитивних целобројних вредности за ψ, у дефинисаном опсегу, формулисан је читав низ различитих мера одстојања, које заједно чине такозвану групу *Minkowski* мера одстојања. Генерално, варирањем вредности параметра ψ мења се релативни значај додељен великим и малим разликама вредности анализираних променљивих у контексту њиховог доприноса формирању вредности мере одстојања (*Timm*, 2002, стр. 517; *Johnson & Wichern*, 2007, стр. 673). Наиме, у општем случају, што је вредност ψ већа, већи релативни значај се приписује великим разликама, а конкретна променљива за коју се координате две јединице посматрања више разликују имаће већи удео у детерминисању финалне вредности одстојања, у односу на остале променљиве које се карактеришу малим разликама, и обратно (*Kovačić*, 1994, стр. 260; *Bartholomew et al.*, 2008, стр. 30). Сходно наведеном, уколико је $\psi = 1$, *Minkowski* одстојање своди се на такозвано *Manhattan* одстојање (L_1-норма), док за $\psi = 2$, наведена мера резултира L_2-нормом, под називом Еуклидско одстојање.[43] Наведене мере, као специјални случајеви опште *Minkowski* мере

[42] Детаљно објашњење *Gower*-ове мере сличности може се наћи у: *Aldenderfer & Blashfield* (1984, стр. 31); *Shmueli, et al.* (2005, стр. 214); *Kaufman & Rousseeuw* (2005, стр. 35–37).

[43] Остале вредности ψ резултирају другим типовима одстојања, која се, генерално, не употребљавају често у контексту анализе груписања (*Sharma*, 1996, стр. 218) и, сходно томе, њихово разматрање није обухваћено излагањем у наставку.

одстојања, представљају најчешће коришћене метричке мере различитости у контексту испитивања блискости између $[n(n–1)/2]$ парова расположивих n мултиваријационих опсервација за потребе реализације циљева анализе груписања. Слика 2.2.2 илуструје разлику између *Manhattan* и Еуклидског одстојања на хипотетичком примеру дводимензионог ($p = 2$) координатног система.



Слика 2.2.2. Еуклидско и *Manhattan* одстојање између хипотетичких опсервација $\mathbf{x}_h$ и $\mathbf{x}_q$.
Извор: Ауторов приказ, прилагођено према *Varmuza & Filzmoser* (2009, стр. 59)

Manhattan одстојање[44], дефинисано изразом (2.2.3), утврђује се као збир апсолутних разлика између опсервација $\mathbf{x}_h$ и $\mathbf{x}_q$ по свакој од анализираних променљивих, односно:

$$d_{Manhattan}(\mathbf{x}_h, \mathbf{x}_q) = \sum_{j=1}^{p} \left| x_{hj} - x_{qj} \right| = \left| x_{h1} - x_{q1} \right| + \left| x_{h2} - x_{q2} \right| + \ldots + \left| x_{hp} - x_{qp} \right| . \qquad (2.2.3)$$

На Слици 2.2.2, ово одстојање кореспондира збиру дужина катета замишљеног правоуглог троугла[45]. С друге стране, Еуклидским одстојањем[46], дефинисаним изразом (2.2.4), мери се праволинијско, односно физичко или стварно геометријско растојање између опсервација у p-димензионом простору (*Kaufman & Rousseeuw*, 2005, стр. 11).

$$d_{Euclidean}(\mathbf{x}_h, \mathbf{x}_q) = \left[(\mathbf{x}_h - \mathbf{x}_q)^{\mathrm{T}} (\mathbf{x}_h - \mathbf{x}_q) \right]^{1/2} = \left[\sum_{j=1}^{p} (x_{hj} - x_{qj})^2 \right]^{1/2} = \sqrt{(x_{h1} - x_{q1})^2 + \ldots + (x_{hp} - x_{qp})^2} \quad (2.2.4)$$

[44] Често се у литератури за означавање L_1-норме, односно *Manhattan* одстојања (енгл. *Manhattan distance*) као синоними користе и следећи термини (*Everitt et al.*, 2011, стр. 50; *Nedović*, 2016, стр. 6): одстојање типа „градског блока“ (енгл. *city-block distance*), такси метрика (енгл. *taxicab metric*), или правоугаоно одстојање (енгл. *rectilinear distance*).

[45] Специфични, алтернативни називи L_1-метрике, предложене од стране *Herman*-а *Minkowski*-ог, потичу од аналогије са начином утврђивања растојања, односно удаљености између две произвољне локације (на пример, у контексту Слике 2.2.2, локације h и q) у граду где је саобраћајна инфраструктура изграђена тако да су улице или паралелне или се секу под правим углом, као што је то случај, на пример, на *Manhattan*-у.
Наиме, уколико се испрекидане црвене линије на Слици 2.2.2, посматрају као суседне улице, односно део мапе једног таквог града на којој су улице приказане у виду хоризонталних и вертикалних линија, „пут“ који такси мора прећи да би превезао корисника такси услуга од локације $\mathbf{x}_h$ до локације $\mathbf{x}_q$ једнак је збиру дужина односних улица које се секу под правим углом (у контексту разматране илустрације, $d_{Manhattan} = \Delta x_1 + \Delta x_2$). Тако утврђено, правоугаоно одстојање, представља „стварну“ удаљеност две локације, у контексту урбанистичког плана града презентираног типа, наспрам краће (праволинијске) удаљености мерене ваздушном линијом.

[46] Као синоними за Еуклидско одстојање (енгл. *Euclidean distance*) често се у литератури и истраживањима користе и називи попут, праволинијско одстојање (*straight-line distance*) и Питагорино одстојање (енгл. *Pythagorean distance*), будући да је формула за израчунавање вредности ове мере у случају када је број коришћених променљивих, $p = 2$, еквивалентна изразу, заснованом на Питагориној теореми, за израчунавање дужине хипотенузе правоуглог троугла (*Sharma*, 1996, стр. 19).

Насупрот *Manhattan* одстојању које узима у обзир само апсолутне разлике вредности p нумеричких променљивих, Еуклидско одстојање засновано је на израчунавању квадратног корена збира квадрата појединачних одступања (односно разлика, у ознаци Δx_j, за j = 1, 2,..., p) вредности p променљивих на нивоу произвољно одабраног пара опсервација, $\mathbf{x}_h$ и $\mathbf{x}_q$. У случају посматрања било које две опсервације у дводимензионом геометријском простору, Еуклидско одстојање између њих детерминисано је дужином праве линије или дужи која повезује тачке којима су опсервације представљене у координатном систему, односно дужином хипотенузе хипотетичког правоуглог трогла, као што је приказано на Слици 2.2.2, и представља најкраће могуће растојање између њих у простору, како наводи *Dinčić* (2016, стр. 26). Израз (2.2.4) представља генерализацију елабориранога дводимензионог концепта, заснованог на примени Питагорине теореме, на случај који подразумева анализирање већег броја променљивих, односно за $p > 2$. Поред наведеног, у емпиријским истраживањима, превасходно заснованим на примени *Ward*-ове или методе центроида у оквиру хијерархијске процедуре груписања, често се препоручује и употреба квадрата Еуклидског одстојања (енгл. *squared Euclidean distance*), као одговарајуће мере блискости (*Kovačić*, 1994, стр. 271; 273; *Hair et al.*, 2010, стр. 494), дефинисане следећим изразом (*Sharma*, 1996, стр. 187):

$$d^2_{Euclidean}(\mathbf{x}_h, \mathbf{x}_q) = [(\mathbf{x}_h - \mathbf{x}_q)^{\mathrm{T}}(\mathbf{x}_h - \mathbf{x}_q)] = \sum_{j=1}^{p}(x_{hj} - x_{qj})^2 = (x_{h1} - x_{q1})^2 + (x_{h2} - x_{q2})^2 + ... + (x_{hp} - x_{qp})^2 \quad .(2.2.5)$$

Иако њен начин израчунавања представља модификацију израза (2.2.4), квадрирано Еуклидско одстојање припада групи мера различитости, будући да не задовољава, претходно наведене, аксиоме метрике (*Timm*, 2002, стр. 517). Генерално, интуитивност и могућност интерпретације различитости као физичког одстојања у простору, допринели су да Еуклидско одстојање, према мишљењу већине истраживача, добије статус најпопуларније и најчешће коришћене мере одстојања у домену анализе груписања. У том смислу, важно је указати на кључна ограничења која карактеришу примену ове, али, између осталог, и већине других мера одстојања изведених на основу L_ψ-норме. Наиме, израчунавање Еуклидске мере одстојања директно на основу матрице оригиналних података углавном није препоручљиво у случају када се анализиране нумеричке променљиве одликују различитим типовима мерних скала, као и, уколико се исказују у различитим мерним јединицима и / или карактеришу различитим распонима забележених вредности. Потврђујући наведено, *Everitt et al.* (2011, стр. 67) напомињу да посматрање и употреба вредности променљивих изражених у различитим мерним јединицама и / или мерним скалама (на пример, ординална *vs*. интервална или рацио скала), односно њихово третирање као еквивалентних елемената у било ком смислу при израчунавању мере одстојања не представља смислен приступ, превасходно из угла логике, али и начина и значења интерпретације добијених вредности. Такође, разлике присутне у погледу мерних јединица и, консеквентно, опсегу у којем се налазе евидентиране вредности променљивих, условиће да променљиве које се одликују већим апсолутним вредностима имају нереално већи утицај на вредност мере одстојања у односу на променљиве које се одликују, у апсолутном смислу, малим опсегом вредности (*Varmuza & Filzmoser*, 2009, стр. 268; *Shmueli et al.*, 2005, стр. 210; *Hair et al.*, 2010, стр. 497). Другим речима, као последица наведених разлика, промене вредности једне или неколико променљивих (исказане на

већој скали вредности) издвајају се као значајније, односно доминантне у поступку израчунавања одстојања, у односу на промене у другим променљивим, које се одликују „мањим“ опсегом вредности (*Berry & Linoff*, 2004, стр. 364; 365). Полазећи од наведеног, али и чињенице да промена мерних јединица може значајно утицати на рангирање израчунатих одстојања између опсервација, а самим тим и на резултирајућу структуру груписања, на шта посебно указују *Shmueli, et al.* (2005, стр. 211), *Rencher* (2002, стр. 453) и *Kaufman & Rousseeuw* (2005, стр. 5–6), *Timm* (2002, стр. 517) истиче да је Еуклидска матрица одстојања најефективнија за нумеричке променљиве које се одликују међусобно сразмерним, односно пропорционалним вредностима. *Berry & Linoff* (2004, стр. 363) иду корак даље, истичући да уопште није важно у којим јединицама мере су исказане променљиве, односно димензије у координатном систему, све док су те јединице исте, подржавајући тако став *Hardle*-а *& Simar*-а (2003, стр. 306) према којем, исказивање вредности анализираних променљивих у истим мерним јединицама представља једну од кључних претпоставки при израчунавању мера одстојања заснованих на L_ψ-норми. Консеквентно, у ситуацији када су наведене различитости присутне, неопходно је, у оквиру фазе припреме података, извршити одговарајуће скалирање нумеричких променљивих, то јест, конверзију њихових вредности на заједничку (упоредиву) мерну скалу (у смислу истоветних јединица мере), а у циљу ублажавања и/или елиминисања ирелевантних разлика у величини вредности, које условљавају нереални однос значаја (или утицаја) појединачних променљивих при израчунавању вредности конкретне мере одстојања, нарочито уколико је, како истичу *Aldenderfer & Blashfield* (1984, стр. 21), реч о Еуклидском одстојању. Скалирање вредности променљивих (енг. *scaling*) може се извршити спровођењем поступка *z*-стандардизације, нормализације, или индексирања (детаљније видети у: *Berry & Linoff*, 2004, стр. 364), при чему се, као најчешће коришћен, издваја приступ заснован на стандардизацији, то јест, трансформацији оригиналних података појединачних променљивих у стандардизоване вредности са аритметичком средином 0 и стандардном девијацијом 1.

Међутим, при стандардизацији вредности променљивих, важно је узети у обзир и „цену“ спроведеног скалирања, односно последице, које су изазване поступком елиминисања нереално високог / ниског значаја појединачних променљивих при израчунавању вредности мере одстојања. Поменуте последице испољавају се у виду промена инхерентног (изворног) варијабилитета оригиналних променљивих, а тиме и њиховог реалног доприноса, односно утицаја на коначни исход анализе. Важност наведене констатације произилази из чињенице да променљиве које карактерише велики варијабилитет, генерално, имају и већи утицај при израчунавању одстојања, а тиме и доминантну улогу у детерминисању броја и структуре изведених група, будући да својим изражено дисперзованим вредностима знатно боље доприносе истицању разлика између јединица посматрања у односу на променљиве које се одликују малим, односно мањим варијабилитетом (*Izenman*, 2008, стр. 413; *Hair, et al.*, 2010, стр. 497; *Manly & Navarro Alberto*, 2017, стр. 84–85). Наиме, стандардизацијом на јединичну варијансу и средину нула, свим променљивим се аутоматски додељује исти значај (пондер) у поступку израчунавања вредности мере одстојања, а (евентуалне) разлике у погледу присутног варијабилитета на нивоу појединачних оригиналних променљивих се елиминишу. На тај начин се, напомињу *Aldenderfer & Blashfield* (1984, стр. 20), *Rencher* (2002, стр. 454) и

Kaufman & Rousseeuw (2005, стр. 9), неоправдано и значајно умањује, евентуално иницијално присутан, велики стварни допринос изразито варијабилних променљивих у погледу раздвајања група, а позитивни ефекти стандардизације у погледу елиминисања зависности резултата анализе груписања од мерних јединица, како истичу *Timm* (2002, стр. 517), *Everitt et al.* (2011, стр. 67) и *Manly & Navarro Alberto* (2017, стр. 168), потиру негативним ефектима минимизирања групних разлика,[47] условљавајући лошији (или мање добар) квалитет коначног исхода анализе.

Сходно наведеном, одлука о начину мерења одстојања условљена је претпоставком, начињеном од стране истраживача, у погледу начина детерминисања и висине релативног значаја (пондера) додељеног појединачним променљивим, у контексту реализације циљева анализе груписање. У том смислу, уколико се претпоставља да су анализиране променљиве, иначе исказане у различитим мерним јединицама, подједнако важне у поступку груписања, односно имају приближно исти утицај у поступку израчунавања одстојања, [48] тада се вредности мере одстојања, углавном, утврђују на основу стандардизованих вредности z_{ij} (за i=1, 2,…, n и j=1, 2,…, p), при чему су пондери, додељени појединачним променљивим, у ознаци w_j, међусобно једнаки, односно $w_j = 1$, за свако j=1, 2,…, p. Формула за израчунавање Еуклидског одстојања на бази стандардизованих података гласи:

$$d_{Euclidean}(\mathbf{z}_h, \mathbf{z}_q) = \left[\sum_{j=1}^{p} (z_{hj} - z_{qj})^2 \right]^{1/2} = \sqrt{(z_{h1} - z_{q1})^2 + (z_{h2} - z_{q2})^2 + ... + (z_{hp} - z_{qp})^2} \,. \qquad (2.2.6)$$

С друге стране, *Berry & Linoff* (2004, стр. 365) и *Kaufman & Rousseeuw* (2005, стр. 11) истичу да уколико истраживач, на основу субјективне процене и знања из подручја примене или разматрања одређених аспеката у вези са оригиналном матрицом података **X**, сматра да већи значај при мерењу одстојања између опсервација треба доделити некој (или неким) променљивој(им) у односу на преостале, тада је могуће извршити прилагођавање израза (2.2.6) у циљу уважавања неједнаког пондерисања важности појединачних променљивих. [49] Заправо, пондерисано Еуклидско одстојање засновано на стандардизованим опсервацијама, може се дефинисати на следећи начин (прилагођено према: *Vercellis*, 2009, стр. 298): [50]

$$d_{weighted\ Euclidean}(\mathbf{z}_h, \mathbf{z}_q) = \left[\sum_{j=1}^{p} w_j (z_{hj} - z_{qj})^2 \right]^{1/2} = \sqrt{w_1 (z_{h1} - z_{q1})^2 + w_2 (z_{h2} - z_{q2})^2 + ... + w_p (z_{hp} - z_{qp})^2} \,, \quad (2.2.7)$$

---

[47] Детаљније о контроверзама у погледу примене стандардизације и ефектима на израчунавање вредности одстојања, видети у *Aldenderfer & Blashfield* (1984, стр. 20–21; 26–28).

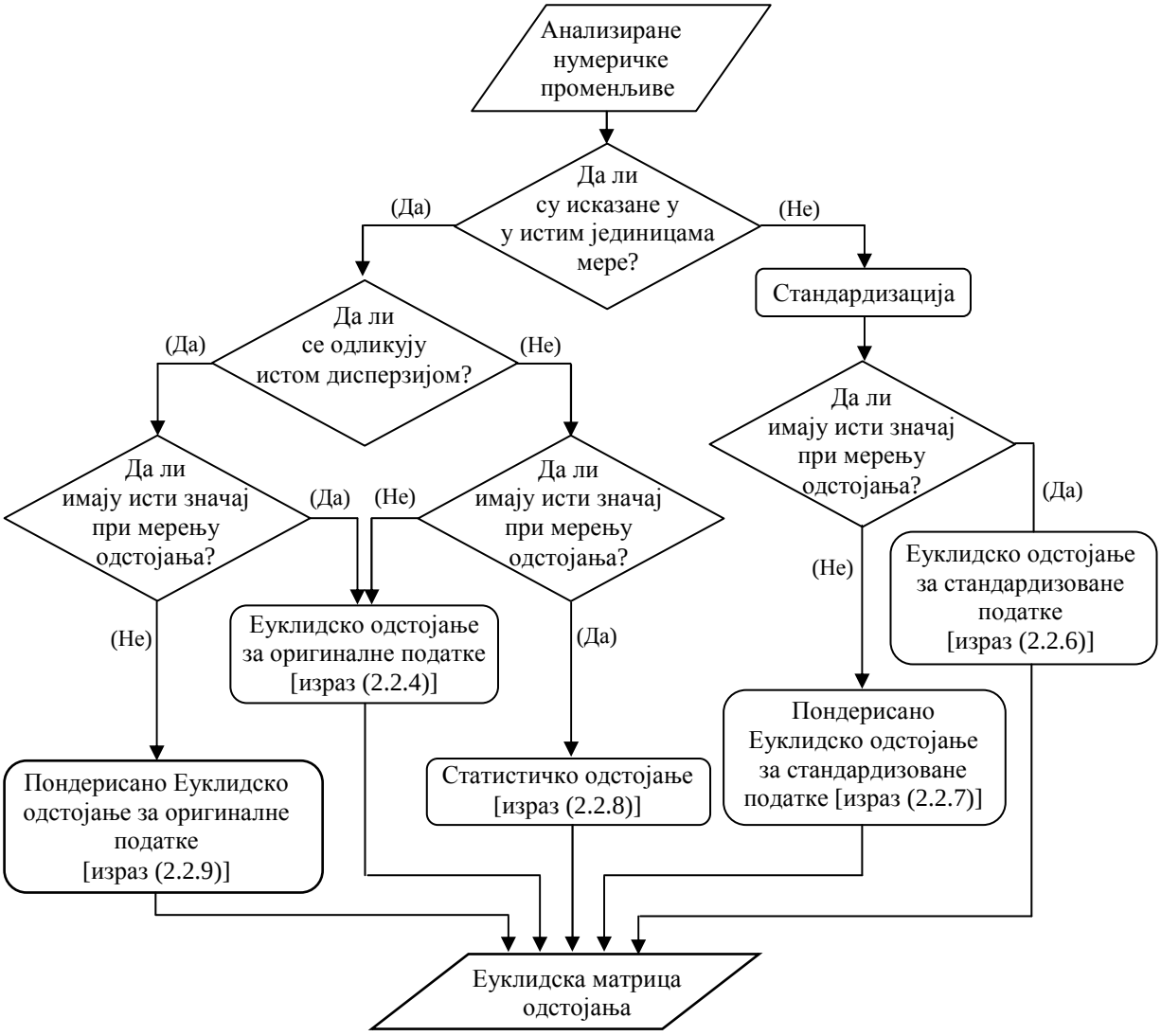
[48] *Everitt et al.* (2011, стр. 67) и *Manly & Navarro Alberto* (2017, стр. 85) истичу да једнако пондерисање значаја појединачних променљивих представља пожељнију и чешће примењивану алтернативу у практичним истраживањима заснованим на примени анализе груписања.

[49] У анализи груписања, пондерисање (енгл. *weighting*) представља активност усмерену на „подешавање“ вредности анализираних променљивих, односно прилагођавање њиховог релативног значаја, у циљу уважавања претпоставке према којој се поједине променљиве сматрају више или мање важним у поступку мерења одстојања између парова расположивих опсервација (*Aldenderfer & Blashfield*, 1984, стр. 21; *Berry & Linoff*, 2004, стр. 363; 365). Такође, иако је схватање суштине концепта пондерисања релативно једноставно, покушај његове имплементације у пракси углавном представља захтеван подухват, чему у значајној мери доприноси и недовољна расположивост конкретних смерница за валидну реализацију истог, наводе *Aldenderfer & Blashfield* (1984, стр. 21).

[50] Све мере одстојања су дефинисане на такав начин да дозвољавају различито пондерисање квантитативних променљивих (*Everitt et al.*, 2011, стр. 63).

где симбол w_j означава пондер приписан j-тој променљивој (за j=1, 2,…, p). Директно сразмерна висини релативног значаја односне променљиве, вредност појединачних пондера се креће у интервалу $[0 \leq w_j \leq 1]$, а њихов збир једнак је јединици, односно $\Sigma w_j = 1$.

Насупрот елабориранim околностима, под којима се стандардизација анализираних променљивих нужно подразумева и обавезно препоручује пре израчунавања мере одстојања (*Hair, et al.*, 2010, стр. 499; *Tufféry*, 2011, стр. 238), у ситуацији када су све променљиве исказане у истим (упоредивим) мерним јединицама, скалирање, односно стандардизација њихових вредности сматра се сувишном и непотребном (*Kaufman & Rousseeuw*, 2005, стр. 9; *Hair, et al.*, 2010, стр. 498). У том смислу, сходно приказу на Слици 2.2.3, израчунавање Еуклидског одстојања на бази оригиналних података (израз 2.2.4) обезбеђује задржавање и истицање неједнаког значаја појединачних променљивих при мерењу блискости, присутног као последица изразитих разлика у погледу варијабилитета њихових вредности.



Слика 2.2.3. *Дијаграм тока процеса доношења одлуке у погледу избора (непристрасног) начина израчунавања Еуклидског одстојања*

Извор: Ауторов приказ

С друге стране, уколико је потребно обезбедити изједначавање релативног значаја изразито варијабилилних променљивих, тада се мера под називом статистичко одстојање

(енгл. *statistical distance*), у ознаци $d_{statistical}$ ($\mathbf{x}_h$, $\mathbf{x}_q$), препоручује као адекватна алтернатива (*Sharma*, 1996, стр. 43). Наиме, статистичко одстојање, дефинисано изразом (2.2.8), заснива се на пондерисању појединачних променљивих одговарајућим вредностима w_j, детерминисаним у виду реципрочне вредности изабране мере (углавном је реч о стандардној девијацији) за квантификовање варијабилитета променљивих, односно (*Johnson & Wichern*, 2007, стр. 33):

$$d_{Statistical}(\mathbf{x}_h,\mathbf{x}_q)=\left[\sum_{j=1}^{p}\frac{1}{s_j}(x_{hj}-x_{qj})^2\right]^{1/2}=\sqrt{\frac{(x_{h1}-x_{q1})^2}{s_1}+\frac{(x_{h2}-x_{q2})^2}{s_2}+...+\frac{(x_{hp}-x_{qp})^2}{s_p}}, \qquad (2.2.8)$$

где s_j представља стандардну девијацију j-те променљиве (за j=1, 2,..., p). Полазећи од чињенице да је значај променљивих директно пропорционалан варијабилитету њихових вредности, евидентно је да се применом израза (2.2.8) постиже елиминисање разлика у варијабилитету, а тиме и уравнотежење релативног значаја појединачних променљивих, будући да се вредности пондера смањују са повећањем варијабилитета и обратно (*Sharma*, 1996, стр. 219). Статистичко одстојање представља посебан облик пондерисаног Еуклидског одстојања за оригиналне податке[51], дефинисан изразом (*Bartholomew et al.*, 2008, стр. 30):

$$d_{weighted\ Euclidean}(\mathbf{x}_h,\mathbf{x}_q)=\left[\sum_{j=1}^{p}w_j(x_{hj}-x_{qj})^2\right]^{\frac{1}{2}}=\sqrt{w_1(x_{h1}-x_{q1})^2+w_2(x_{h2}-x_{q2})^2+...+w_p(x_{hp}-x_{qp})^2}\ . \quad (2.2.9)$$

Специфичност статистичког одстојања огледа се у начину дефинисања вредности пондера заснованом на узимању у обзир информација о дисперзији променљивих и разлика које постоје између њих у том погледу. С друге стране, примена израза (2.2.7) уз детерминисање вредности пондера на бази субјективне процене истраживача и / или знања из подручја примене (енгл. *subject-matter knowledge*) сматра се примереним за потребе апострофирања (претпостављеног и / или анализираним теоријским концептима подржаног) неједнаког значаја појединачних променљивих при мерењу одстојања, иако се исте, исказане у упоредивим мерним јединицама, одликују приближно једнаким варијабилитетом.

Поред наведених специфичности везаних за израчунавање Еуклидског одстојања у контексту реализације циљева анализе груписања, важно је такође истаћи и изражену осетљивост ове мере на присуство нетипичних опсервација (енгл. *outliers*). Сходно наведеном, у таквим ситуацијима углавном се предлаже употреба робустнијих мера одстојања, попут *Manhattan* одстојања (*Kovačić*, 1994, стр. 260; *Shmueli et al.*, 2005, стр. 213; *Varmuza & Filzmoser*, 2009, стр. 268; *Tufféry*, 2011, стр. 237). Наиме, као што је истакнуто на почетку излагања у овом Одељку, *Manhattan* одстојање (то јест, L_1-норма) додељује мањи релативни значај великим разликама између јединица посматрања, будући да се заснива на апсолутним вредностима одстојања пре него њиховом квадрату као што је то случај код Еуклидског одстојања (то јест, L_2-норме). Другим речима, формулисано у контексту хипотетичког приказа на Слици 2.2.2, будући да се дужине катета не квадрирају,

[51] Уколико се анализиране променљиве одликују истим варијабилитетом, симболички $s_1=s_2=...=s_p$, тада се статистичко одстојање своди на стандардно Еуклидско одстојање за оригиналне (нестандардизоване) податке (*Johnson & Wichern*, 2007, стр. 34). У супротном, израз (2.2.8) сматра се еквивалентним изразу (2.2.6) предложеном за израчунавање Еуклидског одстојања на бази стандардизованих вредности променљивих, будући да се пондерисање засновано на варијабилитету, генерално, сматра еквивалентним поступку стандардизације (*Everitt et al.*, 2011, стр. 64).

мања је и вероватноћа да ће велике разлике на нивоу једне димензије (односно променљиве) доминирати укупним одстојањем, наводе *Berry & Linoff* (2004, стр. 363). Такође, наредни „недостатак“ Еуклидског одстојања огледа се и у чињеници да се приликом израчунавања ове мере одстојања не узима у обзир корелација која (евентуално) постоји између променљивих (*Shmueli et al.*, 2005, стр. 213; *Johnson & Wichern*, 2007, стр. 30). Сходно наведеном, у ситуацијима када су анализиране променљиве високо корелисане, већина аутора препоручује израчунавање *Mahalanobis*-овог одстојања (енгл. *Mahalanobis distance*), дефинисано као (*Martinez & Martinez*, 2007, стр. 432):

$$d_{Mahalanobis}(\mathbf{x}_h, \mathbf{x}_q) = \left[(\mathbf{x}_h - \mathbf{x}_q)^{\mathrm{T}} \, \mathbf{S}^{-1} (\mathbf{x}_h - \mathbf{x}_q) \right]^{1/2} = \sqrt{(\mathbf{x}_h - \mathbf{x}_q)^{\mathrm{T}} \, \mathbf{S}^{-1} (\mathbf{x}_h - \mathbf{x}_q)} \,, \tag{2.2.10}$$

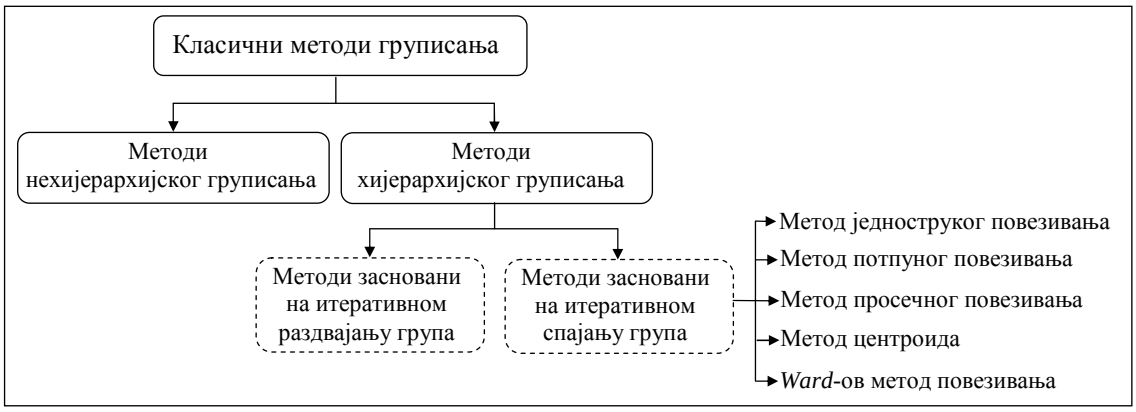
где симбол $\mathbf{S}^{-1}$ означава инверзну форму симетричне ($p \times p$) узорачке коваријационе матрице **S**. Будући да формулом за израчунавање обухвата и коваријациону структуру података (садржану унутар матрице **S**), представљена мера се још назива и мултиваријациона мера одстојања (*Kovačić*, 1994, стр. 261). Такође, *Vercellis* (2009, стр. 298) и *Sharma* (1996, стр. 44; 220) истичу да *Mahalanobis*-ово одстојање представља заправо генерализацију статистичког одстојања (израз (2.2.8)) за случај када се мери одстојање између опсервација описаних вредностима високо корелисаних променљивих које се карактеришу разликама у варијабилитету. Сходно наведеном, за променљиве које нису међусобно корелисане, исти аутори наводе да се *Mahalanobis*-ово одстојање (израз (2.2.10)) своди на израз (2.2.8), а уколико се при томе одликују и једнаким варијансама, тада је израз (2.2.10) еквивалентан изразу (2.2.4), односно Еуклидском одстојању за оригиналне податке.

Избор адекватне мере одстојања представља веома важан корак у имплементацији анализе груписања, будући да од адекватности истог зависи и квалитет, односно репрезентативност резултирајуће матрице блискости, као основе за примену неке од метода хијерархијске процедуре груписања, чије су одлике предмет дискусије у наредном одељку.

2.2.4. Методе груписања

За реализацију дефинисаних циљева анализе груписања, у литератури су предложени бројни методи за класификацију објеката. Иако су сви методи груписања, у суштини, засновани на уважавању концепта међусобне блискости анализираних опсервација, услед изразите варијететности у погледу коришћених статистичких критеријума при креирању група, резултати добијени применом различитих метода на истим мултиваријационим подацима углавном се, у значајној мери, разликују (*Aldenderfer & Blashfield*, 1984, стр. 15; 35; *Varmuza & Filzmoser*, 2009, стр. 265; *Manly & Navarro Alberto*, 2017, стр. 166–167). Као што је претходно истакнуто, категоризација аналитичких метода, осмишљених за спровођење анализе груписања, може се извршити на следеће две, широко дефинисане, групе, односно фамилије метода, и то: ▸ методи хијерархијског груписања и ▸ методи нехијерархијског груписања. Полазећи од чињенице да не постоји, у конкретном смислу, метод који се може сматрати универзалним, односно најпогоднијим за решавање свих могућих проблема у домену анализе груписања (*Izenman*, 2008, стр. 414; *Varmuza & Filzmoser*, 2009, стр. 294; *Manly & Navarro Alberto*, 2017, стр. 166), питање

избора адекватне методе груписања добија на значају. Према мишљењу бројних аутора (попут, на пример: *Aldenderfer & Blashfield*, 1984, стр. 35; *Sharma*, 1996, стр. 211; *Rencher*, 2002, стр. 479; *Kaufman & Rousseeuw*, 2005, стр. 37), правилан избор конкретне методе груписања, у оквиру претходно изабране групе метода, представља захтеван подухват, превасходно условљен типом и карактеристикама расположивих података, дефинисаним циљевима истраживања, мером блискости која је употребљена за сагледавање степена сличности / различитости између опсервација (ако метод захтева примену исте), али и карактеристикама, односно предностима и недостацима појединачних метода чија се могућност примене разматра. У том смислу, у наставку су изложене кључне карактеристике и специфичности везане за имплементацију најчешће коришћених (класичних) метода хијерархијске и нехијерархијске процедуре груписања, обухваћених класификацијом на Слици 2.2.4, при чему је, сходно потребама истраживања у дисертацији, посебна пажња посвећена методима хијерархијског груписања заснованим на итеративном удруживању (спајању) опсервација.



Слика 2.2.4. *Класични методи груписања*
Извор: Ауторов приказ

2.2.4.1. Методе хијерархијског груписања објеката и мере одстојања између група

Узимајући у обзир учесталост њихове употребе у практичним истраживањима, *Everitt et al.* (2011, стр. 110) истичу да хијерархијски методи представљају „кичму“ анализе груписања. Будући да не захтева *a priori* детерминисање (коначног) броја група унутар којих ће расположиве мултиваријационе опсервације бити распоређене (*Sharma*, 1996, стр. 211; *Everitt*, 2010, стр. 241), хијерархијска процедура подразумева сукцесивно проналажење конкретног решења груписања, „оптималног“ са аспекта иницијално дефинисаног критеријума за мерење одстојања између група, у оквиру сваког корака итеративног и иреверзибилног[52] процеса удруживања опсервација и/или група, или процеса поделе група опсервација (*Everitt et al.*, 2011, стр. 71).

[52] Иреверзибилност процеса груписања, у основи хијерархијских метода, огледа се у чињеници да једном реализовано удруживање или подела група у неком од претходних корака, не може бити модификовано или „исправљено“ у каснијим (наредним) корацима итеративне процедуре груписања (*Aldenderfer & Blashfield*, 1984, стр. 37; *Hardle & Simar*, 2003, стр. 308; *Everitt*, 2010, стр. 242; *Rencher*, 2002, стр. 455; *Johnson & Wichern*, 2007, стр. 695). Другим речима, након удруживања две опсервације унутар једне заједничке групе, или њиховог алоцирања у састав различитих група, није могуће извршити реалокацију посматраних јединица посматрања, у неком од наредних корака процеса груписања, којом би се постигло њихово накнадно раздвајање, односно удруживање, респективно (*Kaufman & Rousseeuw*, 2005, стр. 44). Према мишљењу истих аутора (2005, стр. 44–45), немогућност исправке и / или поништавања (потенцијално погрешних) одлука о спајању или раздвајању опсервација, донетих у претходним корацима, представља главно ограничење хијерархијских метода груписања, али уједно и један од кључних фактора који су допринели њиховој

Наведена процедура заснива се на коришћењу матрице оригиналних опсервација $\mathbf{X}_{[n\times p]}$ и / или матрице одстојања $\mathbf{D}_{[n\times n]}$ као улазних компоненти (*Hardle & Simar*, 2003, стр. 302; *Timm*, 2002, стр. 522; *Vercellis*, 2009, стр. 307). У том смислу, методи хијерархијског груписања не резултирају само једним, *k*-тим, решењем сачињеним од унапред специфицираног броја група, g (у ознаци, $k = 1,2,\ldots, g$), већ се њихова примена одликује креирањем одговарајуће хијерархијски уређене структуре распореда опсервација по групама, односно читавог низа могућих решења разматраног проблема груписања (*Timm*, 2002, стр. 522–523; *Dunham*, 2000, стр. 98; *Johnson & Wichern*, 2007, стр. 680). Прецизније, формираном хијерархијском структуром обухваћене су, из угла конкретно примењеног метода, све могуће поделе расположивог скупа / узорка јединица посматрања, односно серија решења у распону[53] од $g = n$ група до $g = 1$ група (*Kaufman & Rousseeuw*, 2005, стр. 44; *Bartholomew et al.*, 2008, стр. 19; *Varmuza & Filzmoser*, 2009, стр. 277). Полазећи од наведеног, може се констатовати да хијерархијско груписање представља аналитички поступак вођен подацима (*Shmueli et al.*, 2005, стр. 220; *Dunham*, 2000, стр. 95), чиме је детерминисан и примарно експлоративни карактер метода који се користе за његову операционализацију (*Sharma*, 1996, стр. 211; *Timm*, 2002, стр. 533). У зависности од тога да ли је развијање хијерархијске структуре распореда јединица посматрања по групама засновано на поступку итеративног спајања, или пак дељења група опсервација, појединачни методи хијерархијског груписања могу бити категорисани унутар једне од следеће две подгрупе метода, и то (*Kovačić*, 1994, стр. 258; 274; *Martinez & Martinez*, 2007, стр. 434; *Everitt et al.*, 2011, стр. 71): ▸ методи удруживања, и ▸ методи раздвајања.

Методи хијерархијског груписања засновани на удруживању

Класификација расположивих јединица посматрања унутар одговарајућих група применом прве категорије хијерархијских метода груписања, познатих и под називом агломеративни методи (енгл. *agglomerative hierarchical clustering methods*), заснована је на приступу груписања „од дна ка врху“ (енгл. *bottom-up clustering approach*) (*Dunham*, 2000, стр. 98; *Izenman*, 2008, стр. 411; *Vercellis*, 2009, стр. 308). Према овом приступу, на првом нивоу хијерархије, свака од n расположивих јединица посматрања третира се као засебна, савршено хомогена, једночлана група. Овако формирана, према речима аутора *Hardle & Simar* (2003, стр. 308; 315), „најфинија“ могућа (али не и практично употребљива) подела разматраног скупа / узорка мултиваријационих опсервација, представља иницијално „решење“ проблема груписања, сачињено од n група (симболички, $C^{(1)} = \{C_1, C_2, C_h,\ldots, C_q,\ldots, C_g\}$, где је $k = 1, 2,\ldots, g$ и $g = n$). У првом кораку процеса груписања, на основу елемената матрице одстојања $\mathbf{D}_{[n\times n]}$, идентификује се „најближи“ пар опсервација, односно две једночлане групе које се одликују најмањим одстојањем, на пример $d(\mathbf{x}_h, \mathbf{x}_q)$, за $h, q \in i = 1, 2,\ldots, n$, након чега се врши њихово спајање, фузија, то јест удруживање у нову (заједничку) групу, у ознаци $C_{hq}=C_h \cup C_q$. Извршеним удруживањем формира се наредно (потенцијално) решење проблема груписања, према којем су опсервације распоређене унутар $g = n–1$ група, односно $C^{(2)} = \{C_1, C_2,\ldots,C_{n-1}\}$. Реализација другог корака

популаризацији и интензивној експлоатацији у практичним истраживањима, будући да наведена ригидност у значајној мери доприноси редукцији времена потребног за пратећа рачунска израчунавања.

[53] Решење анализе груписања за које је број издвојених група једнак броју опсервација, симболички $g = n$, подразумева да свака јединица посматрања формира засебну, једночлану групу. С друге стране, решење сачињено само од једне групе ($g = 1$), означава да су све опсервације алоциране унутар једне, заједничке групе.

хијерархијске агломеративне процедуре груписања, нужно подразумева „ажурирање“ садржаја полазне матрице одстојања $\mathbf{D}_{[n\times n]}$ (алтернативно, у ознаци $\mathbf{D}^{(1)}$, где експонент означава редни број конкретног корака у процесу груписања), односно формирање редуковане $[(n-1)\times(n-1)]$ матрице одстојања, $\mathbf{D}^{(2)}$. Извођење матрице $\mathbf{D}^{(2)}$ заснива се на (а) елиминисању, из матрице одстојања $\mathbf{D}^{(1)}$, редова и колона које одговарају h-тој и q-тој опсервацији чије је удруживање извршено у претходном кораку, а затим и (б) додавању новог реда и колоне (у ознаци $h \cup q$) у којима су садржане вредности одстојања између новоформиране групе (C_{hq}) и свих преосталих ($i \neq h, q$) опсервација, односно једночланих група. Будући да се концепт мерења одстојања између две (појединачне) опсервације и одстојања између две групе опсервација (од којих барем једна није једночлана) не могу сматрати еквивалентним и подједнако очигледним (*Tuffery*, 2011, стр. 254), за утврђивање одстојања између формиране групе (C_{hq}), која се састоји из више опсервација, и свих преосталих појединачних опсервација (или, у каснијим фазама процеса груписања, других група опсервација) неопходно је извршити избор и примену конкретног правила, односно критеријума којим се дефинише начин мерења одстојања између група (енгл. *between-group proximity measure*) (*Sharma*, 1996, стр. 188). Вредности одстојања између група садрже основне информације неопходне за наставак хијерархијског процеса удруживања, а начин њиховог израчунавања, у значајној мери опредељује и детерминише правац у којем ће се даља груписања одвијати (*Varmuza & Filzmoser*, 2009, стр. 277; 294). У литератури је предложен велики број мера за израчунавање одстојања између група које садрже више од једне опсервације, при чему, сходно претходним констатацијама, имплементација сваке од њих резултира дефинисањем другачијег агломеративног хијерархијског метода груписања (*Aldenderfer & Blashfield*, 1984, стр. 38; *Rencher*, 2002, стр. 455–456; *Everitt et al.*, 2011, стр. 73; *Kovačić*, 1994, стр. 275; *Sharma*, 1996, стр. 188). У том смислу, издвајају се, као најчешће коришћене, следеће мере одстојања између група и кореспондентне појединачне методе агломеративне хијерархијске процедуре груписања:

- Метод једноструког повезивања (енгл. *single–linkage method*) – заснован на утврђивању најмањег одстојања између група (енгл. *minimum* / *nearest neighbor distance*).

Коришћењем ове методе удруживања, одстојање између произвољне опсервације $\mathbf{x}_s$ (при чему: $s \in t = 1, 2,..., n-2$, и $t \neq h, q$, тако да је $t \cup h \cup q = i = 1, 2,\ldots, n$), односно једночлане групе C_s (где: $\mathbf{x}_s \in C_s$) и новоформиране групе C_{hq} (где: $\mathbf{x}_h, \mathbf{x}_q \in C_{hq}$), дефинише се и мери као најмања вредност у низу утврђених одстојања између свих могућих парова појединачних опсервација унутар посматране две групе[54] (*Vercellis*, 2009, стр. 307):

$$d(C_{hq}, C_s) = min\left\{d(\mathbf{x}_h, \mathbf{x}_s),\ d(\mathbf{x}_q, \mathbf{x}_s)\right\}, \tag{2.2.11}$$

где су $d(\mathbf{x}_h,\ \mathbf{x}_s)$ и $d(\mathbf{x}_q,\ \mathbf{x}_s)$ елементи полазне матрице одстојања, $\mathbf{D}_{[n\times n]}$, односно $\mathbf{D}^{(1)}$. Поступак мерења одстојања између две групе, у случају када обе садрже више од једне опсервације, спроводи се у потпуности на истоветан начин, а једина разлика огледа се у поређењу већег броја парова опсервација, односно одстојања између њих. Слика 2.2.5. илуструје начин одређивања одстојања између група методом једноструког повезивања.

[54] Будући да се блискост између две групе утврђује као одстојање између најближег пара њима припадајућих опсервација, метод једноструког повезивања се често назива и метод најближих суседа (*Everitt et al.*, 2011, стр. 73).

• Метод потпуног повезивања (енгл. *complete–linkage method*) – заснован на утврђивању највећег одстојања између група (енгл. *maximum / farthest neighbor distance*).

Овај метод хијерархијске агломеративне процедуре груписања у потпуности је супротан претходно елаборираном методу, будући да се одстојање између две посматране групе дефинише се као највећа вредност од свих вредности одстојања утврђених између могућих комбинација појединачних опсервација у саставу две различите групе, то јест, између парова сачињених од по једне опсервације из сваке групе (Слика 2.2.5 (б)). У форми конкретног израза, правило најудаљенијег суседа, може се дефинисати на следећи начин:

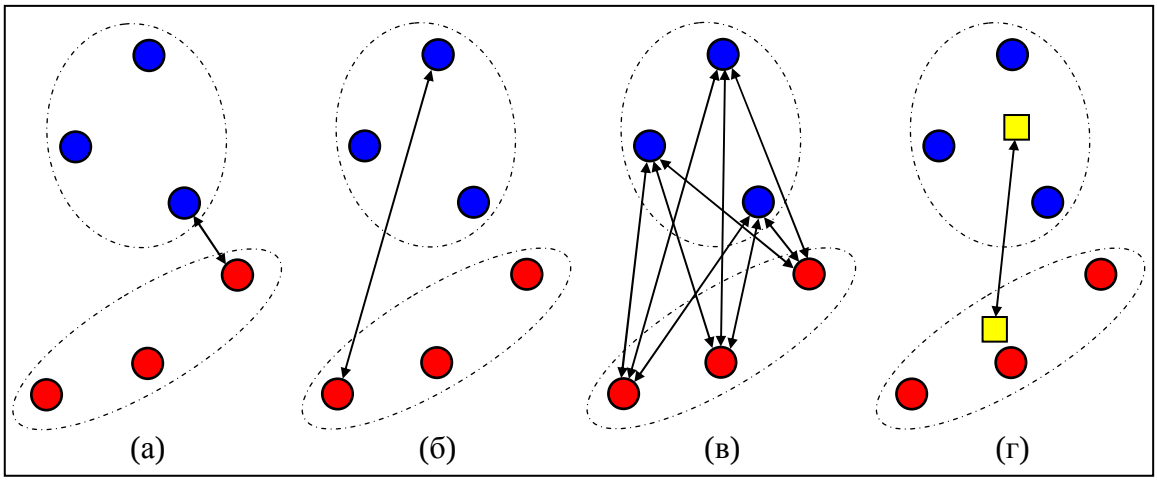
$$d(C_{hq}, C_s) = max\{d(\mathbf{x}_h, \mathbf{x}_s),\ d(\mathbf{x}_q, \mathbf{x}_s)\}\text{, где је: } \mathbf{x}_s \in C_s,\ \text{ а } \mathbf{x}_h, \mathbf{x}_q \in C_{hq}. \qquad (2.2.12)$$

• Метод просечног повезивања (енгл. *average–linkage method*) – заснован на утврђивању просечног одстојања између група (енгл. *average distance*).

Насупрот методама једноструког и потпуног повезивања, код којих је одређивање одстојања између група, а тиме и њихово комбиновање у наредним корацима процеса груписања, у потпуности детерминисано одстојањем између само једног пара опсервација (енгл. „*a single link*“), метод просечног повезивања узима у обзир појединачна одстојања између свих парова опсервација, њиховим инкорпорирањем у поступак израчунавања финалне вредности одстојања између две групе. Прецизније, применом овог метода удруживања, одстојање између група се израчунава као просечна вредност појединачних одстојања (енгл. „*an average link*“) утврђених између свих парова опсервација садржаних унутар посматране две групе, при чему упарене опсервације припадају различитим групама (Слика 2.2.5 (в)), односно:

$$d(C_{hq}, C_s) = \frac{1}{n_{hq} \cdot n_s} \sum_{hq=1}^{n_{hq}} \sum_{s=1}^{n_s} \{d_{(hq,s)}\} = \frac{d(\mathbf{x}_h, \mathbf{x}_s) + d(\mathbf{x}_q, \mathbf{x}_s)}{n_{hq} \cdot n_s}, \qquad (2.2.13)$$

где $\mathbf{x}_s \in C_s$, и $\mathbf{x}_h$, $\mathbf{x}_q \in C_{hq}$, а симболи n_{hq} и n_s означавају величину група C_s и C_{hq}, респективно.



Слика 2.2.5. Графички приказ различитих начина одређивања одстојања између група: минимално (а), максимално (б), просечно (в) и одстојање центроида (г)

Извор: Ауторов приказ

• Метод центроида (енгл. *centroid–linkage method*) – заснован на утврђивању одстојања између центроида група (енгл. *centroid distance*).

Насупрот претходно елабориранim методама удруживања, примена методе центроида ограничена је искључиво на случај када су анализирани мултиваријациони подаци мерени на непрекидној нумеричкој скали (*Kaufman & Rousseeuw*, 2005, стр. 227). Такође, поступак израчунавања одстојања између група у контексту ове методе захтева коришћење елемената матрице оригиналних података $\mathbf{X}_{[n\times p]}$, а не матрице одстојања $\mathbf{D}_{[n\times n]}$ (*Everitt et al.*, 2011, стр. 75). Наведена разлика произлази из промене фокуса са утврђивања минималне, максималне или просечне вредности одстојања између појединачних парова опсервација на мерење одстојања између p-димензионих вектора просечних вредности опсервација у саставу група, који се називају центроиди група, у ознаци $\overline{\mathbf{x}}_k$ (k = 1, 2,…, g) и утврђују на следећи начин:

$$\overline{\mathbf{x}}_k = \frac{\sum_{i=1}^{n_k}\mathbf{x}_i}{n_k} = \begin{bmatrix} \overline{x}_{1k} \\ \overline{x}_{2k} \\ \vdots \\ \overline{x}_{pk} \end{bmatrix}, \text{ где је: } \overline{x}_{jk} = \frac{\sum_{i=1}^{n_k} x_{ijk}}{n_k}, \text{ за } i = 1,..., n_k,\ j = 1, 2,..., p. \qquad (2.2.14)$$

Полазећи од чињенице да је свака група репрезентована припадајућим центроидом, одстојање између две групе (на пример, C_s и C_{hq}) дефинише се као вредност одговарајуће мере одстојања (најчешће квадрат Еуклидског одстојања) између њихових центроида (у ознаци $\overline{\mathbf{x}}_s$ и $\overline{\mathbf{x}}_{hq}$, респективно), односно симболички:

$$d(C_{hq}, C_s) = d^2_{Euclidean}(\overline{\mathbf{x}}_{hq}, \overline{\mathbf{x}}_s) = [(\overline{\mathbf{x}}_{hq} - \overline{\mathbf{x}}_s)^{\mathrm{T}}(\overline{\mathbf{x}}_{hq} - \overline{\mathbf{x}}_s)] = \sum_{j=1}^{p}(\overline{x}_{j(hq)} - \overline{x}_{j(s)})^2 . \qquad (2.2.15)$$

• *Ward*-ов метод (енгл. *Ward's method*) – заснован на утврђивању прираштаја (повећања) укупне суме квадрата одступања (одстојања) између свих опсервација и центроида унутар појединачних група, обухваћених решењем на посматраном нивоу хијерархијске структуре, а који се очекује као исход евентуалног удруживања две конкретне групе. Попут методе центроида, *Ward*-ов метод хијерархијског удруживања (повезивања) намењен је мерењима извршеним на интервалној / скали односа (*Kaufman & Rousseeuw*, 2005, стр. 230; *Bartholomew et al.*, 2008, стр. 24) и захтева употребу квадрата Еуклидског одстојања. Имплементација *Ward*-овог метода иницијално подразумева израчунавање збира квадрата одступања (енгл. *Sum of Squared within-group Errors*, *SSE*) између опсервација унутар сваке појединачне k-те групе и њеног центроида, $\overline{\mathbf{x}}_k$ (k=1,2,...,s,...,hq,…,g)[55], на датом нивоу хијерархије удруживања, путем израза:

$$SSE_k = \sum_{i=1}^{n_k} d^2_{Euclidean}(\mathbf{x}_{i(k)}, \overline{\mathbf{x}}_k) = \sum_{i=1}^{n_k}\left[(\mathbf{x}_{i(k)} - \overline{\mathbf{x}}_k)^{\mathrm{T}}(\mathbf{x}_{i(k)} - \overline{\mathbf{x}}_k)\right] = \sum_{i=1}^{n_k}\sum_{j=1}^{p}(x_{ij(k)} - \overline{x}_{j(k)})^2 . \qquad (2.2.16)$$

Полазећи од наведеног, за издвојено решење проблема груписања $C = \{C_1,...,C_s,...,C_{hq},\ldots,C_g\}$, карактеристично за посматрани ниво хијерархијског удруживања, укупна вредност збира квадрата одступања унутар свих појединачних група k, у ознаци

[55] Збир квадрата одступања свих опсервација унутар конкретне групе од припадајућег центроида (у ознаци SSE_k, за k=1,…,g), представља одговарајућу меру интерне хомогености посматране k-те групе (*Sharma*, 1996, стр. 193).

SSE_C[56], дефинише се следећим изразом (*Johnson & Wichern*, 2007, стр. 693; *Kaufman & Rousseeuw*, 2005, стр. 231):

$$SSE_C = \sum_{k=1}^{g} SSE_k = SSE_1 + SSE_2 + ... + SSE_s + ... + SSE_{hq} + ... + SSE_g \,. \tag{2.2.17}$$

У наставку поступка утврђивања *Ward*-ове мере одстојања између група, разматрају се сви могући парови расположивих група на датом нивоу хијерархије и квантификује допринос њиховог (евентуалног) удруживања повећању вредности SSE_C на наредном нивоу. Наиме, уколико C_{hqs} представља групу насталу удруживањем група C_{hq} и C_s, тада сума квадрата одстојања свих опсервација унутар формиране групе ($n_{hqs}=n_{hq}+n_s$) од центроида, $\bar{\mathbf{x}}_{hqs}$, гласи:

$$SSE_{hqs} = \sum_{i=1}^{n_{hqs}} d^2_{Euclidean}(\mathbf{x}_{i(hqs)}, \bar{\mathbf{x}}_{hqs}) = \sum_{i=1}^{n_{hqs}} \sum_{j=1}^{p} (x_{ij(hqs)} - \bar{x}_{j(hqs)})^2 \,, \tag{2.2.19}$$

при чему је, према правилу, $SSE_{hqs} > (SSE_{hq} + SSE_s)$. Полазећи од непромењених вредности суме квадрата одступања унутар осталих група, очекивано повећање укупне вредности SSE_C, као резултат евентуалног удруживања група C_{hq} и C_s, дефинисано путем израза (*Rencher*, 2002, стр. 466; *Kaufman & Rousseeuw*, 2005, стр. 231):

$$d(C_{hq}, C_s) = \Delta SSE_C = SSE_{hqs} - (SSE_{hq} + SSE_s) \,, \tag{2.2.20}$$

представља, заправо, критеријум за мерење одстојања између група који се налази у основи *Ward*-ове методе хијерархијског удруживања. Алтернативни израз за израчунавање ове мере одстојања гласи (*Tuffery*, 2011, стр. 257):

$$d(C_{hq}, C_s) = \Delta SSE_C = \frac{n_{hq} \cdot n_s}{n_{hq} + n_s} d^2_{Euclidean}(\bar{\mathbf{x}}_{hq}, \bar{\mathbf{x}}_s) = \frac{n_{hq} \cdot n_s}{n_{hq} + n_s} \sum_{j=1}^{p} (\bar{x}_{j(hq)} - \bar{x}_{j(s)})^2 \,. \tag{2.2.21}$$

У контексту претходних разматрања, *Aldenderfer & Blashfield* (1984, стр. 59) напомињу да се резултати груписања, добијени применом различитих метода на истим подацима, могу у значајној мери међусобно разликовати. Такође, полазећи од чињенице да се ниједан метод хијерархијског удруживања не може издвојити као универзално најбољи (*Manly & Navarro Alberto*, 2017, стр. 166; *Kaufman & Rousseeuw*, 2005, стр. 238), избор адекватне процедуре за мерење одстојања између група у конкретним истраживачким околностима је директно условљен дефинисаним циљевима истраживања, својствима расположивог скупа / узорка мултиваријационих података, као и кључним карактеристикама и појединачних метода хијерархијског удруживања[57] (*Aldenderfer & Blashfield*, 1984, стр. 59; *Sharma*, 1996, стр. 211; *Kaufman & Rousseeuw*, 2005, стр. 37).

---

[56] У условима када су све групе, обухваћене конкретним решењем проблема груписања, савршено хомогене, односно када се састоје из само једне опсервације (решење карактеристично за иницијални (нулти) ниво хијерархијске структуре), појединачне вредности SSE_k (за $k = 1,\ldots, g$ и $g = n$, где n означава укупан број опсервација) износе нула (то јест, $\forall\, SSE_k = 0$), а самим тим је и укупна вредност овог показатеља за конкретно решење $C = \{C_1, C_2,\ldots, C_g\}$, односно $SSE_C = 0$. На другом крају екстрема, када су све опсервације удружене унутар једне (заједничке) групе, односно када је $g = 1$, вредност SSE_C је детерминисана следећим изразом (*Johnson & Wichern*, 2007, стр. 693):

$SSE_C = \sum_{i=1}^{n} d^2_{Euclidean}(\mathbf{x}_i, \bar{\mathbf{x}}) = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)^2$, где је: $\bar{\mathbf{x}} = (\sum_{i=1}^{n} \mathbf{x}_i) / n$. (2.2.18)

Између наведених екстрема, вредност SSE_C се, у већој или мањој мери, сукцесивно повећава, као последица удруживања група у сваком кораку хијерархијског процеса груписања, при чему, у ретким, најповољнијим ситуацијама, промена вредности овог показатеља може и изостати.

[57] Детаљан преглед и компарација кључних карактеристика везаних за примену појединачних, најчешће коришћених, метода хијерархијског удруживања, представљен је од стране следећих аутора: *Kaufman & Rousseeuw* (2005, стр. 238–243), *Rencher* (2002, стр. 471–479), *Everitt et al.* (2011, стр. 79; 83), *Sharma* (1996, стр. 211–217).

Међутим, у ситуацијама када *a priori* аргументи нису довољни за сужавање избора на један метод, тада се, у циљу „поједностављења“ процеса доношења одлуке, углавном препоручује експлоративни приступ заснован на имплементацији више различитих процедура за мерење одстојања између група и компарацији резултирајућих (финалних) класификација[58] (*Kaufman & Rousseeuw*, 2005, стр. 37; *Sharma*, 1996, стр. 217), из угла њихове конзистентности са оригиналном матрицом одстојања и степена интерпретабилности. Након избора методе хијерархијске удруживања, односно дефинисања критеријума за сагледавање степена блискости између група, у оквиру другог корака процеса груписања[59], неопходно је одредити одстојања између групе (C_{hq}) формиране у претходном кораку и преосталих (једночланих) група на датом нивоу удруживања. Израчунате вредности одстојања представљају елементе уведеног реда и колоне (у ознаци $h \cup q$) у саставу, такозване, редуковане [(n–1)×(n–1)] матрице одстојања $\mathbf{D}^{(2)}$. Ново „потенцијално“ решење разматраног проблема груписања (то јест, $C^{(3)} = \{C_1, C_2,\ldots,C_{n-2}\}$), обезбеђује се удруживањем две конкретне групе које се одликују најмањим међусобним одстојањем у новоформираној матрици, у односу на међусобну удаљеност свих могућих[60] парова група које постоје на датом нивоу хијерархије. У зависности од структуре идентификованог пара најсличнијих група, *Kovačić* (1994, стр. 274; 275) прецизира следећа два могућа исхода спроведеног поступка удруживања, и то: (а) формирање нове групе (ако су две једночлане групе издвојене као међусобно најближе), или (б) повећање величине већ постојеће вишечлане групе, придруживањем нове опсервације њеном саставу (ако је одстојање неке од једночланих група од раније формиране групе евидентирано као најмање).

Представљени поступак груписања у оквиру II корака, заснован на израчунавању елемената редуковане матрице одстојања[61], идентификовању две међусобно најближе групе (сходно примењеном хијерархијском методу удруживања) и њиховој фузији унутар заједничке групе, итеративно се спроводи у сваком од преосталих n–3 корака започетог процеса агломерације. При томе, из итерације у итерацију, величина претходно формираних група се постепено повећава, а њихов укупан број смањује за 1, у односу на решење остварено у претходном кораку (*Kovačić*, 1994, стр. 274; *Rencher*, 2002, стр. 456). Прецизније, у кораку 1, 2,..., n–1 хијерархијског процеса удруживања, број група обухваћен предложеним решењем биће, респективно, $g = n$–1, n–2,..., 2, 1 (*Sharma*, 1996, стр. 190). Алоцирањем свих n анализираних опсервација унутар једне (заједничке) групе, односно формирањем коначног решења, облика $C^{(n)} = \{C_g\}$, где је $g = 1$, поступак груписања се завршава. У том смислу, хијерархијска процедура груписања, заснована на удруживању опсервација и / или група опсервација резултира читавом серијом (укупно n)

---

[58] Поступак компарације резултата добијених применом различитих хијерархијских метода удруживања, представљен је након елаборирања методолошких одређења на нивоу појединачних корака агломеративне процедуре груписања.

[59] Важно је такође напоменути да се одлука у погледу избора конкретног метода хијерархијског удруживања, донета у оквиру II корака процеса груписања, не може накнадно мењати у неком од наредних корака започетог процеса.

[60] Генерално, у i-том кораку процеса удруживања (за i = 1, 2,..., n–1), укупан број могућих парова расположивих група детерминисан је изразом [(n–i+1)×(n–i)] / 2. Сходно наведеном, у другом кораку хијерархијског процеса груписања (односно, за i = 2), постоји укупно [(n–1)×(n–2)] / 2 могућих парова група.

[61] У сваком од укупно n–1 корака хијерархијског процеса груписања, формира се редукована [(n–i+1)×(n–i+1)] матрица одстојања $\mathbf{D}^{(i)}$, за i = 1,2,..., n–1, (*Izenman*, 2008, стр. 415), као неопходне основе за реализацију удруживања две међусобно најближе групе. На последњем нивоу хијерархијске структуре решења, све јединице посматрања су интегрисане унутар једног заједничког кластера, а одговарајућа (1×1) матрица одстојања је нула матрица, односно $\mathbf{D}^{(n)} = 0$, означавајући завршетак имплементираног поступка груписања.

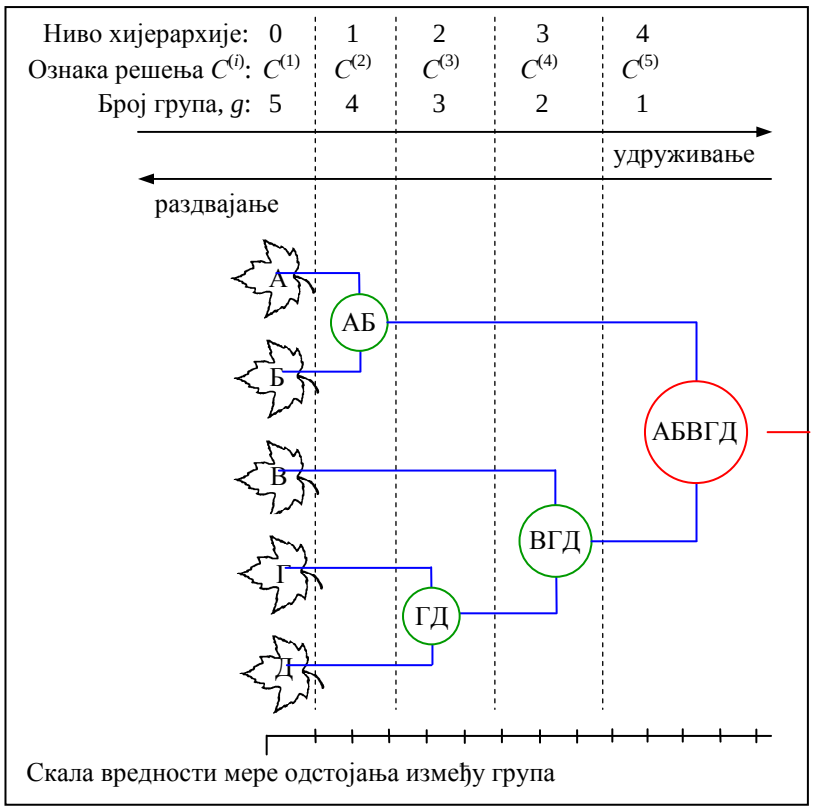
могућих решења проблема груписања, у ознаци $C^{(i)}$ (за i = 1, 2,…, n) сачињеним од одређеног броја (g) издвојених група C_k (за k = 1, 2,.., g), на сваком нивоу развијене хијерархијске структуре. Прецизније, формираном хијерархијском структуром могућих подела расположивог узорка опсервација, величине n, обухваћена су сва сукцесивна решења од првог, $C^{(1)}=\{C_1, C_2,\ldots, C_{g=n}\}$, карактеристичног за иницијални (нулти) ниво хијерархије, другог, $C^{(2)}=\{C_1, C_2,\ldots, C_{g=n-1}\}$, издвојеног на првом нивоу хијерархије, трећег, $C^{(3)}=\{C_1, C_2,\ldots, C_{g=n-2}\}$, на другом нивоу, све до решења изведеног на последњем нивоу хијерарије, облика $C^{(n)} = \{C_{g=1}\}$.

Комплетан процес удруживања по појединачним корацима и резултирајућа хијерархијска класификација анализираних јединица посматрања могу се визуелно представити коришћењем графичког приказа у форми стабло дијаграма (енгл. *tree-diagram*), познатог под називом дендрограм. Реч је о дводимензионом дијаграму који може бити конструисан хоризонтално или вертикално[62], у зависности од тога да ли су информације у погледу вредности коришћене мере одстојања између група представљене дуж апцисе или ординате, респективно. На другој оси садржане су информације о јединицима посматрања чије се груписање спроводи. Слика 2.2.6 илуструје хипотетички пример хоризонталног дендрограма креираног као резултат примене једне од метода хијерархијске процедуре груписања, засноване на удруживању (посматрано са лева на десно) пет јединица посматрања, означених симболима {А, Б, В, Г, Д}. Појединачне опсервације називају се листови (енгл. *leaves*), док хоризонталне праве линије (означене плавом бојом) коришћене за њихово повезивање (удруживање) представљају гране (енгл. *branches*) формираног хијерархијског стабла. Гране су међусобно повезане одговарајућим чворовима (енгл. *nodes*) који репрезентују групе настале удруживањем више опсервација (на датој слици, чворови су уоквирени зеленом бојом). Положај појединачних чворова у односу на (хоризонталну) осу, дуж које су евидентиране вредности коришћене мере одстојања између група, одговара нивоу блискости на којем је извршено удруживање појединачних опсервација и / или група опсервација у сваком кораку процеса груписања (*Johnson & Wichern*, 2007, стр. 682). Завршни чвор (енгл. *terminal node*), уоквирен црвеном бојом на истој слици, репрезентује финално решење, $C^{(n)}=\{C_{g=1}\}$, односно групу која садржи свих n опсервација и назива се корен хијерархијске стабло-структуре (енгл. *root of hierarchical tree-structure*). Главна предност дендрограма огледа се у обезбеђивању јасног и једноставног за тумачење сумарног графичког приказа процеса развоја хијерархијске стабло–структуре могућих решења проблема груписања. Међутим, *Izenman* (2008, стр. 412) напомиње да се у случају коришћења узорака велике величине, наведене погодности дендрограма могу показати неодрживим. *Varmuza & Filzmoser* (2009, стр. 294) допуњују изнету констатацију, наводећи да дендрограм постаје и сувише комплексан за разумевање и тумачење уколико је број, анализом обухваћених, јединица посматрања већи од 100.

Поред дендрограма, изведени резултати и ток хијерархијског процеса груписања могу се алтернативно сумирати и у форми табеларног приказа под називом шема агломерације или распоред удруживања (енгл. *agglomeration schedule*) (*Izenman*, 2008, стр.

[62] Поред наведене две могуће форме дендрограма (хоризонатални и вертикални), у оквиру последњег поглавља, приликом презентовања резултата спроведеног емпиријског истраживања, представљен је и иновативни облик дендрограма, назван комбиновани (хоризонтално-вертикални) дендрограм, нарочито погодан у случају анализирања већег броја опсервација.

415; *Bartholomew et al.*, 2008, стр. 22). Колоне у овој табели садрже информације које се односе на редни број корака у процесу груписања, ознаке опсервација и / или група опсервација удружених у сваком кораку, вредност мере одстојања између група при којој је конкретно спајање извршено, као и укупан број формираних група у сваком кораку.



Слика 2.2.6. *Дендрограм*

Извор: Ауторов приказ

Такође, информације презентоване у форми дендрограма или табеларног приказа распореда удруживања, могу бити употребљене за формирање, такозване, кофенетичке матрице одстојања (енгл. *cophenetic matrix*), у ознаци $\mathbf{D}_C$ (*Jain & Dubes*, 1988, стр. 166). Одређивање елемената симетричне ($n \times n$) матрице $\mathbf{D}_C$ спроводи се тако што се свим паровима опсервација из две различите групе које се (по први пут) спајају у једну заједничку додељује иста вредност одстојања, и то она при којој је удруживање и извршено (*Kovačić*, 1994, стр. 281). Прецизније, hq-ти елемент кофенетичке матрице (у ознаци $d_c(\mathbf{x}_h, \mathbf{x}_q)$, где $h, q \in i = 1, 2, \ldots, n$) представља вредност коришћене мере одстојања између група (у ознаци $d(\mathbf{x}_h, \mathbf{x}_q)$ или $d(\mathrm{C}_h, \mathrm{C}_q)$), при којој је извршено спајање h-те и q-те опсервације (то јест, једночланих група) унутар исте групе по први пут (*Halkidi, et al.*, 2001, стр. 128; *Martinez & Martinez*, 2007, стр. 436), односно „висину“ чвора који репрезентује новоформирану групу C_{hq} (*Everitt et al.*, 2011, стр. 91) на дендрограму. Уколико се у једном од наредних корака врши удруживање опсервације $\mathbf{x}_s$ и претходно формиране групе C_{hq}, вредности sh-тог, sq-тог елемента матрице $\mathbf{D}_C$, симболички се могу представити на следећи начин: $d_c(\mathbf{x}_s, \mathbf{x}_h) = d(\mathbf{x}_s, C_{hq})$, и $d_c(\mathbf{x}_s, \mathbf{x}_q) = d(\mathbf{x}_s, C_{hq})$, респективно. Формирана кофенетичка матрица одстојања представља полазну основу у израчунавању показатеља под називом кофенетички коефицијент корелације (енгл. *CoPhenetic Correlation Coefficient*, *CPCC*), иницијално предложеног од стране *Sokal*-а & *Rohlf*-а (1962).

Према њиховом тумачењу (1962, стр, 38), реч је о статистичком критеријуму за квантитативну оцену (мерење) „величине“ губитка информација узрокованог формирањем хијерархијске стабло–структуре скупа могућих решења (у графичком смислу, дендрограма) применом конкретне методе удруживања опсервација. Утврђени губитак информација, исти аутори дефинишу као дисторзију (или неусаглашеност) између оригиналних и кофенетичких одстојања индивидуалних опсервација, односно елемената ($n \times n$) матрице $\mathbf{D} = [d(\mathbf{x}_s, \mathbf{x}_h)]$ и матрице $\mathbf{D}_C = [d_c(\mathbf{x}_s, \mathbf{x}_h)]$, за $h, s \in i = 1, 2, \ldots, n$. У том смислу, *CPCC* омогућава евалуацију ефикасности појединачних метода удруживања, али и њихову међусобну компарацију у погледу постигнутог степена усклађености резултирајуће хијерархијске структуре груписања (дендрограма) са иницијално идентификованим обрасцем блискости између опсервација, репрезентованим оригиналном матрицом одстојања $\mathbf{D}_{[n \times n]}$ (*Farris*, 1969, стр. 279; *Aldenderfer & Blashfield*, 1984, стр. 62). Вредности *CPCC*-а, утврђене за различита решења процеса груписања добијена применом појединачних метода удруживања, еквивалентне су апсолутним вредностима коефицијента корелације за $[n(n–1)/2]$ парова одстојања између појединачних опсервација (као елемената оригиналне матрице **D**) и вредности одстојања између група при којима је извршено спајање кореспондентних опсервација (као елемената кофенетичке матрице $\mathbf{D}_C$) (*Kovačić*, 1994, стр. 281; *Everitt, et al.*, 2011, стр. 91) и крећу се у интервалу од 0 до +1, укључујући и те две вредности. Сходно наведеном, може се констатовати да кофенетички коефицијент представља меру сличности између елемената оригиналне и кофенетичке матрице одстојања (*Halkidi, et al.*, 2001, стр. 128). Вредности *CPCC*-а блиске јединици указују на чињеницу да примењени метод удруживања резултира хијерархијском структуром група која представља задовољавајуће (квалитетно) решење разматраног класификационог проблема (*Kovačić*, 1994, стр. 281; *Martinez & Martinez*, 2007, стр. 436), и обратно у случају када је $CPCC \approx 0$. Сходно наведеној интерпретацији сасвим је оправдана широка употреба кофенетичког коефицијента као критеријума за компарацију и избор „оптималног“ метода хијерархијског удруживања опсервација (*Farris*, 1969, стр. 279; *Kovačić*, 1994, стр. 282–283).

Методи хијерархијског груписања засновани на раздвајању или деоби

Класификација расположивих јединица посматрања унутар одговарајућих група применом категорије хијерархијских метода груписања заснованих на раздвајању или деоби, познатих и под називом дивизиони методи (енгл. *divisive hierarchical clustering methods*), заснована је на приступу груписању „од врха ка дну“ (енгл. *top-down clustering approach*) (*Dunham*, 2000, стр. 98; *Izenman*, 2008, стр. 411; *Vercellis*, 2009, стр. 310). Наиме, супротно поступку карактеристичном за хијерархијске методе удруживања, применом овог приступа, на првом нивоу хијерархије, свих n расположивих јединица посматрања је распоређено унутар једне заједничке групе. Полазећи од овако формираног иницијалног решења, $C^{(1)} = \{C_{g=1}\}$, којим је, према речима аутора *Hardle & Simar* (2003, стр. 308), презентована „најгрубља“ могућа (али и практично неупотребљива) подела расположивог скупа / узорка мултиваријационих опсервација, у првом кораку процеса раздвајања врши се подела иницијалне групе, величине n, на две нове, мање и међусобно најудаљеније,

групе. Поступак деобе[63] једне од расположивих група, формираних у претходном кораку, итеративно се спроводи у наредним сукцесивним корацима поступка раздвајања, све до тренутка формирања решења које се састоји од n засебних, једночланих група, облика $C^{(n)} = \{C_1, C_2,\ldots, C_{g=n}\}$. Евидентно је да, као и у случају примене агломеративних метода, процес груписања заснован на раздвајању опсервација и / или група такође резултира хијерархијски уређеном серијом могућих решења разматраног проблема, при чему се хијерархијска структура развија у супротном смеру. На Слици 2.2.6, приказана је идеја у основи хијерархијских метода раздвајања (означена стрелицом усмереном са десна на лево), односно ток процеса груписања заснованог на приступу „од корена ка листовима“ у креирању хијерархијске стабло-структуре[64]. У поређењу са методама удруживања, поред наведеног деловања у супротном смеру, методе засноване на раздвајању сматрају се, генерално, рачунски захтевнијим за спровођење (*Kaufman & Rousseeuw*, 2005, стр. 253; *Varmuza & Filzmoser*, 2009, стр. 277; *Everitt et al.*, 2011, стр. 84), будући да исте подразумевају разматрање великог броја могућих комбинација, односно расположивих опција за поделу конкретне групе на две мање, међусобно најудаљеније, (под)групе, нарочито у првим корацима поступка раздвајања.[65] *Kaufman & Rousseeuw* (2005, стр. 49; 253) и *Varmuza & Filzmoser* (2009, стр. 277) идентификују наведено „ограничење“ и комплексност поступка раздвајања, као кључни разлог (релативно) слабије заступљености ове групе хијерархијских метода у релеватној литератури, али и фокуса већине комерцијалних статистичких програмских пакета, посматрано из угла обухвата и расположивости програмских апликација за реализацију анализе груписања, искључиво на агломеративне хијерархијске методе груписања. Сходно наведеном, примена метода хијерархијског груписања на бази раздвајања је у знатно мањој мери присутна у пракси[66], генерално, наспрам прве категорије хијерархијских метода, услед чега исти нису детаљније разматрани у наставку излагања.

2.2.4.2. Методе нехијерархијског груписања објеката

Према иницијално представљеној категоризацији метода за спровођење анализе груписања, означених као класичне методе, поред хијерархијских, другу групу чине нехијерархијски методи (Слика 2.2.4). У начелу, реч је о методама поделе, односно рашчлањавања путем којих се објекти разврставају у одређени, унапред дефинисани број група g, при чему између решења груписања са g и g+1 група не постоји хијерархијска веза. Заправо, нехијерархијска процедура груписања захтева *a priori* детерминисање броја група, с тим што се током спровођења поступка груписања може вршити реалокација (прераспоређивање) расположивих јединица посматрања из једне у другу групу. Другим

[63] Детаљаније о поступку деобе, односно раздвајања опсервација унутар једне од претходно формираних група на две нове, мање групе, видети у: *Rencher* (2002, стр. 479–480), *Izenman* (2008, стр. 420); *Vercellis* (2009, стр. 310-312); *Roux* (2015); *Kaufman & Rousseeuw* (2005, стр. 253–259).

[64] Иако се формиране хијерархијске стабло-структуре могућих решења, представљене као исход примене хијерархијских метода удруживања и метода раздвајања, међусобно поклапају, важно је поновити да је реч о хипотетичком приказу, будући да се исте, у реалним околностима, углавном међусобно значајно разликују *Kaufman & Rousseeuw* (2005, стр. 44).

[65] На пример, у првом кораку процеса груписања, постоји укупно $(2^{n-1}–1)$ комбинација за поделу иницијалне групе сачињене од свих n расположивих опсервација на две мање (под)групе, које је неопходно узети у обзир и квантитативно упоредити.

[66] Детаљније о дистинктивним предностима примене дивизионих хијерархијских метода груписања и ситуацијама у којима се њихов избор сматра пожељнијом опцијом, видети у: *Rencher* (2002, стр. 479); *Kaufman & Rousseeuw* (2005, стр. 49).

речима, насупрот хијерархијском, нехијерархијско груписање представља реверзибилни процес.

Типичан поступак нехијерархијског груписања, упркос различитим алгоритамским варијацијама за разврставање јединица посматрања, обухвата неколико стандардних корака (*Timm*, 2002, стр. 540; *Sharma*, 1996, стр. 202). Након специфицирања броја група, иницијација (покретање) поступка груписања започиње одређивањем g p-димензионих центара група (на основу коришћења матрице оригиналних података), који представљају почетне тачке око којих се формирају групе. У основи, за одређивање иницијалних центара група постоје бројне методе, при чему свака од њих подразумева испуњење захтева да се изабране почетне тачке међусобно довољно разликују. У даљем току имплементације нехијерархијске процедуре, након одређивања центара груписања, следи одређивање одстојања (најчешће квадрата Еуклидског одстојања) између сваког објекта и сваког центра групе, тако да се сви објекти чије је одстојање од центра конкретне групе мање од унапред дефинисаног критеријума класификују као елементи те (најближе) групе. Као што је већ истакнуто, сходно изабраном критеријуму оптималности за формирање група, у спровођењу поступка груписања може доћи до реалокације јединица посматрања по групама како би се остварило побољшање у погледу интерне хомогености и екстерне хетерогености група. Свака реалокација је праћена итеративним одређивањем нових центара група и одређивањем одстојања објеката од нових тачака груписања. Наведени поступак се зауставља када више није могуће постићи нова побољшања са становишта критеријума за формирање група. При томе, за разлику од хијерархијског груписања које се одликује читавим низом могућих решења разматраног проблема, нехијерархијско груписање резултира само једним решењем сачињеним од *a priori* специфицираног (претпостављеног) броја група g. Услед наведеног, насупрот хијерархијским, експлоративним методама, нехијерархијске методе се називају конфирматорним методама (*Timm*, 2002, стр. 533).

Већина метода која припада групи / фамилији метода нехијерархијског груписања међусобно се разликују у погледу начина одређивања центара група и процедура за реалокацију јединица посматрања. (Детаљније информације о наведеним аспектима нехијерархијског груписања видети у *Sharma*, 1996, стр. 202–203) Најчешће коришћена нехијерархијска метода је метода k-средина (енгл. *k-means method*) [67], за коју су карактеристична следећа два својства: ▸ прво, након одређивања почетних тачака груписања и формирања иницијалних група, у наредним итерацијама, нове тачке груписања се одређују као центроиди група, и ▸ друго, процедура груписања се зауставља када се даљом реалокацијом објеката не постиже смањење вредности суме квадрата одступања опсервација од средине на нивоу свих група обухваћених конкретном партицијом. Заправо, ова метода захтева да од стране корисника буду специфицирана три параметра: број група, иницијални центри група и мера одстојања (*Jain*, 2010, стр. 652).

Решења добијена методом k-средина и, генерално, применом других нехијерархијских метода су веома осетљива на иницијалну поделу опсервација. У том контексту, један од често практикованих начина избора улазних параметара за иницијацију нехијерархијске процедуре обухвата, најпре, спровођење хијерархијског

[67] Сходно симболичкој нотацији у овој докторској дисертацији, метод k-средина се може назвати метод g-средина.

груписања за детерминисање броја група и полазних центара група, а затим, на тако припремљене податке, укључује и примену нехијерархијског груписања. Ово комбиновано коришћење усмерено на добијање „квалитетнијих“ резултата груписања, јасно указује да се хијерархијске и нехијерархијске методе пре могу сагледати као комплементарне него конкурентске (*Sharma*, 1996, стр. 211), при чему се недостаци једне групе компензују предностима друге групе метода.

Генерално, истраживачи и практичари, имајући у виду дефинисане циљеве истраживања и каракеристике различитих метода груписања, трагају за решењем које се у конкретној проблемској ситуацији може сматрати „најбољим“, односно оптималним са становишта репрезентовања структуре присутне у подацима. Сходно томе, у циљу обезбеђења најбољег избора, важно питање при спровођењу анализе груписања односи се на оцену различитих аспекта квалитета добијених решења уз коришћење адекватних критеријума, мера и поступака, о чему ће бити речи у наставку текста дисертације.

2.2.5. Евалуација квалитета резултата груписања и избор оптималног броја група

Последњи, а уједно и најкритичнији, корак у анализи груписања односи се на евалуацију издвојене серије могућих решења, настале као резултат примене хијерархијских поступака груписања, у циљу проналажења конкретне поделе (то јест, „оптималног“ броја група, g) за коју се сматра да најбоље репрезентује инхерентну структуру података у расположивом скупу / узорку (*Halkidi, et al.*, 2002, стр. 122; *Varmuza & Filzmoser*, 2009, стр. 284; *Hair et al.*, 2010, стр. 509). Заправо, фундаментална идеја ове финалне активности, усмерене на обезбеђивање практичне употребљивости резултата анализе груписања (*Liu et al.*, 2010, стр. 911), може се дефинисати, у случају примене једне од метода удруживања, на следећи начин: од укупно (n) идентификованих могућих подела n анализираних опсервација, у ознаци $C^{(i)}=\{C_1, C_2,\ldots, C_g\}$ за $i = 1, 2,\ldots, n$, појединачно сачињених од g група C_k (за $k = 1, 2,.., g$), при чему важи тенденција ($C^{(1)}\rightarrow g=n$), ($C^{(2)}\rightarrow g=n-1$),..., ($C^{(n)}\rightarrow g=1$), неопходно је извршити селекцију једног решења $C^{(i)}$, за које се, у поређењу са преосталим решењима и њима припадајућим вредностима параметра g, постиже најбољи однос интерне-хомогености и екстерне хетерогености на нивоу g група. Иако, генерално, не постоји стандардизована (у потпуности објективна) процедура за избор оптималног броја група заснована на теорији статистичког закључивања (*Hair et al.*, 2010, стр. 509), у литератури је предложен велики број критеријума намењених минимизирању субјективности истраживача при решавању овог проблема.[68] Будући да њихова примена омогућава идентификовање оптималне поделе у оквиру развијене хијерархијске структуре могућих решења[69], у литератури се за означавање ових

[68] Веома исцрпна и детаљна компаративна студија перформанси 30 различитих критеријума (статистичких процедура) за избор „оптималног“ броја група спроведена је од стране аутора *Milligan & Cooper* (1985). Поред наведеног извора, који се, према мишљењу бројних аутора (укључујући и аутора ове дисертације), неприкосновено издваја по својој оригиналности, обухвату и информативности, у литератури новијег датума, из угла остваренog доприноса у разматрању анализиране проблематике, као веома корисне, могу се издвојити, на пример, и публикације следећих аутора: *Kovačić* (1994, стр. 283–289); *Sharma* (1996, стр. 197–202); *Halkidi et al.* (2001; 2002); *Timm* (2002, стр. 531–533); *Kaufman & Rousseeuw* (2005, стр. 83–88); *Vercellis* (2009, стр. 312-315); *Charrad, et al.* (2010; 2014); *Liu, et al.* (2010); *Desgraupes, et al.* (2013); *Arbelaitz, et al.* (2013).

[69] Иако су критеријуми за избор оптималног броја група иницијално намењени евалуацији резултата добијених путем метода хијерархијског груписања, уз одговарајућа прилагођавања и модификације, њихова примена такође може бити проширена и на домен нехијерархијских процедура (*Milligan & Cooper*, 1985, стр. 159; 162).

критеријума углавном користе следећи алтернативни називи: ▸ правила за заустављање (енгл. *stopping rules*) (*Milligan & Cooper*, 1985, стр. 159; *Hair et al.*, 2010, стр. 509); ▸ критеријуми оптималности (*Kovačić*, 1994, стр. 283); ▸ индекси за оцену валидности (ваљаности) груписања (енгл. *cluster validity indices*) (*Halkidi et al.*, 2001, стр. 124; *Bouguessa et al.*, 2006; *Rendon et al.*, 2011); ▸ методи за оцену квалитета резултата груписања и избор адекватног броја група (*Martinez & Martinez*, 2007, стр. 458; *Tuffery*, 2011, стр. 244). Описујући основне одлике ових критеријума, *Hair, et al.* (2010, стр. 509) истичу да је реч о *ad hoc* процедурама хеуристичког карактера које најчешће захтевају ручно спровођење, често прилично комплексних, рачунских поступака, с обзиром да њихово израчунавање није у значајној мери подржано статистичким софтверским решењима отвореног кода. Примена ових критеријума, као својеврсних статистичких мера одређеног аспекта „квалитета" појединачних подела у оквиру развијене хијерархијске структуре опсервација, не зависи од избора конкретне методе хијерархијског груписања (*Milligan & Cooper*, 1985, стр. 162). Такође, услед присутних разлика у погледу апликативних одлика расположивих критеријума оптималности разликују се, генерално, два (алтернативна) приступа у идентификовању конкретне поделе која се може сматрати „оптималном", из угла обухваћеног броја група и/или њихове структуре. Према првом приступу, избор оптималне вредности параметра *g*, условљен је идентификовањем поделе за коју коришћени критеријум постиже максималну или пак минималну вредност (у зависности од аспекта квалитета решења чије мерење обезбеђује) у поређењу са кореспондентним вредностима на нивоу преосталих решења проблема груписања (*Everitt et al.*, 2011, стр. 111; *Halkidi et al.*, 2001, стр. 129). С друге стране, уколико се кретање вредности одабраног критеријума одликује растућом или опадајућом, јасно уочљивом тенденцијом током процеса хијерархијског груписања, тада се изражене (и нагле) промене у висини израчунатих вредности на нивоу појединачних решења користе као индикатори тренутка у којем је наступило удруживање две прилично различите групе, а „оптималним" решењем проглашава се подела опсервација која је остварена на нивоу хијерархијске структуре који је непосредно претходио забележеној промени (*Halkidi et al.*, 2001, стр. 129; *Everitt et al.*, 2011, стр. 126). Ради лакшег идентификовања наведених критичних тренутака у процесу груписања, поступак примене ових критеријума углавном је употпуњен конструкцијом графичких приказа вредности изабраних критеријума наспрам кореспондентних вредности параметра *g* (*Tuffery*, 2011, стр. 244–247). У складу са методолошким потребама спроведеног емпиријског истраживања у дисертацији, даљим излагањем обухваћен је приказ најчешће коришћених критеријума оптималности.

• *Критеријум заснован на вредностима одстојања између група које се удружују*

Најједноставнији приступ проблему избора броја група заснован је на праћењу вредности мере одстојања при којој је извршено удруживање одређеног пара група у сваком кораку процеса хијерархијског груписања (*Kovačić*, 1994, стр. 283). Реализацијом појединачних корака у процесу удруживања (од 1 до n–1), број група се постепено смањује (са $g = n$ на $g = 1$), а вредност овог критеријума се сукцесивно повећава, будући да се у првим корацима спајају међусобно више сличне групе, док је у каснијим корацима присутна (по правилу) већа удаљеност између група актера удруживања. Према овом критеријуму, као оптимално решење проблема груписања издваја се број група који је

остварен непосредно пре корака у којем је забележен нагли скок вредности одстојања између удружених група. Често су у употреби и одређене модификације овог критеријума, засноване на израчунавању и праћењу апсолутних и / или релативних износа промена вредности мере одстојања при којима је извршено удруживање група у оквиру два узастопна корака процеса хијерархијског груписања (*Kovačić*, 1994, стр. 284; *Hair et al.*, 2010, стр. 510).

• *Calinski-Harabasz*-ов *критеријум*

Реч је о критеријуму оптималности базираном на израчунавању односа показатеља екстерне-хетерогености (односно, варијабилитета између g група) и показатеља интерне-хомогености (то јест, варијабилитета унутар g група), утврђених за свако појединачно решење у оквиру формиране хијерархијске структуре, путем израза (*Calinski* & *Harabasz*, 1974, стр. 10):

$$F_g^* = \frac{\frac{tr(\mathbf{B}_g)}{g-1}}{\frac{tr(\mathbf{W}_g)}{n-g}} = \frac{\frac{\left(tr(\mathbf{T}_g) - \sum_{k=1}^{g} SSE_k\right)}{g-1}}{\frac{\sum_{k=1}^{g} SSE_k}{n-g}} = \frac{\overbrace{\left(\sum_{j=1}^{p}\sum_{k=1}^{g} n_k(\bar{x}_{jk} - \bar{x}_j)^2\right)}^{\text{сума квадрата одступања између } g \text{ група}} / (g-1)}{\underbrace{\left(\sum_{j=1}^{p}\sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_{ijk} - \bar{x}_{jk})^2\right)}_{\text{сума квадрата одступања унутар } g \text{ група}} / (n-g)}, \text{ за } g=n, n-1,...,1. \quad (2.2.22)$$

У датом изразу[70], укупан и број опсервација унутар k-те групе C_k (k=1,2,.., g) на нивоу конкретне поделе $C^{(i)}$ (i=1,2,…, n) сачињене од g група, представљени су симболима n и n_k, респективно, док $tr(\mathbf{B}_g)$ и $tr(\mathbf{W}_g)$ означавају траг (енгл. *trace*), односно збир елемената на главној дијагонали, матрице суме квадрата и узајамних производа одступања између (у ознаци **B**) и унутар група (у ознаци **W**), на нивоу сваког појединачног решења сачињеног од g група, при чему важи релација $tr(\mathbf{T}_g) = tr(\mathbf{B}_g) + tr(\mathbf{W}_g)$ (*Kovačić*, 1994, стр. 285).

Кретање вредности F_g^* за појединачна решења хијерархијског процеса удруживања у директној је корелацији са кретањем броја група g. Наиме, из корака у корак, вредност *Calinski-Harabasz*-овог критеријума се постепено смањује, будући да сукцесивно смањење броја формираних група условљава опадање вредности бројиоца, праћено постепеним порастом вредности имениоца у изразу (2.2.22). Изразити пад вредности F_g^*, забележен на конкретном нивоу хијерархије, указује на удруживање међусобно прилично удаљених група. Вредност параметра g у основи поделе која је непосредно претходила уоченој „аномалији“ у процесу удруживања, третира се као оптимална (*Timm*, 2002, стр. 532).[71]

• *Критеријум заснован на вредностима коефицијента* R_g^2

Дефинисан као количник суме квадрата одступања између g група (у ознаци $tr(\mathbf{B}_g)$) и укупне суме квадрата одступања (у ознаци $tr(\mathbf{T})$), на нивоу сваког појединачног решења проблема груписања, R_g^2 представља меру степена међусобне раздвојености (екстерне хетерогености) формираних група на датом нивоу хијерархије (*Sharma*, 1996, стр. 198).

---

[70] Због аналогије са статистиком F теста у анализи варијансе, у литератури се, за означавање овог критеријума, најчешће користи назив псеудо-F мера (*Tuffery*, 2011, стр. 246; *Timm*, 2002, стр. 532; *Ковачић*, 1994, стр. 285).

[71] При поређењу два решења са истим бројем група, добијених путем хијерархијског и нехијерархијског груписања, тада се као оптимално, са аспекта структуре формираних група, издваја решење које се одликује већом вредношћу F_g^*.

$$R_g^2 = \frac{tr(\mathbf{T}) - \sum_{k=1}^{g} SSE_k}{tr(\mathbf{T})} = \frac{tr(\mathbf{B})}{tr(\mathbf{T})} = \frac{\overbrace{\left(\sum_{j=1}^{p}\sum_{k=1}^{g} n_k (\bar{x}_{jk} - \bar{x}_j)^2\right)}^{\text{сума квадрата одступања између } g \text{ група}}}{\underbrace{\left(\sum_{j=1}^{p}\sum_{k=1}^{g}\sum_{i=1}^{n_k} (x_{ijk} - \bar{x}_j)^2\right)}_{\text{укупна сума квадрата одступања}}}, \text{ за } g = n, n-1, \ldots, 2, 1. \quad (2.2.23)$$

Наиме, R_g^2 показује пропорцију укупног варијабилитета између опсервација која се може приписати присутном варијабилитету између g група (*Kovačić*, 1994, стр. 286). Током поступка хијерархијског удруживања, вредност R_g^2 сукцесивно опада са смањењем броја група g, сигнализирајући постепено погоршање „квалитета“ добијених решења, будући да се из корака у корак сума квадрата одступања између група смањује (то јест, смањује се екстерна-хетерогеност расположивих група) као последица спајања у већој мери међусобно различитих група. Наведени коефицијент може узети вредности у интервалу од 1 (за $g = n \rightarrow \sum_{k=1}^{g} SSE_k = 0$) до 0 (за $g = 1 \rightarrow tr(\mathbf{T}) - \sum_{k=1}^{g} SSE_k = 0$), односно $R_g^2 \in [1, 0]$. За оптималну поделу узима се број група g у оквиру решења оствареног непосредно пре корака у којем је забележен нагли пад вредности R_g^2 (*Timm*, 2002, стр. 532; *Tuffery*, 2011, стр. 245).[72]

• *Критеријум заснован на вредностима семипарцијалног* R_g^2 *коефицијента* (ΔR_g^2)

Овим показатељем мери се смањење суме квадрата одступања између g група $tr(\mathbf{B}_g)$ (тј. смањење екстерне-хетерогености), или алтернативно[73], повећање суме квадрата одступања унутар g група $tr(\mathbf{W}_g)$ (тј. смањење интерне-хомогености) остварено између два сукцесивна корака у процесу хијерархијског груписања, као последица удруживања две групе у ранијем кораку. Вредност ΔR_g^2 заправо представља (негативни) прираштај вредности коефицијента R_g^2, односно (позитивни) прираштај релативног односа $tr(\mathbf{W}_g)/tr(\mathbf{T}_g)$, забележен при преласку са решења које обухвата g+1 група на решење сачињено од g група, симболички:

$$\Delta R_g^2 = \left| R_g^2 - R_{(g+1)}^2 \right|, \text{ алтернативно } \Delta R_g^2 = \left| \frac{\left(\sum_{k=1}^{g} SSE_k\right) - \left(\sum_{k=1}^{g+1} SSE_k\right)}{\text{tr}(\mathbf{T})} \right|, \text{ за } g = n, n-1, \ldots, 2, 1. \quad (2.2.24)$$

Смањивање броја група g током процеса удруживања, по правилу, праћено је повећањем апсолутне вредности показатеља ΔR_g^2. Већа апсолутна вредност ΔR_g^2 указује да се догодио већи „губитак“ како екстерне-хетерогености тако и интерне-хомогености и обратно (*Sharma*, 1996, стр. 200). Нагли пораст апсолутне вредности овог показатеља сугерише, као оптимално решење, избор броја група који је непосредно претходио уоченој драстичној промени.

---

[72] Уколико се путем овог критеријума пореде решења хијерархијског и нехијерархијског груписања, са истом вредношћу параметра g, тада се као оптимално, у погледу састава група, издваја решење које се одликује већом вредношћу R_g^2.

[73] Термин „алтернативно“ употребљен је на основу релације $tr(\mathbf{T}_g) = tr(\mathbf{B}_g) + tr(\mathbf{W}_g)$ и чињенице да укупна сума квадрата одступања $tr(\mathbf{T}_g)$ остаје непромењена током поступка удруживања, тако да повећање вредности једног од сабирака неминовно подразумева пропорционално (сразмерно) смањење вредности другог сабирка у једначини, и обратно. Наиме, за посматрано решење, већа разлика између формираних група имплицира њихову већу интерну хомогеност и обратно.

• *Критеријум заснован на вредностима коефицијента кохезије*, $coh(C^{(i)})_g$

Заснован на мерењу степена блискости опсервација унутар формираних група, коефицијент кохезије (енгл. *cohesion coefficient*) представља један од најчешће коришћених статистичких показатеља компактности (интерне-хомогености) група обухваћених конкретним решењем и, може се дефинисати на следећи начин (*Vercellis*, 2009, стр. 313):

$$coh(C^{(i)})_g = \sum_{k=1}^{g} coh(C_k), \text{ за } i = 1, 2,\ldots, n \text{ и } g = n, n-1,..., 1. \quad (2.2.25)$$

У датом изразу, $coh(C^{(i)})_g$ представља меру (укупне) кохезије остварене на нивоу посматране поделе $C^{(i)}=\{C_1, C_2,\ldots,C_g\}$, за $i=1, 2,\ldots,n$, док $coh(C_k)$ означава меру интерне-хомогености на нивоу појединачне k-те групе (за $k=1, 2,\ldots, g$), утврђене коришћењем једног од бројних, алтернативних, начина и то као: ▸ збир квадрата одстојања свих опсервација унутар k-те групе од припадајућег центроида[74] (*Chaimontree et al.*, 2010, стр. 51); ▸ збир одстојања између свих парова опсервација унутар исте, k-те групе (*Vercellis*, 2009, стр. 313); ▸ максимално или просечно одстојање између свих појединачних парова опсервација унутар конкретне групе (*Liu et al.*, 2010, стр. 911). Генерално, нижа вредност показатеља $coh(C^{(i)})_g$ сугерише бољу (укупну) компактност формираних група на датом нивоу хијерархије и обратно.[75] Током процеса хијерархијског удруживања, вредност овог показатеља, по правилу, се повећава са смањењем броја група, а нагли скок вредности сугерише да је наступило спајање међусобно значајно различитих група. Консеквентно, број група који је остварен непосредно пре забележене велике промене вредности $coh(C^{(i)})_g$ представља адекватан избор поделе.

• *Критеријум заснован на вредности коефицијента сепарације*, $sep(C^{(i)})_g$

Коефицијент сепарације (енгл. *separation* / *isolation coefficient*) представља меру екстерне-хетерогености, односно степена у којем су формиране групе, у оквиру сваког појединачног решења, међусобно „добро“ раздвојене. Уколико се симболом $sep(C_h,C_q)$ означи сепарација између било које две групе у саставу конкретне, i-те поделе $C^{(i)}=\{C_1, C_2,\ldots,C_g\}$ за $i=1, 2,\ldots, n$, тада се коефицијент укупне сепарације (у ознаци $sep(C^{(i)})_g$) може представити путем следећег израза (*Vercellis*, 2009, стр. 313):

$$sep(C^{(i)})_g = \sum_{\forall(h,q)=1}^{g(g-1)/2} sep(C_h, C_q), \text{ за } \forall(C_h,C_q) \in C^{(i)} \text{ где је } h \neq q \text{ и } g=n, n-1,...,1, \quad (2.2.26)$$

Степен раздвојености две појединачне групе ($sep(C_h,C_q)$) може се утврдити путем једног од следећих приступа, односно израчунавањем: ▸ одстојања између центроида на нивоу група које се пореде (*Liu et al.*, 2010, стр. 912); ▸ минималног одстојања од свих одстојања између појединачних парова опсервација које припадају двема различитим групама (*Liu et al.*, 2010, стр. 912); ▸ збира одстојања између свих (комбинованих) парова опсервација садржаних унутар посматране две, h-те и q-те, групе (*Vercellis*, 2009, стр. 313).

---

[74] Уколико се користи овај начин за одређивање $coh(C_k)$ тада ће вредност $coh(C^{(i)})_g$ бити једнака трагу матрице $\mathbf{W}$, $tr(\mathbf{W}_g)$.

[75] Такође, решење добијено применом једне методе груписања сматра се бољим од истоврсног (са аспекта броја обухваћених група) решења добијеног применом друге методе или поступка груписања, уколико се одликује мањом вредношћу укупног коефицијента кохезије, $coh(C^{(i)})_g$ (*Liu et al.*, 2010, стр. 911; *Vercellis*, 2009, стр. 313).

Алтернативно, уколико се као показатељ изолованости појединачних група (у ознаци $sep(C_k)$) користи збир квадрата одстојања између припадајућег центроида k-те групе величине n_k (у ознаци $\bar{\mathbf{x}}_k$, за k = 1, 2,..., g) и општег центроида (у ознаци $\bar{\mathbf{x}}$) утврђеног за узорак величине n, (*Chaimontree et al.*, 2010, стр. 51), тада је вредност коефицијента укупне сепарације ($sep(C^{(i)})_g$) еквивалентна трагу матрице **B** (у ознаци $tr(\mathbf{B}_g)$), односно симболички:

$$sep(C^{(i)})_g = \sum_{k=1}^{g} sep(C_k) = \sum_{k=1}^{g} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^2 = \sum_{k=1}^{g}\sum_{j=1}^{p} n_k(\bar{x}_{jk} - \bar{x}_j)^2 = tr(\mathbf{B}_g), \text{ за } g=n, n-1,...,1. \qquad (2.2.27)$$

Већа вредност показатеља $sep(C^{(i)})_g$ сугерише бољу (укупну) раздвојеност појединачних група на датом нивоу хијерархије и обратно.[76] Током процеса хијерархијског удруживања, вредност овог показатеља се, по правилу, смањује паралелено са смањењем броја група g, а нагли пад вредности сугерише да је наступило спајање међусобно значајно различитих група. Консеквентно, број група који је остварен непосредно пре забележене велике промене вредности $sep(C^{(i)})_g$ представља оптималан избор поделе расположивих опсервација.

• *Критеријум заснован на вредностима коефицијента силуете*

Предложен од стране *Rousseeuw*-а (1987), коефицијент силуете (енгл. *silhouette coefficient*) представља широко коришћену статистичку меру намењену свеобухватној евалуацији квалитета решења процеса груписања. Обезбеђујући симултано сагледавање и праћење промена интерне-хомогености и екстерне-хетерогености формираних група у саставу појединачних решења, овај коефицијент представља погодно аналитичко средство за давање одговора на следећа, у домену анализе груписања од кључне важности, истраживачка питања (*Rousseeuw*, 1987, стр. 55): ▸ Какав је квалитет (посматрано из угла односа унутаргрупних и међугрупних одстојања) појединачних група у оквиру конкретне поделе?; ▸ Појединачно посматрано, које опсервације се могу сматрати исправно, или (евентуално) погрешно класификованим?; ▸ Да ли постоје опсервације које су подједнако удаљене у односу на две различите групе у оквиру истог решења и, ако постоје, које су то опсервације?; ▸ Какав је квалитет појединачних решења која се одликују различитим вредностима g?; ▸ Које се појединачно решење, односно број група g, може сматрати оптималним избором?; ▸ Које од два решења са истим бројем група, изведених применом хијерархијских и нехијерархијских метода, представља адекватнију поделу опсервација, из угла структуре (односно састава) обухваћених група?. Важна предност коефицијента силуете, која директно произилази из наведених истраживачких питања, у односу на претходно елабориране критеријуме, огледа се у могућности израчунавања његових вредности за: ▸ ниво појединачних опсевација, $\mathbf{x}_i$ (за i =1, 2,…, n), ▸ ниво појединачних група, C_k (за k=1, 2,…, g) у саставу конкретне поделе, и ▸ ниво конкретног решења сачињеног од g група, $C^{(i)}=\{C_1, C_2,\ldots,C_g\}$ (i=2, 3,…,n–1).

Формула за израчунавање вредности коефицијента силуете за (произвољно одабрану) i-ту опсервацију унутар (хипотетичке) групе C_i, гласи (*Rousseeuw*, 1987, стр. 56):

[76] Такође, решење добијено применом једне методе груписања сматра се бољим од истоврсног (са аспекта броја обухваћених група) решења добијеног применом друге методе или поступка груписања, уколико се одликује већом вредношћу укупног коефицијента сепарације, $sep(C^{(i)})_g$ (*Chaimontree et al.*, 2010, стр. 5; *Vercellis*, 2009, стр. 313).

$$silh(\mathbf{x}_i)^g = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}, \text{ за } i = 1, 2,\ldots, n \text{ и } g = n\text{–}1,..., 2. \quad (2.2.28)$$

У представљеном изразу употребљени симболи означавају: ▸ $a(\mathbf{x}_i)$ – просек вредности мере одстојања (најчешће, Еуклидског одстојања) утврђених између i-те опсервације и свих осталих опсервација у саставу (исте) групе C_i. ▸ $b(\mathbf{x}_i)$ – минимум засебно израчунатих аритметичких средина вредности изабране мере одстојања између h-те опсервације ($\mathbf{x}_i \in C_i$) и свих појединачних опсервација у саставу сваке од преосталих група обухваћених конкретним решењем, сачињеним од укупно g група, на датом нивоу хијерархије. Прецизније, за сваку од преосталих (g–1) група у оквиру разматраног решења, израчунава се просечна вредност одстојања опсервације $\mathbf{x}_i$ од појединачних опсервација у саставу сваке од тих група засебно. Минимална вредност, од укупно утврђених (g–1), просека заправо представља вредност $b(\mathbf{x}_i)$, односно $b(\mathbf{x}_i) = min[\bar{d}(\mathbf{x}_i, C_k)]$, за $k = 1, 2,\ldots, g$, уз услов да $C_k \neq C_i$. Група са којом $\mathbf{x}_i$ остварује најмање просечно одстојање назива се „група-сусед“ те опсервације. ▸ $max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}$ – већа вредност од претходно дефинисане две вредности.

Вредност коефицијента силуете на нивоу појединачних опсервација креће се у интервалу од $-1 \leq silh(\mathbf{x}_i)^g \leq 1$, при чему важе следеће релације (*Rousseeuw*, 1987, стр. 56):

$$s(\mathbf{x}_i)^g = \begin{cases} > 0, \text{ уколико је } a(\mathbf{x}_i) < b(\mathbf{x}_i) \\ = 0, \text{ уколико је } a(\mathbf{x}_i) = b(\mathbf{x}_i) \\ < 0, \text{ уколико је } a(\mathbf{x}_i) > b(\mathbf{x}_i) \end{cases}. \quad (2.2.29)$$

Позитивна вредност $silh(\mathbf{x}_i)^g$ указује да је просечно одстојање конкретне опсервације од осталих чланова исте групе мања у односу на просечну удаљеност од опсервација распоређених унутар свих преосталих појединачних група, чиме се потврђује њена исправна алокација. Консеквентно, што је вредност $silh(\mathbf{x}_i)^g$ ближа јединици, јачи су и „докази“ у прилог изнетог закључка. Негативна вредност $silh(\mathbf{x}_i)^g$, с друге стране, условљена обрнутим односом унутаргрупних и међугрупних (просечних) одстојања упућује на закључак другачије садржине. Наиме, што је вредност $silh(\mathbf{x}_i)^g$ ближа –1, јачи су и докази у прилог тврдње о погрешној класификацији опсервације. У таквим околностима, на основу релације $silh(\mathbf{x}_i)^g < 0 \rightarrow a(\mathbf{x}_i) < b(\mathbf{x}_i)$, адекватнија солуција подразумева распоређивање дате опсервације унутар „групе-суседа“, будући да је истој, у просеку, ближа него групи унутар које је иницијално распоређена. Коначно, вредност $silh(\mathbf{x}_i)^g = 0$ интерпретира се као гранични случај. Заправо, у том случају није у потпуности јасно и очигледно да ли дату опсервацију задржати у оквиру исте, или пак распоредити у састав њој најближе групе, будући да је просечна удаљеност опсервације унутар припадајуће групе приближно једнака просечној удаљености од опсервација у саставу „групе-суседа“. Наведена тумачења могу бити од користи при анализирању уочених разлика у структури група на нивоу два различита решења, добијена применом неке од хијерархијских и нехијерархијских метода груписања.

Вредност коефицијента силуете за сваку појединачну, k-ту, групу C_k (за k=1,2,…,g), у саставу конкретне поделе, $C^{(i)}=\{C_1, C_2,\ldots,C_g\}$, израчунава се као просек вредности коефицијената на нивоу свих појединачних опсервација распоређених унутар исте групе (енгл. *average silhouette coefficient*), односно (*Rousseeuw*, 1987, стр. 59):

$$\overline{silh}(C_k)^g = \frac{\sum_{i=1}^{n_k} s(\mathbf{x}_i)^g}{n_k}, \text{ за } k = 1, 2,\ldots, g; \text{ и } g = n\text{–}1,..., 2, \quad (2.2.30)$$

где симбол n_k означава укупан број опсервација унутар *k*-те групе, односно њену величину. Позитивна вредност коефицијента силуете појединачних група представља пожељнији исход наспрам негативног просека индивидуалних коефцијената на нивоу опсервација у њиховом саставу, из угла остваренога односа показатеља интерне хомогености (то јест, кохезије или компактности) и екстерне хетерогености (то јест, сепарације / раздвојености од других група). Другим речима, што је вредност $\overline{silh}(C_k)^g$ ближа 1, *k*-та група се сматра компактнијом и боље раздвојеном од осталих група у оквиру датог решења и обратно.

Коначно, вредност коефицијента силуете за ниво укупног решења (енгл. *overall average silhouette coefficient*), односно конкретне поделе сачињене од *g* група, $C^{(i)}=\{C_1, C_2,\ldots,C_g\}$ (за $i = 2, 3,\ldots, n\text{–}1$) [77], утврђује се као аритметичка средина вредности коефицијента утврђених за сваку од *n* појединачних опсервација, односно (*Rousseeuw*, 1987, стр. 59):

$$\overline{\overline{silh}}(C^{(i)})^g = \frac{\sum_{i=1}^{n} s(\mathbf{x}_i)^g}{n}, \text{ за } i = 2, 3,\ldots, n\text{–}1, \text{ односно } g = n\text{–}1,..., 2. \quad (2.2.31)$$

Уважавајући смисао и суштину претходних тумачења вредности индивидуалних и групних коефцијената, за потребе интерпретације значења добијених вредности коефицијента силуете на нивоу појединачних укупних решења, аутори *Kaufman & Rousseeuw* (1990, цитирано у: *Izenman*, 2008, стр. 408) предложили су следеће смернице:

$$\begin{aligned}
&0,70 < \overline{\overline{silh}}(C^{(i)})^g \le 1,00 \rightarrow \text{формирана (укупна) структура група поседује висок квалитет}\\
&0,50 < \overline{\overline{silh}}(C^{(i)})^g \le 0,70 \rightarrow \text{формирана (укупна) структура група је умереног квалитета}\\
&0,25 < \overline{\overline{silh}}(C^{(i)})^g \le 0,50 \rightarrow \text{формирана (укупна) структура група је слабог квалитета}\\
&0,00 \le \overline{\overline{silh}}(C^{(i)})^g \le 0,25 \rightarrow \text{нема довољно доказа да је квалитет датог решења значајан}
\end{aligned} \quad . (2.2.32)$$

Генерално, доношење одлуке о оптималној подели опсервација у погледу обухваћеног броја група *g*, заснива се на поређењу вредности коефицијента силуете утврђених на нивоу појединачних решења $C^{(i)}$ и идентификовању конкретне поделе за коју је постигнута највећа вредност показатеља $\overline{\overline{silh}}(C^{(i)})^g$ (*Rousseeuw*, 1987, стр. 59). Међутим, поред овог формалног критеријума, независно од тога да ли је реч о компарацији читаве серије решења добијених као резултат хијерархијског груписања, или пак појединачних решења хијерархијског и нехијерархијског груписања са истим бројем (другачије структурираних) група, неопходно је узети у обзир и запажања изведена на основу анализе како вредности коефицијената силуете на нивоу појединачних опсервација и група, тако и тенденција у кретању вредности укупног коефицијента током процеса груписања и смањења броја група.

[77] Иако поступак хијерархијског груписања резултира серијом од укупно *n* могућих подела $C^{(i)}$, израчунавање вредности коефицијента силуете нема смисла у случају када је решењем обухваћено $g = n$ и / или $g = 1$ група, будући да тада тражена вредност износи 0, односно –1, респективно. Сходно наведеном, у наставку излагања о овом коефицијенту разматрају се само преосталих n–2 решења, односно подела $C^{(i)}$, за $i = 2, 3,\ldots, n\text{–}2, n\text{–}1$.

Извођење закључака у погледу „квалитета“ појединачних решења проблема груписања, применом наредна три критеријума, заснива се на поређењу $[n(n–1)/2]$ парова елемената оригиналне матрице одстојања и кореспондентних елемената изведене симетричне $(n×n)$ бинарне матрице, чије вредности (0 или 1), додељене сваком пару појединачних опсервација, указују да ли су или не посматране две опсервације, на датом нивоу хијерархијске структуре, распоређене унутар исте групе, или пак, припадају различитим групама, респективно. Наиме, основни принцип на којем почива њихова примена и итерпретација је да опсервације које су распоређене унутар исте групе у било ком тренутку, треба да се одликују мањим међусобним одстојањем у односу на опсервације унутар различитих група (*Kovačić*, 1994, стр. 286).

• *Критеријум заснован на вредностима бисеријалног коефицијента корелације,* $r_{pb}^{(g)}$

Коефицијент корелације између $[n(n–1)/2]$ парова конституисаних од елемената оригиналне матрице одстојања **D** (односно, вредности мере одстојања између појединачних парова опсервација) и одговарајућих вредности (0 или 1) додељених кореспондентним паровима опсервација (као елемената изведене бинарне матрице) назива се бисеријални[78] коефицијент корелације, у ознаци $r_{pb}^{(g)}$. Вредност овог показатеља израчунава се коришћењем следећег (алтернативног) израза (*Desgraupes, et al.*, 2013, стр. 14; *Charrad, et al.*, 2014):

$$r_{pb}^{g} = \frac{(\overline{d}_b - \overline{d}_w)}{s_d}\sqrt{\frac{n_b n_w}{n_d^2}}\text{, за } g = n, n–1,..., 2, 1, \qquad (2.2.33)$$

где индекси w (енгл. *within*) и b (енгл. *between*) представљају ознаке група парова опсервација којима је додељена вредност 0 и вредност 1 респективно, у складу са претходно наведеним правилом бинарног кодирања сваког од $[n(n–1)/2]$ појединачних парова. Симболима $\overline{d_w}$ и $\overline{d_b}$ означен је просек оригиналних одстојања на нивоу n_w и n_b парова опсервација алоцираних унутар наведене две групе, респективно. Укупан број парова опсервација, у ознаци $n_d = n_w+n_b$, износи $[n(n–1)/2]$, где n представља укупан број јединица посматрања обухваћених анализом. Стандардна девијација оригиналних одстојања n_d парова опсервација од просечне вредности кореспондентних одстојања представљена је симболом s_d. Вредности овог показатеља одговарају апсолутним вредностима коефицијента корелације и налазе се у интервалу $0 \leq r_{pb}^{(g)} \leq +1$, при чему вредности блиске 1 указују на присуство високог степена квантитативног слагања величина оригиналних одстојања и додељених бинарних вредности (односно, припадности појединачних опсервација групама у оквиру конкретног решења) и обрнуто за $r_{pb}^{(g)} \approx 0$. Наиме, висока вредност $r_{pb}^{(g)}$ сугерише да парови опсервација који су кодирани са 1 теже да имају високу вредност одстојања, а парови кодирани вредношћу 0 теже да имају малу вредност одстојања (*Kovačić*, 1994, стр. 287). Консеквентно, максимална вредност $r_{pb}^{(g)}$ се користи као индикатор оптималног броја група g (*Milligan & Cooper*, 1985, стр. 164).

[78] Назив „бисеријални“ коефицијент користи се за означавање корелационе мере која се утврђује између парова вредности две променљиве, од којих је једна непрекидна нумеричка, а друга бинарна (дихотомна) (*Desgraupes et al.*, 2013, стр. 14).

• *Критеријум заснован на* Baker-Hubert-овом γ *коефицијенту конкордансе*

У начелу, примена овог критеријума заснива се на компарацији величине оригиналног одстојања на нивоу сваког, појединачно посматраног, пара опсервација распоређених унутар исте групе (то јест, појединачних одстојања између опсервација чији је пар кодиран вредношћу 0) са сваком од појединачних вредности одстојања на нивоу парова опсервација алоцираних унутар различитих група (односно, појединачних одстојања између опсервација чији је пар кодиран вредношћу 1). На основу исхода спроведених (укупно $n_w \times n_b$) поређења, за свако од могућих решења $C^{(i)}$ (за i=1,2,…, n), утврђује се фреквенција случајева у којима је забележена сагласност (енгл. *concordance*) или несклад (енгл. *discordance*) у погледу односа упоређиваних величина одстојања на нивоу (различито кодираних) парова опсервација. Конкретни пар упоређених одстојања сматра се усаглашеним уколико је одстојање пара опсервација унутар исте групе (у ознаци d_w) стриктно мање од одстојања пара опсервација које припадају различитим групама (у ознаци d_b), и обратно (*Everitt et al.*, 2011, стр. 128). У том смислу, *Baker-Hubert*-ијев γ коефицијент утврђује се путем израза (*Charrad et al.*, 2010, стр. 394; *Milligan & Cooper*, 1985, стр. 163):

$$\gamma_g = \frac{S_{(+)} - S_{(-)}}{S_{(+)} + S_{(-)}} \in [-1, +1], \quad \text{за } g = n, n-1, \ldots, 2, 1, \tag{2.2.34}$$

где симболи $S_{(+)}$ и $S_{(-)}$ означавају укупан број усаглашених (конзистентних) и неусаглашених (неконзистентних) парова одстојања d_w и d_b, то јест, апсолутну фреквенцију исхода поступка компарације у којима је евидентирано присуство релације $d_w < d_b$, односно $d_w > d_b$, респективно, при чему је $S_{(+)} + S_{(-)} = n_w \times n_b$. Конкретна подела опсервација сачињена од g група, за коју је постигнута највећа вредност γ^g показатеља, представља оптимални избор у погледу решења разматраног проблема (*Charrad et al.*, 2014, стр 6; *Milligan & Cooper*, 1985, стр. 163).

• *Критеријум заснован на вредностима* $G_{(+)}$ *коефицијента*

Полазећи од аналогног значења симболике коришћене при разматрању γ коефицијента конкордансе, формула за израчунавање $G_{(+)}$ коефицијента гласи (*Kovačić*, 1994, стр. 287):

$$G_{(+)}^{g} = \frac{S_{(-)}}{n_d (n_d - 1)/2} = \frac{2S_{(-)}}{n_d (n_d - 1)}, \quad \text{за } g = n, n-1, \ldots, 2, 1. \tag{2.2.35}$$

Минимална вредност $G_{(+)}$ коефицијента у серији вредности утврђених за свако појединачно решење проблема груписања (односно, за сваку вредност параметра g), представља индикатор за избор поделе расположивих опсервација која се може сматрати оптималном (*Charrad, et al.*, 2014, стр 8; *Milligan & Cooper*, 1985, стр. 164).

Коначно, посматрано из угла имплементације сваког од наведених, или било којег из широке лепезе расположивих критеријума оптималности, важно је нагласити да не постоји један, универзално најбољи, статистички коефицијент чија се сугестија везана за избор оптималног решења (са аспекта броја или структуре формираних група) може сматрати неприкосновеном и исправном у поређењу са резултирајућим закључцима примене неког(их) другог(их) критеријума. Наиме, будући да се одликују

диференцијалним статистичким својствима, али и заснивају на сагледавању (углавном) различитих аспеката квалитета анализираних решења и, консеквентно, примени различитих поступака израчунавања, назнаке, у погледу избора решења које се може сматрати адекватним за дати узорак / скуп опсервација, добијене њиховом применом, по правилу, ће показивати висок степен варијабилности. У том смислу, већина аутора препоручује коришћење већег броја критеријума за избор оптималне поделе анализираних опсервација (*Everitt et al.*, 2011, стр. 130; *Sharma*, 1996, стр. 202), критички приступ сагледавању генерисаних сугестија (*Milligan & Cooper*, 1985, стр. 160) и, на основама расположивог знања из конкретног подручја примене анализе груписања (*Tuffery*, 2011, стр. 244), њихово синтетизовање у циљу идентификовања најфреквентнијег и „најсмисленијег“, у контексту разматране проблематике, предлога за груписање јединица посматрања. Међутим, иако се учињени избор, генерално, не може третирати као „апсолутна истина“ већ пре као одређена „(статистичка) индикација“ (*Varmuza & Filzmoser*, 2009, стр, 285; *Halkidi et al.*, 2001, стр. 124), поступак опрезне и детаљне валидације резултата представља фундаментални корак у анализи груписања (много више од пуке апликације претходно разматраних критеријума), будући да обезбеђује „чврсте“ (квантитативне) доказе извршеног избора (*Jain & Dubes*, 1988, стр. 137), кроз елиминисање и/или ублажавање ефеката субјективности. Међутим, без обзира на своју неоспорну важност и значај, евалуација квалитета генерисаних резултата груписања анализираних опсервација уједно представља и корак који се најчешће занемарује од стране истраживача при спровођењу анализе груписања (*Aldenderfer & Blashfield*, 1984, стр. 81). У таквим околностима, као главни „статистички“ аргумент за избор одређеног решења наводи се најчешће „интерпретабилност“ поделе. Исти аутори (1984, стр. 81) истичу важност јасног и прецизног навођења статистичке(их) процедуре(а) коришћене(их) за избор оптималног броја и/или структуре група, у циљу елиминисања / ублажавања субјективности у ставу истраживача, а уједно и обезбеђивања компаративних параметара неопходних за евентуални покушај репликације истраживања. Апликативна суштина и статистички значај спровођења поступка евалуације резултата анализе груписања можда се на најбољи начин може исказати следећом констатацијом *Jain*-а *& Dubes*-а (1988, стр. 222): „Валидација резултирајућих структура груписања представља изразито захтеван и врло фрустрирајући сегмент анализе груписања. Међутим, без улагања значајног напора у правцу релаизације исте, анализа груписања ће (по)остати мистериозна (магична) вештина (способност) доступна само истинским верницима који поседују искуство и велику храброст“.

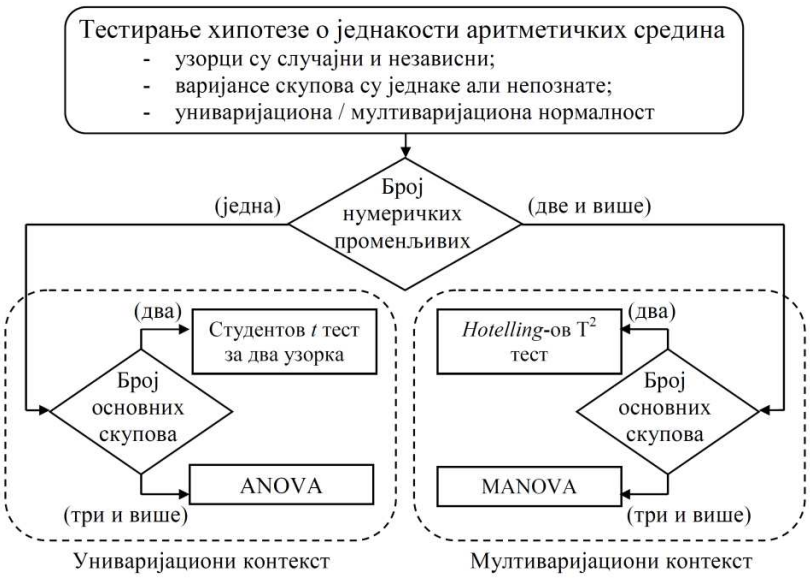
МЕТОДОЛОШКА ОДРЕЂЕЊА ОДАБРАНИХ МУЛТИВАРИЈАЦИОНИХ МЕТОДА ЗАВИСНОСТИ

3.1. МУЛТИВАРИЈАЦИОНА АНАЛИЗА ВАРИЈАНСЕ

У практичним истраживањима, честе су ситуације у којима је од значаја, коришћењем процедуре статистичког закључивања, формулисати одговарајућу одлуку у погледу односа између аритметичких средина две или више популација. Уколико се испитује статистичка значајност разлика аритметичких средина једне променљиве на нивоу два основна скупа, адекватна параметарска статистичка процедура подразумева примену Студентовог *t* теста за два узорка. Униваријационо поређење аритметичких средина више од две популације, заснива се на примени *F* теста, као генерализације поменутог *t* теста на случај три и више скупова, у оквиру једног од најпознатијих статистичких метода под називом анализа варијансе , у ознаци *ANOVA* (енгл. ***AN****alysis* ***O****f* ***VA****riance*). Наведене процедуре статистичког закључивања у литератури су класификоване као униваријационе, због броја, анализом обухваћених, променљивих чије просечне вредности представљају предмет поређења у структури нулте хипотезе. Међутим, често, реализација циљева конципираних истраживања „захтева“ померање аналитичког фокуса са униваријационог на мултиваријациони приступ у испитивању статистичке значајности разлика између посматраних популација, заснованом на мерењу вредности неколико променљивих, а не само једне, и симултаном поређењу њихових аритметичких средина на нивоу (две или више) издвојених група јединица посматрања. Сходно наведеном, а у циљу обезбеђивања адекватних статистичких метода и процедура прилагођених вишедимензионом контексту испитивања разлика између средина две или више популација, почетком 30-их година XX века, захваљујући немерљивом статистичком доприносу *H. Hotelling*–а (1931) и *S.S. Wilks*–а (1932), представљен је концепт мултиваријационе анализе варијансе (енгл. ***M****ultivariate* ***AN****alysis* ***O****f* ***VA****riance*), скраћено *MANOVA*, као вишедимензионог уопштења наведених, еквивалентних униваријационих процедура статистичког закључивања. Релације између униваријационих и мултиваријационих статистичких процедура за анализирање разлика између средина две или више посматраних једнодимензионих и вишедимензионих популација могу се, у општем случају, приказати као на Слици 3.1.

У овом поглављу обезбеђен је детаљан преглед кључних концептуално–методолошких одређења *MANOVA* анализе. Дискусијом је обухваћен само модел једнофакторске *MANOVA*-е, будући да је исти коришћен за потребе реализације емпиријског истраживања презентираног у оквиру последњег поглавља дисертације. Након елаборирања сврхе и циљева *MANOVA*-е, излагање се наставља кратким прегледом специфичних корака карактеристичних за поступак спровођења ове мултиваријационе методе. Посебна пажња посвећена је претпроцесирању мултиваријационих опсервација и, консеквентно, разматрању статистичких претпоставки на којима почива валидна апликација једнофакторске *MANOVA*-е и научна заснованост изведених закључака. У наставку, детаљно је изложен поступак статистичког закључивања у мултиваријационом контексту применом одговарајућих параметарских статистичких тестова, обухватајући и специјални случај *MANOVA*-е, заснован на тестирању једнакости средина на нивоу само две вишедимензионе популације. Такође, за потребе емпиријског истраживања (Поглавље xx), дискусија у овом делу поглавља примарно је усмерена на елаборирање апликативних

могућности *MANOVA*-е у оквиру студија посматрања (енгл. *survey research*). Коначно, навођењем извесних смерницама у контексту спровођења и интерпретације резултата *MANOVA*-е, и накнадне анализе у форми вишеструке компарације, комплетирано је излагање у овом Поглављу.



Слика 3.1. Приказ униваријационих и мултиваријационих параметарских статистичких процедура за анализирање разлика средина две или више популација
Извор: Ауторов приказ

3.1.1. Циљеви и поступак спровођења мултиваријационе анализе варијансе

Генерално, мултиваријациона анализа варијансе (*MANOVA*), представља параметарски, мултиваријациони статистички метод зависности којим се испитује утицај различитих нивоа једне или више „експерименталних“ (независних) променљивих на две или више зависних променљивих (*Kovačić*, 1994, стр. 4). Иако се коришћена терминологија, на први поглед, може учинити нејасном из угла проблематике која се тиче поређења средина вишедимензионих популација, иста је сасвим оправдана и условљена чињеницом да *MANOVA* представља мултиваријационо уопштење *ANOVA*-е. Другим речима, разумевање концепта мултиваријационе анализе варијансе није могуће постићи без потпуног схватања историјског развоја и статистичке филозофије која се налази у основи *ANOVA*-е, као њеног униваријационог статистичког еквивалента. Иницијално, апликативне могућности *ANOVA*-е демонстриране су њеном применом у експерименталним студијама [79] (видети: *Fisher & Mackenzie*, 1923), када су и формулисани специфични термини који су касније преузети и за потребе примене овог метода у оквиру студија посматрања [80] (*Lovrić*, 2009, стр. 252). Терминолошко „наслеђе“ униваријационог

[79] Експерименталне студије (енгл. *experimental design study*) су истраживања у оквиру којих истраживач планира избор експерименталних јединица и директно контролише, односно „манипулише“ једном или више независних променљивих, најчешће категоријских променљивих, у циљу испитивања њихових ефеката на једну (*ANOVA*) или више (*MANOVA*) зависних, углавном нумеричких, променљивих (*Kovačić*, 1994, стр. 95–96; *Hair et al.* 2010, стр. 344).

[80] Студије посматрања (енгл. *survey research*) су истраживања у оквиру којих се мери варијабилитет појава (зависних променљивих) на нивоу јединица посматрања у оквиру, на потпуно случајан начин, генерисаних узорака из више основних скупова, у циљу испитивања валидности нулте хипотезе о једнакости аритметичких средина популација, при чему истраживач ни на који начин не утиче на варијабилитет зависних променљивих које су предмет посматрања (*Lovrić*, 2009, стр. 251).

„рођака“, стављено је на располагање, почетком 30-их година XX века, његовом мултиварицином „наследнику“, *MANOVA* методу.

Посматрано из угла примене *MANOVA*-е у експерименталним студијама, категоријска променљива, чији се утицај на варијабилитет две или више нумеричких променљивих испитује, назива се (контролисани) фактор, експериментална, или независна (објашњавајућа) променљива и обележава симболом *X*. Модалитети фактор променљиве називају се нивои фактора или (експериментални) третмани. Са друге стране, променљиве на којима се испитује дејство контролисаног фактора називају се зависне променљиве и обележавају се симболом *Y*. За разлику од фактора који представља категоријску променљиву мерљиву на номиналној или ординалној скали, зависне променљиве морају бити нумеричке, односно мерљиве на интервалној или скали односа. Такође, насупрот *ANOVA* приступу којим се испитује утицај фактора на варијабилитет само једне зависне променљиве *X*, мултиваријациони карактер *MANOVA*-е детерминисан је укључивањем и симултаним анализирањем две или више зависних нумеричких променљивих. Посматрано из угла броја фактора чије се дејство на варијабилитет анализом обухваћених зависних променљивих испитује, издвајају се следећа два, најчешће коришћена, модела *MANOVA* анализе (*Kovačić*, 1994, стр. 97): ▸ модел *MANOVA* са једним фактором, скраћено једнофакторска *MANOVA* (енгл. *one-way MANOVA*) и ▸ модел *MANOVA* са два фактора, скраћено двофакторска *MANOVA* (енгл. *two-way MANOVA*). Иако се презентиране терминолошке нотације могу на идентичан начин користити при имплементацији *MANOVA*-е у студијама посматрања, поједини аутори (*Rencher*, 2002, стр. 156; *Lovrić*, 2009, стр. 252) наглашавају неопходност опреза при интерпретацији значења експерименталних термина (на пример: експериментална јединица, фактор, нивои фактора, третмани) у студијама посматрања јер се иста могу схватити само у техничком смислу и, на први поглед, могу изазвати конфузију[81].

Наиме, у екпериментални студијама, *MANOVA* омогућава испитивање утицаја (једног или више) фактора на варијабилитет две или више зависних променљивих, провером статистичке значајности оучених разлика између просечних вредности зависних променљивих за које се претпоставља да представљају последицу примене различитих третмана (нивоа фактора) на случајно одабраним експерименталним јединицама. Насупрот наведеном, на јединицама посматрања у студијама посматрања се не примењују експериментални третмани, односно не врши се утицај на варијабилитет зависних променљивих, а „фактор“ има улогу независне, категоријске променљиве (енгл. *grouping variable*) чијим се модалитетима само означава припадност појединачних јединица посматрања конкретној групи или категорији, односно једном од посматраних скупова, или подскупова (стратума), у зависности од нивоа посматрања. Прецизније, у контексту примене једнофакторске *MANOVA*-е у студијама посматрања, само се мери варијабилитет (две или више) зависних променљивих на нивоу јединица посматрања у оквиру генерисаних случајних узорака извучених из три[82] или више скупова (подскупова) који

[81] Према *Lovrić*-у (2009, стр. 252), типичан пример потенцијалне терминолошке конфузије која може настати при спровођењу студија посматрања јесте употреба речи „третман“. Поменути аутор у наставку образлаже изнету констатацију наводећи да у студијама посматрања реч третман означава само један од посматраних скупова (група), а не да су на експерименталним субјектима примењени различити (експериментални) третмани.

[82] Традиционално, примена *MANOVA*-е углавном подразумева поређење аритметичких средина две или више зависних променљивих на нивоу три или више основних скупова (група). Међутим, процедура статистичког закључивања у

репрезентују дефинисане групе у саставу независне променљиве, у циљу испитивања да ли постоји статистички значајна разлика између просечних вредности две или више зависних променљивих на нивоу три или више посматраних скупова, односно група у структури независне променљиве (*Hair et al.*, 2010, стр. 347).

Будући да на постављено истраживачко питање постоје само два могућа одговора, да или не, при чему формулисање истог подразумева коришћење случајних узорака из посматраних популација, јасно је да се реализација циља *MANOVA*-е заснива на поступку тестирања статистичких хипотеза. У општем случају, нулта хипотеза, чија се валидност тестира методом једнофакторске мултиваријационе анализе варијансе, тврди да су аритметичке средине свих p зависних варијабли на нивоу свих g посматраних група (популација) међусобно једнаке, симболички:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_g \quad \rightarrow \quad H_0 : \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{pmatrix} = ... = \begin{pmatrix} \mu_{1g} \\ \mu_{2g} \\ \vdots \\ \mu_{pg} \end{pmatrix} \quad \rightarrow \quad H_0 : \begin{pmatrix} \mu_{11} = \mu_{12} = ... = \mu_{1g} \\ \mu_{21} = \mu_{22} = ... = \mu_{2g} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \mu_{p1} = \mu_{p2} = ... = \mu_{pg} \end{pmatrix}, \quad (3.1.1)$$

где $\boldsymbol{\mu}_k$ означава (p×1) вектор аритметичких средина p зависних променљивих на нивоу k-те популације (групе), при чему је k = 1, 2,…, g и $g \geq 3$, а $p \geq 2$. Представљена хипотеза о једнакости вектора средина подразумева да су g средина сваке од p зависних променљивих једнаке, односно: $\mu_{j1} = \mu_{j2} = ... = \mu_{jg}$ (за j = 1, 2,..., p). Уколико се најмање две аритметичке средине на нивоу само једне од p зависних променљивих међусобно, статистички значајно, разликују (на пример: $\mu_{j1} \neq \mu_{j2}$), тада се нулта хипотеза (израз (3.1.1)) сматра погрешном и исправна одлука подразумева њено одбацивање (*Rencher*, 2002, стр. 159). Другим речима, свих $p(g–1)$ једнакости (израз (3.1.1)-десно) морају бити одрживе како би се H_0 могла сматрати истинитом. У супротном, уз одговарајући ризик грешке I врсте α, врши се њено одбацивање и, консеквентно, усваја алтернативна хипотеза која гласи H_1: постоји статистички значајна разлика између најмање два вектора аритметичких средина зависних променљивих.

Да би се испитала валидност нулте хипотезе, из сваког од g основних скупова формира се по један случајан узорак, величине n_k (k=1, 2,…, g), који репрезентују појединачне категорије у саставу независне променљиве, а прикупљене мултиваријационе опсервације, које ће бити анализиране применом *MANOVA* методе, могу се визуелно приказати у форми (n×p) матрице података (Табела 3.1.1), где n означава укупан број узоркованих опсервација ($n = \Sigma n_k$), а p број зависних променљивих Y_j (j = 1, 2,..., p). Полазећи од симболичких нотација у представљеној матрици мултиваријационих података (Табела 3.1.1), у наставку текста, дефинисан је модел једнофакторске *MANOVA* анализе, као основе за спровођење поступка тестирања формулисане нулте хипотезе и разумевање опште статистичке логике на којој почива примена овог мултиваријационог метода.

основи *MANOVA*-е може се прилагодити и користити и у случају посматрања само две вишедимензионе популације. Реч је о специјалном случају *MANOVA* анализе (*Hair et al.*, 2010, стр. 346–347; *Rencher*, 2002, стр. 130; *Sharma*, 1996, стр. 342–346; *Pituch & Stevens*, 2016, стр. 142; *Coombs, Algina & Olson-Oltman*, 1996, стр. 138), заснованом на примени *Hotelling*-ове T^2 статистике теста, која представља предмет дискусије у оквиру Одељка 3.1.3.

Табела 3.1.1. Матрица мултиваријационих података у једнофакторској *MANOVA* анализи

Јединице посматрања у узорку (n)	Зависне променљиве (Y_j)					Независна променљива X (модалитети $\{x_k\}$)	
	Y_1	Y_2	…	Y_p	векторски запис		
1	y_{111}	y_{121}	…	y_{1p1}	$\mathbf{y}'_{11}$	x_1	(група 1)
2	y_{211}	y_{221}	…	y_{2p1}	$\mathbf{y}'_{21}$	x_1	
⋮	⋮	⋮	…	⋮	⋮	⋮	
n_1	$y_{n_1 11}$	$y_{n_1 21}$	…	$y_{n_1 p1}$	$\mathbf{y}'_{n_1 1}$	x_1	
1	y_{112}	y_{122}	…	y_{1p2}	$\mathbf{y}'_{12}$	x_2	(група 2)
2	y_{212}	y_{222}	…	y_{2p2}	$\mathbf{y}'_{22}$	x_2	
⋮	⋮	⋮	…	⋮	⋮	⋮	
n_2	$y_{n_2 12}$	$y_{n_2 22}$	…	$y_{n_2 p2}$	$\mathbf{y}'_{n_2 2}$	x_2	
⋮	⋮	⋮	…	⋮	⋮	⋮	⋮
1	y_{11g}	y_{12g}	…	y_{1pg}	$\mathbf{y}'_{1g}$	x_g	(група g)
2	y_{21g}	y_{22g}	…	y_{2pg}	$\mathbf{y}'_{2g}$	x_g	
⋮	⋮	⋮	…	⋮	⋮	⋮	
n_g	$y_{n_g 1g}$	$y_{n_g 2g}$	…	$y_{n_g pg}$	$\mathbf{y}'_{n_g g}$	x_g	

Извор: Ауторов приказ

У представљеној табели, коришћени симболи означавају:

X – независна (категоријска) променљива;

x_k – ознака k–тог модалитета (групе) независне променљиве (за k = 1, 2,..., g);

g – укупан број група (категорија) у саставу независне променљиве;

Y_j – j-та зависна променљива (за j = 1, 2,..., p);

p – укупан број зависних променљивих обухваћених анализом;

$y_{ijк}$ – вредност j-те зависне променљиве за i-ту јединицу посматрања у саставу k-те групе (за i = 1, 2,..., n_k, j = 1, 2,..., p и k = 1, 2,..., g);

$\mathbf{y}'_{iк}$ – (1×p) вектор i-те опсервације у саставу k-те групе;

n_k – величина k-тог узорка извученог из k-те популације (групе), за k = 1, 2,..., g;

n – величина укупног узорка, односно, $n = n_1 + n_2 + ... + n_g = \Sigma n_k$, за k = 1, 2,..., g.

Основна идеја наведеног модела огледа се у посматрању и разлагању укупног варијабилитета зависних нумеричких променљивих у циљу откривања, евентуално присутне, систематске тенденције у кретању њихових вредности, која се може приписати утицају контролисаног фактора. Непредвидива одступања од уоченог систематског понашања третирају се као последица случајне грешке. Прецизније, полазећи од варијабилне природе зависних променљивих Y_j, њихов укупан варијабилитет на нивоу појединачних јединица посматрања може се разложити на две адитивне компоненте, од којих је једна ▸ део варијабилитета који се приписује утицају контролисаног фактора, односно припадности конкретној категорији (групи) у саставу независне променљиве, а друга, ▸ део варијабилитета који настаје под утицајем неконтролисаних (резидуалних) фактора који нису директно обухваћени *MANOVA* моделом (*Kovačić*, 1994, стр. 113; 115).

Генерално, једнофакторски *MANOVA* модел за произвољну мултиваријациону опсервацију k-те популације, у ознаци $\mathbf{Y}_{iк}$, може се симболички представити на следећи начин (*Rencher*, 2002, стр. 159):

$$\mathbf{Y}_{ik} = \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}_{ik} = \boldsymbol{\mu} + \boldsymbol{\varphi}_k + \boldsymbol{\varepsilon}_{ik} \rightarrow \begin{bmatrix} y_{i1k} \\ y_{i2k} \\ \vdots \\ y_{ipk} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \varphi_{1k} \\ \varphi_{2k} \\ \vdots \\ \varphi_{pk} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1k} \\ \varepsilon_{i2k} \\ \vdots \\ \varepsilon_{ipk} \end{bmatrix}, \text{ за } i=1,2,..., N_k, \text{ и } k=1,2,\ldots, g. \quad (3.1.2)$$

У представљеном линеарном моделу, $\mathbf{Y}_{iк}$ представља (p×1) вектор вредности p зависних променљивих за i-ту опсервацију унутар k-те популације, $\boldsymbol{\mu}_k$ означава (p×1) вектор аритметичких средина p зависних променљивих на нивоу k-те популације, а $\boldsymbol{\varepsilon}_{ik}$ је (p×1) вектор случајних грешака за који се претпоставља да следи мултиваријациони нормалан распоред са нултим вектором аритметичких средина и коваријационом матрицом $\boldsymbol{\Sigma}$, симболички, $\boldsymbol{\varepsilon}_{ik}$: $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, (*Everitt*, 2010, стр. 271). У развијеном облику израза, симбол $\boldsymbol{\mu}$ представља (p×1) вектор општих (заједничких) средина p зависних променљивих на нивоу свих g популација (група), а $\boldsymbol{\varphi}_k$ означава (p×1) вектор ефеката k-тог третмана (групе) на варијабилитет зависних променљивих.

Будући да вектор $\boldsymbol{\varphi}_k$ показује одступања аритметичких средина p зависних променљивих на нивоу k-те популације, $\boldsymbol{\mu}_k$, од вектора општих средина на нивоу свих популација $\boldsymbol{\mu}$ (односно $\boldsymbol{\varphi}_k = \boldsymbol{\mu}_k - \boldsymbol{\mu}$), а вектор $\boldsymbol{\varepsilon}_{ik}$ случајна, под утицајем неконтролисаних (резидуалних) фактора изазвана, одступања вредности p зависних променљивих, на нивоу i-те опсервације, од кореспондентних средина припадајуће k-те популације, $\boldsymbol{\mu}_k$ (односно $\boldsymbol{\varepsilon}_{ik} = \mathbf{Y}_{ik} - \boldsymbol{\mu}_k$), наведени модел може се представити у следећем облику (*Kovačić*, 1994, стр. 114):

$$\mathbf{Y}_{ik} = \boldsymbol{\mu} + \underbrace{(\boldsymbol{\mu}_k - \boldsymbol{\mu})}_{\boldsymbol{\varphi}_k} + \underbrace{(\mathbf{Y}_{ik} - \boldsymbol{\mu}_k)}_{\boldsymbol{\varepsilon}_{ik}} \rightarrow \begin{bmatrix} y_{i1k} \\ y_{i2k} \\ \vdots \\ y_{ipk} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \mu_{1k} - \mu_1 \\ \mu_{2k} - \mu_2 \\ \vdots \\ \mu_{pk} - \mu_p \end{bmatrix} + \begin{bmatrix} y_{i1k} - \mu_{1k} \\ y_{i2k} - \mu_{2k} \\ \vdots \\ y_{ipk} - \mu_{pk} \end{bmatrix}. \quad (3.1.3)$$

У контексту представљеног *MANOVA* модела, уколико је H_0, формулисана изразом (3.1.1), истинита, односно уколико независна променљива X нема утицај на варијабилитет зависних променљивих Y_j (j = 1, 2,…, p), тада се може констатовати да не постоје статистички значајне разлике између вектора аритметичких средина, $\boldsymbol{\mu}_k$ (k = 1, 2,…, g) и вектора општих средина $\boldsymbol{\mu}$, односно, да је ефекат третмана $\boldsymbol{\varphi}_k$ нулти вектор (симболички: $\boldsymbol{\varphi}_k = \mathbf{0}$) за сваку од g популација (група / третмана). Другим речима, аритметичке средине свих p зависних променљивих, на нивоу свих g популација, у ознаци $\boldsymbol{\mu}_k$, могу се сматрати једнаким, односно H_0: $\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=...=\boldsymbol{\mu}_g=\boldsymbol{\mu}$, а модел дат изразом (3.1.3), консеквентно и аналогно идеји у основи једнофакторског *ANOVA* модела (видети у: *Lovrić*, 2009, стр. 256–257), добија облик: $\mathbf{Y}_{iк} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{ik}$, сугеришући да се варијабилитет вредности зависних променљивих на нивоу i-те опсервације у саставу k-те популације може приписати само случајним, неконтролисаним факторима (*Kovačić*, 1994, стр. 118). Полазећи од наведеног, садржај нулте хипотезе, представљен изразом (3.1.1), може се еквивалентно формулисати на следећи начин: H_0: $\boldsymbol{\varphi}_1 = \boldsymbol{\varphi}_2 = ... = \boldsymbol{\varphi}_g = \mathbf{0}$ (*Johnson & Wichern*, 2007, стр. 302; *Everitt*, 2010, стр. 271), а кореспондентна алтернативна хипотеза, H_1: вектор ефеката најмање једног третмана (групе) се статистички значајно разликује од нултог вектора (*Kovačić*, 1994, стр. 118). Са друге стране, уколико је нулта хипотеза погрешна, тада се може констатовати да

постоје два извора варијабилитета, анализом обухваћених, зависних промењљивих и то: један, под утицајем независне промењљиве (контролисаног фактора), и други, на основу здруженог утицаја неконтролисаних фактора.

Оцењивање представљеног модела *MANOVA* са једним фактором спроводи се на основу случајних узорака сачињених од по n_k елемената $\mathbf{y}_{1k}$, $\mathbf{y}_{2k}$,..., $\mathbf{y}_{n_k k}$, за k = 1, 2,..., g, извучених из g посматраних популација (Табела 3.1.1). Произвољна мултиваријациона опсервација у оквиру k-тог узорка, у ознаци $\mathbf{y}_{ik}$, може се, истоветно описаном поступку, исказати следећом релацијом (*Johnson & Wichern*, 2007, стр. 301):

$$\underset{\left(\begin{array}{c}\text{мултиваријациона}\\ \text{опсервација у } k\text{-том узорку}\end{array}\right)}{\mathbf{y}_{ik}} = \underset{\left(\begin{array}{c}\text{оцена вектора}\\ \text{општих средина, } \hat{\boldsymbol{\mu}}\end{array}\right)}{\bar{\bar{\mathbf{y}}}} + \underset{\left(\begin{array}{c}\text{оцена вектора ефеката}\\ k\text{-тог третмана, } \hat{\boldsymbol{\phi}}_k\end{array}\right)}{(\bar{\mathbf{y}}_k - \bar{\bar{\mathbf{y}}})} + \underset{\left(\begin{array}{c}\text{оцена вектора}\\ \text{резидуала, } \hat{\boldsymbol{\varepsilon}}_{ik}\end{array}\right)}{(\mathbf{y}_{ik} - \bar{\mathbf{y}}_k)}, \qquad (3.1.4)$$

или у развијеном облику

$$\begin{bmatrix} y_{i1k} \\ y_{i2k} \\ \vdots \\ y_{ipk} \end{bmatrix} = \begin{bmatrix} \bar{\bar{y}}_1 \\ \bar{\bar{y}}_2 \\ \vdots \\ \bar{\bar{y}}_p \end{bmatrix} + \begin{bmatrix} \bar{y}_{1k} - \bar{\bar{y}}_1 \\ \bar{y}_{2k} - \bar{\bar{y}}_2 \\ \vdots \\ \bar{y}_{pk} - \bar{\bar{y}}_p \end{bmatrix} + \begin{bmatrix} y_{i1k} - \bar{y}_{1k} \\ y_{i2k} - \bar{y}_{2k} \\ \vdots \\ y_{ipk} - \bar{y}_{pk} \end{bmatrix}, \quad \text{за } i = 1, 2,..., n_k \text{ и } k = 1, 2,\ldots, g. \qquad (3.1.4a)$$

У оцењеном једнофакторском *MANOVA* моделу за $\mathbf{y}_{ik}$ опсервацију, $\bar{\mathbf{y}}_k$ представља (p×1) вектор средина p зависних промењљивих на нивоу k-тог узорка, величине n_k (за k = 1, 2,…, g), а $\bar{\bar{\mathbf{y}}}$ означава (p×1) вектор општих (заједничких) средина p зависних промењљивих на нивоу свих g узорака (група), односно укупног узорка величине $n = \Sigma n_k$ (за k = 1, 2,..., g). Њихови елементи израчунавају се коришћењем израза (3.1.5) и (3.1.6), респективно.

$$\bar{\mathbf{y}}_k = \begin{bmatrix} \bar{y}_{1k} \\ \bar{y}_{2k} \\ \vdots \\ \bar{y}_{pk} \end{bmatrix}, \text{ где је: } \bar{y}_{jk} = \frac{\sum_{i=1}^{n_k} y_{ijk}}{n_k}, \quad \text{за } i = 1, 2,..., n_k, \; j = 1, 2,..., p \text{ и } k = 1, 2,..., g. \qquad (3.1.5)$$

$$\bar{\bar{\mathbf{y}}} = \begin{bmatrix} \bar{\bar{y}}_1 \\ \bar{\bar{y}}_2 \\ \vdots \\ \bar{\bar{y}}_p \end{bmatrix}, \text{ где је: } \bar{\bar{y}}_j = \frac{\sum_{k=1}^{g} n_k \bar{y}_{jk}}{\sum_{k=1}^{g} n_k}, \text{ за } i = 1, 2,..., n_k, \; j = 1, 2,..., p \text{ и } k = 1, 2,..., g. \qquad (3.1.6)$$

Аналогно примењеном поступку у оквиру униваријационе *ANOVA*-е (детаљно објашњење видети, на пример, у: *Lovrić*, 2009, стр. 258), разлагање укупног варијабилитета случајно одабране мултиваријационе опсервације $\mathbf{y}_{ik}' = [y_{i1к}, y_{i2к},\ldots, y_{ipк}]$, који се одређује као разлика између вредности те опсервације и вредности заједничких средина p зависних промењљивих свих g узорака, може се представити следећим изразом (*Kovačić*, 1994, стр. 115):

$$\underbrace{\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}}}_{\substack{\text{укупан варијабилитет}\\ i\text{-те опсервације } k\text{-тог узорка}}} = \underbrace{(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})}_{\substack{\text{факторски}\\ \text{варијабилитет}}} + \underbrace{(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)}_{\substack{\text{резидуални}\\ \text{варијабилитет}}} \rightarrow \begin{bmatrix} y_{i1k}-\bar{\bar{y}}_1 \\ y_{i2k}-\bar{\bar{y}}_2 \\ \vdots \\ y_{ipk}-\bar{\bar{y}}_p \end{bmatrix} = \begin{bmatrix} \bar{y}_{1k}-\bar{\bar{y}}_1 \\ \bar{y}_{2k}-\bar{\bar{y}}_2 \\ \vdots \\ \bar{y}_{pk}-\bar{\bar{y}}_p \end{bmatrix} + \begin{bmatrix} y_{i1k}-\bar{y}_{1k} \\ y_{i2k}-\bar{y}_{2k} \\ \vdots \\ y_{ipk}-\bar{y}_{pk} \end{bmatrix}. \quad (3.1.7)$$

Генерализацијом приказане логике на ниво свих $\mathbf{y}_{ik}$ опсервација g узорака, одређује се укупан варијабилитет свих p зависних променљивих. У том контексту, уз уважавање једне од кључних особина аритметичке средине[83], неопходно је, иницијално, извршити квадрирање обе стране израза (3.1.7) на следећи начин (*Kovačić*, 1994, стр. 115):

$$\begin{aligned} (\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}})^2 &= \left[(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})+(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)\right]^2 \\ (\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}})(\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}})' &= \left[(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})+(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)\right]\left[(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})+(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)\right]' \\ &= (\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})'+(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)'+(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})'+(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)' \end{aligned} \quad (3.1.8)$$

Сумирањем леве и десне стране израза (3.1.8) по индексима i (за i=1,2,...,n_k) и k (за k=1,2,…,g) добија се (*Kovačić*, 1994, стр. 115; *Johnson & Wichern*, 2007, стр. 302):

$$\underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}})(\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}})'}_{\substack{\text{УКУПАН ВАРИЈАБИЛИТЕТ}\\ \text{(матрица укупне суме квадрата и узајамних производа одступања)}}} = \underbrace{\sum_{k=1}^{g}n_k(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})'}_{\substack{\text{ФАКТОРСКИ ВАРИЈАБИЛИТЕТ}\\ \text{(матрица суме квадрата и узајамних производа одступања између група)}}} + \underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)'}_{\substack{\text{РЕЗИДУАЛНИ ВАРИЈАБИЛИТЕТ}\\ \text{(матрица суме квадрата и узајамних производа одступања унутар група)}}}, \quad (3.1.9)$$

$$[\mathbf{T}] \quad = \quad [\mathbf{B}] \quad + \quad [\mathbf{W}]$$

будући да су, у изразу (3.1.8), два средишња сабирка нула матрице, на основу следећих једнакости (*Kovačić*, 1994, стр. 115): $\sum_{k=1}^{g}n_k(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}})=\mathbf{0}$ и $\sum_{i=1}^{nk}(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k)=\mathbf{0}$.

Матрица укупне суме квадрата и узајамних производа одступања свих мултиваријационих опсервација, односно вредности p зависних променљивих свих n јединица посматрања, од кореспондентних вредности заједничких аритметичких средина утврђених на нивоу свих g узорака, у ознаци **T**, представља уопштење укупне суме квадрата у *ANOVA* анализи (Табела 3.1.2). Матрице суме квадрата између и унутар група, у ознаци **B** и **W**, респективно, имају кључну улогу у поступку тестирања валидности *MANOVA* нулте хипотезе, а њихови елементи могу се, у матричној форми, дефинисати на следећи начин:

$$\underset{(p\times p)}{\mathbf{B}} = \begin{bmatrix} \sum_{k=1}^{g}n_k(\bar{y}_{1k}-\bar{\bar{y}}_1)^2 & \sum_{k=1}^{g}n_k(\bar{y}_{1k}-\bar{\bar{y}}_1)(\bar{y}_{2k}-\bar{\bar{y}}_2) & \cdots & \sum_{k=1}^{g}n_k(\bar{y}_{1k}-\bar{\bar{y}}_1)(\bar{y}_{pk}-\bar{\bar{y}}_p) \\ \sum_{k=1}^{g}n_k(\bar{y}_{2k}-\bar{\bar{y}}_2)(\bar{y}_{1k}-\bar{\bar{y}}_1) & \sum_{k=1}^{g}n_k(\bar{y}_{2k}-\bar{\bar{y}}_2)^2 & \cdots & \sum_{k=1}^{g}n_k(\bar{y}_{2k}-\bar{\bar{y}}_2)(\bar{y}_{pk}-\bar{\bar{y}}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{g}n_k(\bar{y}_{pk}-\bar{\bar{y}}_p)(\bar{y}_{1k}-\bar{\bar{y}}_1) & \sum_{k=1}^{g}n_k(\bar{y}_{pk}-\bar{\bar{y}}_p)(\bar{y}_{2k}-\bar{\bar{y}}_2) & \cdots & \sum_{k=1}^{g}n_k(\bar{y}_{pk}-\bar{\bar{y}}_p)^2 \end{bmatrix}, \quad (3.1.10)$$

[83] Збир одступања свих вредности обележја од њихове аритметичке средине увек је једнак нули (*Lovrić*, 2009, стр. 47). Консеквентно, сâм збир одступања не представља адекватну меру варијабилитета посматране променљиве.

$$\underset{(p\times p)}{\mathbf{W}} = \begin{bmatrix} \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{i1k}-\bar{y}_{1k})^2 & \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{i1k}-\bar{y}_{1k})(y_{i2k}-\bar{y}_{2k}) & \cdots & \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{i1k}-\bar{y}_{1k})(y_{ipk}-\bar{y}_{pk}) \\ \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{i2k}-\bar{y}_{2k})(y_{i1k}-\bar{y}_{1k}) & \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{i2k}-\bar{y}_{2k})^2 & \cdots & \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{i2k}-\bar{y}_{2k})(y_{ipk}-\bar{y}_{pk}) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{ipk}-\bar{y}_{pk})(y_{i1k}-\bar{y}_{1k}) & \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{ipk}-\bar{y}_{pk})(y_{i2k}-\bar{y}_{2k}) & \cdots & \sum_{k=1}^{g}\sum_{i=1}^{n_k}(y_{ipk}-\bar{y}_{pk})^2 \end{bmatrix} \quad (3.1.11)$$

Елементи главне дијагонале симетричне ($p\times p$) матрице **B** представљају суме квадрата одступања аритметичких средина за сваку од p појединачних зависних промењљивих унутар сваке од g група (узорка) од опште средине конкретне зависне варијабле Y_j утврђене на нивоу укупног узорка величине n. Вандијагонални елементи представљају аналогне суме производа одступања аритметичких средина на нивоу група од општег просека за сваки пар зависних променљивих (*Rencher*, 2002, стр. 160). Са друге стране, j-ти елемент главне дијагонале ($p\times p$) симетричне матрице **W** је сума квадрата одступања опсервација од аритметичке средине j-те зависне промењљиве k-тог узорка (групи), а вандијагонални елементи представљају суме узајамних производа аналогних одступања (*Kovačić*, 1994, стр. 116). Поред наведеног, матрица **W** може се изразити као пондерисани збир узорачких коваријационих матрица $\mathbf{S}_k$, за k = 1, 2,…, g (*Kovačić*, 1994, стр. 115; *Johnson & Wichern*, 2007, стр. 302), односно:

$$\underset{(p\times p)}{\mathbf{W}} = \sum_{k=1}^{g}\left(n_k-1\right)\mathbf{S}_k = \left(n_1-1\right)\mathbf{S}_1+\left(n_2-1\right)\mathbf{S}_2+\ldots+\left(n_g-1\right)\mathbf{S}_g, \quad (3.1.12)$$

где је

$$\underset{(p\times p)}{\mathbf{S}_k} = \frac{1}{n_k-1}\begin{bmatrix} \sum_{i=1}^{n_k}\left(y_{i1}-\bar{y}_1\right)^2 & \sum_{i=1}^{n_k}\left(y_{i1}-\bar{y}_1\right)\left(y_{i2}-\bar{y}_2\right) & \cdots & \sum_{i=1}^{n_k}\left(y_{i1}-\bar{y}_1\right)\left(y_{ip}-\bar{y}_p\right) \\ \sum_{i=1}^{n_k}\left(y_{i2}-\bar{y}_2\right)\left(y_{i1}-\bar{y}_1\right) & \sum_{i=1}^{n_k}\left(y_{i2}-\bar{y}_2\right)^2 & \cdots & \sum_{i=1}^{n_k}\left(y_{i2}-\bar{y}_2\right)\left(y_{ip}-\bar{y}_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n_k}\left(y_{ip}-\bar{y}_p\right)\left(y_{i1}-\bar{y}_1\right) & \sum_{i=1}^{n_k}\left(y_{ip}-\bar{y}_p\right)\left(y_{i2}-\bar{y}_2\right) & \cdots & \sum_{i=1}^{n_k}\left(y_{ip}-\bar{y}_p\right)^2 \end{bmatrix}. \quad (3.1.13)$$

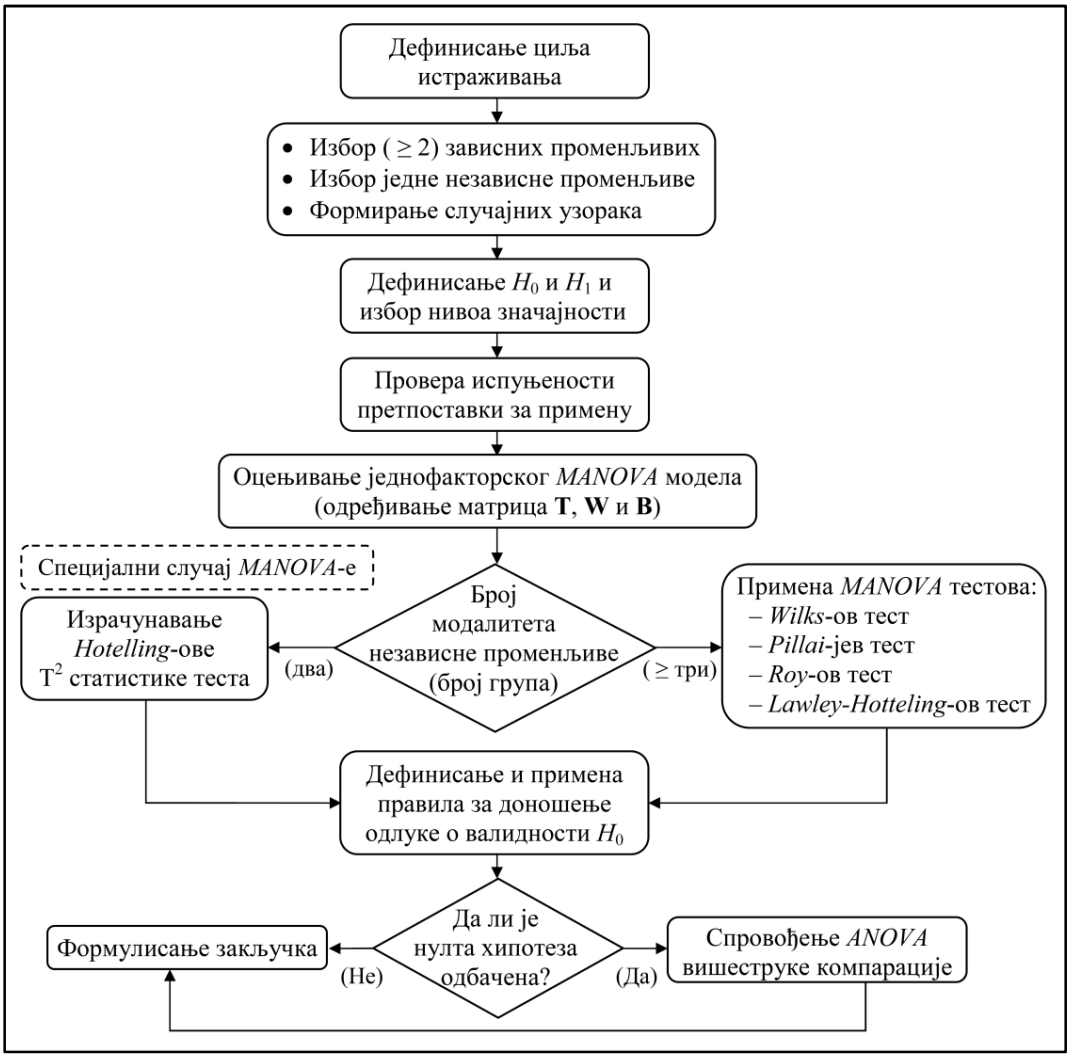
Додатна визуелна потврда констатације да *MANOVA* представља вишедимензионо уопштење *ANOVA* процедуре, обезбеђена је сумарним, упоредним приказом адитивних компоненти укупног варијабилитета зависне(их) променљиве(их) у кореспондентним моделима (Табела 3.1.2).

Аналогно *ANOVA* униваријационом приступу, тестирање валидности H_0 у *MANOVA* анализи заснива се, индиректно, на поређењу варијација између узорака и присутног варијабилитета унутар узорака, при чему, што су варијације између група веће у односу на одступања унутар група јачи су и аргументи против формулисане нулте хипотезе и обрнуто. Шематски приказ поступка спровођења мултиваријационе анализе варијансе представљен је на Слици 3.1.1.

Табела 3.1.2. Упоредни приказ адитивних компоненти укупног варијабилитета у моделу *ANOVA* и *MANOVA* са једним фактором

Извор варијабилитета	Једнофакторска *ANOVA*	Једнофакторска *MANOVA*
	сума квадрата одступања	матрица суме квадрата и узајамних производа одступања
Факторски варијабилитет (између узорака / група)	$B=\sum_{k=1}^{g} n_k\left(\bar{y}_k-\bar{\bar{y}}\right)^2$	$\mathbf{B}=\sum_{k=1}^{g} n_k\left(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}}\right)\left(\bar{\mathbf{y}}_k-\bar{\bar{\mathbf{y}}}\right)'$
Резидуални варијабилитет (унутар узорака / група)	$W=\sum_{k=1}^{g}\sum_{i=1}^{nk}\left(y_{ik}-\bar{y}_k\right)^2$	$\mathbf{W}=\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k\right)\left(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k\right)'$
Укупан варијабилитет	$T=\sum_{k=1}^{g}\sum_{i=1}^{nk}\left(y_{ik}-\bar{\bar{y}}\right)^2$	$\mathbf{T}=\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}}\right)\left(\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}}\right)'$

Извор: Ауторов приказ, прилагођено према *Johnson & Wichern* (2007, стр. 300; 302)



Слика 3.1.1. Шематски приказ поступка спровођења једнофакторске *MANOVA*-е

Извор: Ауторов приказ

Сходно презентираном дијаграму тока поступка спровођења једнофакторске *MANOVA*-е, излагањем у наставку обухваћена је дискусија о кључним статистичким претпоставкама на којима се заснива валидна апликација овог параметарског мултиваријационог статистичког метода и, консеквентно, научна заснованост формулисаних закључака у контексту анализиране проблематике. Такође, иако се примена *MANOVA* методе примарно везује за испитивање статистичке значајности утицаја

независне променљиве, са три или више модалитета, на варијабилитет скупа зависних променљивих, у оквиру Одељка 3.1.3, посебна пажња посвећена је елаборирању поступка статистичког закључивања о једнакости аритметичких средина две мултиваријационе популације, заснован на примени *Hotelling*-ове T^2 статистике теста, као специјалног случаја *MANOVA*-е, у знак поштовања према *Harold*-у. *Hotelling*-у и његовом немерљивом доприносу мултиваријационој генерализацији Студентовог *t* теста за два (униваријациона) узорка. Наиме, упоредо са схватањем *MANOVA*-е, из угла броја анализом обухваћених зависних променљивих, као вишедимензионог уопштења *ANOVA* процедуре, *MANOVA* се такође може сматрати и директном генерализацијом *Hotelling*-ове T^2 процедуре за случај посматрања три или више популација, односно група у структури независне променљиве. Коначно, излагање у оквиру овог поглавља комплетирано је адекватним објашњењем анализе вишеструке компарације, заснованој на примени *post hoc* тестова, у случају када је резултатима *MANOVA* статистичких тестова потврђена одлука о одбацивању нулте хипотезе о одсуству ефеката примењених третмана (група) на варијабилитет зависних променљивих.

3.1.2. Статистичке претпоставке за примену мултиваријационе анализе варијансе

Поступак тестирања нулте хипотезе о једнакости аритметичких средина мултидимензионих популација је знатно комплекснији у односу на аналогни униваријациони контекст анализе. Један од разлога наведене констатације садржан је и у статистичким претпоставкама на којима се заснива валидно спровођење *MANOVA* процедуре и, консеквентно, научна заснованост резултирајућих закључака. У том смислу, у литератури (*Rencher*, 2002, стр. 158; 198; *Johnson & Wichern*, 2007, стр. 296–297; *Pituch & Stevens*, 2016, стр. 220; *Boslaugh & Watters*, 2008, стр. 250; *Hair et al.*, 2010, стр. 361–363; *Sharma*, 1996, стр. 374) углавном се наводе следеће три кључне претпоставке, које морају бити проверене и у одговарајућем степену задовољене, у погледу структуре расположивих мултиваријационих података:

- ✓ Независност мултиваријационих опсервација $\mathbf{y}_{1k}$, $\mathbf{y}_{2k}$,..., $\mathbf{y}_{n_k k}$, *k*-тих узорака (k=1,2,..., g);
- ✓ Нормалност мултиваријационог распореда p зависних променљивих;
- ✓ Хомогеност коваријационих матрица популација из којих су извучени узорци.

Поред ових стриктних статистичких претпоставки, неопходно је такође размотрити и одређена униваријациона и мултиваријациона својства анализираних g узорака опсервација, у складу са значајем, односно утицајем, који имају у обезбеђивању испуњености претходно наведених услова, креирању *MANOVA* модела и реализацији параметарске статистичке процедуре тестирања *MANOVA* нулте хипотезе (*Hair et al.*, 2010, стр. 363), а која могу бити формулисана на следећи начин:

- ✓ Униваријациона нормалност распореда појединачних зависних променљивих, како на нивоу укупног узорка тако и по издвојеним модалитетима независне променљиве;
- ✓ Одсуство униваријационих и мултиваријационих нестандардних опсервација, како на нивоу укупног узорка тако и на нивоу појединачних група јединица посматрања;
- ✓ Присуство статистички значајне линеарне везе између зависних променљивих; и
- ✓ Одсуство мултиколинеарности и сингуларности између зависних променљивих.

Наводећи детаљно образложење, *Rencher* (2002, стр. 198–199) и *Pituch & Stevens* (2016, стр. 220–224) издвајају независност вектора мултиваријационих опсервација на нивоу издвојених узорака као најважнији од претходно наведених предуслова, с обзиром на озбиљност (негативних) последица које, услед нарушености ове претпоставке, могу настати у контексту доношења валидних закључака анализе. Важност испуњености претпоставке о мултиваријационој нормалности распореда опсервација на нивоу популација (у ознаци, $\mathbf{X}_k \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$), произилази из чињенице да се *MANOVA* заснива на коришћењу параметарских статистичких процедура при тестирању нулте хипотезе о једнакости вектора средина. Теоријски, нарушеност претпоставке о мултиваријационој нормалности може имати негативне ефекте на јачину коришћених статистичких тестова, при чему је њихов интензитет детерминисан, између осталог, размером одступања од нормалности, величином појединачних узорака, као и њиховим међусобним односом. Наиме, велики број аутора истиче да се *MANOVA* процедуре могу сматрати у извесној мери робустним у случају мањих одступања од мултиваријационе нормалности, под претпоставком да су коришћени узорци једнаких или приближно једнаких (*Rencher*, 2002, стр. 198; *Manly & Navarro Alberto*, 2017, стр. 56) и / или довољно великих величина (*Johnson & Wichern*, 2007, стр. 297; *Huberty & Petoskey*, 2000, стр. 192; *Hair et al.*, 2010, стр. 363). Такође, мултиваријациона нормалност популација представља веома важан предуслов и при провери претпоставке о једнакости коваријационих матрица на нивоу посматраних популација. Повезаност наведених претпоставки произлази из параметарске природе статистичких тестова који се углавном користе за тестирање нулте хипотезе о хомогености матрица коваријанси две (g=2) или више (g>2) популација (група), која гласи:

$$\underbrace{H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}}_{\text{случај 2 групе}} \quad \wedge \quad \underbrace{H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}}_{\text{случај } g \text{ група (популација)}}, \tag{3.1.14}$$

где је $\boldsymbol{\Sigma}_k$ коваријациона матрица k-те популације (k = 1,2,...,g), а $\boldsymbol{\Sigma}$ претпостављена заједничка коваријациона матрица посматраних популација. Алтернативном хипотезом обухваћена је тврдња да постоји најмање једна популациона коваријациона матрица, од посматраних g, која се разликује од осталих, односно да постоји статистички значајна разлика између коваријационих матрица најмање две популације (*Kovačić*, 1994, стр. 122; *Johnson & Wichern*, 2007, стр. 310). *Box* (1949) је предложио тест, који представља генерализацију *Bartlett*-овог теста универијационе хомогености варијанси, за потребе тестирања нулте хипотезе дате изразом (3.1.14), под називом *Box*-ов *M* тест. Његова статистика, заснована на употреби генерализованих варијанси, односно детерминанти здружене и коваријационих матрица појединачних узорака, гласи (*Box*, 1949, стр. 320):

$$\begin{aligned} M &= -\left[\sum_{k=1}^{g}\left(n_k - 1\right)\right]\log_e\left\{\prod_{k=1}^{g}\left[\frac{|\mathbf{S}_k|}{|\bar{\mathbf{S}}|}\right]^{\frac{n_k-1}{\sum_{k=1}^{g}(n_k-1)}}\right\} \\ &= -\left[\sum_{k=1}^{g}\left(n_k - 1\right)\right]\log_e\left\{\left(\frac{|\mathbf{S}_1|}{|\bar{\mathbf{S}}|}\right)^{\frac{n_1-1}{n-g}} \cdot \left(\frac{|\mathbf{S}_2|}{|\bar{\mathbf{S}}|}\right)^{\frac{n_2-1}{n-g}} \cdot \ldots \cdot \left(\frac{|\mathbf{S}_g|}{|\bar{\mathbf{S}}|}\right)^{\frac{n_g-1}{n-g}}\right\}, \\ &= \left(n-g\right)\log_e|\bar{\mathbf{S}}| - \sum_{k=1}^{g}\left[\left(n_k - 1\right)\log_e|\mathbf{S}_k|\right] \end{aligned} \tag{3.1.15}$$

где симбол n_k означава величину k-тог узорка (k = 1, 2,..., g), n величину укупног узорка (односно: $n = n_1 + n_2 + \ldots + n_g$), $|\mathbf{S}_k|$ детерминанту коваријационе матрице k-тог узорка, а $|\overline{\mathbf{S}}|$ (или $|\mathbf{S}_{pooled}|$) представља детерминанту непристрасне, пондерисане (енгл. *pooled*) оцене заједничке коваријационе матрице датих g популација (симболички: $\mathrm{E}(\mathbf{S}_{pooled}) = \boldsymbol{\Sigma}$), под претпоставком да је H_0: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ истинита. Заједничка коваријациона матрица здружених узорака, у ознаци $\overline{\mathbf{S}}$ (или $\mathbf{S}_{pooled}$) утврђује се као пондерисани просек, изразом (3.1.13) дефинисаних, g узорачких коваријационих матрица (односно, $\mathbf{S}_1$, $\mathbf{S}_2$,..., $\mathbf{S}_g$, при чему је: $\mathrm{E}(\mathbf{S}_1)=\boldsymbol{\Sigma}_1$, $\mathrm{E}(\mathbf{S}_2)=\boldsymbol{\Sigma}_2$,..., $\mathrm{E}(\mathbf{S}_g)=\boldsymbol{\Sigma}_g$), на следећи начин (*Rencher*, 2002, стр. 122–123; *Johnson & Wichern*, 2007, стр. 310):

$$\underset{(p\times p)}{\overline{\mathbf{S}}} = \frac{1}{n-g}\mathbf{W} = \frac{\sum_{k=1}^{g}(n_k-1)\mathbf{S}_k}{\sum_{k=1}^{g}(n_k-1)} = \frac{(n_1-1)\mathbf{S}_1+(n_2-1)\mathbf{S}_2+\ldots+(n_g-1)\mathbf{S}_g}{(n_1+n_2+\ldots+n_g)-g} = \begin{bmatrix} \overline{s}_{11} & \overline{s}_{12} & \ldots & \overline{s}_{1p} \\ \overline{s}_{21} & \overline{s}_{22} & \ldots & \overline{s}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{s}_{p1} & \overline{s}_{p2} & \ldots & \overline{s}_{pp} \end{bmatrix}, \quad (3.1.16)$$

где $\mathbf{W}$ означава ($p\times p$) матрицу суме квадрата и узајамних производа одступања унутар посматраних g узорака, дефинисану изразима (3.1.11) и (3.1.12). У оквиру матрице $\mathbf{S}_{pooled}$, j-ти елемент на главној дијагонали (за j=1, 2,..., p) представља здружену варијансу j-те зависне променљиве на нивоу укупног узорка величине n, док вандијагонални елементи означавају коваријансе појединачних парова зависних променљивих.

У складу са тумачењем логике у основи M статистике, презентираним од стране *Johnson*-а *& Wichern*-а (2007, стр. 311), може се приметити да уколико је H_0 истинита, појединачне узорачке коваријационе матрице ($\mathbf{S}_k$) не би требало да се значајније разликују од опште узорачке коваријационе матрице $\mathbf{S}_{pooled}$, што би резултирало вредностима количника њихових детерминанти блиским јединици и, консеквентно, релативно малим вредностима статистике *Box*-овог M теста. С друге стране, са повећањем размера одступања појединачних $\mathbf{S}_k$ матрица од опште $\mathbf{S}_{pooled}$ матрице, количник детерминанти кореспондентних матрица се приближава нули [84], узрокујући пораст вредности M статистике и одбацивање тврдње садржане у H_0. С обзиром на сложеност и ограничени обухват узорачког распореда *Box*-ове M статистике, на коју указују *Kovačić* (1994, стр. 122) и *Rencher* (2002, стр. 588–589), за потребе сагледавања да ли је добијена вредност M статистике сигнификантно велика, користе се апроксимације *Box*-ове M статистике засноване на *Snedecor*-овом F распореду и χ^2 распореду, предложене од стране *Box*-а (1949). У том смислу, трансформацијом M статистике према изразу (3.1.17) добија се статистика која је апроксимативно распоређена по χ^2 распореду са $\nu = [p(p+1)(g-1) / 2]$ степени слободе, односно (*Box*, 1949, стр. 329):

$$M \rightarrow \chi^2 = \frac{M}{C} \sim (\text{приближно})\chi^2_{\alpha,\underbrace{\frac{p(p+1)(g-1)}{2}}_{\nu}}, \text{ за } \frac{1}{C} = 1 - \frac{2p^2+3p-1}{6(p+1)(g-1)}\left(\sum_{k=1}^{g}\frac{1}{(n_k-1)} - \frac{1}{\sum_{k=1}^{g}(n_k-1)}\right). \quad (3.1.17)$$

[84] У контексту провере елабориране претпоставке, а полазећи од обрасца за статистику *Box*-овог M теста, *Rencher* (2002, стр. 256) и *Tabachnick* & *Fidell* (2013, стр. 252) напомињу да број степени слободе (n_k–1) појединачних узорачких коваријационих матрица ($\mathbf{S}_k$, за k = 1,2,…,g) мора бити већи од броја зависних променљивих, p, јер ће у супротном, вредност детерминанте конкретне коваријационе матрице $\mathbf{S}_k$, за коју наведени услов није задовољен, бити једнака нули ($|\mathbf{S}_k|$=0) као и вредност количника детерминанти у изразу (3.1.15).

Уз дефинисани ниво значајности теста α, нулта хипотеза се одбацује уколико је $(M / C) > \chi^2_{\alpha,\nu}$. Употреба предложене χ^2 апроксимације препоручује се у ситуацијама када је величина појединачних, анализом обухваћених, узорака $n_k \geq 20$, број зависних променљивих $p \leq 5$ и број узорака (група) $g \leq 5$ (*Box*, 1949, стр. 336). У ситуацијама које не задовољавају комбинацију вредности наведених параметара, користи се, предложена такође од стране *Box*-а (1949, стр. 334), у односу на χ^2 апроксимацију, прецизнија апроксимативна F статистика, која приближно следи F распоред са $\nu_1 = [p\,(p+1)\,(g-1) / 2]$ и $\nu_2 = [(\nu_1+2) / |A_2-A_1^2|]$ степени слободе, заснована на примени следеће трансформационе формуле:

$$M \rightarrow F = \frac{M}{b} \sim F_{\alpha,\nu_1,\nu_2}, \text{ где је: } b = \frac{\nu_1}{1-A_1-\nu_1/\nu_2} \tag{3.1.18}$$

Параметри A_1 и A_2, неопходни за израчунавање b, утврђују се на следећи начин:

$$A_1 = \frac{2p^2+3p-1}{6(p+1)(g-1)}\left[\left(\sum_{k=1}^{g}\frac{1}{n_k-1}\right)-\frac{1}{\sum_{k=1}^{g}(n_k-1)}\right] \wedge A_2 = \frac{(p-1)(p+2)}{6(g-1)}\left[\left(\sum_{k=1}^{g}\frac{1}{(n_k-1)^2}\right)-\frac{1}{\sum_{k=1}^{g}(n_k-1)^2}\right]. \tag{3.1.19}$$

Представљена F апроксимација сматра се валидном и прецизнијом уколико је $A_2 - A_1^2 > 0$. У супротном, односно уколико је $A_2 - A_1^2 < 0$, израчунава се апроксимативна F статистика на основу следећег израза (*Box*, 1949, стр. 338):

$$M \rightarrow F = \frac{\nu_2 M}{\nu_1(b-M)} \sim F_{\alpha,\nu_1,\nu_2}, \text{ где је: } b = \frac{\nu_2}{1-A_1+2/\nu_2} \tag{3.1.20}$$

Независно од коришћене алтернативе при одређивању F апроксимације, H_0 се одбацује ако је реализована вредност апроксимативне F статистике већа од критичне вредности, $F_{\alpha,\nu1,\nu2}$. Изражена осетљивост *Box*-овог M теста на нарушеност претпоставке о мултиваријационој нормалности распореда зависних променљивих на нивоу посматраних популација (*Rencher*, 2002, стр. 258) може узроковати „неправедно“ одбацивање H_0 дефинисане изразом (3.1.14). У складу са објашњењем *Johnson*-а & *Wichern*-а (2007, стр. 312), *Pituch*-а & *Stevens*-а (2016, стр. 235) и *Manly*-а & *Navarro Alberto*-а (2017, стр. 61), у таквим околностима, резултати M теста могу сугерисати одбацивање H_0 услед присуства изражене анормалности података, а не стварно присутних, статистички значајних, разлика између коваријационих матрица популација и, сходно томе, иницирати сумњу у истинитост изведених закључака. Из наведених разлога, *Manly & Navarro Alberto* (2017, стр. 61) наводе и детаљно презентују одлике алтернативних робустних процедура тестирања, попут *Van Valen*-овог теста за два узорка и његове генерализације на више од два узорка, а чија се имплементација препоручује у случају одсуства мултиваријационе нормалности распореда популација.

С друге стране, у релевантној литератури, питање значаја и јачине ефеката нарушености претпоставке о хомогености коваријационих матрица популација на ефикасност *MANOVA* тестова који се користе за тестирање H_0 о једнакости вектора средина, карактерише изражена неусаглашеност ставова и варијететност одговора. Наиме, *Sharma* (1996, стр. 351), *Rencher* (2002, стр. 177; 258), *Johnson & Wichern* (2007, стр. 312), *Hair et al.* (2010, стр. 363) и *Manly & Navarro Alberto* (2017, стр. 56) истичу да присуство

умерених, статистички значајних, разлика између коваријационих матрица популација углавном има минималан утицај на резултате *MANOVA* критеријума у случају коришћења узорака једнаких или приближно једнаких величина[85].

Међутим, уколико се величине узорака изразито разликују, претходни став није одржив, а претпоставка о хомогености коваријационих матрица, према мишљењу *Kovačić*-а (1994, стр. 122), сматра се критичним предусловом за валидну примену *MANOVA*-е, због утицаја њене нарушености на ниво значајности *MANOVA* тестова, α. Дати утицај, *Rencher* (2002, стр. 177) и *Tabachnick* & *Fidell* (2013, стр. 254) елаборирају анализом вредности унутар коваријационих матрица на нивоу појединачних група, на следећи начин: ▸ уколико су веће вредности варијанси и коваријанси претежно повезане са узорцима веће величине, стварни ниво значајности (α-ниво) је потцењен, а тестови постају конзервативни, отежавајући одбацивање H_0; ▸ с друге стране, уколико су веће вредности унутар коваријационих матрица иманентне узорцима мање величине, α-ниво биће прецењен („пренадуван“), а коришћени *MANOVA* тестови постају либерални, олакшавајући одбацивање нулте хипотезе о једнакости вектора средина. У циљу избегавања наведених негативних ефеката нарушености претпоставке о хомогености коваријационих матрица *Huberty & Petoskey* (2000, стр. 196) и *Manly & Navarro Alberto* (2017, стр. 56) указују на неопходност коришћења робустних критеријума за тестирање H_0 дефинисане изразом (3.1.1) у условима коришћења узорака различитих величина. У контексту изнетих мишљења, *Coombs et al.* (1996, стр. 162;168) подржавају претходно наведену препоруку, али и проширују исту на околности када су коришћени узорци једнаких величина. Наиме, исти аутори, на основу детаљне компарације перформанси најчешће коришћених *MANOVA* критеријума и великог броја њихових робустних алтернатива, истичу да у ситуацијама када постоје оправдане сумње у одрживост претпоставке о хомогености коваријационих матрица, без обзира да ли су коришћени узорци једнаких или различитих величина, традиционалне *MANOVA* статистике (*Hotelling*-ова T^2, *Wilks*-ова Λ, *Pillai*-јева V, *Lawley-Hotelling*-ова U и *Roy*-ова θ статистика) не треба користити, а ток анализе прилагодити примени неке од робустних алтернативних статистика попут, на пример[86]: *Johansen*-овог, *Nel-van del Merwe*-овог, *Yao*-овог, *Kim*-овог, *James*-овог теста (у случају два мултиваријациона узорка), односно, *James*-овог, *Johansen*-овог, *Coombs–Algina* теста (за три и више узорака).

Коначно, ефикасност *MANOVA*-е, попут већине осталих мултиваријационих метода, може бити у значајној мери условљена величином коришћених узорака (*Hair et al.*, 2010, стр. 356). *Huberty & Petoskey* (2000, стр. 188) разматрање питања „оптималне“ величине узорака сматрају комплексном активношћу која подразумева узимање у обзир и усклађивање читавог спектра различитих фактора, као што су: репрезентативност укупног узорка, број група у саставу независне променљиве, број зависних променљивих, као и могући ефекти на вероватноћу реализације ризика грешке I (α) и II врсте (β). Генерално, иако у литератури не постоје у потпуности прецизне инструкције у домену ове

[85] *Pituch & Stevens* (2016, стр. 210) и *Hair et al.* (2010, стр. 362) под „приближно једнаким“ величинама група подразумевају ситуацију када однос узорака највеће и најмање величине није већи од 1,5, односно $n_{k\,(\max)} / n_{k\,(\min)} < 1{,}5$.

[86] Наведене робустне процедуре превазилазе оквир излагања у овом Поглављу и, сходно томе, нису детаљно елаборирани.

проблематике, кључне (опште) препоруке, које је неопходно уважити, могу се издвојити на следећи начин:

✓ Минимално прихватљив број опсервација унутар појединачних узорака мора бити већи од броја коришћених зависних променљивих (*Hair, et al.*, 2010, стр. 356; *Rencher*, 2002, стр. 256; *Tabachnick* & *Fidell*, 2013, стр. 252), при чему *Hair et al.* (2010, стр. 356) наводе 20 опсервација као препоручену минималну величину појединачних узорака.

✓ Повећање величине појединачних узорака утиче позитивно на репрезентативност узорка и јачину *MANOVA* тестова, при чему је исто, у извесној мери, условљено повећањем броја зависних променљивих.

✓ Према *Rencher*-у (2002, стр. 169) број зависних променљивих, p, не сме бити већи од израза ($n–g$), истичући дати однос као неопходан услов за валидну имплементацију *MANOVA* статистика намењених тестирању хипотезе о једнакости вектора средина.

✓ Иако коришћење већег броја зависних променљивих има позитиван утицај на репрезентативност узорка, резултирајуће повећање димензионалности анализе, генерално, има негативан утицај на јачину и робустност примењених *MANOVA* тестова (*Pituch & Stevens*, 2016, стр. 211; *Rencher*, 2002, стр. 198).

3.1.3. Тестирање статистичке значајности разлика између вектора средина две мултиваријационе популације (групе): *Hotelling*-ов T^2 тест

У овом одељку, представљен је поступак статистичке анализе две групе јединица посматрања у контексту две или више зависних променљивих истовремено, који има за циљ давање одговора на следеће питање: да ли постоји статистички сигнификантна разлика између p-димензионих вектора средина зависних променљивих на нивоу посматране две популације (групе)? Реч је о специјалном случају примене *MANOVA* метода, ограниченом на разматрање мултиваријационог истраживачког проблема који укључује само два модалитета (односно групе / третмана) у оквиру независне променљиве. Уколико се симболом $\boldsymbol{\mu}_k$ означи (p×1) вектор аритметичких средина p зависних променљивих у k-тој популацији, односно центроид k-те популације (*Huberty & Olejnik*, 2006, стр. 23), аналогно логици у основи израза (3.1.1), кореспондентна нулта и алтернативна хипотеза могу се приказати у следећем облику:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \rightarrow \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{pmatrix} \quad \wedge \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \rightarrow \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{pmatrix} \neq \begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{pmatrix}, \tag{3.1.21}$$

при чему ознака μ_{jk} представља аритметичку средину j-те зависне променљиве у k-тој групи (за j=1, 2,..., p, k = 1,…, g, и g = 2). Представљеном нултом хипотезом формално се тврди да су разлике између центроида две групе (популације) једнаке нули (*Sharma*, 1996, стр. 346), односно да су просечне вредности μ_{jk} на нивоу посматране две популације међусобно једнаке, за сваку од p, анализом обухваћених и симултано испитиваних, зависних променљивих Y_j (*Pituch & Stevens*, 2016, стр. 144).

Полазећи од два независна случајна узорка мултиваријационих опсервација (у ознаци $\mathbf{y}_{ik}$, при чему је $\mathbf{y}_{ik}' = [y_{i1k}, y_{i2k},\ldots, y_{ipk}]$), величине n_1 (са елементима $\mathbf{y}_{11}$, $\mathbf{y}_{21}$,..., $\mathbf{y}_{n_11}$) и n_2 (са елементима $\mathbf{y}_{12}$, $\mathbf{y}_{22}$,..., $\mathbf{y}_{n_22}$), извучених из p-димензионих популација које следе мултиваријациони нормалан распоред са векторима аритметичких средина $\boldsymbol{\mu}_1$ и $\boldsymbol{\mu}_2$ и једнаким, али непознатим коваријационим матрицама, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (односно, симболички: $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ и $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, респективно), адекватна мултиваријациона статистичка процедура за тестирање валидности нулте хипотезе о једнакости p-димензионих вектора аритметичких средина на нивоу две популације заснована је на примени *Hotelling*-ове T^2 статистике теста (енгл. *Hotelling's T-square statistic*), као мултиваријационе генерализације Студентовог t теста за два независна узорка. У циљу обезбеђивања потврде наведене аналогије, у наставку текста, приказан је поступак извођења *Hotelling*-ове T^2 статистике из наведеног универијационог еквивалентна, чија статистика, коришћењем симбола у Табели 3.1.1, гласи (*Lovrić*, 2009, стр. 238):

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_{\bar{y}_1 - \bar{y}_2}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\bar{y}_1 - \bar{y}_2}{\bar{s}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}. \tag{3.1.22}$$

У представљеном изразу, коришћени симболи означавају:

n_1, n_2 – величине посматрана два узорка;

s_1^2, s_2^2 – варијансе посматрана два узорка;

$\bar{y}_1, \bar{y}_2$ – аритметичке средине промењиве Y на нивоу два узорка (групе);

$s_{\bar{y}_1 - \bar{y}_2}$ – оцена стандардне грешке разлике аритметичких средина два узорка;

$\bar{s}^2$ – пондерисана (здружена) оцена заједничке варијансе два скупа $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$.

Поступак мултиваријационе генерализације започиње квадрирањем обе стране статистике дате претходним изразом. Сходно наведеном, добија се (*Pituch & Stevens*, 2016, стр. 145):

$$t^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\bar{s}^2(\frac{1}{n_1}+\frac{1}{n_2})} = (\bar{y}_1 - \bar{y}_2)\left[\bar{s}^2(\frac{n_1+n_2}{n_1 n_2})\right]^{-1}(\bar{y}_1 - \bar{y}_2) = \frac{n_1 n_2}{n_1+n_2}(\bar{y}_1 - \bar{y}_2)(\bar{s}^2)^{-1}(\bar{y}_1 - \bar{y}_2). \tag{3.1.23}$$

Представљени образац може се генерализовати за случај p зависних променљивих заменом одговарајућих скалара кореспондентним векторима и матрицама, а резултирајући израз представља *Hotelling*-ову T^2 статистику теста[87] (*Mason & Young*, 2011, стр. 639):

$$t^2 = \frac{n_1 n_2}{n_1+n_2}\underbrace{(\bar{y}_1 - \bar{y}_2)}_{\downarrow (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}\ \underbrace{(\bar{s}^2)^{-1}}_{\downarrow \bar{\mathbf{S}}^{-1}}\ \underbrace{(\bar{y}_1 - \bar{y}_2)}_{\downarrow (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'} \Bigg\} \rightarrow T^2 = \frac{n_1 n_2}{n_1+n_2}\left[(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\,\bar{\mathbf{S}}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\right], \tag{3.1.24}$$

где $\bar{\mathbf{y}}_1$ и $\bar{\mathbf{y}}_2$ означавају $(p\times 1)$ векторе аритметичких средина p зависних променљивих на нивоу посматрана два узорка (групе), који се израчунавају коришћењем израза (3.1.5), а

[87] Статистика T^2 и њен узорачки распоред представљени су иницијално од стране *Harold*-а *Hotelling*-а (1931) и уобичајено се назива *Hotelling*-ова T^2 статистика. Прва званична употреба ове статистике везује се за тестирање нулте хипотезе о једнакости вектора средина, $\boldsymbol{\mu}$, мултиваријационе нормалне популације и хипотетичког (специфичног) вектора средина $\boldsymbol{\mu}_0$, симболички H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ (*Mason & Young*, 2011, стр. 639).

матрица $\bar{\mathbf{S}}$ (или $\mathbf{S}_{pooled}$) представља непристрасну, пондерисану (енгл. *pooled*) оцену (претпостављене) заједничке коваријационе матрице посматрана два скупа, $\boldsymbol{\Sigma}$ (израз (3.1.16)).

Под претпоставком да је H_0 (израз (3.1.21)) истинита, а полазне претпоставке задовољене, *Hotelling*-ова T^2 статистика, дата изразом (3.1.24), има *Hotelling*-ов T^2 распоред[88], који је у потпуности одређен параметром p (број зависних променљивих) и бројем степени слободе $\nu = n_1 + n_2 - 2$ (*Rencher*, 2002, стр. 122–123).

Кључна својства *Hotelling*-ове T^2 статистике за два независна узорка, могу се представити у форми следеће листе:

✓ Број степени слободе карактеристичан за T^2 статистику, идентичан је броју степени слободе кореспондентне униваријационе Студентове t статистике теста (израз 3.1.22), односно $v = n_1 + n_2 - 2$ (*Rencher*, 2002, стр. 123);

✓ Израчунавање вредности T^2 статистике условљено је претпоставком да број степени слободе v мора бити већи од броја зависних променљивих, односно, $n_1+n_2-2 > p$. У случају да наведена претпоставка није задовољена, матрица $\mathbf{S}_{pooled}$ била би јединична матрица, што би онемогућило реализацију израза (3.1.24), (*Rencher*, 2002, стр. 123);

✓ Велике вредности T^2 статистике сугеришу да H_0 није тачна и да је треба одбацити, док мале вредности не обезбеђују довољно емпиријских доказа за одбацивање H_0 (*Manly & Navarro Alberto*, 2017, стр. 56). Наиме, што је већа вредност T^2 статистике, то су уверљивији докази против H_0, и обрнуто (*Rencher*, 2002, стр. 123). *Sharma* (1996, стр. 346) и *Huberty & Olejnik* (2006, стр. 40) објашњавају наведено својство кроз повезаност квадрата *Mahalanobis*-овог одстојања, у ознаци MD^2, са изразом за T^2 статистику:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left[(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\,\bar{\mathbf{S}}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\right] \rightarrow T^2 = \frac{n_1 n_2}{n_1 + n_2} MD^2. \qquad (3.1.25)$$

Наиме, део израза (3.1.25) у угластим заградама, назива се квадрат *Mahalanobis*-овог одстојања и представља стандардизовану квадрирану меру одстојања између вектора средина $\bar{\mathbf{y}}_1$ и $\bar{\mathbf{y}}_2$ посматрана два узорка. Јасно је да са повећањем удаљености између центроида две групе долази до пропорционалног повећања *Hotelling*-ове T^2 статистике.

✓ Иако је алтернативна хипотеза двосмерна (H_1: $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$), критична област одбацивања H_0 налази се искључиво на десној страни T^2 распореда ($T^2 > T^2_{\alpha,\, p,\, n_1 + n_2 - 2}$), тако да је T^2 тест за два мултиваријациона узорка, генерално једносмеран тест[89] (*Rencher*, 2002, стр. 124);

✓ Статистика T^2 теста је инваријантна за све несингуларне линеарне трансформације мултиваријационих опсервација ($\mathbf{X}$), облика: $Z = \mathbf{A'X} + b$, уколико су задовољене полазне статистичке претпоставке за примену теста. Прецизније, у контексту наведеног, *Mason & Young* (2011, стр. 639) наводе да су резултати T^2 теста независни од промене мерне скале у којима се исказују подаци, односно да је реализована вредност T^2

[88] T^2 распоред је асиметричан (удесно) непрекидни теоријски распоред вероватноћа, попут F распореда, и дефинисан је за све реалне вредности T^2 статистике од 0 до $+\infty$ (*Rencher*, 2002, стр. 123).

[89] Иако нулта и алтернативна хипотеза не сугеришу примену једносмерног теста, представљено својство T^2 теста је типично за многе мултиваријационе тестове статистичке значајности (*Rencher*, 2002, стр. 124), будући да се у мултиваријационом простору не разматрају једносмерне алтернативне хипотезе (детаљније у: *Lovrić*, 2009, стр. 266).

статистике заснована на вредностима новоформиране променљиве Z, идентична вредности T^2 статистике утврђеној на бази оригиналних вредности мултиваријационих опсервација **X**. Поред наведеног, исти аутори (2011, стр. 639) истичу да се, у поређењу са свим осталим инваријантним тестовима, статистички тест заснован на T^2 статистици одликује највећом јачином теста, односно вероватноћом детектовања и одбацивања нулте хипотезе која није истинита, за фиксни ниво значајности α.

Доношење коначне одлуке о одбацивању или неодбацивању H_0 дефинисане изразом (3.1.21), на основу класичног, *Neyman-Pearson*-овог приступа заснива се на поређењу реализоване вредности T^2 статистике (израз 3.1.24) са одговарајућим вредностима из таблице критичних вредности *Hotelling*-овог T^2 распореда, у ознаци $T^2_{\alpha,\ p,\ n_1\ +\ n_2\ -\ 2}$, и примени правила одлучивања које гласи: уколико је $T^2 > T^2_{\alpha, p, n_1 + n_2 - 2} \rightarrow H_0$ се одбацује и усваја алтернативна, H_1. У супротном, може се констатовати, уз ниво значајности α, да не постоји довољно аргумената за одбацивање H_0.

Примена „савременог“ приступа, заснованог на одређивању p-вредности, у поступку одлучивања о валидности H_0, омогућена је трансформацијом *Hotelling*-ове T^2 статистике у F статистику, која следи егзактни *Snedecor*-ов F распоред са $v_1=p$ и $v_2=n_1+n_2-p-1$ степени слободе, уколико је H_0 истинита, а полазне претпоставке о независности мултиваријационих опсервација, мултиваријационој нормалности популација и хомогености коваријационих матрица популација задовољене (*Huberty & Olejnik*, 2006, стр. 40), путем следеће формуле (*Mason & Young*, 2011, стр. 639):

$$T^2 \sim T^2_{\alpha, p, n_1+n_2-2} \rightarrow F = \frac{(n_1+n_2-p-1)}{(n_1+n_2-2)p} T^2 \sim F_{(p,\ n_1+n_2-p-1)} \tag{3.1.26}$$

Коначно, ако је реализовани ниво значајности (односно, p-вредност), утврђен за израчунату вредност F статистике, мањи од дефинисаног нивоа значајности α, H_0 се одбацује и прихвата алтернативна хипотеза која гласи: постоји статистички значајна разлика између центроида посматране две популације (групе).

3.1.4. Тестирање статистичке значајности разлика између вектора средина три или више мултиваријационих популација (група): *MANOVA* тестови

За разлику од претходно представљеног специјалног случаја *MANOVA* анализе заснованог на примени *Hotelling*-ове T^2 статистике, при спровођењу поступка тестирања нулте хипотезе о једнакости вектора аритметичких средина на нивоу три или више популација (израз (3.1.1)), односно када је анализом обухваћено више од два модалитета независне променљиве, ситуација се компликује чињеницом да тада не постоји јединствена статистика теста која се одликује највећом јачином у контексту детекције статистички значајних одступања од претпоставке садржане у оквиру H_0, већ су на располагању следеће четири, углавном коришћене, алтернативне статистике:

- ✓ *Wilks*-ова Λ статистика (енгл. *Wilks' lambda statistic*);
- ✓ *Pillai*-јева V статистика (енгл. *Pillai's trace statistic*);
- ✓ *Lawley-Hotelling*-ова U статистика (енгл. *Lawley-Hotelling trace statistic*); и
- ✓ *Roy*-ова θ статистика (енгл. *Roy's largest root statistic*);

На основама матрице мултиваријационих података (Табела 3.1.1), изведених компоненти оцењеног једнофакторског *MANOVA* модела у одељку 3.1.1 и статистичких претпоставки на којима почива примена овог мултиваријационог метода, дискусија у наставку излагања усмерена је на елаборирање кључних карактеристика, поступака израчунавања и компарације наведених *MANOVA* статистика.

***Wilks*-ова статистика теста (Λ)**

Предложена од стране *Wilks*-а (1932), за тестирање валидности *MANOVA* нулте хипотезе, која гласи H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_g$ (израз 3.1.1), најчешће се користи „најстарија“ међу наведеним статистикама, *Wilks*-ова Λ статистика (позната и под називом *Wilks*-ова *U* статистика), дефинисана следећим количником (*Johnson & Wichern*, 2007, стр. 303):

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{\left|\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k\right)\left(\mathbf{y}_{ik}-\bar{\mathbf{y}}_k\right)'\right|}{\left|\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}}\right)\left(\mathbf{y}_{ik}-\bar{\bar{\mathbf{y}}}\right)'\right|}. \qquad (3.1.27)$$

У суштини, датим изразом пореди се варијабилитет унутар посматраних узорака са укупним варијабилитетом, будући да $|\mathbf{W}|$ представља детерминанту матрице суме квадрата и узајамних производа одступања унутар група (израз (3.1.11)), $|\mathbf{T}|$ детерминанту матрице укупне суме квадрата и узајамних производа одступања мултиваријационих опсервација од опште средине, а **B** матрицу суме квадрата и узајамних производа одступања између група (израз (3.1.10)), при чему, на основу оцењеног *MANOVA* модела (израз (3.1.9)), важи релација $\mathbf{T} = \mathbf{B} + \mathbf{W}$, односно $|\mathbf{T}| = |\mathbf{B} + \mathbf{W}|$. Представљена *Wilks*-ова Λ статистика узима вредности у интервалу $0 \leq \Lambda \leq 1$ и под претпоставком да је H_0 истинита, а полазне претпоставке задовољене, следи *Wilks* Λ распоред[90] који је у потпуности одређен следећим параметрима (*Rencher*, 2002, стр. 161): ▸ (p) број зависних променљивих, ▸ (v_1=g–1) број степени слободе за факторски варијабилитет и, ▸ (v_2=n–g) број степени слободе за резидуални варијабилитет.

За дефинисани ризик грешке I врсте α, H_0 се одбацује уколико је реализована вредност статистике теста (Λ) мања или једнака од критичне вредности теста, у ознаци $\Lambda_{\alpha, p, v_1, v_2}$. Симболички, правило за одлучивање о валидности нулте хипотезе гласи: $\Lambda \leq \Lambda_{\alpha, p, v_1, v_2} \rightarrow H_1$ се прихвата. Консеквентно, може се констатовати да је тест заснован на *Wilks* Λ статистици инверзан тест, будући да се одбацивање H_0 спроводи за мале вредности Λ (*Pituch & Stevens*, 2016, стр. 178). Другим речима, уколико су вектори узорачких средина једнаки, тада би матрица суме квадрата и узајамних производа одступања између узорака била заправо нулта матрица (симболички: $\mathbf{B} = \mathbf{0}$), а резултирајућа вредност Λ била би једнака јединици, односно $\Lambda = |\mathbf{W}| / |\mathbf{B}+\mathbf{W}| = |\mathbf{W}| / |\mathbf{0}+\mathbf{W}| = 1$, обезбеђујући тиме јаке доказе у прилог претпоставке садржане у H_0 (*Rencher*, 2002, стр. 162). Супротно, уколико је

[90] Таблица критичних вредности *Wilks* Λ распореда (*Kres*, 1983, стр. 20) одликује се ограниченим обухватом, дефинисаним за следеће вредности кореспондентних параметара (*Kres*, 1983, стр. 15):
- за α = 0,01 и α = 0,05;
- за p = 1(1)8, односно p = 1, 2,..., 8, уз услов да је $p \leq v_2$; *Rencher* (2002, стр. 162) истиче важност наведеног услова у контексту обезбеђивања позитивних вредности детерминанти матрица у изразу за израчунавање Λ статистике;
- за v_1 (g–1) = 1(1)15, 18(3)30, 40(20)120; и
- за v_2 (n–g) = 1(1)30, 40(20)140, 170, 200, 240, 320, 440, 600, 800, 1000, +∞.

варијабилитет унутар узорака (**W**) мање изражен у поређењу са варијабилитетом присутним између узорака (**B**), вредност статистике Λ се приближава нули, сугеришући тиме да постоје јаки докази против H_0, односно докази у прилог алтернативне тврдње о присуству статистички значајних разлика између вектора средина посматраних g популација (група). Такође, при примени овог теста, важно је знати да се критичне вредности *Wilks* Λ распореда ($\Lambda_{\alpha,\ p,\ v1,\ v2}$) смањују са повећањем броја зависних променљивих које су обухваћене анализом. У том контексту, *Rencher* (2002, стр. 162) упозорава да се укључивање нових зависних променљивих може негативно одразити на јачину *Wilks*-овог теста, осим уколико исте у значајној мери не доприносе интензивирању разлика између вектора узорачких средина посматраних група и, сходно томе, јачању доказа за одбацивање H_0. Међутим, будући да се коришћење узорачког распореда вредности *Wilks*-ове Λ статистике може, у извесној мери, сматрати компликованим, из угла доступности кореспондентне таблице критичних вредности и њеног ограниченог обухвата, употреба одговарајућих апроксимација Λ статистике се често у пракси сматра неопходном (*Pituch & Stevens*, 2016, стр. 179). У том контексту, истраживачима су на располагању две апроксимације Λ статистике: 1) *Bartlett*-ова χ^2 и 2) *Rao*-ова F апроксимација.

Пре елаборирања поступка извођења наведених апроксимација, неопходно је указати на два специјална случаја, у којима је могуће извршити трансформацију статистике Λ у егзактну F статистику која следи *Snedecor*-ов F распоред са v_1 и v_2 степени слободе. Услови под којима се поменута трансформација спроводи у мултиваријационом контексту, трансформационе формуле и начин одређивања броја степени слободе v_1 и v_2, представљени су у Табели 3.1.3. За разлику од поступка одлучивања о валидности H_0 на бази Λ статистике, за извршену трансформацију правило одлучивања гласи: $F > F_{v1,\ v2} \rightarrow H_0$ се одбацује. Важно је уочити да се број степени слободе, v_1 и v_2, засебно одређује за сваку од представљених ситуација и да су условљене коришћеним вредностима параметара *Wilks* Λ распореда, p и v_1.

Табела 3.1.3. Трансформација Λ статистике у егзактну F статистику (специјални случајеви)

Вредности параметара *Wilks* Λ распореда		Формуле за трансформацију
p	$v_1 = g - 1$	($\Lambda \rightarrow$ егзактна F статистика)
$p = 2$	$v_1 \geq 2$ односно ($g \geq 3$)	$\Lambda \rightarrow F = \left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)\left(\frac{n-g-1}{g-1}\right) \sim F_{\underbrace{2(g-1)}_{v_1},\ \underbrace{2(n-g-1)}_{v_2}}$
$p \geq 2$	$v_1 = 2$ односно ($g = 3$)	$\Lambda \rightarrow F = \left(\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\right)\left(\frac{n-p-2}{p}\right) \sim F_{\underbrace{2p}_{v_1},\ \underbrace{2(n-p-2)}_{v_2}}$

Напомена: симбол p означава број зависних променљивих, а g број узорака (група) у *MANOVA* анализи.
Извор: Ауторов приказ (адаптирано према: *Rencher*, 2002, стр. 163; *Johnson & Wichern*, 2007, стр. 303)

Трансформација Λ статистике у апроксимативну F статистику, односно статистику која приближно (апроксимативно) следи *Snedecor*-ов F распоред са v_1 и v_2 степени слободе, нарочито погодна у случајевима који нису обухваћени Табелом 3.1.3 и када су коришћени

узорци релативно мале величине, спроводи се путем следеће формуле (*Manly & Navarro Alberto*, 2017, стр. 69):

$$\Lambda \sim \Lambda_{\alpha,p,(g-1),(n-g)} \rightarrow F = \left(\frac{1-\sqrt[t]{\Lambda}}{\sqrt[t]{\Lambda}} \right) \left(\frac{\nu_2}{\nu_1} \right) \sim (\text{приближно}) F_{\nu_1,\, \nu_2}\,, \tag{3.1.28}$$

где је $v_1 = p(g-1)$, $v_2 = wt - ½\, v_1 + 1$, за $w = n - 1 - ½\,(p+g)$ и $t = [\{p^2(g-1)^2 - 4\} / \{p^2 + (g-1)^2 - 5\}]^{½}$. Апроксимативна вредност F статистике, дата претходним изразом, своди се на егзактну вредност F статистике из Табеле 3.1.3, уколико је $p = 2$, или $g = 3$ (*Rencher*, 2002, стр. 163). Поред дате *Rao*-ове F апроксимације, у ситуацијама када су узорци средње и велике величине, модификација Λ статистике заснована на употреби *Bartlett*-ове χ^2 апроксимације такође може бити коришћена за тестирање *MANOVA* нулте хипотезе. *Bartlett*-ова χ^2 апроксимација *Wilks*-ове Λ статистике дата је следећим изразом (*Kovačić*, 1994, стр. 119):

$$\Lambda \sim \Lambda_{\alpha,p,(g-1),(n-g)} \rightarrow \chi^2 = -\left(n - 1 - \frac{p+g}{2} \right) \ln \Lambda \sim (\text{приближно}) \chi^2_{\alpha,\, p(g-1)}\,. \tag{3.1.29}$$

Уколико је H_0 истинита, представљена трансформисана вредност апроксимативно следи χ^2 распоред са $p(g–1)$ степени слободе. Консеквентно, за велике узорке и дефинисани ниво значајности теста α, нулта хипотеза се одбацује ако је израчуната вредност χ^2 апроксимације већа од одговарајуће критичне вредности χ^2 распореда са $p(g–1)$ степени слободе, у ознаци $\chi^2_{\alpha,\, p(g-1)}$ (*Johnson & Wichern*, 2007, стр. 304).

***Pillai*-јева статистика теста (V)**

Наредна мултиваријациона статистика теста која се уобичајено користи за тестирање нулте хипотезе у *MANOVA* анализи јесте *Pillai*-јева статистика V, заснована на трагу (енгл. *trace*) матрице $\mathbf{BT^{-1}}$, односно (*Everitt*, 2010, стр. 271; *Rencher*, 2002, стр. 166):

$$\mathrm{V} = trace\left[\frac{\mathbf{B}}{\mathbf{B}+\mathbf{W}} \right] = \mathrm{tr}\left[\frac{\mathbf{B}}{\mathbf{T}} \right] = \mathrm{tr}\left[\mathbf{B}\mathbf{T}^{-1} \right]. \tag{3.1.30}$$

Нулта хипотеза, H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g$, се одбацује уколико је реализована вредност статистике V једнака или већа од критичне вредности теста, у ознаци $\mathrm{V}_{\alpha,s,m,N}$, односно симболички: $\mathrm{V} \geq \mathrm{V}_{\alpha,s,m,N} \rightarrow H_1$ се прихвата. Параметри распореда *Pillai*-јеве статистике V, неопходни за одређивање критичне вредности $\mathrm{V}_{\alpha,s,m,N}$, дефинисани су на следећи начин (*Kres*, 1983, стр. 136): $s = min(g–1,\, p)$; $m = (|(g–1)–p| – 1) / 2$; и $N = (n–g–p–1) / 2$. Насупрот Λ статистици, високе вредности *Pillai*-јеве статистике обезбеђују јаче доказе у прилог тврдње да посматрани мултиваријациони узорци потичу из популација које се одликују различитим векторима средина (*Manly & Navarro Alberto*, 2017, стр. 68). За вредности параметара α, s, m, N, које нису обухваћене таблицом критичних вредности $\mathrm{V}_{\alpha,s,m,N}$ (*Kres*, 1983, стр. 140), спроводи се трансформација *Pillai*-јеве статистике у апроксимативну F статистику, коришћењем једне од следеће три, алтернативне, формуле (*Rencher*, 2002, стр. 166–167):

$$\mathrm{V} \sim \mathrm{V}_{\alpha,s,m,N} \rightarrow F_1 = \frac{(2N+s+1)\mathrm{V}}{(2m+s+1)(s-\mathrm{V})} \sim (\text{приближно}) F_{\underbrace{s(2m+s+1)}_{\nu_1},\, \underbrace{s(2N+s+1)}_{\nu_2}}\,, \tag{3.1.31a}$$

$$\mathrm{V} \sim \mathrm{V}_{\alpha,s,m,N} \rightarrow F_2 = \frac{s\left[(n-g)-(g-1)+s\right]\mathrm{V}}{p(g-1)(s-\mathrm{V})} \sim (\text{приближно}) F_{\underbrace{p(g-1)}_{\nu_1},\ \underbrace{s[(n-g)-(g-1)+s]}_{\nu_2}}, \quad (3.1.31б)$$

$$\mathrm{V} \sim \mathrm{V}_{\alpha,s,m,N} \rightarrow F_3 = \frac{(n-g-p+s)\mathrm{V}}{d(s-\mathrm{V})} \sim (\text{приближно}) F_{\underbrace{sd}_{\nu_1},\ \underbrace{s(n-g-p+s)}_{\nu_2}}, \text{ где је } d = max(p, g-1). \quad (3.1.31в)$$

Значај наведених F апроксимација огледа се и у чињеници да њихова употреба омогућава израчунавање реализованог нивоа значајности (*Everitt*, 2010, стр. 271) и, сходно томе, одлучивање о валидности H_0 на бази поређења p-вредности и ризика грешке I врсте α.

***Lawley–Hotelling*-ова статистика теста (*U*)**

Трећа статистика теста, која се користи у оквиру *MANOVA*-е за тестирање нулте хипотезе о једнакости вектора аритметичких средина посматраних g популација, јесте *Lawley–Hotelling*-ова U статистика[91], дефинисана следећим изразом (*Johnson & Wichern*, 2007, стр. 336):

$$U = trace\left[\frac{\mathbf{B}}{\mathbf{W}}\right] = \mathrm{tr}\left[\mathbf{B}\mathbf{W}^{-1}\right]. \quad (3.1.32)$$

У суштини, слично поступку израчунавања *Pillai*-јеве V статистике, представљеним изразом одређује се траг матрице $\mathbf{BW}^{-1}$ добијене као резултат односа факторског (**B**) и резидуалног варијабилитета (**W**), дефинисаним у оквиру оцењеног *MANOVA* модела (израз 3.1.9). Будући да матрица $\mathbf{BW}^{-1}$ представља мултиваријациони аналог односа мере варијабилитета између и унутар узорака, који је у основи израчунавања статистике F теста у *ANOVA* анализи (израз (3.1.33)), *Pituch & Stevens* (2016, стр. 210) истичу да *Lawley–Hotelling*-ова мултиваријациона U статистика представља природну генерализацију униваријационе F статистике.

$$F = \frac{SSE_b}{SSE_w} = \frac{\left[\sum_{k=1}^{g} n_k \left(\bar{y}_k - \bar{\bar{y}}\right)^2\right]/(g-1)}{\left[\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(y_{ik} - \bar{y}_k\right)^2\right]/(n-g)} = \frac{(n-g)}{(g-1)} \frac{\overbrace{\sum_{k=1}^{g} n_k \left(\bar{y}_k - \bar{\bar{y}}\right)^2}^{\text{факторски варијабилитет}}}{\underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(y_{ik} - \bar{y}_k\right)^2}_{\text{резидуални варијабилитет}}}. \quad (3.1.33)$$

Нулта хипотеза се одбацује за високе вредности *Lawley–Hotelling* статистике, односно уколико је реализована, а коригована[92], вредност U статистике већа од одговарајуће табличне вредности дефинисане следећим параметрима: α, p, $\nu_1 = g-1$ и $\nu_2 =$

[91] Због директне повезаности која постоји између *Hotelling*-ове T^2 статистике (израз (3.1.24)) и U статистике (израз (3.1.32)) у случају када су анализом обухваћене само две групе (тзв. специјални случај примене *MANOVA*-е), *Lawley-Hotelling*-ова статистика се често назива и *Hotelling*-ова генерализована T^2 статистика (*Rencher*, 2002, стр. 170), (видети израз (3.1.46)).

[92] *Rencher* (2002, стр. 167) истиче да је доношење одлуке о валидности H_0 на бази критичних вредности *Lawley–Hotelling* теста условљено израчунавањем прилагођене вредности U статистике, коришћењем следећег израза:

$$U^* = \frac{\nu_2}{\nu_1} U = \frac{n-g}{g-1} U \quad (3.1.34)$$

Исти аутор (2002, стр. 167), анализирајући обухват и однос између вредности параметара за које су дефинисани елементи у Таблици критичних вредности разматраног теста, напомиње да је, при одређивању табличних вредности, неопходно, уместо уобичајеног редоследа параметара (p, ν_1, ν_2) користити следећи измењени редослед (ν_1, p, $\nu_2 + \nu_1 - p$), у случајевима када је $p > \nu_1$, будући да је однос обухваћених вредности датих параметара у таблици: $p \leq \nu_1$ и $p \leq \nu_2$.

n–g. За доношење одлуке о валидности H_0, у случају када вредности параметара α, p, ν_1 и ν_2 нису обухваћене таблицом критичних вредности, користи се апроксимативна вредност F статистике, која приближно следи F распоред са ν_1 и ν_2 степени слободе, применом једне од следеће три, алтернативне, трансформационе формуле (*Rencher*, 2002, стр. 167–168):

$$U \to F_1 = U \frac{[\frac{(n-p-2)(n-g-1)}{(n-g-p-3)(n-g-p)} - 3](n-g-p-1)}{p(g-1)\frac{6+pg-p}{[\frac{(n-p-2)(n-g-1)}{(n-g-p-3)(n-g-p)} - 3]}} \sim (\text{приближно}) F_{\underbrace{p(g-1)}_{\nu_1},\ \underbrace{\frac{(6+pg-p)}{\left[\frac{(n-p-2)(n-g-1)}{(n-g-p-3)(n-g-p)}-1\right]}}_{\nu_2}}, \quad (3.1.35а)$$

$$U \to F_2 = \frac{2(sN+1)U}{s^2(2m+s+1)} \sim (\text{приближно}) F_{\underbrace{s(2m+s+1)}_{\nu_1},\ \underbrace{2(sN+1)}_{\nu_2}}, \quad (3.1.35б)$$

$$U \to F_3 = \frac{[s(n-2g)+2]U}{sp(g-1)} \sim (\text{приближно}) F_{\underbrace{p(g-1)}_{\nu_1},\ \underbrace{s(n-2g)}_{\nu_2}}, \quad (3.1.35в)$$

где је s=*min*(g–1, p), m=(|(g–1)–p| –1) / 2, и N=(n–g–p–1) / 2. Ако је $p \le (g-1)$ тада је $F_3 \equiv F_2$.

***Roy*-ова статистика теста (θ)**

Логика у основи извођења и примене *Roy*-ове статистика теста огледа се у проналажењу линеарне комбинације p зависних променљивих, облика: $Z = a_1Y_1 + a_2Y_2 + \ldots + a_pY_p = \mathbf{a'Y}$, која максимизира рацио факторског (**B**) и резидуалног варијабилитета (**W**) анализом обухваћених група мултиваријационих опсервација. Формулисан у контексту вредности променљиве Z, односно $z_{ik} = a_1y_{1k} + a_2y_{2k} + \ldots + a_py_{pk} = \mathbf{a'y}_{ik}$, наведени проблем подразумева идентификовање вектора тежинских коефицијената $\mathbf{a'} = [a_1, a_2, \ldots, a_p]$ за који се остварује максимизирање универијационог F односа који, модификацијом израза (3.1.33), гласи:

$$F_{(z)} = \frac{(SSE_b)_z}{(SSE_w)_z} = \frac{\left[\sum_{k=1}^{g} n_k\left(\overline{z}_k - \overline{\overline{z}}\right)^2\right]/(g-1)}{\left[\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(z_{ik} - \overline{z}_k\right)^2\right]/(n-g)} = \frac{(n-g)}{(g-1)} \frac{\overbrace{\sum_{k=1}^{g} n_k\left(\overline{z}_k - \overline{\overline{z}}\right)^2}^{\text{факторски варијабилитет}}}{\underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(z_{ik} - \overline{z}_k\right)^2}_{\text{резидуални варијабилитет}}}. \quad (3.1.36)$$

Будући да матрице **W** и **B** представљају мултиваријациону генерализацију универијационих сума квадрата одступања између и унутар група, а матрица $\mathbf{BW}^{-1}$ мултиваријациони аналог универијационог односа наведених компоненти укупног варијабилитета, претходни израз може се, коришћењем вектора коефицијената **a**, записати као (*Rencher*, 2002, стр. 165):

$$F_{(z)} = \frac{(n-g)}{(g-1)} \frac{\mathbf{a'Ba}}{\mathbf{a'Wa}}. \quad (3.1.37)$$

Проблем одређивања вектора **a**, који обезбеђује максимизирање датог односа за све могуће линеарне комбинације Z, своди се на проналажење карактеристичних вредности (енгл. *eigenvalue*) λ од несиметричне, ($p \times p$) матрице $\mathbf{BW}^{-1}$ (израз (3.1.38а)), односно решавање карактеристичне једначине p-тог реда по λ, дате изразом (3.1.38б).

$$\left|\mathbf{BW}^{-1}-\lambda\mathbf{I}\right|=\left|\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}-\lambda\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}\right|=\left|\begin{bmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp}-\lambda \end{bmatrix}\right|=0 \qquad (3.1.38а)$$

$$\left|\mathbf{BW}^{-1}-\lambda\mathbf{I}\right|=(-1)^{p}\left[\lambda^{p}+c_{1}\lambda^{p-1}+c_{2}\lambda^{p-2}+\ldots+c_{p}\lambda+c_{p+1}\right]=0 \qquad (3.1.38б)$$

Израз (3.1.38а) представља детерминанту матрице [$\mathbf{BW}^{-1}-\lambda\mathbf{I}$], која након сређивања поприма форму карактеристичног полинома реда p (израз (3.1.38б)). Укупан број могућих решења за формирани полином p-тог реда једнак је, генерално, броју зависних променљивих p, при чему ће нека од њих, у математичком смислу, бити тривијална решења, односно $\lambda = 0$. *Klecka* (1980, стр. 34) наводи да ће, када је $p > (g–1)$, увек бити $(p–g+1)$ решења за која је $\lambda = 0$, односно $(g–1–p)$ таквих решења у случају да је $p < (g–1)$. У том контексту, максималан број позитивних карактеристичних вредности λ који може бити одређен, увек износи: $s = min(p, g–1) = p – ((p–g+1)$ или $(g–1–p))$. Свакој од утврђених карактеристичних вредности λ_r (за $r = 1,2,\ldots, s$, и $s = min(p, g–1)$), при чему је $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_s > 0$, одговара по један карактеристични вектор (енгл. *eigenvector*) $\mathbf{a}_r$, на бази којих се може креирати s линеарних комбинација облика: $Z_r = \mathbf{a}_r'\mathbf{Y}$. Међутим, максимизирање односа представљеног изразом (3.1.36), односно изразом (3.1.37), постиже се само линеарном комбинацијом Z_1, заснованој на употреби вектора $\mathbf{a}_1$, који представља карактеристични вектор придружен највећој од свих s карактеристичних вредности матрице $\mathbf{BW}^{-1}$, у ознаци λ_1, односно (*Rencher*, 2002, стр. 165):

$$\underbrace{\max_{\mathbf{a}} F_{(z)}}=\frac{(n-g)}{(g-1)}\frac{\mathbf{a}_1'\mathbf{B}\mathbf{a}_1}{\mathbf{a}_1'\mathbf{W}\mathbf{a}_1}=\frac{(n-g)}{(g-1)}\lambda_1. \qquad (3.1.39)$$

Полазећи од изнетих релација, карактеристична вредност λ_1 може се, у контексту вредности формиране линеарне комбинације Z_1, изразити као:

$$\lambda_1=\frac{\sum_{k=1}^{g} n_k\left(\overline{z}_k-\overline{\overline{z}}\right)^2}{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(z_{ik}-\overline{z}_k\right)^2}, \text{ где је } z_{ik} = \mathbf{a}_1'\mathbf{y}_{ik}. \qquad (3.1.40)$$

На основу аналогије са универијационом F статистиком (израз 3.1.33), представљена максимална вредност карактеристичног корена λ_1 може се употребити као критеријум за тестирање претпоставке да је варијабилитет између узорака сигнификантно висок, односно да посматрани узорци не потичу из популација које се одликују једнаким векторима средина (*Manly & Navarro Alberto*, 2017, стр. 68). У том контексту, мултиваријациони статистички тест који се користи за тестирање *MANOVA* нулте хипотезе, H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 =... = \boldsymbol{\mu}_g$, а чија је статистика заснована на употреби највеће карактеристичне вредности λ_1 матрице $\mathbf{BW}^{-1}$, назива се *Roy*-ов тест највећег (карактеристичног) корена (енгл. *Roy's largest root test*). У релевантној литератури, али и комерцијалним статистичким програмским пакетима, разликују се и равноправно примењују следећа два приступа у израчунавању статистике *Roy*-овог теста, θ. Према једном приступу, коришћеном од стране већине аутора (попут: *Kovačić*, 1994; *Johnson &*

Wichern, 2007; *Izenman*, 2008; *Everitt*, 2010; *Pituch & Stevens*, 2016; *Manly & Navarro Alberto*, 2017), израчунавање статистике θ заснива на њеном изједначавању са λ_1, односно: $\theta = \lambda_1$. С друге стране, *Kres* (1983, стр. 52), *Rencher* (2002, стр. 165) и *Manly & Navarro Alberto* (2017, стр. 68) указују на алтернативни приступ израчунавању статистике θ, путем израза:

$$\theta = \frac{\lambda_1}{1+\lambda_1} \sim \theta_{\alpha,s,m,N}\,. \tag{3.1.41}$$

Нулта хипотеза, H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_g$, се одбацује уколико је реализована вредност статистике θ већа од критичне вредности теста, у ознаци $\theta_{\alpha,s,m,N}$, симболички: $\theta > \theta_{\alpha,s,m,N} \rightarrow H_1$ се прихвата. Параметри распореда *Roy*-ове θ статистике, неопходни за одређивање критичне вредности $\theta_{\alpha,s,m,N}$, дефинисани су на следећи начин (*Kres*, 1983, стр. 53): $s=min(p, g–1)$; $m=(|(g–1)–p|–1)/2$; и $N=(n–g–p–1)/2$. За вредности параметара које нису обухваћене таблицом критичних вредности $\theta_{\alpha,s,m,N}$ (*Kres*, 1983, стр. 56), трансформација *Roy*-ове θ статистике у апроксимативну F статистику, може се спровести путем следеће формуле (*Rencher*, 2002, стр. 165; *Manly & Navarro Alberto*, 2017, стр. 69):

$$\theta \sim \theta_{\alpha,s,m,N} \rightarrow F = \frac{n-g-d-1}{d}\lambda_1 \sim (\text{приближно}) F_{\underbrace{d}_{\nu_1},\ \underbrace{n-g-d-1}_{\nu_2}},\ \text{где је } d = max(p, g–1). \tag{3.1.42}$$

Међутим, за приказану трансформацију θ статистике, *Rencher* (2002, стр. 165) истиче да иста не представља у потпуности задовољавајућу F апроксимацију јер је резултирајућа вредност F (израз 3.1.42) увек већа од „стварне“ F вредности, односно $F > F_{d,\ n-g-d-1}$, што може имати негативне ефекте на јачину теста и, сходно томе, валидност формулисаних закључака, нарочито уколико је донета одлука о одбацивању H_0.

Такође, важно је напоменути да извођење закључака у погледу распореда центроида (вектора аритметичких средина, $\boldsymbol{\mu}_k$), у p-димензионом простору (за $p \geq 2$) на основу карактеристичних вредности (λ) матрице $\mathbf{BW}^{-1}$ није својствено само примени *Roy*-ове статистике θ. Претходне три статистике могу се такође исказати преко карактеристичних корена матрице $\mathbf{BW}^{-1}$, на следећи начин (*Manly & Navarro Alberto*, 2017, стр. 68-69):

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \prod_{r=1}^{s}\frac{1}{(1+\lambda_r)} = \frac{1}{1+\lambda_1}\times\frac{1}{1+\lambda_2}\times\ldots\times\frac{1}{1+\lambda_s}; \tag{3.1.43}$$

$$\mathrm{V} = trace\left[\frac{\mathbf{B}}{\mathbf{T}}\right] = \mathrm{tr}\left[\mathbf{B}\mathbf{T}^{-1}\right] = \sum_{r=1}^{s}\frac{\lambda_r}{(1+\lambda_r)}; \tag{3.1.44}$$

$$\mathrm{U} = trace\left[\frac{\mathbf{B}}{\mathbf{W}}\right] = \mathrm{tr}\left[\mathbf{B}\mathbf{W}^{-1}\right] = \sum_{r=1}^{s}\lambda_r\,. \tag{3.1.45}$$

За разлику *Roy*-ове статистике θ, чије је израчунавање засновано на коришћењу само највеће карактеристичне вредности (λ_1), приказани изрази представљају различите функције свих s карактеристичних вредности λ_r (за $r = 1, 2,..., s$, и $s = min(p, g–1)$) (*Pituch & Stevens*, 2016, 210). Наведена специфичност је од посебне важности у реализацији *MANOVA*-е, будући да и информације садржане у преосталим карактеристичним вредностима ($\lambda_2, \lambda_3,..., \lambda_s$), а не само највећој ($\lambda_1$), могу бити од значаја са аспекта доношења одлуке о валидности H_0 и извођење закључака о „положају“ центроида у p-димензионом простору (*Rencher*, 2002, стр. 176–177). Такође, представљени приступ у

тестирању *MANOVA* нулте хипотезе, заснован на употреби карактеристичних вредности матрице $\mathbf{BW}^{-1}$, директно је повезан са поступком тестирања статистичке значајности појединачних дискриминационих функција, изведених применом дискриминационе анализе, која је предмет дискусије у Поглављу 3.2.

3.1.5. Компарација перформанси *MANOVA* тестова и њихова повезаност са T^2 тестом

У оквиру овог одељка представљен је компаративни преглед перформанси четири најчешће коришћене статистике у *MANOVA* анализи, посматрано у зависности од броја, анализом обухваћених, модалитета независне промењљиве као и међусобног односа карактеристичних вредности матрице $\mathbf{BW}^{-1}$. Наиме, иако се, према мишљењу *Everitt*-а (2010, стр. 271), у већини случајева може очекивати одређена еквивалентност и компатибилност одлука донетих на бази примене наведених тестова у поступку тестирања *MANOVA* нулте хипотезе, постоје оправдани аргументи који, у одређеним истраживачким ситуацијама, дају извесну предност избору једне или неколико статистика, наспрам осталих (*Huberty & Petoskey*, 2000, стр. 196).

У ситуацији када су истраживањем обухваћене само две групе (специјални случај примене *MANOVA*-е), као што је већ истакнуто, статистички тест заснован на *Hotelling*-овој T^2 статистици одликује се највећом јачином теста, у ознаци $1 - \beta$, где β означава вероватноћу неодбацивања погрешне H_0. Посматрано из угла четири *MANOVA* теста, иако се њихова примена генерално везује за поступак тестирања H_0 у случају посматрања три или више група, *Rencher* (2002, стр. 130) и *Hair et al.* (2010, стр. 347) наглашавају да се претходно описане процедуре тестирања могу прилагодити и за имплементацију у случају када је $g = 2$, а $s = g–1=1$ односно, када постоји само један ненулти карактеристични корен (λ_1). У таквим околностима, сва четири *MANOVA* теста се сматрају не само међусобно еквивалентним (будући да њихова примена резултира идентичним одлукама у погледу одрживости H_0), већ и еквивалентним са *Hotelling* T^2 тестом (будући да су вредности њихове F трансформације идентичне егзактној F статистици добијеној путем израза (3.1.26)). Наведене релације, могу се приказати на следећи начин (*Rencher*, 2002, стр. 130):

$$\left.\begin{aligned}
\Lambda &= \frac{1}{1+\lambda_1} \rightarrow \mathrm{T}^2 = (n_1+n_2-2)\frac{1-\Lambda}{\Lambda} \\
\mathrm{V} &= \frac{\lambda_1}{1+\lambda_1} \rightarrow \mathrm{T}^2 = (n_1+n_2-2)\frac{\mathrm{V}}{1-\mathrm{V}} \\
\mathrm{U} &= \lambda_1 \quad \rightarrow \mathrm{T}^2 = (n_1+n_2-2)\mathrm{U} \\
\theta &= \frac{\lambda_1}{1+\lambda_1} \rightarrow \mathrm{T}^2 = (n_1+n_2-2)\frac{1-\theta}{\theta}
\end{aligned}\right\} \underbrace{\mathrm{T}^2 = (n_1+n_2-2)\lambda_1}_{\cong} \rightarrow \underbrace{F = \frac{(n-p-1)}{p}\lambda_1 \sim F_{\underbrace{p}_{\nu_1},\underbrace{n-p-1}_{\nu_2}}}_{\cong} . \quad (3.1.46)$$

Израз (3.1.25) — Израз (3.1.26)

С друге стране, при тестирању нулте хипотезе о једнакости центроида три или више група, ситуација је нешто комплекснија, будући да се не може издвојити конкретна статистика, од наведене четири, која се може сматрати супериорном и неприкосновеном, односно са највећом јачином теста, у односу на остале у свим истраживачким околностима (*Kovačić*, 1994, стр. 120; *Rencher*, 2002, стр. 176; *Everitt*, 2010, стр. 271; *Pituch & Stevens*, 2016, стр. 210). Другим речима, иако се сматрају егзактним тестовима, будући

да имају исту вероватноћу одбацивања истините H_0 (у ознаци α), наведена четири теста нису међусобно еквивалентна, јер се одликују различитим вероватноћама одбацивања H_0 уколико је иста погрешна, односно јачином теста. У том смислу, може се очекивати да примена различитих тестова, на истом узорку мултиваријационих опсервација, резултира различитим одлукама у погледу одбацивања / неодбацивања H_0, или пак истим одлукама, али са различитом јачином емпиријских доказа у прилог, односно против тестиране тврдње. Као главни разлог описаног стања, *Rencher* (2002, стр. 176–177) наводи мултиваријациону природу (*s*-димензионог) простора унутар којег су вектори аритметичких средина ($\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, ..., $\boldsymbol{\mu_g}$) лоцирани и распоређени[93].

Наиме, за разлику од *MANOVA* случаја за две групе и само једним ненултим латентним кореном λ, када су центроиди посматране две групе распоређени унутар једнодимензионог (под)простора *p*-димензионог простора, односно дуж само једне ($s=min(p, g–1)=1$) димензије, у ситуацијама када је $s=min(p, g–1) > 1$, вектори средина су лоцирани унутар простора који је сачињен од *s* димензија, са широким спектром могућих комбинација у погледу њиховог међусобног распореда. Заправо, релативно схватање јачине теста разматраних статистика у конкретним ситуацијама, превасходно зависи од начина на који су центроиди распоређени у *s*-димензионом простору и степена у којем су појединачне статистике прилагођене анализирању те конфигурације. У том смислу, исти аутор (2002, стр. 176–177) наводи следеће две опште ситуације у погледу распореда центроида у *s*-димензионом простору уз пратеће рангирање посматраних статистика са аспекта јачине теста којом се одликују у конкретним околностима:

- Уколико су вектори средина распоређени приближно дуж једне, замишљене, праве линије, може се закључити да су они, генерално, колинеарни, односно да су разлике између центроида посматраних група претежно обухваћене и објашњене само једним, највећим, карактеристичним кореном (λ_1). У таквим околностима, *Roy*-ова статистика θ, за чије израчунавање се користи само највећа карактеристична вредност (λ_1) матрице $\mathbf{BW}^{-1}$, одликује се највећом јачином теста и, сходно томе, најбољим перформансама, у поређењу са преостале три статистике (*Pituch & Stevens*, 2016, стр. 210). Редослед посматраних статистика, са аспекта јачине теста, може се приказати следећом релацијом: $\theta \geq U \geq \Lambda \geq V$.

- У случају када су центроиди група распоређени дуж неколико димензија, односно када је њихова конфигурација у *s*-димензионом простору „дифузна“ (распршена), редослед посматране четири статистике, са аспекта јачине теста којом се одликују, је инверзан претходно извршеном рангирању, и гласи: $V \geq \Lambda \geq U \geq \theta$. Наиме, *Pillai*-јева статистика одликује се, генерално, највећом, мада незнатно већом у поређењу са *Wilks*-овом Λ и *Lawley–Hotelling*-овом *U* статистиком, јачином теста, у случају када вектори средина нису колинеарни (*Pituch & Stevens*, 2016, стр. 210). У оваквим условима, перформансе *Roy*-овог теста, сматрају се инфериорним у поређењу са јачином претходних статистика, тако да је потпуно оправдан став *Rencher*-а (2002, стр, 177) и *Johnson &*

[93] Карактеристичне вредности (λ) матрице $\mathbf{BW}^{-1}$ обезбеђују важне назнаке у погледу распоређености вектора аритметичких средина у *p*-димензионом простору мултиваријационих опсервација. Наиме, у случају када је $s = min(p, g–1) > 1$ тада постоји само једна велика карактеристична вредност λ а преостале су веома мале, тада су центроиди група распоређени у *s*-димензионом (под)простору *p*-димензионог простора приближно у форми праве линије, односно претежно дуж само једне димензије. С друге стране, ако постоје два изразита, са аспекта вредности, карактеристична корена λ, у поређењу са осталим, тада су вектори средина, по правилу, распоређени у две димензије, и слично (*Rencher*, 2002, стр. 176).

Wichern (2007, стр. 336), према којем се употреба θ статистике не препоручује, осим у случају када су вектори средина посматраних група колинеарни, под важећим претпоставкама.

Такође, сва четири теста се сматрају прилично робустним на извесну нарушеност неке од претпоставки на којима се заснива њихова примена, уколико су узорци једнаких или приближно једнаких величина (*Rencher*, 2002, стр, 177; 198; *Huberty & Petoskey*, 2000, стр. 192; *Manly & Navarro Alberto*, 2017, стр. 70). Уколико се величине посматраних група знатно разликују, а постоје оправдане сумње у погледу одрживости било претпоставке о мултиваријационој нормалности или хомогености матрица коваријанси популација из којих су узорци извучени, према мишљењу *Johnson*-а & *Wichern*-а (2007, стр. 336), *Finch*-а & *French*-а (2013, стр. 41) и *Manly*-а & *Navarro Alberto*-а (2017, стр. 70), *Pillai*-јева V статистика сматра се робустнијом у поређењу са преосталим тестовима.

3.1.6. Интерпретација резултата и специфичности везане за примену *MANOVA* методе

Генерално, резултати примењених *MANOVA* тестова могу сугерисати, сходно претходно представљеним процедурама тестирања, доношење једне од следеће две одлуке: (1) не постоји довољно емпиријских доказа за одбацивање нулте хипотезе дефинисане изразом (3.1.1), чиме се потврђује да анализирани узорци потичу из популација које се одликују једнаким векторима аритметичких средина, или пак, (2) нулта хипотеза се може одбацити и, консеквентно, прихватити алтернативна хипотеза којом се тврди, уз *a priori* детерминисани ниво значајности α, да постоји статистички значајна разлика између најмање две популације (групе) у контексту вредности припадајућих вектора аритметичких средина p зависних променљивих обухваћених анализом. У случају доношења одлуке (2), неопходно је усмерити истраживачке напоре у правцу прецизирања, односно ближег одређења, уопштено формулисане тврдње у оквиру алтернативне хипотезе. Наиме, иако се одбацивањем нулте хипотезе, H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_g = \boldsymbol{\mu}$, сугерише присуство статистички сигнификантних разлика између просечних вредности посматраних зависних променљивих на нивоу најмање две популације, резултати *MANOVA*-е не садрже директан одговор нити на једно од следећих, са становишта циљева ове анализе (на први поглед) секундарних, али у контексту комбиноване примене са другим мултиваријационим методама, важних истраживачких питања:

✓ Да ли је идентификована статистички сигнификантна разлика између посматраних група, односно њима припадајућих вектора аритметичких средина $\boldsymbol{\mu}_k$, последица присуства статистички значајних разлика између просечних вредности једне, неколико, или пак свих, анализом обухваћених, p зависних променљивих? Другим речима, која(е) зависна(е) променљива(е) је(су) допринела(е) постизању статистички сигнификантних резултата мултиваријационих *MANOVA* тестова?

✓ Уколико је одговором на претходно питање обухваћено две и више зависних променљивих, на који начин су исте рангиране са аспекта евидентираног доприноса у контексту обезбеђивања статистички значајних разлика између вектора аритметичких средина зависних променљивих на нивоу издвојених група независне променљиве?

✓ Између којих група постоје статистички сигнификантне разлике аритметичких средина по свакој од анализом обухваћених, појединачних зависних променљивих?

Обезбеђивање одговора на наведена питања, у условима када је применом *MANOVA* тестова, односно *Hotelling*-овог T^2 теста, одбачена H_0 о једнакости вектора средина променљивих Y_j на нивоу посматраних g популација, нужно подразумева одвојено тестирање валидности p униваријационих хипотеза о једнакости аритметичких средина појединачних зависних променљивих на нивоу g група, засновано на спровођењу једнофакторске *ANOVA*-е и примени униваријационог F теста (израз (3.1.33)), у случају када је $g \geq 3$, односно статистике Студентовог t теста за два независна узорка (израз (3.1.22)) уколико је $g = 2$. Такође, за разлику од случаја када се испитују само две мултиваријационе популације, при анализи три и више модалитета независне променљиве, сигнификантност резултата *MANOVA* тестова не „гарантује“ присуство статистички значајних разлика између свих ($g \geq 3$) група. У таквим околностима, у циљу идентификовања конкретних парова група између којих постоје статистички сигнификантне разлике у погледу средина појединачних зависних променљивих, чији је допринос потврђен серијом униваријационих F тестова, неопходно је спровести одговарајућу *post hoc* анализу, засновану на вишеструкој компарацији свих могућих [g (g–1) / 2] парова група, применом једне од следећих *post hoc* статистичких процедура[94]: (1) *Scheffé*-ов тест, (2) *Tukey*-јев *HSD* (енгл. *Honestly Significant Difference*) тест, (3) *Tukey-Kramer*-ов тест, (4) *Fisher*-ов *LSD* (енгл. *Least Significant Difference*) тест, (5) *Duncan*-ов *MR* (енгл. *Multiple Range*) тест, (6) *Newman-Keuls*-ов тест. У контексту наведене вишеструке имплементације униваријационог F или t теста, а у циљу минимизирања могућности доношења погрешних закључака, важну улогу има ефикасно контролисање укупног ризика грешке I врсте, α, будући да се вероватноћа идентификовања најмање једне статистички сигнификантне разлике, у условима када је H_0 истинита, вишеструко увећава са повећањем броја реализација датог теста (односно броја зависних променљивих).

Наиме, при примени униваријационог теста (F или t тест, сходно броју група) за испитивање статистичке сигнификантности разлика између средина конкретне зависне променљиве на нивоу g популација ($g \geq 2$), уз ниво значајности $\alpha = 0{,}05$, вероватноћа неодбацивања истините H_0 износи 0,95, односно 95%. Прецизније, за изабрани ризик грешке I врсте, истраживач прихвата могућност доношења погрешне одлуке, односно идентификовања статистички значајних разлика иако реално узорци потичу из популација које се одликују једнаким аритметичким срединама, једном у двадесет покушаја тестирања. Међутим, ова констатација није одржива у ситуацији када се врши спровођење p засебних тестова под истим околностима. Наиме, у случају вишеструке имплементације истог теста, вероватноћа одбацивања истините H_0, у најмање једном случају, вишеструко се увећава а њен приближан износ за серију униваријационих тестова, под претпоставком да су посматране зависне променљиве међусобно независне може се утврдити на следећи начин[95] (*Manly & Navarro Alberto*, 2017, стр. 59; *Pituch & Stevens*, 2016, стр. 143):

---

[94] Детаљне информације о тестовима вишеструке компарације (енгл. *post hoc tests*) видети у: *Čobanović i drugi* (2003); *Pituch & Stevens* (2016, стр. 184–192).

[95] На пример, као резултат примене серије p *ANOVA F* тестова, у случају анализирања 3, 4, 5 и 10 зависних променљивих, уз појединачни ниво значајности $\alpha = 0{,}05$, укупан (прецењени) ниво (енгл. *inflated*) ризика грешке I врсте, сходно изразу (3.1.47), налазиће се (приближно) унутар следећих интервала: (1) за $p = 3 \rightarrow \alpha \in$ (0,05–0,14), (2) за $p = 4 \rightarrow \alpha \in$ (0,05 – 0,18), (3) за $p = 5 \rightarrow \alpha \in$ (0,05–0,23) и (4) за $p = 10 \rightarrow \alpha \in$ (0,05–0,40).

$$\text{P}(\textit{одбацивање истините } H_0) = 1 - \text{P}(\textit{неодбацивање истините } H_0)^{\,p}, \qquad (3.1.47)$$

где је p број зависних променљивих (односно, број појединачних униваријационих тестова). Прецизније, вероватноћа доношења исправне одлуке, под наведеном претпоставком, се смањује са 0,95 на $0{,}95^p$, а вероватноћа доношења погрешне одлуке се увећава са 0,05 на $1–(0{,}95)^p$, што се, генерално, сматра неприхватљиво високим укупним ризиком грешке I врсте. Међутим, будући да је између зависних променљивих, у мултиваријационом контексту, углавном присутан одређени степен корелације, услед чега претпоставка о њиховој независности не може бити у потпуности уважена, исти аутори наводе да се укупан ниво значајности, у случају вишеструке примене униваријационих тестова, по правилу креће у опсегу од 0,05 до $1–(0{,}95)^p$. Другим речима, применом серије појединачних униваријационих тестова укупан ризик грешке α постаје прецењен, будући да је увек изнад дефинисаног номиналног нивоа α=0,05.

Полазећи од наведеног, у условима вишеструке примене униваријационих тестова, важно је, са аспекта ефикасног контролисања укупног ризика грешке α, извршити одговарајуће прилагођавање појединачних нивоа значајности. Расположиве процедуре прилагођавања разликују се у зависности од тога да ли је анализом обухваћено два[96] или више модалитета независне променљиве. Најједноставнији приступ у контролисању укупног ризика грешке α заснива се на спровођењу *Bonferonni*-јеве корекције, односно процедуре прилагођавања појединачних нивоа значајности броју спроведених тестова[97]. Наиме, дељењем нивоа значајности за сваки појединачни тест са бројем зависних променљивих врши се њихова индивидуална корекција са нивоа α на (прилагођени) ниво α/p, а укупан ризик грешке I врсте (енгл. *overall* α), за серију униваријационих тестова, биће генерално једнак или мањи од α (*Manly & Navarro Alberto*, 2017, стр. 60; *Pituch & Stevens*, 2016, стр. 151). За разлику од серије униваријационих тестова чије спровођење резултира проблемом израженог повећања укупног ризика грешке I врсте, испитивање статистичке значајности разлика вектора средина свих зависних променљивих (истовремено) на нивоу посматраних група коришћењем свеобухватних *MANOVA* тестова, врши се уз прецизну контролу (одржавање) нивоа значајности на изабраном нивоу α, независно од броја зависних променљивих. Строга контрола нивоа значајности, карактеристична за *MANOVA* тестове, представља један од кључних статистичких аргумената за преферирање мултиваријационог приступа наспрам вишеструке примене кореспондентних униваријационих тестова, посебно када је анализом обухваћен велики број зависних променљивих (*Pituch & Stevens*, 2016, стр. 143; *Manly & Navarro Alberto*, 2017, стр. 59–60; *Hair et al.*, 2010, стр. 354). Додатни аргументи у корист примене *MANOVA*-е садржани су и у чињеници да мултиваријациони тестови омогућавају симултано испитивање више зависних променљивих уз уважавање њихове међузависности, услед чега се, према мишљењу истих аутора, одликују већом јачином теста у поређењу са приступом заснованим на серији униваријационих тестова, који у

[96] Детаљан приказ поступка прилагођавања нивоа значајности при спровођењу серије униваријационих *t* тестова, у случају посматрања две популације (групе) видети у *Hair et al.* (2010, стр. 376).

[97] Као главни недостатак *Bonferonni*-јеве корекције, *Pituch & Stevens* (2016, стр. 151) и *Manly & Navarro Alberto* (2017, стр. 60) истичу изражени пад јачине теста до којег може доћи у условима коришћења релативно великог броја зависних променљивих, конкретно када је $p > 7$, будући да тада ниво значајности појединачних тестова постаје веома низак ($\alpha/p \approx 0$).

потпуности „игнорише“ информације у погледу корелације присутне између зависних променљивих. Захваљујући наведеним предностима, *MANOVA* тестови могу обезбедити доказе у корист алтернативне хипотезе, чак и уколико су резултати свих спроведених униваријационих *F* или *t* тестова несигнификантни, будући да омогућавају детектовање „укупних“ разлика, насталих комбиновањем (здруживањем) малих разлика присутних на нивоу свих (или само неких) зависних променљивих појединачно, а које иначе могу бити неопажене засебним испитивањем варијабли у оквиру униваријационе анализе (*Hair et al.*, 2010, стр. 354; *Pituch & Stevens*, 2016, стр. 143; 156; *Manly & Navarro Alberto*, 2017, стр. 59). Обратно, у ситуацији када су докази о присуству разлика, обезбеђени од стране сигнификантних променљивих, „преплављени“ доказима о одсуству разлика, обезбеђени другим несигнификантним променљивим, резултати мултиваријационих тестова могу се показати статистички несигнификантним, упркос закључцима појединих униваријационих тестова (*Everitt*, 2010, стр. 263; *Manly & Navarro Alberto*, 2017, стр. 59).

3.2. ДИСКРИМИНАЦИОНА АНАЛИЗА

Средином 30-их година XX века, *R.A. Fisher* (1936) је развио и предложио технику за креирање линеарне комбинације варијабли за обезбеђивање максималног разликовања три врсте цвета перунике на основу четири мерне карактеристике. Наведена линеарна композитна комбинација анализом обухваћених мерних карактеристика, креирана за потребе класификације засноване на двема групама (енгл. *two-group classification*), названа је *линеарна дискриминациона функција*, што уједно представља и прву званичну (формалну) употребу наведеног термина у писаној форми (*Huberty*, 2011, стр. 390). Иако се од стране појединих аутора, за одређене идеје, попут истраживања вишедимензионог међугрупног одстојања (енгл. *multivariable intergroup distance*) спроведених од стране, превасходно, *K. Pearson*-а и *P.C. Mahalanobis*-а током 1920-их и почетком 1930-их година, респективно, сматра да су у извесној мери подстакле и претходиле *Fisher*-овом формулисању концептуално-методолошких одређења дискриминационе анализе (*Huberty & Olejnik*, 2006, стр. 3), *Fisher*-ов рад (1936), објављен у часопису *Annals of Eugenics*, под насловом „*The use of multiple measurements in taxonomic problems*" изазвао је изражени пораст интересовања бројних истраживача из различитих научних области у контексту испитивања апликативних и могућности за теоријско–методолошка унапређења и даљи развој дискриминационе анализе. Детаљан хронолошки приказ кључних достигнућа која се односе на даљи развој дискриминационе анализе представљен је од стране *Huberty & Olejnik* (2006, стр. 3–5) и *Klecka* (1980, стр. 14–15). У том контексту, посебно се издваја проширење могућности примене дискриминационе анализе са класификације засноване на двема групама на решавање истраживачких проблема заснованих на разликовању више од две групе јединица посматрања, предложено од старне *C.R. Rao*-а (1948).

У овом поглављу представљена је детаљна дискусија у погледу природе, филозофије, логике која карактерише и услова на којима почива адекватна и статистички оправдана примена дискриминационе анализе. Излагање започиње циљевима и кратким прегледом специфичних корака у поступку имплементације, након чега следи њихово, појединачно, детаљније објашњење, уз елаборирање кључних статистичких предуслова за њихово извршавање. Исправна интерпретација добијених резултата, праћена провером статистичке значајности истих, као и разматрање предиктивних могућности изведеног дискриминационог модела опредељују садржај остатка излагања.

3.2.1. Циљеви и поступак спровођења дискриминационе анализе

Дискриминациона анализа (акроним: ДА) представља мултиваријациони статистички метод зависности који омогућава испитивање разлика између две или више, међусобно искључивих, група јединица посматрања, на основу симултаног сагледавања и анализе измерених вредности две или више међусобно сродних али независних нумеричких карактеристика, за потребе извођења одговарајућег броја (једне или више) линеарних латентних варијабли (енгл. *variate*) које на најбољи начин раздвајају *a priori* дефинисане групе мултиваријационих опсервација.

Наведена, на први поглед комплексна, формулација суштинских одређења ДА садржи одређене елементе који су кључни за јасно диференцирање овог

мултиваријационог метода у односу на друге, по свом карактеру сличне, статистичке методолошке поступке. Наиме, ДА представља мултиваријациону статистичку процедуру чија се примена везује за решавање истраживачких проблема који укључују једну категоријску (номиналну или ординалну) зависну променљиву Y_k (где k означава редни број издвојених модалитета, класа, категорија и / или група, за k=1, 2,..., g) и неколико (две или више) нумеричких независних променљивих X_j (за j=1, 2,..., p), најчешће мерених на интервалној или скали односа. Уколико је зависна променљива Y_k дихотомна (односно, g = 2), примењена варијанта дискриминационе анализе назива се *ДА заснована на двема групама* (енгл. *two-group discriminant analysis*). Када је анализом обухваћена мултихотомна категоријска променљива Y_k (симболички, g > 2) тада је реч о примени *вишегрупне ДА* (енгл. *multigroup discriminant analysis*). У случају оба типа анализе, ДА омогућава испитивање да ли, на бази расположивог узорка мултиваријационих опсервација, између издвојених модалитета категоријске и скупа нумеричких објашњавајућих променљивих постоји статистички значајна квантитативна зависност, кроз формирање одређеног броја латентних линеарних комбинација независних променљивих, које се називају дискриминационе функције (енгл. *discriminant functions*) или дискриминанте (енгл. *discriminants*). Засноване на максимизирању односа варијансе мултиваријационих опсервација између и унутар група, формиране дискриминационе функције обезбеђују најбољу дискриминацију (односно, раздвајање) *a priori* дефинисаних категорија јединица посматрања, из угла коришћених независних променљивих. Сходно томе, оне представљају погодну основу за описивање карактеристика појединачних категорија зависне променљиве, разумевање разлика између њих, али и за предвиђање припадности (односно, алокацију) нових опсервација једној од расположивих категорија променљиве Y_k, уз минимизирање вероватноће погрешне класификације. Полазећи од наведеног, дискриминациона анализа представља погодан мултиваријациони статистички метод за реализацију следећих истраживачких циљева (*Sharma*, 1996, стр. 237; *Hair et al.*, 2010, стр. 246–247; *Izenman*, 2008, стр. 237; *Brown & Wicker*, 2000, стр. 209–210):

- Идентификовање (под)скупа независних променљивих (дискриминатора) које у значајној мери доприносе сепарацији (разликовању) анализираних опсервација унутар једне од *a priori* дефинисаних категорија зависне променљиве;
- Рангирање дискриминатора (енгл. *discriminator variables*) на основу њиховог појединачног релативног доприноса дискриминацији и идентификовање независне променљиве која поседује највећу дискриминациону моћ (енгл. *discriminatory power*) у процесу раздвајања посматраних група;
- Креирање одговарајућег броја (једне или више) дискриминационих функција, односно пондерисаних линеарних комбинација (под)скупа одабраних независних променљивих које на најбољи начин и недвосмислено репрезентују диференцијалне карактеристике јединица посматрања распоређених унутар различитих, *a priori* дефинисаних, категорија зависне променљиве;
- Класификација нових јединица посматрања, које нису коришћене у поступку оцењивања дискриминационог модела и чија припадност конкретној групи није позната, унутар једног од *a priori* дефинисаних модалитета зависне променљиве, коришћењем изведене дискриминационе функције, одговарајућег класификационог правила и

припадајућих вредности (под)скупа независних променљивих, али тако да се минимизира вероватноћа погрешне класификације.

Представљена листа циљева у опусу дискриминационе анализе отвара могућности за детаљније разумевање концептуално-методолошке природе овог статистичког метода. Наиме, прва три циља усмерена су на обезбеђивање објективне оцене разлика између одговарајућих категорија зависне променљиве које могу бити присутне из угла вредности разматраних независних променљивих. Прецизније, за потребе разумевања, описивања и интерпретације разлика које постоје између посматраних категорија јединица посматрања, ДА обезбеђује увид у релативни допринос појединачних независних променљивих процесу дискриминације и пружа могућност њиховог пондерисаног линеарног комбиновања у форми одговарајуће дискриминационе функције. Креирана дискриминациона функција и на бази ње изведено класификационо правило, омогућавају истраживачу да квантитативно објасни припадност јединица посматрања једној од дефинисаних категорија зависне променљиве, из угла расположивих вредности дискриминатора и припадајућих пондера. Важно је напоменути да је за реализацију наведених настојања неопходно располагати одговарајућим узорком мултиваријационих опсервација у којем је за сваку јединицу посматрања позната информација о припадности конкретној категорији зависне променљиве. Изражена повезаност и наглашена (примарно) дескриптивна природа издвојена прва три циља утицали су на њихово, у литератури често присутно, обједињавање у оквиру једног заједничког ДА циља, за чије означавање се равноправно користе термини попут (*Johnson & Wichern*, 2007, стр. 575): сепарација, раздвајање група и / или дискриминација. ДА спроведена за потребе сепарације и описивање утврђених разлика између две или више категорија зависне променљиве, односно реализације овог „обједињеног“ циља, назива се дескриптивна дискриминациона анализа (акроним: ДДА) (*Huberty*, 2011, стр. 391).

Међутим, насупрот разматраним дескриптивним аспектима дискриминационе анализе, у околностима када истраживач располаже одређеним јединицама посматрања чија припадност једној од *a priori* дефинисаних категорија није позната, при чему је веома важно или пак пожељно идентификовати исту, предиктивни аспект дискриминационе анализе обухваћен формулацијом последњег циља на листи, долази до изражаја. Аспекти који се односе на реализацију овог циља углавном су у литератури означени терминима (*Johnson & Wichern*, 2007, стр. 575): класификација, алокација и / или разврставање. Спровођење предиктивне ДА (акроним: ПДА) захтева употребу креираних дискриминационих функција и изведених класификационих правила за потребе алокације „нових“ јединица посматрања.[98]

На основу наведеног, може се констатовати да је реализација ПДА директно условљена резултатима ДДА, односно да остваривање предиктивних циљева нужно подразумева претходно остваривање дескриптивних циљева ДА. Прецизније, ПДА захтева употребу класификационих правила и дискриминационих функција изведених у оквиру претходно спроведене ДДА на расположивом узорку мултиваријационих опсервација у којем је припадност појединачних јединица посматрања конкретној категорији зависне

[98] Придев „нових“ употребљен је у циљу јасног означавања да је реч о јединицама посматрања које нису коришћене у поступку оцењивања дискриминационих функција и за које није позната припадност конкретној категорији анализом обухваћене зависне променљиве.

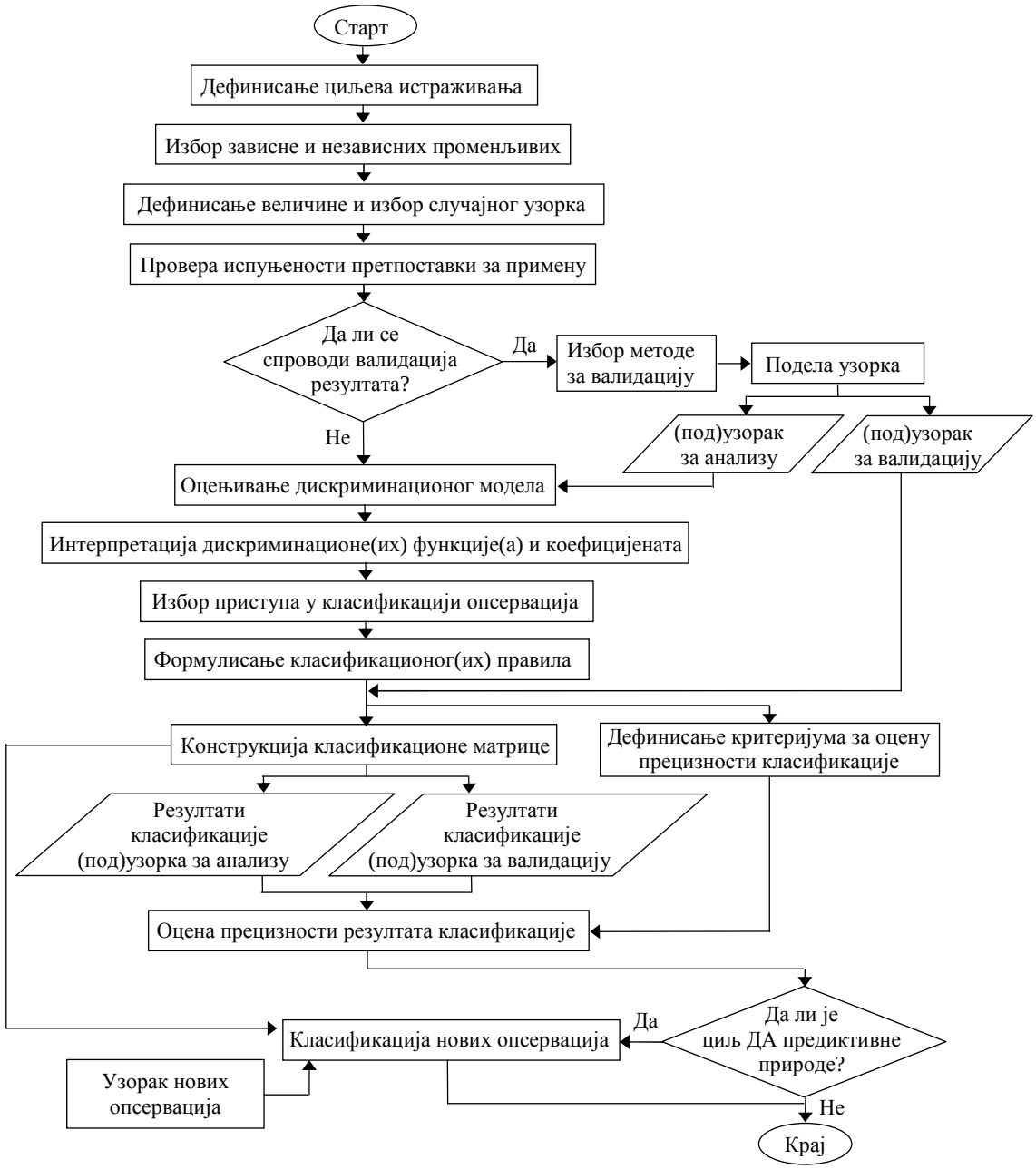
променљиве позната. Разликовање ДДА и ПДА приступа представља прилично важно питање из угла правилног дефинисања истраживачког проблема, циљева истраживања, интерпретације и представљања резултата ДА. Међутим, услед занемаривања специфичности и разлика, али и међусобног преклапања дискутованих дескриптивних и предиктивних циљева ДА, нажалост, у литератури и пракси је прилично уобичајено поистовећивање ДДА и ПДА, а разликовање термина сепарације од алокације и дискриминације од класификације, често је замагљено (*Huberty*, 2011, стр. 390–391; *Johnson & Wichern*, 2007, стр. 575).

Генерално, уз уважавање истакнутих разлика у концептуалној основи између ДДА и ПДА, оправдано је закључити да се дискриминациона анализа може посматрати и користити и као специфични тип дескриптивне анализе и као предиктивна аналитичка техника, у зависности од формулисаних циљева истраживања и конкретног истраживачког проблема. Поступак спровођења дискриминационе анализе може бити представљен у форми вишеетапног дијаграма процеса изградње дискриминационог модела (Слика 3.2.1).

Слично већини осталих мултиваријационих статистичких метода, поступак спровођења ДА у полазним етапама подразумева јасно и прецизно формулисање циљева анализе у контексту планираног истраживања, као и избор једне категоријске променљиве са одговарајућим модалитетима, и добро образложен избор скупа независних нумеричких променљивих које ће бити коришћене у анализи. Након разматрања кључних питања која се односе на дизајн истраживања, пажња се усмерава на комплексан поступак претпроцесирања прикупљених мултиваријационих података и њихову адекватну припрему за даљи ток анализе. Доминантна улога у овој етапи припада незаобилазној и свеобухватној провери степена испуњености конкретних статистичких претпоставки на којима почива валидна апликација ДА и обезбеђивање научне заснованости добијених резултата и изведених закључака. Анализа даље подразумева разматрање и избор методе за вредновање резултата формираног дискриминационог модела и поступка класификације. На основама изабране методе, врши се подела иницијалног узорка мултиваријационих опсервација на (под)узорак за анализу (енгл. *analysis sample*), који ће бити коришћен у оцењивању параметара дискриминационог модела, и (под)узорак за валидацију (енгл. *validation / holdout sample*), који ће бити употребљен за вредновање резултата, односно оцену прецизности класификације спроведене на основу изведеног дискриминационог модела.

Централна улога у поступку спровођења ДА, без обзира да ли је реч о ДДА или ПДА, управо припада креирању одговарајућег дискриминационог модела, сачињеног од једне или више дискриминационих функција, будући да представља заједнички именитељ успешне примене оба типа ДА. Другим речима, кључни корак у реализацији циљева дискриминационе анализе јесте пондерисање и линеарно комбиновање информација из скупа независних варијабли X_j (за $j = 1,2,...,p$) на начин који обезбеђује што је могуће боље раздвајање g дефинисаних категорија зависне променљиве Y_k. Важно је истаћи да дискриминациони модел није сам по себи циљ, већ средство неопходно за реализацију претходно дефинисаних циљева истраживања. Овом етапом обухваћене су и активности усмерене на испитивање статистичке и практичне значајности креираних дискриминационих функција. Интерпретација и рангирање вредности дискриминационих

коефицијената утврђених за сваку појединачну независну промењљиву у саставу дискриминационе функције омогућавају идентификовање дискриминатор промењљиве која се одликује најизраженијом дискриминационом снагом.



Слика 3.2.1. Дијаграм тока поступка спровођења дискриминационе анализе

Извор: Ауторов приказ

Након креирања ДА модела, идентификовања статистички значајних дискриминационих функција и њихове интерпретације, спроводи се оцена, односно вредновање резултата формираног дискриминационог модела у контексту прецизности класификације јединица посматрања унутар појединачних категорија зависне промењљиве, као и предиктивних могућности модела. Овај сложени поступак, као што је и приказано на Слици 3.2.1, обухвата неколико критичних активности, и то:

(1) избор конкретног приступа у класификацији јединица посматрања, односно избор између приступа заснованог на израчунавању дискриминационих *Z* вредности (енгл.

discriminant Z scores) и критичне *Z* вредности (енгл. *cutoff value*) и приступа заснованог на дефинисању *Fisher*-ових линеарних дискриминационих функција, такође познатих под називом класификационе функције појединачних категорија зависне променљиве;

(2) формулисање одговарајућих класификационих правила, односно класификатора (енгл. *classifiers*) у зависности од изабраног приступа процесу класификације;

(3) дефинисање критеријума за оцену прецизности класификације; и

(4) класификација расположивих јединица посматрања за које је позната припадност једној од категорија зависне променљиве и конструкција класификационе матрице.

Важно је истаћи, да се за потребе оцене прецизности класификације и предикције модела, у оквиру ове фазе, на бази дефинисаних правила врши класификација мултиваријационих опсервација (под)узорка за анализу, коришћених за оцењивање дискриминационог модела, као и, мада одвојено, јединица посматрања у саставу (под)узорка за валидацију у циљу спровођења додатне екстерне евалуације класификационих способности и карактеристика модела уз минимизирање пристрасности у примени модела. Међусобном компарацијом резултата класификације (под)узорака за анализу и валидацију и њиховим поређењем са претходно дефинисаним критеријумима, изводе се закључци у погледу статистичке значајности оцене класификационе прецизности модела. Све претходно елаборирane етапе заједно детерминишу поступак спровођења ДДА. Међутим, уколико је дефинисани ДА циљ истраживања предиктивне природе, поступак се наставља и подразумева коришћење претходно формулисаних класификационих правила за потребе класификације нових опсервација унутар једне од дефинисаних група зависне променљиве.

3.2.2. Избор променљивих и статистичке претпоставке за примену ДА

Спровођење дискриминационе анализе захтева располагање репрезентативним узорком јединица посматрања које су, на основу одговарајућих вредности две или више независних променљивих, распоређене унутар две или више међусобно искључивих група, односно модалитета зависне променљиве. Након формулисања циља истраживања, адекватан избор зависне и скупа одговарајућих независних променљивих представља иницијални кључни корак у постизању успешне имплементације ДА, којем је потребно посветити посебну пажњу.

Зависна променљива у дискриминационој анализи је променљива којом се означава припадност сваке појединачне јединице посматрања конкретној групи или категорији. Прецизније, модалитети променљиве овог типа, представљају могуће, међусобно искључиве категорије (односно, групе) унутар којих јединице посматрања могу бити алоциране (*Brown & Wicker*, 2000, стр. 211). Термин „међусобно искључиве“ употребљен је како би се нагласила чињеница да свака јединица посматрања може бити распоређена унутар једне и само једне групе (*Hair et al.*, 2010, стр. 247). Полазећи од карактеристика „вредности“ које може узети, као и чињенице да се исте утврђују углавном коришћењем одговарајуће номиналне или ординалне мерне скале, зависна променљива у ДА представља категоријску променљиву. За потребе спровођења ДА, зависна променљива мора поседовати најмање два модалитета, док ограничења у погледу горње границе

коришћеног броја категорија мултихотомне зависне променљиве нису дефинисана у теоријском контексту. Наиме, теоретски посматрано, ДА може бити примењена на неограниченом броју модалитета у оквиру зависне променљиве, међутим, питања практичности и комплексности резултирајуће анализе морају бити предмет разматрања при избору модалитета зависне променљиве. У том контексту, *Hair et al.* (2010, стр. 247–248) апострофирају неопходност уважавања следећих препорука при одређивању броја модалитета и структуре зависне променљиве:

- Категорије зависне променљиве треба дефинисати на начин који обезбеђује, у што је могуће већој мери, њихову међусобну различитост и појединачну јединственост, посматрано из угла скупа независних променљивих које ће бити употребљене за оцењивање дискриминационог модела. Наиме, уколико се две или више категорија одликују релативно сличним профилима, ДА неће бити у могућности да прецизно и јасно моделира карактеристике и разлике између њих, у контексту коришћених независних променљивих, што ће коначно резултирати лошијим перформансама изведеног дискриминационог модела како у погледу објашњавања разлика тако и са аспекта прецизности у реализацији предиктивних циљева анализе.
- Комплексност спровођења и интерпретације резултата дискриминационе анализе директно је пропорционална повећању броја категорија зависне променљиве. Сходно наведеном, *Hair et al.* (2010, стр. 248) истичу да истраживачи при спровођењу ДА треба да теже ка избору мањег пре него већег броја категорија у оквиру зависне променљиве.

Претходна дискусија сугерише неопходност перманентног балансирања између жеље за постизањем прецизније категоризације јединица посматрања и потребе за повећањем ефикасности и репрезентативности резултирајућег дискриминационог модела. С друге стране, избор независних променљивих директно је условљен дефинисаним категоријама зависне променљиве и не може се спроводити као независан и изолован процес, иако се одлукује одређеним специфичностима. Независне променљиве у ДА, такође познате и под називом дискриминатор променљиве, представљају нумеричке карактеристике јединица посматрања које се користе за описивање и моделирање разлика које су присутне између издвојених категорија зависне променљиве. *Klecka* (1980, стр. 9) истиче да дискриминатор променљиве морају бити мерене искључиво на интервалној или скали односа, како би, будући да исте представљају праве нумеричке скале, примена неопходних математичких операција на њиховим вредностима имала смисла. *Brown & Wicker* (2000, стр. 213) допуњују и проширују претходни став, напомињући да би независне променљиве у ДА требале да задовоље барем захтеве који се односе на ординални ниво мерења.[99]

Полазећи од чињенице да ДА оцењује степен у којем анализиране независне променљиве доприносе разумевању и објашњавању дистинкције између категорија зависне променљиве, логично је закључити да перформансе и репрезентативност креираног дискриминационог модела у великој мери зависи од извршеног избора дискриминатор променљивих. У литератури се препоручује појединачна и, у већој мери,

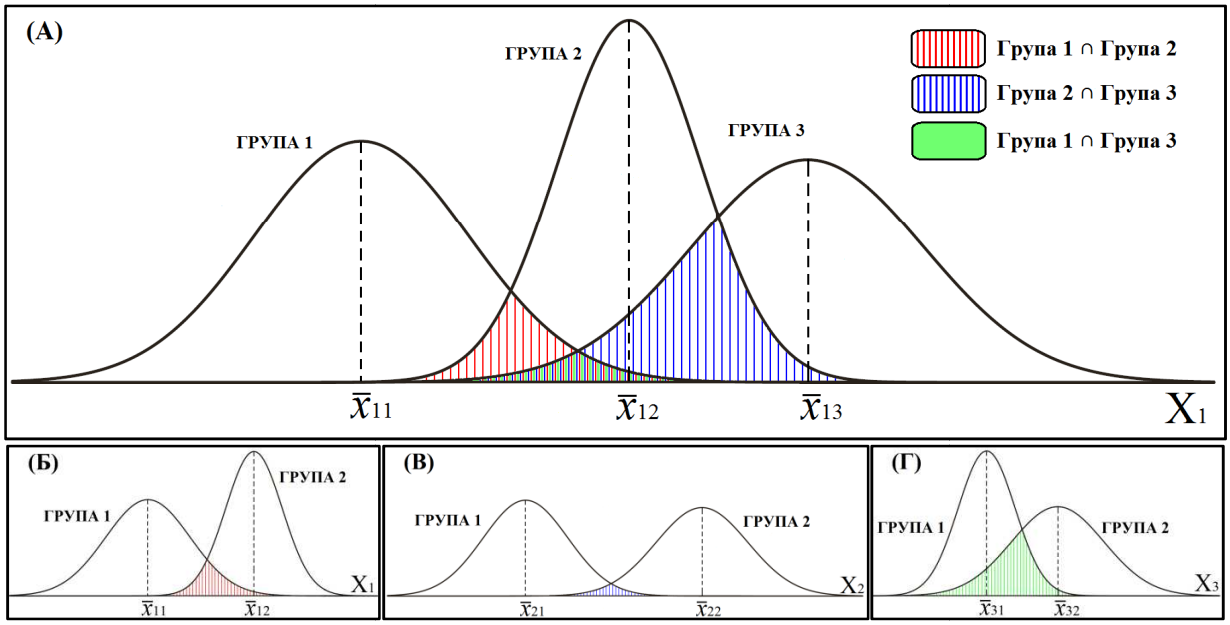
[99] Наиме, иако бројеви на ординалној скали показују само редослед рангова јединица посматрања али не и степен њиховог разликовања (*Lovrić*, 2009), као што је то углавном случај код променљивих чије су вредности прибављене коришћењем анкетног упитника и *Likert*-ове скале, ординалност ипак омогућава описивање релација које постоје између издвојених група јединица посматрања у контексту посматраних независних променљивих *Brown & Wicker* (2000, стр. 213).

комбинована примена следећа два приступа при избору потенцијалних независних променљивих (*Brown & Wicker*, 2000, стр. 212; *Hair et al.*, 2010, стр. 249):

- Идентификовати и размотрити променљиве које су се показале релевантним за дискриминацију, на основу анализе релеватних теоријских постулата, постојећих теоријских модела и претходних емпиријских истраживања.
- Заснованост избора независних променљивих које нису обухваћене претходним емпиријским и / или теоријским истраживањима, а које се, логички посматрано, одликују пожељним и адекватним дискриминационим потенцијалом у контексту разматраног истраживачког проблема, на интуицији и експертском знању истраживача.

Независно од тога којем приступу је поверена доминантна улога при спровођењу поступка избора, коначна одлука о томе које ће независне променљиве бити укључене у анализу доноси се на основу резултата прелиминарне статистичке анализе којој су подвргнуте променљиве-кандидати за укључивање у ДА. Статистички скрининг, односно прелиминарна анализа потенцијалних независних променљивих, подразумева употребу одговарајућих како графичких тако и нумеричких статистичких метода заснованих на униваријационом и мултиваријационом поређењу просечних вредности разматраних независних променљивих по издвојеним групама зависне променљиве, а у циљу давања одговора на следећа питања: ▸ Да ли независне променљиве, појединачно и заједно, обезбеђују разликовање посматраних група јединица посматрања? ▸ У којој мери посматране променљиве, појединачно и заједно, доприносе дискриминацији између издвојених група, наравно, уколико је одговор на претходно питање потврдан? ▸ Између којих група јединица посматрања се обезбеђује најбоље раздвајање коришћењем конкретних независних променљивих, посматрано из униваријационе и мултиваријационе перспективе? ▸ Која независна променљива се може сматрати најбољим дискриминатором? ▸ Како су независне променљиве рангиране према прелиминарно утврђеној, појединачној дискриминационој моћи?

Приступ заснован на коришћењу униваријационих и мултиваријационих графичких метода, иако се закључци донети на основу њега могу сматрати у већој или мањој мери субјективним, представља погодно средство за стицање прелиминарног увида у појединачну и / или заједничку дискриминациону моћ независних променљивих и, сходно томе, обезбеђивању оквирних одговора на претходно наведена питања. Униваријациони графички прикази који могу бити коришћени у наведеном контексту, у зависности од тога да ли је реч о прекидним или непрекидним нумеричким вредностима независних променљивих су: дијаграм ордината тачака, линијски дијаграм, хистограм фреквенција, полигон фреквенција и теоријска крива фреквенција. С друге стране, за визуелну процену степена у којем две или три независне променљиве заједно доприносе дискриминацији између издвојених група, углавном се користи дијаграм распршености и његова тродимензионална верзија (енгл. *3D scatter plot*), респективно. У циљу ближег објашњења изнетих тврдњи, на Слици 3.2.2.(А) представљен је хипотетички пример употребе униваријационог графичког приказа у форми теоријске криве фреквенција за променљиву X_1 за три групе јединица посматрања, док је на Слици 3.2.2.(Б, В, Г) илустрован хипотетички истраживачки проблем дискриминације између две групе јединица посматрања коришћењем три независне променљиве (X_1, X_2, X_3).



Слика 3.2.2. Хипотетички примери универијационе визуелне оцене дискриминационе моћи потенцијалних независних променљивих

Извор: Ауторов приказ

Графички прикази распореда фреквенција за променљиву X_1 по свакој од издвојене три групе (Слика 3.2.2.(А)) упућује на закључак да ова променљива обезбеђује одлично раздвајање група 1 и 3, будући да је површина преклапања двеју кривих фреквенција најмања, односно знатно мања у поређењу са површинама пресека осталих парова кривих. Поред улоге доброг дискриминатора група 1 и 3, представљену променљиву карактерише и мања али ипак запажена дикриминациона моћ када је реч о раздвајању група 1 и 2. Велика површина преклапања кривих фреквенција које се односе на групе 2 и 3 сугеришу да променљива X_1 не обезбеђује њихову довољно прецизну и јасну сепарацију. Консеквентно, придруживање променљивој X_1 неке(их) друге(их) променљиве(их) која(е) ће обезбедити прецизнију дистинкцију група 2 и 3, резултираће погодном основом за креирање репрезентативног дискриминационог модела.

С друге стране, на основу графичких приказа на Слици 3.2.2.(Б,В,Г), који се односе на случај када истраживачки проблем подразумева сагледавање дискриминационе моћи три независне променљиве у контексту две групе опсервација, могу се формулисати следећи, прелиминарни закључци: ▸ променљива X_2 је „најбољи“ дискриминатор посматране две групе; ▸ у поређењу са променљивом X_1 и X_2, променљива X_3 обезбеђује најмање прецизну класификацију јединица посматрања унутар посматраних група, односно, одликује се најмањом дискриминационом моћи у раздвајању посматраних група, тако да иста вероватно не треба да буде укључена у поступак оцењивања дискриминационог модела; ▸ претходне закључке потврђују и визуелно уочљиве разлике (односно, удаљеност) просечних вредности посматраних променљивих на нивоу издвојених група ▸ променљиве X_1 и X_2, односно њихова комбинација, обезбеђују погодну основу за ефикасно описивање и моделирање разлика које постоје између група 1 и 2.

Верификација закључака донетих на бази графичких приказа подразумева спровођење одговарајућих рачунских статистичких метода усмерених на испитивање да ли постоје статистички значајне разлике између просечних вредности независних

променљивих на нивоу издвојених група у структури зависне променљиве, посматрано из униваријационе и мултиваријационе перспективе. Такође, важно је истаћи да многи статистичари не препоручују да се мултиваријационим аспектима прелиминарне статистичке анализе обухватају променљиве за које није претходно потврђена униваријациона статистичка значајност разлика између просечних вредности на нивоу различитих група, без обзира на евентуално присутну, изражену корелацију са променљивима које се одликују запаженим дискриминационим својствима (*Huberty*, 1975, стр. 555). У складу са наведеним, *Hair et al.* (2010, стр. 249) напомињу да улогу дискриминатора треба поверити оним независним променљивим чије се просечне вредности статистички значајно разликују на нивоу барем две групе јединица посматрања, и истичу да су променљиве које не задовољавају овај услов од малог значаја за реализацију ДА циљева.

Посебно методолошко питање које је неопходно размотрити односи се на максимални број независних променљивих које је могуће користити у оцењивању дискриминационог модела. Наиме, и док за доњу границу постоји општа сагласност око употребе најмање две независне променљиве (*Brown & Wicker*, 2000, стр. 213) у релевантној литератури присутна је изражена неусаглашеност ставова у погледу максималног броја независних променљивих. За разлику од става према којем генерално не постоје ограничења у погледу коришћених независних променљивих све док је исти мањи од укупног броја опсервација умањеног за два (*Klecka*, 1980, стр. 11; *Savić i drugi*, 2008, стр. 30), *Brown & Wicker* (2000, стр. 213) истичу да број независних променљивих не би требао да буде већи од броја категорија у структури зависне променљиве. Без прецизирања горње границе, *Izenman* (2008, стр. 240) апострофира приступ у реализацији циљева ДА који је заснован на коришћењу што је могуће мањег броја независних променљивих. У контексту предмета полемисања, *Huberty & Olejnik* (2006, стр. 11) предлажу ограничење које сматрају применљивим и одрживим за већину студија у домену примене ДА, а према којем максимални број коришћених независних променљивих треба да се налази у распону од 10 до 12, осим у случају када постоје довољно уверљиви и оправдани разлози за укључивање додатних дискриминатора. До сличног закључка долазе и *Varmuza & Filzmoser* (2009, стр. 151), али на основу разматрања идентичног проблема у случају модела вишеструке регресије, наводећи да је интерпретација мултиваријационог модела изводљива само уколико је истим обухваћено приближно 10 независних променљивих. Повезаност претходна два става потврђују и *Brown & Tinsley* (1983; цитиран у *Brown & Wicker*, 2000, стр. 213) који истичу да је комплексност инетрпретације резултата дискриминационе анализе директно сразмерна повећању броја дискриминатор променљивих изабраних за оцењивање дискриминационог модела. Такође, али посматрано из угла репрезентативности дискриминационог модела и његове предиктивне прецизности, важно је напоменути да се повећање количника броја независних променљивих и укупне величине узорка одликује тенденцијом смањења прецизности дискриминације у условима када се дискриминациона функција утврђена на једном узорку примењује за потребе класификације јединица посматрања другог узорка извученог из исте популације (*Horst*, 1941; цитиран у *Huberty*, 1975, стр. 555).

Полазећи од издвојених, у извесној мери контрадикторних ставова, евидентно је да у релевантној литератури, још увек, није формулисан „оптималан“ одговор на разматрано

методолошко питање. Начелно, потребно је анализом обухватити довољно велики број, теоријски и / или емпиријски релевантних, променљивих који ће обезбедити поуздане резултате, али тако да тај број буде „довољно мали“ како интерпретација истих не би била превише комплексна.

Поред наведеног, питање избора конкретних независних променљивих, како са аспекта њиховог броја тако и њихове структуре, у значајној мери је детерминисано резултатима испитивања специфичних универијационих и мултиваријационих статистичких својстава којима независне променљиве, као „кандидати“ за укључивање у поступак спровођења ДА, морају да се одликују. Другим речима, као и у случају већине осталих мултиваријационих метода, валидна апликација ДА заснива се на одговарајућим статистичким претпоставкама. У циљу обезбеђивања научне заснованости и генерализације резултата линеарне ДА, неопходно је проверити и осигурати одговарајући степен испуњености следећих претпоставки на нивоу узорка опсервација у контексту независних променљивих:

- ✓ Независност опсервација у узорку;
- ✓ Присуство статистички значајне линеарне повезаности између променљивих;
- ✓ Одсуство мултиколинеарности и сингуларности између независних променљивих;
- ✓ Одсуство универијационих и мултиваријационих нестандардних опсервација, како на нивоу укупног узорка тако и на нивоу појединачних група јединица посматрања;
- ✓ Универијациона нормалност распореда појединачних независних променљивих, како на нивоу укупног узорка тако и издвојених група у саставу зависне променљиве;
- ✓ Нормалност мултиваријационог распореда разматраних независних променљивих;
- ✓ Хомогеност матрица варијанси / коваријанси независних променљивих на нивоу група опсервација у структури зависне променљиве;
- ✓ Обезбеђивање „адекватне“ величине појединачних група у саставу узорка и његове укупне величине.

Будући да припада групи параметарских мултиваријационих метода, једна од кључних претпоставки ДА подразумева да расположиве мултиваријационе опсервације потичу из популације која следи мултиваријациону нормалну расподелу у контексту разматраних независних променљивих. У условима одрживости наведене претпоставке, параметарски тестови коришћени за тестирање статистичке значајности различитих оцена параметара у ДА, одликују се највећом јачином у поређењу са било којим другим статистичким тестовима. Теоријски, њена нарушеност угрожиће валидност резултата тестирања статистичке значајности појединачних дискриминационих функција, а самим тим и резултата класификације опсервација, из угла прецизности (*Betz*, 1987, стр. 401; *Sharma*, 1996, стр. 263; *Hair et al.*, 2010, стр. 251). Детаљно објашњење наведених негативних ефеката анормалности мултиваријационог распореда опсервација представљено је у *Klecka* (1980, стр. 60). Иако постоје ставови и истраживања који сугеришу робустност наведених аспеката ДА у случају мање изражене нарушености ове претпоставке (видети у: *Tabachnick & Fidell*, 2013, стр. 384), у литератури ипак није прецизно дефинисан степен мултиваријационе анормалности распореда који се може сматрати прихватљивим и без значајнијег утицаја на ефикасност ДА и прецизност класификације (*Sharma*, 1996, стр. 263–264). Генерално, уколико је претпоставка о мултиваријационој нормалности одржива, линеарно класификационо правило изведено на

основу формиране дискриминационе функције представља оптимално средство за дискриминацију издвојених група у саставу зависне променљиве по критеријуму минималних очекиваних трошкова погрешне класификације (*Klecka*, 1980, стр. 61; *Kovačić*, 1994, стр. 167). У том смислу, уколико постоје оправдани разлози и јасни докази који сугеришу нарушеност наведене претпоставке, а спроведени приступи за ублажавање идентификоване мултиваријационе анормалности се покажу неефикасним, неопходно је размотрити могућност примене неке од алтернативних, непараметарских мултиваријационих метода попут, на пример, логистичке регресионе анализе (*Sharma*, 1996, стр. 264; *Hair et al.*, 2010, стр. 251; *Pituch & Stevens*, 2016, стр. 425).[100]

Поред наведеног, важност елабориране претпоставке, у контексту примене ДА, огледа се и њеном значају за поступак провере одрживости претпоставке о једнакости популационих матрица варијанси / коваријанси дефинисаних група у саставу зависне променљиве. Наиме, готово сви статистички тестови који се уобичајено користе за испитивање ове претпоставке изузетно су осетљиви на евентуално присутну мултиваријациону анормалност. Прецизније, у таквим околностима, одбацивање нулте хипотезе о хомогености матрица коваријанси може уследити као последица анормалности распореда, пре него постојања стварних неједнакости (*Huberty & Olejnik*, 2006, стр. 280; *Timm*, 2009, стр. 431). Дискриминациона анализа је веома осетљива на нарушеност ове претпоставке, нарочито аспекти који се односе на процес оцењивања дискриминационе функције и прецизност класификације (*Hair et al.*, 2010, стр. 251). Наиме, како се израчунавање дискриминационих коефицијената у саставу линеарне дискриминационе функције заснива на коришћењу оцене опште (заједничке и једнаке) коваријационе матрице свих група, идентификовање статистички значајних разлика између матрица коваријанси на нивоу популације, резултираће погрешним вредностима дискриминационих коефицијената, тако изведена функција неће обезбеђивати максималну сепарацију између група, а прецизност класификације биће озбиљно нарушена (*Klecka*, 1980, стр. 60). У условима нарушености разматране претпоставке, *Varmuza & Filzmoser* (2009, стр. 213) наглашавају да резултирајуће класификационо правило, засновано на коришћењу линеарне дискриминационе функције, више не представља оптимално средство за дискриминацију издвојених група, посматрано из угла минимизирања укупне вероватноће погрешне класификације јединица посматрања. Алтернативни приступ који се у таквим околностима показао оптималним, мада у извесној мери и компликованијим за примену у поређењу са претходним, подразумева формирање квадратне дискриминационе функције и на њој заснованих квадратних класификационих правила (*Sharma*, 1996, стр. 264; *Kovačić*, 1994, стр. 167; *Timm*, 2009, стр. 439; *Hair, et al.*, 2010, стр. 251).[101]

Коначно, веома важно питање, посматрано из угла репрезентативности оцењеног ДА модела и могућности генерализације добијених резултата, односи се на обезбеђивање адекватне величине узорка, n, као и величине појединачних група у његовом саставу, n_k (за k = 1, 2,..., g, где g означава број модалитета зависне променљиве). Дефинисање „пожељних“ и минимално прихватљивих величина наведених категорија у ДА, углавном

[100] *Izenman* (2008, стр. 256) обезбеђује детаљан компаративни приказ карактеристика линеарне дсикриминационе анализе и логистичке (дискриминационе) анализе.

[101] Специфичности везане за избор између линеарне или квадратне ДА, посматрано из угла односа величине појединачних група у саставу узорка и броја коришћених независних променљивих, презентиране су од стране *Huberty & Olejnik* (2006, стр. 281).

је засновано на уважавању броја коришћених независних променљивих, p, и консеквентно, исказивању вредности њиховог релативног односа у форми n / p, односно n_k / p. Наиме, велики број аутора (*Hair et al.*, 2010, стр. 249; *Pituch & Stevens*, 2016, стр. 396), позивајући се на резултате спроведених студија, сугерише да би величина узорка (n), који ће бити коришћен за спровођења ДА, требало да задовољи критеријум представљен количником n / $p \approx 20 : 1$. Исти аутори напомињу да ће употреба узорка чија је укупна величина, посматрано из угла броја независних променљивих, значајно мања од датог критеријума резултирати израженом нестабилношћу дискриминационих коефицијената и немогућношћу генерализације добијених резултата класификације. Другим речима, дискриминациона функција и класификационо правило утврђено на једном таквом узорку неће бити практично применљиви, из угла остварене прецизности класификације, на другом (то јест, новом) узорку извученом из исте популације. Уз уважавање изнетих ставова, мада у извесној мери дискутабилно и можда контрадикторно, *Hair et al.* (2010, стр. 249) наводе однос 5:1, у корист броја јединица посматрања, као минимално прихватљиву вредност коју рацио n/p може узети. С друге стране, *Brown & Tinsley* (1983; цитирано у: *Brown & Wicker*, 2000, стр. 214) предлажу да доњу границу вредности овог рациа треба поставити на нивоу 10:1. Сублимација изложене дискусије садржана је у препоруци коју су дали *Brown & Wicker* (2000, стр. 214), а према којој се адекватном сматра величина узорка за коју је рацио n / p између 10:1 и 20:1.

Поред наведеног, будући да ДА не захтева нужно да (под)узорци који одговарају појединачним категоријама зависне променљиве (n_k) унутар узорка буду једнаке величине, неопходно је узети у обзир и препоруке у погледу броја јединица посматрања које оне морају обухватити. У том контексту, *Poulsen & French* (2008, стр. 3), *Huberty & Olejnik* (2006, стр. 309), *Hair et al.* (2010, стр. 249) и *Pituch & Stevens* (2016, стр. 396) препоручују да најмања група у саставу зависне променљиве садржи минимум 20 јединица посматрања, односно, исказано у односу на број независних променљивих, да релација $min(n_k) \geq 5p$, буде задовољена. У случају практичне немогућности испуњавања наведене релације, *Tatsuoka* (1970, цитирано у: *Huberty*, 1975, стр. 558), *Brown & Tinsley* (1983; цитирано у: *Brown & Wicker*, 2000, стр. 214), *Pituch & Stevens* (2016, стр. 396) и *Hair, et al.* (2010, стр. 249), истичу, као апсолутни минимум, да број јединица посматрања унутар најмање групе не сме бити мањи од броја коришћених дискриминатора, односно, $min(min(n_k)) \geq p$.

Питање обезбеђивања адекватне и/или минимално прихватљиве величине узорка и појединачних група у његовом саставу, не може се посматрати одвојено од поступка спровођења екстерне валидације дикриминационог модела и резултата класификације (који је предмет дискусије у оквиру Одељка 3.2.6), заснованом на креирању одговарајућег узорка за валидацију. Сходно наведеном, важно је нагласити да изложене препоруке морају бити задовољене не само на нивоу узорка за анализу већ и узорка који ће бити коришћен за валидацију перформанси изведеног дискриминационог модела (*Hair et al.*, 2010, стр. 250).

С тим у вези, полазећи од претпоставке да ће (под)узорци за анализу и валидацију бити случајним путем издвојени из једног укупног узорка, *Huberty* (1975, стр. 558) формулише минимално прихватљив критеријум у погледу величине појединачних група у структури укупног узорка према којем број јединица посматрања у најмањој групи не би требао да буде мањи од троструке вредности броја коришћених независних променљивих,

односно, $min(min(n_k)) \geq 3p$. Минимална укупна величина узорка у том случају износила би $min(n) = 3pg$, где p представља број независних променљивих, а g број група у структури зависне променљиве. *Huberty & Olejnik* (2006, стр. 309) у разматрање укључују и коришћени конкретни метод валидације и истичу да је претходни критеријум, $min(min(n_k)) \geq 3p$, погодан у случају примене метода унакрсне валидације (енгл. *cross-validation method*), док за примену метода задржавања (енгл. *holdout method*) као погоднију, предлажу минимално прихватљиву величину најмање групе која задовољава релацију $min(min(n_k)) \geq 5p$, уз минималну укупну величину узорка $min(n) = 5pg$.

3.2.3. Поступак оцењивања линеарног дискриминационог модела

Након избора зависне променљиве Y сачињене од одговарајућих категорија k (за k = 1, 2,..., g), селекције скупа независних променљивих X_j (за j = 1, 2, ..., p), формирања узорка за анализу и провере испуњености елабориранихпретпоставки, централна статистичка активност у поступку спровођења ДА односи се на креирање одговарајућег оцењеног дискриминационог модела. Наиме, полазећи од расположивог узорка мултиваријационих опсервација величине n, представљених у матричној форми Табелом 3.2.1, основна идеја ове етапе огледа се у формирању једне или више одговарајућих линеарних дискриминационих варијата, које представљају пондерисани збир две или више независних променљивих, за које је прелиминарном анализом утврђено да располажу израженом дискриминационом снагом, односно потенцијалом за раздвајање јединица посматрања у саставу конкретних, *a priori* издвојених категорија зависне променљиве. Изведене линеарне латентне комбинације, у контексту ДА такође познате под називом (каноничке) дискриминационе функције, или дискриминанте, представљају статистичку мултиваријациону основу за описивање разлика између издвојених група, идентификовање дискриминационог доприноса појединачних независних променљивих, али и предвиђање припадности нових јединица посматрања једној од конкретних група.

Табела 3.2.1. Матрица мултиваријационих података у дискриминационој анализи

Јединице посматрања у узорку (n)	Независне променљиве (X_j)				Зависна променљива
	X_1	X_2	…	X_p	Y_k
1	x_{111}	x_{121}	…	x_{1p1}	1
2	x_{211}	x_{221}	…	x_{2p1}	1
⋮	⋮	⋮	…	⋮	⋮
n_1	$x_{n_1 11}$	$x_{n_1 21}$	…	$x_{n_1 p1}$	1
1	x_{112}	x_{122}	…	x_{1p2}	2
2	x_{212}	x_{222}	…	x_{2p2}	2
⋮	⋮	⋮	…	⋮	⋮
n_2	$x_{n_2 12}$	$x_{n_2 22}$	…	$x_{n_2 p2}$	2
1	x_{11g}	x_{12g}	…	x_{1pg}	g
2	x_{21g}	x_{22g}	…	x_{2pg}	g
⋮	⋮	⋮	…	⋮	⋮
n_g	$x_{n_g 1g}$	$x_{n_g 2g}$	…	$x_{n_g pg}$	g

Извор: Ауторов приказ

У табели коришћени симболи означавају:

Y_k – зависна променљива чији су модалитети означени симболом k (за k = 1, 2,..., g);

g – укупан број група (категорија, или модалитета) у саставу зависне променљиве;

X_j – j-та независна променљива (за j = 1, 2,..., p);

p – укупан број независних променљивих обухваћених анализом;

x_{ijk} – вредност j-те независне променљиве за i-ту јединицу посматрања у оквиру k-те групе зависне променљиве (за i = 1, 2,..., n_k и k = 1, 2,..., g);

n_k – величина појединачних група у саставу зависне променљиве (за k = 1, 2,..., g);

n – величина укупног узорка, при чему је $n = n_1 + n_2 + ... + n_g = \Sigma n_k$ (за k = 1, 2,..., g).

Генерално, математички облик оцењене линеарне дискриминационе функције може се представити у форми линеарне једначине која гласи (*Brown & Wicker*, 2000, стр. 219):

$$Z_m = b_{0m} + b_{1m}X_1 + b_{2m}X_2 + \ldots + b_{pm}X_p, \tag{3.2.1}$$

где коришћени симболи означавају:

Z_m – вредност m-те дискриминационе функције која се још назива и дискриминациони Z скор (енгл. *discriminant Z scor*), при чему је m = 1,..., r, а $r = min(g–1, p)$;

b_{0m} – оцењена вредност параметра одсечка у оквиру m-те дискриминационе функције;

b_{jm} – вредност дискриминационог коефицијената (пондера) уз j-ту независну променљиву у структури m-те дискриминационе функције;

X_j – j-та независна променљива (за j = 1, 2,..., p).

Максимални број дискриминационих функција који може бити изведен у оквиру дискриминационог модела директно зависи од броја група у саставу зависне променљиве и броја коришћених независних променљивих, будући да се одређује као $min(g–1, p)$. Број дискриминационих функција, одређених у циљу обезбеђивања максималне раздвојености група јединица посматрања на бази измерених вредности независних променљивих, назива се ранг или димензионалност дискриминације (енгл. *rank / dimensionality of discrimination*).

Извођење оцењене дискриминационе функције заснива се на израчунавању оцењених вредности дискриминационих коефицијената b_j, (енгл. *discriminant coefficients / weights*) односно емпиријских тежинских коефицијената којима се пондерише степен у којем свака појединачна независна променљива доприноси детерминисању припадности јединица посматрања конкретној групи. Прецизније, поступак одређивања дискриминационих функција заснива се на одређивању m (p×1) вектора дискриминационих коефицијената $\mathbf{b}_m$ (за m = 1,..., r), тако да се обезбеди максимизирање релативног односа варијација између и унутар анализом обухваћених група, познатог под називом *Fisher*-ов дискриминациони критеријум, који се може представити коришћењем следећег израза (*Everitt*, 2010, стр. 266):

$$\frac{\mathbf{b}_m' \mathbf{B} \mathbf{b}_m}{\mathbf{b}_m' \mathbf{W} \mathbf{b}_m}. \tag{3.2.2}$$

У наведеном изразу, мера интергрупне хомогености представљена је (p×p) матрицом суме квадрата одступања мултиваријационих опсервација од просечних вредности

независних променљивих унутар група и узајамних производа, у ознаци **W**, која се утврђује путем следећег израза:

$$\underset{(p\times p)}{\mathbf{W}} = \sum_{k=1}^{g}\left(n_k - 1\right)\mathbf{S}_k = \left(n_1 - 1\right)\mathbf{S}_1 + \left(n_2 - 1\right)\mathbf{S}_2 + \ldots + \left(n_g - 1\right)\mathbf{S}_g \,. \tag{3.2.3}$$

Симбол $\mathbf{S}_k$ означава одговарајућу ($p\times p$) коваријациону матрицу на нивоу појединачних група k (за k = 1,2,..., g) у структури узорка за анализу, односно:

$$\underset{(p\times p)}{\mathbf{S}_k} = \frac{1}{n_k - 1}\begin{bmatrix} \sum_{i=1}^{n_k}\left(x_{i1} - \overline{x}_1\right)^2 & \sum_{i=1}^{n_k}\left(x_{i1} - \overline{x}_1\right)\left(x_{i2} - \overline{x}_2\right) & \cdots & \sum_{i=1}^{n_k}\left(x_{i1} - \overline{x}_1\right)\left(x_{ip} - \overline{x}_p\right) \\ \sum_{i=1}^{n_k}\left(x_{i2} - \overline{x}_2\right)\left(x_{i1} - \overline{x}_1\right) & \sum_{i=1}^{n_k}\left(x_{i2} - \overline{x}_2\right)^2 & \cdots & \sum_{i=1}^{n_k}\left(x_{i2} - \overline{x}_2\right)\left(x_{ip} - \overline{x}_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n_k}\left(x_{ip} - \overline{x}_p\right)\left(x_{i1} - \overline{x}_1\right) & \sum_{i=1}^{n_k}\left(x_{ip} - \overline{x}_p\right)\left(x_{i2} - \overline{x}_2\right) & \cdots & \sum_{i=1}^{n_k}\left(x_{ip} - \overline{x}_p\right)^2 \end{bmatrix}. \tag{3.2.4}$$

У изразу (3.2.2), симболом **B** представљена је мера међугрупне хетерогености (односно, мера варијација између група), у форми ($p\times p$) матрице суме квадрата одступања просечних вредности независних променљивих на нивоу појединачних група од кореспондентних просечних вредности на нивоу укупног узорка за анализу и узајамних производа, односно:

$$\underset{(p\times p)}{\mathbf{B}} = \sum_{k=1}^{g} n_k \left(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}\right)', \tag{3.2.5}$$

где $\overline{\mathbf{X}}_k$ означава ($p\times 1$) вектор просечних вредности p независних променљивих у k-тој групи, односно $\overline{\mathbf{X}}_k' = [\overline{x}_{1k},\ \overline{x}_{2k},\ldots,\ \overline{x}_{pk}]$, док $\overline{\mathbf{X}}$, представља ($p\times 1$) вектор опште средине p независних променљивих на нивоу узорка за анализу, симболички $\overline{\mathbf{X}}' = [\overline{x}_1,\ \overline{x}_2,\ldots,\ \overline{x}_p]$. Еквивалентно, матрични приказ израза (3.2.5) гласи:

$$\underset{(p\times p)}{\mathbf{B}} = \sum_{k=1}^{g} n_k \begin{bmatrix} \left(\overline{x}_{1k} - \overline{x}_1\right)^2 & \left(\overline{x}_{1k} - \overline{x}_1\right)\left(\overline{x}_{2k} - \overline{x}_2\right) & \cdots & \left(\overline{x}_{1k} - \overline{x}_1\right)\left(\overline{x}_{pk} - \overline{x}_p\right) \\ \left(\overline{x}_{2k} - \overline{x}_2\right)\left(\overline{x}_{1k} - \overline{x}_1\right) & \left(\overline{x}_{2k} - \overline{x}_2\right)^2 & \cdots & \left(\overline{x}_{2k} - \overline{x}_2\right)\left(\overline{x}_{pk} - \overline{x}_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\overline{x}_{pk} - \overline{x}_p\right)\left(\overline{x}_{1k} - \overline{x}_1\right) & \left(\overline{x}_{pk} - \overline{x}_p\right)\left(\overline{x}_{2k} - \overline{x}_2\right) & \cdots & \left(\overline{x}_{pk} - \overline{x}_p\right)^2 \end{bmatrix}, \tag{3.2.6}$$

где је:

$$\overline{x}_{jk} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ijk}\,, \qquad \text{за } i = 1,2,\ldots, n_k,\ j = 1,2,\ldots, p \text{ и } k = 1,2,\ldots, g; \text{ и} \tag{3.2.7}$$

$$\overline{x}_j = \frac{1}{\sum_{k=1}^{g} n_k}\sum_{k=1}^{g} n_k \overline{x}_{jk}\,, \qquad \text{за } j = 1,2,\ldots, p \text{ и } k = 1,2,\ldots, g. \tag{3.2.8}$$

Генерално, математички проблем израчунавања вектора дискриминационих коефицијената $\mathbf{b}_m$, у контексту максимизирања израза (3.2.2), своди се на проналажење карактеристичних вредности (енгл. *eigenvalues*) λ_m (за m = 1,..., r) и придружених ($p\times 1$)

карактеристичних вектора (енгл. *eigenvectors*) $\mathbf{b}_m^*$ несиметричне ($p\times p$) матрице $\mathbf{W}^{-1}\mathbf{B}$, утврђене на основу израза (3.2.3) и (3.2.6). Наиме, решавањем карактеристичне једначине p-тог реда (енгл. *eigenequation* или *characteristic equation*) квадратне ($p\times p$) матрице $\mathbf{W}^{-1}\mathbf{B}$, представљене изразом (3.2.9), могуће је одредити максимално $min(g–1, p)$ позитивних (ненултих) карактеристичних вредности λ [102] за које је следећа једнакост задовољена (*Kovačić*, 1994, стр. 148):

$$\left|\mathbf{W}^{-1}\mathbf{B}-\lambda\mathbf{I}\right| = \left|\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}\right| = \left|\begin{bmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp}-\lambda \end{bmatrix}\right| = 0 \,. \quad (3.2.9)$$

$$\left|\mathbf{W}^{-1}\mathbf{B}-\lambda\mathbf{I}\right| = (-1)^p\left[\lambda^p + c_1\lambda^{p-1} + c_2\lambda^{p-2} + \ldots + c_p\lambda + c_{p+1}\right] = 0 \,. \quad (3.2.9a)$$

Израз (3.2.9) представља детерминанту новоформиране матрице $\mathbf{[W^{-1}B}–\lambda\mathbf{I]}$, односно карактеристични полином реда p по λ (израз (3.2.9a)), симболи $\{a_{ij}\}$ означавају елементе ($p\times p$) матрице $\mathbf{W^{-1}B}$, а λ вредност каракатеристичног корена матрице $\mathbf{W^{-1}B}$. Након одређивања карактеристичних вредности λ_m, решавањем хомогеног система једначина представљеног изразом (3.2.10), врши се одређивање сваког од r придружених карактеристичних вектора $\mathbf{b}_m^*$ матрице $\mathbf{W}^{-1}\mathbf{B}$, односно (*Huberty & Olejnik*, 2006, стр. 83):

$$\underset{(p\times p)}{\left(\mathbf{W}^{-1}\mathbf{B}-\lambda_m\mathbf{I}\right)}\underset{(p\times 1)}{\mathbf{b}_m^*} = \begin{bmatrix} a_{11}-\lambda_m & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22}-\lambda_m & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp}-\lambda_m \end{bmatrix}\begin{bmatrix} b_{1m}^* \\ b_{2m}^* \\ \vdots \\ b_{pm}^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (3.2.10)$$

Коначно, нормализацијом карактеристичних вектора $\mathbf{b}_m^*$ за сваку од r карактеристичних вредности λ, путем израза (3.2.11), одређују се елементи ($p\times 1$) вектора $\mathbf{b}_m$, односно оцењене вредности дискриминационих коефицијената b_j, придружених свакој од p независних променљивих, X_j (за j = 1,2,..., p), у структури m-те дискриминационе функције (за m = 1,..., r), односно (прилагођено према *Kovačić*, 1994, стр. 73):

$$\underset{(p\times 1)}{\mathbf{b}_m} = \begin{bmatrix} b_{1m} \\ b_{2m} \\ \vdots \\ b_{pm} \end{bmatrix} = \frac{1}{\sqrt{\mathbf{b}_m^{*\mathrm{T}}\,\overline{\mathbf{S}}\,\mathbf{b}_m^*}}\mathbf{b}_m^* \,. \quad (3.2.11)$$

[102] Број могућих решења за утврђени карактеристични полином p-тог реда једнак је заправо броју независних променљивих, p. Међутим, нека од тих решења ће бити математички тривијална решења, односно $\lambda = 0$, која се сматрају неупотребљивим у контексту дискриминационе анализе. Прецизније, за p независних променљивих и g група јединица посматрања, када је $p > (g–1)$, увек ће бити $(p – g + 1)$ решења карактеристичног полинома која имају карактеристичну вредност λ једнаку нули, односно, када је $p < (g–1)$, тада ће бити $(g–1–p)$ тривијалних решења карактеристичног полинома. Максималан број дискриминационих функција, m, који може бити изведен, односно број решења за које је $\lambda > 0$, јесте мањи број од следећа два броја: p и $(g–1)$. Наиме, када се од броја могућих решења (p) одузме број тривијалних решења ($p–g+1$, односно, $g–1–p$), максимални број решења која ће се одликовати позитивним вредностима λ увек ће бити $min(p, g–1)$ (*Klecka*, 1980, стр. 34).

У претходном изразу, симбол $\bar{\mathbf{S}}$ означава непристрасну оцену опште коваријационе матрице $\boldsymbol{\Sigma}$, односно заједничку узорачку коваријациону матрицу, која се одређује као пондерисани просек збира реализованих вредности коваријационих матрица појединачних група ($\mathbf{S}_1$, $\mathbf{S}_2$,..., $\mathbf{S}_g$) у саставу узорка за анализу (израз (3.2.12)). Логика њеног израчунавања заснована је на испуњености полазне претпоставке о хомогености коваријационих матрица популација из којих су случајним путем формирани (под)узорци односно групе у структури зависне променљиве (*Johnson & Wichern*, 2007, стр. 310):

$$\underset{(p\times p)}{\bar{\mathbf{S}}} = \mathbf{S}_{pooled} = \frac{1}{n-g}\mathbf{W} \tag{3.2.12}$$

Поред тежинских коефицијената b_j, саставна, али генерално не и „неизоставна“ компонента оцењеног модела линеарне дискриминационе функције, представљене изразом (3.2.1), јесте и оцењена вредност параметра одсечка, b_0. Наведена константа уводи се из практичних разлога како би се обезбедило изједначавање просечне вредности дискриминационих Z скорова за све јединице посматрања узорка за анализу са нулом (*Huberty & Olejnik*, 2006, стр. 100), односно за потребе прилагођавања и подешавања мерне скале новоформиране, латентне варијабле Z (*Sharma*, 1996, стр. 252; *Brown & Wicker*, 2000, стр. 219), а израчунава, за m-ту дискриминациону функцију, путем следећег израза (*Huberty & Olejnik*, 2006, стр. 100):

$$b_0 = -\sum_{j=1}^{p} b_j \bar{x}_j \tag{3.2.13}$$

Оцењене вредности дискриминационих коефицијената b_j, утврђене презентованим поступком, називају се нестандардизовани дискриминациони коефицијенти [103], а резултирајућа варијата (израз (3.2.1)) нестандардизована оцена линеарне дискриминационе функције. Њихова вредност показује просечну апсолутну промену (повећање или смањење) вредности дискриминационе функције Z, односно дискриминационог скора z_i, уколико се вредност конкретне независне променљиве повећа за једну своју јединицу мере, под претпоставком да су вредности преосталих променљивих у функцији непромењене. Иако неопходни, заједно са одсечком b_0, за израчунавање дискриминационог скора z_i (за i = 1, 2,..., n) и, сходно томе, за класификацију и предвиђање припадности опсервација датим групама, нестандардизовани коефицијенти b_j, не могу бити коришћени за рангирање независних променљивих према јачини дискриминационе моћи у контексту сепарације група. Разлог томе налази се у различитом варијабилитету којим се одликују, али и различитим мерним јединицама у којима се исказују вредности независних променљивих. Решење је садржано у трансформацији нестандардизованих коефицијената у њихове стандардизоване супституенте, коришћењем следећег израза (*Huberty & Olejnik*, 2006, стр. 88; *Sharma*, 1996, стр. 253):[104]

---

[103] Термин „нестандардизовани“ употребљен је како би се истакла чињеница да су за оцену дискриминационе функције коришћени оригинални подаци анализираних независних променљивих, исказани у припадајућим, углавном различитим, мерним јединицама, односно који нису преведени у стандардизован облик (*Klecka*, 1980, стр. 23).

[104] Алтернативни начин одређивања стандардизованих дискриминационих коефицијената јесте и путем израза (3.2.11) под условом да су за потребе одређивања карактеристих вектора $\mathbf{b}_m$ и свих пратећих матрица, коришћене

$$\underset{(p\times 1)}{\mathbf{b}_m^{std}} = \underset{(p\times p)}{\left(diag\,\bar{\mathbf{S}}\right)^{1/2}}\,\underset{(p\times 1)}{\mathbf{b}_m} \Rightarrow \begin{bmatrix} b_{1m}^{std} \\ b_{2m}^{std} \\ \vdots \\ b_{pm}^{std} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(x_{i1}-\bar{x}_1\right)^2}{n-g} & 0 & \cdots & 0 \\ 0 & \frac{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(x_{i2}-\bar{x}_2\right)^2}{n-g} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(x_{i2}-\bar{x}_2\right)^2}{n-g} \end{bmatrix}^{1/2} \begin{bmatrix} b_{1m} \\ b_{2m} \\ \vdots \\ b_{pm} \end{bmatrix}. \quad (3.2.14)$$

У представљеном изразу, симбол b_{jm} (елемент вектора $\mathbf{b}_m$) означава j-ти нестандардизовани коефицијент у оквиру m-те дискриминационе функције (за j = 1, 2,..., p; и m = 1,..., r), b_{jm}^{std} одговарајући j-ти стандардизовани коефицијент у оквиру m-те дискриминационе функције (елемент вектора $\mathbf{b}_m^{std}$), а елементи матрице $(diag\,\bar{\mathbf{S}})^{1/2}$ представљају квадратне корене дијагоналних елемената опште коваријационе матрице $\bar{\mathbf{S}}$, (израз (3.2.12)).

Стандардизовани дискриминациони коефицијенти, нису под утицајем мерних јединица независних променљивих и, сходно томе, испитивањем њихових вредности, поређењем и рангирањем, могуће је остварити јасан увид у релативни значај и допринос појединачних независних променљивих дискриминационој снази изведене функције која обезбеђује максимално раздвајање група јединица посматрања.

У том контексту, независне променљиве које се одликују већим апсолутним вредностима стандардизованих пондера више доприносе, односно имају већи релативни допринос, дискриминационој снази формиране функције него дискриминатор променљиве са нижим вредностима коефицијената. Прецизније, већа апсолутна вредност стандардизованог коефицијента репрезентује јачи дискриминациони допринос конкретне независне променљиве моделираној дискриминацији група, и обратно. Предзнак вредности стандардизованих коефицијената није од интереса за сагледавање јачине дискриминационог доприноса појединачних променљивих, будући да исти указује само на смер (позитиван или негативан) поменутог доприноса.

Поред наведеног приступа, релативни значај независних променљивих у дискриминационој функцији може се испитати и на основу вредности структурних коефицијената корелације (енгл. *structure correlation coefficients*), који се називају и дискриминациона оптерећења (енгл. *discriminant loadings*). Ови релативни показатељи, у ознаци $r_{X_jZ_m}$, представљају меру јачине просте линеарне корелације између варијација појединачно посматраних независних променљивих и вредности дискриминационих z_i скорова сваке од m формираних дискриминационих функција. Њихово израчунавање спроводи се применом једног од следећа два алтернативна приступа, еквивалентна са аспекта исхода, заснована на коришћењу стандардизованих (израз (3.2.15)) или нестандардизованих дискриминационих коефицијената (израз (3.2.17)), тим редом.

стандардизоване вредности независних променљивих, чија средња вредност износи нула а стандардна девијација је једнака јединици.

$$\underset{(p\times 1)}{\mathbf{r}_{XZ}} = \underset{(p\times p)}{\mathbf{R}}\ \underset{(p\times 1)}{\mathbf{b}_m^{std}} \Rightarrow \begin{bmatrix} r_{X_1Z_m} \\ r_{X_2Z_m} \\ \vdots \\ r_{X_pZ_m} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{bmatrix} \begin{bmatrix} b_{1m}^{std} \\ b_{2m}^{std} \\ \vdots \\ b_{pm}^{std} \end{bmatrix}. \tag{3.2.15}$$

У представљеном изразу (*Sharma*, 1996, стр. 254), симболом **R** је означена ($p\times p$) симетрична узорачка корелациона матрица анализом обухваћених независних променљивих X_j, чији елементи $\{r_{js}\}$ означавају биваријационе коефицијенте линеарне корелације између j-те и s-те независне променљиве, утврђене коришћењем следеће формуле:

$$r_{js} = \frac{\overline{s}_{js}}{\sqrt{\overline{s}_{jj}}\sqrt{\overline{s}_{ss}}}, \text{ за } j, s = 1,2,..., p, \tag{3.2.16}$$

у којој су елементи у бројиоцу и имениоцу одговарајући (j, s)-ти, (j, j)ти и (s, s)-ти елементи опште узорачке коваријационе матрице $\overline{\mathbf{S}}$. Алтернативно, приступ у израчунавању дискриминационих оптерећења $r_{X_jZ_m}$, заснован на употреби нестандардизованих дискриминационих коефицијената спроводи се путем следећег израза:

$$\underset{(p\times 1)}{\mathbf{r}_{XZ}} = \left(diag\,\overline{\mathbf{S}}\right)^{-1/2}_{(p\times p)}\ \underset{(p\times p)}{\overline{\mathbf{S}}}\ \underset{(p\times 1)}{\mathbf{b}_m} \Rightarrow \begin{bmatrix} r_{X_1Z_m} \\ r_{X_2Z_m} \\ \vdots \\ r_{X_pZ_m} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\overline{s}_{11}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\overline{s}_{22}}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\overline{s}_{pp}}} \end{bmatrix} \begin{bmatrix} \overline{s}_{11} & \overline{s}_{12} & \dots & \overline{s}_{1p} \\ \overline{s}_{21} & \overline{s}_{22} & \dots & \overline{s}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{s}_{p1} & \overline{s}_{p2} & \dots & \overline{s}_{pp} \end{bmatrix} \begin{bmatrix} b_{1m} \\ b_{2m} \\ \vdots \\ b_{pm} \end{bmatrix}. \tag{3.2.17}$$

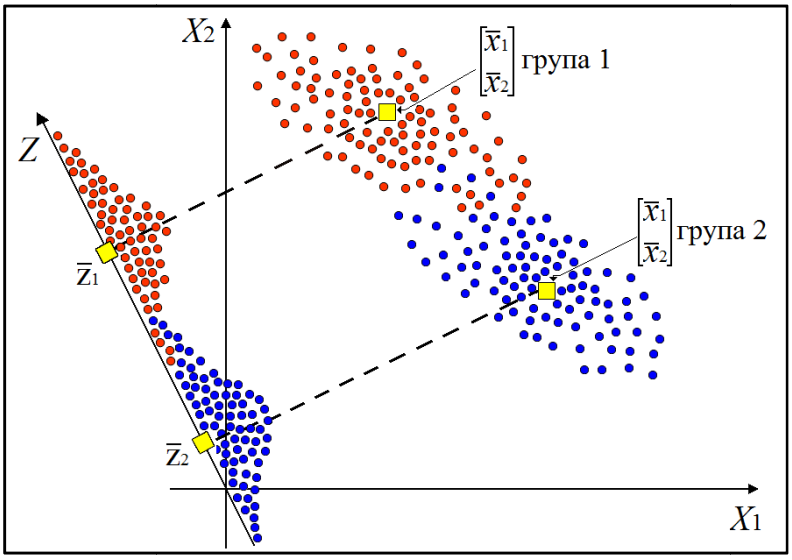
Вредности дискриминационих оптерећења налазе се у опсегу од –1 до +1, при чему већа апсолутна вредност указује на већи дискриминациони допринос конкретне независне променљиве у поређењу са дискриминаторима који се одликују нижим вредностима овог показатеља, док предзнак, као и у случају стандардизованих коефицијената, показује смер датог доприноса при израчунавању дискриминационих z_i скорова. Када су апсолутне вредности ових коефицијената прилично високе, $|r_{X_jZ_m}| \approx 1$, дискриминациона функција поседује готово идентична дискриминациона својства као и конкретна независна променљива, и обратно (*Klecka*, 1980, стр. 31). Слично факторским оптерећењима, квадрат структурних коефицијената корелације рефлектује удео укупног варијабилитета конкретне независне променљиве, присутног између група зависне променљиве, који је објашњен, односно обухваћен m-том дискриминационом функцијом. Посматрано из угла дефинисања минимално прихватљиве вредности коју дискриминациона оптерећења треба да досегну како би се дата независна променљива могла сматрати значајним дискриминатором, а њено укључивање у модел оправдано, у литератури су присутни блиски, углавном субјективно детерминисани, али ипак неусаглашени ставови. Најчешће предложене доње границе дискриминационих оптерећења су: $|r_{X_jZ_m}| \geq 0{,}50$ (*Sharma*, 1996, стр. 254) и $|r_{X_jZ_m}| \geq 0{,}40$ (*Hair et al.*, 2010, стр. 266).

У литератури су присутни опречни ставови у погледу оправданости употребе и предности која се додељује једном од ова два коефицијента у поступку сагледавања

степена релативног доприноса појединачних независних променљивих дискриминацији група. Наиме, дискусија је примарно усмерена на анализу и компарацију њихове осетљивости на евентуално присуство мултиколинеарности али и стабилност њихових вредности у условима коришћења малих узорака за анализу и, на тим основама, фаворизовању једног од њих за потребе интерпретације формиране дискриминационе функције. [105] Међутим, у условима када су у потпуности испуњене статистичке претпоставке на којима се заснива апликација ДА, важно је истаћи да ће вредности стандардизованих и структурних коефицијената, по правилу, сугерисати сличне, а у идеалном случају и идентичне закључке, због чега се углавном врши њихова симултана употреба и компарација сугерисаних закључака.

3.2.4. Испитивање статистичке и практичне значајности оцењеног ДА модела

Полазећи од елаборираног правила у погледу максималног броја дискриминационих функција који може бити изведен у контексту одређеног истраживачког проблема, може се констатовати да ће ДА заснована на двема групама увек резултирати само једном дискриминационом функцијом ($r = 1$), док ће, у случају примене вишегрупне ДА, број изведених дискриминационих функција у саставу модела бити $r \geq 2$. Свака од изведених r дискриминационих варијата обезбеђује пројекцију i-тих оригиналних мултиваријационих опсервација на новоформирану(е) Z_m осу(е). Наведене пројекције називају се дискриминациони z_{im} скорови. Графички приказ презентиране идеје дискриминационе анализе представљен је на Слици 3.2.3. Истраживачки проблем раздвајања g група у p-димензионом простору, своди се, представљеним поступком креирања пондерисаних линеарних комбинација p независних променљивих, на проблем раздвајања група у саставу зависне променљиве у r-димензионом дискриминационом простору који конституишу формиране Z_m композитне променљиве (за $m = 1,..., r$ и $r = min(p, g–1)$), при чему је свака од њих засебна димензија дискриминације, некорелисана у односу на остале.



Слика 3.2.3. Илустративни приказ исхода хипотетитичког раздвајања две групе јединица посматрања за две независне променљиве у контексту дискриминационе анализе

Извор: Ауторов приказ (прилагођено према *Johnson & Wichern* (2007, стр. 592))

[105] Детаљнија разматрања овог методолошког питања презентована су од стране, на пример, следећих аутора: *Rencher* (2002, стр. 288-291), *Pituch & Stevens* (2016, стр. 395-396), *Hair et al.* (2010, стр. 266).

Посматрано у контексту примене вишегрупне ДА, неопходно је истаћи да се издвојене дискриминације функције међусобно разликују према дискриминационој снази којом се одликују, односно према доприносу у процесу раздвајања издвојених група јединица посматрања. Величина карактеристичне вредности λ матрице $\mathbf{[W^{-1}B]}$ директно је повезана са дискриминационом снагом функције изведене на основу придруженог карактеристичног вектора. Наиме, дискриминациона функција која се одликује највећом вредношћу λ представља најснажнију дискриминанту, док се функција изведена на основу најмање λ вредности сматра најслабијом дискриминантом (*Klecka*, 1980, стр. 34). По правилу, прва дискриминанта формира се на основу највеће вредности λ и, сходно томе, њоме је обухваћен и објашњен највећи удео варијабилитета који постоји између група у p-димензионом систему. Другим речима, прва дискриминациона функција у r-дискриминационом простору одликује се најповољнијим (то јест, највећим) односом варијабилитета између и унутар група. Сваком наредном функцијом обухваћен је мањи део међугрупних разлика у поређењу са претходном, али уједно већи у односу на преостале функције које се одликују мањом карактеристичном вредношћу. Наведене констатације подржане су чињеницом да карактеристична вредност m-те функције, λ_m, представља количник суме квадрата одступања између и унутар група израчунатих на основу кореспондентних дискриминационих z_i скорова функције Z_m, (*Brown & Wicker*, 2000, стр. 223; *Sharma*, 1996, стр. 293; *Rencher*, 2002, стр. 278) односно:

$$\lambda_m = \frac{\left(SS_b\right)_{Zm}}{\left(SS_w\right)_{Zm}} = \frac{\sum_{k=1}^{g} n_k \left(\overline{z}_{km} - \overline{z}_m\right)^2}{\sum_{k=1}^{g}\sum_{i=1}^{n_k} \left(z_{ikm} - \overline{z}_{km}\right)^2}, \text{ за } m=1,\ldots, r. \qquad (3.2.18)$$

Генерално, редослед изведених функција према дискриминационој снази, односно доприносу у поступку раздвајања група може се симболички приказати на следећи начин, рангирањем њима припадајућих карактеристичних вредности λ_m (за m=1,...,r), односно: $\lambda_1 > \lambda_2 > \ldots > \lambda_r$. Међутим, иако ће дискриминационом анализом бити генерисан максимално могући број (r) дискриминационих функција, и то у опадајућем низу, посматрано из угла њихове важности и доприноса сепрацији група, сасвим је могуће да поједине међу њима не резултирају статистички и / или практично значајним разликама просечних вредности припадајућих дискриминационих z_i скорова на нивоу дефинисаних група. Уколико се у обзир узме и допунски циљ вишегрупне ДА, који подразумева редукцију дискриминационог простора на најмањи могући број димензија довољних за квалитетно репрезентовање разлика које постоје између група, статистички је сасвим оправдано и неопходно, након спроведеног поступка оцењивања линеарних дискриминационих функција, извршити тестирање статистичке значајности њихове дискриминационе моћи. Поменутим поступком тестирања испитује се способност изведене(их) дискриминационе(их) функције(а) да резултира(ју) дискриминационим скоровима између чијих ће просечних вредности на нивоу дефинисаних група постојати статистички значајне разлике. Наиме, тестира се статистичка значајност карактеристичних вредности λ_m, како би се обезбедила емпиријска потврда да њихова позитивна (ненулта) вредност није последица узорачке или одређене неузорачке грешке. Иако се за спровођење поступка тестирања статистичке значајности дискриминационе снаге функције(а), генерално, може користити било који од претходно елаборираних *MANOVA*

мултиваријационих тестова (Поглавље 3.1), у наставку је указано на специфичности примене, како у домену спровођења ДА засноване на двема групама, тако и вишегрупне ДА, *Bartlett*-ове V статистике теста, која представља χ^2 апроксимацију *Wilks*–ове Λ статистике теста[106].

У контексту дискриминације две групе јединица посматрања, поменута статистика теста, која следи χ^2 распоред са $p(g-1)$ степени слободе, има следећи облик:

$$\chi^2 = -\left[n-1-(p+g)/2\right]\ln\Lambda\,. \tag{3.2.19}$$

У датом изразу, симбол Λ означава *Wilks*-ову Λ статистику теста, одређену изразом:

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{T}|} = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{1}{1+\lambda}, \tag{3.2.20}$$

где $|\cdot|$ представљају детерминанте респективних матрица, а симболом **T** означена је матрица укупне суме квадрата одступања и узајамних производа, која представља збир матрица **W** и **B**, а λ је карактеристични корен матрице $\mathbf{W^{-1}B}$, који одговара изведеној дискриминационој функцији. Кореспондента нулта и алтернативна хипотеза *Bartlett*-ове статистике теста гласе:

$$H_0: \begin{bmatrix} \mu^1_{X_1} \\ \mu^1_{X_2} \\ \vdots \\ \mu^1_{X_p} \end{bmatrix} = \begin{bmatrix} \mu^2_{X_1} \\ \mu^2_{X_2} \\ \vdots \\ \mu^2_{X_p} \end{bmatrix} \Rightarrow H_0: \lambda = 0; \quad \text{и} \quad H_1: \begin{bmatrix} \mu^1_{X_1} \\ \mu^1_{X_2} \\ \vdots \\ \mu^1_{X_p} \end{bmatrix} \neq \begin{bmatrix} \mu^2_{X_1} \\ \mu^2_{X_2} \\ \vdots \\ \mu^2_{X_p} \end{bmatrix} \Rightarrow H_1: \lambda \neq 0 \tag{3.2.21}$$

Уколико је реализовани ниво значајности мањи од дефинисаног ризика грешке α, H_0 се одбацује и, сходно томе, може се закључити да између вектора средина $\boldsymbol{\mu_1}$ и $\boldsymbol{\mu_2}$, односно просечних вредности p независних променљивих на нивоу посматране две групе у популацији, постоји статистички значајна разлика, односно, потврђена је статистичка значајност оцене карактеристичне вредности λ. Будући да изведена дискриминациона функција Z представља пондерисану линеарну комбинацију посматраних независних променљивих, може се такође закључити да је и формирана дискриминациона функција статистички значајна, односно, да се просечне вредности резултирајућих дискриминационих скорова z_i за посматране две групе међусобно статистички значајно разликују.[107]

С друге стране, примена вишегрупне ДА резултира извођењем више од једне дискриминационе функције, при чему се свака одликује одговарајућим карактеристичним кореном матрице $\mathbf{W^{-1}B}$, у ознаци λ_m, за које важи релација: $\lambda_1 > \lambda_2 > ... > \lambda_r > 0$. Полазећи од чињенице да свака од r карактеристичних вредности репрезентује једну од r идентификованих димензија могуће сепарације вектора средина унутар дефинисаних

[106] У поређењу са преосталим *MANOVA* статистикама, *Wilks'* Λ статистика се сматра у извесној мери погоднијом, нарочито при оцени сигнификантности функција изведених као резултат вишегрупне ДА (*Rencher*, 2002, стр. 285).

[107] Неуспех у поступку одбацивања H_0 сугерише да нема довољно емпиријских доказа којима би се потврдило постојање статистички значајних разлика вектора просечних вредности на нивоу посматране две групе у контексту датих независних променљивих, односно да нема довољно доказа да изведена дискриминациона функција обезбеђује сигнификантно раздвајање посматране две групе.

група, неопходно је испитати да ли је иједна од њих статистички сигнификантна, и ако јесте, која(е) је(су) то дискриминациона(е) функција(е). За разлику од случаја када је само једна функција изведена, поступак тестирања статистичке значајности у вишегрупној ДА не заснива се на испитивању сигнификантности сваке појединачне функције већ статистичке значајности свих изведених функција, посматраних заједно. Сходно наведеном, *MANOVA* статистика теста заснована на *Wilks*-овој Λ статистици, која се користи за спровођење поменуте процедуре тестирања, може се представити као следећа функција карактеристичних вредности λ_m (*Klecka*, 1980, стр. 39):

$$\Lambda_m = \prod_{m=1}^{r} \frac{1}{1+\lambda_m} = \frac{1}{1+\lambda_1}\frac{1}{1+\lambda_2}\ldots\frac{1}{1+\lambda_r}, \text{ за } m = 1,2,\ldots, r. \qquad (3.2.22)$$

Коришћењем горње формуле, резултирајућа *Bartlett*-ова χ^2 апроксимација са $(p–m+1)(g–m)$ степени слободе, има следећи облик (*Klecka*, 1980, стр. 40; *Sharma*, 1996, стр. 300):

$$\chi^2 = -\left[n-1-\left(p+g\right)/2\right]\ln\Lambda_m . \qquad (3.2.23)$$

Представљена χ^2 статистика користи се при спровођењу вишеетапног, итеративног поступка тестирања чији је сумарни приказ презентиран Табелом 3.2.2. Наиме, у полазној етапи, врши се испитивање истинитости нулте хипотезе која, у општем облику, гласи: H_{0_1}: $\lambda_1 = \lambda_2 = \ldots = \lambda_r = 0$ (*Timm*, 2009, стр. 436). Тако формулисана нулта хипотеза еквивалентна је *MANOVA* хипотези о једнакости вектора средина g група у популацији, односно, H_0: $\boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \ldots = \boldsymbol{\mu_g}$. Будући да је кореспондентном *Wilks*' Λ_1 статистиком обухваћено свих r карактеристичних корена λ_m јасно је да се не испитује статистичка значајност само једне функције већ свих могућих r функција заједно, што потврђује и текстуална формулација хипотеза у поменутој табели. Уколико резултати тестирања сугеришу одбацивање H_0, може се закључити да се најмање једна од карактеристичних вредности λ_m статистички значајно разликује од нуле и да, сходно томе, постоји барем једна димензија дискриминације, односно дискриминациона функција, која раздваја векторе средина на нивоу дефинисаних група. Прецизније, будући да λ_1 представља највећу карактеристичну вредност, изведени закључак односи се на њену статистичку значајност, а самим тим и дискриминациону функцију Z_1, изведену на основу коефицијената карактеристичног вектора придруженог карактеристичном корену λ_1. Након спроведене прве етапе, поставља се питање: Да ли је само функција Z_1 довољна за описивање разлика између група, или је можда још нека од карактеристичних вредности преосталих функција, λ_2, λ_3,..., λ_r, статистички значајна? У том контексту, за потребе испитивања тачности новоформиране нулте хипотезе, облика: H_{0_2}: $\lambda_2 = \lambda_3 = \ldots = \lambda_r = 0$, израчунава се вредност *Wilks*' Λ_2 статистике, путем формуле која представља модификацију претходне, јер не укључује карактеристичну вредност λ_1, и њој придружене *Bartlett*-ове апроксимације која следи χ^2 распоред са $(p–1)(g–2)$ степени слободе. Уколико резултати потврде статистичку значајност карактеристичне вредности λ_2, као највеће у низу посматраних вредности λ_m (за $m = 2,\ldots, r$), представљени поступак се итеративно спроводи, тестирањем статистичке значајности преосталих карактеристичних вредности λ_m, до тренутка када добијени резултати не сугеришу да не постоји довољно доказа за одбацивање H_0, или пак до тренутка потврде статистичке значајности дискриминационог доприноса свих изведених функција.

Табела 3.2.2. Сумарни приказ поступка тестирања статистичке значајности дискриминационих функција у случају примене вишегрупне ДА

Дефинисање нулте и алтернативне хипотезе		*Wilks'* Λ_m статистика
симболички	текстуално	број степени слободе χ^2 апроксимације
H_{0_1}: $\forall\lambda_m \in \{\lambda_1, \lambda_2,\ldots, \lambda_r\}$: $\lambda_m = 0$ H_{a_1}: $\exists\lambda_m \in \{\lambda_1, \lambda_2,\ldots, \lambda_r\}$: $\lambda_m \neq 0$	H_{0_1}: Не постоје статистички значајне разлике између средина z_i скорова посматраних група ни за једну од ф-ја; H_{a_1}: Најмање једна функција је потребна за интерпретирање разлика између група;	$\Lambda_1 = \prod_{m=1}^{r} \frac{1}{1+\lambda_m}$ $v = p(g-1)$
H_{0_2}: $\forall\lambda_m \in \{\lambda_2, \lambda_3,\ldots, \lambda_r\}$: $\lambda_m = 0$ H_{a_2}: $\exists\lambda_m \in \{\lambda_2, \lambda_3,\ldots, \lambda_r\}$: $\lambda_m \neq 0$	H_{0_2}: Највише једна функција је потребна за интерпретирање разлика између група; H_{a_2}: Најмање две функције су потребне за интерпретирање разлика између група;	$\Lambda_2 = \prod_{m=2}^{r} \frac{1}{1+\lambda_m}$ $v = (p-1)(g-2)$
⋮	⋮	⋮
$H_{0_{r-1}}$: $\forall\lambda_m \in \{\lambda_{r-1}, \lambda_r\}$: $\lambda_m = 0$ $H_{a_{r-1}}$: $\exists\lambda_m \in \{\lambda_{r-1}, \lambda_r\}$: $\lambda_m \neq 0$	$H_{0_{r-1}}$: Највише r-2 функције су потребне за интерпретирање разлика између група; $H_{a_{r-1}}$: Најмање r-1 функција је потребно за интерпретирање разлика између група	$\Lambda_{r-1} = \prod_{m=r-1}^{r} \frac{1}{1+\lambda_m}$ $v = [p-(r-1)+1][g-(r-1)]$
H_{0_r}: $\lambda_r = 0$ H_{a_r}: $\lambda_r \neq 0$	H_{0_r}: Највише r-1 функција је потребно за интерпретирање разлика између група; H_{a_r}: Најмање r функција је потребно за интерпретирање разлика између група;	$\Lambda_r = \prod_{m=r}^{r} \frac{1}{1+\lambda_m}$ $v = (p-r+1)(g-r)$

Извор: Ауторов приказ

Будући да се представљеном формалном статистичком процедуром одређивања статистичке сигнификантности, индиректно, одређује и укупан број статистички значајних дискриминационих димензија које могу бити употребљене за описивање разлика између дефинисаних група, односно детерминише димензионалност дискриминационог простора за потребе интерпретирања резултата ДА, ова процедура се назива и *тест димензионалности* (енгл. *test of dimensionality*) или *анализа редукције димензија* (*Timm*, 2009, стр. 436). Важно је такође истаћи да се интерпретација али и утврђивање прецизности класификације и предикције, заснива превасходно на употреби статистички значајних дискриминационих функција, идентификованих овом процедуром. Наведени критеријум јесте неопходан, међутим, не и довољан услов који дискриминационе димензије треба да задовоље.

Наиме, полазећи од изражене осетљивости тестова статистичке значајности на величину узорка, али и могућности допунске редукције димензионалности дискриминационог простора, корисно је испитати практични значај дискриминационог доприноса функција чија је статистичка значајност у поступку сепарације група претходно потврђена. У условима када су коришћени узорци велике величине, могуће је да се, чак и не толико изражене, разлике присутне између вектора средина посматраних група идентификују као статистички значајне (*Sharma*, 1996, стр. 302). Консеквентно, не морају све дискриминационе функције које објашњавају статистички значајан „део“ варијабилитета између група уједно бити и од практичног значаја у поступку раздвајања група (*Brown & Wicker*, 2000, стр. 222).

Један од начина за испитивање практичне значајности дискриминационих димензија подразумева израчунавање и интерпретирање вредности коефицијената каноничке корелације r^*, који показују степен повезаности дефинисаних група у саставу зависне

променљиве и појединачних, статистички значајних, дискриминационих функција, односно варијација њима припадајућих дискриминационих z_i скорова (*Brown & Wicker*, 2000, стр. 224). Њихово израчунавање заснива се на коришћењу карактеристичне вредности λ_m конкретно посматране функције Z_m, путем израза (*Rencher*, 2002, стр. 282; *Klecka*, 1980, стр. 36):

$$r_m^* = \sqrt{\frac{\lambda_m}{1+\lambda_m}}, \tag{3.2.24}$$

алтернативно, укључивањем израза (3.2.18) (*Sharma*, 1996, стр. 253):

$$r_m^* = \sqrt{\frac{(SS_b)_{Zm}}{(SS_t)_{Zm}}} = \sqrt{\frac{\sum_{k=1}^{g} n_k \left(\overline{z}_{km} - \overline{z}_m\right)^2}{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(z_{ikm} - \overline{z}_m\right)^2}}, \tag{3.2.25}$$

где симболи SS_t и SS_b означавају укупну и суму квадрата одступања дискриминационих скорова функције Z_m између група, респективно. Узимајући у обзир опсег вредности које овај коефицијент може узети, $0 \leq r^* \leq 1$, високе вредности (блиске јединици) сугеришу да између линеарне комбинације независних променљивих, представљене у форми дискриминационе функције, и дефинисаних категорија зависне променљиве постоји висок ниво квантитативног слагања, односно повезаности, и обратно уколико је $r^* \approx 0$ (*Klecka*, 1980, стр. 36).

Такође, за потребе сагледавања практичног значаја и „величине“ дискриминационог доприноса појединачних функција, од користи је и интерпретација квадриране вредности коефицијента каноничке корелације, у ознаци $(r^*)^2$. Наведена вредност показује удео, односно пропорцију укупне суме квадрата одступања дискриминационих скорова функције Z_m, која се може приписати варијацијама скорова између група, што се јасно може уочити на основу садржаја израза (3.2.25). Вредности које овај релативни показатељ може узети крећу се у интервалу 0 до 1, при чему већа вредност имплицира и већу практичну значајност посматране функције у контексту њеног доприноса сепарацији група, и обратно. Генерално, не постоје јасно дефинисане величине ова два показатеља које би имале улогу граничних вредности приликом доношења одлуке о задржавању или елиминисању конкретне дискриминационе функције из даљег разматрања. Наведена граница углавном је условљена специфичном природом проблема који се истражује (*Kovačić*, 1994, стр. 92). *Sharma* (1996, стр. 253) допуњује претходно изнети став и предлаже коришћење кореспондентних вредности добијених у студијама сличног карактера као основе за поређење и закључивање у погледу репрезентативности дискриминационих функција које су предмет анализе.

У циљу обезбеђивања провере и/или додатне верификације закључака о репрезентативности појединачних димензија дискриминације, изведених на бази претходних показатеља, често се користе и, такозвани, показатељи апсолутног и релативног процента објашњене варијансе (енгл. *absolute / relative proportion of variance*), засновани на λ_m вредностима (*Brown & Wicker*, 2000, стр. 224). Заправо, стављањем у однос збира карактеристичних вредности свих појединачних функција ($\Sigma\lambda_m$) са бројем независних променљивих коришћених у анализи, p, (лева страна израза (3.2.26)), одређује

се релативна величина укупног варијабилитета између група која је генерално обухваћена и објашњена изведеним дискриминационим функцијама. Заступљеност, односно допринос појединачних функција у акумулирању, овако одређеног, укупног варијабилитета између група утврђује се дељењем вредности сваког појединачног карактеристичног корена λ са бројем дискриминатор променљивих. Резултирајућа вредност назива се *апсолутни проценат објашњене варијансе* на нивоу појединачних функција (десна страна израза (3.2.26)) и показује релативну, будући да се исказује у процентима, величину међугрупног варијабилитета која је обухваћена и објашњена конкретном дискриминационом функцијом у односу на укупну количину варијабилитета који је присутан између група. Сходно начину на који се дефинише, дискриминациона функција која се одликује већом вредношћу овог показатеља има већи практични значај и важност за раздвајање група у поређењу са функцијом коју карактерише мања кореспондентна вредност. Однос појединачних и укупног апсолутног процента објашњене варијансе може се приказати следећом једначином:

$$\frac{\sum_{m=1}^{r}\lambda_m}{p}=\frac{\lambda_1}{p}+\frac{\lambda_2}{p}+\ldots+\frac{\lambda_r}{p}. \qquad (3.2.26)$$

За разлику од појединачних форми овог показатеља које су погодне за компарацију и сагледавање практичног значаја дискриминационог доприноса појединачних функција у структури дискриминационог модела, укупан апсолутни проценат објашњене варијансе представља користан показатељ за оцену ефективности независних променљивих у поступку сепарације. Разумевање статистичке логике у основи овог показатеља предуслов је за разумевање, израчунавање и интерпретацију мере репрезентативности дискриминационих функција, под називом *релативни проценат објашњене варијансе*, а који се одређује, полазећи од израза (3.2.26), на следећи начин (*Rencher*, 2002, стр, 278; *Sharma*, 1996, стр. 302):

$$\frac{\frac{\lambda_m}{p}}{\frac{\sum_{m=1}^{r}\lambda_m}{p}}=\frac{\lambda_m}{\sum_{m=1}^{r}\lambda_m} \quad\Rightarrow\quad \frac{\lambda_1}{\sum_{m=1}^{r}\lambda_m},\ \frac{\lambda_2}{\sum_{m=1}^{r}\lambda_m},\ldots,\ \frac{\lambda_r}{\sum_{m=1}^{r}\lambda_m}. \qquad (3.2.27)$$

Сходно презентираном изразу, ова мера показује релативно учешће дела, (λ_m/p), укупног варијабилитета између група обухваћеног дискриминационим моделом, ($\Sigma\lambda_m/p$), који је објашњен конкретном дискриминационом функцијом у саставу модела. Генерално, веће вредности овог показатеља повезане су са већим практичним доприносом конкретне функције и обратно. Термин „генерално“ употребљен је са разлогом како би се истакла важност интерпретирања вредности овог показатеља у контексту утврђене вредности укупног апсолутног процента објашњене варијансе и, сходно томе, избегло доношење погрешних закључака у погледу практичног значаја појединачних функција. Наиме, могуће је да статистички значајна функција, која се одликује високим релативним процентом објашњене варијансе, заправо објашњава релативно мали део „реалног“ варијабилитета који постоји између група, уколико је релативна величина укупног варијабилитета између група која је обухваћена изведеним дискриминационим

функцијама, односно моделом ($\Sigma\lambda_m/p$), мала (*Brown & Wicker*, 2000, стр. 224). Наиме, релативни проценат објашњене варијансе, уколико се посматра изоловано од апсолутног процента, може резултирати искривљеном сликом дискриминационих доприноса појединачних функција у саставу модела.

Полазећи од наведеног, за потребе спровођења редукције ранга димензионалности, користан је приступ заснован на постепеном кумулирању релативних процената објашњене варијансе појединачних вредности λ_m, у опадајућем низу, посматрано из угла величине њихове вредности, односно (*Huberty & Olejnik*, 2006, стр. 91):

$$\frac{\lambda_1}{\sum_{i=1}^{r}\lambda_i} \rightarrow \frac{\lambda_1+\lambda_2}{\sum_{i=1}^{r}\lambda_i} \rightarrow \frac{\lambda_1+\lambda_2+\lambda_3}{\sum_{i=1}^{r}\lambda_i} \ldots \qquad (3.2.28)$$

Кумулативна вредност којом је обухваћен „знатан“ удео укупног варијабилитета између група који је објашњен дискриминационим моделом, може послужити као аргумент за смањење броја дискриминационих функција у оквиру модела, односно искључивање оних које нису садржане у израчунатом кумулативу, превасходно услед чињенице да се одликују релативно ниским карактеристичним вредностима (*Huberty & Olejnik*, 2006, стр. 91).

Комплексност презентиране дискусије указује на неопходност уважавања великог броја показатеља и њихову комбиновану примену у циљу извођења квалитетних и веродостојних закључака о практичном значају појединачних дискриминационих функција изведених за потребе сепарације посматраних група јединица посматрања. Као и у случају коефицијената каноничке корелације, тако и у контексту апсолутног и релативног процента објашњене варијансе, не постоје дефинисани општи критеријуми у погледу минимално прихватљиве вредности поменутих, чије би незадовољавање иницирало изостављање конкретне функције из даље дискриминационе процедуре. Генерално, питање броја дискриминационих функција које треба задржати, након евентуално извршене редукције током поступка испитивања статистичке значајности, повезано је са субјективним ставом истраживача, формулисаним циљевима анализе и природом проблема који се истражује. Свакако, прецизно и исправно тумачење елабориираних показатеља у значајној мери олакшава доношење такве одлуке.

3.2.5. Процедуре за класификацију јединица посматрања

У складу са дефинисаним циљевима и извршеном поделом на дескрипривни и предиктивни аспект квантитативних активности у оквиру дискриминационе анализе, након дискусије усмерене на поступак и статистичку сврху извођења одговарајућег броја дискриминационих функција Z_m и испитивање њихове статистичке и практичне значајности за потребе идентификовања и/или описивања разлика које постоје између одговарајућих група јединица посматрања у контексту анализом обухваћених независних променљивих, излагање у наставку текста посвећено је предиктивним аспектима ДА. Наведени аспект ДА, често у литератури означен терминима класификација, алокација или разврставање, има важну улогу у ситуацијама када за одређене јединице посматрања информација о њиховој припадности једној од конкретних група обухваћених анализом није расположива, али је ипак од значаја одредити, односно предвидети категорију

зависне променљиве којој највероватније исте припадају. У оквирима ДА, класификација нових, некатегорисаних опсервација подразумева претходно спровођење адекватне и комплексне, статистичке сумаризације мултиваријационе структуре података, засноване на јединицама посматрања за које је позната припадност конкретној групи, у циљу извођења одговарајућих класификационих правила (енгл. *classification rules*), такозваних класификатора (енгл. *classifiers*), који ће представљати основу за реализацију циљева ПДА. Другим речима, примарни циљ класификационе анализе јесте успостављање, на бази расположивог узорка груписаних мултиваријационих опсервација, одговарајућих класификационих правила, која ће бити коришћена, са одговарајућом поузданошћу, за предвиђање припадности „нових“, односно „некатегорисаних“ јединица посматрања, које се одликују истим скупом мерења у контексту независних променљивих, једној од анализом обухваћених група. Сходно наведеном, може се закључити да класификација у домену ДА представља надгледану технику, будући да извођење класификатора захтева располагање информацијама о припадности сваке јединице посматрања у узорку за анализу одређеној категорији зависне променљиве. Изведена класификациона правила треба да буду, у што је могуће већој мери, поуздана, како би се осигурала ефикасна реализација предиктивних циљева ДА, посматрано из угла минимизирања вероватноће настанка погрешних класификација и/или трошкова изазваних погрешном класификацијом опсервација[108]. За спровођење класификационе процедуре и извођење класификатора, у оквиру линеарне дискриминационе анализе, најчешће се користе следећа два приступа: ▸ метод заснован на коришћењу дискриминационих функција; и ▸ метод заснован на коришћењу линеарних класификационих функција на нивоу појединачних категорија зависне променљиве.

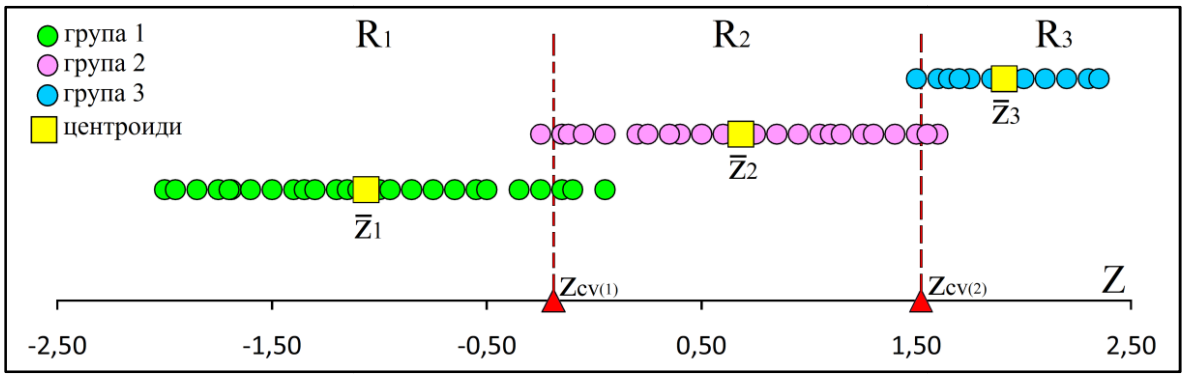
Класификациона процедура заснована на коришћењу дискриминационих Z функција

Употреба дискриминационе Z функције, изведене у оквиру ДДА (израз (3.2.1)), за потребе дефинисања класификационог правила намењеног алоцирању нових опсервација, заснива се на примени фундаменталне логике према којој, јединице посматрања које се одликују сличним дискриминационим скоровима треба да буду распоређене унутар исте групе и обратно, у случају постојања изразитих разлика између кореспондентних дискриминационих z_i вредности појединачних опсервација (*Babin*, 2011, стр. 446). Класификационом процедуром, заснованом на дикриминационим функцијама, врши се сагледавање позиције дискриминационог z_i скора сваке појединачне опсервације узорка за анализу (коришћеног за извођење дискриминационе функције) у односу на центроиде група[109] у дискриминационом простору и, сходно томе, распоређивање истих у састав

[108] Под очекиваним трошковима погрешне класификације (енгл. *expected costs of misclassification*, *ECM*) подразмевају се практичне импликације алокације јединице посматрања у састав групе којој реално не припада, изражене у материјалном, емотивном или неком другом облику (*Klecka*, 1980, стр. 46). Будући да су у пракси трошкови погрешне класификације углавном непознати (*Timm*, 2009, стр. 424), у оквиру излагања у овом одељку, елаборирање поступка извођења класификационих правила у ДА засновано је искључиво на претпоставци о једнакости трошкова погрешне класификације, према којој последице класификовања јединице посматрања из, на пример, групе 1 у групу 2, нису ништа више или пак мање озбиљније у односу на обратну ситуацију. Другим речима, представљена класификациона процедура заснована је на коришћењу класификатора који обезбеђује минимизирање укупне вероватноће погрешних класификација (енгл. *total probability of misclassification*, *TPM*).

[109] Центроид, у ознаци $\overline{z_k}$, представља просечну вредност дискриминационих скорова опсервација унутар конкретне групе, односно k-те категорије зависне променљиве Y_k (за k = 1, 2,…, g), (*Brown & Wicker*, 2000, стр. 219). Полазећи од чињенице да дискриминациона функција представља линеарну функцију p независних променљивих X_j (израз

конкретне групе чијем центроиду је њихова позиција „најближа“ (*Klecka*, 1980, стр. 42). У том смислу, неопходно је формирати класификационо правило које ће обезбедити „оптималну“ поделу дискриминационог простора на g међусобно искључивих области, у ознаци R_k (где k представља број група, k=1,2,…, g) чија унија покрива комплетан простор узорка (*Sharma*, 1996, стр. 244), тако да, уколико оригинална опсервација $\mathbf{x}_i=[x_{i1}, x_{i2},\ldots, x_{ip}]$, то јест, њена дискриминациона пројекција z_i, припада области R_k, (симболички, $z_i \in R_k$) тада се може констатовати да конкретна јединица посматрања припада групи y_k (симболички, $\mathbf{x}_i \in y_k$). Слика 3.2.4. илуструје изложену идеју у основи класификационе процедуре, у форми хипотетичког примера вишегрупне ДА засноване на посматрању три групе и само једној, статистички значајној, дискриминационој Z функцији.



Слика 3.2.4. Хипотетички једнодимензиони дијаграм дискриминационих скорова, центроида група и R_k области у случају примене вишегрупне ДА

Извор: Ауторов приказ

Заправо, дискриминациони простор, детерминисан једном дискриминационом варијаблом Z, подељен је испрекиданим црвеним линијама, које се зову линије пресека (енгл. *cutoff lines*), на области R_1, R_2 и R_3. Вредност дискриминационе Z функције која репрезентује граничну линију између две R_k области представља, такозвану, вредност пресека (енгл. *cutoff value / cutting score value / critical Z score*). Вредност пресека, у ознаци z_{cv}, представља критеријум са којим се упоређује дискриминациони скор сваке појединачне опсервације (то јест, z_i пројекција p-димензионе i–те опсервације $\mathbf{x}_i$) у циљу идентификовања групе у коју дата опсервација највероватније треба да буде класификована (*Hair et al.*, 2010, стр. 257).

Израчунавање критичне вредности дискриминационе Z функције која обезбеђује, из угла минимизирања броја погрешних класификација, „оптимално“ раздвајање било које две k-те групе, од укупно g група у структури зависне променљиве, засновано је на испуњености полазних статистичких претпоставки које се односе на хомогеност матрица коваријанси група и мултиваријациону нормалност распореда популације из које је случајним путем извучен узорак јединица посматрања за анализу (*Hair et al.*, 2010, стр. 259). Општи поступак за израчунавање вредности пресека за било који пар, од укупно g група, може се представити на следећи начин (*Babin*, 2011, стр. 446):

(3.2.1)), центроид k-те групе, за конкретно посматрану дискриминациону функцију Z_m, еквивалентан је вредности линеарне функције просечних вредности независних променљивих на нивоу k-те групе, односно: $\bar{z}_k = b_0 + \sum b_j \bar{x}_{jk}$ (*Pituch & Stevens*, 2016, стр. 397).

$$z_{cv(1)} = \frac{n_2\overline{z}_1 + n_1\overline{z}_2}{n_1 + n_2}, \quad z_{cv(2)} = \frac{n_3\overline{z}_2 + n_2\overline{z}_3}{n_2 + n_3}, \ldots, \quad z_{cv(g-1)} = \frac{n_g\overline{z}_{g-1} + n_{g-1}\overline{z}_g}{n_{g-1} + n_g}, \tag{3.2.29}$$

где $z_{cv\,(1)}$, $z_{cv\,(2)}$,…, $z_{cv\,(g\text{-}1)}$, означавају одговарајуће вредности пресека између групе 1 (y_1) и групе 2 (y_2), групе y_2 и y_3, односно $y_{g\text{-}1}$ и y_g, респективно. Симболи n_1, n_2,..., n_g показују број јединица посматрања унутар група y_1, y_2,..., y_g, а симболи $\overline{z}_1$, $\overline{z}_2$,..., $\overline{z}_g$ њихове центроиде.

На основу приказаних формула, евидентна је условљеност „позиције“ пресека између конкретне две групе од њихових појединачних величина n_k и припадајућих вредности центроида. У том смислу, при спровођењу класификационе процедуре засноване на коришћењу дискриминационих Z функција, важно је указати на улогу коју имају величине посматраних група и важност разматрања њиховог релативног односа на нивоу узорка у комбинацији са припадајућим *a priori* вероватноћама на нивоу популације.[110] Наиме, кључно питање на које треба пружити одговор гласи: „да ли релативни однос величина две посматране групе на нивоу узорка адекватно репрезентује њихов однос у популацији из које потичу?“ Могући одговори различито усмеравају поступак израчунавања вредности пресека z_{cv}. Прецизније, уколико не постоји довољан степен сигурности у репрезентативност релативног односа величина посматране две групе опсервација у узорку, или се пак, на бази искуства и / или претходних истраживања, може прихватити одрживост претпоставке о једнакости *a priori* вероватноћа посматране две групе у популацији ($\pi_1 = \pi_2$) независно од односа величина тих група на нивоу узорка (односно, независно од тога да ли је: $n_1 \neq n_2 \rightarrow n_1 / n_2 \neq 1 \rightarrow p_1 \neq p_2$, или, $n_1 = n_2 \rightarrow n_1 / n_2 \approx 1 \rightarrow p_1 = p_2$), за израчунавање вредности пресека између те две групе користи се модификована верзија формуле дате изразом (3.2.29), која гласи (*Hair et al.*, 2010, стр. 258):

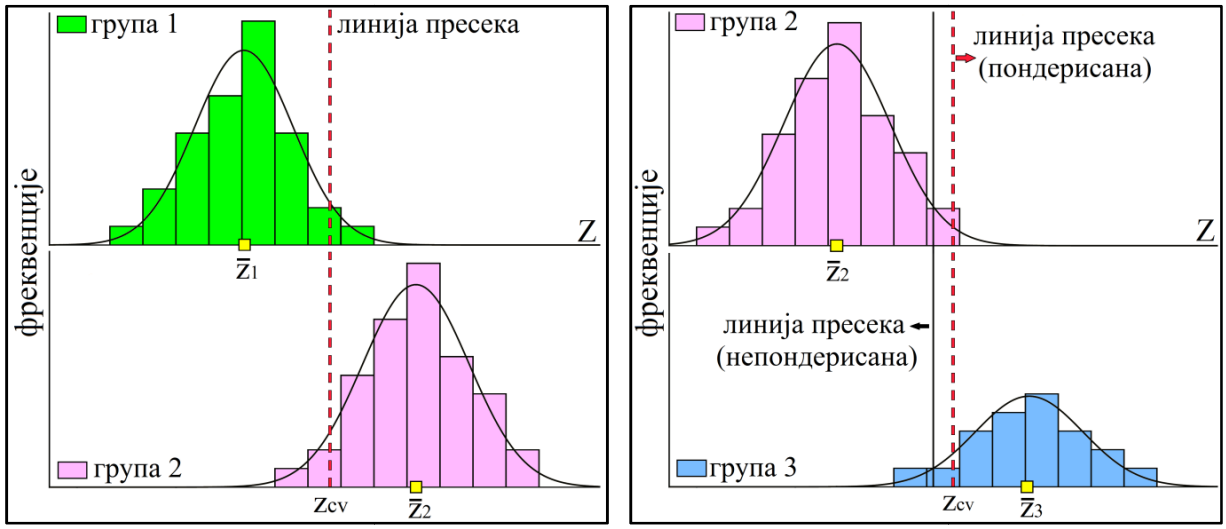
$$z_{cv} = \frac{\overline{z}_1 + \overline{z}_2}{2}. \tag{3.2.30}$$

Насупрот овим ситуацијама у којима се врши израчунавање, такозване, непондерисане вредности пресека (израз (3.2.30)), уколико се у формираном случајном узорку стварне величине група, иако међусобно различите, сматрају репрезентативним у контексту описивања њиховог релативног односа на нивоу популације, односно, ако су p_1 и p_2 непристрасне оцене *a priori* вероватноћа π_1 и π_2, тада се величине посматране две групе n_1 и n_2 користе директно за израчунавање пондерисане вредности пресека путем израза (3.2.29).

У контексту хипотетичког примера на Слици 3.2.4, а у циљу додатног истицања важности адекватног уважавања оцењених вредности *a priori* вероватноћа група, на Слици 3.2.5, илустрован је поступак одређивања критичних вредности дискриминационе Z функције у случају коришћења група једнаких и неједнаких величина, уз претпоставку да величине група n_k адекватно репрезентују однос *a priori* вероватноћа група у популацији. Наиме, будући да су групе y_1 и y_2 једнаких величина (Слика 3.2.5.-лево) вредност пресека,

[110] У контексту ДА класификационе процедуре, *a priori* вероватноћа сваке групе у популацији, у ознаци π_k, представља вероватноћу да било која на случај изабрана опсервација припадне тој конкретној групи (*Sharma*, 1996, стр. 256; *Hardle & Simar*, 2003, стр. 325), без претходног разматрања вредности те мултиваријационе опсервације. У условима када су анализом обухваћене само две групе, тако да је $n_1 + n_2 = n$, *Hair et al.* (2010, стр. 257) и *Izenman* (2008, стр. 245) указују на могућност одређивања оцењених вредности *prior* вероватноћа група израчунавањем релативног учешћа сваке групе у укупном узорку, односно: $p_1 = n_1 / n$ и $p_2 = n_2 / n$. Дакле, групе једнаких величина одликоваће се једнаким вредностима оцена *prior* вероватноћа, и обратно.

z_{cv}, која обезбеђује њихово оптимално раздвајање, одређује се као једноставан просек њихових центроида, односно, коришћењем израза (3.2.30). Сепарација група остварена на овај начин, сматра се симетричном, будући да је вредност пресека на једнакој удаљености од центроида обе групе. Насупрот одређивању непондерисане вредности пресека, приликом израчунавања вредности пресека за сепарацију група y_2 и y_3, неопходно је применити формулу дату изразом (3.2.29) која уважава и директно инволвира, евидентно различите, величине појединачних група (Слика 3.2.5.-десно). У циљу обезбеђивања оптималне сепарације, тако утврђена вредност пресека пондерисана је у корист мање групе. Другим речима, линија раздвајања алоцирана је ближе центроиду групе која се одликује мањом величином, односно мањом *a priori* вероватноћом у популацији. Класификационо правило засновано на тако одређеном пондерисаном просеку центроида група резултираће, евидентно, мањим бројем погрешних класификација, него у случају коришћења поступка заснованог на претпоставци о једнаким *a priori* вероватноћама група.



Слика 3.2.5. Илустративни приказ поступка одређивања пондерисане и непондерисане вредности пресека дискриминационе *Z* функције

Извор: Ауторов приказ

Класификациона правила, формирана на основу утврђених вредности пресека z_{cv} распореда дискриминационих z_i вредности на нивоу појединачних група опсервација, у контексту хипотетичког примера приказаног на Слици 3.2.4 и Слици 3.2.5, могу се приказати у следећој форми (адаптирано према: *Sharma*, 2006, стр. 257):

$$\begin{aligned} z_i < z_{cv(1)} &\Rightarrow \text{класификуј опсервацију } \mathbf{x}_i \text{ у групу 1, } (z_i \in R_1 \rightarrow \mathbf{x}_i \in y_1) \\ z_{cv(1)} \le z_i < z_{cv(2)} &\Rightarrow \text{класификуј опсервацију } \mathbf{x}_i \text{ у групу 2, } (z_i \in R_2 \rightarrow \mathbf{x}_i \in y_2), \\ z_i \ge z_{cv(2)} &\Rightarrow \text{класификуј опсервацију } \mathbf{x}_i \text{ у групу 3, } (z_i \in R_3 \rightarrow \mathbf{x}_i \in y_3) \end{aligned} \qquad (3.2.31)$$

где z_{cv} означава одговарајућу вредност пресека две конкретне групе, а z_i представља дискриминациони скор *i*-те мултиваријационе опсервације $\mathbf{x}_i$, односно *i*-ту вредност нестандардизоване, статистички и практично значајне, линеарне дискриминационе функције *Z*, изведене претходним спровођењем ДДА.

Поред наведеног, уз истакнуту улогу у реализацији циљева ПДА, елаборирани методолошки оквир одређивања класификационих правила и на њима заснована оцена

класификационе / предиктивне прецизности коришћених дискриминационих *Z* функција, такође представља и финални корак у разматрању генералне корисности издвојених дискриминанти у оквиру ДДА (*Brown & Wicker*, 2000, стр. 225). Изложену констатацију потврђују *Johnson & Wichern* (2007, стр. 593) истичући да статистички сигнификантна сепарација група, иако неопходан услов за приступање поступку извођења класификационих правила, не гарантује нужно и високу ефикасност резултирајуће класификације. Прецизније, што су групе боље и више раздвојене дискриминантом већа је и вероватноћа извођења квалитетних класификационих правила. Користећи приказ на Слици 3.2.5, може се рећи да хипотетичка дискриминациона функција обезбеђује практично значајну сепарацију група, с обзиром на релативно мале површине хистограма фреквенција појединачних група које су „одсечене“ линијом пресека. Поређењем пресека приказаних парова група, такође се може приметити да коришћена дискриминанта боље раздваја другу од треће групе, будући да су „одсечене“ површине њихових хистограма мање, у односу на кореспондентне површине детерминисане линијом пресека између прве и друге групе. Површине издвојене линијом пресека од остатка хистограма фреквенција репрезентују опсег вредности *Z* функције на нивоу појединачних опсервација које ће, формираним класификаторима, највероватније бити погрешно класификоване у састав неадекватне групе.

Класификациона процедура заснована на коришћењу дискриминационих *Z* функција посебно је погодна за реализацију ПДА циљева у случају примене ДА засноване на двема групама као и вишеструке ДА, али када је оценом статистичке и практичне значајности утврђено да је, од *r* изведених дискриминанти, само једна дискриминациона функција потребна за адекватно репрезентовање, односно описивање разлика између посматраних *g* група (*Sharma*, 2006, стр. 293). Извођење класификатора и реализација представљене класификационе процедуре у оквирима вишеструке ДА значајно се компликује када постоји више дискриминационих функција, у ознаци *s* (при чему важи релација: $1 < s \leq r$), чија је статистичка и практична значајност, у обезбеђивању дискриминације између група, потврђена. У таквим условима, за потребе извођења класификационих правила и предвиђање групне припадности „нових“ опсервација, углавном се практикује имплементација класификационе процедуре засноване на линеарним класификационим функцијама, чије елаборирање следи у наставку.

Класификациона процедура заснована на линеарним класификационим функцијама

Алтернативни приступ класификацији опсервација у дискриминационој анализи заснива се на израчунавању и примени одговарајућих класификационих функција, такође познатих под називом *Fisher*-ове линеарне класификационе функције. За разлику од линеарне дискриминационе функције *Z*, класификационе функције се утврђују за сваку од *g* група, односно категорија зависне променљиве Y_k (k = 1,2,…, g), користе се искључиво за потребе класификације / предикције, и одликују се следећим обликом:

$$\left.\begin{array}{l} \text{за групу 1 } (y_1) \rightarrow C_1 = b_{01} + b_{11}X_1 + b_{21}X_2 + \ldots + b_{p1}X_p \\ \text{за групу 2 } (y_2) \rightarrow C_2 = b_{02} + b_{12}X_1 + b_{22}X_2 + \ldots + b_{p2}X_p \\ \vdots \qquad\qquad \vdots \\ \text{за групу } g \ (y_g) \rightarrow C_g = b_{0g} + b_{1g}X_1 + b_{2g}X_2 + \ldots + b_{pg}X_p \end{array}\right\} C_k = b_{0k} + b_{1k}X_1 + b_{2k}X_2 + \ldots + b_{pk}X_p \quad (3.2.32)$$

У датом изразу, симболом C_k означена је вредност класификационе функције на нивоу k-те групе, или класификациони скор, док b_{jk} представља оцењену вредност класификационог коефицијента j-те независне променљиве у k-тој класификационој функцији, утврђене путем израза (*IBM*, 2011, стр. 286):

$$\underset{(p\times 1)}{\mathbf{b}_k} = \underset{(p\times p)}{\overline{\mathbf{S}}^{-1}}\,\underset{(p\times 1)}{\overline{\mathbf{X}}_k} \;\Rightarrow\; \begin{bmatrix} b_{1k} \\ b_{2k} \\ \vdots \\ b_{pk} \end{bmatrix} = \begin{bmatrix} \overline{s}_{11} & \overline{s}_{12} & \dots & \overline{s}_{1p} \\ \overline{s}_{21} & \overline{s}_{22} & \dots & \overline{s}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{s}_{p1} & \overline{s}_{p2} & \dots & \overline{s}_{pp} \end{bmatrix}^{-1} \begin{bmatrix} \overline{x}_{1k} \\ \overline{x}_{2k} \\ \vdots \\ \overline{x}_{pk} \end{bmatrix}. \tag{3.2.33}$$

Оцењена вредност одсечка k-те класификационе функције, представљена ознаком b_{0k} унутар израза (3.2.32), утврђује се коришћењем следеће формуле (*IBM*, 2011, стр. 287):

$$b_{0k} = \log p_k - \frac{1}{2}\left(\underset{(1\times p)}{\overline{\mathbf{X}}_k'}\,\underset{(p\times 1)}{\mathbf{b}_k}\right) = \log\frac{n_k}{\sum_{k=1}^{g} n_k} - \frac{1}{2}\begin{bmatrix}\overline{x}_{1k} & \overline{x}_{2k} & \dots & \overline{x}_{pk}\end{bmatrix}\begin{bmatrix} b_{1k} \\ b_{2k} \\ \vdots \\ b_{pk} \end{bmatrix} \tag{3.2.34}$$

У изразима (3.2.33) и (3.2.34), као што је већ речено, елементи (p×1) вектора $\mathbf{b}_k$, представљају оцењене вредности коефицијената b_j придружених свакој од p независних променљивих у оквиру k-те класификационе функције, $\overline{\mathbf{S}}^{-1}$ је инверзна матрица (p×p) заједничке узорачке коваријационе матрице дефинисане изразом (3.2.12), $\overline{\mathbf{X}}_k$ означава (p×1) вектор просечних вредности p независних променљивих у k-тој групи, чији се елементи одређују коришћењем израза (3.2.7), док симбол p_k показује релативно учешће величине k-те групе у укупном узорку за анализу, величине $n = \Sigma n_k$ (за k=1, 2,…, g).

Класификациона процедура, заснована на примени овог метода, подразумева израчунавање вредности g класификационих функција C_k за сваку појединачну јединицу посматрања у оквиру узорка n, заменом њима припадајућих вредности независних променљивих у линеарним функцијама дефинисаним изразом (3.2.32). Наиме, свака опсервација се одликује посебним класификационим скором израчунатим за сваку од g група, у ознаци: C_1, C_2,..., C_g. Према класификационом правилу, карактеристичним за овај метод, свака појединачна јединица посматрања се распоређује, односно класификује, у састав k-те групе за коју је постигнута вредност припадајуће класификационе функције највећа (*Hair et al.*, 2010, стр. 257). Класификационо правило за опсервацију $\mathbf{x}_i' = [x_{i1}, x_{i2},\dots, x_{ip}]$ у векторском запису, гласи:

$$\text{ако је } \left[C_k(\mathbf{x}) = c_{0k} + \mathbf{b}_k'\mathbf{x}\right] \max \;\Rightarrow\; \text{класификуј опсервацију } \mathbf{x}_i \text{ у } k\text{-ту групу}. \tag{3.2.35}$$

Иако овај метод класификације у контексту ДА подразумева коришћење свих g линеарних класификационих функција, једноставност његове примене, као што је претходно истакнуто, посебно долази до изражаја у случају реализације вишегрупне ДА, у поређењу са, у таквим околностима знатно компликованијом, процедуром заснованом на израчунавању одговарајућих (g–1) вредности пресека z_{cv}, на нивоу сваке од s изведених, али статистички и практично значајних, дискриминационих функција Z. Такође, примена ове класификационе процедуре у потпуности је независна од спровођења ДДА, будући да није условљена претходним извођењем дискриминационих Z функција. Стога, може се

констатовати да је статистичка сврха извођења линеарних класификационих правила искључиво и директно повезана са предиктивним аспектом ДА.

3.2.6. Статистичко закључивање о прецизности класификационих правила

Пре покушаја класификације нових јединица посматрања, односно предвиђања групне припадности некатегорисаних опсервација, неопходно је извршити евалуацију прецизности претходно изведених класификационих правила (изрази (3.2.31) и / или (3.35)), сагледавањем, тестирањем и интерпретирањем резултата њихове примене на јединицама посматрања, чија је припадност конкретној групи позната. Наиме, врши се појединачна (ре)класификација већ категорисаних опсервација коришћењем изабране класификационе методе [111] . Сумарни приказ резултата поређења стварне групне припадности коришћених јединица посматрања и евидентираних исхода њихове (ре)класификације презентирају се у форми класификационе матрице (енгл. *classification* / *confusion* / *assignment* / *prediction matrix*), као у Табели 3.2.3.

Табела 3.2.3. Класификациона матрица

Стварна припадност конкретној групи	Предвиђена припадност конкретној групи				Стварна величина група (n_k)	Пропорција исправних класификација
	група y_1	група y_2	…	група y_g		
група y_1	$f_{1\to1}$	$f_{1\to2}$	…	$f_{1\to g}$	$\sum_{k=1}^{g} f_{1\to k} = n_1$	$p_1^{HR} = \frac{f_{1\to1}}{n_1}$
група y_2	$f_{2\to1}$	$f_{2\to2}$	…	$f_{2\to g}$	$\sum_{k=1}^{g} f_{2\to k} = n_2$	$p_2^{HR} = \frac{f_{2\to2}}{n_2}$
⋮	⋮	⋮	⋱	⋮	⋮	⋮
група y_g	$f_{g\to1}$	$f_{g\to2}$	…	$f_{g\to g}$	$\sum_{k=1}^{g} f_{g\to k} = n_g$	$p_g^{HR} = \frac{f_{g\to g}}{n_g}$
Предвиђена величина група	$\sum_{k=1}^{g} f_{k\to1}$	$\sum_{k=1}^{g} f_{k\to2}$	…	$\sum_{k=1}^{g} f_{k\to g}$	$\sum_{k=1}^{g} n_k = n$	$p^{HR} = \frac{\sum_{k=1}^{g} f_{k\to k}}{n}$

Извор: Ауторов приказ

Елементи на главној дијагонали класификационе матрице, $f_{k\to k}$, означавају фреквенцију јединица посматрања у саставу k-те групе које су распоређене на основу примењених класификатора у исту, k-ту групу. Заправо, дијагонални елементи матрице показују број исправно класификованих опсервација на нивоу појединачних група, а њихов збир укупан број исправних класификација у узорку величине n. Логично, елементи ван главне дијагонале ($f_{k\to \neq k}$) садрже фреквенције јединица посматрања, по свакој од g група, које су применом класификационих правила распоређене у неку од група којој оне не припадају. Прецизније, од укупно n_1 јединица посматрања које заиста

[111] У оквиру ДА засноване на двема групама, оба класификациона метода резултираће, генерално, истим исходима (ре)класификације уколико су претходно задовољене претпоставке о мултиваријационој нормаланости распореда популације из које потиче узорак коришћених опсервација и хомогености коваријационих матрица појединачних група у популацији, као и претпоставка о једнакости трошкова погрешне класификације и *a priori* вероватноћа појединачних група (*Sharma*, 1996, стр. 261). Уколико нека(е) од наведених претпоставки није(су) задовољена(е), добијени резултати могу се, у већој или мањој мери, међусобно разликовати, као и у случају примене вишеструке ДА. У том контексту, избор конкретне класификационе процедуре условљен је субјективним ставом истраживача, специфичностима конкретног истраживања и, коначно, бољим резултатима класификације.

припадају групи y_1, примењени класификатор погрешно распоређује $f_{1\to2}$ опсервација у групу y_2, односно $f_{1\to g}$ опсервација у групу y_g. Вредности у колони – стварна величина група, показују број јединица посматрања које стварно припадају свакој од посматраних g група, док вредности у последњем реду матрице показују укупан број јединица посматрања распоређених путем класификационог правила у сваку од g група. Уколико не би било грешака у алокацији опсервација тада би класификациона матрица била дијагонална матрица. Представљени елементи чине основу за ефикасну оцену перформанси и квалитета дикриминационог модела, формираног на бази линеарних дискриминационих *Z* функција и / или *Fisher*-ових линеарних класификационих функција, у контексту класификације / предикције групне припадности „некатегорисаних", (нових) опсервација.

Формирана класификациона матрица представља адекватно средство за израчунавање одговарајуће, директне мере ефективности, односно предиктивне прецизности, изведених класификационих правила и имплементиране класификационе процедуре[112]. При давању одговара на питање: „Колика се прецизност класификационог правила, заснованог на узорку који је предмет анализе, може очекивати у поступку класификације јединица посматрања будућих узорака?", посматрано у контексту оцењивања предиктивне прецизности у ДА, у литератури су присутне „значајне" разлике у погледу коришћених мера, њихових дефиниција, симбола и напомена. Најчешће коришћена(е) мера(е) предиктивних перформанси класификатора је(су):

- ✓ стопа грешака у класификацији, односно, стопа погрешних класификација (енгл. *classification error rate, misclassification rate*) и / или
- ✓ стопа погодака у класификацији, односно, стопа исправних класификација (енгл. *classification hit rate, correct classification rate*).

Полазећи од чињенице да су ове две мере комплементарне, односно да је њихов збир једнак јединици (*Timm*, 2009, стр. 427), за потребе излагања у наставку текста, усвојен је приступ заснован на сагледавању пропорције исправних класификација, као мере предиктивне прецизности класификатора. Израчунавање вредности овог показатеља за сваку појединачну групу врши се коришћењем формула представљеним у последњој колони класификационе матрице (Табела 3.2.3). На нивоу узорка, укупна стопа исправних класификација дефинише се као количник збира фреквенција исправно класификованих опсервација по појединачним групама и величине укупног узорка, *n*. Логично, већа вредност укупне стопе погодака показује боље предиктивне способности изведених линеарних класификатора и обратно.

Израчуната стопа исправних класификација на нивоу коришћеног узорка представља оцењену вредност стварне стопе погодака (енгл. *Actual Hit Rate, AHR*[113]). Особине

---

[112] Значај и важност информација које мера предиктивних перформанси изведених класификатора садржи нису ограничене само на оквире ПДА. Индиректно, поменута мера пружа додатан увид у квалитет дискриминације, односно сепарације група која је остварена претходно изведеним дискриминационим Z функцијама, у оквиру имплементиране ДДА (*Klecka*, 1980, стр. 50). Другим речима, директна мера предиктивне прецизности класификатора представља уједно и додатак дискусији о статистичкој и практичној значајности дискриминационих функција, изложеној у оквиру Одељка 3.2.4.

[113] *AHR* означава стопу погодака остварену применом класификационог правила, изведеног коришћењем конкретног узорка за анализу, на будућим узорцима некатегорисаних опсервација, прибављеним из исте популације (*Huberty & Olejnik*, 2006, стр. 296). Директно израчунавање њене вредности није могуће спровести, будући да је условљено

утврђене оцене, а тиме и поузданост изведених закључака у погледу предиктивних перформанси дискриминационог модела, у великој мери зависе од узорка категорисаних опсервација који је употребљен за спровођење поступка евалуације класификационе прецизности изведених правила. Наиме, уколико је класификациона матрица формирана (ре)класификацијом истих јединица посматрања које су директно коришћене за извођење класификационих правила, тада вредност пропорције исправних класификација, утврђена на бази ње, представља пристрасну оцену од стварне пропорције погодака, *AHR*. Такав поступак евалуације предиктивних перформанси изведених правила назива се интерна класификациона анализа (енгл. *internal classification analysis*) (*Huberty & Olejnik*, 2006, стр. 300), а резултирајућа оцена *AHR*-а, заснована на опсервацијама узорка за анализу, очигледна стопа погодака (енгл. *Apparent Hit Rate*), у ознаци *ApHR* (*Rencher*, 2002, стр. 307). Будући да су класификациона правила оптимизирана за конкретни узорак који је коришћен за њихово извођење, резултати интерне класификационе анализе биће повољнији, у контексту евалуиране прецизности модела, у односу на резултате класификације засноване на случајности, нарочито у условима када су групе у саставу узорка за анализу веома мале. Прецизније, за утврђену оцену *ApHR*, каже се да је пристрасна „навише“, односно оптимистичка, у смислу да прецењује вредност стварне стопе погодака *AHR*, која се реално може очекивати при примени узорачких класификационих правила на будућим узорцима. Генерално, и поред става *Rencher*–а (2002, стр. 309) да питање позитивне пристрасности *ApHR*-а није толико изражено у случају коришћења великих узорака, третирање *ApHR* оцене као опште оцене стварне пропорције погодака *AHR*, може резултирати извођењем неодрживих и погрешних закључака о предиктивној прецизности дискриминационог модела (*Huberty & Olejnik*, 2006, стр. 300).

Приступ којим се обезбеђује кориговање и / или ублажавање присутне пристрасности оцене *ApHR*, подразумева спровођење, такозване, екстерне класификационе анализе (енгл. *external classification analysis*). Наведени приступ, за разлику од претходно описане интерне класификације, подразумева евалуацију предиктивних перформанси дискриминационог модела, изведеног на бази једног узорка, применом резултирајућих класификационих правила на другом, посебном, узорку сачињеном од, такође, категорисаних јединица посматрања, али који није коришћен за извођење класификатора који се примењују. Резултирајућа класификациона матрица представља реалнију основу за израчунавање оцењене вредности стварне стопе исправних класификација. Другим речима, тако утврђена оцена од *AHR*, сматра се знатно мање пристрасном, односно приближно непристрасном оценом (*Timm*, 2009, стр. 428), у ознаци *ApHR*', у поређењу са оценом *ApHR*, добијеном као резултат интерне класификације, представљајући тиме и погоднију основу за формулисање статистички валидних закључака у контексту предиктивних способности оцењеног дискриминационог модела. Важност и значај поступка екстерне валидације потврђен је и његовом позицијом у дијаграму тока поступка спровођења ДА, представљеном на Слици 3.2.1. Најчешће коришћене методе за валидацију, то јест, екстерно вредновање предиктивних перформанси дискриминационог модела и правила за класификацију, су:

непознатим вредностима одговарајућих параметара популације, због чега се приступа одређивању њене оцене на нивоу расположивог узорка категорисаних јединица посматрања.

• *Метод задржавања* (енгл. *holdout method*)

Примена методе задржавања подразумева поделу иницијалног (укупног) узорка, случајним путем, на два дела, и то: ▸ (под)узорак за анализу (енгл. *analysis* / *training* / *estimation sample*) који се користи за оцењивање дискриминационих *Z* функција и / или *Fisher*-ових линеарних класификационих функција и креирање одговарајућих класификационих правила на бази истих; и ▸ (под)узорак за валидацију (енгл. *holdout* / *validation sample*) који је намењен превасходно за евалуацију предиктивне прецизности претходно изведених класификационих правила. Класификацијом опсервација садржаних унутар „задржаног" узорка, који није коришћен за извођење примењених класификационих правила, елиминише се пристрасност при израчунавању оцењене вредности мере класификационе прецизности, односно *ApHR*'.

Прецизне инструкције у погледу релативног односа величина наведена два (под)узорка нису дефинисане у литератури и предмет су неусаглашености статистичара. Сходно наведеном, иако истичу могућност поделе иницијалног узорка у односу 60 : 40 %, или пак, 75 : 25 %, у корист (под)узорка за анализу, *Hair, et al.* (2010, стр. 250), као најпопуларнији приступ, наводе поделу узорка на два једнака дела. С друге стране, *Brown & Wicker* (2000, стр. 227) и *Huberty* (1975, стр. 558), као подразумевану, наводе поделу укупног узорка према којој ће 2/3 опсервација бити распоређене у састав (под)узорка за анализу, а преостала, 1/3 јединица посматрања у оквиру (под)узорка за валидацију. Генерално, размера поделе узорка на његове партиције примарно зависи од величине расположивог иницијалног узорка, што такође подразумева и узимање у обзир препорука датих у погледу величине узорка који ће бити коришћен у анализи као и величине појединачних група опсервација у његовом саставу, а које су изложене у оквиру Одељка 3.2.2. У том контексту, *Klecka* (1980, стр. 52) напомиње да (под)узорком за анализу мора да буде обухваћен довољно велики број опсервација како би се обезбедила статистичка стабилност дискриминационих коефицијената. Такође, важно је истаћи да се формирање (под)узорака врши се применом процедуре пропорционално стратификованог случајног узорковања из иницијалног (укупног) узорка.

• *„Изостави-по-једну-опсервацију" метод* (енгл. *„ Leave-One-Out"* (*L-O-O*) *method*)

Овај метод унакрсне валидације, нарочито погодан када је величина укупног узорка недовољно велика за примену метода задржавања, спроводи се кроз следеће кораке (*Timm*, 2009, стр. 428; *Brown & Wicker*, 2000, стр. 227):

1. Из узорка величине *n*, извученог из популације *π*, изоставља се једна случајно одабрана опсервација и креира класификационо правило на основу осталих *n*–1 опсервација;

2. Коришћењем изведеног правила спроводи се класификација задржане опсервације;

3. Врши се *n*–1 понављања претходна два описана корака све док и последња опсервација не буде издвојена и класификована у једну од група зависне променљиве;

4. Формира се класификациона матрица на основу прибављених резултата распоређивања и израчунава оцењена вредност стварне пропорције погодака, *ApHR*'.

Поступак примене *L-O-O* методе еквивалентан је спровођењу *n* дискриминационих анализа и класификацији *n* задржаних опсервација и, према мишљењу *Sharma*-е (1996, стр.

274), резултира готово у потпуности непристрасном оценом прецизности класификационих правила.[114]

Спровођење екстерне евалуације предиктивних перформанси узорачких класификационих правила, према речима *Johnson*-а *& Wichern*-а (2007, стр. 595), представља морални чин, неопходан у циљу обезбеђивања генерализације добијених резултата, будући да елиминише могућности доношења закључака о репрезентативности дискриминационог модела који су применљиви само на коришћеном узорку за анализу. У контексту изложених настојања, важну улогу има и испитивање статистичке и практичне значајности примарно, описаним поступком утврђене, (непристрасне) оцене стварне пропорције исправних класификација.

Статистичка и практична значајност оцене показатеља прецизности класификације

Додатна статистичка верификација предиктивне прецизности изведених класификационих правила и репрезентативности дискриминационог модела заснива се на поређењу емпиријске вредности пропорције погодака и очекиване стопе исправних класификација која би на датом (истом) узорку била остварена али у случају спровођења класификације на бази случајности, без употребе, односно, „без помоћи и подршке“ формираних класификационих правила. Евалуација статистичке значајности оцењене вредности стварне пропорције исправних класификација, утврђене на нивоу узорка за валидацију, подразумева спровођење поступка тестирања статистичких хипотеза о пропорцији популације, применом одговарајуће статистике *Z* теста, која, заједно са кореспондентном нултом и алтернативном хипотезом, има следећи облик (*Brown & Wicker*, 2000, стр. 228; *Lovrić*, 2009, стр. 230):

$$\left.\begin{array}{l} H_0 : \pi^{HR} \le \pi_0 \\ H_1 : \pi^{HR} > \pi_0 \end{array}\right\} \rightarrow Z = \frac{p^{HR} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \qquad (3.2.36)$$

У датом изразу, коришћени симболи означавају[115]: π^{HR} – стварна пропорција исправних класификација која се очекује од примене узорачких класификационих правила на некатегорисаним опсервацијама будућих узорака (ранија ознака - *AHR*); π_0 – хипотетичка вредност пропорције погодака, односно пропорција исправно класификованих јединица за коју се очекује да ће бити реализована у случају класификације засноване на случајности; p^{HR} – оцењена вредност (укупне) стварне стопе погодака у узорку за валидацију (претходна ознака – *ApHR*'); n – величина узорка за валидацију; и Z – статистика коришћеног теста која следи стандардизован нормалан распоред, симболички Z ~ N(0, 1), уколико су испуњени следећи услови за примену (*Lovrić*, 2009, стр. 230): ▸ $n \ge 30$; ▸ $n \cdot \pi_0 \ge 5$; и ▸ $n \cdot (1 - \pi_0) \ge 5$.

Кључни корак у примени датог теста односи се на одређивање хипотетичке вредности π_0, која представља одговарајући стандард или критеријум наспрам којег се врши поређење и евалуација емпиријски утврђене прецизности класификације

[114] Детаљна разматрања метода унакрсне валидације (енгл. *cross-validation methods*) представљена су од стране *Han et al.* (2012, стр. 370) и *Sharma* (1996, стр. 273–274).

[115] У представљеном изразу, ради боље разумљивости и усклађивања са логиком поступка тестирања хипотеза о пропорцији, извршено је извесно прилагођавање симбола појединих елемената израза, у односу на њихово означавање у оквиру претходног Одељка.

опсервација у саставу узорка за валидацију. Резултати класификације добијени у оквиру претходно спроведених, упоредивих истраживања сличног карактера, могу се користити као хипотетичка вредност π_0, у поступку испитивања сигнификантности резултата актуелног истраживања (*Brown & Wicker*, 2000, стр. 228). Међутим, у одсуству таквих могућности, тестирање претпоставки о статистичкој значајности добијених резултата предиктивне прецизности класификатора захтева одређивање претпостављене пропорције исправних класификација за коју се очекује да може бити остварена у случају када би се класификација опсервација у узорку за валидацију извршила на бази случајности, односно независно од изведених класификатора. Генерално, начин одређивања стопе погодака на бази случајности детерминисан је односом величина појединачних група у саставу узорка, односно оцењеним вредностима *a priori* вероватноћа појединачних група у узорку, у ознаци p_k (k=1, 2,…, g). Наиме, када су величине коришћених група једнаке, очекивана укупна пропорција исправно класификованих опсервација на бази случајности израчунава се као реципрочна вредност укупног броја група, односно 1 / g. Овако одређена хипотетичка стопа погодака базира се на претпоставци да се, у условима спровођења класификације на бази случајности, свака јединица посматрања у узорку одликује једнаком вероватноћом распоређивања унутар било које од посматраних група. *Huberty & Olejnik* (2006, стр. 316) дају детаљно објашњење логике одређивања елабориараног стандарда за поређење, које се може приказати на следећи начин:

- нека је симболима n_k и p_k означена величина и оцењена *prior* вероватноћа k-те групе (за k=1,2,…, g), тада се очекивани број исправних класификација на бази случајности за ту групу одређује као $f_{k\to k} \approx p_k\, n_k$;
- уколико је $n_1 = n_2 = ... = n_g = n^*$, тада се величина укупног узорка (n) може одредити као $n = g \cdot n^*$, а оцењена *prior* вероватноћа k-те групе ($p_k = n_k / n$) изразом $p_k = n^* / n$;
- консеквентно, укупна фреквенција погодака на бази случајности у узорку (n), биће:

$$\sum_{k=1}^{g} f_{k\to k} = \sum_{k=1}^{g} p_k n_k = \sum_{k=1}^{g} \frac{n^*}{n} n^* = g \frac{n^*}{n} n^* = g \frac{n^*}{gn^*} n^* = n^*; \qquad (3.2.37)$$

- коначно, укупна стопа погодака на бази случајности, биће: $\pi_0 = \frac{\sum_{k=1}^{g} f_{k\to k}}{n} = \frac{n^*}{n} = \frac{n^*}{gn^*} = \frac{1}{g}$.

У случају када су групе у саставу узорка различитих величина, односно када се не може усвојити претпоставка о једнакости њихових *a priori* вероватноћа, израчунавање хипотетичке стопе погодака π_0, подразумева примену једног од следећа два приступа (*Huberty & Olejnik*, 2006, стр. 316–319; *Hair et al.*, 2010, стр. 261–262): ▸ приступ заснован на критеријуму максималне случајности (енгл. *maximum chance criterion*) и ▸ приступ заснован на критеријуму пропорционалне случајности (енгл. *proportional chance criterion*).

Критеријум максималне случајности може се применити у ситуацијама када је основни и једини циљ дискриминационе анализе максимизирање укупне пропорције исправних класификација на нивоу узорка (p^{HR}), при чему не постоји заинтересованост за оствареним резултатима на нивоу појединачних група, у ознаци p_k^{HR}, и насталим типовима грешака при класификацији (*Brown & Wicker*, 2000, стр. 228). Максимизирање укупне стопе погодака могуће је остварити уколико би све јединице посматрања случајним путем

биле алоциране у највећу групу. Према овом критеријуму, хипотетичка пропорција исправних класификација једнака је оцењеној вредности *a priori* вероватноће највеће групе, $\pi_0 = max(p_1, p_2,\ldots, p_g)$, (*Huberty & Olejnik*, 2006, стр. 319) односно, представља количник величина највеће групе и укупног узорка (*Brown & Wicker*, 2000, стр. 228).

Међутим, у ситуацијама када је циљ истраживања постизање максималне стопе погодака на нивоу сваке појединачне групе, а не само највеће као у претходном случају, тада се *приступ заснован на критеријуму пропорционалне случајности* сматра најадекватнијим за одређивање хипотетичке стопе погодака. Кореспондентна формула за утврђивање π_0, која узима у обзир *a priori* вероватноће свих група а не само највеће, гласи (*Huberty & Olejnik*, 2006, стр. 317):

$$\pi_0 = \frac{\sum_{k=1}^{g} f_{k\to k}}{n} = \frac{\sum_{k=1}^{g} p_k n_k}{n} \quad . \tag{3.2.38}$$

У представљеном изразу, n је величина посматраног узорка, n_k величина појединачних група, а симбол $f_{k\to k}$ означава очекивани број исправних класификација на бази случајности за k-ту групу (за k=1,2,…, g), који се може исказати, коришћењем оцењених вредности *a priori* вероватноћа (p_k) релацијом $f_{k\to k} \approx p_k\, n_k$ (видети објашњење израза (3.2.37)).

Независно од поступка тестирања статистичких хипотеза (израз (3.2.36)), могуће је, на бази изабраног приступа и израчунате хипотетичке стопе погодака која се очекује да ће бити реализована на бази случајности, π_0, извршити „грубу“ процену прихватљивости и репрезентативности пропорције погодака у узорку за валидацију. У том контексту, *Hair, et al.* (2010, стр. 262) наводи да се остварена класификациона прецизност може сматрати „условно“ прихватљивом уколико је стопа погодака забележена у узорку за валидацију (p^{HR}) најмање за ¼ већа од стопе исправних класификација која се очекује на бази случајности (π_0), као одговарајућег стандарда за поређење. Посматрано из супротног угла, *Huberty & Olejnik* (2006, стр. 317) истичу да нема потребе спроводити тестирање хипотеза о статистичкој значајности уколико је укупна пропорција погодака p^{HR} у узорку за валидацију, мања од или једнака очекиваној пропорцији исправних класификација на бази случајности π_0.

Коначан и много поузданији одговор у погледу репрезентативности и предиктивне прецизности изведених класификационих правила обезбеђује се резултатима тестирања статистичке сигнификантности. Сходно наведеном, уколико је реализовани ниво значајности (p-вредност) мањи од постављеног ризика грешке I врсте α, H_0 се одбацује и, консеквентно, уз поменути ризик прихвата се алтернативна хипотеза која гласи: „стварна стопа погодака која се очекује од примене узорачких класификационих правила на некатегорисаним опсервацијама будућих узорака (π^{HR}) је статистички сигнификантно већа од хипотетичке вредности пропорције исправно класификованих јединица (π_0) за коју се очекује да ће бити реализована у случају класификације засноване на случајности“. Алтернативно, неуспех у одбацивању H_0 сугерише, уз ризик грешке α, да не постоји довољно емпиријских доказа којима би се потврдила статистичка значајност оцене стварне укупне пропорције погодака утврђене применом изведених класификационих правила, односно да нема довољно доказа за прихватање тврдње према којој прецизност

класификације засноване на примени класификационих правила надмашује очекивану стопу погодака на бази случајности.

Међутим, чак и уколико је прихваћена H_1 у представљеном поступку тестирања статистичке значајности укупне пропорције погодака (p^{HR}) на нивоу узорка за валидацију, неопходно је добијеним резултатима приступити критички, будући да је могуће да се једна или више мањих група одликују неприхватљиво ниским пропорцијама исправних класификација, иако је укупна пропорција статистички прихватљива и значајна *Hair et al.* (2010, стр. 263). У циљу отклањања наведене сумње неопходно је извршити додатну проверу статистичке значајности остварене пропорције исправно класификованих опсервација за сваку групу појединачно, у ознаци $p_{(k)}^{HR}$. Статистика теста за оцену статистичке значајности стопе погодака за било коју групу *k*, представљена је формулом (*Sharma*, 1996, стр. 258):

$$\left.\begin{array}{l} H_0 : \pi_{(k)}^{HR} \le \pi_{0(k)} \\ H_1 : \pi_{(k)}^{HR} > \pi_{0(k)} \end{array}\right\} \rightarrow Z_{(k)} = \frac{p_{(k)}^{HR} - \pi_{0\,(k)}}{\sqrt{\frac{\pi_{0\,(k)}(1-\pi_{0\,(k)})}{n_k}}}, \tag{3.2.39}$$

где симбол n_k означава број јединица посматрања у *k*-тој групи у узорку за валидацију, $\pi_{(k)}^{HR}$ је стварна пропорција исправних класификација за *k*-ту групу која се очекује од примене узорачких класификационих правила на будућим, некатегорисаним опсервацијама, $p_{(k)}^{HR}$ представља оцењену вредност стопе погодака за конкретну, *k*-ту групу (елементи последње колона у Табели 3.2.3), а $\pi_{0(k)}$ означава хипотетичку стопу погодака која се очекује да ће бити остварена при класификацији на бази случајности. Одређивање вредности $\pi_{0(k)}$ доста је једноставније него у случају тестирања статистичке значајности укупне стопе погодака, будући да је увек једнака оцењеној вредности *a priori* вероватноће посматране *k*-те групе, односно: $\pi_{0(k)} = p_k = n_k / n$, за $k = 1,2,\ldots, g$. Спровођење поступка тестирања статистичких хипотеза датих изразом (3.2.39) у потпуности је идентичан претходно елаборираном за укупну стопу исправних класификација π^{HR} (израз (3.2.36)).

Поред наведеног, важно је направити и разлику између практичне и статистичке сигнификантности предиктивне прецизности коришћених класификатора. Практична значајност класификационих резултата омогућава сагледавање степена побољшања прецизности у предвиђању групне припадности нових опсервација који се постиже применом изведених, на бази дискриминационе анализе, класификатора у односу на резултате који би били остварени класификацијом истих опсервација али на бази случајности. Предложен од стране *Huberty* (1994, стр. 107, цитирано у *Huberty & Lowman*, 2000, стр. 547), показатељ који се користи као мера поменутог побољшања, израчунава се путем формуле:

$$I = \frac{p^{HR} - \pi_0}{1 - \pi_0}. \tag{3.2.40}$$

У представљеном изразу, нумеричка вредност *I* показује проценат побољшања пропорције исправних класификација на бази случајности, који се постиже употребом изведених класификационих правила у оквиру дискриминационе анализе. Примера ради,

за I = 0,8512, добијена вредност показује да је спровођењем класификације коришћењем линеарног класификационог правила, приближно 85% више опсервација исправно класификовано (односно, приближно 85% мање грешака у класификацији је начињено), него што би то био случај да је њихова класификација спроведена искључиво на бази случајности. Логично, множењем вредности I са очекиваним бројем исправно класификованих опсервација на бази случајности, могуће је одредити фреквенцију јединица посматрања која је одговорна за идентификовану величину, односно степен побољшања прецизности наспрам случајности. За потребе прецизнијег интерпретирања вредности овог показатеља, *Huberty & Lowman* (2000, стр. 558) предлажу следећу скалу степена практичне значајности у зависности од вредности I:

У случају ДА засноване на двема групама (g=2):

висок ниво → 0,35 < I
средњи ниво → 0,20 < I < 0,30
низак ниво → I < 0,15

У случају вишегрупне ДА ($g \geq 3$):

висок ниво → 0,30 < I
средњи ниво → 0,15 < I < 0,25
низак ниво → I < 0,10.

Дакле, за разлику од поступка провере статистичке значајности који пружа одговор на питање: Да ли су, уз одговарајући ризик грешке α, резултати предвиђања на бази класификационих правила бољи од резултата који би били остварени класификацијом заснованој на случајности?, испитивање практичне значајности омогућава давање одговора на питање: Колико су заправо резултати класификације засноване на дискриминационим правилима, чија је статистичка сигнификантност потврђена, бољи у односу на очекиване резултате класификације засноване искључиво на случајности? Генерално, сви претходно разматрани аспекти евалуације прецизности изведених класификационих правила у контексту класификације / предикције групне припадности опсервација, јасно указују на комплексност и значај питања оцене квалитета диксриминационог модела, формираног било на бази линеарних дискриминационих Z функција и / или *Fisher*-ових класификационих функција.

РЕГИОНАЛНЕ НЕРАВНОМЕРНОСТИ КАО ИЗАЗОВ ЗА ПРИМЕНУ МУЛТИВАРИЈАЦИОНЕ АНАЛИЗЕ ПОДАТАКА

4.1. Значај концепта регионализације и регионалног развоја у контексту развоја националне економије

Неравномерности у развоју присутне између дефинисаних административно-територијалних јединица у саставу државе[116] представљају једно од најважнијих, али и најкомплекснијих друштвено-економских проблема са којим се креатори развојних политика и представници државе данас, генерално, суочавају (*Rovan & Sambt*, 2003, стр. 265; *Maletić & Bucalo-Jelić*, 2016, стр. 13). Оправдање за изнету констатацију садржано је у чињеници да изражене разлике у погледу степена развијености региона[117] могу имати озбиљан и значајан (негативан) утицај на друштвено-политичку стабилност једне државе (*Goletsis & Chletsos*, 2011, стр. 174) као и перформансе, односно резултате националне економије у целини (*Obradović et al.*, 2016, стр. 162). Апострофирајући „зависност" државе и ефикасност целокупне привреде од економске структуре и стабилности њених региона, *Jakopin* (2015, стр. 109) економски развој региона сматра основом за реализацију националних економских циљева. Сличног мишљења је и Вуковић (2009, стр. 190), истичући да је конкурентност једне привреде генерисана и условљена капацитетима њених региона да осигурају и подрже ниво привредне активности неопходан за остваривање динамичног привредног раста. У том смислу, остваривање интензивног раста и одрживог економског развоја земље и њене привреде нужно подразумева уважавање концепта регионалне једнакости (у што је могуће већој мери), односно предузимање активности усмерених на уравнотежење и / или повећање степена развијености свих њених региона, а тиме и благостања свих њених становника (*Rovan & Sambt*, 2003, стр. 265; *Istrate & Horea-Serban*, 2016, стр. 209). Сходно наведеном, стварање привредно стабилних и конкурентних региона, како у националном оквиру тако и на међународном тржишту, кроз решавање питања друштвено-економског „баласта" недовољно развијених подручја, као „уског грла" укупног развоја и конкурентности целе привреде (Вуковић, 2009, стр. 190) и стварање услова за успостављање равномерног регионалног развоја, представља, као што је претходно наведено, приоретни задатак сваке државе и кључни корак у настојањима да се обезбеди успешна интеграција националне економије у глобалне економске токове (РЗР, 2009, стр. 88; *Despotović & Cvetanović*, 2017, стр. 112; *Krstić & Vukadinović*, 2011, стр. 554). Додатна потврда значаја и важности концепта равномерног регионалног развоја у контексту подстицања и повећања степена развијености националне економије и њене

[116] Неравномеран развој, како наводе *Mohiuddin & Hashia* (2012, стр. 86) и Милетић и други (2009, стр. 153), представља општу законитост свеукупног развоја која у појединим етапама посебно долази до изражаја, а манифестује се у централизацији и / или поларизацији привредних активности и демографских карактеристика, у смислу да поједине територијалне целине у саставу државе „остају на периферији", слабо или незнатно захваћене развојем. *Jansky* (2016, стр. 65) регионалне диспаритете дефинише као друштвено непожељне разлике које постоје између појединих региона у погледу достигнутог степена (односно, нивоа) развијености, посматрано из угла економске, друштвене (социјалне), еколошке, или пак неке друге димензије развоја. Присуство регионалних диспаритета у погледу достигнутог степена друштвено-економске развијености својствено је како развијеним тако и, мада у већем интензитету, земљама у развоју (*Mohiuddin & Hashia*, 2012, стр. 86; *Obradović et al.*, 2016, стр. 162; *Miljačić & Paunović*, 2011, стр. 380).

[117] Појам „регион" (изведен од латинске речи *regio*), по дефиницији, означава област, рејон, зону, крај, суседство, површину, округ, подручје, поље, простор, место, предео. У оквиру овог Одељка, употреба наведеног термина извршена је за потребе означавања територијално заокруженог дела простора у саставу једне конкретне државе, независно од њене величине, који се одликује, с једне стране, скупом карактеристика који га повезују са целином државе и скупом (другачијих) карактеристика који опредељују његову специфичност, с друге стране (*Polednikova*, 2014, стр. 497). Прецизније, термин „регион" се користи за означавање различитих нивоа територијалних целина унутар националне државе (односно, регионе, субрегионе (или области) и / или јединице локалне самоуправе).

конкурентности, садржана је и у чињеници да се и политика развоја Европске уније (ЕУ) темељи на регионалном принципу, који подразумева да развој региона доприноси развоју држава чланица, а самим тим и ЕУ у целини (РЗР, 2009, стр. 88). Наиме, усмерена на ублажавање изражених регионалних диспаритета који могу озбиљно угрозити целокупну структуру европске интеграције (РЗР, 2009, стр. 51), подршка равномерном регионалном развоју и активностима усмереним на јачање националне и регионалне конкурентности држава чланица издвојила се као један од кључних циљева политике развоја ЕУ (*Melecky*, 2014, стр. 581; *Polednikova*, 2014, стр. 496; *Pintilescu*, 2011, стр. 46; *Goletsis & Chletsos*, 2011, стр. 174; *Amendola et al.*, 2004, стр. 7; *Lukić & Anđelković-Stoilković*, 2017, стр. 66). У контексту претходних разматрања, отклањање или ублажавања регионалних развојних неравномерности нужно подразумева дефинисање институционалног оквира регионалног развоја на нивоу конкретне државе, као основе за успостављање и реализацију јасног и целовитог концепта регионалне политике, заснованом, у суштини, на перманентној институционалној изградњи капацитета јавне администрације на свим нивоима управљања, односно стварању услова за усклађену примену формулисаних принципа и механизама управљања регионалним развојем на свим деловима њене територије.

4.2. Институционални оквир актуелне регионализације и политике регионалног развоја у Републици Србији

Потврђујући резултате спроведених студија и мишљења бројних аутора (попут, *Abdollahzade & Sharifzadeh*, 2012, стр. 9; *Lukić & Anđelković-Stoilković*, 2017, стр. 66; *Puljiz & Maleković*, 2007, стр. 1), према којима се транзиционе и земље у развоју издвајају као посебно осетљиве на интензивирање проблема регионалних неједнакости, Република Србија се одликује веома израженим међурегионалним и унутаррегионалним диспропорцијама и асиметричностима у погледу степена развијености територијалних јединица у њеном саставу, са тенденцијом њиховог континуираног повећања (Влада РС, 2007а, стр. 1; *Krstić & Vukadinović*, 2011, стр. 554; *Miljačić & Paunović*, 2011, стр. 387–388; *Winkler*, 2012, стр. 82; *Vukmirović*, 2013, стр. 39; *Molnar*, 2013, стр. 68).[118] Забрињавајуће размере регионалних диспаритета, манифестоване у вишеслојној регионалној и унутаррегионалној поларизацији, примарно у правцу „развијени север – неразвијени југ“ (*Čoček*, 2010, стр. 15; *Miljačić & Paunović*, 2011, стр. 387; *Manić i drugi*, 2017, стр. 287), су последица вишедеценијског маргинализовања питања регионализације у економској политици, а делимично и научним круговима (*Molnar*, 2013, стр. 66), односно непостојања конзистентне и координиране регионалне политике, недовршеног институционалог оквира који системски уређује регионални развој Србије, али и наслеђених и транзиционих неусклађности привредних активности, интензивног процеса депопулације, као и дугогодишњег економског заостајања (РЗР, 2009, стр. 1).

У начелу, темељ изградње постојећег система спровођења регионалне политике у Републици Србији постављен је 2006. год., доношењем *Устава Републике Србије* (Народна скупштина РС, 2006) којим је, у члану 94, дефинисана обавеза државе да се

[118] Детаљна анализа стања и целовит динамички приказ регионалних неравномерности из угла различитих и бројних фактора релевантних за сагледавање достигнутог степена развијености територијалних јединица различитог нивоа посматрања у Републици Србији може се видети у: РЗР (2009), НАРР (2012), Министарство привреде, Сектор за регионални развој и стратешке анализе привреде (2014), Министарство привреде (2015), *Jakopin* (2014); (2015).

стара о равномерном и одрживом регионалном развоју, у складу са законом. Први стратешки кораци у реализацији уставних надлежности државе учињени су доношењем *Стратегије регионалног развоја Републике Србије за период од* 2007. *до* 2012. *године* (Влада РС, 2007а) и *Закона о регионалном развоју* (Влада РС, 2009а), чиме је створена институционална основа за увођење и примену принципа и механизама управљања развојем заснованим на поштовању савремених (европских) концепција. Како је наведено у њеном уводном делу: „Стратегија представља први стратешки развојни документ из области регионалног развоја који на конзистентан и целовит начин дефинише основне развојне приоритете регионалног развоја земље и начине њиховог остваривања“ (Влада РС, 2007а, стр. 1), а у циљу подстицања равномерног регионалног развоја на територији РС. Операционализација формулисаног циља усмерена је пре свега на следеће активности (Влада РС, 2007а, стр. 2): (1) подизање регионалне конкурентности; (2) смањење регионалних неравномерности и сиромаштва и (3) изградња институционалне регионалне инфраструктуре. У контексту дефинисаних циљева дисертације, односно верификације њихове актуелности и значаја, важно је истаћи да се као један од три кључна стуба у основи Стратегије управо наводи утврђивање степена развијености – категоризација и типологија подручја, док се процес континуираног прикупљања података и информација за мерење успешности предузетих активности издваја као веома важан корак у имплементацији исте (Влада РС, 2007а, стр. 2–3).

Закон о регионалном развоју, с друге стране, представља разраду комплексног система подршке процесу спровођења и имплементације политике управљања регионалним развојем Републике Србије, како на државном, тако и регионалном, субрегионалном и општинском нивоу (РЗР, 2009, стр. 102). Прецизније, њиме се уређују циљеви и начела подстицања регионалног развоја, развојни документи, субјекти (институционални органи) надлежни за управљање регионалним развојем, мере, подстицаји и финансијски извори за спровођење политике регионалног развоја, као и надзор и евалуација имплементације исте. Такође, овим законом одређују се називи развојних региона, уређује начин одређивања области и јединица локалне самоуправе, које чине регион, односно област, респективно, дефинишу показатељи степена развијености региона и јединица локалне самоуправе, као и њихово рангирање према степену развијености (Влада РС, 2009а, измене и допуне 2010 и 2015, чланови 7–13.).

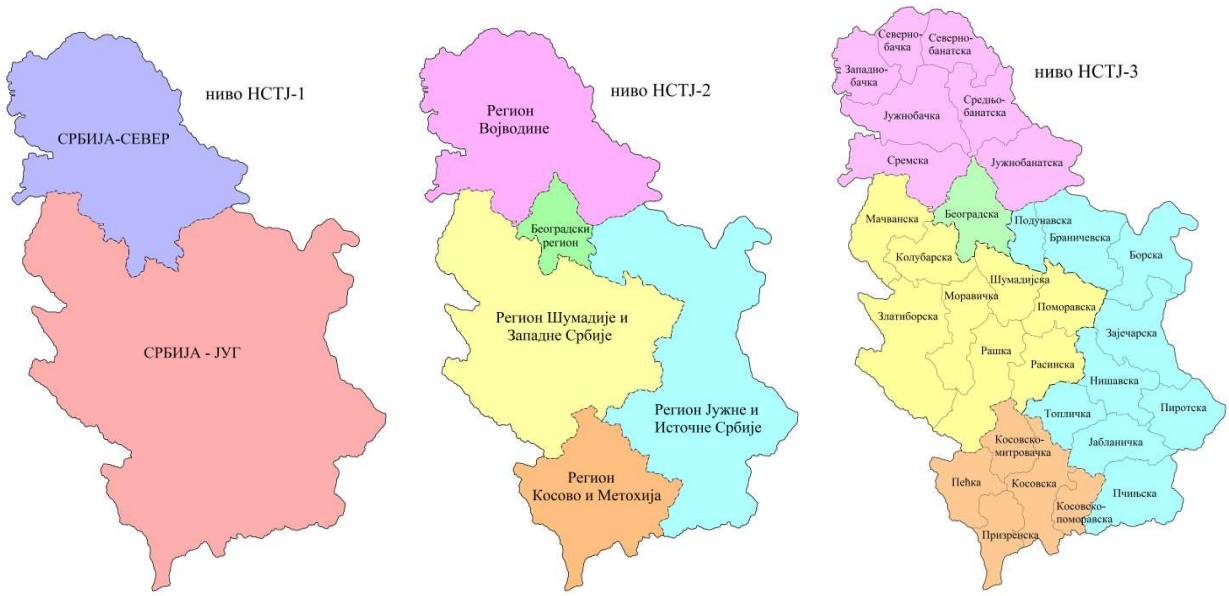
Поред наведеног, у складу са чланом 5., претходно наведеног закона, пратећег подзаконског акта под називом *Уредба о номенклатури статистичких територијалних јединица* (Влада РС, 2009б, измене и допуне 2010) и *Законом о територијалној организацији РС* (Влада РС, 2007б) успостављена је административно–статистичка подела територије РС (Табела 4.2.1), која представља важан инструмент успешног спровођења политике регионалног развоја. *Уредбом о номенклатури статистичких територијалних јединица* (акроним: *НСТЈ*), целокупна територија Републике Србије подељена је, према сваком од три (основна) нивоа *НСТЈ* хијерархијске класификације [119], на: ▸ 2 функционалне целине нивоа *НСТЈ*-1: Србија–север и Србија–југ; ▸ 5 територијалних целина нивоа *НСТЈ*-2 (региони); и ▸ 30 територијалних области, нивоа *НСТЈ*-3.

[119] „Номенклатура статистичких територијалних јединица представља скуп појмова, назива и симбола који описују групе територијалних јединица са нивоима груписања и која садржи критеријуме по којима је извршено груписање који су уређени у складу са стандардима Европске уније“ (*Закон о регионалном развоју*, 2009, члан 4.).

Табела 4.2.1. Актуелна регионализација Републике Србије према НСТЈ класификацији

Ниво НСТЈ-1	Ниво НСТЈ-2 (региони)	Ниво НСТЈ-3 (области)	ЈЛС – локални ниво (локалне самоуправе)
Србија-север	Регион Војводине	7 области: 1. Западнобачки управни округ 2. Јужнобанатски управни округ 3. Јужнобачки управни округ 4. Севернобанатски управни округ 5. Севернобачки управни округ 6. Средњобанатски управни округ 7. Сремски управни округ	8 градова и 37 општина
	Београдски регион	1 област: Београдска област (територија Града Београда)	Град Београд са својих 17 градских општина
	Регион Шумадије и Западне Србије	8 области: 1. Златиборски управни округ 2. Колубарски управни округ 3. Мачвански управни округ 4. Моравички управни округ 5. Поморавски управни округ 6. Расински управни округ 7. Рашки управни округ 8. Шумадијски управни округ	10 градова и 42 општине
Србија-југ	Регион Јужне и Источне Србије	9 области: 1. Борски управни округ 2. Браничевски управни округ 3. Зајечарски управни округ 4. Јабланички управни округ 5. Нишавски управни округ 6. Пиротски управни округ 7. Подунавски управни округ 8. Пчињски управни округ 9. Топлички управни округ	9 градова са градским-општинама и 38 општина
	Регион Косово и Метохија	5 области: 1. Косовски управни округ 2. Косовско-митровачки управни округ 3. Косовскопоморавски управни округ 4. Пећки управни округ 5. Призренски управни округ	1 град и 28 општина

Извор: Ауторов приказ



Слика 4.2.1. Картографски приказ територијалне организације Републике Србије према различитим нивоима *НСТЈ* (енгл. *NUTS*) класификације

Представљеном статистичком регионализацијом територијалног простора државе (Слика 4.2.1), заснована на критеријумима *NUTS* класификационе методологије [120], креирана је неопходна основа за ефикасно прикупљање упоредивих података, праћење, анализу и утврђивање степена развијености статистичких функционалних територијалних целина различитих нивоа *HCTJ*, као и праћење и евалуацију резултата примене дефинисаних мера и подстицаја политике регионалног развоја.

4.3. Значај и одлике поступка мерења степена развијености територијалних јединица

Генерално, објективна оцена и „мерење“ достигнутог степена развијености територијалних јединица на различитом административно-статистичком нивоу посматрања, [121] представља веома важну аналитичку активност у сложеном процесу успостављања равномерног регионалног развоја на целој територији државе. Наиме, прецизна евалуација нивоа развијености појединачних територија и резултати њихове међусобне компарације, представљају важан извор информација у погледу размера присутних развојних (регионалних) диспаритета, али и аналитичку основу за идентификовање развојних потенцијала и ограничења, којима се појединачне територијалне јединице одликују у конкретном временском тренутку и / или периоду посматрања (*Zhao et al.*, 2006, стр. 1; *Melecky*, 2012, стр. 979; 2014, стр. 583). Наведене информације су од фундаменталног значаја и имају кључну улогу у поступку стратешког планирања равномерног регионалног развоја и ефикасно / ефективно спровођење политике регионалног развоја (*Wishlade & Yuill*, 1997, стр. 3; *Coombes & Wong*, 1994, стр. 1297; *Rovan & Sambt*, 2003, стр. 265; *Ozaslan et al.*, 2006, стр. 5; *Goletsis & Chletsos*, 2011, стр. 174, 175; *Abdollahzade & Sharifzadeh*, 2012, стр. 9; *Perišić & Wagner*, 2015, стр. 206; *Angelis et al.*, 2016, стр. 71; *Molnar*, 2013, стр. 66). Наиме, детаљно сагледавање развојних специфичности појединачних територијалних јединица у значајној мери олакшава и чини ефективнијим избор, односно формулисање и прецизирање, адекватних развојних планова, програма и подстицајних мера намењених конкретном региону / области / општини или граду, чиме се такође, с обзиром на њихову прилагођеност индивидуалним диференцијалним предностима / недостацима разматраних територија, повећава и ефикасност реализације истих (*Bromley*, 1971, стр. 319; *Alasia*, 1996, стр. 1; *Wishlade & Yuill*, 1997, стр. 3; *Zhao et al.*, 2006, стр. 1; *Lipshitz & Raveh*, 1998, стр. 747; *Krstić & Vukadinović*, 2011, стр. 553). Поред наведеног, мерење степена развијености и сагледавање размера регионалних диспаритета омогућава ефикасну евалуацију и мониторинг успешности предложених / спроведених програма и мера политике равномерног

[120] Акроним *NUTS* (фра. *Nomenclature des Unités Territoriales Statistiques*; односно, енгл. *Nomenclature of Territorial Units for Statistics*) означава заједничку (јединствену) методологију за статистичку класификацију субнационалних територијалних јединица држава чланица Европске уније, креиране у Статистичком уреду Европске уније (енгл. *Statistical office of the EU – EUROSTAT*). Према *NUTS* класификационој методологији, територија сваке државе чланице дели се на пет нивоа, и то: (1) три (основна) нивоа: *NUTS* I (еквивалентно, *HCTJ*-1), *NUTS* II (*HCTJ*-2) и *NUTS* III (*HCTJ*-3) ниво; и (2) два додатна (нижа) нивоа, *LAU*-1 и *LAU*-2 (енгл. *Local Administrative Units*), који нису директно обухваћени *NUTS* класификацијом. Наведени нижи нивои класификације односе се, у случају Србије, на локални ниво (*ЈЛС* ниво), односно јединице локалне самоуправе (градове и општине), које сачињавају територијалне јединице – области, нивоа *HCTJ*-3.

[121] У складу са *HCTJ* методологијом одређивања статистичких региона и актуелној територијалној организацији Републике Србије, истраживања везана за феномен регионалне развијености конкретних територијалних јединица и диспаритета између њих могу бити класификована, према нивоу посматрања, на следећи начин: регионална (ниво НСТЈ-2), унутаррегионална (ниво НСТЈ-3) и унутар-субрегионална или локална (ниво локалних самоуправа, ЈЛС ниво (градови / општине)) (прилагођено према *Molnar*, 2013, стр. 66).

регионалног развоја, усмерених на ублажавање и / или отклањање регионалних диспаритета (Влада РС, 2007а, стр. 3; *Munandar & Azhari*, 2015, стр. 137), чиме се омогућава и њихова (евентуално потребна) корекција, током времена. Додатни аспект на који је неопходно указати, у контексту изнетих разматрања, односи се на чињеницу да је сврсисходност мера регионалне политике, а тиме и ефикасност њиховог спровођења, у великој мери условљена и детерминисана квалитетом и поузданошћу статистичких података који се користе приликом реализације регионалних истраживања усмерених на анализу нивоа развијености и размера регионалних диспаритета (*Molnar*, 2013, стр. 69).

Генерално, иако од изузетно великог значаја за формулисање одговарајућих стратегија, политика развоја и адекватну алокацију подстицајних средстава и других инструмената подршке од стране надлежних државних органа и институција, оцена достигнутог нивоа развијености појединачних статистичких региона и локалних самоуправа представља, у концептуално-методолошком смислу, веома захтеван и „тежак задатак“ (*Meyer et al.*, 2016, стр. 101; HAPP, 2012, стр. 10), чија реализација не подразумева рутински приступ и „лака решења“ (*Rovan et al.*, 2009, стр. 93). Примарни узрок наведене комплексности садржан је у мултидимензионој природи концепта регионалне развијености и, консеквентно, регионалних диспаритета (*Das*, 1999, стр. 313; *Milenković et al.*, 2014, стр. 604; *Nielsen*, 2011, стр. 5; *Maletić & Bucalo-Jelić*, 2016, стр. 13) [122] као специфичних, у начелу, латентних феномена који представљају предмет истраживања, односно (индиректног) „мерења“ (HAPP, 2012, стр. 10; Влада РС, 2007а, стр. 86). Дефинишући феномен регионалног развоја као комплексни скуп процеса који се дешавају унутар територије појединачних региона и при томе доводе до појаве већих или мање изражених позитивних / негативних, квантитативних / квалитативних промена у домену економских, друштвених (социјалних) карактеристика, односно квалитета животне средине (еколошких одлика) или пак неких других димензија и развојних околности на нивоу конкретног региона, *Polednikova* (2014, стр. 497) потврђује и образлаже претходно изнету констатацију. Другим речима, (општи) степен развијености територија заправо представља резултат „експлоатације“ регионалних специфичности и развојних потенцијала, детерминисаних деловањем динамичких процеса и интензивних промена у домену друштвено-политичког, привредног и социјалног амбијента, али и географских одлика и расположивих природних ресурса (HAPP, 2012, стр. 10; *Braičić & Lončar*, 2011, стр. 95). У релевантној литератури се као главне и најчешће коришћене, за потребе оцене достигнутог степена развијености појединачних територија и њиховог развојног потенцијала / лимитираности, углавном издвајају следеће димензије регионалног развоја (*Das*, 1999, стр. 313; *Ozaslan et al.*, 2006, стр. 7; Влада РС, 2007а, стр. 86; РЗР, 2009, стр. 14; *Goletsis & Chletsos*, 2011, стр. 174; *Bahovec et al.*, 2011, стр. 87; HAPP, 2012, стр. 11–12; *Polednikova*, 2014, стр. 497; *Istrate & Horea-Serban*, 2016, стр. 202): економска, друштвена (или социјална), еколошка, инфраструктурна, демографска (и/или образовна) итд.

Аспект који додатно доприноси повећању комплексности поступка квантификовања степена регионалне развијености и размера утврђених асиметричности, односи се на чињеницу да свака од наведених димензија регионалног развоја заправо може бити

[122] Детаљније о кључним одликама различитих типова регионалних диспаритета и њиховој класификацији у зависности од апострофиране развојне димензије, видети, на пример, у: *Melecky* (2012, стр. 980) и *Polednikova* (2014, стр. 498).

схваћена као засебна мултидимензиона латентна променљива, чије се „мерење“ по правилу спроводи индиректно, на бази симултане анализе вредности и веза које постоје између више (из угла конкретне димензије) репрезентативних (опсервабилних, мерљивих) нумеричких показатеља (Влада РС, 2007а, стр. 86; *Nielsen*, 2011, стр. 7; *Meyer et al.*, 2016, стр. 102; *Melecky*, 2012, стр. 980; *Polednikova*, 2014, стр. 498).

Такође, уз неизоставно уважавање значаја свих појединачних, за анализу расположивих, димензија регионалне развијености, неопходно је и истаћи, сагласно мишљењу великог броја аутора у литератури (попут, на пример: *Rovan & Sambt*, 2003, стр. 265; *Das*, 1999, стр. 313; *Zhao et al.*, 2006, стр. 1; *Pintilescu*, 2011, стр. 46; *Polednikova*, 2014, стр. 496), кључну улогу економске димензије и њој карактеристичних показатеља у оцени степена развијености територијалних јединица, с обзиром на доминантну позицију економских односа у процесу „стварања једнаких шанси за све регионе“, односно постизања равномерног регионалног развоја (Влада РС, 2007а, стр. 73; *Bojović*, 2010, стр. 10; *Meyer et al.*, 2016, стр. 102). Изнето мишљење потврђују *Obradović et al.* (2016, стр. 162) наводећи да су регионални диспаритети у највећој мери условљени присутним разликама у погледу регионалне економске структуре. Посматрано у националном контексту, у *Извештају о регионалном развоју за 2009. год.* (РЗР, 2009, стр. 14), апострофирано је да су регионалне асиметричности у Србији у директној зависности од економске развијености, уз напомену да се размере регионалних неравномерности најбоље илуструју путем репрезентативних економских индикатора. Дакле, уз неоспоран значај у процесу ублажавања социјалних неједнакости, уравнотежења демографских кретања и заштите животне средине, равномерни регионални развој, који у суштини представља својеврсни структурни проблем једне државе, има превасходно економски значај (НАРР, 2012, стр. 10), будући да „без економске једнакости нема ни националне (односно, регионалне) равноправности“ (Влада РС, 2007а, стр. 72).

Логично и очекивано, апострофирани мултидимензиони карактер концепта регионалне развијености, односно регионалних диспаритета и појединачних развојних димензија, условио је „померање“ аналитичког оквира од (традиционалног) једнодимензионог праћења вредности великог броја појединачних показатеља различитих развојних димензија ка развоју и примени разноврсних софистицираних мултидимензионих методолошких поступака заснованих на експлоатацији апликативних потенцијала метода мултиваријационе статистичке анализе података (*Polednikova*, 2014, стр. 499; *Melecky*, 2012, стр. 979; *Meyer et al.*, 2016, стр. 101) у домену истраживања регионалних карактеристика и квантификовању присутних асиметричности. Употребљени појединачно или у комбинацији, наведени статистички методи представљају погодно аналитичко средство за „мерење“ степена регионалне развијености конкретних територијалних јединица (углавном засновано на конструкцији композитних показатеља), њиховој класификацији у интерно-хомогене / екстерно-хетерогене групе, сходно расположивим потенцијалима и / или ограничењима, или пак креирање одговарајућих статистичких модела намењених предвиђању развојне групе (категорије) конкретних територија на бази појединачних вредности, анализом обухваћених, развојних показатеља. Узимајући у обзир изражену варијететност расположивих метода мултиваријационе анализе, како из угла њихових појединачних специфичности, тако и апликативних потенцијала њихове комбиноване примене, али и приступа у имплементацији истих,

сасвим је оправдано закључити, односно потврдити мишљење *Melecky*-а (2012, стр. 979), да не постоји један универзални и „најбољи“ методолошки приступ мерењу регионалне развијености и територијалних диспаритета. Ову констатацију потврђују *Mazzocchi & Montresor* (2000, стр. 31) наводећи како различити приступи у имплементацији наведених метода често могу резултирати међусобно некомпатибилним (другачијим) исходима и закључцима, иако су спроведени на истим подацима.

Генерално, питање броја и избора конкретне(их) димензије(а) развоја, као и прецизирање њихове структуре из угла појединачних репрезентативних показатеља,[123] у великој мери зависи од дефинисаних циљева истраживања, расположивости података, а посебно од територијалног приступа, односно, конкретног *НСТЈ* или *ЈЛС*, нивоа на којем се налазе територијалне јединице чија се развијеност „мери“ и сагледава (НАРР, 2012, стр. 12). У том смислу, а полазећи од правилности да се регионалне асиметрије повећавају што је ниво посматрања / „мерења“ нижи, важно је истаћи да регионална истраживања спроведена на нижим нивоима територијалне агрегације, конкретно локалном нивоу (градови / општине), најбоље указују на величину асиметрије и дубину проблематике регионалне (не)развијености (НАРР, 2012, стр. 23; *Alasia*, 1996, стр. 1; *Mohiuddin & Hashia*, 2012, стр. 87). У контексту разматране проблематике, посматрано на нивоу Републике Србије, одређивање степена развијености региона и јединица локалне самоуправе спроводи се на основу смерница дефинисаних *Уредбом о утврђивању методологије за израчунавање степена развијености региона и јединица локалне самоуправе* (*Vlada RS*, 2015). Прецизније, на основу критеријума дефинисане методологије, одређивање степена развијености региона заснива се на сагледавању вредности бруто домаћег производа по глави становника (БДП *per capita*) појединачних региона у односу на републички просек, након чега се врши њихова категоризација у групу развијених или недовољно развијених у зависности од тога да ли је регионална вредност БДП *per capita* изнад или испод поменуте просечне вредности. Насупрот наведеном универијационом (једнодимензионом) приступу, оцењивање степена развијености јединица локалне самоуправе спроводи се на основу израчунатих вредности композитног показатеља, који се, у начелу, утврђује као пондерисани просек одступања вредности пет друштвено–економских показатеља [124] од републичког просека (*Vlada RS*, 2015, стр. 2). У складу са одредбама *Закона о регионалном развоју* (Влада РС, 2009а, чланови 11–13), на основу утврђених вредности *индекса развијености* врши се разврставање јединица локалних самоуправа у једну од дефинисаних категорија према степену опште развијености. Кључна карактеристика поменутог композитног показатеља садржана је у релативно субјективном приступу одређивању висине појединачних пондера. У наведеном контексту, а полазећи од јасно уочљиве сличности индекса развијености који се користи у Србији и Хрватској, детаљан критички осврт на методолошки приступ у израчунавању поменутог индекса, представили су *Perišić & Wagner* (2015), предлажући уједно и релативно објективнији приступ за одређивање пондера заснован на примени метода мултиваријационе анализе.

---

[123] Генерално, поступак избора и алоцирање појединачних показатеља унутар конкретних димензија развоја карактерише се израженом арбитрарношћу (односно, субјективношћу) истраживача (*Rovan et al*., 2009, стр. 93; *Milenković et al.*, 2014, стр. 604; *Krishnan*, 2010, стр. 11). У прилог наведеном, *Wishlade & Yuill* (1997, стр. 10) наводе да је, у пракси, готово немогуће извршити класификацију показатеља на начин који неће бити подложан критичкој дискусији.

[124] Показатељи који се користе за израчунавање индекса развијености су: стопа незапослености, доходак по становнику, изворни приход по становнику, степен образовања и стопа пада или раста становништва.

4.4. Преглед досадашњих истраживања у примени мултиваријационих метода за мерење регионалних диспаритета

Проблематика неравномерног регионалног развоја, односно потрага за теоријским моделима његових узрока, квантитативним приступима за оцену развијености територијалних јединица и сагледавање ефикасности предложених мера за ублажавање регионалних диспаритета, припада оним истраживачким темама чија се апликативна вредност и друштвени значај подразумевају сами по себи (*Braičić & Lončar*, 2011, стр. 93; *Goletsis & Chletsos*, 2011, стр. 174). У том смислу, посебно атрактивну истраживачку нишу представника, како домаће тако и иностране, научне / стручне заједнице чини „мерење“ степена развијености територијалних јединица у саставу конкретне земље или групе земаља коришћењем различитих комбинација репрезентативних показатеља једне или више димензија развоја и креирање резултирајућих класификација разматраних територија у релативно хомогене групе (*Stamenković & Savić*, 2017, стр. 104; *Sokolovska i drugi*, 2014, стр. 9). Наведену констатацију потврђује значајан број, али и разноврсност, публикованих научних радова и спроведених емпиријских истраживања (*Villaverde & Maza*, 2009, стр. 6). Већина предложених приступа у оцени нивоа регионалне развијености се заснива на појединачној и / или комбинованој имплементацији одређених мултиваријационих метода. Ради потврде апликативних могућности ове групе статистичких метода, а у складу са опредељеним предметом и циљевима дисертације, у Табели 4.4.1, представљене су кључне методолошке одреднице одабраних релевантних приступа, сличног карактера истраживању презентираном у наредном Поглављу.

Представљена мултиваријациона истраживања одликују се израженом варијететношћу и то, превасходно, у погледу аналитичких питања, која се односе на:

- *просторни обухват анализе* – усмереност истраживачког фокуса на анализу регионалне развијености и диспаритета између територијалних јединица различитог нивоа посматрања (*NUTS I, II, III, LAU*) у оквиру различитих држава или група држава;
- *временски обухват анализе* – коришћење података за различите (конкретне) године, израчунавање просека за одређени период и сл.;
- *селекцију појединачних показатеља и димензија развијености* – условљеност избора дефинисаним циљевима истраживања, доступношћу података за дати ниво територијалног обухвата, уз изражену субјективност истраживача при алокацији појединачних показатеља унутар апостофираних димензија развоја;
- *коришћени метод мултиваријационе анализе* – условљеност избора креативношћу истраживача у погледу осмишљавања начина за експлоатацију апликативних потенцијала појединачне и / или комбиноване примене мултиваријационих метода, у контексту циљева истраживања. Управо овај аспект варитетности мултиваријационих истраживања представља предмет разматрања у наставку излагања.

Кључне методолошке одреднице мултиваријационих истраживања у домену мерења нивоа регионалне развијености и размера регионалних диспаритета, издвојене мета-анализом садржаја радова у Табели 4.4.1, могу се сумирати у форми следећих запажања:

✓ Посматрано из угла појединачне и / или комбиноване примене са неком од других метода, анализа груписања и факторска анализа представљају најчешће коришћене методе мултиваријационе анализе у домену разматране проблематике.

Табела 4.4.1. Компаративни преглед релевантних мултиваријационих истраживања

Аутор(и) / (година публикације)	Временски обухват	Територијалне јед. NUTS / LAU ниво	Држава(е)	Димензије развоја (показатељи)	Коришћени МВА метод(и)
Maletić & Bucalo-Jelić (2016)	2012.год.	LAU	Србија	Ек./Соц./Агр.	FA / CA
Lepojević i drugi (2015)	2012.год.	LAU ∈ (1) NUTS II	Србија	Ек./Дем./Обр.	CA
Winkler (2012)	2009.год.	LAU	Србија	Ек./Дем./Обр./Соц./Здр.	FA
Rašić-Bakarić (2006)	2001.год.	LAU ∈ (4) NUTS III	Хрватска	Ек./Дем./Обр.	FA / CA
Rašić-Bakarić (2005)	2001.год.	LAU ∈ (3) NUTS III	Хрватска	Ек./Дем./Обр.	FA / CA
Rimac i drugi (1992)	1991.год.	LAU	Хрватска	Ек./Дем./Соц.	FA / CA
Rovan et al. (2009)	2005.год.	LAU	Словенија	Ек./Дем./Соц./Екол.	PCA / CA
Rovan & Sambt (2003)	2001.год.	LAU	Словенија	Ек./Дем./Соц./Обр.	CA
Brauksa (2013)	неуједначено	LAU	Летонија	Ек./Соц.	CA
Pastor et al. (2009)	неуједначено	LAU ∈ (1) NUTS II	Шпанија	Ек./Дем./Соц.	FA / CA / DA
Soares et al. (2003)	1995.год.	LAU	Португалија	Ек./Дем./Соц./Обр.	FA / CA
Mazzocchi & Montresor (2000)	1990.год.	LAU ∈ (1) NUTS II	Италија	Ек./Дем./Соц./Агр.	PCA / CA
Pintilescu (2011)	2008.год.	NUTS II	Румунија	Ек./Соц.	PCA
Michaelides et al. (2006)	2001.год.	NUTS II	Грчка	Ек.	CA
Goletsis & Chletsos (2011)	1995 / ’00 /’07.	NUTS II	Грчка	Ек./Соц./Обр./Здр.	FA / CA
Polednikova (2014)	2010.год.	NUTS II	Вишеградска 4	Ек./Соц.	CA
Melecky (2014)	2004–'08–'12.	NUTS II	ЕУ–28	Ек./Соц./Инфр./Обр./Тех.	FA / CA
del Campo et al. (2008)	2003.год.	NUTS II	ЕУ–25	Ек./Дем./Обр.	FA / CA
Avram & Postoiu (2016)	2007/2012.год.	NUTS II	ЕУ–27	Ек./Обр.	CA
Aumayr (2006)	2002.год.	NUTS III	ЕУ–25	Ек./Дем.	CA
Aumayr (2007)	2003.год.	NUTS III	ЕУ–25	Ек./Соц./Дем.	MR / FA / CA
Stamenković & Savić (2017)	2013.год.	NUTS III	Србија	Ек.	FA / CA
Стаменковић и други (2017)	2011.год.	NUTS III	Србија	Дем./Обр.	CA
Игић (2014)	2011.год.	NUTS III	Србија	Ек./Дем./Тех./Екол.	FA
Janković-Milić et al. (2013)	2011.год.	NUTS III	Србија	Ек./Дем.	CA
Čoček (2010)	2007.год.	NUTS III	Србија	Ек./Соц./Дем.	PCA
Jurun & Ratković (2012)	2010.год.	NUTS III	Хрватска	Предузетништво	MR / CA
Kurnoga-Ž. (2007)	2006.год.	NUTS III	Хрватска	Ек./Соц.	CA / FA / DA
Kurnoga-Ž. & Sorić (2008)	неуједначено	NUTS III	Хрватска	Апсорпција сред. фонд	CA
Kvičalova et al. (2014)	2011.год.	NUTS III	Чешка	Ек./Соц.	CA
Vydrova & Novotna (2012)	2010.год.	NUTS III	Чешка	Ек./Обр./Соц./Здр.	FA / CA
Jansky (2016)	/	NUTS III	Чешка	Ек./Соц./Екол.	FA
Istrate & Horea-Serban (2016)	2014.год.	NUTS III	Румунија	Ек.	CA
Jaba et al. (2007)	2003.год.	NUTS III	Румунија	Ек./Соц./Обр.	DA
Kronthaler (2003)	2000.год.	NUTS III	Немачка	Ек.	CA / DA
Capriati (2005)	2001.год.	NUTS III	Италија	ИР димензија	CA
Perišić (2014)	2006 –‘08.год	NUTS III & LAU	Хрватска	Ек./Дем./Соц./Обр.	CA / DA
Ozaslan et al. (2006)	2000.год.	NUTS I / II / III	Турска	Ек./Дем./Соц./Обр.	PCA
Jun (2010)	2008.год.	31 провинција	Кина	Ек.	PCA / CA
Krishnan (2010)	2006.год.	Подручја дисеминације	Канада	Ек./Соц./Дем.	FA
Alasia (1996)	1996.год.	Окрузи (NUTS III-CDs)	Канада	Ек./Соц./Дем.	FA
Das (1999)	1990-1991.год.	16 држава	Индија	Ек./Соц./Здр./Тех.	PCA

Напомене везане за значење коришћених скраћеница:

Колона „Коришћени МВА метод(и): *анализа груписања* (CA), *факторска анализа* (FA), *анализа главних компоненти* (PCA), *вишеструка регресија* (MR), *дискриминациона анализа* (DA);

Колона „Димензије развоја (показатељи): *економска* (Ек.), *социјална* (Соц.), *демографска* (Дем.), *образовна* (Обр.), *здравствена* (Здр.), технолошка (Тех.), квалитет животне средине (Екол.), *инфраструктурна* (Инфр.), *аграрна* (Агр.), *истраживање и развој* (ИР), *апсорпција средстава из фондова ЕУ* (Апсорпција сред. фонд);

Колона NUTS / LAU ниво: LAU ∈ (1) NUTS II, на пример, означава да су анализиране јединице локалне самоуправе (градови / општине) у саставу једног (1) конкретног статистичког региона, нивоа NUTS II.

Извор: Ауторов приказ

✓ У зависности од дефинисаног циља и сврхе њеног спровођења, анализи груписања се углавном додељује једна од следеће две „аналитичке улоге“: ▸ *секундарна* (или *споредна*) *улога*, ако се њени резултати користе за проверу прецизности класификације територијалних јединица утврђене на бази њиховог рангирања сходно вредностима претходно креираног композитног показатеља (*Goletsis & Chletsos*, 2011; *Rovan et al.*, 2009; *Soares et al.*, 2003, *Stamenković & Savić*, 2017), и ▸ *примарна* (или *главна*) *улога*, у условима када њен резултат, односно класификација територијалних јединица у одређене интерно–хомогене / екстерно–хетерогене групе, на основу утврђеног степена сличности (или различитости) између њих, представља основни циљ истраживања (сви остали радови у Табели 4.4.1, који у последњој колони садрже ознаку *CA*, независно да ли је реч о самосталној или комбинованој примени).

✓ Насупрот истраживањима заснованим на појединачној примени анализе груписања када се матрица различитости, као полазна основа у извођењу класификационе структуре, утврђује на бази (углавном стандардизованих) вредности изворних показатеља различитих димензија развијености, у случају додељене примарне улоге и комбиноване примене са факторском или анализом главних компонената, груписање територијалних јединица заснива се (доминантно) на коришћењу вредности претходно издвојених латентних променљивих (заједничких фактора / главних компоненти).[125]

✓ Уз, генерално, подједнаку заступљеност појединачне примене хијерархијске (*Maletić & Bucalo-Jelić*, 2016; *Rimac et al.*, 1992; *Melecky*, 2014; *Vydrova & Novotna*, 2012; *Jun*, 2010; *Istrate & Horea-Serban*, 2016; *Aumayr*, 2006; *Kvičalova et al.*, 2014; *Kronthaler*, 2003; *Goletsis & Chletsos*, 2011) и нехијерархијске процедуре груписања (*Rašić-Bakarić*, 2005; 2006; *Mazzocchi & Montresor*, 2000; *Lepojević i drugi*, 2015; *Stamenković & Savić*, 2017; *Avram & Postoiu*, 2016; *Michaelides et al.*, 2006; *Brauksa*, 2013; *Janković-Milić et al.*, 2013), као доминантни приступ у имплементацији апликативних потенцијала анализе груписања издваја се њихова комплементарна примена (*del Campo et al.*, 2008; *Soares et al.*, 2003; *Rovan & Sambt*, 2003; *Kurnoga-Živadinović*, 2007; *Aumayr*, 2007; *Pastor et al.*, 2009; *Rovan et al.*, 2009; *Kurnoga-Živadinović & Sorić*, 2008; *Perišić*, 2014; Стаменковић и други, 2017; *Jurun & Ratković*, 2012; *Polednikova*, 2014). Наведени приступ подразумева коришћење резултата хијерархијске процедуре, као улазних параметара при спровођењу нехијерархијског груписања, због чега је веома погодан за поређење добијених класификација, сагледавање одступања и, консеквентно, избор „квалитетнијих" резултата груписања.

✓ При коришћењу нехијерархијске процедуре искључиво је коришћен метод *k*-средина. С друге стране, имплементација хијерархијске процедуре искључиво се заснива на примени метода удруживања, при чему се, као доминантна, издваја заступљеност *Ward*-овог метода[126], након чега следи прилично скромна фреквентност употребе методе просечног повезивања (*Jurun & Ratković*, 2012; *Goletsis & Chletsos*, 2011). Такође, конкретно изабрани метод хијерархијског удруживања, у највећем броју случајева, уједно представља и једини метод чије су апликативне могућности и прилагођеност расположивим мултиваријационим опсервацијама, у контексту калсификације, разматране и испитиване. Изузетак, у том смислу, представљају истраживања спроведена од стране следећих аутора: *Kurnoga-Živadinović & Sorić* (2008), *Kurnoga-Živadinović* (2007), *Perišić* (2014). Генерално, увидом у анализиране радове, стиче се утисак да је избор конкретне методе хијерархијског груписања, код већине аутора, примарно извршен на бази субјективног мишљења и/или (недовољно информативних) смерница садржаних у актуелним „приручницима" за примену мултиваријационих метода коришћењем неке од расположивих софтверских платформи.

✓ Закључак, сличног садржаја претходно изнетом, може се формулисати и у погледу доминантно уоченог начина доношења одлуке која се тиче избора оптималног броја група у резултирајућим класификационим структурама, било да је реч о примени

[125] У случају када је анализи груписања додељена „споредна" улога, подједнако су заступљена оба приступа у извођењу жељених класификација, у зависности од формулисаних циљева, осмишљеног дизајна истраживања, али и структуре креираног композитног показатеља.

[126] У више од ¾ анализираних радова у којима је спроведена хијерархијска процедура груписања коришћен је *Ward*-ов метод.

једне / више метода хијерархијског груписања, или пак поређењу структуре класификација добијених путем хијерархијске и нехијерархијске процедуре. Наиме, поступак евалуације квалитета резултата хијерархијског груписања и, консеквентно, избора „оптималног“ броја група, доминантно се заснива на субјективном утиску и избору (такозваног) „најинтерпретабилнијег“ решења, у структури свих могућих решења, приказаних у форми дендрограма, или, евентуално, сагледавањем размера промена вредности мере одстојања између група током процеса удруживања (попут, на пример, *Polednikova*, 2014). Сходно наведеном, може се констатовати да је у анализираним радовима присутно изражено занемаривање употребе статистички заснованих (објективних) критеријума при евалуацији квалитета добијених резултата анализе груписање.

✓ У контексту претходног запажања, важно је указати и на методолошке приступе заснованим на примени дискриминационе анализе у функцији валидације, односно провере прецизности претходно извршене класификације на бази анализе груписања, демонстриране од стране следећих аутора: *Kronthaler* (2003), *Kurnoga-Živadinović* (2007), *Jaba et al.* (2007), *Pastor et al.* (2009), *Perišić* (2014). Наиме, поменути приступи подразумевају коришћење конкретног броја и структуре група, претходно издвојених путем анализе груписања, за дефинисање модалитета (категоријске) зависне променљиве у дискриминационој анализи, док улога независних променљивих може бити додељена оригиналним показатељима или пак, такозваним, факторским скоровима издвојених димензија развоја, у зависности од коришћене комбинације мултиваријационих метода у истраживању.

✓ У случају када је анализом обухваћен велики број показатеља различитих димензија регионалне развијености често практикован мултиваријациони приступ у истраживању разматране проблематике подразумева примену факторске или анализе главних компонената. Кључна апликативна предност наведених метода огледа се у редукцији димензионалности оригиналног истраживачког проблема која се њиховом применом постиже, будући да иста омогућава издвајање мањег броја заједничких фактора (односно, главних компоненти), који заправо представљају „главне“ димензије регионалног развоја посматраног територијалног простора око којих су груписани појединачни развојни показатељи. Вредности факторских скорова (или издвојених главних компоненти), као што је претходно истакнуто, могу бити коришћени као улазни елементи анализе груписања (за *FA+CA* видети: *Maletić & Bucalo-Jelić*, 2016; *Rimac et al.*, 1992; *Melecky*, 2014; *Vydrova & Novotna*, 2012; *del Campo et al.*, 2008; *Soares et al.*, 2003;[127] *Pastor et al.*, 2009; *Aumayr*, 2007; *Rašić-Bakarić*, 2005, 2006; односно, за *PCA+CA*: *Mazzocchi & Montresor*, 2000; *Jun*, 2010).

✓ Насупрот претходно наглашеној („споредној“) улози факторске и анализе главних компонената, у ситуацијама када је истраживачки фокус усмерен на оцену достигнутог степена развијености сваке територијалне јединице, из угла једне или више развојних димензија (као засебних мултидимензионих феномена), а не само на њихову

[127] Будући да (факторска) редукција димензионалности нужно подразумева „известан губитак“ информација садржаних у полазној матрици опсервација, *Soares et al.* (2003), наведени мултиваријациони приступ допуњује паралеленим спровођењем анализе груписања како на оригиналним тако и редукованим подацима, односно факторским скоровима, у циљу сагледавања евентуално присутних одступања између резултирајућих класификација и њихових размера.

класификацију унутар различитих група и резултирајућу интерпретацију одлика истих као мање или више развијених, апликативне могућности и својства наведених метода могу бити употребљене за креирање одређених синтетичких (композитних) показатеља (или индекса). Различити приступи употребе факторске анализе у контексту одређивања тежинских коефицијената у структури композитног индекса демонстрирани су од стране следећих аутора: *Alasia* (1996), *Krishnan* (2010), *Goletsis & Chletsos* (2011), *Winkler* (2012), Игић (2014), *Stamenković & Savić* (2017), *Jansky* (2016). Кореспондентни приступи, али засновани на употреби анализе главних компонената, предложени су у радовима следећих аутора: *Das* (1999), *Ozaslan et al.* (2006), *Rovan et al.* (2009), *Čoček* (2010), *Pintilescu* (2011).

Поред претходно наведених запажања, неопходно је указати и на следећи, веома важан, методолошки аспект, карактеристичан за анализиране радове. Наиме, према мишљењу аутора дисертације, један од недостатака значајног броја анализираних истраживања, односи се на занемаривање (изостављање), или пак, некомплетно испитивање испуњености одговарајућих статистичких претпоставки (на пример: мултиколинеарност, сингуларност, мултиваријациона нормалност, присуство мултиваријационих нестандардних опсервација, хомогеност матрица варијанси / коваријанси и сл.), на којима се заснива њихова статистички валидна имплементација. Нарушеност појединих статистичких претпоставки у значајној мери умањује могућност генерализације добијених решења (*Cziraky et al.*, 2005, стр. 2) и повећава ризик у погледу извођења погрешних закључака, чиме се, у крајњој мери, угрожава квалитет, на њима заснованих, подстицајних мера за ублажавање регионалних разлика, као и успешност њихове примене. Такође, недовољна транспарентност и спецификација спроведених аналитичких поступка, као и неујединачена и често конфузна употреба статистичке терминологије иманентне коришћеним методама, карактеристични за поједине радове, у великој мери онемогућавају и / или отежавају јасно разумевање статистичке логике појединих корака у структури примењеног концептуално-методолошког оквира, чиме се умањује могућност (евентуалне) репликације истраживања, за потребе обезбеђивања упоредивих резултата, као и могућност адекватног сагледавања практичног значаја и важности презентираних решења и формулисаних закључака.

Иако усмерена на испитивање разлика у погледу различитих димензија развијености између појединачних држава у саставу Европске уније или одређених регија на тлу Европе, будући да пружају користан преглед и демонстрацију апликативних могућности појединачних мултиваријационих метода, или пак, њихове комбиноване примене, у релеватној литератури посебно се могу издвојити и истраживања представљена у Табели 4.4.2.

Табела 4.4.2. Преглед додатних мултиваријационих истраживања – национални ниво

Аутор(и) / (година публикације)	Временски обухват	Територијални обухват	Димензије развоја (показатељи)	Коришћени MBA метод(и)
Žmuk (2015)	2015.год.	34 државе Европе	Квалитет живота	CA
Savić & Zubović (2015)	2002 / 2010.год.	8 држава Југоисточне Европе	Запосленост	FA
Melecky (2012)	2000 /'05./'10.год.	Вишеградска 4, Немачка и Аустрија	Ек./Соц./Екол.	CA
Bahovec et al. (2011)	2009.год.	ЕУ-27 (без Грчке), Хрватска и Турска	Економски	FA / CA
Kurnoga-Ž. et al. (2009)	2007.год.	ЕУ-27 и Хрватска	Економски	FA / CA
Lovrić & Stamenković (2016)	2006-'10. / 2011-15.	ЕУ-28 и државе ЕУ кандидати	Економски	CA

Извор: Ауторов приказ

Емпиријски део

ИСТРАЖИВАЊЕ МОГУЋНОСТИ И ИЛУСТРАЦИЈА ПРИМЕНЕ МУЛТИВАРИЈАЦИОНИХ МЕТОДА У САГЛЕДАВАЊУ СТЕПЕНА ЕКОНОМСКЕ РАЗВИЈЕНОСТИ ОПШТИНА У РЕПУБЛИЦИ СРБИЈИ

5.1. РАЗВОЈ МУЛТИВАРИЈАЦИОНОГ МОДЕЛА ЗА МЕРЕЊЕ СТЕПЕНА ЕКОНОМСКЕ РАЗВИЈЕНОСТИ ОПШТИНА У РЕПУБЛИЦИ СРБИЈИ

Сходно излагањима садржаним у претходним Поглављима, која се тичу концептуално–методолошких одређења одабраних метода мултиваријационе статистичке анализе података и важности обезбеђивања објективног квантификовања достигнутог нивоа развијености територијалних јединица у саставу државе из угла сагледавања размера присутних регионалних диспаритета, уз апострофирање мултидимензионе природе концепта регионалне развијености и појединачних развојних димензија, у овом Поглављу представљено је емпиријско истраживање засновано на мултиваријационој анализи одабраних показатеља економске димензије регионалне развијености јединица локалних самоуправа у Републици Србији. У том смислу, конципиран је и приказан иновативни методолошки оквир истраживања заснован на интегрисаној примени факторске анализе, анализе груписања и мултиваријационе анализе варијансе (*MANOVA*) у функцији развоја мултиваријационог модела за мерење степена економске развијености градова / општина у Републици Србији.

5.1.1. Дефинисање истраживачког проблема

У контексту дефинисаног предмета и основног циља дисертације, у фокусу овог дела емпиријског истраживања је испитивање и презентовање могућности статистички валидне комбиноване примене факторске, *MANOVA* и анализе груписања за потребе креирања и евалуације практичне значајности композитног показатеља намењеног мерењу степена економске развијености јединица локалне самоуправе у Србији.

Полазећи од наведеног, а сагласно дефинисаном сету посебних циљева дисертације, дефинисан је и циљ овог дела емпиријског истраживања који гласи: развој мултидимензионог статистичког модела у форми композитног показатеља, под називом индекс економске развијености (акроним, *ИЕР*), као погодне основе за класификацију јединица локалних самоуправа према достигнутом нивоу економске развијености и сагледавање размера диспаритета присутних између појединачних и / или група *ЈЛС*-а.

У том смислу, сврха спроведеног истраживања огледа се у обезбеђивању иновативног проширења, у релевантној литератури предложених, методолошких приступа у анализи проблематике регионалне развијености, конкретно, у домену утврђивања нивоа развијености, разврставања и типологије разматраних подручја.

Изложени предмет и дефинисани циљ у овом емпиријском делу су у непосредној вези са првом посебном хипотезом дисертације, формулисане на следећи начин:

Хипотеза 1: Примена комбинације одабраних метода мултиваријационе анализе омогућава развој статистичког модела (у форми одговарајућег композитног индекса) који може допринети успешном и прецизном мерењу достигнутог степена економске развијености посматраних територијалних јединица у саставу државе.

5.1.2. Методолошки оквир креирања композитног индекса економске развијености

За потребе реализације дефинисаног циља и проверу формулисане хипотезе дизајниран је оквир емпиријског истраживања који обухвата следеће кључне кораке: ▸ избор променљивих, ▸ формирање узорка мултиваријационих опсервација (дефинисање јединица посматрања, просторно-временског обухвата и извора података), ▸ претпроцесирање оригиналних података, ▸ обрада претпроцесираних података, и ▸ евалуација, приказ и интерпретација резултата. Наведени кораци детаљно су елаборирани у наставку.

Сходно наведеном, полазећи од актуелне статистичке регионализације територијалног простора Републике Србије, заснованој на *НСТЈ* (енгл. *NUTS*) класификацији, за потребе емпиријског истраживања прикупљени су и анализирани секундарни, у тренутку реализације истог, последње доступни подаци одабрана четири показатеља различитих аспеката економске развијености (Табела 5.1.1) за сваку од 165 територијалних јединица нивоа *ЈЛС* (енгл. *LAU*), односно јединица локалне самоуправе (градова / градских општина / општина) у саставу следећа четири (од укупно пет) региона (односно територијалних целина нивоа *НСТЈ*-2) у Републици Србији, и то: Београдски регион, Регион Војводине, Регион Јужне и Источне Србије и Регион Шумадије и Западне Србије. Подаци су прибављени из електронске базе података Агенције за привредне регистре [*http://www.apr.gov.rs/*] и публикације под насловом *Општине и региони у Републици Србији, 2016.*, Републичког завода за статистику (*РЗС*, 2016). Сви прикупљени подаци односе се на исту, 2015. годину. Такође, будући да *РЗС* од 1999. године не располаже подацима који се односе на територијалне јединице у оквиру Региона АП Косово и Метохија, исти нису садржани у обухвату података и спроведеним ауторским израчунавањима. Детаљан приказ коришћених извора и поступак израчунавања вредности одабраних економских показатеља, представљен је у форми напомена у Табели 5.1.1.

Табела 5.1.1. *Листа коришћених показатеља економске развијености општина*

Ознака	*Назив економског показатеља*	*Јединица мере*
X_1	Број предузећа (привредних друштава и предузетника) на 1000 становника	број предузећа
X_3	Стопа запослености	у %
X_4	Број незапослених на 1000 становника	број незапослених
X_5	Просечна зарада по запосленом	у хиљадама РСД

Напомене у вези са начином утврђивања вредности појединачних променљивих и коришћених извора:

Променљива X_1: количник (*збир броја активних привредних друштава и предузетника у 2015. години*) и (*процењеног броја становника за 2015. годину*) на нивоу конкретне општине, помножен са 1000;
Извор података: електронска база података АПР-а [*http://www.apr.gov.rs/*] и (*РЗС*, 2016).

Променљива X_2: количник вредности променљивих *Број запослених* (*укупно*) и *Контигент радно способног становништва* (*старости од 15 до 64 године*) у 2015. години, на територији посматране општине, помножен са 100;
Извор података: Публикација *Општине и региони у Републици Србији, 2016.* (*РЗС*, 2016).

Променљива X_3: директно преузете вредности из коришћеног извора података;
Извор података: Публикација *Општине и региони у Републици Србији, 2016.* (*РЗС*, 2016).

Променљива X_4: директно преузете вредности из коришћеног извора података;
Извор података: Публикација *Општине и региони у Републици Србији, 2016.* (*РЗС*, 2016).

Извор: Ауторова систематизација табеларног приказа

Приликом израчунавања вредности одабраних показатеља економске развијености општина, уместо коришћења апсолутних вредности (на пример, број предузећа, број запослених, број незапослених и сл.) извршено је њихово исказивање у форми вредности (или броја) по глави становника или одговарајућег релативног учешћа, како би се неутралисао или ублажио утицај укупне демографске масе појединачних територијалних јединица на исход мултиваријационе анализе и резултирајућу класификацију, сагласно препорукама бројних аутора истраживања сличног карактера (попут: *Alasia*, 1996; *Aumayr*, 2006; *Brauksa*, 2013; *Kronthaler*, 2003; *Ozaslan et al.*, 2006).

Оправданост извршеног избора наведених променљивих подржана је чињеницом да исте представљају најчешће коришћене показатеље економске развијености у оквиру репрезентативних публикација и истраживања усмерених на анализу нивоа развијености територијалних јединица у саставу Републике Србије (РЗР, 2009; НАРР, 2013; *Jakopin*, 2014).

Свеобухватан методолошки оквир за развој композитног (мултиваријационог) модела за мерење степена економске развијености јединица локалних самоуправа у Републици Србији са током спроведеног истраживања представљен је на Слици 5.1.1.

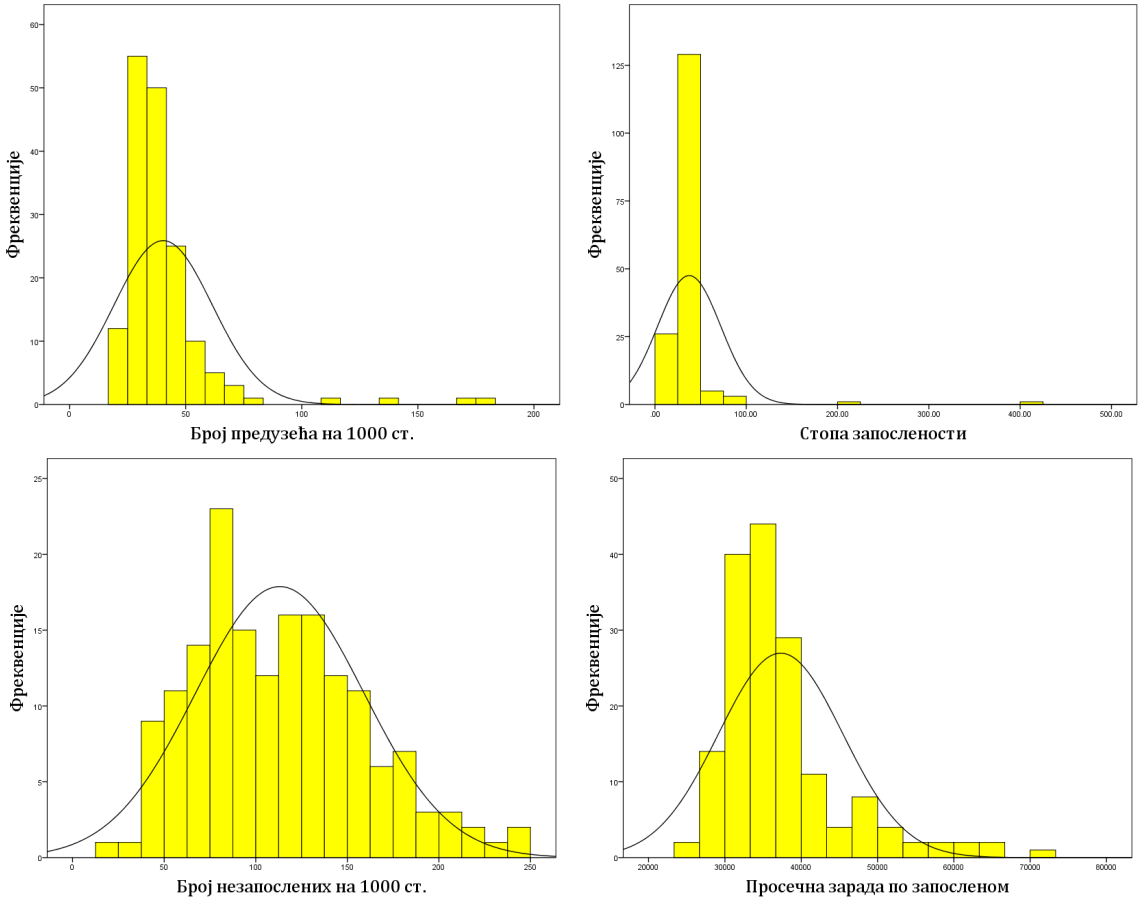
Слика 5.1.1. *Шематски приказ концептуално-методолошког оквира истраживања*

Извор: Ауторов визуелни приказ

5.1.3. Испитивање испуњености претпоставки у контексту примењених метода

Полазећи од чињенице да време, напор и ресурси утрошени током процеса (иницијалне) припреме расположивих података, заправо, представљају „улагање у мултиваријационо осигурање“, у складу са разматрањима изложеним у Одељку 2.1.2, пре примене изабране методе мултиваријационе анализе, конкретно факторске анализе, спроведено је детаљно испитивање степена испуњености статистичких претпоставки на којима се заснива њена валидна имплементација. У том смислу, испитивање универијационе и мултиваријационе нормалности распореда одабраних показатеља економске развијености општина у Србији, представља важан (први) корак у обезбеђивању бољег разумевања статистичке природе расположивих, анализом обухваћених, променљивих.

Универијациона нормалност распореда променљивих анализирана је путем графичког приказа хистограма фреквенција са нормалном кривом (Слика 5.1.2), као и поступка тестирања статистичких хипотеза заснованог на *Anderson-Darling*-овом, *Shapiro-Wilk*-овом, *Kolmogorov-Smirnov*-ом тесту, *Z* тесту симетричности (енгл. $Z_{skewness}$) и заобљености распореда (енгл. $Z_{kurtosis}$), чији су резултати представљени у Табели 5.1.2.



Слика 5.1.2. *Хистограми фреквенција са нормалном кривом за појединачне променљиве*
Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Прикази коришћеног метода визуелизације сугеришу присуство израженог одступања облика распореда анализираних променљивих од нормалног распореда, како у погледу (а)симетричности, тако и заобљености (издужености) распореда, будући да се све

четири променљиве одликују позитивно асиметричним и издуженим распоредом. Резултати тестирања хипотеза о униваријационој нормалности распореда, спроведеног уз ризик грешке $\alpha = 0{,}05$, квантитативно потврђују претходне, визуелно изведене, закључке. Наиме, полазећи од израчунатих вредности статистика коришћених тестова, за сваку од разматраних променљивих појединачно, може се закључити, уз дефинисани ниво ризика грешке, да постоји довољно емпиријских доказа за одбацивање кореспондентне нулте хипотезе [128], будући да су резултирајуће *p*-вредности значајно мање од $\alpha = 0{,}05$. Такође, изразито великим и позитивним вредностима статистика $Z_{skewness}$ и $Z_{kurtosis}$ потврђена је позитивна асиметрија и издуженост распореда (у односу на висину нормалног распореда), анализом обухваћених, променљивих.

Табела 5.1.2. *Резултати тестирања униваријационе нормалности променљивих*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
X_1	16,124	0,000	0,583	0,000	0,218	0,000	22,733	H_1	62,147	H_1
X_2	31,681	0,000	0,295	0,000	0,329	0,000	45,786	H_1	230,746	H_1
X_3	0,966	0,015	0,974	0,004	0,072	0,035	2,685	H_1	-0,425	H_0
X_4	7,707	0,000	0,838	0,000	0,169	0,000	9,476	H_1	10,155	H_1

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*, *EduStat 4.05* и *QI Macros for Excel*.

На основу изолованих хистограма на десним крајевима графичких приказа распореда, може се претпоставити да потврђена одступања од нормалног распореда највероватније представљају последицу присуства једне или више нестандардних опсервација. Сходно наведеној претпоставци, на основу графичких приказа појединачних *box–plot* дијаграма (Слика 5.1.3), извршена је провера присуства униваријационих нетипичних вредности. Нестандардне вредности[129] (означене звездицама на *box-plot* дијаграму), идентификоване том приликом, на нивоу појединих оригиналних променљивих, рангиране по висини, су:

(X_1): Стари Град, Црна Трава, Савски венац, Врачар;

(X_2): Савски венац, Стари Град, Нови Београд, Црна Трава, Врачар, Медијана; и

(X_4): Сурчин, Нови Београд, Лазаревац, Стари Град, Лајковац.

Чињеница да су приказане нетипичне опсервације идентификоване код променљивих које се одликују најјачим доказима (велика вредност статистике теста и мала *p*-вредност) за одбацивање нулте хипотезе о нормалности, потврђује претходно изнету претпоставку.

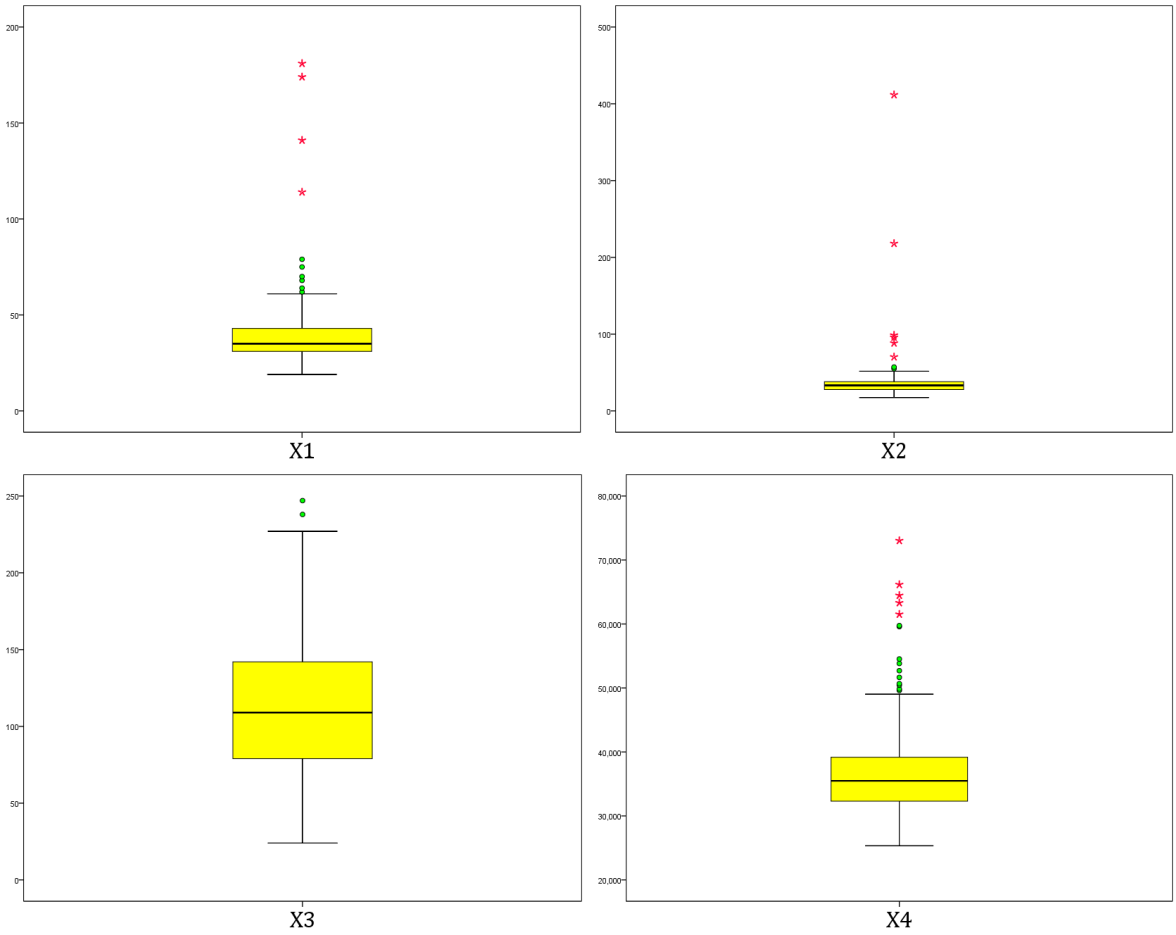
Полазећи од сагледаних карактеристика и природе анализом обухваћених променљивих, у циљу ублажавања утицаја идентификованих нетипичних опсервација и обезбеђивања испуњености претпоставке о униваријационој нормалности, као неопходног (али не и довољног) услова за постизање мултиваријационе нормалности *p*-димензионог распореда, извршена је *Box-Cox* трансформација оригиналних променљивих, коришћењем израза (1.4.1). [130] Ради провере ефеката примењене трансформације на испуњеност

[128] При примени *Anderson-Darling*-овог, *Shapiro-Wilk*-овог, *Kolmogorov-Smirnov*-ог теста нормалности, нулта хипотеза садржи претпоставку да анализирана променљива следи нормалан распоред, док се истом, у случају $Z_{skewness}$ и $Z_{kurtosis}$ тестова, сугерише да униваријациони распоред има нормалну симетричност, односно заобљеност, респективно.

[129] Издвојене нетипичне опсервације превазилазе граничну вредност утврђену на нивоу три интерквартилне разлике изнад трећег квартила, $Q_3+3*IQR$.

[130] Конверзија оригиналних променљивих (X_1, X_2, X_3, X_4) у њихове трансформисане супституенте ($T–X_1$, $T–X_2$, $T–X_3$, $T–X_4$) спроведена је на основу следећих, оптималних вредности трансформационог параметра (λ), утврђених за сваку променљиву појединачно: $(X_1) \rightarrow \lambda_1 = -0{,}88275$; $(X_2) \rightarrow \lambda_2 = -0{,}78666$; $(X_3) \rightarrow \lambda_3 = 0{,}44511$; и $(X_4) \rightarrow \lambda_4 = -2{,}03016$.

претпоставке о универијационој нормалности, на трансформисаним вредностима променљивих *T*-X_1, *T*-X_2, *T*-X_3, *T*-X_4, спроведени су претходно коришћени статистички тестови нормалности, уз ниво значајности теста $\alpha = 0{,}05$, а резултати су дати у Табели 5.1.3.



Слика 5.1.3. *Box-plot дијаграми оригиналних променљивих*

Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Табела 5.1.3. *Резултати тестирања универијационе нормалности распореда након извршене Box-Cox-ове трансформације*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
T-X_1	0,901	0,021	0,983	0,037	0,079	0,013	-0,000003	H_0	2,3703	H_1
T-X_2	2,643	0,000	0,953	0,000	0,092	0,002	-0,000005	H_0	4,2844	H_1
T-X_3	0,322	0,526	0,993	0,583	0,053	0,200	0,000005	H_0	-1,3215	H_0
T-X_4	0,353	0,462	0,992	0,492	0,047	0,200	0,000002	H_0	0,1730	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *EduStat 4.05*.

За разлику од променљивих *T*-X_3 и *T*-X_4, за које се, уз ризик грешке од 5%, закључује да не постоји довољно емпиријских доказа за одбацивање H_0 претпоставке о нормалности универијационог распореда (будући да су добијене *p*-вредности, за сваки од коришћених тестова, веће од $\alpha = 0{,}05$), у случају променљивих *T*-X_1 и *T*-X_2, при дефинисаном нивоу значајности теста, може се констатовати да тврдња о универијационој нормалности распореда није истинита, имајући у виду чињеницу да добијени резултати поступка тестирања, и након спроведене трансформације података, поново сугеришу одбацивање H_0, односно, усвајање алтернативне H_1 хипотезе.

Како трансформациони поступак није у довољној мери ублажио утицај свих нетипичних вредности, спроведен је поступак њиховог постепеног искључења из расположивог узорка намењеног даљој анализи[131]. Приоритет је додељен нестандарданим опсервацијама уоченим на нивоу променљиве X_2, будући да су код исте забележене најмање вредности реализованог нивоа значајности (Табела 5.1.3), односно најизраженије одступање од нормалности из угла облика распореда. Наведена променљива одликује се уједно и највећим бројем забележених нетипичних вредности (укупно шест), при чему је важно истаћи да је реч о општинама које су, такође, као нетипичне опсервације издвојене и на нивоу променљиве X_1, као друге променљиве код које није потврђена униваријациона нормалност распореда након трансформације. Сходно наведеном образложењу, из расположивог узорка мултиваријационих опсервација (величине, n = 165), искључене су, поменутим итеративним поступком, следеће општине (носиоци нетипичних вредности променљиве X_1 и X_2): *Савски венац*, *Стари Град*, *Нови Београд*, *Црна Трава*, *Врачар* и *Медијана*. Кључне карактеристике искључених општина из угла вредности променљивих и основне дескриптивне мере узорка, приказане су у Табели 5.1.4. Резултати тестирања статистичких хипотеза о униваријационој нормалности коришћењем оригиналних, а затим и трансформисаних вредности променљивих[132] у узорку (редуковане) величине, n = 159 општина, представљени су у Табели 5.1.5 и Табели 5.1.6, респективно.

Табела 5.1.4. *Карактеристике искључених шест општина и основне дескриптивне мере*

Променљиве / Искључене општине	Број предузећа на 1000 ст. (X_1)	Стопа запослености (X_2)	Број незапослених на 1000 ст. (X_3)	Просечна зарада по запосленом (X_4)
Савски венац	141	411,67	54	49,031
Стари Град	181	218,04	63	63,304
Нови Београд	79	98,63	52	66,154
Црна Трава	174	95,81	148	33,460
Врачар	114	88,31	55	59,591
Медијана	61	70,30	110	39,478
Дескриптивне статистичке мере (n = 165)				
Аритметичка средина ($\bar{x}$)	40,27	37,49	113,10	37,302
Минимална вредност (x_{min})	19	17,43	24	25,344
Максимална вредност (x_{max})	181	411,67	247	73,027
Q_3+3**IQR*	79	69,75	338	60,026

Извор: Ауторов прорачун

Табела 5.1.5. *Резултати тестирања униваријационе нормалности распореда оригиналних променљивих након искључења 6 (атипичних) општина*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
X_1	3,111	0,000	0,933	0,000	0,131	0,000	5,5132	H_1	3,6988	H_1
X_2	0,458	0,261	0,984	0,057	0,053	0,200	1,6524	H_0	0,8417	H_0
X_3	0,882	0,023	0,976	0,008	0,071	0,048	2,5842	H_1	-0,3681	H_0
X_4	6,132	0,000	0,850	0,000	0,154	0,000	9,5954	H_1	12,6255	H_1

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *EduStat 4.05*.

[131] Поменути итеративни поступак подразумева постепено искључење једне по једне нетипичне опсервације, редом, од најекстремније ка мање израженим, при чему се, након сваког (појединачног) искључења, врши тестирање хипотеза о нормалности униваријационог распореда, путем претходно поменутих тестова, како на оригиналним, тако и, уколико је забележено присуство нестандардних опсервација, на трансформисаним подацима, све до тренутка док се резултатима поступка тестирања не обезбеди потврда униваријационе нормалности распореда оригиналних и / или трансформисаних променљивих, обухваћених анализом.

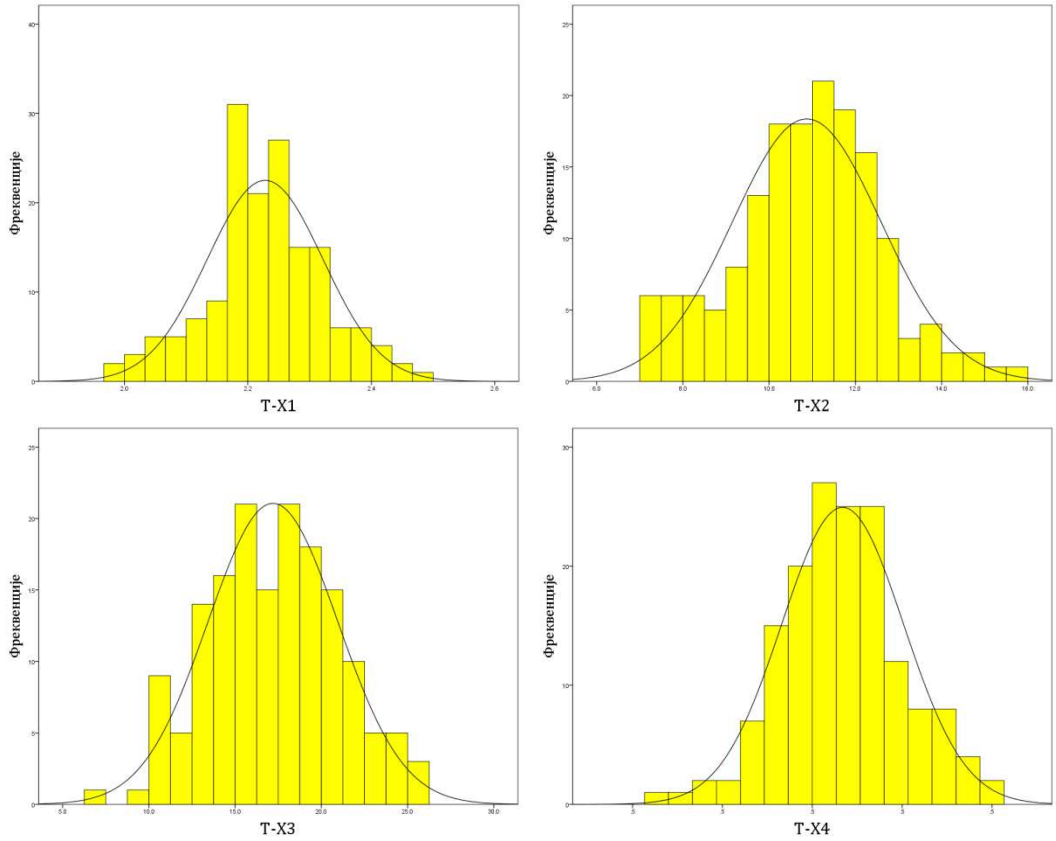
[132] Оптималне вредности трансформационог параметра (λ), на основу којих је спроведена *Box-Cox*-ова трансформација појединачних променљивих, износе: (X_1) → λ_1 = – 0,2885; (X_2) → λ_2 = 0,5662; (X_3) → λ_3 = 0,4697; и (X_4) → λ_4 = – 1,917.

Табела 5.1.6. *Резултати тестирања универијационе нормалности распореда трансформисаних променљивих након искључења 6 (атипичних) општина*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
T-X_1	0,572	0,136	0,990	0,350	0,064	0,200	0,000	H_0	0,412	H_0
T-X_2	0,522	0,182	0,988	0,186	0,043	0,200	0,000	H_0	0,106	H_0
T-X_3	0,259	0,711	0,994	0,751	0,051	0,200	0,000	H_0	-1,125	H_0
T-X_4	0,276	0,653	0,994	0,796	0,048	0,200	0,000	H_0	0,358	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *EduStat 4.05*.

На основу резултата представљених у Табели 5.1.6, може се закључити да, у случају све четири (трансформисане) променљиве, не постоји довољно емпиријских доказа за одбацивање претпоставке о универијационој нормалности, будући да су добијене *p*-вредности, на нивоу коришћених тестова, значајно веће од *a priori* дефинисаног нивоа значајности теста, α = 0,05, чиме је потврђена испуњеност испитиване претпоставке (Слика 5.1.4).

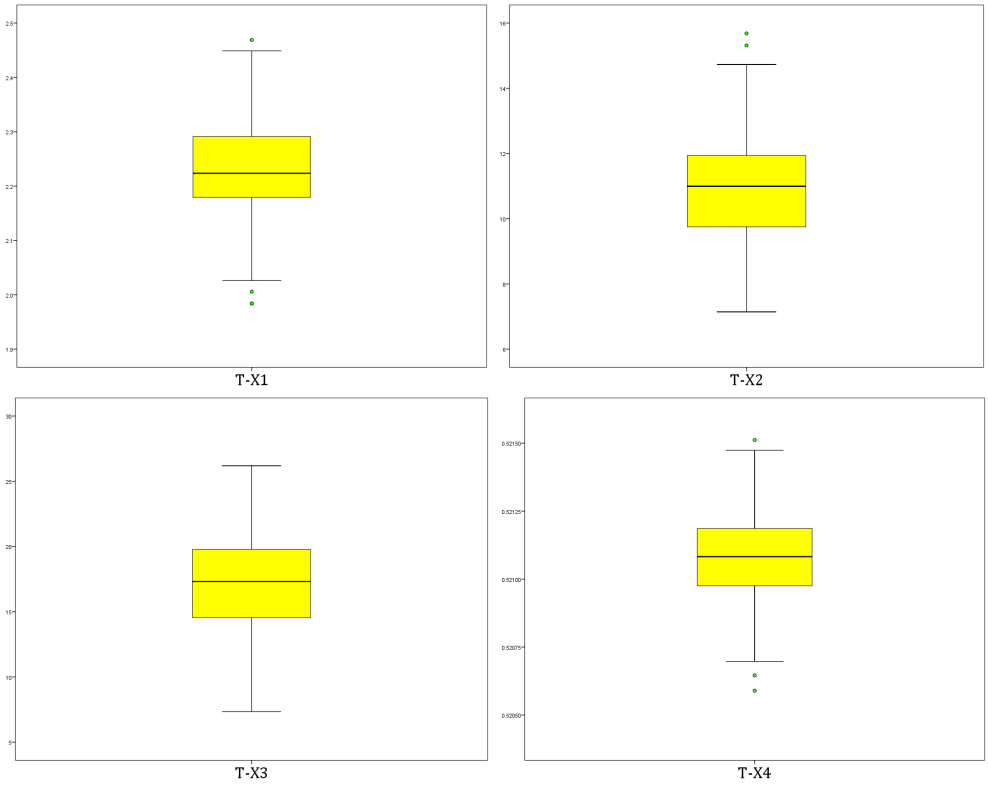


Слика 5.1.4. *Хистограми фреквенција за трансформисане променљиве*
Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Такође, коришћењем графичких приказа у форми појединачних *box-plot* дијаграма, није установљено присуство (правих) универијационих нестандарданих вредности на нивоу трансформисаних променљивих (Слика 5.1.5).

Будући да је остварена универијациона нормалност за сва четири показатеља економске развијености општина, у наредном кораку, коришћењем трансформисаних променљивих, извршено је испитивање испуњености претпоставке о нормалности мултиваријационог распореда, путем *Mardia тестова симетричности и заобљености*

мултиваријационог распореда [133] и *Henze–Zikler*–овог *теста мултиваријационе нормалности* [134], а резултати спроведеног поступка тестирања дати су у Табели 5.1.7.



Слика 5.1.5. *Box-plot дијаграми трансформисаних променљивих*
Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Табела 5.1.7. *Резултати тестирања статистичких хипотеза о нормалности мултиваријационог распореда трансформисаних променљивих*

Мултиваријациони тестови нормалности	Статистика теста	*p*-вредност	Одлука
Mardia тест симетричности	47,211	0,001	H_1
Mardia тест заобљености	2,123	0,034	H_1
Henze–Zirkler–ов тест	1,195	0,002	H_1

Извор: Ауторов прорачун коришћењем програма *SYSTAT 13.1*.

Представљени резултати тестирања недвосмислено сугеришу, уз ризик грешке α = 0,05, одбацивање тврдње о нормалности мултиваријационог распореда, односно усвајање алтернативне хипотезе према којој заједнички распоред анализираних (p) променљивих, на нивоу популације, није нормалан, будући да су добијене p-вредности, у случају сва три теста, далеко испод постављеног нивоа значајности теста, α = 0,05. Сходно изнетом закључку, извршено је провера (евентуалног) присуства мултиваријационих нетипичних опсервација, заснована на поређењу израчунатих вредности *Mahalanobis*-овог одстојања за сваку појединачну општину, као мере која приближно следи χ^2 распоред са p степени слободе (где је p број променљивих) и вредности 97,5 перцентила χ^2 распореда ($\chi^2_{p,\ 0{,}975}$), као критичне (граничне) вредности за идентификовање *outlier*-а. Реализацијом описане процедуре, у узорку сачињеном од 159 општина, идентификовано је присуство шест

[133] *Mardia* тест симетричности → H_0: мултиваријациони распоред (p променљивих) има нормалну симетричност;
Mardia тест заобљености → H_0: мултиваријациони распоред (p променљивих) има нормалну заобљеност;
[134] *Henze-Zikler*-ов тест мултиваријационе нормалности → H_0: мултиваријациони (p-димензиони) распоред је нормалан.

мултиваријационих нетипичних опсервација, односно општина, за које су вредности *Mahalanobis*-ове мере одстојања биле веће од утврђене критичне вредности, $\chi^2_{p,\ 0{,}975}$ = 11,14329, и то: *Жабари* [*MD* = 19,122], *Мало Црниће* [*MD* = 16,111], *Рача* [*MD* = 12,649], *Жагубица* [*MD* = 12,008], *Медвеђа* [*MD* = 11,279] и *Трговиште* [*MD* = 11,236].

У даљем току анализе, спроведен је итеративни поступак постепеног („корак-по-корак“) искључења (из узорка) појединачних општина, за које је установљено да представљају мултиваријационе нестандардне опсервације, заснован на реализацији следећих етапа:

- Избор најекстремније мултиваријационе нестандарадне опсервације (општине), мерено величином израчунатих вредности *Mahalanobis*-ове мере одстојања;
- Елиминисање изабране општине из узорка мултиваријационих опсервација;
- Тестирање статистичких хипотеза о универијационој нормалности, коришћењем оригиналних вредности променљивих у редукованом узорку, величине n–1;
- Испитивање (евентуалног) присуства универијационих нетипичних вредности;
- У случају нарушености претпоставке о универијационој нормалности, примена *Box-Cox*-ове трансформације вредности оригиналних променљивих;
- Тестирање статистичких хипотеза о универијационој нормалности, коришћењем трансформисаних вредности променљивих у редукованом узорку, величине n–1;
- Тестирање статистичких хипотеза о нормалности мултиваријационог распореда p трансформисаних, универијационо нормално распоређених, променљивих;
- Уколико резултати тестирања у претходној етапи сугеришу одбацивање H_0, врши се провера присуства мултиваријационих нетипичних опсервација, уз итеративно спровођење претходних етапа, до тренутка док се не осигура испуњеност претпоставки о мултиваријационој нормалности и одсуству мултиваријационих нетипичних опсервација.

Имплементацијом наведеног поступка, из расположивог узорка мултиваријационих опсервација (n = 159), искључено је следећих 14 општина (наведених према редоследу њиховог елиминисања), и то: *Жабари*, *Мало Црниће*, *Жагубица*, *Рача*, *Гроцка*, *Голубац*, *Трговиште*, *Медвеђа*, *Гаџин Хан*, *Раковица*, *Сремски Карловци*, *Сопот*, *Ариље*, и *Нови Сад*. Резултати тестирања хипотеза о универијационој нормалности оригиналних, а затим и трансформисаних променљивих[135], приказани су у Табели 5.1.8 и 5.1.9, респективно.

Табела 5.1.8. *Резултати тестирања универијационе нормалности распореда оригиналних променљивих након искључења додатних 14 (укупно 20) општина*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
X_1	2,169	0,000	0,952	0,000	0,117	0,000	4,0659	H_1	1,9076	H_0
X_2	0,349	0,472	0,988	0,249	0,045	0,200	0,3392	H_0	-0,1327	H_0
X_3	0,647	0,089	0,976	0,012	0,061	0,200	2,3402	H_1	-0,2359	H_0
X_4	6,503	0,000	0,828	0,000	0,158	0,000	9,7000	H_1	12,7458	H_1

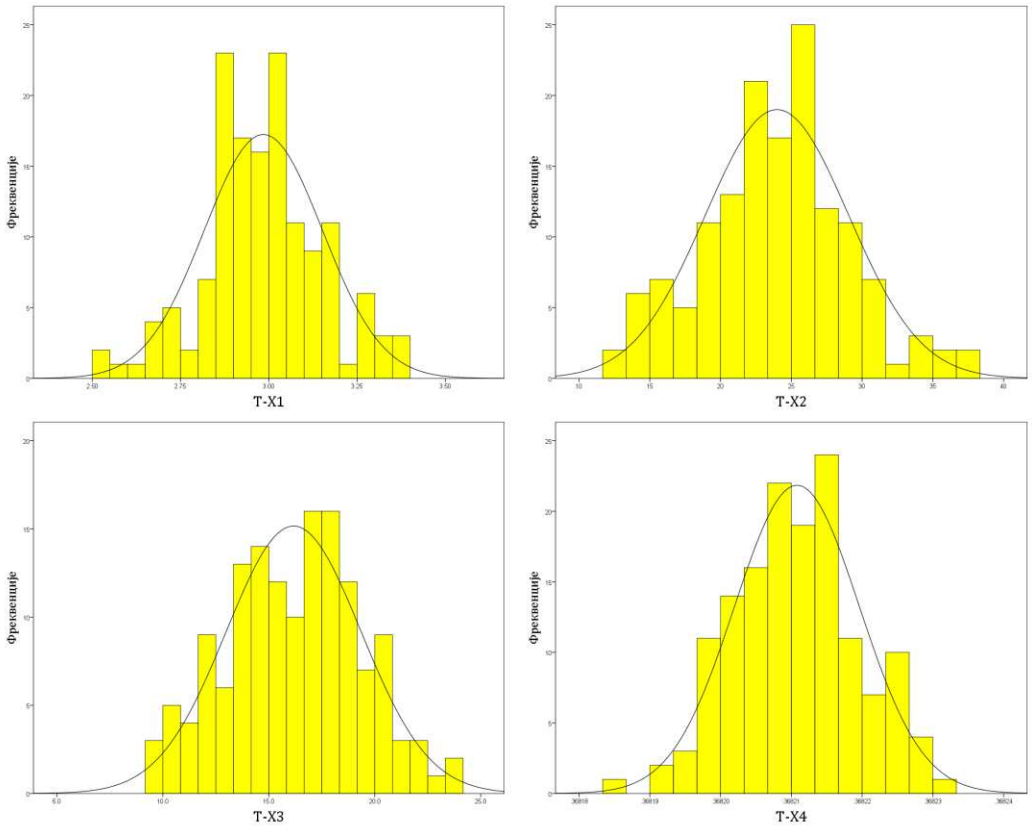
Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *EduStat 4.05*.

[135] Оптималне вредности трансформационог параметра (λ), на основу којих је спроведена *Box-Cox*-ова трансформација појединачних променљивих, износе: (X_1) → λ_1 = – 0,1018; (X_2) → λ_2 = 0,8933; (X_3) → λ_3 = 0,4469; и (X_4) → λ_4 = –2,7157.

Табела 5.1.9. *Резултати тестирања универијационе нормалности распореда трансформисаних променљивих након искључења додатних 14 (укупно 20) општина*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
T-X_1	0,591	0,121	0,988	0,244	0,066	0,200	0,000	H_0	0,543	H_0
T-X_2	0,388	0,383	0,988	0,253	0,044	0,200	0,000	H_0	-0,209	H_0
T-X_3	0,322	0,525	0,989	0,323	0,060	0,200	0,000	H_0	-1,286	H_0
T-X_4	0,251	0,737	0,993	0,667	0,052	0,200	0,000	H_0	-0,789	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *EduStat 4.05*.

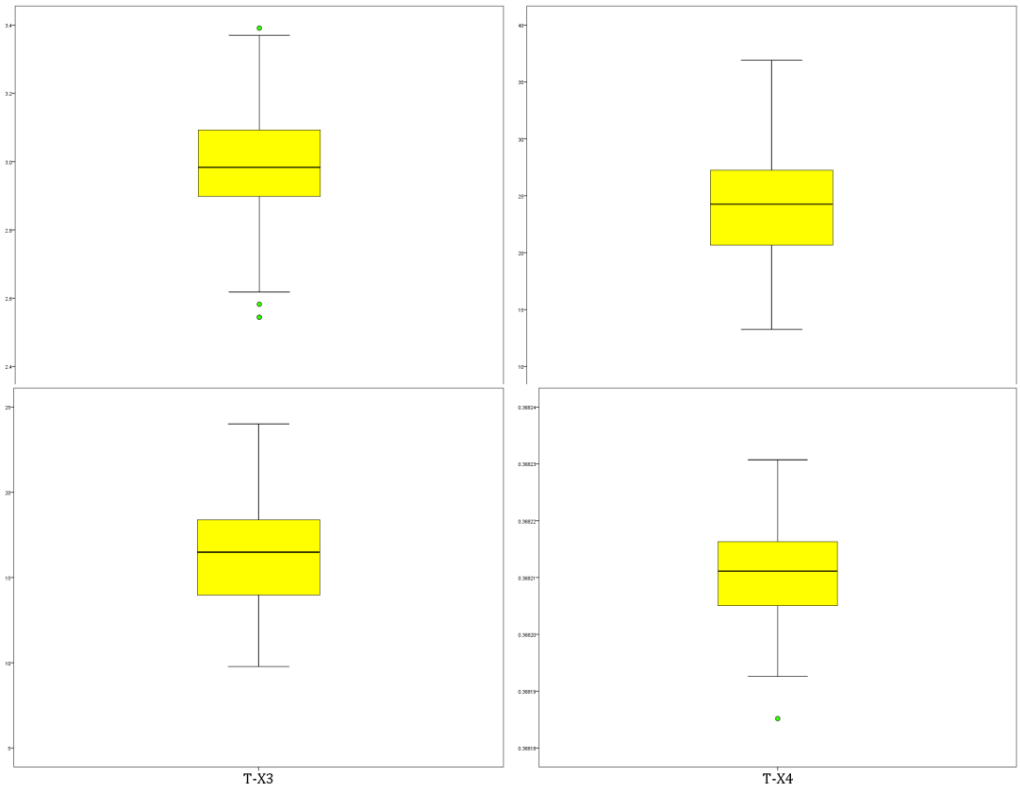


Слика 5.1.6. *Хистограми фреквенција за трансформисане променљиве*
Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

На Слици 5.1.6, представљен је графички приказ хистограма фреквенција са нормалном кривом појединачних трансформисаних променљивих, док су Сликом 5.1.7, приказани њима кореспондентни *box-plot* дијаграми, у циљу потврде одсуства универијационих нетипичних опсервација у „новоформираном" узорку од 145 општина[136].

Резултати поступка тестирања статистичких хипотеза о нормалности мултиваријационог распореда анализираних (трансформисаних) променљивих, представљени Табелом 5.1.10, недвосмислено сугеришу да, уз ризик грешке α = 0,05, не постоји довољно емпиријских доказа за одбацивање претпоставке о нормалности заједничког распореда анализираних *p* променљивих на нивоу популације, будући да су добијене *p*-вредности за сва три теста значајно веће од постављеног нивоа значајности теста, α. Наиме, уз дати ризик грешке, може се констатовати да је претпоставка о нормалности распореда мултиваријационих опсервација на нивоу популације, испуњена.

[136] Посматрано на нивоу трансформисаних вредности појединачних променљивих издвојено је присуство следећих умерених екстрема (енгл. *suspected outliers*), који немају значајан утицај на испуњеност претпоставке о нормалности, и то: Земун, Нова Црња, Чока и Житорађа (за променљиву *T-*X_1), односно, Бела Паланка (за променљиву *T-*X_4).



Слика 5.1.7. *Box-plot дијаграми трансформисаних променљивих*
Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Табела 5.1.10. *Резултати тестирања статистичких хипотеза о нормалности мултиваријационог распореда трансформисаних променљивих*

Мултиваријациони тестови нормалности	Статистика теста	*p*-вредност	Одлука
Mardia тест симетричности	27,066	0,133	H_0
Mardia тест заобљености	-1,315	0,189	H_0
Henze–Zirkler–ов тест	0,984	0,141	H_0

Извор: Ауторов прорачун коришћењем програма *SYSTAT 13.1*.

Такође, поређењем вредности *Mahalanobis*-ове мере одстојања на нивоу појединачних општина са кореспондентном критичном вредношћу χ^2 распореда ($\chi^2_{4,\ 0,975}$ = 11,1433), у узорку од 145 општина, нису уочене мултиваријационе нетипичне опсервације.

Након провере степена испуњености претпоставки које се односе на униваријациону и мултиваријациону нормалност, као и одсуство униваријационих и мултиваријационих нестандардних опсервација, у наставку текста, пажња је усмерена на испитивање кључних одлика међузависности анализираних променљивих, односно на провери претпоставки о присуству линеарне повезаности и одсуству мултиколинеарности. У том смислу, утврђене су вредности *Pearson*-ових коефицијената просте линеарне корелације између свих парова (трансформисаних) променљивих, након чега је извршено тестирање хипотеза о статистичкој значајности оцењених вредности коефицијената корелације, уз ризик грешке α = 0,05. Будући да се наведени поступак тестирања заснива на вредности статистике параметарског теста, чија примена претпоставља нормалност заједничког распореда парова променљивих, у Табели 5.1.11, дати су резултати тестирања хипотеза о нормалности дводимензионих распореда. Коначни резултати просте линеарне корелационе анализе представљени су у форми корелационе матрице (Табела 5.1.12).

Табела 5.1.11. *Резултати тестирања статистичких хипотеза о нормалности дводимензионог распореда парова трансформисаних променљивих*

Парови променљивих	*Mardia* тест симетричности		*Mardia* тест заобљености		*Henze-Zirkler*-ов тест нормалности		Одлука
	статистика	*p*-вредност	статистика	*p*-вредност	статистика	*p*-вредност	
T-X_1 и T-X_2	0,063	1,000	-0,708	0,479	0,742	0,463	H_0
T-X_1 и T-X_3	2,918	0,572	-0,959	0,337	0,712	0,540	H_0
T-X_1 и T-X_4	0,524	0,971	-0,963	0,336	0,414	0,313	H_0
T-X_2 и T-X_3	10,492	0,033	-0,463	0,644	1,097	0,057	H_0
T-X_2 и T-X_4	5,834	0,212	-1,281	0,200	0,636	0,785	H_0
T-X_3 и T-X_4	4,581	0,333	-1,535	0,125	0,636	0,785	H_0

Извор: Ауторов прорачун коришћењем програма *SYSTAT 13.1*.

Табела 5.1.12. *Корелациона матрица за трансформисане променљиве*

Променљиве	T-X_1	T-X_2	T-X_3	T-X_4
T-X_1	1,000	0,475 [**]	–0,405 [**]	0,233 [**]
T-X_2	0,475 [**]	1,000	–0,541 [**]	0,446 [**]
T-X_3	–0,405 [**]	–0,541 [**]	1,000	–0,353 [**]
T-X_4	0,233 [**]	0,446 [**]	–0,353 [**]	1,000

Напомена: Симбол [**] означава статистичку значајност израчунатих оцена уз ризик грешке, α=0,05

Извор: Ауторов прорачун

Израчунате вредности коефицијената просте линеарне корелације (*r*), као и резултати тестирања хипотеза о њиховој статистичкој значајности, сугеришу да између свих парова (трансформисаних вредности) показатеља економске развијености постоји статистички значајна линеарна веза на нивоу популације, чиме је потврђена испуњеност претпоставке о линеарности. Такође, доминантно је присуство директне корелационе везе на нивоу већине анализираних парова променљивих. Изузетак представља променљива *Број незапослених на 1000 ст.* (означена симболом *T-X*$_3$), која је негативно корелисана са свим преосталим показатељима. У представљеној корелационој матрици нису евидентиране вредности коефицијената веће од 0,80 или 0,90 будући да се исте, у апсолутном износу, крећу у интервалу од $|r_{min}| = 0{,}233$ до $|r_{max}| = 0{,}541$. Сходно наведеном, може се констатовати да у изабраном „скупу“ показатеља нема високо корелисаних променљивих, чиме је уједно и потврђена испуњеност претпоставке о одсуству мултиколинеарности.

Следећи важан корак у оцени степена међусобне повезаности анализираних променљивих укључује интерпретацију вредности *KMO* мере адекватности, израчунате како на нивоу комплетне корелационе матрице тако и појединачних променљивих, као и примену *Bartlett*-овог теста сферичности. Конкретно, израчуната путем израза (2.1.4), вредност *KMO* мере адекватности узорковања износи 0,732 и сугерише просечан ниво адекватности извршеног избора четири конкретне променљиве, посматрано из угла (укупне) „јачине“ присутне корелације између њих. Вредности овог показатеља на нивоу појединачних показатеља економске развијености, утврђене коришћењем израза (2.1.5), представљене су у Табели 5.1.13. Међусобно приближне величине и евидентно изнад минимално прихватљивог нивоа од 0,50, добијене вредности указују да је корелисаност сваке појединачне променљиве са преосталим индикатор-променљивим, генерално, на просечном, задовољавајућем нивоу, чиме је додатно потврђена оправданост њиховог избора, као улазних компоненти у имплементацији факторске анализе.

Табела 5.1.13. *Вредности КМО мере адекватности појединачних променљивих*

Индикатор-променљиве	Ознака	$KMO\text{-}MSA_j$ (за j=1,...,4)
Број предузећа на 1000 ст.	T-X_1	0,76009
Стопа запослености	T-X_2	0,68656
Број незапослених на 1000 ст.	T-X_3	0,75147
Просечна зарада по запосленом	T-X_4	0,76453

Извор: Ауторов прорачун

За тестирање истинитости тврдње (садржане у нултој хипотези) према којој узорачка корелациона матрица (Табела 5.1.12) потиче из скупа у којем анализиране променљиве нису међусобно статистички значајно корелисане (односно, H_0: $|\mathbf{R}|=|\mathbf{I}|=1$), коришћењем израза (2.1.6), израчуната је вредност статистике *Bartlett*-овог теста сферичности, χ^2 = 125,788. Будући да је реализовани ниво значајности теста (p-вредност < 0,00001), утврђен за добијену вредност статистике теста и број степени слободе χ^2 распореда, v = 6, мањи од постављеног нивоа значајности теста (α=0,05), може се закључити да постоје довољно јаки докази за одбацивање H_0 и, консеквентно, прихватање алтернативне хипотезе, којом се тврди да се корелациона матрица **R**, на нивоу популације из које је узорак извучен, статистички значајно разликује од јединичне матрице **I**. Наведеном одлуком додатно је потврђена оправданост примене факторске анализе на елементима разматране матрице.

Коначно, имајући у виду чињеницу да величина узорка у значајној мери може утицати на прецизност статистичких поступака намењених оцењивању непознатих параметара модела ФА, неопходно је нагласити да (поступком провере статистичких претпоставки редукована) величина узорка у потпуности задовољава искуствене препоруке које се тичу обезбеђивања „пожељног“ и / или прихватљивог броја опсервација у примени ФА. Наиме, расположивим узорком обухваћено је више од 100 опсервација (n = 145), док је број општина по свакој променљивој већи од 30, односно n / p = 145 / 4 ≈ [36 : 1] > [30 : 1].

Након провере испуњености претпоставки за валидну примену факторске анализе, а имајући у виду чињеницу да се анализиране променљиве исказују у различитим мерним јединицама, у оквиру последњег корака припреме података, спроведен је поступак нормализације њихових вредности. Нормализација вредности анализом обухваћених појединачних показатеља економске развијености извршена је коришћењем метода *min-max* трансформације (израз (1.4.2) и (1.4.3)), чиме се њихове трансформисане вредности (T-X_1, T-X_2, T-X_3, T-X_4) конвертују у одговарајуће, нормализоване вредности (X_1', X_2', X_3', X_4') на скали од 0 до 1. У циљу прецизнијег сагледавања нормализованих вредности показатеља, њихове компарације по општинама, као и олакшавања интерпретације добијених резултата анализе и вредности композитног показатеља који ће бити развијен, извршено је проширење опсега у којем се нормализоване вредности крећу, превођењем скале од 0 до 1 на скалу вредности од 1 до 10. При спровођењу поступка нормализације, уз уважавање смера корелационе везе између посматраних променљивих, инверзно кодирање (израз (1.4.3)) извршено је само за променљиву *Број незапослених на 1000 становника*, пошто већа вредност овог показатеља имплицира мањи степен економске развијености конкретне територијалне јединице, и обратно.

5.1.4. Развој композитног модела за мерење степена економске развијености општина у Републици Србији

На основу презентираних резултата провере испуњености разматраних претпоставки и констатоване адекватности расположивог узорка мултиваријационих опсервација из угла обезбеђивања статистички валидне имплементације факторске анализе, у овом Одељку, представљен је поступак развоја композитног индекса за мерење степена економске развијености општина у Републици Србији, заснован на оцењеном моделу ФА.

Примена ФА у наведеном контексту, заснива се на претпоставци да у основи међусобне квантитативне повезаности издвојених појединачних показатеља економске развијености анализираних општина постоји, најмање једна, неопсервабилна димензија, која се може сматрати главним „узроком“ и / или „најодговорнијим“ фактором присутних корелација између њих. Изведени модел ФА представља основу за развој композитног показатеља којим се обезбеђује индиректно „мерење“ латентне променљиве од интереса (односно, степена економске развијености општина). Прецизније, елементи факторског модела употребљени су при одређивању вредности пондера који су додељени појединачним индикаторима у структури предложеног композитног индекса.

Сходно наведеном, коришћењем података корелационе матрице као улазних елемената, извршено је оцењивање непознатих параметара модела експлоративне факторске анализе, применом методе главних компонената, сагласно поступку изложеном у Одељку 2.1.3. Иницијална форма оцењеног (p-димензионог) модела ФА, изведена коришћењем израза (2.1.17) – (2.1.19), гласи:

$$\begin{aligned} X_1' &= (0{,}700)\, F_1 + (-0{,}546)\, F_2 + (0{,}421)\, F_3 + (0{,}184)\, F_4 \\ X_2' &= (0{,}843)\, F_1 + (-0{,}001)\, F_2 + (-0{,}041)\, F_3 + (-0{,}536)\, F_4 \\ X_3' &= (0{,}782)\, F_1 + (-0{,}082)\, F_2 + (-0{,}559)\, F_3 + (0{,}264)\, F_4 \\ X_4' &= (0{,}656)\, F_1 + (0{,}682)\, F_2 + (0{,}270)\, F_3 + (0{,}178)\, F_4 \end{aligned} \qquad (5.1.1)$$

Вредности карактеристичних корена корелационе матрице, λ_f (за f = 1, 2, 3, 4), у основи представљеног иницијалног модела ФА, и њима кореспондентне вредности објашњеног удела укупне варијансе индикатор-променљивих на нивоу узорка, сваким појединачним (заједничким) фактором F_f, приказане су у Табели 5.1.14.

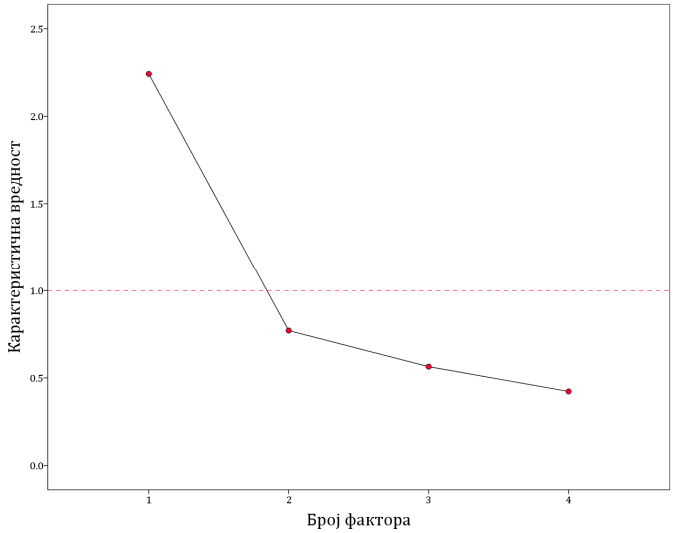
Табела 5.1.14. *Резултати спроведеног поступка издвајања заједничких фактора*

Заједнички фактори (F_f)	Карактеристичне вредности (λ_f)	Објашњени удео укупног иницијалног варијабилитета (у %)	Кумулатив пропорције објашњеног дела укупног варијабилитета у узорку
F_1	$\lambda_1 = 2{,}242$	56,053	56,053
F_2	$\lambda_2 = 0{,}771$	19,267	75,320
F_3	$\lambda_3 = 0{,}564$	14,110	89,430
F_4	$\lambda_4 = 0{,}423$	10,570	100,000

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.

Будући да постоји само један карактеристични корен, конкретно λ_1, чија је вредност већа од трага анализиране корелационе матрице, односно просека четири карактеристичне вредности ($\bar{\lambda}$=1), према *Kaiser-Guttman*-овом правилу, у циљу редефинисања иницијално развијеног модела (израз 5.1.1) и редукцију његове полазне димензионалности, донета је одлука о избору само једног, првог, заједничког фактора (F_1) и његовом задржавању у

редукованом моделу ФА. Избор једнофакторског решења као „оптималног“ из угла броја издвојених фактора, потврђен је и критеријумом заснованим на *Cattell*-овом дијаграму превоја (Слика 5.1.8), али и критеријумом објашњеног удела укупне узорачке варијансе, будући да је издвојеним фактором објашњено више од половине укупног варијабилитета анализираних променљивих на нивоу узорка (односно, ≈ 56%), док је кореспондентни износ код преосталих фактора знатно мањи (F_2→ ≈ 20%, F_3→ ≈ 14% и F_4→ ≈ 10%).



Слика 5.1.8. *Cattell-ов дијаграм „превоја“*

Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Оцењене вредности параметара (редукованог) издвојеног једнофакторског модела (израз 5.1.2) представљене су у Табели 5.1.14а.

$$\begin{aligned} X_1' &= 0{,}700 \times F_1 + e_1 \\ X_2' &= 0{,}843 \times F_1 + e_2 \\ X_3' &= 0{,}782 \times F_1 + e_3 \\ X_4' &= 0{,}656 \times F_1 + e_4 \end{aligned} \qquad (5.1.2)$$

Табела 5.1.14а. *Оцењене вредности параметара* (*редукованог*) *факторског модела*[137]

Променљиве	Иницијални варијабилитет (s_j^2)	Оцена факторских оптерећења	Оцењене вредности комуналитета	Удео објашњене (појединачне) варијансе заједничким фактором	Оцена специфичне варијансе
X_1'	1,000	0,700	0,490	49,10	0,510
X_2'	1,000	0,843	0,711	71,10	0,289
X_3'	1,000	0,782	0,611	61,10	0,389
X_4'	1,000	0,656	0,430	43,00	0,570
∑	4,000	/	2,242 = λ_1	≈ 56,053 %	1,758

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.

Будући да се израчунате вредности оцена факторских оптерећења крећу у интервалу од 0,656 до 0,843, указујући на присуство изражене (у случају променљиве X_4' – просечна зарада по запосленом) или (у случају осталих променљивих) јаке линеарне корелације између варијација појединачних променљивих и издвојеног заједничког фактора (F_1), исте

[137] У Табели 5.1.14а, представљене су оцењене вредности неротираних факторских оптерећења, будући да је на основу критеријума за избор „оптималног“ броја фактора издвојен само један заједнички фактор, услед чега није постојала потреба за спровођењем неког од метода ротације кореспондентне матрице.

се могу сматрати практично сигнификантним, из угла обезбеђивања адекватне апроксимације међусобне повезаности одабраних показатеља економске развијености општина, односно анализиране корелационе структуре. Такође, с обзиром да је величина коришћеног узорка n = 145,[138] може се констатовати да су израчунате оцене факторских оптерећења уједно и статистички значајне. Поређењем оцењених вредности заједничке и специфичне варијансе на нивоу појединачних променљивих, приметно је да променљиве *стопа запослености* и *број незапослених на 1000 становника*, представљају, у начелу, боље и поузданије мере издвојеног заједничког фактора, будући да је истим објашњено приближно 71%, односно 61% њиховог иницијалног варијабилитета, респективно. Удео варијабилитета који преостале две променљиве „деле“ са осталим показатељима је евидентно мањи, што се може уочити и на основу величине оцена њихових специфичних варијанси, али ипак на прихватљивом нивоу (односно ≈ 50%).

Заједничком фактору (F_1), као доминантно израженом у креираној факторској структури, додељен је назив *степен економске развијености*, будући да све четири нумеричке променљиве представљају појединачне показатеље неког од различитих, али међусобно повезаних, аспеката економске развијености општина.

Посматрајући издвојени фактор као неопсервабилни, мултидимензиони феномен од интереса, а индикатор-променљиве као „помоћна средства“ неопходна за индиректно „мерење“ његовог нивоа, коришћењем израза (2.1.17), креиран је композитни показатељ степена економске развијености општина – *Индекс економске развијености* (акроним, *ИЕР*). Одређивање пондера (w_j), додељених појединачним променљивим, засновано је на анализи структуре удела укупне узорачке варијансе који је објашњен издвојеним једнофакторским решењем. Заправо, релативни значај појединачних променљивих утврђен је израчунавањем релативног учешћа појединачних оцењених вредности комуналитета у укупном објашњеном уделу узорачке варијансе, односно коришћењем израза (2.1.18). Симболички, креирани композитни индекс има следећи облик:

$$ИЕР_{(i)} = \sum_{j=1}^{4} w_j x'_{ij} = 0,219 \times x'_{i1} + 0,317 \times x'_{i2} + 0,272 \times x'_{i3} + 0,192 \times x'_{i4}, \text{ за } i = 1, 2, ..., n. \qquad (5.1.3)$$

Очекивано, највећи релативни значај при израчунавању агрегатне вредности издвојеног заједничког фактора додељен је променљивој X_2 – *стопа запослености*, обзиром да се иста одликује и највећим факторским оптерећењем, након чега следе *број незапослених на 1000 становника* (0,272), *број предузећа на 1000 становника* (0,219) и *просечна зарада по запосленом* (0,192). Такође, будући да је композитни показатељ развијен коришћењем редукованог узорка величине n = 145, примена истог за потребе мерења нивоа економске развијености свих 165 општина (укључујући и општине које су током поступка припреме података искључене из узорка), захтева поновно спровођење *min-max* трансформације вредности четири променљиве на полазном узорку (n = 165), а у циљу обезбеђивања упоредивости њихових нормализованих вредности, x_{ij}'.

[138] Узорак величине 150 јединица посматрања углавном се сматра довољним за обезбеђивање статистичке значајности Уколико израчунате вредности оцена факторских оптерећења износе 0,45, да би се исте могле сматрати статистички значајним, неопходно је располагати узорком величине, $n \approx 150$ (*Hair et al.*, 2010, стр. 116).

Дескриптивне статистичке мере израчунатих вредности композитног индекса ИЕР

Вредности *Индекса економске развијености* (*ИЕР*), израчунате за сваку од 165 јединица локалне самоуправе у Републици Србији, представљене су у Табели 5.1.15. Вредности *ИЕР*, сагласно извршеној нормализацији појединачних показатеља, крећу се у интервалу од 1 до 10, при чему веће вредности *ИЕР* указују на виши ниво економске развијености и обратно. Кључне карактеристике распореда вредности *ИЕР*, илустроване су на Слици 5.1.9 и представљене у Табели 5.1.16.

Табела 5.1.15. *Преглед општина / градова према израчунатим вредностима ИЕР*

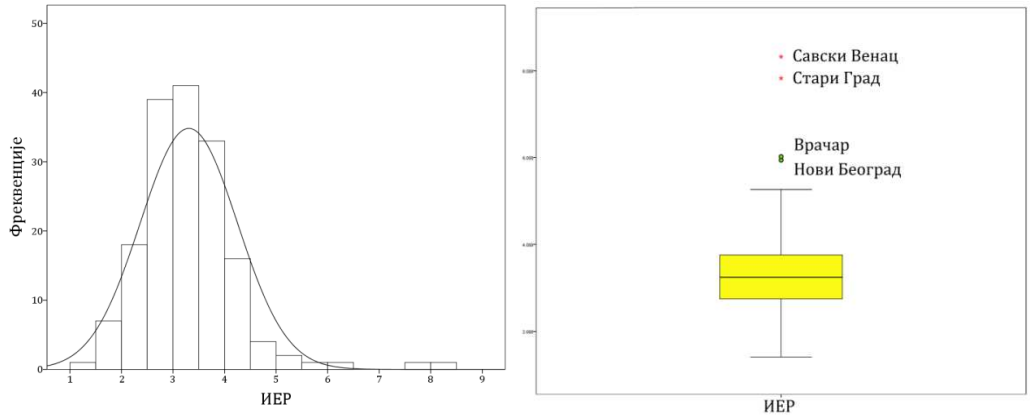
Интервали вредности *ИЕР*	Број ЈЛС	Назив јединице локалне самоуправе [вредност *ИЕР*]			
од 9 до 9,99	0	/			
од 8 до 8,99	1	Савски Венац [8,31]			
од 7 до 7,99	1	Стари Град [7,82]			
од 6 до 6,99	1	Врачар [6,02]			
од 5 до 5,99	3	Нови Београд [5,94]	Сурчин [5,26]	Палилула [5,00]	
од 4 до 4,99	20	Лазаревац [4,89]	Вождовац [4,37]	Пећинци [4,22]	Панчево [4,16]
		Црна Трава [4,83]	Звездара [4,36]	Сента [4,20]	Суботица [4,11]
		Земун [4,77]	град Пожаревац [4,34]	град Ужице [4,20]	Беочин [4,08]
		град Нови Сад [4,71]	Стара Пазова [4,23]	Чајетина [4,19]	Жагубица [4,03]
		Лајковац [4,46]	Косјерић [4,22]	Чукарица [4,18]	Ада [4,00]
од 3 до 3,99	74	Зрењанин [3,97]	Ариље [3,72]	Деспотовац [3,48]	Аранђеловац [3,24]
		Мионица [3,95]	Кањижа [3,71]	Бачка Топола [3,48]	Ковин [3,23]
		Обреновац [3,91]	Бач. Паланка [3,69]	Варварин [3,46]	Бечеј [3,22]
		Медијана [3,91]	С. Митровица [3,66]	Сокобања [3,44]	Владимирци [3,22]
		Г. Милановац [3,90]	Неготин [3,66]	Уб [3,42]	Пирот [3,20]
		Кучево [3,89]	Свилајнац [3,62]	Кикинда [3,38]	Бољевац [3,18]
		Гроцка [3,88]	Бајина Башта [3,56]	Краљево [3,36]	Крушевац [3,17]
		Инђија [3,87]	Лучани [3,56]	Крагујевац [3,35]	град Врање [3,15]
		Сопот [3,85]	В. Градиште [3,54]	Смед. Паланка [3,34]	Оџаци [3,13]
		Вршац [3,80]	Кладово [3,53]	Шид [3,33]	Кнић [3,13]
		Петр. на Млави [3,79]	Бор [3,52]	Александровац [3,32]	В. Бања [3,12]
		Ваљево [3,79]	Шабац [3,51]	Апатин [3,32]	Ражањ [3,12]
		Бач. Петровац [3,78]	Барајево [3,50]	Топола [3,32]	Опово [3,08]
		С. Карловци [3,77]	Темерин [3,50]	Црвени крст [3,31]	Тител [3,07]
		Љиг [3,76]	Мајданпек [3,50]	Н. Кнежевац [3,30]	Зајечар [3,07]
		Раковица [3,76]	Сомбор [3,50]	Рача [3,28]	Ћуприја [3,03]
		Пожега [3,75]	Жабари [3,50]	Рума [3,26]	Рашка [3,00]
		Чачак [3,74]	Мало Црниће [3,50]	Осечина [3,25]	
		Смедерево [3,74]	Велика Плана [3,50]	Голубац [3,24]	
од 2 до 2,99	57	Трстеник [2,99]	Жабаљ [2,80]	Алексинац [2,67]	Владич. Хан [2,35]
		Кула [2,97]	Ириг [2,80]	Сурдулица [2,66]	Мерошина [2,33]
		Бач [2,96]	Нови Бечеј [2,79]	Пријепоље [2,66]	Нова Црња [2,26]
		Лапово [2,91]	Лесковац [2,78]	Палилула [2,64]	М. Зворник [2,23]
		Ивањица [2,91]	Богатић [2,77]	Житиште [2,61]	Нишка Бања [2,23]
		Коцељева [2,90]	Параћин [2,77]	Пландиште [2,59]	Власотинце [2,19]
		Младеновац [2,89]	Брус [2,76]	Блаце [2,59]	Мали Иђош [2,15]
		Врбас [2,88]	Дољевац [2,75]	Крупањ [2,58]	Босилеград [2,10]
		Чока [2,88]	Ћићевац [2,75]	Гаџин Хан [2,58]	Рековац [2,10]
		Пантелеј [2,87]	Ковачица [2,74]	Сечањ [2,50]	Баточина [2,09]
		Бабушница [2,86]	Јагодина [2,74]	Куршумлија [2,44]	Бела Паланка [2,08]
		Књажевац [2,85]	Нова Варош [2,74]	Нови Пазар [2,42]	Сјеница [2,03]
		Бујановац [2,85]	Србобран [2,70]	Прешево [2,42]	
		Љубовија [2,82]	Лозница [2,70]	Алибунар [2,40]	
		Сврљиг [2,81]	Димитровград [2,68]	Прокупље [2,39]	
од 1 до 1,99	8	Прибој [1,99]	Житорађа [1,94]	Трговиште [1,86]	Бојник [1,59]
		Бела Црква [1,95]	Медвеђа [1,93]	Тутин [1,61]	Лебане [1,41]

Извор: Ауторов прорачун и систематизација табеларног приказа

Табела 5.1.16. *Дескриптивне статистичке мере вредности ИЕР за n=165*

Аритметичка средина	Медијана	Модални интервал	Минимална вредност	Максимална вредност	Просечно одступање	Коефцијент варијације
3,31	3,25	3,00–3,99	1,41	8,31	0,95	28,70 %

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.



Слика 5.1.9. *Хистограм фреквенција и box-plot дијаграм вредности ИЕР за n=165*

Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Класификација градова / општина у Србији на основу израчунатих вредности ИЕР

Класификација јединица локалних самоуправа (градова / општина) у Републици Србији, према „измереном“ степену економске развијености спроведена је на основу поређења појединачних вредности *ИЕР* и њихове (адекватне) средње вредности на нивоу узорка (односно, такозване, републичке средње вредности). С тим у вези, иако се као основа за наведена поређења, у истраживањима сличног карактера, углавном користи аритметичка средина,[139] у овом случају изабрана је медијана, као прикладнија мера средње вредности за улогу упоредне основе, будући да је резултатима дескриптивне (графичке) анализе серије израчунатих вредности *ИЕР* потврђено присуство две нестандардне опсервације, које одговарају општинама Савски венац и Стари Град (Слика 5.1.9). Сходно наведеном, класификација ЈЛС-а према вредностима *ИЕР*, извршена је на бази сагледавања позиције појединачних градова / општина у односу на „републички медијални ниво“ економске развијености, а не, уобичајено коришћен ниво, „републичког просека“. Наведени приступ захтева исказивање апсолутних вредности $ИЕР_i$, то јест, утврђеног степена економске развијености појединачних ЈЛС-а, као процентуалног износа медијане (m_e=3,25), односно као % остварења републичког медијалног нивоа економске развијености (у ознаци $[\%m_e]_i$) коришћењем следећег израза:

$$ИЕР_i \rightarrow [\%m_e]_i = \frac{ИЕР_i}{m_e} \times 100 \text{, за } i = 1, 2, \ldots, n. \qquad (5.1.4)$$

Компаративном анализом серије конвертованих вредности ($\%m_e$) на нивоу појединачних *ЈЛС*-а, издвојена су следећа запажања дескриптивног карактера:

✓ На нивоу 83 градова / општина (≈ 50% укупног броја *ЈЛС*-а), „измерени“ степен економске развијености налази се испод (републичког) медијалног нивоа ($ИЕР_i < m_e$,

[139] Методолошки приступ, коришћен од стране Националне Агенције за Регионални Развој (НАРР), при класификацији региона и јединица локалне самоуправе (градова / општина) у Републици Србији према степену развијености заснива се на поређењу вредности конкретног индикатора на нивоу појединачних јединица посматрања у односу на вредност републичког просека (Влада РС, 2009а) и сагледавању истих као % остварења нивоа дефинисаног просеком.

односно, $[\%m_e]_i$ < 100%). Преостале 82 *ЈЛС*-а одликују се достигнутим степеном економске развијености који је изнад медијалног, као упоредне основе.

✓ Степен економске развијености на нивоу изнад 150% републичког медијалног нивоа евидентиран је код свега 7 градова / општина. Поред општина које су издвојене као нестандардне опсервације (Савски венац и Стари Град), овој категорији *ЈЛС*-а припадају и следеће београдске општине: Врачар, Нови Београд, Сурчин, Палилула и Лазаревац.

✓ Приближно 91% укупног броја *ЈЛС*-а чија је вредност *ИЕР* изнад медијалне, одликују се степеном економске развијености на нивоу од 100% до 150% републичке медијалне вредности (симболички, 150% m_e > $ИЕР_i \geq$ 100% m_e).

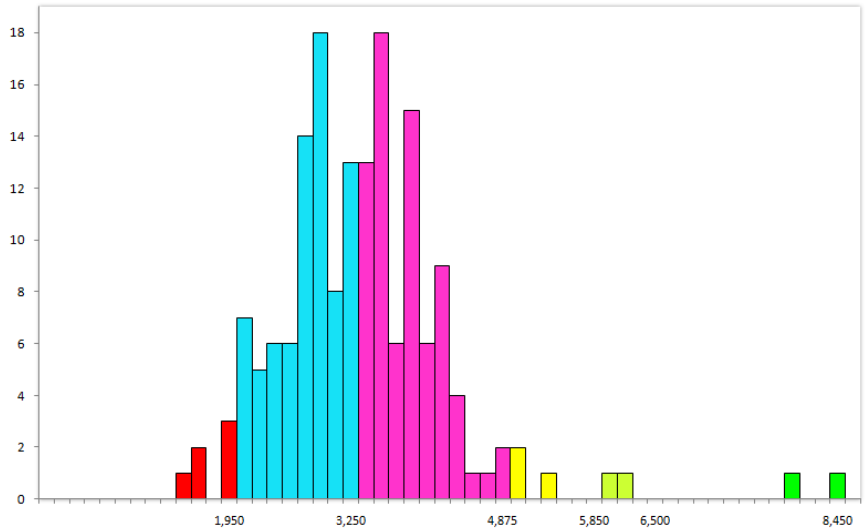
✓ Свега 6 општина (претежно општине у саставу управних округа унутар Региона Јужне и Источне Србије) има вредност *ИЕР* која је испод 60% медијане ($ИЕР_i$ < 60% m_e).

На темељу изнетих запажања извршено је формулисање конкретних критеријума за класификацију, као основе за разврставање *ЈЛС*-а у једну од издвојених шест категорија. Поменута класификациона правила дефинисана су у форми одговарајућих интервала вредности % остварења медијане (у ознаци, $[\%m_e]_i$). Граничне вредности ових интервала прецизиране су уважавањем појединих аспеката начина разврставања *ЈЛС*-а према степену развијености који је прописан Законом о регионалном развоју (*Влада РС*, 2009), као и одлика распореда вредности *ИЕР* (Слика 5.1.10). Предложена класификација *ЈЛС*-а у Републици Србији, према степену економске развијености, представљена је у Табели 5.1.17. Илустрација исте, заснована на приказу општина које се одликују најмањом и највећом вредношћу $[\% m_e]_i$ по свакој од издвојених група, дата је на Слици 5.1.11.

Табела 5.1.17. *Предложена ИЕР класификација градова / општина у Републици Србији*

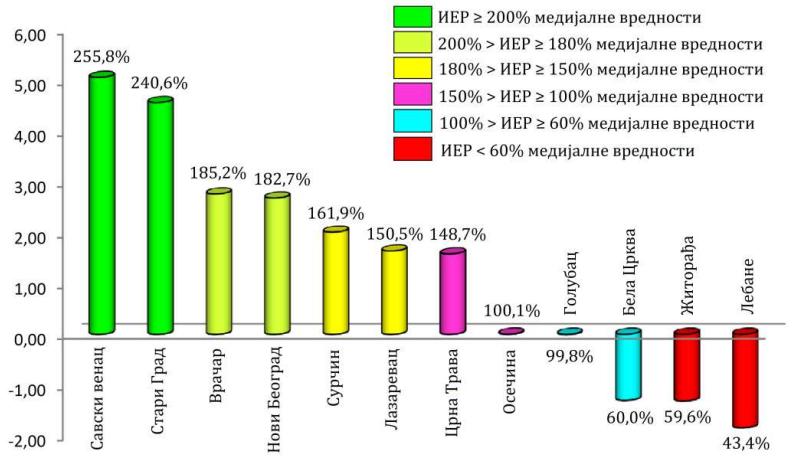
Категорија	Број ЈЛС-а	Критеријуми за класификацију	Интервалне границе група (у вредностима *ИЕР*)	Просек *ИЕР* по групи
Група 1	2	(% $m_e)_i \geq$ 200% m_e	> 6,500	8,07
Група 2	2	200% m_e > (% $m_e)_i \geq$ 180% m_e	[5,850 – 6,500)	5,98
Група 3	3	180% m_e > (% $m_e)_i \geq$ 150% m_e	[4,875 – 5,850)	5,05
Група 4	75	150% m_e > (% $m_e)_i \geq$ 100% m_e	[3,250 – 4,875)	3,77
Група 5	77	100% m_e > (% $m_e)_i \geq$ 60% m_e	[1,950 – 3,250)	2,72
Група 6	6	60% m_e > (% $m_e)_i$	< 1,950	1,72

Извор: Ауторов прорачун и систематизација табеларног приказа



Слика 5.1.10. *Хистограм фреквенција вредности ИЕР по издвојеним групама*

Извор: Ауторов визуелни приказ коришћењем програма *Excel*.



Слика 5.1.11. *Графички приказ општина са најмањом и највећом вредношћу* $[\% \, m_e]_i$ *унутар издвојених категорија ЈЛС-а*

Табела 5.1.18. *Преглед општина / градова по издвојеним класификационим категоријама*

Категорија	Број ЈЛС-а	Назив јединице локалне самоуправе [вредност *ИЕР*]			
Група 1	2	Савски Венац [8,31]	Стари Град [7,82]		
Група 2	2	Врачар [6,02]	Нови Београд [5,94]		
Група 3	3	Сурчин [5,26]	Палилула [5,00]	Лазаревац [4,89]	
Група 4	75	Црна Трава [4,83]	Зрењанин [3,97]	Ариље [3,72]	Деспотовац [3,48]
		Земун [4,77]	Мионица [3,95]	Кањижа [3,71]	Бачка Топола [3,48]
		град Нови Сад [4,71]	Обреновац [3,91]	Бач. Паланка [3,69]	Варварин [3,46]
		Лајковац [4,46]	Медијана [3,91]	С. Митровица [3,66]	Сокобања [3,44]
		Вождовац [4,37]	Г. Милановац [3,90]	Неготин [3,66]	Уб [3,42]
		Звездара [4,36]	Кучево [3,89]	Свилајнац [3,62]	Кикинда [3,38]
		град Пожаревац [4,34]	Гроцка [3,88]	Бајина Башта [3,56]	Краљево [3,36]
		Стара Пазова [4,23]	Инђија [3,87]	Лучани [3,56]	Крагујевац [3,35]
		Косјерић [4,22]	Сопот [3,85]	В. Градиште [3,54]	Смед. Паланка [3,34]
		Пећинци [4,22]	Вршац [3,80]	Кладово [3,53]	Шид [3,33]
		Сента [4,20]	Петр. на Млави [3,79]	Бор [3,52]	Александровац [3,32]
		град Ужице [4,20]	Ваљево [3,79]	Шабац [3,51]	Апатин [3,32]
		Чајетина [4,19]	Бач. Петровац [3,78]	Барајево [3,50]	Топола [3,32]
		Чукарица [4,18]	С. Карловци [3,77]	Темерин [3,50]	Црвени крст [3,31]
		Панчево [4,16]	Љиг [3,76]	Мајданпек [3,50]	Н. Кнежевац [3,30]
		Суботица [4,11]	Раковица [3,76]	Сомбор [3,50]	Рача [3,28]
		Беочин [4,08]	Пожега [3,75]	Жабари [3,50]	Рума [3,26]
		Жагубица [4,03]	Чачак [3,74]	Мало Црниће [3,50]	Осечина [3,25]
		Ада [4,00]	Смедерево [3,74]	Велика Плана [3,50]	
Група 5	77	Голубац [3,24]	Бач [2,96]	Дољевац [2,75]	Прешево [2,42]
		Аранђеловац [3,24]	Лапово [2,91]	Ћићевац [2,75]	Алибунар [2,40]
		Ковин [3,23]	Ивањица [2,91]	Ковачица [2,74]	Прокупље [2,39]
		Бечеј [3,22]	Коцељева [2,90]	Јагодина [2,74]	Владич. Хан [2,35]
		Владимирци [3,22]	Младеновац [2,89]	Нова Варош [2,74]	Мерошина [2,33]
		Пирот [3,20]	Врбас [2,88]	Србобран [2,70]	Нова Црња [2,26]
		Бољевац [3,18]	Чока [2,88]	Лозница [2,70]	М. Зворник [2,23]
		Крушевац [3,17]	Пантелеј [2,87]	Димитровград [2,68]	Нишка Бања [2,23]
		град Врање [3,15]	Бабушница [2,86]	Алексинац [2,67]	Власотинце [2,19]
		Оџаци [3,13]	Књажевац [2,85]	Сурдулица [2,66]	Мали Иђош [2,15]
		Кнић [3,13]	Бујановац [2,85]	Пријепоље [2,66]	Босилеград [2,10]
		В. Бања [3,12]	Љубовија [2,82]	Палилула [2,64]	Рековац [2,10]
		Ражањ [3,12]	Сврљиг [2,81]	Житиште [2,61]	Баточина [2,09]
		Опово [3,08]	Жабаљ [2,80]	Пландиште [2,59]	Бела Паланка [2,08]
		Тител [3,07]	Ириг [2,80]	Блаце [2,59]	Сјеница [2,03]
		Зајечар [3,07]	Нови Бечеј [2,79]	Крупањ [2,58]	Прибој [1,99]
		Ћуприја [3,03]	Лесковац [2,78]	Гаџин Хан [2,58]	Бела Црква [1,95]
		Рашка [3,00]	Богатић [2,77]	Сечањ [2,50]	
		Трстеник [2,99]	Параћин [2,77]	Куршумлија [2,44]	
		Кула [2,97]	Брус [2,76]	Нови Пазар [2,42]	
Група 6	6	Жагорађа [1,94]	Трговиште [1,86]	Бојник [1,59]	
		Медвеђа [1,93]	Тутин [1,61]	Лебане [1,41]	

Извор: Ауторов прорачун и систематизација табеларног приказа

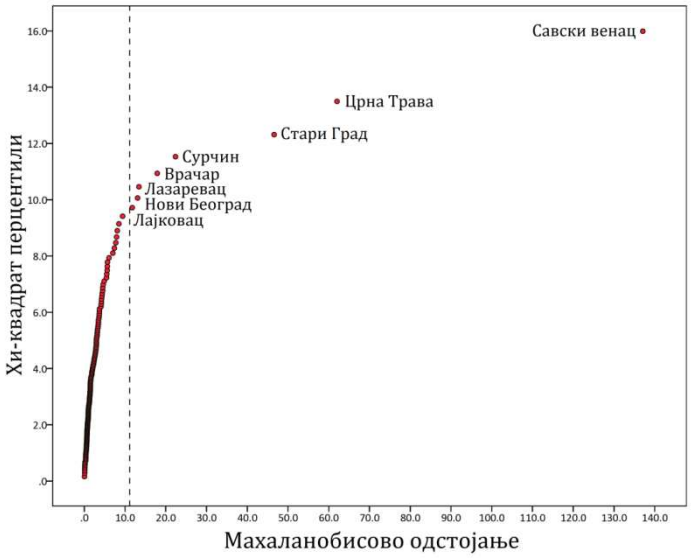
Резултати разврставања 165 градова / општина, применом дефинисаних критеријума за класификацију, у једну од шест издвојених категорија достигнутог степена економске развијености, представљени су у Табели 5.1.18.

5.1.5. Оцена валидности изведеног мултиваријационог модела *ИЕР* применом анализе груписања

У оквиру овог Одељка, извршена је провера одрживости и валидности формулисаних претпоставки у погледу броја и структуре издвојених група на нивоу предложене *ИЕР* класификације *ЈЛС*-а у Републици Србији према степену економске развијености, а у циљу извођења одговарајућих закључака везаних за употребну вредност и практични значај, на бази ФА модела развијеног, композитног показатеља *ИЕР*, чије су вредности у основи поменуте класификације. За потребе наведене провере, креирана је компарабилна класификациона структура, применом (конфирматорне) анализе груписања, конкретно, метода хијерархијске агломеративне процедуре груписања.

Анализа груписања спроведена је на узорку сачињеном од 165 *ЈЛС*-а (градова / општина), коришћењем нормализованих вредности следећа четири показатеља степена економске развијености *ЈЛС*-а: *број предузећа на 1000 становника* (X_1'), *стопа запослености* (X_2'), *број незапослених на 1000 становника* (X_3') и *просечна зарада по запосленом* (X_4'). Поступак нормализације спроведен је путем методе *min–max* трансформације уз проширење опсега у којем се нормализоване вредности крећу, са скале од 0 до 1, на скалу од 1 до 10. Нормализација оригиналних вредности променљивих, извршена је с обзиром на чињеницу да се исте исказују у различитим мерним јединицама и, сходно томе, одликују различитим распонима вредности.

Такође, у оквиру фазе претпроцесирања и припреме података, извршено је и испитивање (евентуалног) присуства универијационих и мултиваријационих нетипичних опсервација, коришћењем графичких приказа представљених у форми *box-plot* дијаграма (Слика 5.1.3) и мултиваријационог *Q-Q* дијаграма (Слика 5.1.12), респективно.



Слика 5.1.12. *Мултиваријациони Q-Q дијаграм за 165 ЈЛС-а*[140]

Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

[140] На датом дијаграму, вредност 97,5 перцентила χ^2 распореда ($\chi^2_{p,\,0{,}975} = 11{,}143$) означена је испрекиданом линијом.

На основу визуелне инспекције приказаног дијаграма и резултата поређења вредности *Mahalanobis*-ове мере одстојања на нивоу појединачних *ЈЛС*-а и критичне вредности 97,5 перцентила χ^2 распореда ($\chi^2_{p,\ 0,975}$ = 11,143), издвојено је присуство 8 мултиваријационих нестандардних опсервација, које одговарају следећим општинама: *Савски венац* [*MD* = 137,080], *Црна Трава* [*MD* = 62,029], *Стари Град* [*MD* = 46,586], *Сурчин* [*MD* = 22,386], *Врачар* [*MD* = 17,901], *Лазаревац* [*MD* = 13,414], *Нови Београд* [*MD* = 13,083] и *Лајковац* [*MD* = 11,764]. Датим списком општина уједно је обухваћено и 8 од укупно 9 идентификованих униваријационих екстрема, уз додатак општине *Медијана*. Будући да идентификоване нестандардне опсервације репрезентују, реално мали (≈4,8%), али важан део анализираног узорка, неопходан за извођење компрабилног модела груписања намењеног евалуацији квалитета *ИЕР* класификације, исте су задржане у узорку и нису искључене из даљег тока анализе.

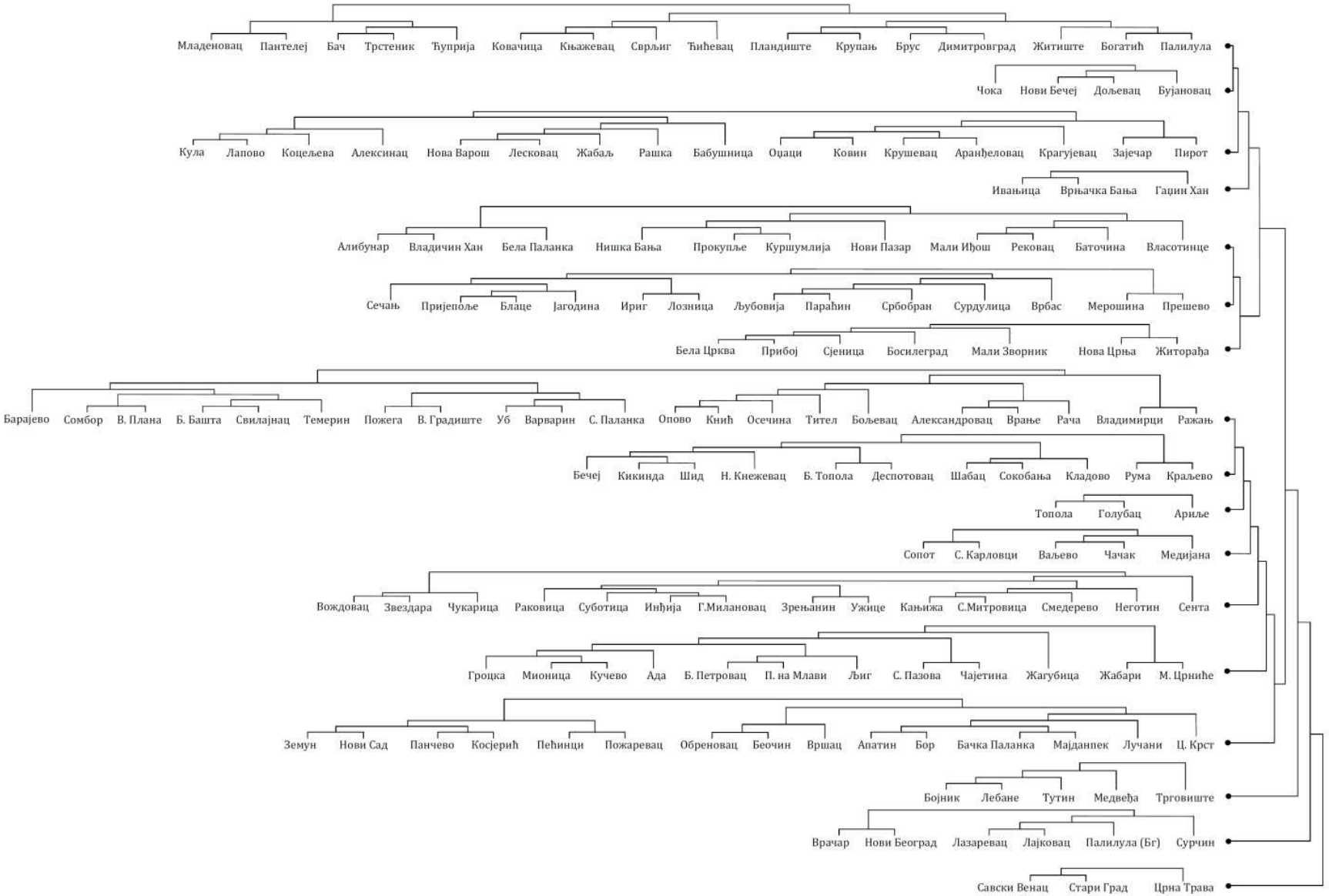
Заснован на коришћењу квадрата Еуклидског одстојања, као адекватне мере блискости између опсервација, поступак груписања расположивих мултиваријационих опсервација спроведен је, иницијално, коришћењем следећих метода хијерархијске агломеративне процедуре груписања: ▸ метод једноструког повезивања; ▸ метод потпуног повезивања; ▸ метод просечног повезивања; ▸ метод центроида; и ▸ *Ward*-ов метод. Избор оптималне мере одстојања између група, односно, на њој засноване хијерархијске методе, извршен је на основу вредности кофенетичког коефицијента корелације (*CPCC*), као статистичког критеријума за квантитативну оцену „величине губитка“ информација узрокованог формирањем хијерархијске структуре скупа могућих решења. Поређењем израчунатих вредности кофенетичког коефицијента на нивоу појединачних метода (Табела 5.1.19), као „оптимална“ (односно, „најквалитетнија“) хијерархијска структура разматраног класификационог проблема, издваја се она добијена применом методе просечног повезивања, будући да исту карактерише највећа вредност *CPCC* коефицијента.

Табела 5.1.19. *Вредности кофенетичког коефицијента корелације за примењене методе хијерархијске агломеративне процедуре груписања*

Примењени метод хијерархијског груписања	Кофенетички коефицијент корелације (*CPCC*)
Метод једноструког повезивања	0,73899
Метод потпуног повезивања	0,55267
Метод просечног повезивања	0,87495
Ward-ов метод	0,32839
Метод центроида	0,86875

Извор: Ауторов прорачун коришћењем програма *Excel*

Сходно наведеном, у наставку излагања, пажња је усмерена на детаљну анализу хијерархијске структуре груписања добијене методом просечног повезивања, која је, коришћењем иновативне форме графичког приказа стабло–дијаграм (овом приликом предложена и названа *комбиновани дендрограм*), приказана на Слици 5.1.13.

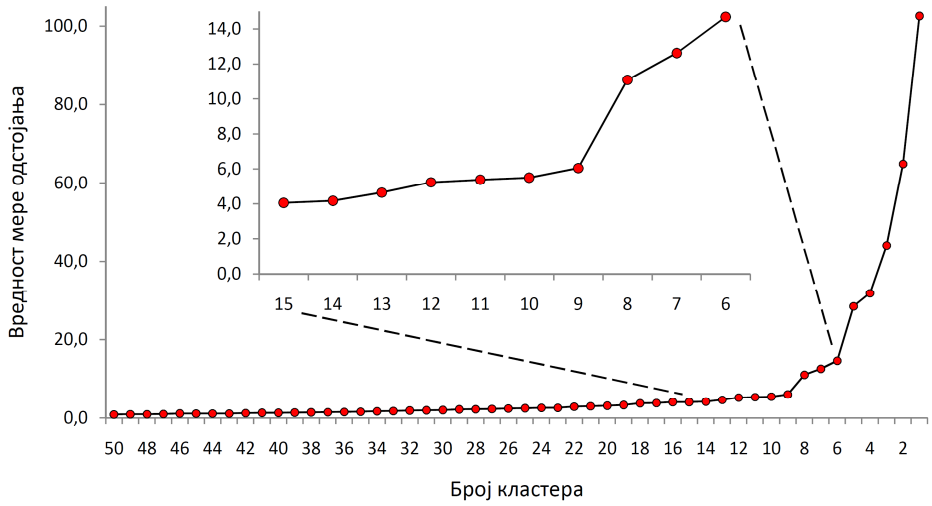


Слика 5.1.13. *Комбиновани-дендрограм – метод просечног повезивања*
Извор: Ауторов (иновативни) визуелни приказ

Будући да приказана хијерархијска структура садржи комплетну серију (укупно n–2) могућих решења проблема груписања, доношење одлуке о избору „оптималног“ броја група (g) и, сходно томе, проналажење конкретне поделе која најбоље репрезентује инхерентну („природну“) структуру *ЈЛС*-а из угла коришћених показатеља економске развијености, извршено је на основу следећих критеријума оптималности:

- ✓ Критеријум заснован на вредностима одстојања између група које се удружују;
- ✓ *Calinski–Harabasz*-ов критеријум оптималности (*псеудо-F мера*);
- ✓ Критеријум заснован на вредностима коефицијента R_g^2;
- ✓ Критеријум заснован на вредностима семипарцијалног R_g^2 коефицијента, ΔR_g^2;
- ✓ Критеријум заснован на вредностима коефицијента кохезије, $coh(C^{(i)})_g$;
- ✓ Критеријум заснован на вредностима коефицијента сепарације, $sep(C^{(i)})_g$;
- ✓ Критеријум заснован на вредностима коефицијента силуете, $\overline{\overline{silh}}(C^{(i)})_g$;

Графички приказ кретања вредности мере одстојања између група, евидентираних при формирању серије могућих решења у распону од g = 50, оствареног у 115. кораку, до g = 2 групе, забележеног у 163. кораку поступка хијерархијског удруживања, представљен је на Слици 5.1.14. Вредности преосталих коефицијената оптималности, израчунате за низ хијерархијских решења у интервалу од g = 13 до g = 3 групе, приказане су у Табели 5.1.20, а њихове промене током обухваћеног (репрезентативног) сегмента поступка удруживања илустроване су Сликама 5.1.16, 5.1.17 и 5.1.18. Ширина изабраног (анализираног) опсега хијерархијских решења, при избору „оптималног“ броја група у решењу, дефинисана је у зависности од одлика коришћеног(их) критеријума, а у циљу јаснијег уочавања присутне тенденције и, консеквентно, наглих / изражених промена (пораста / смањења) у кретању њихових вредности.

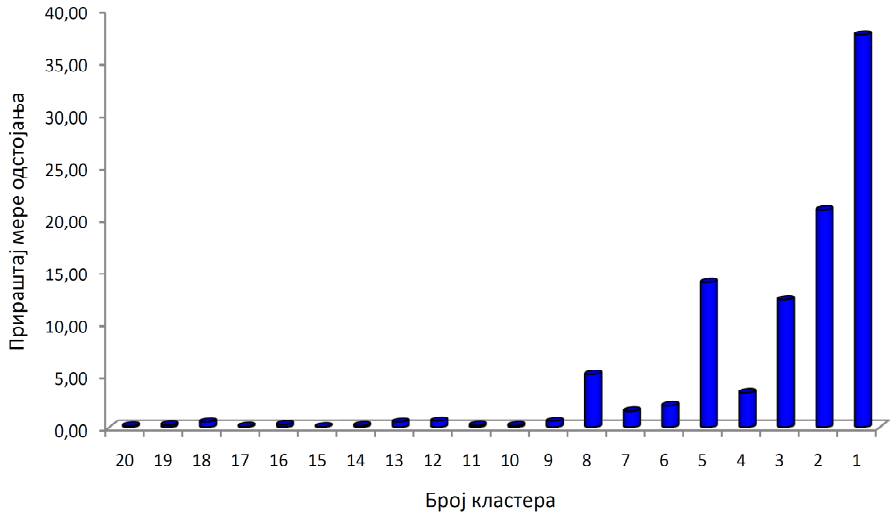


Слика 5.1.14. *Графички приказ кретања вредности мере одстојања између група током процеса удруживања*

Извор: Ауторов визуелни приказ коришћењем програма *Excel*.

Праћењем висине просечног одстојања између група (појединачних *ЈЛС*-а и / или група *ЈЛС*-а) чије се удруживање врши у приказаним, на Слици 5.1.14, сукцесивним корацима хијерархијског процеса груписања, може се приметити да, након доминантно присутне тенденције изразито благог и постепеног раста у иницијалним корацима, при

формирању решења које обухвата 8 група долази до првог, у односу на претходне кораке, значајног повећања вредности мере одстојања и, консекветно, њеног прираштаја (Слика 5.1.15), а затим и њиховог другог „скока“ у 159. кораку, при решењу g = 6 група, након чега приказана крива вредности поприма вертикални облик. На основу изнетих запажања, може се констатовати да се, према овом критеријуму, као „оптималан“, сугерише избор решења са g = 9 или, алтернативно, g = 7 група.



Слика 5.1.15. *Графички приказ кретања прираштаја вредности мере одстојања између група током процеса удруживања*

Извор: Ауторов визуелни приказ коришћењем програма *Excel*.

Табела 5.1.20. *Вредности коефицијената оптималности за хијерархијска решења са различитим бројем група*

Број група (g)	*Pseudo-F*	Δ*Pseudo-F*	R_g^2	ΔR_g^2	$coh(C^{(i)})_g$	$sep(C^{(i)})_g$	$\overline{silh}(C^{(i)})_g$
13	60,83	–	0,828	–	24546,98	398801,54	0,4481
12	65,55	4,72	0,825	-0,003	24578,46	398769,57	0,4441
11	71,54	5,98	0,823	-0,002	24589,23	398758,79	0,4405
10	68,32	-3,21	0,799	-0,024	31721,74	391626,29	0,4752
9	71,95	3,63	0,787	-0,012	32376,81	390971,23	0,4978
8	56,69	-15,26	0,716	-0,071	54334,68	369013,35	0,5312
7	24,28	-32,41	0,480	-0,236	197591,44	225756,59	0,4869
6	27,89	3,61	0,467	-0,013	197826,42	225521,61	0,4958
5	24,53	-3,36	0,381	-0,086	240897,50	182450,53	0,7196
4	31,21	6,68	0,368	-0,013	240961,23	182386,81	0,7157
3	21,48	-9,73	0,210	-0,158	323421,98	99926,05	0,8124

Извор: Ауторов прорачун коришћењем програма *Excel*

Констатације изведене анализом уочених тенденција у кретању вредности коефицијената у основи преосталих критеријума оптималности (Табела 5.1.20), могу бити сумиране, уз уважавање њихових (појединачних) специфичности, на следећи начин:

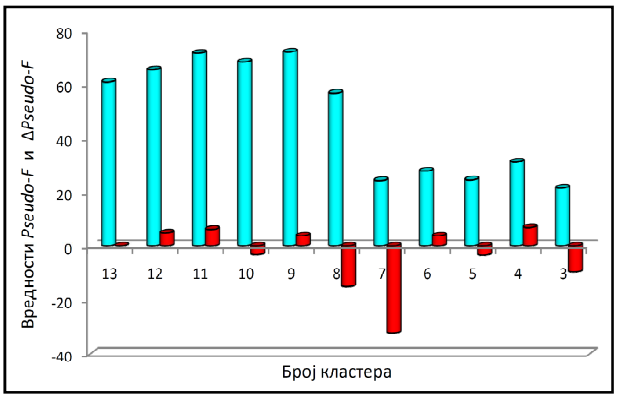
✓ При формирању решења g = 7 група, вредност псеудо-F мере бележи драстичан пад у односу на претходне кораке хијерархијског удруживања, о чему сведочи и висина забележеног (негативног) прираштаја исте (у ознаци, Δ*Pseudo-F*) (Слика 5.1.16).

✓ Након прилично уравнотеженог и благог смањења евидентираног у претходним корацима процеса удруживања, при издвајању решења са g = 7 група, долази до јасно уочљивог пада вредности коефицијента R_g^2, о чему сведочи и висина њеног прираштаја, односно, вредност семипарцијалног R_g^2 коефицијента (симболички, ΔR_g^2) (Слика 5.1.17).

✓ Драстичан пораст вредности показатеља интерне хомогености формираних група, односно коефицијента кохезије, забележен при издвајању решења које садржи 7 група, сугерише да је у том кораку хијерархијског процеса груписања извршено удруживање међусобно знатно „удаљених“ група *ЈЛС*-а (Слика 5.1.18).

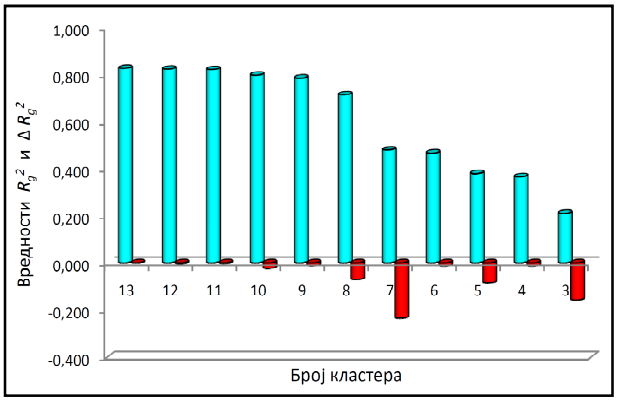
✓ Претходно изнета констатација подржана је и кореспондентним (из угла размере) падом вредности коефицијента сепарације, као мере екстерне хетерогености формираних група, забележеним у истом кораку процеса хијерархијске агломерације (Слика 5.1.18).

✓ Значајан пад вредности коефицијента силуете бележи се, аналогно претходним случајевима, такође при формирању решења које обухвата g=7 група.



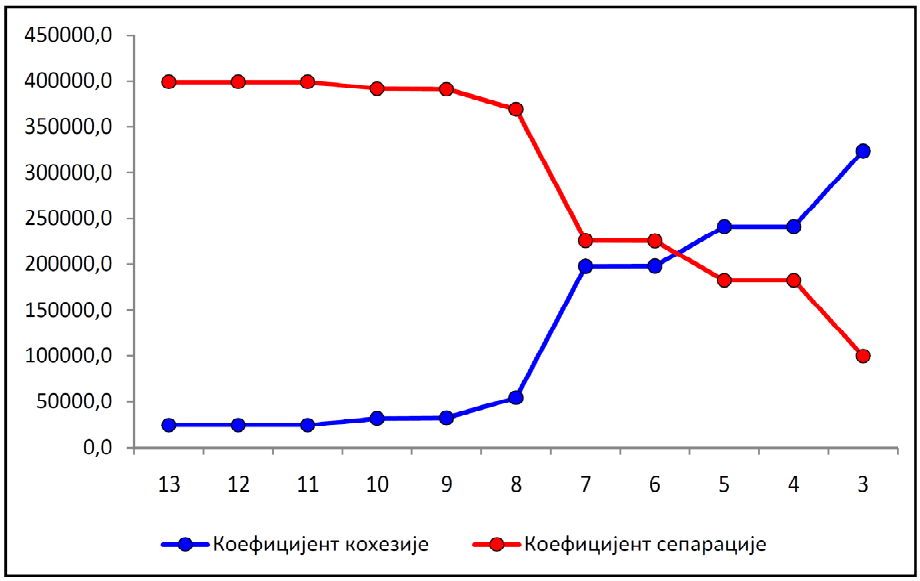
Слика 5.1.16. *Вредности псеудо-F мере и Δпсеудо-F за решења од g=13 до g=3*

Извор: Ауторов визуелни приказ



Слика 5.1.17. *Вредности коефицијената R_g^2 и ΔR_g^2 за решења од g=13 до g=3*

Извор: Ауторов визуелни приказ



Слика 5.1.18. *Графички приказ вредности коефицијената кохезије и сепарације за решења која садрже од g=13 до g=3 групе*

Извор: Ауторов визуелни приказ коришћењем програма *Excel*.

Насупрот препоруке у погледу избора оптималног решења (g=9 или g=7), добијене по основу критеријума заснованом на кретању вредности мере одстојања између група, преосталих шест критеријума оптималности, сагласно претходно изнетим запажањима, једногласно сугеришу избор хијерархијског решења сачињеног од g = 8 група, као најприхватљивије („оптималне“) алтернативе или предлога за груписање анализираних *ЈЛС*-а, у контексту разматраног класификационог проблема. Додатни аргумент у корист (евидентно) најфреквентнијег, из угла примењених критеријума оптималности, предлога за избор решења анализе груписања које се састоји од 8 група, обезбеђује и утврђена вредност коефицијента силуете, као статистичке мере намењене свеобухватној (општој) евалуацији квалитета појединачних решења из угла постигнуте интерне-хомогености и екстерне-хетерогености обухваћених група. Наиме, за разлику од решења са 9, 7 и / или 6 група, за која се може рећи да се одликују слабим (ниским) квалитетом формиране структуре група [$\overline{\overline{sılh}}\,(C^{(i)})_g$ < 0,50], однос остваренog степена интерне-хомогености и екстерне-хетерогености на нивоу формиране класификационе структуре за решење са g = 8 група, може се протумачити као задовољавајући, односно умереног квалитета, будући да је израчуната вредност коефицијента силуете [$\overline{\overline{sılh}}\,(C^{(157)})_{g\,=\,8}$ = 0,5312] већа од 0,50. На основу формулисаних закључака и резултата примене серије критеријума оптималности, решење анализе груписања којим се посматране *ЈЛС*-а класификују у 8 група, изабрано је као „оптимално“. Распоред анализираних *ЈЛС*-а (градова / општина) према изабраном „оптималном“ решењу хијерархијске агломеративне процедуре груписања, по свакој од издвојених осам група (у ознаци, *А*, *Б*, *В*, *Г*, *Д*, *Ђ*, *Е* и *Ж*), представљен је у Табели 5.1.21.

Интерпретација резултирајуће класификације и поређење са ИЕР класификацијом

Првих пет група (у ознаци, *А*, *Б*, *В*, *Г*, *Д*) представљају, такозване, нестандардне кластере, будући да је њима обухваћено свих 8 општина за које је, у оквиру фазе претпроцесирања података, констатовано да представљају како униваријационе тако и мултиваријационе нестандардне опсервације. Очекивано, спроведени процес груписања резултирао је њиховим издвајањем у виду изолованих, једночланих (*група А*, *Б* и *Д*) или пак, група са изразито малим бројем елемената (*група В* и *Г*). Разлика у броју издвојених група на нивоу компарираних класификација (8 : 6 група) представља управо последицу примене анализе груписања на узорку опсервација са задржаним нетипичним мултиваријационим опсервацијама. Наиме, *ИЕР* класификацијом *ЈЛС*-а, општине Савски венац и Стари град, сврстане су у састав једне, заједничке, *групе 1* (у складу са изразито високим % остварења медијалног нивоа економске развијености који их карактерише), док су исте, у случају хијерархијске класификације, издвојене у виду две засебне (једночлане) *групе*, *А* и *Б*. Оправданост њиховог разврставања у састав заједничке групе, као што је учињено *ИЕР* класификацијом, потврђује и дендрограм на Слици 5.1.13, на којем се јасно може уочити да током једног од наредних корака процеса груписања, врло брзо, долази до њиховог (првенствено) међусобног удруживања. Слично запажање и закључак могу се извести и у случају општине Црна Трава, која је, изабраним решењем анализе груписања, распоређена унутар (једночлане) *групе Д*. Наиме, такав исход процеса груписања може се сматрати очекиваним будући да је реч о мултиваријационој нетипичној опсервацији која се и у *ИЕР* класификацији издваја као „најистуренија“ општина *групе 4*, будући да се одликује вредношћу *ИЕР* (4,83) која је

ближа доњој граници *групе 3* (*ИЕР* вредност 4,875) него *ИЕР* вредности првог наредног елемента, њој припадајуће, *групе 4* (општина Земун, *ИЕР* = 4,77). Уважавањем изнетих образложења, елиминише се разлика у погледу броја издвојених група на нивоу разматраних класификација, чиме се потврђује оправданост и валидност предложене *ИЕР* класификације засноване на разврставању *ЈЛС*-а у шест група или категорија према достигнутом степену економске развијености.

Табела 5.1.21. *Класификација општина / градова према решењу анализе груписања заснованом на издвајању осам група ЈЛС-а*

Категорија	Број ЈЛС-а	Назив јединице локалне самоуправе				
Група А	1	*Савски Венац*				
Група Б	1	*Стари Град*				
Група В	2	*Врачар* и *Нови Београд*				
Група Г	4	*Сурчин*, Палилула, *Лазаревац* и *Лајковац*				
Група Д	1	*Црна Трава*				
Група Ђ	81	Земун град Нови Сад Вождовац Звездара град Пожаревац Стара Пазова Косјерић Пећинци Сента град Ужице Чајетина Чукарица Панчево Суботица Беочин Жагубица Ада	Зрењанин Мионица Обреновац Медијана Г. Милановац Кучево Гроцка Инђија Сопот Вршац Петр. на Млави Ваљево Бач. Петровац С. Карловци Љиг Раковица Пожега	Чачак Смедерево Ариље Кањижа Бач. Паланка С. Митровица Неготин Свилајнац Бајина Башта Лучани В. Градиште Кладово Бор Шабац Барајево Темерин Мајданпек	Сомбор Жабари Мало Црниће Велика Плана Деспотовац Бачка Топола Варварин Сокобања Уб Кикинда Краљево Смед. Паланка Шид Александровац Апатин Топола Црвени крст	Н. Кнежевац Рача Рума Осечина Голубац Бечеј Владимирци Бољевац град Врање Кнић Ражањ Опово Тител
Група Е	70	Крагујевац Аранђеловац Ковин Пирот Крушевац Оџаци В. Бања Зајечар Ћуприја Рашка Трстеник Кула Бач Лапово	Ивањица Коцељева Младеновац Врбас Чока Пантелеј Бабушница Књажевац Бујановац Љубовија Сврљиг Жабаљ Ириг Нови Бечеј	Лесковац Богатић Параћин Брус Дољевац Ћићевац Ковачица Јагодина Нова Варош Србобран Лозница Димитровград Алексинац Сурдулица	Пријепоље Палилула Житиште Пландиште Блаце Крупањ Гаџин Хан Сечањ Куршумлија Нови Пазар Прешево Алибунар Прокупље Владич. Хан	Мерошина Нова Црња М. Зворник Нишка Бања Власотинце Мали Иђош Босилеград Рековац Баточина Бела Паланка Сјеница Прибој Бела Црква Житорађа
Група Ж	5	Медвеђа, Трговиште, Тутин, Бојник и Лебане				

Напомене: *Црвеном бојом означене су општине код којих је забележено одступање у погледу утврђене припадности конкретној групи, при поређењу са резултатима ИЕР класификације.*

Општине које представљају мултиваријационе нетипичне опсервације приказане су курзивом.

Извор: Ауторов прорачун и систематизација табеларног приказа

Запажања и закључци у погледу карактеристика структура преосталих група формираних на бази анализе груписања (конкретно, *група В*, *Г*, *Ђ*, *Е* и *Ж*) и извршеног поређења са, претходно предложеном, *ИЕР* класификацијом, могу се формулисати на следећи начин:

✓ *Група В* обухвата општине Врачар и Нови Београд, и у потпуности је идентична *групи 2* у *ИЕР* класификацији.

✓ *Група Г* садржи исте општине које су и *ИЕР* класификацијом разврстане унутар, кореспондентне, *групе 3*, уз додатак општине Лајковац, која је, на основу вредности композитног индекса (*ИЕР* = 4,46), првобитно распоређена у састав *групе 4*. Општина

Лајковац представља још једну од мултиваријационих нетипичних опсервација (Слика 5.1.12), што највероватније представља и „главни“ узрок њене „погрешне“, или (боље речено) другачије алокације у контексту компарираних класификација. Поред наведеног, ова општина се може сматрати и „граничним случајем“ у *ИЕР* класификацији, будући да је, сходно својој *ИЕР* вредности, лоцирана у самом врху *групе 4*, ближе доњој граници групе 3 (4,875) него просечној вредности *ИЕР* на нивоу припадајуће *групе 4* ($\bar{x}_{иер}$= 3,77).

✓ У случају структуре *групе Ђ* остварена је подударност у износу од 96% са, кореспондентном, *групом 4* у *ИЕР* класификацији. Наиме, свега три општине (Лајковац, Црна Трава и Крагујевац [141]) распоређене су хијерархијским решењем у неку од суседних група (конкретно, *групу Г*, *Д* и / или *Е*, респективно). Поред наведеног, *група Ђ* својим саставом обухвата и девет „нових“ општина (Табела 5.1.21), које су, иницијалном разврставањем на бази вредности *ИЕР* индекса, алоциране унутар *групе 5*. Наведених 9 „погрешно“ разврстаних општина могу се третирати као „гранични случајеви“, обзиром на близину њихових *ИЕР* вредности (у опсегу од 3,07 до 3,245) у односу на доњу границу *групе 4* у *ИЕР* класификацији, на нивоу 3,25 индексних поена. Будући да се њихов степен економске развијености креће у интервалу од 94,4% до 99,8% медијалне вредности, не изненађује значајно њихова, анализом груписања сугерисана, алокација у састав *групе Ђ*.

✓ Степен усклађености и подударности структуре *групе Е* и, њој кореспондентне, *групе 5* према *ИЕР* класификацији, износи ≈ 88%. Разлог лежи у „погрешној“ алокацији претходно разматраних девет општина унутар *групе Ђ*, уместо у састав *групе Е*. Такође, овој групи *ЈЛС*-а, сагласно резултатима анализе груписања, додељена је и општина Житорађа, која (иницијално), према *ИЕР* класификацији, представља члана *групе 6*, заједно са општинама Медвеђа, Трговиште, Тутин, Бојник и Лебане, односно групе најслабије развијених локалних самоуправа, у економском смислу. Аналогно претходним образложењима, општина Житорађа представља очигледан пример „граничног случаја“, с обзиром на изразиту близину њене *ИЕР* вредности (1,94) доњој граници *групе 5* (1,95).

Сходно изложеној интерпретацији формиране класификације на бази резултата анализе груписања и запажањима везаним за њену компарацију са, претходно предложеном и, као предмет евалуације из угла квалитета, овом приликом коришћеном, такозваном, *ИЕР* класификацијом, може се закључити да је валидност и одрживост класификације *ЈЛС*-а на бази вредности композитног индекса *ИЕР*, генерално потврђена и верификована, како у погледу броја тако и структуре, односно састава предложених шест група градова / општина. Прецизније, поређењем наведене две класификације, од укупно 165 *ЈЛС*-а, само је у случају 12 општина (или 7,27% узорка) евидентирано одступање у погледу припадности конкретној групи, уз претпоставку прихватања образложења исхода класификације општина Црна Трава и Стари Град. Сходно наведеном, остварени укупан степен подударности *ИЕР* класификације и, њој компаративне, поделе *ЈЛС*-а, добијене имплементацијом хијерархијске агломеративне процедуре груписања, из угла анализираних структура формираних група, износи 92,73%, чиме се, више него недвосмислено, потврђује квалитет предложене *ИЕР* класификације, али и употребна вредност креираног композитног показатеља, у њеној основи.

[141] Имајући у виду чињеницу да се границе *групе 4* у *ИЕР* класификацији крећу у интервалу 3,25 ≤ *ИЕР* < 4,875, општина Крагујевац може се третирати као „гранични случај“, будући да је њена *ИЕР* вредност (3,35) ближа доњој граници (разлика износи свега 0,10) него просечној вредности *ИЕР* на нивоу припадајуће *групе 4* ($\bar{x}_{иер}$= 3,77).

5.1.6. Оцена валидности изведеног композитног показатеља *ИЕР* на основу *MANOVA* модела

Испитивањем одрживости претпоставке о статистичкој значајности (евентуално присутних) разлика између (вектора) просечних вредности одабраних показатеља економске развијености градова / општина на нивоу различитих група *ЈЛС*-а, издвојених у оквиру предложене *ИЕР* класификације, извршена је финална провера валидности и практичног значаја предложеног композитног показатеља *ИЕР*. За потребе реализације наведеног циља, имплементирана је једнофакторска *MANOVA*, коришћењем, на следећи начин дефинисаног, скупа расположивих променљивих (Табела 5.1.22):

✓ Ниво економске развијености *ЈЛС*-а (утврђен *ИЕР* класификацијом) представља независну (категоријску) променљиву X_k, са шест модалитета који одговарају издвојеним групама *ЈЛС*-а.

✓ Четири појединачна показатеља економске развијености *ЈЛС*-а (коришћених при развоју композитног показатеља *ИЕР*) представљају зависне (нумеричке) променљиве (у ознаци, Y_j, (за j = 1, 2, 3 и 4)), чије се аритметичке средине пореде по издвојеним групама.

Табела 5.1.22. *Вредности аритметичких средина зависних променљивих по издвојеним групама ЈЛС-а (модалитетима независне променљиве)*

Независна променљива / Зависне променљиве		Број предузећа на 1000 ст.	Стопа запослености	Број незапослених на 1000 ст.	Просечна зарада по запосленом
модалитети	Величина(n_k)	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$
Група 1	(n_1 = 2)	161,00	314,86	58,50	56,17
Група 2	(n_2 = 2)	96,50	93,47	53,50	62,87
Група 3	(n_3 = 3)	41,67	48,24	56,57	65,75
Група 4	(n_4 = 75)	43,40	37,56	81,05	39,26
Група 5	(n_5 = 77)	33,42	29,28	140,92	33,47
Група 6	(n_6 = 6)	29,33	25,47	223,00	32,93
Σ	165	40,27	37,49	113,10	37,30

Извор: Ауторов прорачун коришћењем програма *Excel*

У складу са излагањем у оквиру Одељка 3.1.2, пре спровођења поступка детаљне провере испуњености претпоставки на којима се заснива валидна имплементација наведене мултиваријационе методе, извршена је редукција иницијално дефинисаних модалитета независне променљиве са аспекта њиховог броја. Наиме, будући да се одликују изразито малим бројем елемената (Табела 5.1.22), прве три групе не задовољавају искуствене препоруке у погледу минимално дозвољеног и прихватљивог броја елемената по групи. У начелу, реч је о групама унутар којих су алоцирани градови / општине високог и изразито високог степена економске развијености (односно, на нивоу изнад 150% медијалне вредности *ИЕР*). Увидом у Табелу 5.1.22, могу се уочити јасне и евидентне разлике у погледу просечних вредности сваког појединачног оригиналног показатеља на нивоу прве три групе у односу на преостале, односно групу 4, 5 и 6, између којих су те разлике мање изражене. Сходно наведеном, обзиром да је њихова величина (n_k) мања од броја коришћених зависних променљивих (p = 4), извршено је њихово искључивање из даљег тока анализе,[142] а пажња је фокусирана на испитивање статистичке

[142] Додатни аргумент у корист елиминисања прве три групе из даље анализе садржан је у чињеници да *ЈЛС* обухваћене њиховим саставом представљају универијационе и / или мултиваријационе нетипичне опсервације, које би свакако,

значајности разлика између вектора аритметичких средина анализираних показатеља на нивоу преостале три, знатно бројније групе, односно групе 4 (n_4=75), 5 (n_5=77) и групе 6 (n_6=6). Дакле, редукована величина узорка коришћеног у *MANOVA* анализи је $n = 158$ *ЈЛС*.

Испитивање испуњености статистичких претпоставки за креирање MANOVA модела

Резултати тестирања статистичких хипотеза о униваријационој нормалности распореда (оригиналних) зависних променљивих представљени су у Табели 5.1.23.

Табела 5.1.23. *Резултати тестирања униваријационе нормалности распореда зависних променљивих (n = 158)*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
Y_1	7,999	0,000	0,658	0,000	0,158	0,000	25,658	H_1	106,763	H_1
Y_2	2,667	0,000	0,853	0,000	0,104	0,000	11,495	H_1	31,621	H_1
Y_3	0,754	0,049	0,979	0,017	0,067	0,081	2,458	H_1	-0,275	H_0
Y_4	3,883	0,000	0,912	0,000	0,108	0,000	6,517	H_1	5,168	H_1

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0, EduStat 4.05* и *Excel*.

Будући да, уз ризик грешке 0,05, униваријациона нормалност променљивих не може бити констатована, извршена је провера присуства униваријационих нетипичних опсервација, поређењем вредности свих показатеља са граничном вредношћу на нивоу $Q_3 \pm 3 * IQR$ и, том приликом, издвојене су следеће „нестандардне“ општине по појединачним зависним променљивима: (Y_1) – Црна Трава; (Y_2) – Црна Трава, Медијана; и (Y_4)–Лајковац. Да би се избегло уклањање истих, примењена је *Box-Cox* трансформација зависних променљивих, [143] а резултати тестирања хипотеза о униваријационој нормалности, добијени на трансформисаним вредностима, приказани су Табелом 5.1.24.

Табела 5.1.24. *Резултати тестирања униваријационе нормалности распореда зависних променљивих након Box-Cox трансформације (n = 158)*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
T-Y_1	0,654	0,0877	0,986	0,103	0,074	0,033	0,000	H_0	2,381	H_1
T-Y_2	1,531	0,0006	0,967	0,001	0,073	0,037	0,000	H_0	3,356	H_1
T-Y_3	0,219	0,8384	0,995	0,848	0,047	0,200	0,000	H_0	-0,921	H_0
T-Y_4	0,275	0,6614	0,994	0,754	0,048	0,200	0,000	H_0	0,123	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0, EduStat 4.05* и *Excel*.

Спроведеним поступком трансформације ублажен је утицај општине Лајковац у случају зависне променљиве (Y_4), али не и у случају променљиве (Y_1) и (Y_2), због чега су општине Црна Трава и Медијана елиминисане из узорка за анализу. Резултати поновљеног поступка униваријационог тестирања нормалности представљени су у Табели 5.1.25.

током поступка припреме података и провере статистичких претпоставки, биле уклоњене из узорка за анализу, што је и потврђено поступком примене факторске анализе.

[143] *Box-Cox*-ова трансформација зависних променљивих извршена је за следеће оптималне вредности трансформационог параметра (λ): (Y_1) → λ_1 = –0,6363; (Y_2) → λ_2 = –0,1335; (Y_3) → λ_3 = 0,5084; и (Y_4) → λ_4 = –1,5928.

Табела 5.1.25. *Резултати тестирања универијационе нормалности распореда зависних променљивих (n = 156)*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
Y_1	3,054	0,0000	0,932	0,000	0,127	0,000	5,543	H_1	3,834	H_1
Y_2	0,413	0,3377	0,984	0,075	0,047	0,200	1,367	H_0	1,005	H_0
Y_3	0,809	0,0364	0,978	0,014	0,070	0,058	2,483	H_1	-0,303	H_0
Y_4	3,901	0,0000	0,911	0,000	0,113	0,000	6,471	H_1	5,058	H_1

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*, *EduStat 4.05* и *Excel*.

Анализом нестандардних опсервација идентификовано је присуство само једне (праве) нетипичне опсервације – општина Лајковац, на нивоу променљиве (Y_4), док је у случају осталих променљивих забележено присуство неколико умерених екстремних вредности. Трансформациони поступак је поновљен уз изузимање променљиве (Y_2),[144] будући да је код исте потврђена универијациона нормалност распореда на оригиналним вредностима, а резултати накнадно спроведеног поступка тестирања дати су у Табели 5.1.26.

Табела 5.1.26. *Резултати тестирања универијационе нормалности распореда зависних променљивих након Box-Cox трансформације (n = 156)*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
T-Y_1	0,542	0,1644	0,991	0,383	0,064	0,200	0,000	H_0	0,416	H_0
Y_2	0,413	0,3377	0,984	0,075	0,047	0,200	1,367	H_0	1,005	H_0
T-Y_3	0,235	0,7933	0,994	0,824	0,048	0,200	0,000	H_0	-0,951	H_0
T-Y_4	0,266	0,6912	0,994	0,760	0,046	0,200	0,000	H_0	0,087	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*, *EduStat 4.05* и *Excel*.

Представљени резултати уз ризик грешке α = 0,05, потврђују универијациону нормалност распореда све четири зависне променљиве, а испитивањем нетипичних опсервација није утврђено присуство истих у трансформисаним вредностима. Сходно наведеном, у наредном кораку, коришћењем трансформисаних вредности променљивих *T*-Y_1, *T*-Y_3, *T*-Y_4 и оригиналних вредности променљиве Y_2, извршено је испитивање претпоставке о нормалности мултиваријационог распореда, а резултати тестирања представљени су у Табели 5.1.27.

Табела 5.1.27. *Резултати тестирања статистичких хипотеза о нормалности мултиваријационог распореда зависних променљивих (n = 156)*

Мултиваријациони тестови нормалности	Статистика теста	*p*-вредност	Одлука
Mardia тест симетричности	37,756	0,008	H_1
Mardia тест заобљености	2,237	0,062	H_0
Henze–Zirkler–ов тест	1,143	0,002	H_1

Извор: Ауторов прорачун коришћењем програма *SYSTAT 13.1*.

Иако је представљеним резултатима тестирања, уз ризик грешке α = 0,05, потврђена нормална заобљеност мултиваријационог распореда зависних променљивих на нивоу популације, добијене *p*-вредности код преостала два теста, будући да су значајно мање од дефинисаног нивоа значајности (0,05), сугеришу одбацивање нулте хипотезе и усвајање алтернативних хипотеза према којима заједнички распоред *p* променљивих, на нивоу

[144] *Box-Cox*-ова трансформација зависних променљивих извршена је за следеће оптималне вредности трансформационог параметра: (Y_1) → λ_1 = –0,297; (Y_3) → λ_3 = 0,4972; и (Y_4) → λ_4 = –1,605.

популације, није симетричан, односно (генерално) није нормално распоређен, респективно. Полазећи од претпоставке да су дати закључци последица „деловања“ мултиваријационих нетипичних опсервација, извршена је провера присуства истих у расположивом узорку од 156 градова / општина. Том приликом, издвојено је 6 општина којима је додељен статус мултиваријационих нетипичних опсервација, будући да су утврђене, њима припадајуће, вредности *Mahalanobis*-ове мере одстојања веће од утврђене критичне вредности, $\chi^2_{p,\,0{,}975}$ = 11,14329, и то: *Жабари* [*MD* = 18,494], *Рача* [*MD* = 15,510], *Мало Црниће* [*MD* = 14,937], *Нови Сад* [*MD* = 13,236], *Жагубица* [*MD* = 11,862] и општина *Медвеђа* [*MD* = 11,375]. У даљем току анализе, спроведен је итеративни поступак постепеног („корак-по-корак“) искључења (из узорка) појединачних општина, за које је установљено да представљају мултиваријационе нестандардне опсервације, заснован на реализацији следећих етапа:

- Избор најекстремније мултиваријационе нестандарадне опсервације (општине), мерено величином израчунатих вредности *Mahalanobis*-ове мере одстојања;
- Елиминисање изабране општине из узорка мултиваријационих опсервација;
- Тестирање статистичких хипотеза о универијационој нормалности, коришћењем оригиналних вредности променљивих у редукованом узорку, величине n–1;
- Испитивање (евентуалног) присуства универијационих нетипичних вредности;
- У случају нарушености претпоставке о универијационој нормалности, примена *Box-Cox*-ове трансформације вредности оригиналних променљивих;
- Тестирање статистичких хипотеза о универијационој нормалности, коришћењем трансформисаних вредности променљивих у редукованом узорку, величине n–1;
- Тестирање статистичких хипотеза о нормалности мултиваријационог распореда p трансформисаних, универијационо нормално распоређених, зависних променљивих;
- Уколико резултати тестирања у претходној етапи сугеришу одбацивање H_0, врши се провера присуства мултиваријационих нетипичних опсервација, уз итеративно спровођење претходних етапа, до тренутка док се не осигура испуњеност претпоставки о мултиваријационој нормалности и одсуству мултиваријационих нетипичних опсервација.

Имплементацијом представљеног поступка, из узорка мултиваријационих опсервација, величине n = 156, искључено је укупно следећих 14 општина (наведених према редоследу њиховог елиминисања), и то: *Жабари*, *Рача*, *Мало Црниће*, *Нови Сад*, *Жагубица*, *Медвеђа*, *Гроцка*, *Трговиште*, *Голубац*, *Сремски Карловци*, *Раковица*, *Гаџин Хан*, *Ариље* и *Сопот*. Резултати тестирања хипотеза о универијационој нормалности оригиналних, а затим и трансформисаних зависних променљивих[145], дати су у Табели 5.1.28 и 5.1.29, респективно.

Табела 5.1.28. *Резултати тестирања универијационе нормалности распореда зависних променљивих (n = 142)*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
Y_1	2,054	0,0000	0,952	0,000	0,112	0,000	4,072	H_1	2,055	H_1
Y_2	0,400	0,3625	0,988	0,249	0,051	0,200	-0,370	H_0	-0,542	H_0
Y_3	0,588	0,1254	0,978	0,021	0,060	0,200	2,282	H_1	-0,131	H_0
Y_4	4,282	0,0000	0,892	0,000	0,119	0,000	6,748	H_1	5,356	H_1

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*, *EduStat 4.05* и *Excel*.

[145] *Box-Cox*-ова трансформација зависних променљивих (изузев променљиве Y_2) извршена је за коришћењем следећих, оптималних вредности трансформационог параметра: $(Y_1) \rightarrow \lambda_1 = -0{,}1026$; $(Y_3) \rightarrow \lambda_3 = 0{,}4681$; и $(Y_4) \rightarrow \lambda_4 = -2{,}4466$.

Табела 5.1.29. *Резултати тестирања униваријационе нормалности распореда зависних променљивих након Box-Cox трансформације (n = 142)*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
T-Y_1	0,549	0,1579	0,988	0,287	0,064	0,200	0,000	H_0	0,557	H_0
Y_2	0,400	0,3625	0,988	0,249	0,051	0,200	-0,370	H_0	-0,542	H_0
T-Y_3	0,263	0,7021	0,991	0,469	0,057	0,200	0,000	H_0	-1,075	H_0
T-Y_4	0,277	0,6530	0,992	0,592	0,054	0,200	0,000	H_0	-0,898	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*, *EduStat 4.05* и *Excel*.

Међу оригиналним (у случају промењљиве Y_2) и трансформисаним вредностима осталих зависних променљивих није идентификовано присуство униваријационих нестандардних опсервација. Резултати поступка тестирања хипотеза о мултиваријационој нормалности распореда зависних променљивих представљени су у Табели 5.1.30.

Табела 5.1.30. *Резултати тестирања хипотеза о нормалности мултиваријационог распореда зависних променљивих (n = 142)*

Мултиваријациони тестови	Статистика теста	*p*-вредност	Одлука
Mardia тест симетричности	22,619	0,308	H_0
Mardia тест заобљености	-1,422	0,155	H_0
Henze–Zirkler–ов тест	0,989	0,128	H_0

Извор: Ауторов прорачун коришћењем програма *SYSTAT 13.1*.

Презентовани резултати, уз ризик грешке α = 0,05, недвосмислено сугеришу да не постоји довољно емпиријских доказа за одбацивање претпоставке о нормалности заједничког распореда четири зависне променљивих на нивоу популације, будући да су добијене *p*-вредности за сва три теста значајно веће од постављеног нивоа значајности теста, α. Другим речима, за дати ризик грешке, може се закључити да је претпоставка о нормалности распореда мултиваријационих опсервација на нивоу популације, испуњена. У расположивом (редукованом) узорку, величине n=142, није идентификовано присуство мултиваријационих нестандардних опсервација.

Испитивање одрживости претпоставке у погледу линеарне повезаности и одсуства мултиколинеарности између зависних променљивих, спроведено је израчунавањем вредности *Pearson*-ових коефицијената просте линеарне корелације између свих парова променљивих и тестирањем њихове статистичке значајности, уз ризик грешке α = 0,05. Будући да се валидна примена поступка тестирања заснива на потврђеној нормалности заједничког распореда разматраних парова променљивих, у Табели 5.1.31, дати су резултати тестирања хипотеза о нормалности кореспондентних дводимензионих распореда. Коначни резултати просте линеарне корелационе анализе представљени су у форми одговарајуће корелационе матрице (Табела 5.1.32).

Табела 5.1.31. *Резултати тестирања статистичких хипотеза о нормалности дводимензионог распореда парова зависних променљивих*

Парови променљивих	*Mardia* тест симетричности		*Mardia* тест заобљености		*Henze-Zirkler*-ов тест нормалности		Одлука
	статистика	*p*-вредност	статистика	*p*-вредност	статистика	*p*-вредност	
T-Y_1 и Y_2	0,220	0,994	-1,011	0,312	0,767	0,400	H_0
T-Y_1 и T-Y_3	3,314	0,507	-0,770	0,442	0,677	0,639	H_0
T-Y_1 и T-Y_4	0,691	0,952	-0,915	0,360	0,408	0,297	H_0
Y_2 и T-Y_3	9,364	0,053	-0,517	0,605	1,023	0,089	H_0
Y_2 и T-Y_4	2,423	0,658	-1,574	0,115	0,631	0,797	H_0
T-Y_3 и T-Y_4	3,118	0,538	-1,647	0,099	0,586	0,969	H_0

Извор: Ауторов прорачун коришћењем програма *SYSTAT 13.1*.

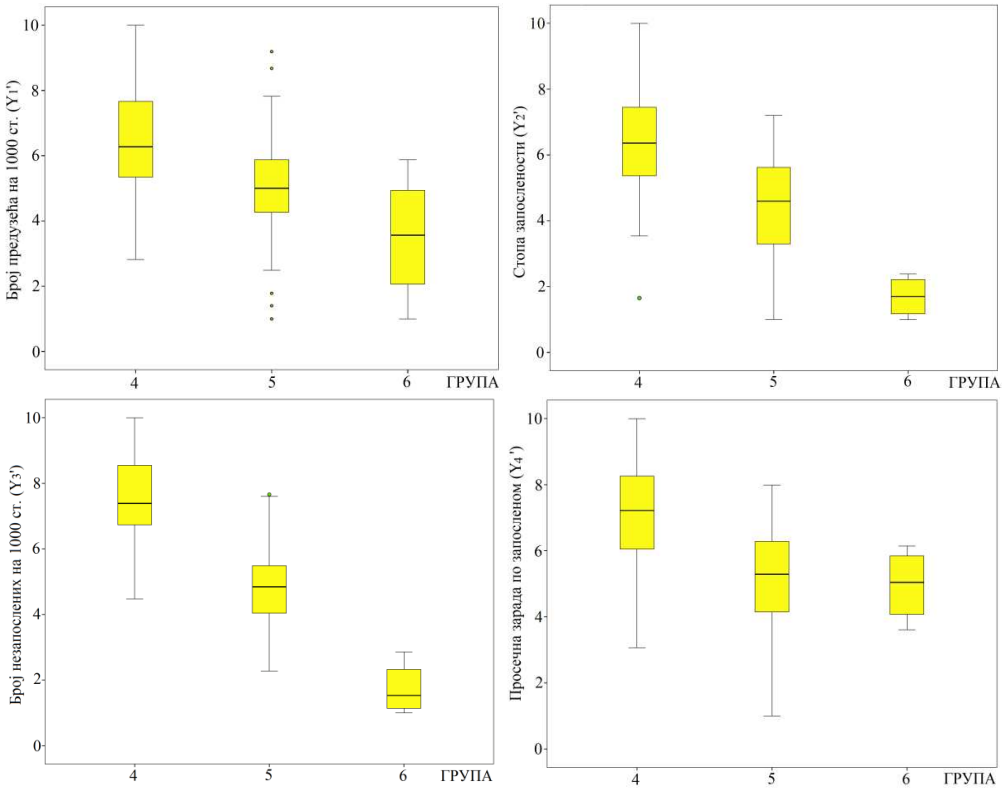
Табела 5.1.32. *Корелациона матрица анализираних зависних промењивих*

Зависне промењиве	T-Y_1	Y_2	T-Y_3	T-Y_4
T-Y_1	1,000	0,474 [**] *p*-вред. = 0,000	-0,399 [**] *p*-вред. = 0,000	0,220 [**] *p*-вред. = 0,009
Y_2	0,474 [**] *p*-вред. = 0,000	1,000	-0,515 [**] *p*-вред. = 0,000	0,395 [**] *p*-вред. = 0,000
T-Y_3	-0,399 [**] *p*-вред. = 0,000	-0,515 [**] *p*-вред. = 0,000	1,000	-0,309 [**] *p*-вред. = 0,000
T-Y_4	0,220 [**] *p*-вред. = 0,009	0,395 [**] *p*-вред. = 0,000	-0,309 [**] *p*-вред. = 0,000	1,000

Напомена: Симбол [**] означава статистичку значајност израчунатих оцена уз ризик грешке (I врсте), α=0,05

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.

Израчунате вредности коефицијената просте линеарне корелације (*r*), као и резултати тестирања хипотеза о њиховој статистичкој значајности, сугеришу да између свих парова зависних промењивих постоји статистички значајна линеарна веза на нивоу популације, чиме је потврђена испуњеност претпоставке о линеарној повезаности. У корелационој матрици нису евидентиране вредности коефицијената веће од 0,80 или 0,90 будући да се исте, у апсолутном износу, крећу у интервалу од $|r_{min}| = 0{,}220$ до $|r_{max}| = 0{,}515$. Сходно наведеном, може се констатовати да међу анализираним показатељима економске развијености *ЈЛС*-а нема високо корелисаних промењивих, чиме је уједно и потврђена испуњеност претпоставке о одсуству мултиколинеарности. Претпоставке које се односе на одсуство универијационих нестандардних опсервација и универијациону нормалност распореда зависних промењивих (*T*-Y_1, Y_2, *T*-Y_3 и *T*-Y_4) на нивоу појединачних модалитета независне промењиве (група 4, 5 и 6), такође су проверене, а потврде њихове испуњености представљене су на Слици 5.1.19 и у Табели 5.1.33.



Слика 5.1.19. *Box-plot дијаграми зависних промењивих по групама ЈЛС-а*

Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Табела 5.1.33. *Резултати тестирања универијационе нормалности распореда зависних променљивих по појединачним групама ЈЛС-а*

Зависне променљиве	Група	Величина групе (n_k)	*Shapiro-Wilk* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
			статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
Y_1'	4	63	0,978	0,315	0,78741	H_0	-0,59137	H_0
	5	75	0,987	0,660	-0,37123	H_0	0,83439	H_0
	6	4	0,995	0,981	-0,14370	H_0	0,26046	H_0
Y_2'	4	63	0,989	0,838	-0,29487	H_0	0,47633	H_0
	5	75	0,962	0,023	-1,36472	H_0	-1,26926	H_0
	6	4	0,946	0,689	-0,01306	H_0	-1,32599	H_0
Y_3'	4	63	0,981	0,417	0,19766	H_0	-0,59137	H_0
	5	75	0,988	0,723	0,18385	H_0	-0,16087	H_0
	6	4	0,920	0,538	0,94632	H_0	0,37028	H_0
Y_4'	4	63	0,975	0,233	-1,33827	H_0	-0,63511	H_0
	5	75	0,981	0,302	-1,32583	H_0	-0,10783	H_0
	6	4	0,975	0,869	-0,25883	H_0	-0,74301	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *Excel*.

На основу представљених графичких приказа у форми *box-plot* дијаграма, може се такође приметити да је извршена и нормализација вредности зависних променљивих на којима је потврђена универијациона и мултиваријациона нормалност распореда, коришћењем метода *min-max* трансформације (израз (1.4.2) и (1.4.3)). Наиме, наведеним поступком вредности зависних променљивих (T-Y_1, Y_2, T-Y_3, T-Y_4) конвертоване су у одговарајуће нормализоване вредности (Y_1', Y_2', Y_3', Y_4') на скали од 0 до 1, након чега је извршено њихово рескалирање на интервал од 1 до 10, идентично раније описаном поступку, спроведеном при имплементацији факторске анализе.

Коначно, провера одрживости претпоставке о хомогености коваријационих матрица на нивоу разматране три групе (односно популација из којих су групе „узорковане“), дефинисана изразом (3.1.14), извршена је применом *Box*-овог *M* теста, а резултати поступка тестирања представљени су у Табели 5.1.34.

Табела 5.1.34. *Резултати примене Box-овог M теста*

	Вредност статистике	Бр. степени слободе		*p*-вред.	Одлука
		v_1	v_2		
Box-ов *M* тест	15,884				
F апроксимација	1,537	10	82540,845	0,119	H_0
χ^2 апроксимација	12,787	20		> 0,10	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *Excel*.

Добијена вредност статистике *Box*-овог *M* теста (M = 15,884) и, на бази ње изведене вредности апроксимативне *F* статистике ($F_{(\alpha;\ 10;\ 82540,845)}$ = 1,537) и χ^2 апроксимације ($\chi^2_{(\alpha;\ 20)}$ = 12,787), уз ниво значајности теста α = 0,05, сугеришу да не постоји довољно емпиријских доказа за одбацивање H_0 претпоставке о хомогености коваријационих матрица три групе мултиваријационих опсервација, будући да су резултирајуће *p*-вредности, утврђене за обе апроксимативне статистике, веће од ризика грешке I врсте, α.

Изведени закључак потврђен је и резултатима *Levene*-овог теста хомогености варијанси појединачних зависних променљивих за издвојене групе на нивоу популација (Табела 5.1.35).

Табела 5.1.35. *Резултати Levene-овог теста једнакости варијанси промењљивих*

Зависне промењљиве	Вредност F статистике	Бр. степени слободе		p-вред.	Одлука
		v_1	v_2		
Y_1'	0,299	2	139	0,742	H_0
Y_2'	1,391	2	139	0,252	H_0
Y_3'	0,764	2	139	0,468	H_0
Y_4'	1,175	2	139	0,312	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.

Прецизније, у случају све четири зависне промењљиве није евидентирана статистичка значајност добијених вредности F статистике *Levene*-овог теста, будући да су утврђене p-вредности, у сва четири случаја, значајно веће од ризика грешке I врсте, $\alpha = 0{,}05$, сугеришући да не постоји довољно доказа за одбацивање претпоставке о хомогености варијанси појединачних зависних промењљивих за издвојене групе, на нивоу популације.

Оцена једнофакторског MANOVA модела и примена MANOVA тестова

Након комплетне провере и обезбеђене потврде испуњености статистичких претпоставки на којима се заснива валидна примена *MANOVA* методе, у складу са презентованим поступком у Одељку 3.1.1, извршено је оцењивање једнофакторског *MANOVA* модела, као основе за тестирање кореспондентне нулте хипотезе (изразом (3.1.1)). Заснован на идеји разлагања укупног варијабилитета свих p зависних (нумеричких) промењљивих на део који је узрокован припадношћу *ЈЛС*-а конкретном модалитету независне промењљиве и део изазван дејством резидуалних фактора, оцењени једнофакторски *MANOVA* модел (израз (3.1.19)) може се представити на следећи начин:

$$\underbrace{\begin{bmatrix} 447,73 & 235,01 & 198,41 & 96,81 \\ 235,01 & 549,38 & 238,84 & 192,99 \\ 198,41 & 283,84 & 551,97 & 151,16 \\ 96,81 & 192,99 & 151,16 & 434,34 \end{bmatrix}}_{\substack{\text{УКУПАН ВАРИЈАБИЛИТЕТ} \\ \text{(матрица укупне суме квадрата и узајамних производа одступања)} \\ [\mathbf{T}]}} = \underbrace{\begin{bmatrix} 97,50 & 139,79 & 182,37 & 105,56 \\ 139,79 & 201,20 & 261,79 & 148,80 \\ 182,37 & 261,79 & 341,25 & 196,39 \\ 105,56 & 148,80 & 196,39 & 122,64 \end{bmatrix}}_{\substack{\text{ФАКТОРСКИ ВАРИЈАБИЛИТЕТ} \\ \text{(матрица суме квадрата и узајамних производа одступања између група)} \\ [\mathbf{B}]}} + \underbrace{\begin{bmatrix} 350,23 & 95,22 & 16,04 & -8,75 \\ 95,22 & 348,18 & 22,05 & 44,19 \\ 16,04 & 22,05 & 210,72 & -45,23 \\ -8,75 & 44,19 & -45,23 & 311,70 \end{bmatrix}}_{\substack{\text{РЕЗИДУАЛНИ ВАРИЈАБИЛИТЕТ} \\ \text{(матрица суме квадрата и узајамних производа одступања унутар група)} \\ [\mathbf{W}]}}. \quad (5.1.5)$$

Оцењивање представљеног модела *MANOVA* са једним фактором (у ознаци X_k, $k = 1, 2, 3$), спроведено је коришћењем нормализованих вредности четири зависне промењљиве (у ознаци Y_j', $j = 1, 2, 3$ и 4) расположивих на нивоу узорка од 142 *ЈЛС*-а, распоређених у три групе или категорије према „измереном“ степену економске развијености. На основама изведеног модела и његових компоненти, конкретно, матрице суме квадрата између и унутар група (симболички, **B** (израз (3.1.10)) и **W** (израз (3.1.11)), респективно) извршено је тестирање статистичке значајности разлика између вектора средина четири показатеља економске развијености, утврђених на нивоу три групе (издвојене *ИЕР* класификацијом) јединица локалних самоуправа, коришћењем следећих статистика *MANOVA* тестова: *Wilks*-ова статистика (Λ), *Pillai*-јева статистика (V), *Lawley–Hotelling*-ова статистика (U) и *Roy*-ова статистика теста (θ).

Табела 5.1.36. *Резултати примене MANOVA тестова*

Статистике *MANOVA* тестова	Вредност статистике теста	Вредност *F* статистике (апроксимација)	Бр. степени слободе		*p*-вред.	Одлука
			v_1	v_2		
Wilks-ова статистика (Λ)	0,260	32,651	8	272	0,000	H_1
Pillai-јева статистика (V)	0,761	21,051	8	274	0,000	H_1
Lawley–Hotelling-ова статистика (*U*)	2,760	46,577	8	270	0,000	H_1
Roy-ова статистика (θ)	2,730	93,495	4	137	0,000	H_1

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *Excel*.

Добијени резултати примене четири *MANOVA* статистике, представљени у Табели 5.1.36, недвосмислено и „једногласно“ сугеришу одбацивање нулте хипотезе, уз ризик грешке $\alpha = 0{,}05$, и усвајање алтернативне, којом се тврди да постоји статистички значајна разлика између најмање две групе *ЈЛС*-а у погледу припадајућих вектора просечних вредности четири показатеља економске развијености, будући да је реализовани ниво значајности (*p*-вредност) апроксимативних *F* статистика, у случају све четири *MANOVA* статистике, мањи од ризика грешке, α. Другим речима, добијеним износима резултирајућих *p*-вредности обезбеђени су јаки (емпиријски) докази за прихватање тврдње да најмање два од посматрана три „узорка“ (групе) мултиваријационих опсервација потичу из (засебних, у погледу степена економске развијености) популација које се одликују различитим векторима средина анализираних зависних променљивих.

У циљу бољег разумевања добијених (статистички сигнификатних) резултата *MANOVA* анализе, спроведена је једнофакторска *ANOVA* за појединачне показатеље економске развијености (Табела 5.1.37), а затим и *post hoc* анализа, заснована на вишеструкој компарацији свих парова група *ЈЛС*-а, применом *Tukey*-јевог *HSD* теста (Табела 5.1.38).

На основу резултата спроведене *ANOVA* процедуре, може се закључити да, у случају све четири посматране зависне променљиве, постоји довољно аргумената за усвајање *ANOVA* алтернативне хипотезе којом се тврди да постоји статистички значајна разлика између најмање две (од посматране три) групе *ЈЛС*-а у погледу просечних вредности појединачних показатеља економске развијености, будући да су добијене *p*-вредности, за сваку од четири вредности $F_{(2;139)}$ статистике теста, мање од прилагођеног (*Bonferonni*-јевим приступом коригованог) нивоа значајности, $\alpha^* = 0{,}0125$.

Табела 5.1.37. *Резултати једнофакторске ANOVA анализе*

Зависне променљиве		Статистика *F* теста	Број степени слободе		*p*-вред.	Кориговани ризик грешке α^*	Парцијални ета-квадрат (η_p^2)
			v_1	v_2			
Број предузећа на 1000 ст.	Y_1'	19,347	2	139	0,000	0,0125	0,218
Стопа запослености	Y_2'	40,162	2	139	0,000	0,0125	0,366
Незапосленост на 1000 ст.	Y_3'	112,551	2	139	0,000	0,0125	0,618
Просечна зарада по запосленом	Y_4'	27,344	2	139	0,000	0,0125	0,282

Напомена: Прилагођавање *a priori* дефинисаног ризика грешке $\alpha = 0{,}05$, извршено је путем *Bonferonni*-јеве корекције на следећи начин: $\alpha^* = \alpha / p = 0{,}05 / 4 = 0{,}0125$.

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.

Такође, поређењем величина израчунатих вредности *F* статистика и показатеља η_p^2, приметно је да се зависна променљива - *незапосленост на 1000 становника*, одликује најповољнијим односом факторског и резидуалног варијабилитета из угла разматраних

група *ЈЛС*-а, а тиме и највећим доприносом у контексту обезбеђивања статистичке значајности резултата *MANOVA*-е. Остале зависне промењљиве карактерише у извесној мери мањи (*стопа запослености*) или пак знатно мањи допринос (*број предузећа на 1000 становника* и *просечна зарада по запосленом*) у претходно наведеном контексту.

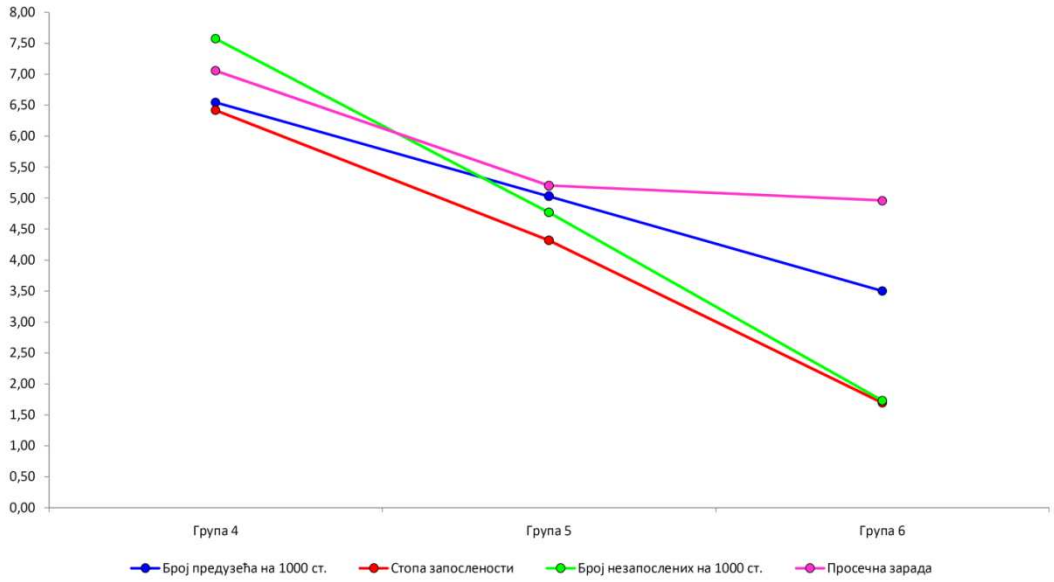
Табела 5.1.38. *Резултати Tukey-јевог HSD теста универијационих накнадних поређења за појединачне варијабле по паровима посматране три групе ЈЛС-а*

Зависне промењљиве	Парови група које се пореде		Разлика средина	Стандардна грешка	*p*-вред.	Одлука
Број предузећа на 1000 ст. (Y_1')	4	5	1,5127*	0,27127	0,000	H_1
		6	3,0416*	0,81848	0,001	H_1
	5	4	-1,5127*	0,27127	0,000	H_1
		6	1,5289	0,81456	0,149	H_0
	6	4	-3,0416*	0,81848	0,001	H_1
		5	-1,5289	0,81456	0,149	H_0
Стопа Запослености (Y_2')	4	5	2,1022*	0,27048	0,000	H_1
		6	4,7251*	0,81608	0,000	H_1
	5	4	-2,1022*	0,27048	0,000	H_1
		6	2,6229*	0,81217	0,004	H_1
	6	4	-4,7251*	0,81608	0,000	H_1
		5	-2,6229*	0,81217	0,004	H_1
Број незапослених на 1000 ст. (Y_3')	4	5	2,8020*	0,21042	0,000	H_1
		6	5,8394*	0,63486	0,000	H_1
	5	4	-2,8020*	0,21042	0,000	H_1
		6	3,0374*	0,63183	0,000	H_1
	6	4	-5,8394*	0,63486	0,000	H_1
		5	-3,0374*	0,63183	0,000	H_1
Просечна зарада по запосленом (Y_4')	4	5	1,8566*	0,25592	0,000	H_1
		6	2,0979*	0,77215	0,020	H_1
	5	4	-1,8566*	0,25592	0,000	H_1
		6	0,2413	0,76845	0,947	H_0
	6	4	-2,0979*	0,77215	0,020	H_1
		5	-0,2413	0,76845	0,947	H_0

Напомена: Симбол [*] коришћен у III колони табеле означава статистичку значајност разлика средина упарених група у контексту конкретне зависне промењљиве при нивоу значајности, α = 0,05.

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.

Изнето запажање потврђено је и резултатима *post hoc* вишеструке компарације појединачних парова разматране три групе (Табела 5.1.38). Наиме, из угла промењљивих *број незапослених на 1000 становника* и *стопа запослености*, потврђена је, уз ризик грешке α = 0,05, статистичка значајност разлика њихових просечних вредности између све три групе *ЈЛС*, односно на нивоу свих парова разматраних група (групе 4, 5 и групе 5). У случају преостала два показатеља економске развијености, верификована је статистичка значајност разлика средина утврђених на нивоу следећих парова група: група 4 – група 5 и група 4 – група 6. При тестирању хипотезе о једнакости аритметичких средина ове две промењљиве засебно, између групе 5 и групе 6, може се закључити да не постоји довољно доказа за одбацивање H_0, будући да су добијене *p*-вредности значајно веће од дефинисаног ризика грешке α = 0,05. Упоредни приказ (евидентно присутних и претходним анализама потврђених разлика) просечних вредности зависних промењљивих по издвојеним групама *ЈЛС*-а илустрован је на Слици 5.1.20.



Слика 5.1.20. *Упоредни приказ просечних вредности зависних променљивих по издвојеним групама ЈЛС-а*

Извор: Ауторов визуелни приказ коришћењем програма *Excel*.

Потврдом статистичке значајности разлика присутних између вредности аритметичких средина показатеља економске развијености *ЈЛС*-а распоређених унутар група 4, 5 и 6, у универијационом и мултиваријационом контексту, осигурана је додатна верификација практичне значајности развијеног композитног показатеља и, на њему засноване, *ИЕР* класификације *ЈЛС*-а. Наиме, резултати *MANOVA* тестова и пратећих универијационих анализа потврђују да удео укупног узорачког варијабилитета оригиналних показатеља економске развијености од „свега“ 56%, објашњен и обухваћен креираним композитним индексом *ИЕР*, представља „више него довољно“ квалитетну основу за класификацију *ЈЛС*-а у одговарајуће групе, будући да исте верно одражавају разлике (утврђене на бази *ИЕР* вредности) у достигнутом степену економске развијености *ЈЛС*-а, и из угла оригиналних показатеља.

5.1.7. Интерпретација резултата истраживања

Интерпретација предложене, а резултатима *MANOVA* и анализе груписања потврђене, *ИЕР* класификације *ЈЛС*-а у Републици Србији према „измереном“ степену економске развијености (Табела 5.1.18), заснована је на анализи оригиналних вредности економских показатеља, коришћених у поступку развоја композитног индекса *ИЕР*, по појединачним групама *ЈЛС*-а. Сумарни приказ квантитативних карактеристика издвојених група *ЈЛС*-а (Табела 5.1.39) подржан је одговарајућим картографским приказом резултата *ИЕР* класификације (Слика 5.1.21) и мултиваријационим методима визуелизације у форми *Chernoff*-ових лица (Слика 5.1.22) и *Andrews*-ових кривих (Слика 5.1.23 и 5.1.24).

Група 6, својом структуром обухвата 6 општина (или, ≈ 4% укупног броја посматраних *ЈЛС*-а), чије су вредности *ИЕР* на нивоу испод 60% медијалне вредности (m_e = 3,25), односно мање од *ИЕР* = 1,95, и то: Тутин (Рашки округ), Житорађа (Топлички округ), Трговиште (Пчињски округ), Медвеђа, Бојник и Лебане (Јабланички округ). Изузев Тутина који припада Региону Шумадије и Западне Србије, свих пет преосталих општина у

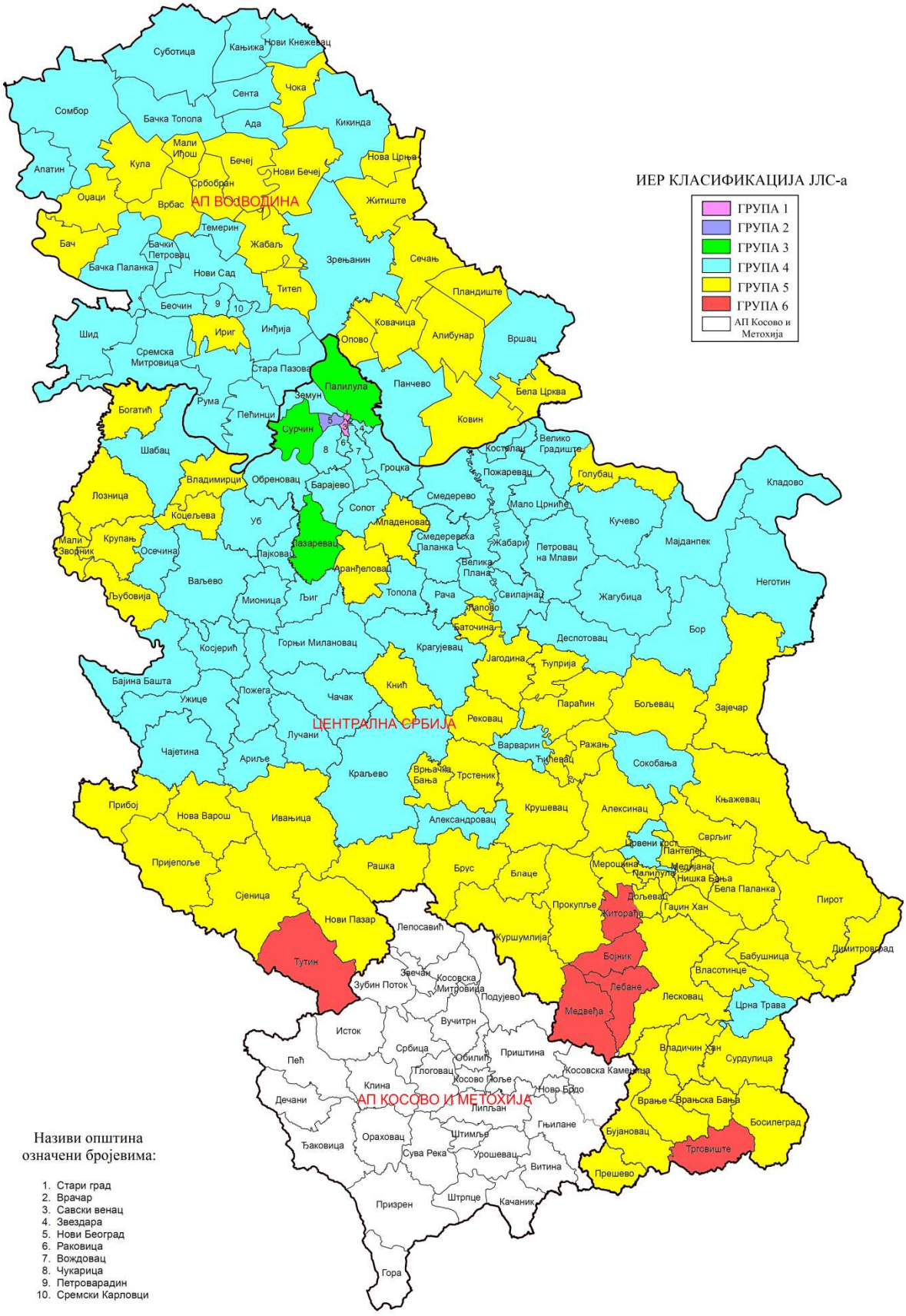
саставу је Региона Јужне и Источне Србије. Посматрано из угла вредности коришћених економских показатеља, приближно двоструко већи број незапослених лица на 1000 ст., за 30% мањи број предузећа на 1000 становника и исто толико нижа стопа запослености и оквирно 10 % нижи износ просечне зараде по запосленом, у просеку, у поређењу са кореспондентним вредностима републичког просека (Табела 5.1.39), недвосмилено указују на забрињавајуће стање привреде и целокупне економске ситуације у општинама унутар ове групе. У том смислу, додељен јој је следећи описни назив: *Група 6 – изразито низак степен економске развијености*. Оправданост додељеног назива потврђује и запажање да су и максималне вредности појединачних показатеља општина у *Групи 6* на нивоу од свега 4–5 % (у случају *стопе запослености* и *просечне зараде по запосленом*) или 12% (*број предузећа на 1000 ст.*) до запањујућих 118% изнад вредности републичког просека, али у случају променљиве *број незапослених на 1000 становника*. У контексту последње променљиве, чак је и минимална вредност овог показатеља приближно 70% изнад просечног броја незапослених лица на нивоу 165 *ЈЛС*-а. Такође, рацио ≈ 4,7 : 1, добијен поређењем просечних вредности *ИЕР* на нивоу *Групе 1* и *Групе 6*, јасно указује на озбиљне размере присутних диспаритета у погледу економске развијености јединица локалне самоуправе у Републици Србији. Посматрано по појединачним економским показатељима, однос просечних вредности ове две групе варира од 1,71 : 1 (из угла просечне зарада по запосленом), односно 5,5 : 1 и 0,26 : 1 (у случају броја предузећа и незапослених лица на 1000 становника, респективно) до 12,4 : 1 (у погледу стопе запослености), наравно, у корист општина у саставу *Групе 1*. Озбиљност забележеног стања, из угла економске развијености, у општинама *Групе 6*, можда најбоље илуструју њихови „тужни“ изрази *Chernoff*-ових лица на „ивици плача“, односно, спуштене обрве, веома широк нос и мала, на доле окренута, уста (Слика 5.1.22).

Табела 5.1.39. *Просек и min-max вредности економских показатеља по групама ЈЛС-а*

Показатељи → / Класификација		Број предузећа на 1000 ст.		Стопа запослености (у %)		Број незапослених на 1000 ст.		Просечна зарада по запосленом		Просек *ИЕР* по групи
групе	Број ЈЛС	$\bar{x}_1$	*min–max*	$\bar{x}_2$	*min–max*	$\bar{x}_3$	*min–max*	$\bar{x}_4$	*min–max*	$\overline{ИЕР}$
1	2	161,00	141-181	314,86	218-412	58,50	54-63	56,17	49-63	8,07
2	2	96,50	79-114	93,47	88-99	53,50	52-55	62,87	60-66	5,98
3	3	41,67	34-54	48,24	41-52	56,57	52-61	65,75	60-73	5,05
4	75	43,40	21-174	37,56	20-96	81,05	24-148	39,26	25-62	3,77
5	77	33,42	19-58	29,28	17-39	140,92	79-207	33,47	26-42	2,72
6	6	29,33	19-45	25,47	17-39	223,00	190-247	32,93	27-39	1,72
Σ	n = 165	40,27	/	37,49	/	113,10	/	37,30	/	3,31

Извор: Ауторов прорачун коришћењем програма *Excel*

Група 5 обухвата 77 градова / општина (или, ≈ 47% укупног броја посматраних *ЈЛС*-а), за које су утврђене вредности *ИЕР* на нивоу од 60% до 100% медијалне вредности, односно у интервалу $1{,}95 \leq ИЕР < 3{,}25$. Насупрот ситуацији у претходној, *Групи 6*, коју је карактерисало доминантно присуство општина у саставу Региона Јужне и Источне Србије, у овој групи присутан је прилично уједначен распоред општина по припадајућим територијалним јединицама нивоа *НСТЈ*-2, односно: Регион Војводине – 21 *ЈЛС*-а (или ≈ 27%), Регион Шумадије и Западне Србије – 27 *ЈЛС*-а (или ≈ 35%), Регион Јужне и Источне Србије – 28 градова / општина (или ≈ 36%) и Београдски регион – само једна општина (Младеновац).

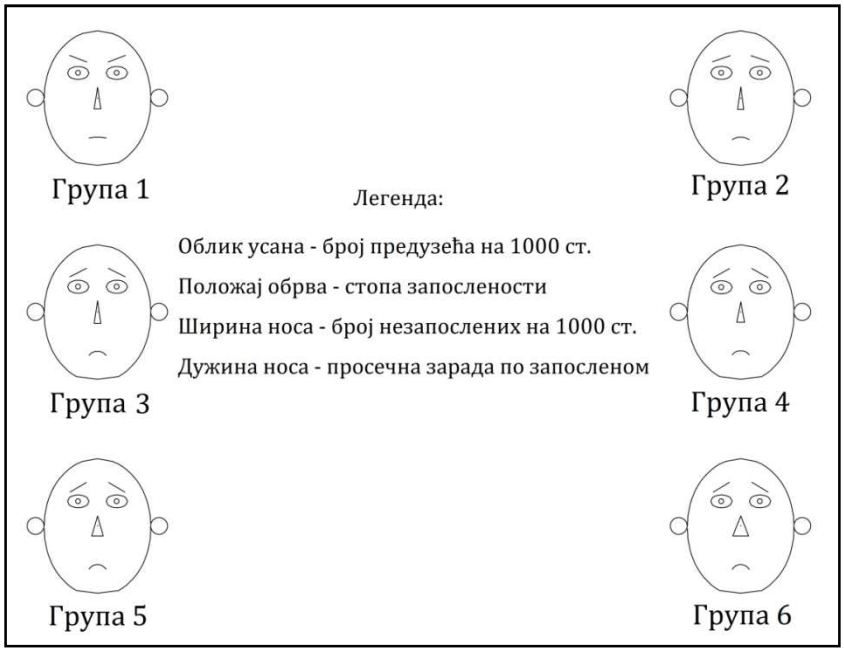


Слика 5.1.21. *Картографски приказ ИЕР класификације ЈЛС-а у Републици Србији*

Извор: Ауторов визуелни приказ

На основу Слике 5.1.21, може се такође уочити да је претежно реч о општинама које су, у односу на остале *ЈЛС*-а у саставу припадајућих региона, генерално, више „удаљене“ од Београдског региона и његових општина. У односу на републички просек, просечне вредности свих економских показатеља на нивоу групе, очекивано, показују негативна одступања, при чему су иста знатно мањих размера него у случају претходне *Групе 6*, нарочито код променљиве – *број незапослених на 1000 ст*.

Резултати поређења са просечним вредностима наведених показатеља на нивоу *Групе 6*, указују да се општине у саставу *Групе 5*, сходно исходу *ИЕР* класификације, одликују у извесној мери повољнијом економском ситуацијом, о чему сведочи и за 14 – 15% већи број предузећа на 1000 становника и већа стопа запослености, односно за 63% мањи број незапослених лица на 1000 становника. Изнето запажање додатно је потврђено односом просечних вредности *ИЕР* на нивоу *Група 1* и *5* који износи, приближно, 3 : 1, у корист прве групе. У том смислу, *Групи 5* додељен је следећи описни назив: *Група 5 – степен економске развијености испод просека*. Такође, будући да обухвата највећи број *ЈЛС*-а у поређењу са осталим групама, може се рећи да *Група 5* представља модалну групу у статистичком смислу речи. Сагласно презентираним карактеристикама ове групе, изрази *Chernoff*-ovih *lica* не показују значајније разлике у односу на лице *Групе 6*, изузев очигледно мање ширине носа.

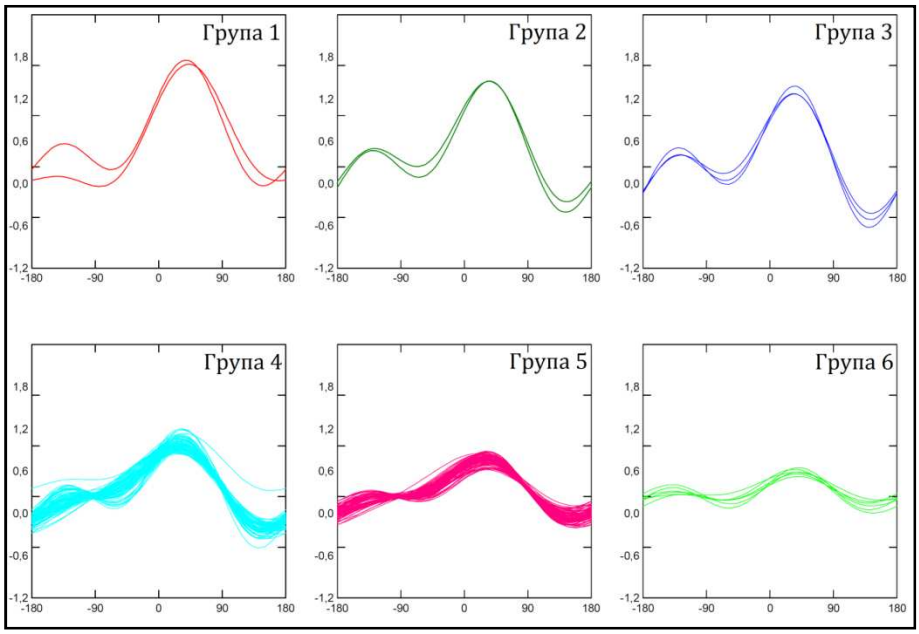


Слика 5.1.22. *Chernoff-ова лица издвојених група ЈЛС-а према ИЕР класификацији*

Извор: Ауторов визуелни приказ коришћењем програма *SYSTAT 13.1*.

Група 4 обухвата 75 *ЈЛС*-а (или, ≈ 45% укупне величине узорка) чије се вредности *ИЕР* налазе на нивоу од 100% до 150% медијалне вредности, односно у интервалу $3{,}25 \leq$ *ИЕР* $< 4{,}875$. Посматрано из угла распореда *Групом 4* обухваћених општина по конкретним регионима у Републици Србији, уочава се, на први поглед, слична тенденција као и у случају претходно описане *Групе 5*, односно присуство релативно равномерне алокације. Прецизније: Регион Војводине – 25 *ЈЛС*-а (или ≈ 33%), Регион Шумадије и Западне Србије – 23 *ЈЛС*-а (или ≈ 31%), Регион Јужне и Источне Србије – 18 градова / општина (или ≈ 24%, при чему њих ≈ 80% припада територији Источне Србије) и коначно,

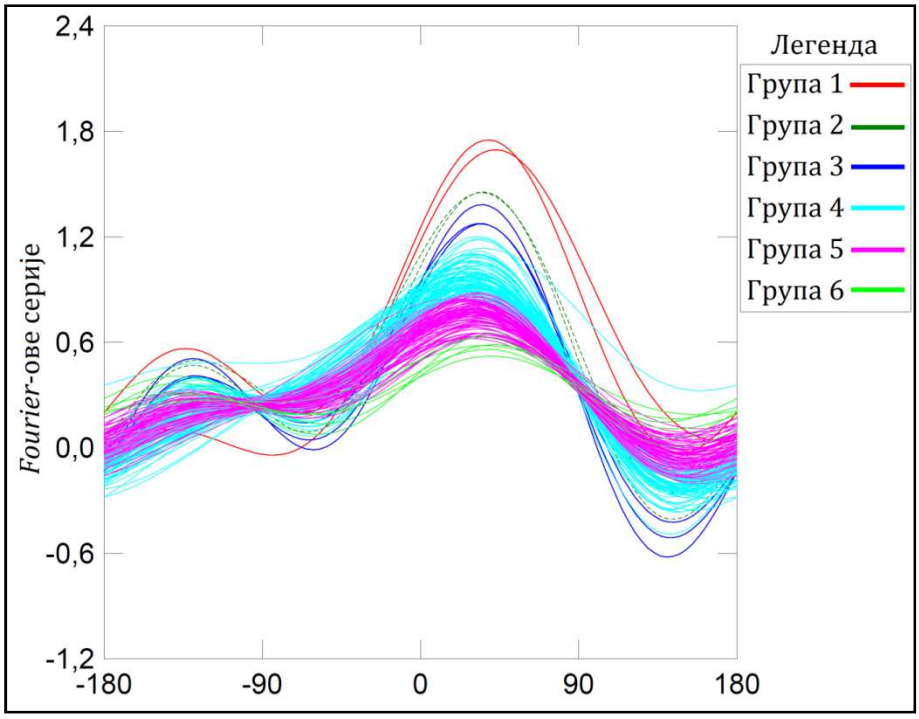
9 београдских градских општина (од укупно 17). Међутим, детаљнијим посматрањем може се уочити, генерално у структури ове групе, повећање броја и учешћа општина које су у саставу Београдског и Региона Војводине, у односу на ситуацију у *Групи 5*. Такође, евидентан је и пораст броја општина из Источне Србије у односу на број општина из Јужне Србије, у структури кореспондентног Региона Јужне и Источне Србије на нивоу ове групе, у односу на обрнуту ситуацију у случају претходне групе (Слика 5.1.21). Готово идентичне (у случају *стопе запослености*) или благо веће (за променљиве *број предузећа на 1000 становника* и *просечна зарада по запосленом*) просечне вредности на нивоу групе у односу на кореспондентне републичке просеке за појединачне економске показатеље, сугеришу додељивање следећег описног назива овој групи *ЈЛС*-а: *Група 4 – просечан степен економске развијености*. Очигледније одступање, у позитивном смислу, забележено је само у случају променљиве *број незапослених на 1000 становника*, чија је просечна вредност, приближно, за 30% мања од републичког просека. Такође, сходно броју обухваћених општина, интервалу *ИЕР* вредности који је дефинише, као и односу просечних вредности економских показатеља на нивоу *Групе 4* и републичких просека, за ову групу се може рећи да представља „модалну“, „медијалну“, али и „просечну групу“, у статистичком смислу. Однос просечних вредности *ИЕР* на нивоу *Групе 1* и *4* који износи, приближно, 2 : 1, у корист прве групе, наравно, указује на присуство израженог јаза у погледу економске развијености на нивоу компарираних група *ЈЛС*-а.



Слика 5.1.23. *Појединачни прикази Andrews-ове кривих по издвојеним групама ЈЛС-а*
Извор: Ауторов визуелни приказ коришћењем програма *SYSTAT 13.1*.

Преостале три групе, *Група 3*, *Група 2* и *Група 1*, заједно обухватају свега 7 од укупно 165 класификацијом обухваћених *ЈЛС*-а, или приближно 4% величине узорка, слично *Групи 6*. Реч је искључиво о општинама на територији града Београда, и то: Савски Венац, Стари Град (*Група 1*), Врачар, Нови Београд (*Група 2*) и Сурчин, Палилула и Лазаревац (*Група 3*). Одликују се вредностима композитног индекса *ИЕР* које су изнад 150% медијалне вредности (3,25), односно, које су веће од 4,875 ≤ *ИЕР*. У том смислу, може се констатовати да је достигнути, односно индексом *ИЕР* „измерени“, степен

економске развијености ових општина евидентно већи у поређењу са члановима претходно описаних група. Овакав исход *ИЕР* класификације, у погледу састава група које се одликују најповољнијим вредностима економских показатеља, је у потпуности очекиван, будући да су управо ове општине, током поступка припреме података и провере статистичких претпоставки, идентификоване углавном као мултиваријационе и / или униваријационе нетипичне опсервације. Резултати међусобног поређења њихових (на нивоу припадајућих група) просечних вредности појединачних економских показатеља, као и са кореспондентним републичким просецима, определили су им следеће описне називе: *Група 3 – степен економске развијености изнад просека*; *Група 2 – висок степен економске развијености* и *Група 1 – изразито висок степен економске развијености*.



Слика 5.1.24. *Здружени приказ Andrews-ових кривих по издвојеним групама ЈЛС-а*
Извор: Ауторов визуелни приказ коришћењем програма *SYSTAT 13.1*.

Сличност кључних одлика ове три групе у погледу вредности разматраних економских показатеља најбоље илуструју, њима кореспондентне, *Andrews*-ове криве приказане на Слици 5.1.23, а њихову „позицију“ у односу на остале, претходно описане групе, из угла степена економске развијености Слика 5.1.24. „Препотентни“, „надмени“ израз *Chernoff*-овог лица које одговара *Групи 1*, недвосмислено указује на супериорност ове, нажалост само, две (београдске) општине са аспекта достигнутог степена економске развијености у односу на преосталих 158 градова / општина (или 96% анализираних *ЈЛС*-а) обухваћених групама изразито ниске, испод просечне и просечне економске развијености.

Генерално, на основу извршене интерпретације кључних карактеристика појединачних група *ЈЛС*-а, издвојених предложеном *ИЕР* класификацијом, може се, нажалост, потврдити присуство изражених регионалних неједнакости и асиметричности међу анализираним територијалним јединицама у Републици Србији са аспекта достигнутог и / или „измереног“ степена економске развијености, у 2015. години. Изнету констатацију најбоље потврђује рацио вредности композитног показатеља *ИЕР*

најразвијеније (Савски Венац, *ИЕР* = 8,31) и најнеразвијеније општине (Лебане, *ИЕР* = 1,41), који износи 5,89 : 1, уз напомену да се вредности *ИЕР* крећу у интервалу од 1 до 10. Додатна потврда размера поменутих диспаритета обезбеђена је израчунатим односима просечних вредности *ИЕР* на нивоу парова појединачних група *ЈЛС*-а, представљеним у Табели 5.1.40.

Табела 5.1.40. *Однос просечних вредности ИЕР па нивоу парова група ЈЛС-а*

Просек ИЕР	Ознака групе	Група 1	Група 2	Група 3	Група 4	Група 5	Група 6
8,07	Група 1	1 : 1	1,35 : 1	1,60 : 1	2,14 : 1	2,97 : 1	4,69 : 1
5,98	Група 2	0,74 : 1	1 : 1	1,18 : 1	1,59 : 1	2,20 : 1	3,48 : 1
5,05	Група 3	0,63 : 1	0,84 : 1	1 : 1	1,34 : 1	1,86 : 1	2,94 : 1
3,77	Група 4	0,47 : 1	0,63 : 1	0,75 : 1	1 : 1	1,39 : 1	2,19 : 1
2,72	Група 5	0,34 : 1	0,45 : 1	0,54 : 1	0,72 : 1	1 : 1	1,58 : 1
1,72	Група 6	0,21 : 1	0,29 : 1	0,34 : 1	0,46 : 1	0,63 : 1	1 : 1

Извор: Ауторов прорачун коришћењем програма *Excel*

У начелу, спроведена класификација *ЈЛС*-а потврђује често навођену констатацију о присуству изражене регионалне и унутаррегионалне поларизације у Републици Србији, примарно на релацији „развијени север – неразвијени југ“. Наиме, приближно 91% општина у саставу Јужне Србије припада *Групи 5* или *6* (односно групи изразито ниске и испод просечне економске развијености), док је тај проценат нешто нижи, али и даље довољно висок, у случају посматрања свих општина Региона Јужне и Источне Србије и износи ≈ 65%. С друге стране, приближно 55% општина Региона Војводине распоређено је унутар *Групе 4*, док су преостале *ЈЛС*-а чланови *Групе 5*, али претежно са вредностима *ИЕР* изнад кореспондентног просека за ту групу. Такође, као што је претходно истакнуто, 9 од 17 београдских општина формирају *Групе 1*, *2* и *3*, док су преостале доминантно присутне на водећим позицијама у *Групи 4* (Табела 5.1.18).

Коначно, важно је нагласити да упоредивост презентираних резултата мерења степена економске развијености *ЈЛС*-а у Републици Србији и њихове класификације са резултатима претходно спроведених истраживања сличног карактера и циљева није могућа, услед изразито присутних разлика у погледу просторног и временског обухвата података, избора развојних димензија и њихових карактеристичних појединачних показатеља, као и коришћеног методолошког приступа, како из угла примењених метода тако и самог карактера анализе (униваријациони или мултиваријациони приступ). Независно од претходног става, предложена класификација *ЈЛС*-а према степену економске развијености може послужити као погодна основа за даљу и „дубљу“ анализу стања и тенденција у погледу осталих важних димензија регионалне развијености и, сходно томе, за извођење закључака о њиховој међусобној зависности.

5.2. РАЗВОЈ МУЛТИВАРИЈАЦИОНОГ МОДЕЛА ЗА КЛАСИФИКАЦИЈУ ОПШТИНА У РЕПУБЛИЦИ СРБИЈИ ПРМА СТЕПЕНУ ЕКОНОМСКЕ РАЗВИЈЕНОСТИ

Као логичан наставак емпиријског истраживања реализованог у оквиру Поглавља 5.1, у овом Поглављу демонстриран је поступак извођења класификационих (предиктивних) правила намењених разврставању анализираних *ЈЛС*-а у једну од, *ИЕР* класификацијом претходно предложених, категорија (група) према степену економске развијености. У том смислу, конципиран је и приказан иновативни методолошки оквир истраживања заснован на примени дискриминационе анализе у функцији развоја мултиваријационог класификационог модела и детаљној евалуацији његове предиктивне прецизности.

5.2.1. Дефинисање истраживачког проблема

У контексту дефинисаног предмета и основног циља дисертације, у фокусу овог дела емпиријског истраживања је испитивање и презентовање могућности комбиноване примене дискриминационе анализе и пратећих метода за припрему мултиваријационих података и оцену статистичке и практичне значајности изведених дискриминационих функција и резултирајућих класификационих правила.

Полазећи од наведеног, а сагласно дефинисаном сету посебних циљева дисертације, опредељен је и циљ овог дела емпиријског истраживања, формулисан на следећи начин: креирање мултиваријационог класификационог модела намењеног за предикцију групне припадности или разврставање анализираних *ЈЛС*-а у одговарајуће интерно хомогене и екстерно хетерогене групе (категорије), на основу припадајућих вредности одабраних показатеља степена економске развијености *ЈЛС*-а.

Изложени предмет и дефинисани циљ у овом емпиријском делу су у непосредној вези са другом посебном хипотезом дисертације, формулисане на следећи начин:

Хипотеза 2: Примена комбинације одабраних метода мултиваријационе анализе омогућава развој одговарајућег класификационог статистичког модела који се може користити за прецизно и објективно разврставање посматраних територијалних јединица у одговарајуће интерно хомогене и екстерно хетерогене групе према достигнутом степену економске развијености.

5.2.2. Методолошки оквир креирања мултиваријационог класификационог модела

За потребе реализације претходно дефинисаног циља и проверу формулисане хипотезе, дизајниран је концептуално-методолошки оквир истраживања, заснован, у начелу, на идентичним (кључним) корацима као и претходни део истраживања, уз уважавање одговарајућих специфичности које директно произилазе из примењене мултиваријационе методе, конкретно, дискриминационе анализе. Свеобухватан преглед планираних и реализованих активности обухваћених структуром концептуално-методолошког оквира илустрован је на Слици 5.2.1.

Слика 5.2.1. *Шематски приказ концептуално-методолошког оквира истраживања*
Извор: Ауторов визуелни приказ

При реализацији овог дела истраживања, полазећи од већ прецизираног просторног и временског обухвата података, избор независних променљивих извршен је из иницијално дефинисаног скупа променљивих (Табела 5.1.1) коришћених у поступку креирања *ИЕР* композитног индекса економске развијености и, на њему засноване, *ИЕР* класификације. Заправо, од полазне четири, издвојене су, и за потребе спровођења дискриминационе анализе означене као независне, следеће три променљиве: *Број предузећа на 1000 становника* (X_1), *Стопа запослености* (X_2) и *Број незапослених на 1000 становника* (X_3). Одлука о редукцији броја коришћених економских показатеља, односно, изостављању променљиве *Просечна зарада по запосленом* (X_4), донета је на основу резултата *MANOVA* анализе и *post hoc* процедуре (Табела 5.1.38), као и препорука у

погледу броја независних променљивих (елаборираних у Одељку 3.2.2). Прецизније, накнадним тумачењем резултата *Tukey*-овог *HSD* теста, установљено је да се променљиве X_1 и X_4 одликују сличном дискриминационом снагом, односно доприносом у раздвајању анализираних група *ЈЛС*-а, будући да су обе означене као слаби дискриминатори групе 5 и групе 6 (предложене *ИЕР* класификацијом), насупрот показатељима X_2 и X_3 код којих је потврђена статистичка значајност разлика њихових просечних вредности између свих могућих парова разматраних група *ЈЛС*-а. Сходно наведеном, а у циљу ублажавања комплексности оцењеног дискриминационог модела и његове интерпретације, донета је одлука о изостављању променљиве X_4 из даље анализе, будући да се одликује знатно већим износима резултирајућих *p*-вредности при спровођењу поменутог теста, у односу на кореспондентне износе добијене на нивоу променљиве X_1.

С друге стране, улога зависне променљиве, у развоју ДА модела и класификационих правила, додељена је категоријској променљивој под називом *Степен економске развијености*, са следећим модалитетима: Група 1 (*изразито висок*), Група 2 (*висок*), Група 3 (*изнад просека*), Група 4 (*просечан*), Група 5 (*испод просека*) и Група 6 (*изразито низак*). Реч је о „новоформираној" променљивој (у ознаци, Y_k, за k = 1, 2,..., 6, где k означава редни број издвојених модалитета), која одражава структуру предложене класификације *ЈЛС*-а спроведене на бази вредности *ИЕР* композитног показатеља (Одељак 5.1.7), као једног од резултата претходног дела истраживања (Поглавље 5.1).

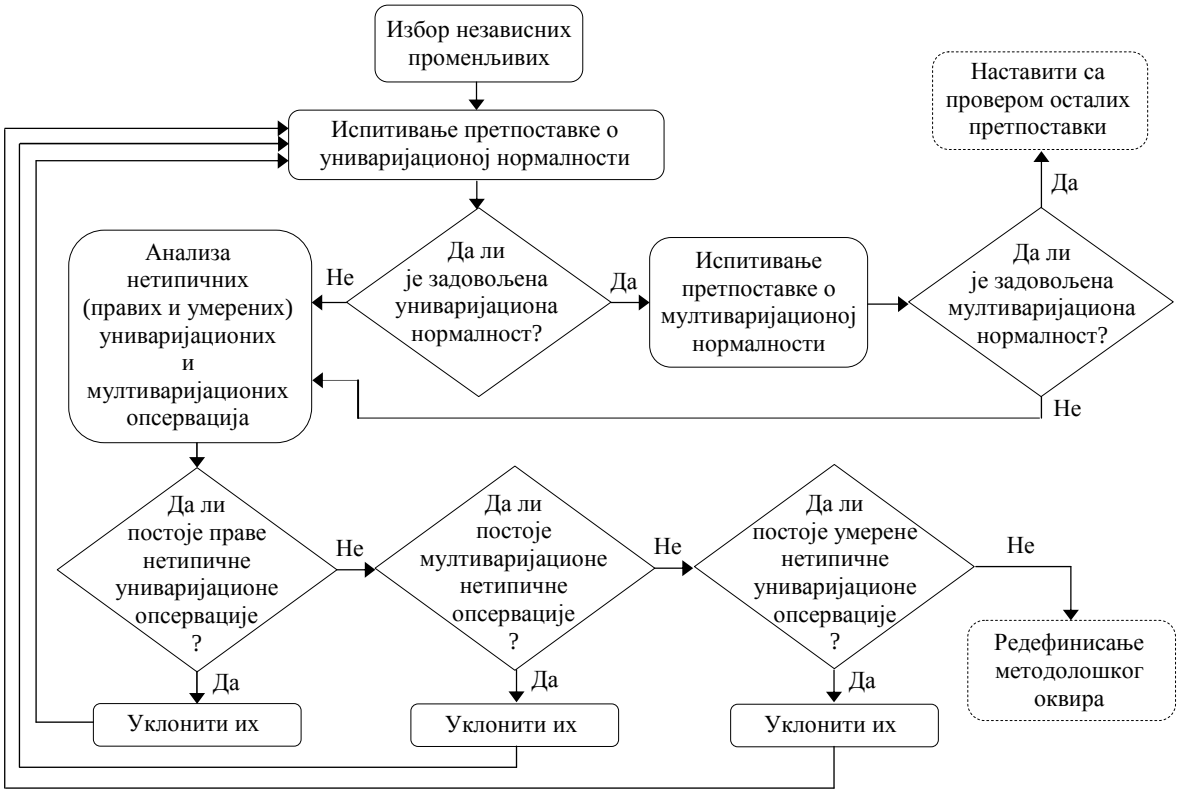
5.2.3. Испитивање претпоставки за примену дискриминационе анализе

Полазећи од презентираног методолошког оквира на којем се заснива реализација овог дела емпиријског истраживања, а у складу са излагањем садржаним у Одељку 3.2.2, у наставку текста представљени су резултати детаљно спроведене провере испуњености статистичких претпоставки на којима се заснива валидна имплементација ДА и развој статистички значајног ДА модела, као основе за извођење класификационих правила намењених одређивању индивидуалне припадности *ЈЛС*-а конкретној категорији према достигнутом нивоу економске развијености, односно једној од претходно дефинисаних група на бази предложене *ИЕР* класификације.

Како би се избегла употреба трансформисаних вредности економских показатеља за извођење дискриминационог модела, при тестирању универијационе и мултиваријационе нормалности кореспондентних распореда није коришћена *Box*-*Cox*-ова трансформациона процедура за ублажавање и / или елиминисање нарушености наведених претпоставки. Због изражене комплексности и рачунске интензивности спроведеног (комплементарног) поступка провере, презентирани су само коначни резултати (ефекти) њиме обухваћених аналитичких корака, илустрованих одговарајућим дијаграмом тока на Слици 5.2.2.

Крајњи резултати тестирања статистичких хипотеза о нормалности универијационог и мултиваријационог распореда анализираних променљивих, представљени Табелама 5.2.1 и 5.2.2 респективно, уз ризик грешке *I* врсте α = 0,05, недвосмислено сугеришу да не постоји довољно емпиријских доказа за одбацивање кореспондентних H_0, односно претпоставки о универијационој нормалности и нормалности заједничког распореда анализиране три променљиве на нивоу популације, будући да су *p*-вредности, утврђене за одговарајуће статистике коришћених (универијационих и мултиваријационих) тестова,

веће од постављеног нивоа значајности теста, α. Другим речима, уз дати ризик грешке, може се закључити да је претпоставка о униваријационој и мултиваријационој нормалности распореда мултиваријационих опсервација на нивоу популације, испуњена.



Слика 5.2.2. *Итеративни поступак провере претпоставки о нормалости распореда*
Извор: Ауторов визуелни приказ

Табела 5.2.1. *Резултати тестирања хипотеза о униваријационој нормалности распореда променљивих (n=135)*

Вар.	*Anderson-Darling* тест нормалности		*Shapiro-Wilk* тест нормалности		*Kolmogorov-Smirnov* тест нормалности		*Z* тест симетричности ($Z_{skewness}$)		*Z* тест заобљености ($Z_{kurtosis}$)	
	статистика	*p*-вред.	статистика	*p*-вред.	статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
X_1	0,631	0,100	0,982	0,080	0,077	0,049	0,356	H_0	-1,060	H_0
X_2	0,394	0,374	0,987	0,240	0,044	0,200	-0,640	H_0	-0,579	H_0
X_3	0,341	0,495	0,989	0,386	0,052	0,200	1,338	H_0	-0,610	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*, *EduStat 4.05* и *QI Macros for Excel*

Табела 5.2.2. *Резултати тестирања хипотеза о нормалности мултиваријационог распореда променљивих (n=135)*

Мултиваријациони тестови нормалности	Статистика теста	*p*-вредност	Одлука
Mardia тест симетричности	15,977	0,100	H_0
Mardia тест заобљености	-0,686	0,493	H_0
Henze–Zirkler–ов тест	1,002	0,103	H_0

Извор: Ауторов прорачун коришћењем програма *SYSTAT 13.1*.

Као последица спровођења описаног поступка провере претпоставки о нормалности, будући да су идентификоване као мултиваријационе или, пак, униваријационе (праве или умерене) нетипичне опсервације, из полазног узорка (величине, n = 165) искључено је 30 јединица посматрања. При томе, важно је нагласити да у редукованом узорку (величине, n

= 135) није забележено присуство наведених категорија „проблематичних“ опсервација. Посматрано из угла дефинисаних модалитета зависне промењиве – *степен економске развијености*, имплементираним поступком провере (Слика 5.2.2), поред појединачних случајева општина у саставу *Групе 4*, *5* или *6*, очекивано, искључене су све општине распоређене унутар прва три модалитета (односно, *Групе 1*, *2* и *Групе 3*) променљиве Y_k. Полазећи од наведене констатације, распоред 135 задржаних *ЈЛС*-а у узорку према преосталим категоријама економске развијености има следећи изглед: *Група 4* = 57 *ЈЛС*-а, *Група 5* = 74 *ЈЛС*-а и *Група 6*→4 *ЈЛС*-а.

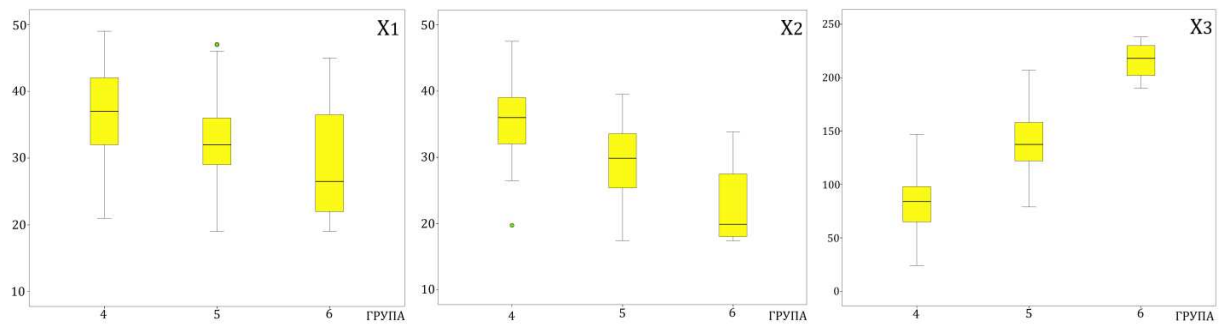
Без обзира на извршену редукцију броја општина, расположиви узорак (n = 135), више него довољно, испуњава искуствене препоруке које се тичу његове величине n, потребне за адекватну имплементацију ДА, будући да садржи 45 пута више опсервација од броја независних променљивих, p, односно: n / p = (45 : 1) > (20 : 1). Посматрано из угла величина појединачних група (категорија) *ЈЛС*-а, *Група 6* не задовољава препоручени минимум од 20 опсервација, међутим, будући да је број општина обухваћен њеном структуром већи од броја независних променљивих (n_6 = 4 > p = 3) може се констатовати да ова група испуњава препоручени апсолутни минимум у погледу њене величине, услед чега је иста задржана у даљој анализи.

Провера претпоставки које се тичу одсуства универијационих нетипичних опсервација и нормалности распореда независних променљивих (X_1, X_2 и X_3) на нивоу појединачних модалитета зависне (категоријске) променљиве (*Група 4*, *5* и *Група 6*) такође је извршена, а потврда њихове испуњености илустрована је на Слици 5.2.3 и представљена у Табели 5.2.3, респективно.

Табела 5.2.3. *Резултати тестирања универијационе нормалности распореда независних променљивих по појединачним групама ЈЛС-а*

Независне променљиве	Ознака групе	Величина групе (n_k)	*Shapiro-Wilk* тест нормалности		Z тест симетричности ($Z_{skewness}$)		Z тест заобљености ($Z_{kurtosis}$)	
			статистика	*p*-вред.	статистика	Одлука	статистика	Одлука
X_1	4	n_4 = 57	0,970	0,167	-0,213	H_0	-0,846	H_0
	5	n_5 = 74	0,986	0,595	0,667	H_0	-0,125	H_0
	6	n_6 = 4	0,905	0,457	1,081	H_0	0,915	H_0
X_2	4	n_4 = 57	0,987	0,809	-0,453	H_0	0,307	H_0
	5	n_5 = 74	0,968	0,060	-1,187	H_0	-1,215	H_0
	6	n_6 = 4	0,802	0,105	1,435	H_0	1,272	H_0
X_3	4	n_4 = 57	0,988	0,857	0,068	H_0	-0,120	H_0
	5	n_5 = 74	0,984	0,499	0,671	H_0	-0,430	H_0
	6	n_6 = 4	0,982	0,911	-0,457	H_0	0,379	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *Excel*.



Слика 5.2.3. *Box-plot дијаграми независних променљивих по групама ЈЛС-а*

Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

Испитивање одрживости претпоставке у погледу линеарне повезаности и одсуства мултиколинеарности између независних променљивих, спроведено је израчунавањем вредности *Pearson*-ових коефицијената просте линеарне корелације између свих парова променљивих и тестирањем њихове статистичке значајности, уз ризик грешке α = 0,05. Резултати тестирања хипотеза о нормалности дводимензионих распореда свих парова независних променљивих представљени су у Табели 5.2.4, а резултати просте линеарне корелационе анализе, у форми одговарајуће корелационе матрице, дати су у Табели 5.2.5.

Табела 5.2.4. *Резултати тестирања статистичких хипотеза о нормалности дводимензионог распореда парова независних променљивих*

Парови независних променљивих	*Mardia* тест симетричности		*Mardia* тест заобљености		*Henze-Zirkler*-ов тест нормалности		Одлука
	статистика	*p*-вредност	статистика	*p*-вредност	статистика	*p*-вредност	
X_1 и X_2	2,321	0,677	-1,602	0,109	0,751	0,424	H_0
X_1 и X_3	3,648	0,456	-0,517	0,605	0,569	0,980	H_0
X_2 и X_3	8,397	0,078	-0,532	0,595	0,947	0,137	H_0

Извор: Ауторов прорачун коришћењем програма *SYSTAT 13.1*.

Табела 5.2.5. *Корелациона матрица анализираних независних променљивих*

Независне променљиве	X_1	X_2	X_3
X_1	1,000	0,371 [0,000]	-0,240 [0,005]
X_2	0,371 [0,000]	1,000	-0,419 [0,000]
X_3	-0,240 [0,005]	-0,419 [0,000]	1,000

Напомена: Вредност у [ ] означава *p*-вредност добијену при тестирању статистичке значајности израчунатих оцена

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.

Израчунате вредности коефицијената просте линеарне корелације (*r*), као и резултати тестирања хипотеза о њиховој статистичкој значајности, сугеришу да између свих парова независних променљивих постоји статистички значајна линеарна веза на нивоу популације, чиме је потврђена испуњеност претпоставке о линеарној повезаности. Такође, у корелационој матрици нису евидентиране вредности коефицијената веће од 0,80 или 0,90 будући да се исте, у апсолутном износу, крећу у интервалу од $|r_{min}|$ = 0,240 до $|r_{max}|$ = 0,419. Сходно наведеном, може се констатовати да међу анализираним показатељима економске развијености *ЈЛС*-а нема високо корелисаних променљивих, чиме је уједно и потврђена испуњеност претпоставке о одсуству мултиколинеарности.

Коначно, провера одрживости претпоставке о хомогености коваријационих матрица на нивоу разматране три групе (односно популација из којих су групе „узорковане“), дефинисана изразом (3.1.14), извршена је применом *Box*-овог *M* теста, а резултати поступка тестирања представљени су у Табели 5.2.6.

Табела 5.2.6. *Резултати примене Box-овог M теста*

	Вредност статистике	Бр. степени слободе		*p*-вред.	Одлука
		v_1	v_2		
Box-ов *M* тест	21,966	-	-	-	-
F апроксимација	1,389	12	252,219	0,171	H_0
χ^2 апроксимација	17,715	12	-	> 0,10	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0* и *Excel*.

Добијена вредност статистике *Box*-овог *M* теста (M = 21,966) и, на бази ње изведене вредности апроксимативне *F* статистике ($F_{(\alpha;\ 12;\ 252{,}219)}$ = 1,389) и χ^2 апроксимације ($\chi^2_{(\alpha;\ 12)}$ = 17,715), уз ниво значајности теста α = 0,05, сугеришу да не постоји довољно емпиријских доказа за одбацивање H_0 претпоставке о хомогености коваријационих матрица три групе мултиваријационих опсервација, будући да су добијене *p*-вредности, за обе апроксимативне статистике, веће од ризика грешке I врсте, α. Изведени закључак потврђен је и резултатима *Levene*-овог теста хомогености варијанси појединачних независних променљивих за издвојене групе на нивоу популација (Табела 5.2.7).

Табела 5.2.7. *Резултати Levene-овог теста једнакости варијанси променљивих*

Зависне променљиве	Вредност *F* статистике	Бр. степени слободе		*p*-вред.	Одлука
		v_1	v_2		
X_1	1,248	2	132	0,290	H_0
X_2	0,207	2	132	0,813	H_0
X_3	1,368	2	132	0,258	H_0

Извор: Ауторов прорачун коришћењем програма *IBM SPSS Statistics 20.0*.

Наиме, за анализиране независне променљиве, није евидентирана статистичка значајност добијених вредности *F* статистике *Levene*-овог теста, будући да су утврђене *p*-вредности у сва три случаја значајно веће од ризика грешке I врсте, α = 0,05, сугеришући да нема довољно емпиријских доказа за одбацивање претпоставке о хомогености варијанси појединачних независних променљивих за издвојене групе, на нивоу популације.

Након извршене провере и обезбеђене потврде испуњености статистичких претпоставки, за потребе екстерне валидације резултата и класификационе прецизности оцењеног модела ДА и на њему заснованих класификационих правила, извршена је подела редукованог узорка (величине 135 *ЈЛС*-а) на (под)узорак за анализу (чија је величина означена симболом, n_a) и (под)узорак за валидацију (величине, n_b), у међусобном односу 60% : 40% у корист (под)узорка за анализу, коришћењем процедуре пропорционално стратификованог случајног узорковања (Табела 5.2.8). При томе, важно је нагласити да, у начелу, оба (под)узорка испуњавају критеријуме у погледу њихове препоручене или минимално прихватљиве величине, детаљно елабориране у Одељку 3.2.2.

Табела 5.2.8. *Структура иницијалног узорка и (под)узорака за анализу и валидацију*

Ознака групе	Укупан узорак (*n*) структура			(под)узорак за анализу структура			(под)узорак за валидацију структура		
	Ознака величине	број *ЈЛС*-а	(у %)	Ознака величине	број *ЈЛС*-а	(у %)	Ознака величине	број *ЈЛС*-а	(у %)
Г-4	n_4	57	42	n_{4a}	33	41	n_{4b}	24	44
Г-5	n_5	74	55	n_{5a}	44	54	n_{5b}	30	56
Г-6	n_6	4	3	n_{6a}	4	5	n_{6b}	(4)[146]	-
Укупно	n	135	100	n_a	81 (60%)	100	n_b	54 (40%)	100,0

Извор: Ауторов прорачун и систематизација табеле коришћењем програма *Excel*.

[146] Будући да садржи свега 4 општине у свом саставу, *Групу 6* није могуће смислено (за потребе ДА) поделити, услед чега је иста у целости прикључена структури (под)узорка за анализу. Разлог представља обезбеђивање адекватне (приближне) процентуалне заступљености све три разматране групе у структури наведеног (под)узорка, који ће бити коришћен за извођење модела ДА. Такође, у циљу обезбеђивања основе за оцену прецизности класификационих правила из угла предикције припадности општина овој групи, њени елементи су распоређени и у састав (под)узорка за валидацију, при чему исти нису евидентирани у збирном реду, због чега је њихов број приказан у загради.

5.2.4. Развој мултиваријационог класификационог модела

Полазећи од формираног (под)узорка за анализу, сачињеног од n_a = 81 мултиваријационе опсервације, распоређене унутар једне од три разматране групе (Г-4, Г-5 и Г-6), коришћењем израза (3.2.3) – (3.2.8), утврђени су елементи (3×3) матрице суме квадрата и узајамних производа одступања унутар (у ознаци, **W**) и између група (у ознаци, **B**), као полазне основе за оцену непознатих параметара дискриминационог модела. Наведене матрице имају следећи изглед:

$$\mathbf{B}=\begin{bmatrix} 495,812 & 672,766 & -7254,437 \\ 672,766 & 919,887 & -9986,504 \\ -7254,437 & -9986,504 & 109046,038 \end{bmatrix} \quad \text{и} \quad \mathbf{W}=\begin{bmatrix} 3419,546 & 671,808 & 1283,564 \\ 671,808 & 2836,412 & 235,499 \\ 1283,564 & 235,499 & 63325,45 \end{bmatrix}. \qquad (5.2.1)$$

Коришћењем елемената матрица **W** и **B**, а утврђена је матрица $\mathbf{W}^{-1}\mathbf{B}$, која гласи:

$$\mathbf{W}^{-1}\mathbf{B}=\begin{bmatrix} 0,14776 & 0,20084 & -2,16896 \\ 0,21202 & 0,29026 & -3,15470 \\ -0,11834 & -0,16285 & 1,77769 \end{bmatrix}. \qquad (5.2.2)$$

У складу са изразом (3.2.9) и (3.2.9а) детерминанта новоформиране матрице [$\mathbf{W}^{-1}\mathbf{B}$ – $\lambda\mathbf{I}$], гласи: $-\lambda^3 + 2{,}21571 \times \lambda^2 - 0{,}00855287 \times \lambda = 0$. Решавањем карактеристичног полинома *III* степена по λ, добијене су следеће две (ненулте) карактеристичне вредности λ квадратне, несиметричне (3×3) матрице $\mathbf{W}^{-1}\mathbf{B}$, и то: λ_1 = 2,2118 и λ_2 = 0,0039. Решавањем хомогеног система једначина дефинисаног изразом (3.2.10) утврђени су елементи карактеристичних вектора $\mathbf{b}_1^*$ и $\mathbf{b}_2^*$, матрице $\mathbf{W}^{-1}\mathbf{B}$, следећег облика:

$$\mathbf{b}_1^{*(\mathrm{T})}=[-1{,}224 \quad -1{,}777 \quad 1] \quad \text{и} \quad \mathbf{b}_2^{*(\mathrm{T})}=[9{,}034 \quad 4{,}327 \quad 1]. \qquad (5.2.3)$$

Нормализацијом наведених карактеристичних вектора, придружених карактеристичним вредностима λ_1 и λ_2 респективно, израчунате су оцењене вредности (нестандардизованих) дискриминационих коефицијената b_j, придружених свакој од три независне променљиве X_j, у структури изведене две дискриминационе функције Z_1 и Z_2, док су кореспондентне вредности оцена параметра одсечка $b_{0(1)}$ и $b_{0(2)}$, утврђене коришћењем израза (3.2.13). Оцењени дискриминациони модел има следећи облик:

$$\begin{aligned} &\textit{Дискриминациона функција } Z_1 \rightarrow z_{1i} = -0,659 - 0,039 \cdot x_{i1} - 0,057 \cdot x_{i2} + 0,032 \cdot x_{i3} \\ &\textit{Дискриминациона функција } Z_2 \rightarrow z_{2i} = -7,213 + 0,116 \cdot x_{i1} + 0,056 \cdot x_{i2} + 0,013 \cdot x_{i3} \end{aligned}. \qquad (5.2.4)$$

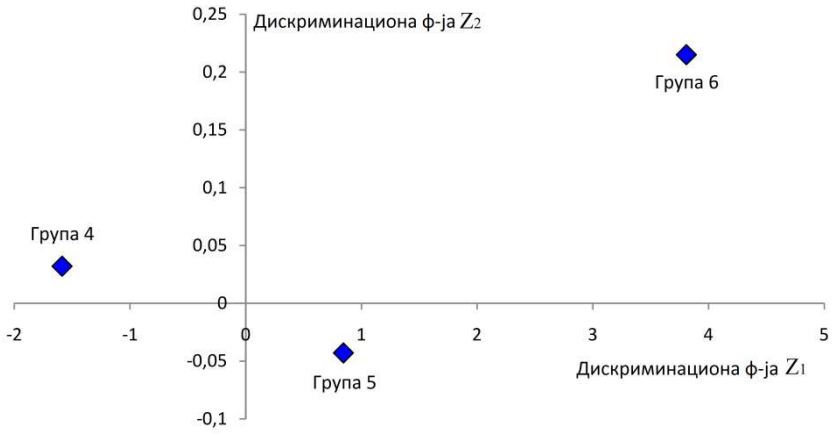
У представљеном дискриминационом моделу, већом дискриминационом снагом одликује се функција Z_1, будући да је иста изведена на основу карактеристичног вектора $\mathbf{b}_1^*$, придруженог знатно већој карактеристичној вредности λ_1 = 2,2118, у односу на другу, Z_2, функцију. У циљу обезбеђивања формалне потврде изнете констатације, извршено је тестирање статистичке значајности дискриминационе моћи појединачних функција у погледу раздвајања разматраних група *ЈЛС*-а у контексту коришћених економских показатеља. Тестирање статистичке значајности дискриминационих функција извршено је путем *Wilks*-ове Λ статистике и, на њој засноване, *Bartlett*-ове χ^2 апроксимације, а резултати су, заједно са кореспондентним H_0 и H_1 хипотезама, приказани у Табели 5.2.9.

Табела 5.2.9. *Сумарни приказ резултата поступка тестирања статистичке значајности дискриминационих функција у формираном ДА моделу*

Дефинисање нулте и алтернативне хипотезе		*Wilks'* Λ статистика
симболички	текстуално	χ^2 апроксимација (број степени слободе) [*p*-вредност]
H_{0_1}: $\lambda_1 = \lambda_2 = 0$	H_{0_1}: Не постоје статистички значајне разлике средина z_i скорова посматраних група ни за једну од ф-ја;	$\Lambda_1 = 0{,}310$
H_{a_1}: $\exists\lambda_m \in \{\lambda_1, \lambda_2\}: \lambda_m \neq 0$	H_{a_1}: Најмање једна функција је потребна за интерпретирање разлика између три групе *ЈЛС*-а;	$\chi^2 = 90{,}145$ ($v = 6$) [*p*-вредност = 0,000]
H_{0_2}: $\lambda_2 = 0$	H_{0_2}: Највише једна функција је потребна за интерпретирање разлика између три групе *ЈЛС*-а;	$\Lambda_2 = 0{,}996$
H_{a_2}: $\lambda_2 \neq 0$	H_{a_2}: Најмање две функције су потребне за интерпретирање разлика између три групе *ЈЛС*-а;	$\chi^2 = 0{,}298$ ($v = 2$) [*p*-вредност = 0,862]

Извор: Ауторов прорачун коришћењем програма *Excel*.

На основу презентираних резултата, уз ризик грешке $\alpha = 0{,}05$, може се закључити да је за описивање разлика између група довољна само једна, и то, дискриминациона функција Z_1, будући да резултирајуће *p*-вредности извршеног тестирања сугеришу само усвајање алтернативне хипотезе H_{a1}. Другим речима, може се прихватити претпоставка којом се тврди да постоји статистички значајна разлика између средина дискриминационих z_{1i} вредности функције Z_1, на нивоу појединачних група *ЈЛС*-а. Будући да статистичка значајност друге функције није потврђена, може се констатовати да иста не доприноси својом дискриминационом снагом у значајнијој мери раздвајању разматраних група и, сходно томе, побољшању резултата класификације *ЈЛС*-а у једну од група, услед чега је донета одлука о њеном искључивању из даље анализе и ДА модела. Изнети закључци у погледу доприноса појединачних функција укупној дискриминационој моћи формираног ДА модела, потврђени су и графичким приказом распореда центроида анализираних група у дводимензионом дискриминационом простору (Слика 5.2.4).



Слика 5.2.4. *Графички приказ распореда центроида појединачних група ЈЛС-а у дводимензионом дискриминационом простору*

Извор: Ауторов визуелни приказ коришћењем програма *Excel*.

Оправданост одлуке о задржавању само дикриминационе функције Z_1 у ДА моделу додатно је потврђена и вредностима одговарајућих показатеља практичне значајности појединачних дискриминационих функција, представљених у Табели 5.2.10. Наиме, за разлику од функције Z_2 која обухвата свега 0,1%, дискриминационом функцијом Z_1 је

објашњено 73,7% укупног варијабилитета присутног између разматраних група у контексту коришћених економских показатеља. Наиме, дискриминациону функцију Z_1 карактерише удео објашњеног међугрупног варијабилитета у укупно обухваћеном уделу узорачког међугрупног варијабилитета изведеним моделом ДА (приближно, 73,8%) у износу од приближно 99,8%, док је остатак, свега 0,2%, обухваћен функцијом Z_2. Дакле, функција Z_1 објашњава приближно 560 пута већу „количину" варијабилитета присутног између дате три групе *ЈЛС*-а у структури зависне променљиве у контексту вредности три независне променљиве, у поређењу са дискриминационим доприносом функције Z_2. Изнета тврдња подржана је и кореспондентним вредностима коефицијента каноничке корелације, утврђеним за појединачне функције. У случају дискриминационе функције Z_1, добијена вредност поменутог коефицијента, $r_1^* = 0{,}8298$, указује да постоји висок ниво квантитативног слагања између модалитета (група) у структури зависне променљиве и варијација њима припадајућих дискриминационих z_{1i} вредности, односно варијација резултирајућих вредности линеарне комбинације три независне променљиве.

Табела 5.2.10. *Показатељи практичне значајности дискриминационих ф-ја у ДА моделу*

	Карактеристична вредност λ	Апсолутни % објашњене варијансе (израз (3.2.26))	Релативни % објашњене варијансе (израз (3.2.27))	Коефицијент каноничке корелације (израз (3.2.24))
Дискр. функција Z_1	2,2118	0,7373	0,9982	$r_1^* = 0{,}8298$
Дискр. функција Z_2	0,0039	0,0013	0,0018	$r_2^* = 0{,}0623$
Модел ДА (израз 5.2.4)	-	73,86 %	100,00 %	-

Извор: Ауторов прорачун коришћењем програма *Excel*.

Сагласно изнетим аргументима, уз поновно истицање да је 99,82% дискриминационе моћи ДА модела (представљеног изразом (5.2.4)) обухваћено првом дискриминационом функцијом, може се констатовати да функција Z_1 представља довољно добру основу за креирање класификационих правила намењених разврставању анализираних општина у једну од дефинисане три категорије према достигнутом степену економске развијености. Након потврђене статистичке и практичне значајности дискриминационе функције Z_1, у наставку је извршена интерпретација кореспондентних вредности нестандардизованих и стандардизованих дискриминационих коефицијената, представљених у Табели 5.2.11, а у циљу сагледавања доприноса појединачних показатеља „измереној" дискриминационој моћи функције Z_1.

Табела 5.2.11. *Нестандардизовани и стандардизовани дискриминациони и коефицијенти каноничке корелације појединачних економских показатеља у функцији* Z_1

Независне променљиве \ Коефицијенти		Нестандардизовани коефицијенти	Стандардизовани коефицијенти	Дискриминациона оптерећења
Број предузећа на 1000 становника	X_1	$b_1 = -0{,}039$	$b_1^{std} = -0{,}259$	$r_{X_1Z_1} = -0{,}253$
Стопа запослености	X_2	$b_2 = -0{,}057$	$b_2^{std} = -0{,}342$	$r_{X_2Z_1} = -0{,}382$
Број незапослених на 1000 становника	X_3	$b_3 = 0{,}032$	$b_3^{std} = 0{,}911$	$r_{X_3Z_1} = 0{,}882$

Извор: Ауторов прорачун коришћењем програма *Excel*.

Вредности појединачних нестандардизованих дискриминационих коефицијената могу се интерпретирати на следећи начин:

✓ Оцењена вредност коефицијента $b_1 = -0{,}039$ показује да повећање броја предузећа на 1000 становника за једно предузеће условљава смањење вредности дискриминационог скора z_{1i} за 0,039, у просеку, под претпоставком да су остале променљиве непромењене.

✓ Оцењена вредност коефицијента $b_2 = -0{,}057$ показује да повећање стопе запослености за 1% условљава смањење вредности дискриминационог скора z_{1i} за 0,057, у просеку, под претпоставком да су вредности осталих променљивих непромењене.

✓ Оцењена вредност $b_3 = 0{,}032$ показује да повећање броја незапослених на 1000 становника за једно незапослено лице, доводи до повећања вредности дискриминационог скора z_{1i} за 0,032, у просеку, при непромењеним вредностима осталих показатеља.

Утврђене коришћењем израза (3.2.14), вредности стандардизованих коефицијената, b_j^{std}, показују да највећи релативни значај, из угла појединачног доприноса дискриминационој моћи функције Z_1, поседује променљива *број незапослених на 1000 ст.* ($b_3^{std} = 0{,}911$) након чега, са знатно мањим доприносом, следе *стопа запослености* ($b_2^{std} = -0{,}342$) и *број предузећа на 1000 становника* ($b_1^{std} = -0{,}259$). Вредности структурних коефицијената корелације, израчунатих коришћењем израза (3.2.15), потврђују извршено рангирање појединачних економских показатеља у погледу њиховог релативног значаја и доприноса дискриминационој моћи функције Z_1, будући да је управо најјачи степен линеарног квантитативног слагања остварен између варијација променљиве *број незапослених на 1000 становника* и дискриминационих z_{1i} скорова функције Z_1, односно, $r_{X_3 Z_1} = 0{,}882$.

Извођење класификационих правила на бази оцењеног модела ДА

Након оцењивања модела ДА и провере статистичке и практичне значајности њиме обухваћених дискриминанти, спроведен је поступак извођења класификационих правила намењених разврставању анализираних општина у једну од три (одабране), *ИЕР* класификацијом предложене, групе на основу припадајућих вредности коришћених економских показатеља, заснован на експлоатацији дискриминационе моћи задржане, Z_1, дискриминанте, будући да је њена статистичка и практична значајност верификована.

Сходно наведеном, у циљу „оптималне“ поделе (једнодимензионог) дискриминационог простора, детерминисаног само функцијом Z_1, након одређивања дискриминационих скорова z_i за сваку појединачну опсервацију узорка за анализу, извршено је одређивање одговарајућих вредности пресека z_{cv}, између појединачних парова, анализом обухваћених, суседних група *ЈЛС*-а, коришћењем израза (3.2.29) и (3.2.30), на следећи начин:

$$\begin{aligned} z_{cv(\text{Г-4}/\text{Г-5})} &= \frac{\overline{z}_{(\text{Г-4})} + \overline{z}_{(\text{Г-5})}}{2} = \frac{-1{,}587 + 0{,}844}{2} = \frac{-0{,}743}{2} = -0{,}3715 \\ z_{cv(\text{Г-5}/\text{Г-6})} &= \frac{n_{(\text{Г-6})} \times \overline{z}_{(\text{Г-5})} + n_{(\text{Г-5})} \times \overline{z}_{(\text{Г-6})}}{n_{(\text{Г-5})} + n_{(\text{Г-6})}} = \frac{4 \times 0{,}844 + 44 \times 3{,}808}{44 + 4} = \frac{170{,}928}{48} = 3{,}561 \end{aligned} \quad . \qquad (5.2.5)$$

Полазећи од утврђених вредности пресека између група Г-4 и Г-5 (у ознаци, $z_{cv\,(\text{Г-4}/\text{Г-5})}$) и група Г-5 и Г-6 (у ознаци, $z_{cv\ (\text{Г-5}\ /\ \text{Г-6})}$) формулисана су дискриминациона (класификациона) правила (или класификатори), која могу бити представљена у следећем облику:

$$\begin{aligned} &\textit{уколико је } z_i < z_{cv(\Gamma\text{-}4\,/\,\Gamma\text{-}5)} \Rightarrow i-\text{та } \textit{општина} \in \textit{група}\ \Gamma-4 \\ &\textit{уколико је } z_{cv(\Gamma\text{-}4\,/\,\Gamma\text{-}5)} < z_i < z_{cv(\Gamma\text{-}5\,/\,\Gamma\text{-}6)} \Rightarrow i-\text{та } \textit{општина} \in \textit{група}\ \Gamma-5\,. \\ &\textit{уколико је } z_{cv(\Gamma\text{-}5\,/\,\Gamma\text{-}6)} < z_i \Rightarrow i-\text{та } \textit{општина} \in \textit{група}\ \Gamma-6 \end{aligned} \qquad (5.2.6)$$

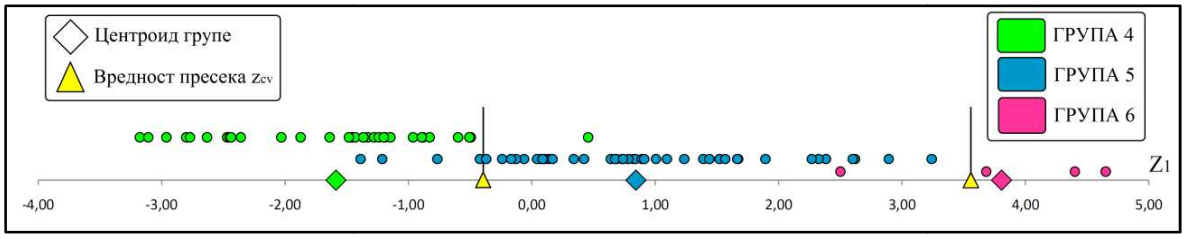
У циљу прелиминарне евалуације предиктивне (класификационе) прецизности изведеног дискриминационог модела и на бази њега формулисаних класификационих правила, извршена је (ре)класификација јединица посматрања (општина) у саставу (под)узорка за анализу. Резултати интерне класификационе анализе, засновани на поређењу стварне групне припадности 81 општине и евидентираних исхода њихове (ре)класификације на бази изведених класификатора, приказани су, у форми одговарајуће класификационе матрице, у Табели 5.2.12 и илустровани на Слици 5.2.5 и 5.2.6.

Интерпретацијом елемената презентиране класификационе матрице може се приметити да је изведеним класификаторима остварена исправна алокација 32 општине у саставу групе Г-4 (од укупно 33, односно 96,97%), 40 општина у саставу групе Г-5 (од укупно 44 општина, односно 90,91%) и, коначно, 3 општине у саставу групе Г-6 (од укупно 4, или 75%). Наиме, од укупно 81 општине у структури (под)узорка за анализу, 75 општина је исправно (ре)класификовано у састав групе којој заиста и припада, док је погрешна класификација забележена у случају свега 6 општина.

Табела 5.2.12. *Резултати (ре)класификације ЈЛС-а у (под)узорку за анализу*

(под)узорак за анализу (n_a = 81)	Ознака групе	Предвиђена припадност општина конкретној групи (резултат класификације)						Укупно	
		Г-4		Г-5		Г-6			
		f_i	у %	f_i	у %	f_i	у %	f_i	у %
Стварна припадност групи	Г-4	32	**96,67**	1	3,03	0	0,00	33	100,0
	Г-5	4	9,09	40	**90,91**	0	0,00	44	100,0
	Г-6	0	0,00	1	25,00	3	**75,00**	4	100,0
Предвиђена величина група		36		42		3		81	100,0

Извор: Ауторов прорачун коришћењем програма *Excel*.



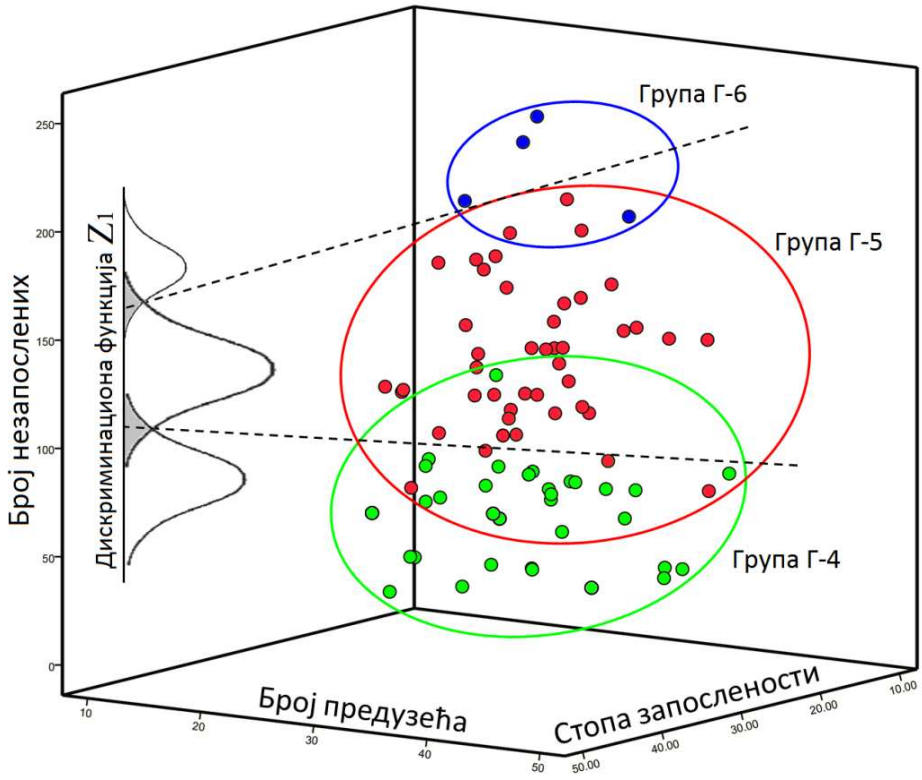
Слика 5.2.5. *Графички приказ (ре)класификације ЈЛС-а у (под)узорку за анализу*
Извор: Ауторов визуелни приказ коришћењем програма *Excel*.

Укупан проценат исправно класификованих општина у (под)узорку за анализу, као одговарајућа (сумарна) мера предиктивних перформанси развијених класификатора и, сходно томе, дискриминационе функције Z_1, износи:

$$\textit{очигледна стопа погодака}\ (ApHR) = \frac{32+40+3}{81}\times 100 = \frac{75}{81}\times 100 = 92,59\%\,. \qquad (5.2.7)$$

Изразито висока пропорција исправних класификација сугерише да се класификациона правила одликују веома добрим предиктивним способностима. Међутим,

будући да иста представља пристрасну оцену стварне пропорције погодака (*AHR*), за потребе извођења статистички валидних закључака у погледу предиктивне прецизности оцењеног модела ДА и изведених класификатора, у наставку је извршена њихова екстерна валидација.



Слика 5.2.6. *3D графички приказ (ре)класификације ЈЛС-а у (под)узорку за анализу*

Извор: Ауторов визуелни приказ коришћењем програма *IBM SPSS Statistics 20.0*.

5.2.5. Оцена валидности креираног мултиваријационог класификационог модела

Будући да третирање *ApHR* оцене као опште оцене стварне пропорције погодака *AHR*, може резултирати извођењем неодрживих закључака о предиктивној прецизности модела ДА, у наставку је спроведен поступак екстерне валидације, заснован на примени методе задржавања, у циљу обезбеђивања непристрасне евалуације квалитета и прецизности (дискриминационих) класификационих правила, формулисаних изразом (5.2.6).

У том смислу, полазећи од извршене поделе иницијалног узорка, величине n = 135 *ЈЛС*-а, представљене Табелом 5.2.8, изведени класификатори примењени су на (под)узорку за валидацију, односно на 58 општина, од којих 54 нису коришћене у поступку оцењивања модела ДА, али чија је припадност конкретној категорији економске развијености позната. Резултати извршене класификације, организовани у форми класификационе матрице, представљени су у Табели 5.2.13.

Табела 5.2.13. *Резултати класификације ЈЛС-а у саставу (под)узорка за валидацију*

(под)узорак за валидацију (n_b = 58)	Ознака групе	Предвиђена припадност општина конкретној групи (резултат класификације)						Укупно	
		Г-4		Г-5		Г-6			
		f_i	у %	f_i	у %	f_i	у %	f_i	у %
Стварна припадност групи	Г-4	20	**83,33**	4	16,67	0	0,00	24	100,00
	Г-5	1	3,33	29	**96,67**	0	0,00	30	100,00
	Г-6	0	0,00	1	25,00	3	**75,00**	4	100,00
Предвиђена величина група		21		34		3		58	100,00

Извор: Ауторов прорачун коришћењем програма *Excel*.

Стопа погодака, односно проценат исправно класификованих општина у (под)узорку за валидацију, у ознаци *ApHR*', износи:

$$ApHR' = \frac{20+29+3}{58}\times 100 = \frac{52}{58}\times 100 = 89,65\% . \qquad (5.2.8)$$

Иако, очекивано, нешто нижа у односу на кореспондентну вредност добијену на основу (под)узорка за анализу (*ApHR* = 92,59%), изразито висока стопа исправних класификација обезбеђена на нивоу (под)узорка за валидацију, потврђује, у начелу, претходно изнет прелиминарни (иницијални) закључак у погледу квалитета и предиктивне прецизности изведених класификационих правила.

У циљу обезбеђивања формалне потврде изведеног закључка, извршено је тестирање статистичке значајности утврђене (непристрасне) оцене стварне пропорције исправних класификација, коришћењем статистике једносмерног-десностраног *Z*-теста, дефинисане изразом (3.2.36). Наведеним изразом прецизиран је и садржај кореспондентних хипотеза, H_0 и H_1. При одређивању хипотетичке вредности пропорције исправних класификација за коју се очекује да ће бити остварена у случају класификације засноване на случајности, коришћен је приступ заснован на критеријуму пропорционалне случајности. Извршени избор детерминисан је присутним разликама у величини појединачних група у структури узорка, али и самим циљем истраживања, усмереним на формирање класификационог модела који обезбеђује максималне стопе погодака на нивоу сваке појединачне групе *ЈЛС*-а, а не само укупне, на нивоу узорка. Сходно наведеном, применом израза (3.2.38), утврђена је хипотетичка вредност пропорције погодака, π_0 = 0,4435. Након провере услова за примену, израчуната је вредност статистике *Z*-теста на следећи начин:

$$\left.\begin{array}{r} n_b = 58 > 30 \\ n\pi_0 = 25,72 > 5 \\ n(1-\pi_0) = 32,28 > 5 \end{array}\right\} Z = \frac{0,8965 - 0,4435}{\sqrt{\frac{0,4435\times(1-0,4435)}{58}}} = \frac{0,453}{0,06523} = 6,94 . \qquad (5.2.9)$$

Будући да је резултирајућа *p*-вредност = 0,000, мања од ризика грешке α = 0,05, може се констатовати да постоји довољно емпиријских доказа за одбацивање H_0 и усвајање H_1 којом се потврђује статистичка значајност оцене стварне пропорције погодака, утврђене на нивоу (под)узорка за валидацију, *ApHR*' = 89,65%. Додатна верификација изведеног закључка, обезбеђена је провером статистичке значајности евидентираних пропорција исправно класификованих општина на нивоу сваке појединачне групе *ЈЛС*-а у структури

(под)узорка за валидацију, коришћењем прилагођене статистике $Z_{(\kappa)}$-теста (израз (3.2.39)), а резултати тестирања представљени су у Табели 5.2.14. Очекивано, у случају све три групе, уз ризик грешке α = 0,05, потврђена је статистичка значајност остварених пропорција исправно класификованих општина.

Табела 5.2.14. *Резултати тестирања хипотеза о статистичкој значајности пропорција погодака појединачних група ЈЛС-а у саставу (под)узорка за валидацију*

Ознака групе	Величина групе	Пропорција погодака у групи	π_0	Статистика $Z_{(\kappa)}$-теста	*p*-вредност	Одлука
Г-4	24	0,8333	0,414	4,17	0,000	H_1
Г-5	30	0,9667	0,517	4,93	0,000	H_1
Г-6	4	0,7500	0,069	5,37	0,000	H_1

Извор: Ауторов прорачун коришћењем програма *Excel*.

Практична значајност добијених (статистички значајних) резултата класификације, спроведене на бази предложеног линеарног класификационог модела (израз (5.2.6)), проверена је израчунавањем и интерпретацијом вредности *Huberty*-јевог критеријума *I*. Коришћењем израза (3.2.40), на нивоу (под)узорка за валидацију, вредност поменутог критеријума износи, $I = 0{,}814$. Добијена вредност показује да се класификацијом општина у једну од разматране три категорије економске развијености, заснованом на изведеним линеарним класификационим правилима (израз (5.2.6)) обезбеђује приближно 81% више исправних исхода класификације (односно, приближно 81% мање грешака у предикцији припадности конкретној групи), него што би то био случај да је њихова класификација спроведена искључиво на бази случајности. Другим речима, побољшање прецизности у предвиђању групне припадности „нових“ општина, који се постиже применом изведених класификационих правила, износи (приближно) 20 исправно класификованих јединица посматрања више, у односу на поступак класификације заснован на случајности. Сходно наведеном, може се констатовати да се предложени мултиваријациони класификациони модел одликује високим нивоом практичне значајности.

Коначно, додатна провера практичне значајности изведених класификатора спроведена је анализом резултата њихове примене на општине, које су током поступка провере статистичких претпоставки искључене из разматраног узорка, услед чињенице да су идентификоване као носиоци универијационих или мултиваријационих нетипичних опсервација. Реч је о посебном (под)узорку сачињеном од 30 општина распоређених унутар следећих категорија зависне променљиве:

- ✓ *Група 1* – 2 општине (Савски Венац и Стари Град);
- ✓ *Група 2* – 2 општине (Врачар и Нови Београд);
- ✓ *Група 3* – 3 општине (Сурчин, Палилула и Лазаревац);
- ✓ *Група 4* – 18 општина (Гроцка, Медијана, Сопот, Нови Сад, Рача, Раковица, Црна трава, Сремски Карловци, Земун, Звездара, Вождовац, Чајетина, Чукарица, Ариље, Стара Пазова, Чачак, Панчево и Ваљево;
- ✓ *Група 5* – 3 општине (Гаџин Хан, Ивањица и Врњачка Бања); и
- ✓ *Група 6* – 2 општине (Медвеђа и Тутин).

Поред 23 *ЈЛС*-а које припадају једној од три, дискриминационим моделом обухваћене, групе (*Група 4*, *5* и *Група 6*), предложена класификациона правила употребљена су и за алокацију преосталих 7 општина, иако исте не припадају некој од наведене три групе. Наиме, њихово укључивање у поступак додатне екстерне валидације заснован је на логици према којој би исте, будући да припадају групама изразито високе, високе и изнад просечне економске развијености, требале да буду распоређене унутар њима најближе расположиве, односно доступне групе, конкретно, Групе 4. Класификациона матрица добијена као резултат екстерног вредновања квалитета и предиктивне прецизности линеарних (дискриминационих) класификатора представљена је у Табели 5.2.15.

Табела 5.2.15. *Резултати класификације 30 општина искључених током фазе припреме података*

посебан (под)узорак (n_c = 30)	Ознака групе	Предвиђена припадност општина датој групи (резултат класификације)						Укупно	
		Г-4		Г-5		Г-6			
		f_i	у %	f_i	у %	f_i	у %	f_i	у %
Стварна припадност групи	Г-4	18 + (7)	**100,00**	0	0,00	0	0,00	18 + (7)	100,0
	Г-5	1	33,33	2	**66,67**	0	0,00	3	100,0
	Г-6	0	0,00	0	0,00	2	**100,00**	2	100,0
Предвиђена величина група		26		2		2		23 + (7)	100,0

Извор: Ауторов прорачун коришћењем програма *Excel*.

Остварени проценат исправних исхода класификације у анализираном (под)узорку од 30 општина, износи 96,67%. Добијена вредност је виша у односу на кореспондентне износе утврђене на нивоу (под)узорка за анализу (92,59%) и (под)узорка за валидацију (89,65%) и недвосмислено потврђује претходно констатован, висок ниво практичне значајности и употребне вредности предложених класификационих правила, заснованих на оцењеном ДА моделу, у контексту предикције групне припадности *ЈЛС*-а, на бази оригиналних вредности коришћених економских показатеља.

5.3. Основна ограничења и могући правци будућег научно-истраживачког рада

Сваком истраживању, у начелу, својствена су извесна ограничења која је потребно узети у обзир приликом интерпретације, разумевања и дубље анализе добијених резултата, али и дефинисања могућих праваца будућег научно-истраживачког рада.

Основна ограничења којима се одликује спроведено емпиријско истраживање у овој дисертацији, директно произлазе из његове, више него очигледне, изразите комплексности присутне како у погледу осмишљеног и реализованог оригиналног методолошког приступа, тако и мултидимензионе природе економског феномена од интереса на којем је имплементација истог демонстрирана. Наиме, полазећи од чињенице да је реч о анализи вођеној подацима, операционализација појединих корака, претежно оних усмерених на обезбеђивање објективне евалуације добијених резултата и ублажавање субјективности истраживача при усмеравању даљег тока анализе, захтева спровођење изразито временски захтевних и комплексних израчунавања и / или улагање значајних финансијских средстава у набавку погодних комерцијалних софтверских решења, будући да иста није у потпуности подржана расположивим софтверским платформама отвореног кода.

Такође, извесна ограничења могу бити идентификована и у погледу ширине развојног обухвата предложеног композитног показатеља за мерење степена економске развијености, али и броја анализом обухваћених појединачних економских показатеља коришћених при његовом развоју. Ограничења наведеног типа, директно су везана за доступност, ажурност и квалитет статистичких података о показатељима економске и других димензија развијености јединица локалне самоуправе у Републици Србији, али и апострофирану методолошку комплексност.

Поред наведеног, извршени избор конкретних показатеља економске развијености, фокус искључиво на економску димензију развоја, а делимично и просторно-временски обухват података, као и примењени методолошки оквир, заснован на интегралној примени широког спектра одабраних мултиваријационих метода, у извесној мери ограничавају могућности за компарацију добијених резултата са кореспондентним исходима других истраживања сличног карактера и циља. Такође, независно од афирмативних резултата детаљно спроведене евалуације валидности добијених резултата класификације и разврставања *ЈЛС*-а у поједине категорије према степену економске развијености, за дубљу анализу и интерпретацију резултата праћену идентификовањем и образложењем узрока позиционирања појединих *ЈЛС*-а унутар неке од издвојених категорија економске развијености, неопходна је адекватна подршка екперата са компетенцијама у домену регионалног развоја.

Коначно, треба напоменути да наведена ограничења нису у функцији оспоравања употребне вредности предложеног (иновативног) мултиваријационог методолошког оквира или, пак, одбацивања добијених резултата истраживања, већ њиховог критичког преиспитивања и дефинисања могућих праваца будућег научно-истраживачког деловања.

Полазећи од наведених ограничења, правци будућих истраживачких настојања примарно ће бити усмерени на репликацију предложеног мултиваријационог методолошког оквира у циљу обезбеђивања континуираног праћења и компарације добијених резултата у погледу достигнутог степена економске развијености *ЈЛС*-а у Републици Србији, током времена. Узимајући у обзир флексибилност и апликативне могућности конципираног методолошког поступка, будућим истраживањима биће обухваћена и његова имплементација за потребе сагледавања развојних потенцијала и ограничења јединица локалних самоуправа или, пак, територијалних јединица виших нивоа територијалне организације у Републици Србији из угла осталих димензија (регионалног) развоја, било у појединачном или интегралном контексту. Проширење просторног обухвата података и померање истраживачког фокуса са територијалне организације Републике Србије на ниво одабраних или, пак, свих европских земаља, односно територијалних јединица различитог *NUTS* нивоа у њиховом саставу, такође представља потенцијални правац истраживачког рада у предстојећем периоду. Разматрање могућности проширења предложеног методолошког поступка кроз укључивање нових метода мултиваријационе анализе представља, у начелу, још једно од могућих истраживачких усмерења аутора дисертације.

З А К Љ У Ч А К

У складу са дефинисаним предметом истраживања, у дисертацији су анализирани кључни концептуално-методолошки аспекти одабране групе мултиваријационих статистичких метода и сагледани њихови апликативни потенцијали у функцији моделирања мултидимензионих економских феномена од интереса. На основу извршене систематизације досадашњих теоријских и практичних сазнања, као и резултата конципираног и реализованог оригиналног емпиријског истраживања, могуће је формулисати извесне закључне констатације, чију валидност треба, превасходно, посматрати из перспективе верификације формулисаних хипотеза и степена остварења дефинисаних циљева дисертације.

Генерално, иако је (р)еволуција у домену компјутерске технологије у значајној мери убрзала и олакшала примену мултиваријационих метода у контексту решавања конкретних проблемских ситуација у скоро свим научним областима и доменима, успешно спровођење мултиваријационе анализе података представља изузетно захтеван подухват који превазилази проблем избора методе(а) и пуке апликације „*user-friendly point-and-click*“ софтверских платформи. Истраживањем је потврђено да, упркос значају коју софтверска подршка има при реализацији фаза процеса мултиваријационе анализе података, ипак, доминанатна улога припада истраживачким знањима и вештинама у спровођењу исте.

Наиме, мултидимензионално моделирање и, консеквентно, креирање статистички валидних модела, као исхода мултиваријационе анализе података, нужно подразумева дефинисање јасног процедуралног оквира у форми процесног модела за идентификовање скривених релација у подацима и оцену могућности њихове генерализације у контексту проблематике у разматрању. Истина, мултиваријациони процесни модел садржи низ универзалних (апликативно неутралних) фаза, којима су обухваћене серије различитих процесних активности, са инкорпорираним повратним спрегама и итерацијама. При томе, свака фаза захтева низ одлука којима се детерминише правац даље анализе, а чије доношење није могуће аутоматизовати и, коначно, спровести без активног учешћа истраживача, уз ублажавање утицаја субјективности кроз употребу објективних статистичких критеријума. Другим речима, број, врста и структура активности које чине садржај сваке фазе зависи не само од комплексности истраживаног проблема већ и општег нивоа знања о проблему, способности и специфичних (статистичких) вештина истраживача.

Сходно наведеном, може се констатовати да у домену начина експлоатације апликативних могућности појединачних метода мултиваријационе анализе или њихове комбинације долази до изражаја креативност истраживача у смислу да ће добро обучен и статистички едукован истраживач применити сложеније методе, и то на методолошки исправан начин, што имплицира богатији садржај активности у свакој фази. У том смислу, полазећи од констатације да свеобухватно разумевање својстава конкретне методе представља битан услов за њену успешну примену, у дисертацији су детаљно анализирана општа и специфична концептуално-методолошких одређења одабраних, у истраживањима разматраног економског феномена најчешће коришћених, мултиваријационих метода и то:

факторске анализе, анализе груписања, *MANOVA* (мултиваријационе анализе варијансе) и дискриминационе анализе. При томе, у оквиру сваке од њих појединачно, елаборирани су циљеви, типови и поступак спровођења, уз јасно разграничење истраживачких околности под којима се њихова примена сматра прикладном и статистички оправданом.

Полазећи од тога да квалитет података директно опредељује резултат мултиваријационе анализе, али могућност генерализације изведених закључака, детаљним образложењем статистичких претпоставки, начина провере њихове испуњености, као и последица нарушености истих у контексту примене конкретних метода и научне заснованости добијених резултата, апострофирана је и потврђена фундаментална улога фазе претпроцесирања у процесу креирања мултиваријационог модела и обезбеђивања његове статистичке и практичне значајности. У вези са истакнутим и верификованим значајем, констатовано је да је реч о најзахтевнијој фази у процесу креирања статистичког модела како са становишта времена, тако и у погледу ангажовања непоходних ресурса (људских, финансијских, информатичких и слично).

Апликативни потенцијал наведене четири мултиваријационе методе демонстриран је у контексту сагледавања степена економске развијености јединица локалне самоуправе у Републици Србији, као мултидимензионог (економског) феномена од интереса. Истраживање проблематике неравномерног регионалног развоја примарно је засновано на имплементацији постојећих или предлагању иновативних квантитативних приступа за мерење степена развијености територијалних јединица у саставу конкретне земље или групе група земаља коришћењем различитих комбинација репрезентативних показатеља једне или више димензија развоја и креирању резултирајућих класификација разматраних територија у релативно хомогене групе. Детаљним прегледом досадашњих истраживања латентног феномена констатовано је присуство изражене методолошке варијететности у погледу низа аналитичких питања, попут просторног-временског обухвата, изабраних показатеља и димензија развијености, као и коришћених метода мултиваријационе анализе, и истовремено потврђена изражена комплексност и атрактивност ове истраживачке нише међу представницима домаће и иностране научне / стручне заједнице. У том смислу, конципирано је и реализовано оригинално емпиријско истраживање сачињено од два међусобно комплементарна дела.

У оквиру првог дела емпиријског истраживања испитана је и презентована могућност статистички валидне интегрисане примене факторске анализе, анализе груписања и мултиваријационе анализе варијансе у функцији развоја композитног показатеља (у ознаци, *ИЕР*) намењеног мерењу степена економске развијености јединица локалне самоуправе у Србији и и евалуације његове практичне значајности. За потребе анализе прикупљени су и коришћени секундарни, у тренутку спровођења исте, последње доступни, подаци следећих показатеља различитих аспеката економске развијености, за сваку од 165 територијалних јединица нивоа *ЈЛС*, и то: Број предузећа (привредних друштава и предузетника) на 1000 становника, Стопа запослености, Број незапослених на 1000 становника и Просечна зарада по запосленом. У методолошком смислу, централни део реализованог истраживања чини објективно и статистички оправдано одређивање пондера појединачних променљивих у структури композитног *ИЕР* индекса, на основу елемента оцењеног факторског модела. Опсежна евалуација практичне значајности креираног *ИЕР* индекса и квалитета, на њему засноване, *ИЕР* класификације *ЈЛС*-а у шест

међусобно искључивих категорија економске развијености, извршена је применом хијерархијске агломеративне процедуре груписања (засноване на коришћењу квадрата Еуклидске мере одстојања између опсервација и методе просечног повезивања), а затим и једнофакторске мултиваријационе анализе варијансе. Недостатак адекватних, довољно сличних, студија за компарацију и извођење закључака у погледу његове употребне вредности, додатно повећава корисност и значај спроведеног поступка мултиваријационе евалуације квалитета предложеног (иновативног) композитног индекса. Посебно се истиче чињеница да су у сваком кораку истраживања, заснованог на предложеном мултиваријационом методолошком оквиру, инкорпорирани елементи статистичког начина размишљања и, сходно току подацима вођене анализе, примењени одговарајући статистички критеријуми, у циљу ублажавања и / или елиминисања субјективности при одлучивању. Генерално, на основу резултата спроведеног истраживања, могуће је констатовати да презентовани методолошки оквир обезбеђује генерисање валидног статистичког модела (у форми композитног индекса) који доприноси успешном и прецизном мерењу достигнутог степена економске развијености *ЈЛС*-а, с тим што је исти могуће имплементирати и за потребе сагледавања развојних потенцијала и ограничења из угла осталих димензија развоја, било у појединачном или интегралном контексту, на истом или вишим нивоима територијалне организације.

У оквиру другог дела реализованог емпиријског истраживања испитана је и презентована могућност комбиноване примене дискриминационе анализе и пратећих метода за припрему мултиваријационих података и оцену статистичке и практичне значајности изведених дискриминационих функција и резултирајућих класификационих правила. У том смислу, конципирањем и реализацијом одговарајућег методолошког оквира, креиран је мултиваријациони класификациони модел намењен за предикцију групне припадности анализираних ЈЛС-а у одговарајуће категорије сличних карактеристика у погледу припадајућих вредности одабраних показатеља степена развијености ЈЛС-а. За потребе анализе, из иницијалног дефинисаног скупа променљивих (коришћених у оквиру претходног дела истраживања), на основу резултата *MANOVA* анализе, *post hoc* процедуре, као и препорука у погледу броја независних променљивих предложених у релевантној литератури, извршен је избор следећих дискриминатор променљивих: Број предузећа (привредних друштава и предузетника) на 1000 становника, Стопа запослености, Број незапослених на 1000 становника. Улога зависне променљиве у развоју дискриминационог модела додељена је категоријској променљивој под називом *Степен економске развијености*, следећих модалитета: Група 1 (*изразито висок*), Група 2 (*висок*), Група 3 (*изнад просека*), Група 4 (*просечан*), Група 5 (*испод просека*) и Група 6 (*изразито низак*). Реч је о „новоформираној“ променљивој која одражава структуру предложене класификације *ЈЛС*-а спроведене на бази вредности *ИЕР* композитног показатеља, као једног од резултата претходно спроведеног дела истраживања. У методолошком смислу, централни део реализованог истраживања обухвата детаљна провера претпоставки на којима се заснива развој дискриминационог модела и извођење класификационих правила на основу статистички значајних дискриминационих функција. Резултатима спроведене екстерне валидације потврђена је статистичка и висока практична значајност забележене предиктивне прецизности креираног класификационог модела, у интервалу од 89,65% до 92,59%. Генерално, на основу резултата спроведеног

истраживања, могуће је констатовати да презентовани методолошки оквир обезбеђује генерисање валидног статистичког класификационог модела (у форми скупа класификационих правила), који омогућава прецизно и објективно разврставање анализираних *ЈЛС*-а у једну од обухваћених категорија економске развијености, утврђених *ИЕР* класификацијом.

Полазећи од изнетих закључака, формулисаних на основу резултата истраживања теоријског материјала и спроведене оригиналне емпиријске студије, са правом се може констатовати да су све три посебне хипотезе потврђене, а које су директно повезане са дефинисаним основним циљем и прецизираним сетом посебних циљева ове дисертације.

Имајући у виду дефинисане циљеве и примењену методологију истраживања, научно-истраживачки допринос ове дисертације огледа се у обезбеђивању систематског и свеобухватног приказа релевантних налаза и сазнања из научне и стручне грађе у области примењене статистике и мултиваријационе анализе података, како из угла концептуално-методолошких одређења метода и модела, тако и њиховог потенцијала примене у домену истраживања и моделирања комплексних економских феномена, конкретно, у сагледавању степена регионалних економских неравномерности територијалних јединица у саставу државе. У том смислу, на основу детаљног увида у различита виђења проблематике у разматрању, систематизације истраживачких ставова и идеја, њиховог сучељавања, аргументованог и критичког преиспитивања, развијени одговарајући мултиваријациони статистички модели представљају круцијални практични допринос ове докторске дисертације, имајући у виду њихову корисност у мерењу степена економске развијености општина у Републици Србији и, сходно томе, потенцијалне позитивне импликације на прецизирање мера и препорука релевантних за подстицање и управљање равномерним регионалним развојем.

Л И Т Е Р А Т У Р А

1. Abdollahzade, Gh. & Sharifzadeh, A. (2011). Classifying regional development in Iran (Application of Composite Index Approach). *Urban–Regional Studies and Research Journal*, 4(13): 9-14.
2. Alasia, A. (1996). Mapping the socio-economic diversity of rural Canada: A Multivariate analysis. *Agriculture and Rural Working Paper Series, Working Paper No. 67*. Доступно на: *http://www.publications.gc.ca/Collection/Statcan/21-601-MIE/21-601-MIE2004067.pdf*
3. Aldenderfer, M.S. & Blashfield, R.K. (1984). *Cluster Analysis.* Sage University Paper, series on Quantitative Applications in the Social Sciences, No. 07-044. Newbury Park, California: Sage Publications, Inc.
4. Amendola, A., Caroleo, F.E. & Coppola, G. (2004). Regional Disparities in Europe. *Discussion Paper No. 78*, Università degli Studi di Salerno, Centro di Economia del Lavoro e di Politica Economica. Доступно на:*https://core.ac.uk/download/pdf/6293559.pdf*.
5. Andrews, D.F. (1972). Plots of High-Dimensional Data. *Biometrics*, 28(1), Special multivariate issue: 125-136.
6. Andritsos, P. (2002). Data cLustering Techniques. *Qualifying Oral Examination Paper*. Доступно на: *http://www.cs.toronto.edu/~periklis/pubs/depth.pdf*.
7. Angelis, V., Angelis-Dimakis, A. & Dimaki, K. (2016). Identifying Clusters of Regions in the European South, based on their Economics, Social and Environmental Characteristics. *Region*, 3(2):71-102.
8. Arbelaitz, O., Gurrutxaga, I., Muguerza, J. Perez, J.M. & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46: 243-256.
9. Aumayr, C.M. (2006). European Region Types: A Cluster analysis of European NUTS 3 Regions. In: *Proceedings of the 46th Congress of the European Regional Science Association: "Enlargement, Southern Europe and the Mediterranean"*, August 30th-September 3rd, 2006, Volos, Greece. Доступно на: *http://www-sre.wu-wien.ac.at/ersa/ersaconfs/ersa06/papers/56.pdf*
10. Aumayr, C.M. (2007). European Region Types in EU-25. *The European Journal of Comparative Economics*, 4(2): 109-147.
11. Avram, M. & Postoiu, C. (2016). Territorial patterns of development in the European Union. *Theoretical and Applied Economics*, XXIII(1): 77-88.
12. Babin, B.J. (2011). Discriminant Analysis: An Overview. In: Lovrić, M. (Ed.), *International Encyclopedia of Statistical Science*: 388-390. Berlin: Springer-Verlag.
13. Bahovec, V., Dumičić, K. & Palić, I. (2011). Multivarijatna analiza pokazatelja društveno-ekonomskog razvoja u odabranim evropskim zemljama. *Zbornik Ekonomskog fakulteta u Zagrebu*, 9(1): 85-103.
14. Bailey, K.D. (1975). Cluster Analysis. *Sociological Methodology*, 6: 59–128.
15. Bartholomew, D., Steele, F., Moustaki, I. & Galbraith, J. (2008). *Analysis of multivariate social science data*, 2nd edition. Boca Raton: CRC Press, Taylor & Francis Group.
16. Bartholomew, D. (2011). Factor analysis and latent variable modelling. In: Lovrić, M. (Ed.), *International Encyclopedia of Statistical Science*: 501–503. Berlin: Springer-Verlag.
17. Bartlett, M.S. (1954). A note on the Multiplying Factors for Various χ^2 Approximations. *Journal of the Royal Statistical Society*. *Series B* (*Methodological*), 16(2): 296–298.
18. Berry, M. & Linoff, G. (2004). *Data Mining Techniques: For Marketing, sales and Customer Relationship Management*, 2nd edition. Indianapolis: Wiley Publishing, Inc.
19. Betz, N.E. (1987). Use of Discriminant Analysis in Counseling Psychology Research. *Journal of Counseling Psychology*, 34(4): 393-403.
20. Bijnen, E.J. (1973). *Cluster Analysis: Survey and Evaluation of Techniques*. The Netherlands: Tilburg University Press.

21. Bock, H.-H. (2008). Origins and extensions of the *k*-means algorithm in cluster analysis. *Electronic Journal for History of Probability and Statistics*, 4(2): 1-18.

22. Bojović, J. (2010). *Lokalni ekonomski razvoj, priručnik za praktičare*. Beograd: The Urban Institute. Доступно на: *https://www.slideshare.net/NALED/lokalni-ekonomski-razvoj-u-srbiji-prirunik-za-praktiare-12127476*.

23. Boslaugh, S. & Watters, P.A. (2008). *Statistics in a nutshell*, 1st edition. Sebastopol: O'Reilly Media, Inc.

24. Bouguessa, M., Wang, S. & Sun, H. (2006). An Objective Approach to Cluster Validation. *Pattern Recognition Letters*, 27(13): 1419-1430.

25. Box, G.E.P. (1949). A General Distribution Theory for a Class of Likelihood Criteria. *Biometrika*, 36(3/4): 317-346.

26. Braičić, Z. & Lončar, J. (2011). Intra-regional disparities in Sisak-Moslavina County. *Geoadria*, 16(1): 93-118.

27. Brauksa, I. (2013). Use of cluster analysis in exploring economic indicator differences among Regions: the case of Latvia. *Journal of Economics, Business and Management*, 1(1): 42-45.

28. Bromley, D.W. (1971). The use of discriminant analysis in selecting rural development strategies. *American Journal of Agricultural Economics*, 53(2): 319-322.

29. Brown, T.A. (2015). *Confirmatory Factor Analysis for Applied Research*, 2nd edition. New York: The Guilford Press.

30. Brown, M. T. & Wicker, L. R. (2000). In: Tinsley, H. & Brown, S. (Eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling*: 209–235. Massachusetts: Academic Press.

31. Calinski, T. & Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in statistics*, 3(1): 1-27.

32. Capriati, M. (2005). Expenditure in R&D and local development: an analysis of Italian provinces. *ERSA conference papers, ersa05p222*, European Regional Science Association. Доступно на:*https://ideas.repec.org/p/wiw/wiwrsa/ersa05p222.html*

33. Cattell, R.B. (1983). This Week's Citation Classic – The scree test for the number of factors. *Citation Classic, CC/Soc.Behav.Sci*(5):16–16.

34. Chaimontree, S., Atkinson, K. & Coenen, F. (2010). Best Clustering Configuration Metrics: Towards Multiagent Based Clustering. In: *Proceedings of the International Conference on Advanced Data Mining and Applications, ADMA* 2010: 48-59.

35. Chan, W.W.-Y. (2006). *A survey on multivariate data visualization*. Доступно на: *https://www.saedsayad.com/docs/multivariate_visualization.pdf*

36. Charrad, M., Lechevallier, Y., Ahmed, M.B. & Saporta, G. (2010). On the number of clusters in block clustering algorithms. In: *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference* (*FLAIRS* 2010): 392-397.
Доступно на: *https://cedric.cnam.fr/fichiers/art_2048.pdf*

37. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of statistical software*, 61(6): 1-36.

38. Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(No.342): 361-368.

39. Chernoff, H. (2011). Chernoff Faces. In: Lovrić, M. (Ed.), *International Encyclopedia of Statistical Science*: 243-244. Berlin: Springer-Verlag.

40. Choi, S.S., Cha, S.-H. & Tappert, C.C. (2010). A survey of Binary Similarity and Distance Measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1): 43-48.

41. Coombes, M. & Wong, C. (1994). Methodological steps in the development of multivariate indexes for urban and regional policy analysis. *Environment and Planning A*, 26: 1297-1316.

42. Coombs, W.T., Algina, J. & Olson-Oltman, D. (1996). Univariate and Multivariate Omnibus Hypothesis Tests Selected to Control Type I Error rates When Population Variances are not Necessarily Equal. *Review of Educational Research*, 66(2): 137-179.

43. 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.

44. Cziraky, D., Sambt, J., Rovan, J. & Puljiz, J. (2005). Regional development assessment: A structural equation approach. *European Journal of Operational Research*. Доступно на: *http://stats.lse.ac.uk/ciraki/EJOR.pdf*

45. Čobanović, K., Nikolić-Đorić, E. & Mutavdžić, B. (2003). Testovi višestrukih upoređenja. *Letopis naučnih radova*, 27(1): 66-73.

46. Čoček, L. (2010). Patterns of Regional Development in Serbia: A Multivariate Statistical Analysis. *Geographica Pannonica*, 14(1): 14-22.

47. Dalbelo-Bašić, B. (2011). Distance Measures. In: Lovrić, M. (Ed.), *International Encyclopedia of Statistical Science*: 397-398. Berlin: Springer-Verlag.

48. Das, A. (1999). Socio-economic development in India: A regional Analysis. *Development and Society*, 28(2): 313-345.

49. del Campo, C., Monteiro, C.M.F. & Soares, J.O. (2008). The European regional policy and the socio-economic diversity of European regions: a multivariate analysis. *European Journal of Operational Research*,187(2): 600-612.

50. Desgraupes, B. (2013). Clustering Indices. *University of Paris Ouest-Lab Modal'X*. Доступно на: *https://cran.biodisk.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf*.

51. Despotović, D. & Cvetanović, S. (2017). Teorijska eksplikacija faktora regionalnog rasta i ekonomske konvergencije (divergencije) regiona. *Ekonomski horizonti*, 19(2): 109-123.

52. Dinčić, N. (2016). Neki primeri metrika. *Matematika i informatika*, 3(3): 23-36.

53. Driver, H.E. & Kroeber, A.L. (1932). Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology*, 31(4): 211–256.

54. Dunham, M.H. (2000). *Data Mining Techniques and Algorithms*. New Jersey: Prentice Hall.

55. Đorđević, V., Lepojević, V. & Janković-Milić, V. (2011). *Primena statističkih metoda u istraživanju tržišta*. Niš: Ekonomski fakultet Univerziteta u Nišu.

56. Everitt, B. (2010). *Multivariable modeling and multivariate analysis for the behavioral sciences*. Boca Raton: CRC Press, Taylor & Francis Group.

57. Everitt, B.S., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster Analysis*, 5th edition. Chichester (UK): John Wiley & Sons Ltd.

58. Farris, J.S. (1969). On the Cophenetic Correlation Coefficient. *Systematic Zoology*, 18(3): 279-285.

59. Fernando, M.A.C.S.S., Samita, S. & Abeynayake, R. (2012). Modified Factor Analysis to Construct Composite Indices: Illustration on Urbanization Index. *Tropical Agricultural Research*, 23(4): 327–337.

60. Finch, H. & French, B. (2013). A Monte Carlo Comparison of Robust MANOVA Test Statistics. *Journal of Modern Applied Statistical Methods*, 12(2): 35-81.

61. Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179-188.

62. Fisher, R. A. & Mackenzie, W.A. (1923). Studies in crop variation. II The manurial responseof different potato varieties. *Journal of Agricultural Science*, 13: 311-320.

63. Foa, R. & Tanner, J.C. (2012). Methodology of the Indices of Social Development. *ISD Working Paper Series*, N0 2012-04. The Hague: Institute of Social Studies of Erasmus University Rotterdam. Доступно на: *https://repub.eur.nl/pub/50510/ISD-WP-2012-4.pdf*

64. Goletsis, Y. & Chletsos, M. (2011). Measurement of development and regional disparities in Greek periphery: A multivariate approach. *Socio-Economic Planning Sciences*, 45(4): 174-183.

65. Gore, P.A. Jr. (2000). Cluster Analysis. In: Tinsley, H. & Brown, S. (Ed.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling*: 297–321. Massachusetts: Academic Press.

66. Hair, J.F.Jr, Black, W., Babin, B. & Anderson, R. (2010). *Multivariate data analysis*, 7th edition. New Jersey: Pearson Prentice Hall.

67. Hair, J.F.Jr, Black, W., Babin, B. & Anderson, R. (2014). *Multivariate data analysis*, 7th edition. Harlow: Pearson Education Limited.

68. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2/3): 107-145.

69. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002). Clustering validity checking methods: part II. *ACM Sigmod Record*, 31(3): 19-27.

70. Han, J., Kamber, M. & Pei, J. (2012). *Data Mining: Concepts and Techniques*, 3rd edition. Waltham: Morgan Kaufmann Publishers.

71. Hardle, W. & Simar, L. (2003). *Applied Multivariate Statistical Analysis*. Доступно на: *http://www.leg.ufpr.br/lib/exe/fetch.php/wiki:internas:biblioteca:applied_multivariate_statistics.pdf*

72. Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2(3): 360-378.

73. Huberty, C.J. (1975). Discriminant Analysis. *Review of Educational Research*, 45(4): 543-598.

74. Huberty, C.J. (2011). Discriminant Analysis: Issues and Problems. In: Lovrić, M. (Ed.), *International Encyclopedia of Statistical Science*: 390-392. Berlin: Springer-Verlag.

75. Huberty, C.J. & Lowman, L.L. (2000). Group Overlap as a Basis for Effect Size. *Educational and Psychological Measurement*, 60(4): 543-563.

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

77. Huberty, C.J. & Petoskey, M.D. (2000). In: Tinsley, H. & Brown, S. (Eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling*: 183–208. Massachusetts: Academic Press.

78. Hudrlikova, L. (2013). Composite indicators as a useful tool for international comparison: The Europe 2020 example. *Prague Economic Papers*, 22(4): 459-473.

79. *IBM* (2011). *IBM SPSS Statistics 20 Algorithms*. Доступно на: *ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf*

80. Игић, В. (2014). *Примена мултиваријационе анализе у формирању композитног индекса развијености округа у Србији*. Докторска дисертација. Ниш: Економски факултет Универзитета у Нишу.

81. Istrate, M. & Horea-Serban, R.-I. (2016). Economic Growth and Regional Inequality in Romania. *Analele Universitatii din Oradea, Seria Geografie*, XXVI(2): 201-209.

82. Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques*. New York: Springer Science+Business Media, LLC.

83. Jaba, E., Jemna, D.V., Viorica, D. & Balan, C.B. (2007). Discriminant Analysis in the Study of Romanian Regional Economic Development. *Scientific Annals of the Alexandru Ioan Cuza University of Iasi*, LIV: 147-153.

84. Jain, A. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31:651-666.

85. Jain, A. & Dubes, R. (1988). *Algorithms for Clustering Data*. New Jersey: Prentice Hall.

86. Janković-Milić, V., Marković, I. & Igić, V. (2013). Cluster analysis of the districts in Serbia according to social development indicators. In: *Proceedings of the 2nd International scientific conference – Post Crisis Recovery*, Belgrade Banking Academy & Institute of Economic Sciences, Belgrade, Serbia: 676-693.

87. Jansky, J. (2016). Analysis of the Disparities Between the Regions of the Czech Republic. *Littera Scipta*, 9(2):59–67.

88. Jakopin, E. (2014). Regional Inequalities and Transition: The Case of Serbia. *Ekonomika preduzeća*, LXII(januar-februar): 117-133.

89. Jakopin, E. (2015). Regional drivers of economic growth. *Ekonomika preduzeća*, LXIII(1-2): 99-113.

90. Johnson, R.A. & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th edition. New Jersey: Pearson Prentice Hall.

91. Joreskog, K.G. (1969). A General Approach to Confirmatory Factor Analysis. *Psychometrika*, 34(2): 183–202.

92. Jun, T. (2010). Ranking and Clustering of the Economic Status of Rural Residents in 31 Provinces and Regions in China. *Asian Agricultural Research*, 2(12): 34-36.

93. Jurun, E. & Ratković, N. (2012). A Multivariate analysis of Croatian Counties Entrepreneurship. *Croatian Operational Research Review* (*CRORR*), 3: 310-320.

94. Kasper, D. & Unlu, A. (2013). On the relevance of assumptions associated with classical factor analytic approaches. *Frontiers in Psychology*, 4 (*Article* 109): 1–20 .

95. Kaufman, L. & Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Chichester (UK): John Wiley & Sons Ltd.

96. Klecka, W.R. (1980). *Discriminant Analysis*. Beverly Hills, California: Sage Publications, Inc.

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

98. Kramer, C.Y. (1978). An Overview of Multivariate Analysis. *Journal of Dairy Science*, 61: 848-854.

99. Kres, H. (1983). *Statistical tables for Multivariate Analysis*. New York: Springer-Verlag.

100. Krishnan, V. (2010). Constructing an area-based socioeconomic status index: A principal components analysis approach. *Paper presentedat the Early Childhood Intervention Australia (ECIA) 2010 Conference*. Доступно на: *https://pdfs.semanticscholar.org/b115/2cdc9e9ca217434db492a9e1c4ba519640ce.pdf?_ga=2.237820615.653746795.1554916748-1150936846.1533839343*

101. Kronthaler, F. (2003). A study of the Competitiveness of Regions based on a Cluster Analysis: The Example of East Germany. *IWH Discussion Papers*, No. 179. Доступно на: *http://nbn-resolving.de/urn:nbn:de:gbv:3:2-20453*

102. Krstić, B. & Vukadinović, D. (2011). Determinante konkurentnosti MSPP – pretpostavke za ravnomerni regionalni razvoj. U: *Zbornik radova XVI Naučnog skupa Regionalni razvoj i demografski tokovi zemalja Jugoistočne Evrope*: 553-568. Niš: Ekonomski fakultet Univerziteta u Nišu.

103. Kurnoga-Živadinović, N. (2007). Multivariate Classification of Croatian Counties. *Zbornik Ekonomskog fakulteta u Zagrebu*, 5(1): 1-15.

104. Kurnoga Živadinović, N., Dumičić, K. & Čeh Časni, A. (2009). Cluster and Factor analysis of Structural Economic Indicators for Selected European Countries. *WSEAS Transactions on Business and Economics*, 7(6): 331-341.

105. Kurnoga-Živadinović, N. & Sorić, P. (2008). Klaster analiza Županija Hrvatske prema sredstvima dobivenim iz programa Evropske unije. *Zbornik Ekonomskog fakulteta u Zagrebu*, 6(1): 193-207.

106. Kvičalova, J., Mazalova, V. & Široky, J. (2014). Identification of the differences between the regions of the Czech Republic based on the economic characteristics. *Procedia Economics and Finance*, 12: 343–352.

107. Lawley, D.N. & Maxwell, A.E. (1962). Factor Analysis as a Statistical Method. *Journal of the Royal Statistical Society, Series D* (*The Statistician*), 12(3): 209–229.

108. Lepojević, V., Bošković, G. & Janković-Milić, V. (2015). Klaster analiza pokazatelja razvijenosti opština u Regionu Južne i Istočne Srbije. U: *Zbornik radova XX Naučnog skupa – Regionalni razvoj i demografski tokovi zemalja Jugositočne Evrope*, Niš, Ekonomski fakultet Univerziteta u Nišu: 151-162.

109. Lipshitz, G. & Raveh, A. (1998). Socio-economic Differences among Localities: A New Method of Multivariate Analysis. *Regional Studies*, 32(8): 747-757.

110. Liu, Y., Li, Z., Xiong, H., Gao, X. & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. In: *Proceedings of the 2010 IEEE International Conference on Data Mining*, Sydney, Australia: 911-916. Доступно на: *http://datamining.rutgers.edu/publication/internalmeasures.pdf.*

111. Lovrić, M. (2009). *Osnovi statistike*. Kragujevac: Ekonomski fakultet Univerziteta u Kragujevcu.

112. Lovrić, M. & Stamenković, M. (2016). Analiza pokazatelja razvijenosti privrede Republike Srbije i zemalja u okruženju. U: Marinković, V, Janjić, V. i Mićić, V. (Redaktori), *„Unapređenje konkurentnosti privrede Republike Srbije"*: 479-491. Kragujevac: Ekonomski fakultet Univerziteta u Kragujevcu.

113. Lukić, V. & Anđelković Stoilković, M. (2017). Interrelation of spatial disparities in development and migration patterns in transition economy: Serbia – Case study. *Human Geographies – Journal of Studies and Research in Human Geography*, 11(1): 65-76.

114. Maletić, R. & Bucalo-Jelić, D. (2016). Definition of Homogeneous and Narrower Areas of the Republic of Serbia. *Agroekonomika*, 45(69): 13-24.

115. Manić, E., Mitrović, Đ. & Popović, S. (2017). Regional Disparity Analysis of Business Conditions: The Case of Serbia. *Ekonomika preduzeća*, LXV(mart-april): 275-293.

116. Manly, B.F.J. & Navarro Alberto, J.A. (2017). *Multivariate statistical methods: a primer*, 4th edition. Boca Raton: CRC Press, Taylor & Francis Group.

117. Martinez, W.L. & Martinez, A. R. (2007). *Computational Statistics Handbook with MATLAB*, 2nd edition. Boca Raton: CRC Press, Taylor & Francis Group.

118. Mason, R.L. & Young, J.C. (2011). Hotelling's T^2 Statistic. In: Lovrić, M. (Ed.), *International Encyclopedia of Statistical Science*: 638-640. Berlin: Springer-Verlag.

119. Mazzocchi, M. & Montresor, E. (2000). A Multivariate Statistical Approach to the Analysis of Rural Development. *Agricultural Economics Review*, 1(2): 31–45.

120. Melecky, L. (2012). Evaluation of Cohesion in Visegrad Countries in Comparison with Germany and Austria by Multivariate Methods for Disparities Measurement. *International Journal of Mathematical Models and Methods in Applied Sciences*, 8(6): 979–989.

121. Melecky, L. (2014). Spatial Analysis of NUTS 2 Regions based on Competitiveness Factors and Their Regional Variablity. In: *Proceedings of the 5th Central European Conference in Regional Science, CERS-2014*: 581–591.

122. Meyer, D.F., Jongh, J. De & Meyer, N. (2016). The formulation of a Composite Regional Development Index. *International Journal of Business and Management Studies*, 8(1): 100-116.

123. Michaelides, P.G., Economakis, G. & Lagos, D. (2006). Clustering Analysis Methodology for employment and Regional Planning in Greece. *MPRA Paper No. 74468*. Доступно на: *https://mpra.ub.uni-muenchen.de/74468/.*

124. Милановић, М. (2018). *Извођење законитости из економских података применом ata Mining приступа*. Докторска дисертација. Ниш: Економски факултет Универзитета у Нишу.

125. 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.

126. Милетић, Р., Тодоровић, М. и Миљановић, Д. (2009). Приступ неразвијеним подручјима у регионалном развоју Србије. *Зборник радова Географског Института „Јован Цвијић" САНУ*, 59(2): 149-171.

127. Miljačić, D. & Paunović, B. (2011). Regional Disparities in Serbia. *Ekonomika preduzeća*, LIX(7-8): 379–389.

128. Milligan, G.W. & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2): 159-179.

129. Министарство привреде (2015). *Извештај о привредном развоју Србије у 2014*. Доступно на: *http://www.privreda.gov.rs/wp-content/uploads/2015/10/IZVESTAJ-O-PRIVREDNOM-RAZVOJU-SRBIJE-U-2014-_SAJT.pdf*.

130. Министарство привреде, Сектор за регионални развој и стратешке анализе привреде (2014). *Извештај о Регионалном развоју Србије 2013*. Доступно на: *https://privreda.gov.rs/izvestaj-o-regionalnom-razvoju-srbije-za-2013-godinu/*

131. Molnar (2013). Osvrt 3. Činjenice o regionalnim razlikama u Srbiji. *Kvartalani monitor ekonomskih trendova i politika u Srbiji*, 32(januar-mart): 66-72.

132. Mohiuddin, S. & Hashia, H. (2012). Regional socio-economic disparities in the Kashmir Valley (India) – a geographical approach. *Bulletin of Geography. Socio-Economic Series*, 18: 85-98.

133. Moustafa, R.E. (2011). Andrews curves. *Advanced Review*, 3(July/August): 373-382.

134. Munandar, Tb. Ai & Azhari, SN. (2015). Analysis of Regional Development Disparity with Clustering Technique Based Perspective. *International Journal of Advanced Research in Computer Science*, 6(1): 137-141.

135. Murtagh, F. (2016). A Brief History of Cluster Analysis. In: Henning, C., Meila, M. Murtagh, F. & Rocci, R. (Eds.), *Handbook of Cluster Analysis*: 21–30. Boca Raton: CRC Press, Taylor & Francis Group.

136. Murtagh, F. & Heck, A. (1987). *Multivariate Data Analysis*. Dordrecht: D.Reidel Publis. Company.

137. Национална Агенција за Регионални Развој–НАРР (2012). *Извештај о Регионалном Развоју, 2012*. Београд: НАРР.

138. Nardo, M., Saisana, M., Saltelli, A. & Tarantola, S. (2005). *Tools for Composite Indicators Building*. Brussels: Joint Research Centre, European Commission. Доступно на: *http://publications.jrc.ec.europa.eu/repository/bitstream/JRC31473/EUR%2021682%20EN.pdf*.

139. Narodna Skupština Republike Srbije (2006). *Ustav Republike Srbije*. Beograd: Službeni Glasnik RS (br. 98/2006).

140. Nedović, Lj. (2016). *Neki tipovi rastojanja i fazi mera sa primenom u obradi slika*. Doktorska disertacija. Novi Sad: Fakultet Tehničkih Nauka Univerziteta u Novom Sadu.

141. Nielsen, L. (2011). Classifications of Countries Based on Their Level of Development: How it is Done and How it Could be Done. *IMF Working Paper*. Доступно на: *https://www.imf.org/external/pubs/ft/wp/2011/wp1131.pdf*

142. Obradović, S., Lojanica, N. & Janković, O. (2016). The influence of economic growth on regional disparities: Empirical evidence from OECD countries. *Zbornik radova Ekonomskog fakulteta u Rijeci*, 24(1): 161-186.

143. OECD (2008). *Handbook on constructing composite indicators: methodology and user guide*. Доступно на: *www.oecd.org/publishing/corrigenda*.

144. Osborne, J.W. (2010). Improving your data transformations: applying the Box-Cox transformation. *Practical assessment, research and evaluation*, electronic journal, 15(12): 1-9.

145. Ozaslan, M., Dincer, B. & Ozgur, H. (2006). Regional disparities and territorial indicators in Turkey: socio-economic development index (SEDI). In: *Proceedings of the 46th Congress of the European Regional Science Association: "Enlargement, Southern Europe and the Mediterranean"*, August 30th – September 3rd, 2006, Volos, Greece. Доступно на: *http://hdl.handle.net/10419/118537*

146. Pastor, E.M.C., Garcia, J.de H. & Gavilan, M.D.S. (2009). Statistic analysis of the socioeconomic context in Andalusia: an approach at municipal level. *Investigaciones Regionales*, 18: 107-138.

147. Perišić, A. (2014). Multivarijatna klasifikacija jedinica lokalne i regionalne samouprave prema socioekonomskoj razvijenosti. *Društvena istraživanja*, 23(2): 211-231.

148. Perišić, A. & Wagner, V. (2015). Development Index: Analysis of the basic instrument of Croatian regional policy. *Financial Theory and Practice*, 39(2): 205-236.

149. Pintilescu, C. (2011). Regional economic disparities in Romania. In: *Recent Researches in Applied Economics –Proceedings of the 3rd World Multiconference on Applied Economics, Business and Development* (*AEBD'*11) Iasi, Romania: 43-47. Доступно на: *http://www.wseas.us/e-library/conferences/2011/Iasi/AEBD/AEBD-06.pdf*

150. Pituch, K.A. & Stevens, J.P. (2016). *Applied Multivariate Statistics for the Social Sciences*, 6th edition. New York: Routledge, Taylor & Francis Group.

151. Polednikova, E. (2014). Comparing Regions' Ranking by MCDM methods: the Case of Visegrad Countries. *WSEAS Transactions on Business and Economics*, 11: 496-509.

152. Poulsen, J. & French, A. (2008). *Discriminant function analysis*. Доступно на: *http://userwww.sfsu.edu/efc/classes/biol710/discrim/discrim.pdf*

153. Puljiz, J. & Maleković, S. (2007). Regional income and unemployment disparities in Croatia. In: *Proceedings of the 7th International Conference „Enterprise in transition"*: 1280-1298. Split: Faculty of Economics in Split. Доступно на: *https://bib.irb.hr/datoteka/332497.Full_Paper_Disparities.pdf*

154. Raciborski, R. (2009). Graphical representation of multivariate data using Chernoff faces. *The Stata Journal*, 9(3): 374-387.

155. Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B*, 10: 159-193.

156. Rašić-Bakarić, I. (2005). Primjena faktorske i klaster analize u otkrivanju regionalnih nejednakosti. *Privredna kretanja i ekonomska politika*, 105/2005: 53-76.

157. Rašić-Bakarić, I. (2006). Methods of multivariate analysis to uncover socio-economic differences among spatial-economics entities. In: *Proceedings of the 46th Congress of the European Regional Science Association: "Enlargement, Southern Europe and the Mediterranean"*, August 30th-September 3rd, 2006, Volos, Greece. Доступно на: *http://www-sre.wu-wien.ac.at/ersa/ersaconfs/ersa06/papers/56.pdf*

158. Rencher, A. C. (2002). *Methods of Multivariate Analysis*, 2nd edition. New York: John Wiley & Sons, Inc.

159. Rendon, E., Abundez, I., Arizmendi, A. & Quiroz, E.M. (2011). Internal versus External cluster validation indexes. *International Journal of Computers and Communications*, 5(1): 27-34.

160. Републички завод за развој–РЗР (2009). *Регионални развој Србије 2009*. Београд: РЗР.

161. Републички завод за статистику, РЗС (2016). *Општине и региони у Републици Србији*. Београд: РЗС.

162. Rimac, I., Rihtar, S. & Oliveira-Roca, M. (1992). Multivarijatna klasifikacija općina Hrvatske kao moguća metoda regionalizacije Republike. *Društvena istraživanja*, 1(1): 87-99.

163. Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53-65.

164. Roux, M. (2015). A comparative study of divisive hierarchical clustering algorithms. *arXiv preprint, arXiv*: 1506.08977. Доступно на: *https://arxiv.org/abs/1506.08977*.

165. Rovan, J., Malešić, K. & Bregar, L. (2009). Well-being of the municipalities in Slovenia. *Geodetski Vestnik*, 53(1): 92-113.

166. Rovan, J. & Sambt, J. (2003). Socio-economic Differences Among Slovenian Municipalities: A Cluster Analysis Approach. In: Ferligoj, A. & Mrvar, A. (Eds.) *Developments in Applied Statistics, Metodološki zvezki*, 19: 265-278. Доступно на: *http://ams.sisplet.org/uploadi/editor/rovan.pdf*.

167. Sachinopoulou, A. (2001). Multidimensional Visualization. *VTT Tiedotteita – Research Notes 2114*. Доступно на: *https://www.vtt.fi/inf/pdf/tiedotteet/2001/T2114.pdf*

168. Sakia, R.M. (1992). The Box-Cox transformation technique: a review. *The Statistician*, 41:169-178.

169. Salzman, J. (2003). *Methodological choices encountered in the construction of composite indices of economic and social well-being*. Center for the Study of Living Standards. Доступно на: *http://www.csls.ca/events/cea2003/salzman-typol-cea2003.pdf*

170. Savić, M., Brcanov, D. & Dakić, S. (2008). Discriminant Analysis – Applications and Software Support. *Management Information Systems*, 3(1): 29-33.

171. Savić, M. & Zubović, J. (2015). Comparative Analysis of Labour Markets in South East Europe. *Procedia Economics and Finance*, 22: 388-397.

172. Sharma, S. (1996). *Applied Multivariate Techniques*. New York: John Wiley & Sons, Inc.

173. Shmueli, G., Patel, N. & Bruce, P. (2005). *Data Mining in Excel: Lecture Notes and Cases*. Arlington: Resampling Stats, Inc.

174. Soares, J.O., Marques, M.M.L. & Monteiro, C.M.F. (2003). A multivariate methodology to uncover regional disparities: A contribution to improve European Union and governmental decisions. *European Journal of Operational Research*, 145: 121–135.

175. Sokal, R.R. & Rohlf, F.J. (1962). The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2): 33-40.

176. Sokal, R.R. & Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy*. San Francisco: W.H.Freeman & Co.

177. Sokolovska, V., Nikolić-Đorić, E. & Lazar, Ž. (2014). Regionalne razlike u Republici Srbiji. U: Sokolovska, V. & Lazar, Ž. (Urednici), *Regioni i Regionalizacija - sociološki aspekti 3*: 9-21. Novi Sad: Filozofski fakultet Univerziteta u Novom Sadu.

178. Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2): 201-292.

179. Spinelli, J.G. & Zhou,Yu (2004). Mapping Quality of Life with Chernoff Faces. In: *Proceedings of 24th ESRI International User Conference*. Доступно на: *http://proceedings.esri.com/library/userconf/educ04/papers/pap5000.pdf*

180. Stamenković, M. & Savić, M. (2017). Measuring regional economic disparities in Serbia: Multivariate statistical approach. *Industrija*, 45(3): 101–130.

181. Стаменковић, М., Веселиновић, П. и Милановић, М. (2017). Демографски ресурси округа у Републици Србији: Анализа груписања. *TEME*, XLI(4): 873-897.

182. Timm, N.H. (2002). *Applied Multivariate Analysis*. New York: Springer-Verlag.

183. Thompson, B. (2014). *Exploratory and confirmatory factor analysis: understanding concepts and applications*. Washington: American Psychological Association.

184. Thurstone, L.L. (1931). Multiple Factor Analysis. *Psychological Review*, 38(5): 406–427.

185. Tobachnick, B.G. & Fidell, L.S. (2013). *Using Multivariate Statistics*, 6th edition. New Jersey: Pearson Education, Inc.

186. Tryon, R.C. (1939). *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*. Ann Arbor: Edwards Brothers.

187. Tuffery, S. (2011). *Data Mining and Statistics for Decision Making*. Chichester: John Wiley & Sons.

188. Varmuza, K. & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton: CRC Press, Taylor & Francis Group.

189. Vasić, A. (2014). Charles E. Spearman (1863–1945) i faktorska analiza posle 110 godina. *Civitas*, 4(1): 56–85.

190. Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. Chichester (UK): John Wiley & Sons Ltd.

191. Villaverde, J. & Maza, A. (2009). Measurement of regional economic disparities. *UNU-CRIS Working Papers, W-2009/12*. Доступно на: *http://cris.unu.edu/measurement-regional-economic-disparities*

192. Влада Републике Србије (2007а). *Стратегија Регионалног Развоја Републике Србије за период од 2007. до 2012. године*. Београд: Службени Гласник РС. Доступно на: *http://www.gs.gov.rs/lat/strategije-vs.html*

193. Влада Републике Србије (2007б). *Закон о територијалној организацији Републике Србије*. Београд: Службени Гласник РС (бр. 129/2007).

194. Влада Републике Србије (2009а). *Закон о Регионалном развоју*. Београд: Службени Гласник РС (бр. 51/2009 и 30/2010).

195. Влада Републике Србије (2009б). *Уредба о номенклатури статистичких територијалних јединица*. Београд: Службени Гласник РС (бр. 109/2009 и 46/2010).

196. Vlada Republike Srbije (2015). *Uredba o utvrđivanju metodologije za izračunavanje stepena razvijenosti regiona i jedinica lokalne samouprave*. Доступно на: *https://www.paragraf.rs/dnevne-vesti/150715/150715-vest10.html*

197. Вуковић, Д. (2009). Ниска конкурентност неразвијених подручја: „уско грло“ привреде Србије. *Зборник радова Географског Института „Јован Цвијић“ САНУ*, 59(2): 189-204.

198. Vukmirović, J. (2013). Regionalni razvoj kao preduslov za izlazak iz krize. *Makroekonomske analize i trendovi*, 219: 39-43.

199. Vydrova, H.V. & Novotna, Z. (2012). Evaluation of disparities in living standards of regions of the Czech Republic. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, XL(4): 407-414.

200. Watkins, M.W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, 44(3): 219–246.

201. Wilks, S.S. (1932). Certain Generalizations in the Analysis of Variance. *Biometrika*, 24(3/4):471-494.

202. Wilson, P. & Cooper, C. (2008). Finding the Magic Number. *Methods*, 21(10): 866–867.

203. Winkler, A. (2012). Measuring regional inequality: an index of socio-economic pressure for Serbia. *Zbornik radova–Geografski Fakultet Univerziteta u Beogradu*, 60: 81-102.

204. Wishlade, F. & Yuill, D. (1997). Measuring Disparities for Area Designation Purposes: Issues for the European Union. *Regional and Industrial Policy Research Paper, Number 24*. Glasgow: European Policies Research Centre.

205. Yong, A. G. & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2):79–94.

206. Zhao, Y., Zhang, H. & Wang, C. (2006). Factor analysis and the measurement of economic growth: A comparison of Chinese provinces. *C-level essay in Statistics*. Dalarna University. Доступно на: *http://www.statistics.du.se/essays/C06C.Wang_Zhang_Zhao.pdf*

207. Zubin, J.A. (1938). A technique for measuring likemindedness. *Journal of Abnormal and Social Psychology*, 33(4): 508-516.

208. Zygmont, C. & Smith, M. (2014). Robust factor analysis in the presence of normality violations, missing data, and outliers: Empirical questions and possible solutions. *The Quantitative Methods for Psychology*, 10(1): 40–55.

209. Žmuk, B. (2015). Quality of life indicators in selected European countries: Statistical hierarchical cluster analysis approach. *Croatian Review of Economic, Business & Social Statistics*, 1(1-2):42-54.

Интернет извор података: [*http://www.apr.gov.rs/*] – Агенција за привредне регистре

Образац 1

ИЗЈАВА АУТОРА О ОРИГИНАЛНОСТИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Ја, Милан (Драган) Стаменковић, изјављујем да докторска дисертација под насловом:

„Мултиваријационо статистичко моделирање у функцији мерења степена економске развијености територијалних јединица“,

која је одбрањена на Економском факултету Универзитета у Крагујевцу представља *оригинално ауторско дело* настало као резултат *сопственог истраживачког рада*.

Овом Изјавом такође потврђујем:

- да сам *једини аутор* наведене докторске дисертације,
- да у наведеној докторској дисертацији *нисам извршио/ла повреду* ауторског нити другог права интелектуалне својине других лица,
- да умножени примерак докторске дисертације у штампаној и електронској форми у чијем се прилогу налази ова Изјава садржи докторску дисертацију истоветну одбрањеној докторској дисертацији.

У Крагујевцу, 11.4.2019. године,

Милан Стаменковић

потпис аутора

Образац 2

ИЗЈАВА АУТОРА О ИСКОРИШЋАВАЊУ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Ја, Милан (Драган) Стаменковић ,

☑ дозвољавам

☐ не дозвољавам

Универзитетској библиотеци у Крагујевцу да начини два трајна умножена примерка у електронској форми докторске дисертације под насловом:

"Мултиваријационо статистичко моделирање у функцији мерења степена економске развијености територијалних јединица",

која је одбрањена на Економском факултету

Универзитета у Крагујевцу, и то у целини, као и да по један примерак тако умножене докторске дисертације учини трајно доступним јавности путем дигиталног репозиторијума Универзитета у Крагујевцу и централног репозиторијума надлежног министарства, тако да припадници јавности могу начинити трајне умножене примерке у електронској форми наведене докторске дисертације путем *преузимања*.

Овом Изјавом такође

☑ дозвољавам

☐ не дозвољавам[1]

[1] Уколико аутор изабере да не дозволи припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од *Creative Commons* лиценци, то не искључује право припадника јавности да наведену докторску дисертацију користе у складу са одредбама Закона о ауторском и сродним правима.

припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од следећих *Creative Commons* лиценци:

1) Ауторство

2) Ауторство - делити под истим условима

3) Ауторство - без прерада

4) Ауторство - некомерцијално

5) Ауторство - некомерцијално - делити под истим условима

6) Ауторство - некомерцијално - без прерада[2]

У Крагујевцу, 11.4.2019. године,

Милан Стаменковић

потпис аутора

[2] Молимо ауторе који су изабрали да дозволе припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од *Creative Commons* лиценци да заокруже једну од понуђених лиценци. Детаљан садржај наведених лиценци доступан је на: http://creativecommons.org.rs/