УНИВЕРЗИТЕТ У ПРИШТИНИ СА ПРИВРЕМЕНИМ
СЕДИШТЕМ У КОСОВСКОЈ МИТРОВИЦИ
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА

Драгиша Р. Миљковић

КОРИШЋЕЊЕ МЕТОДА ИСТРАЖИВАЊА ПОДАТАКА ЗА ПРЕПОЗНАВАЊЕ РАСПОДЕЛА СИГНАЛА И ЊИХОВИХ ПАРАМЕТАРА У РЕАЛНИМ ПРОЦЕСИМА

Докторска дисертација

Косовска Митровица, 2024.

UNIVERSITY OF PRISTINA TEMPORARY SETTLED IN KOSOVSKA MITROVICA

FACULTY OF TECHNICAL SCIENCES

Dragiša R. Miljković

USING DATA ANALYSIS METHODS FOR IDENTIFYING SIGNAL PROBABILITY DISTRIBUTIONS AND THEIR PARAMETERS IN REAL-WORLD PROCESSES

Doctoral dissertation

Kosovska Mitrovica, 2024.

Идентификациона страница докторске дисертација

I — Аутор	
Име и презиме:	Драгиша Р. Миљковић
Датум и место рођења:	20.12.1989. године, К. Митровица
Садашње запослење:	асистент
II — Докторска дисертација	
Наслов:	Коришћење метода истраживања података за препознавање расподела сигнала и њихових параметара у реалним процесима
Број страница:	162
Број слика:	48
Број библиографских података:	108
Установа и место где је рад израђен:	Универзитет у Приштини, Факултет техничких наука Косовска Митровица
Научна област (УДК):	Вештачка интелигенција - УДК 004.8
Ментор:	др Синиша Илић, ред. проф. Универзитета у Приштини са привременим седиштем у Косовској Митровици
III — Оцена и одбрана	
Датум пријаве теме:	28.12.2022. године
Број одлуке и датум прихватања заснованости теме докторске дисертације:	428/3-2 06.04.2023. године
Комисија за оцену подобности теме и кандидата:	1. др Драгана Радосављевић, ванред. проф. ФТН у К. Митровици – председник комисије 2. др Синиша Илић, ред. проф. ФТН у К. Митровици – ментор 3. др Бошко Николић, ред. проф. Електротехничког факултета у Београду – члан
Комисија за оцену и одбрану докторске дисертације:	
Датум одбране:	

САЖЕТАК

У овој дисертацији је представљено коришћење метода истраживања података за препознавање расподеле вероватноће у сигналима реалних процеса и за одређивање вредности параметара тих расподела. Развијени алгоритам омогућава брзо и прецизно одређивање расподела вероватноће и идентификацију параметара сигнала без априори знања о њему. Такође, описане су и технике за генерисање и предобраду псеудослучајних сигнала са унапред познатим расподелама вероватноће, што доприноси квантификацији грешака у препознавању.

У дисертацији је извршена свеобухватна анализа резултати препознавања расподела вероватноће и вредности њихових параметара за гама, Рејлијев, Рајсов, Накагамијев и Вејбулов модел фединга у бежичним каналима, уз оцену утицаја броја узорака сигнала и тачака нормализоване дискретне густине на тачност препознавања.

Такође, дат је и приказ коришћења техника машинског учења директно на подацима у базама података,. Развијени алгоритам је примењен над великим скуповима интернет саобраћаја у циљу моделовања њихових карактеристика и откривања могућности препознавања DDoS напада на основу препознатих расподела вероватноће.

У раду је представљена и софтверска апликација развијена на основу изложеног алгоритма, која омогућава генерисање псеудослучајних сигнала различите дужине и зашумљености, те препознавање расподела вероватноће и њихових параметара у реалним процесима.

Кључне речи: *истраживање података, нелинеарна регресија, расподела вероватноће, фединг, DDoS напад*

ABSTRACT

This dissertation presents data analysis methods for identifying probability distributions in real-world signals and estimating their parameters. The developed algorithm allows for the fast and precise recognition of these distributions and signal parameters without prior knowledge of the signal characteristics. Techniques for generating and preprocessing pseudorandom signals with known distributions are described, aiding in error quantification.

A comprehensive analysis of probability distributions and their parameters for Gamma, Rayleigh, Rice, Nakagami, and Weibull models in wireless channels is conducted, and the impact of sample size and normalized discrete cumulative distribution function on recognition accuracy is assessed.

Moreover, this dissertation demonstrates the application of machine learning techniques on database data. The algorithm was applied to extensive internet traffic datasets to model characteristics and explore the detection of DDoS attacks through recognized probability distributions.

The work also introduces a software application based on this algorithm, facilitating the generation of pseudorandom signals of varying lengths and noise levels, and the recognition of probability distributions in various processes.

Keywords: *data analysis, nonlinear regression, probability distribution, fading, DDoS attack*

Садржај

1. УВОД

Истраживање података је процес примене рачунарских метода над великим количинама података како би се открише нове корисне и релевантне информације. Истраживање података се не користи само за проналажење занимљивих образаца из података, већ и за претраживање великих скупова података, за изградњу модела који описују релевантна својства података и за прављење предвиђања заснованих на подацима [1].

Многи аспекти нашег свакодневног живота се бележе и чувају. Многе компаније, државна тела и организације бележе разне информације. Нпр. метеоролошке станице чувају податке о температури и притиску, надзорне камере посматрају саобраћај, а скоро све што се одвија електронским путем се на неки начин бележи: телекомуникације, финансијске трансакције, интеракције на друштвеним мрежама, и многе друге. Гомилање података је довело до, може се рећи, *big data* ере. Данашње време је јединствено по томе што по први пут имамо огромну количину података којима се може приступати преко рачунара. Ово богатство информација има потенцијал да пружи систематични поглед на податке.

Машинско учење је грана рачунарства која се бави развојем алгоритама који омогућавају рачунарима да уче из података и на основу тога доносе интелигентне одлуке или предузимају акције. Развој ове области омогућен брзим растом и унапређењем рачунарских перформанси, повећањем доступности великих количина података и напретком статистичких метода за њихову обраду. Оно се укратко може дефинисати као процес откривања корисног знања из велике количине података [1]. Истраживање података је блиско везано са машинским учењем и постоји извесно преклапање међу њима, али иако скоро сви процеси истраживања података укључују машинско учење, има области машинског учења код којих нема истраживања података. Може се рећи да је главна разлика у томе што се код машинског учења покушава научити рачунар како да на основу података реши неки проблем,

док се код истраживања података покушава научити рачунар како да идентификује обрасце у подацима који ће потом помоћи људима да реше неки проблем.

У добу у којем живимо човек је у могућности да мери и анализира мерене податке у реалном времену. То је омогућено развојем квалитетних сензора који промене посматране величине претварају у промене одговарајуће електричне величине формирајући тако електрични сигнал. Одмеравањем сигнала добијају се његове дискретне вредности које се, употребом различитих алгоритама, могу анализирати и обрађивати.

1.1. Предмет и циљеви истраживања

Ова докторска дисертација се бави применом метода истраживања података за препознавање расподела сигнала и њихових параметара у реалним процесима. Циљ овог истраживања јесте развијање нових и бољих алгоритама за бржу и прецизнију анализу података, чиме се унапређује процес анализе сигнала у одговарајућим реалним процесима.

Основни циљ ове дисертације јесте проучавање параметара фединга са различитим утицајима на канал преноса коришћењем нелинеарне регресије, и пружање информација о фединг каналима. Секундарни циљ ове дисертације јесте примена развијеног алгоритма за анализу великих скупова интернет саобраћаја, како би се моделовањем расподела вероватноће одабраних карактеристика овог саобраћаја стекли нови увиди.

Код пријема бежичних комуникационих сигнала веома је важно да примљени сигнал буде што квалитетнији, тј. да има што боље мере перофрмансе. У том циљу развијено је више технологија које омогућавају повећање мере перформанси система. Једна од њих су диверзити технике, тј. систем са више пријемних антена који на основу алгоритма рада бирају ''најбољи'' сигнал са једне од антена и даље га процесирају. ''Најбољи'' сигнал може бити изабран

на основу више критеријума, односно параметара који фигуришу у статистичким моделима за описивање простирања сигнала и феднига.

Одабрани приступ омогући ће процену параметара феднига из одмерака сигнала без априори знања о параметрима или о самом сигналу. У дисертацији се такође проучава утицај броја одмерака сигнала на резултате препознавања. Поред тога, проучава се тачност препознавања сигнала.

1.2. Основне хипотезе од којих ће се полазити у истраживању:

Хипотезе представљају претпостављено објашњење чињеница и појава које се проверавају предложеним истраживањем. Главна хипотеза која ће бити тестирана у раду гласи:

Употребом метода истраживања података могуће је са великом тачношћу извршити препознавање расподеле анализираног сигнала и вредности њених параметара.

На основу дефинисаног предмета и циљева истраживања може се издвојити неколико посебних хипотеза:

- За већину теоријских расподела од интереса познати су математички модели (формуле) и за њихове кумулативне расподеле вероватноће,
- За већину теоријских расподела могуће је генерисати одмерке псеудо-случајног сигнала са минималним одступањем од теоријске расподеле,
- Релативно брже препознавања расподеле анализираног сигнала и њених параметара могуће је постићи оптимизацијом алгоритма односно филтрирањем очекиваних расподела и њених параметара,
- Понављањем експеримената препознавања са различитим вредностима параметара расподеле у генерисаном сигналу велики број пута добијају се тачнији резултати успешности препознавања,

- Понављањем експеримената препознавања са различитим дужинама сигнала и израчунавањем нормализоване дискретне кумулативне расподеле вероватноће (НДКР) сигнала са различитим бројем тачака, на основу добијене грешке у препознавању може се установити оптимална дужина сигнала и број тачака НДКР за брзо и тачно препознавање.

1.3. Оквирни опис садржаја дисертације:

У Уводу дисертације представљени су предмет и циљеви истраживања, полазне хипотезе, затим методологија и коришћени алати, као и структура и организација рада.

У другом поглављу представљене су методе истраживања података које су коришћене за остваривање циљева дисертације, као и тренутног стања у датој области. Такође су представљене технике регресије за проналажење односа између независних и зависних модела, као и предности и мане тестова адекватности модела и метода за одабир модела.

У трећем поглављу је дат опис појављивања варијација у јачини сигнала (односно фединга), најчешће коришћених расподела вероватноће за моделовање овог процеса, као и поступак препознавања расподела вероватноће у каналима фединга и њихових параметара.

У четвртом поглављу су представљени процес генерисања псеудо-случајних сигнала, његове обраде и препознавања расподела и њихових параметара за тај сигнал. Статистичко моделовање улазних параметара се врши преко тестова за испитивање адекватности модела (енгл. *goodness-of-fit*) и метода за одабир модела.

У петом поглављу су дати резултати анализе генерисаних псеудослучајних сигнала фединга и приказано је моделовање процеса помоћу нелинеарне регресије. С обзиром на то да је један од основних проблема у бежичним комуникацијама зашумљеност сигнала, резултати су представљени за три различита нивоа шума.

У шестом поглављу је дат опис коришћења нелинеарне регресије за моделовање интернет саобраћаја, као и коришћење OML4R модула у оквиру Оракл серверу за управљање базама података помоћу којег је вршено моделовање две класе интернет саобраћаја (нормалног и саобраћаја DDoS напада). Овим приступ, где се методе истраживања података извршавају директно у серверу базе података, елиминише се потреба за слањем великих количина података, већ се ка клијенту прослеђују само резултати анализе.

У седмом поглављу је описан развијени софтвер за аутоматско генерисање сигнала, његову пред-обраду, препознавање, документовање резултата препознавања и статистичку анализу.

У закључку су наведени остварени резултати, њихов допринос, као и план будућих истраживања.

У последњим поглављима су наведене (редом) листа литературе коришћене у истраживању, списак табела и списак слика.

2. МЕТОДЕ ИСТРАЖИВАЊА ПОДАТАКА

2.1. Општи принципи и област примене истраживања података

Истраживање података је мултидисциплинарно поље које подразумева коришћење рачунарских техника за издвајање вредних информација из великих скупова података. Термин „истраживање података“ је настао 1990-их година, а своје корене вуче из поља статистике, машинског учења и вештачке интелигенције. Формална дефиниција машинског учења која се приписује Тому М. Мичелу каже да *„машина које може да учи јесте она која је у стању да са проласком времена аутоматски унапређује своје перформансе на основу претходног искуства“* [2, p. 1].

Примарни циљ истраживања података јесте откривање до тада непознатих образаца, веза и нових увида унутар великих и сложених скупова података. Сирови подаци добијају на вредности када се применом различитих алгоритама и техника из њих могу одредити трендови, корелације и одступања унутар података. Овако добијене информације могу обезбедити веома корисне увиде за предузећа, истраживаче и доносиоце одлука и могу бити искоришћене за доношење одлука, предвиђање и откривање знања, чиме се обезбеђује предност над конкурентнима на глобалном тржишту.

Истраживање података има широк спектар примена у различитим доменима. Веома је коришћено у секторима пословања и финансија за анализу понашања купаца, идентификацију трендова на тржишту и оптимизацију маркетиншких стратегија [3]. Даље, користи се за откривање превара кроз идентификовање неуобичајених образаца и аномалија унутар трансакционих података [4]. У области здравства, технике истраживања података се користе за анализу електронских здравствених записа, медицинских слика и података о геномима, а све у циљу откривања образаца и веза које могу побољшати негу

и исходе лечења пацијената [5]. Истраживање података све више се користи у друштвеним наукама за анализу великих друштвених мрежа и проучавање људског понашања, друштвених динамика и ширења информација [6]. Користи се и за анализу текстуалних података, као што су новински чланци и академске публикације, у циљу идентификације трендова и образаца. У области сајбер безбедности, истраживање података се користе за откривање и спречавање сајбер претњи анализом мрежног саобраћаја, системских логова и понашања корисника у циљу идентификације образаца и аномалија које могу указивати на злонамерну активност [7].

У данашњем свету се свакодневно генеришу огромне количине података из различитих извора, као што су друштвене мреже, сензори, трансакционе базе података, и друго. Експоненцијални раст количине доступних података непосредно ствара потребу за развојем напредних техника и алгоритама за њихову обраду. Ово истовремено ствара нове изазове, али и отвара нове могућности у области истраживања података.

Постоји неколико стандардних модела [8] који дефинишу секвенцијалне кораке које треба извршити у процесу истраживања података. Једна од широко коришћених методологија јесте CRISP-DM (енгл. *Cross-Industry Standard Process for Data Mining*). Ова методологија процес истраживања података разложе на шест фаза: разумевање пословања, разумевање података, припремање података, моделовање, евалуацију и имплементацију [9].

2.1.1. Историјски контекст и развој

Историја истраживања података почиње од раног 20. века, са развојем статистичких метода за анализу података, као што су регресиона анализа, тестирање хипотеза и Бајесовско закључивање. Ове методе обезбедиле су основу за ране технике истраживања података, као што су анализа кластера и стабла одлучивања. Међутим, до истинског процвата истраживања података долази са развојем машинског учења и вештачке интелигенције. Рани

алгоритми машинског учења, као што су перцептрони и стабла одлучивања су били дизајнирани да уче шаблоне и везе унутар података, чиме су отворене могућности и за напредније технике истраживања података. Развој вештачке интелигенције је довео до стварања експертских система који користе правила одлучивања да реше комплексне проблеме и дају предвиђања на основу великих скупова података.

Развој система база података и складиштења података у другој половини 20. века обезбедио је инфраструктуру неопходну за рударење података над огромним количинама података. Системи база података су омогућили ефикасно складиштење и приступ великим количинама структурираних података, док су технике складиштења података омогућиле интеграцију и анализу података из више извора. Ово је положило темеље за појаву истраживања података као посебног поља деведесетих година 20. века.

У другој половини двадесетог века су истраживачи и програмери почели да препознају потенцијал рачунарских техника за извлачење вредних информација из великих скупова података. Овај процес је 1989. године првобитно назван „откривање знања у базама података“ (енгл. *Knowledge Discovery in Databases*) [10], да би се потом, током деведесетих година, проширио назив „истраживање података“, односно „рударење података“ (енгл. *data mining*).

Појава великих података (енгл. *big data*) на почетку 21. века, а као последица експоненцијалног раста података генерисаних на разне начине (на друштвеним мрежама, помоћу сензора, у трансакционим базама података, итд), довео је до развоја напредних техника и алгоритама за обраду великих и сложених скупова података. Неки од кључних трендова и достигнућа у овој области укључују све већу употребу дубоког учења и неуронских мрежа, интеграција са другим дисциплинама (као што су обрада природног језика и рачунарски вид), али и развој техника за обраду токова података и података временских серија. Све указује на то да ће будућност рударења података вероватно бити обликована напретком у рачунарској снази и складишном капацитету, као и растом количина података из различитих извора.

2.2. Фазе истраживања података

Процес истраживања података може се поделити на неколико фаза:

- прикупљање и предобрада података,
- трансформација и редукција података,
- моделовање процеса,
- одабир и издвајање карактеристика,
- оптимизација и оцењивање модела.

У наставку је дат преглед наведених фаза, при чему су истакнути кључни концепти и технике повезане са сваким кораком. Овај процес је итеративан и углавном захтева неколико пролазака кроз различите фазе да би се модели усавршили и резултати побољшали. Разумевање сваке фазе и међузависности између њих омогућава да се донесу исправне одлуке и помаже развијању ефикасних стратегија за издвајање вредних увида из података.

2.2.1. Сакупљање и предобрада података

Попут људског учења, и машинско учење није могуће без података из којих могу извлачити закључке. Људска бића могу без проблема да обраде неструктуриране податке, попут текста, слике или звука слободног обрасца. С друге стране, рачунари захтевају да улазни подаци буду структурирани, тј. потребно је да сваки пример има исте карактеристике, при чему и те карактеристике треба да буду организоване у облику који рачунар може да разуме. Дакле, пуко складиштење података није довољно, већ је потребно обезбедити виши ниво разумевања података, односно увид у шири контекст који они носе. Неопходна је селекција и одабир репрезентативних информација, као и разумевање међусобних веза тих информација и на који се начин донети закључци могу интерполирати на претходно невиђене

ситуације. На овај начин је могуће препознати шире обрасце који постоје у подацима, а тиме се смањује количина података која је неопходна за анализу.

Прва фаза у процесу обраде података је прикупљање и предобрада података. Предобрадом података се може осигурати да су они високог квалитета и прикладни за анализу, што је кључно за успех процеса истраживања података. Ова фаза укључује сакупљање сирових података из различитих извора и њихову припрему за анализу. Прикупљање података се може обавити на различите начине, као што је прикупљање података са Веба (енгл. *web scraping*), екстраховање података из база података, или прикупљање података са сензора и разних уређаја.

Након што су подаци сакупљени, потребно их је обрадити како би се осигурао њихов квалитет и прикладност за анализу. Предобрада података се састоји од неколико корака, укључујући чишћење података, интеграцију података, трансформацију података и редукцију података [11]. Чишћење података је процес идентификације и исправљања грешака, недоследности и нетачности у подацима. Интеграција података је процес комбиновања података из више извора у јединствени, конзистентни скуп података. Ово може укључивати решавање конфликата у представљању података, спајање скупова података који имају различите шеме или агрегацију података.

2.2.2. Одабир и екстракција карактеристика

Избор и екстракција карактеристика имају за циљ идентификацију најрелевантнијих карактеристика или променљивих у скупу података.

Избор карактеристика је процес претварања сирових података у податке који се могу користити за обучавање машинског учења. Овај процес укључује избор подскупа најрелевантнијих карактеристика из оригиналног скупа података. Ово се може учинити користећи различите методе, као што су филтер методе, методе омотача или уграђене методе (енгл. *embedded methods*). Филтер методе

процењују релевантност сваке карактеристике независно, на основу њене корелације са циљном променљивом или њеног односа са осталим карактеристикама. Методе омотача користе тај подскуп да би проценили релевантност подскупа карактеристика изградњом и оценом модела. Уграђене методе комбинују процес одабира карактеристика са процесом учења модела, одабирајући најрелевантније карактеристике током процедуре фитовања модела. Од процеса издвајања карактеристика зависи квалитет података који се уводе, а самим тиме и квалитет исхода истраживања података.

2.2.3. Трансформација и редукција података

Фаза трансформације и редукције података има задатак превођења претходно обрађених података у формат погодан за анализу и смањење њихове димензионалности. Трансформација података подразумева пребацивање у формат који лако могу да разумеју и анализирају алгоритми за рударење података. Овде је неопходно апстраховати податке тако да се сировим подацима представе неке шире и апстрактније идеје и концепти. Овај процес може укључивати нормализацију података, дискретизацију континуалних променљивих или кодирање категоријских података [12].

Нормализација података је процес скалирања података тако да се налазе унутар одређеног опсега, као што су нпр. [0, 1] или [-1, 1]. Ово је важно јер су алгоритми за истраживање података често осетљиви на скалу улазних променљивих. Дискретизација континуалних променљивих подразумева њихово претварање у дискретне променљиве тако што се њихов опсег подели на коначан број интервала. Ово може помоћи у смањењу сложености података и учинити их погоднијим за анализу одређеним алгоритмима за истраживање података. Кодирање категоријских података подразумева њихово претварање у нумеричке вредности које лако могу бити обрађене алгоритмима за рударење података; пример ових вредности су бинарне или редоследне тј.

ординалне вредности (тј. променљиве са категоријама које припадају уређеној листи).

Редукција података је процес смањења величине скупа података, при чему се води рачуна да се одржи њихов интегритет и употребљивост за анализу. Током овог процеса рачунар сумира ускладиштене податке користећи одабрани модел за описивање образаца унутар података. Редукција се може постићи различитим техникама, као што су одабир карактеристика, редукција димензионалности, или компресија података [12]. Одабир карактеристика подразумева избор подскупа најрелевантнијих карактеристика из оригиналног скупа података. Технике смањења димензионалности, као што су анализа главних компоненти (енгл. *Principal Component Analysis* (PCA)) или сингуларна декомпозиција вредности (енгл. *Singular Value Decomposition* (SVD)), подразумевају трансформацију оригиналног скупа података у простор смањене димензионалности, при томе се води рачуна о очувању његове суштинске структуре и веза. Ове технике укључују проналажење линеарне или нелинеарне трансформације оригиналног скупа података која чува важне, а уклања нерелевантне информације. Технике компресије података, као што су компресије са губицима (енгл. *lossy*) и без губитака (енгл. *lossless*), подразумевају смањење величине скупа података уклањањем редундантних информација.

2.2.4. Моделовање процеса

Фаза моделовања процеса обухвата сагледавање проблема у склопу математичког оквира и рачунарских алгоритама. Модел је математичка представа реалног процеса, има задатак да идентификује шаблоне у улазним подацима у циљу предвиђања или доношења одлука.

Одабир одговарајућег алгоритма је суштински предуслов за успешно истраживање података и зависи од природе проблема који се посматра (нпр. класификација, регресија, кластеровање), од сложености података (нпр.

структуирани и неструктуирани, континуални и категоријски), и доступних рачунарских ресурса. Након одабира алгоритма, следећи корак је изградња модела. У оквиру изградње модела, одабрани алгоритам се конфигурише пажљивим одабиром неопходних параметара (нпр. коефицијенти полинома у нелинеарној регресији или тежине и пристрасност у неуронској мрежи) и хиперпараметара (нпр. брзина учења или избор активационе функције неуронске мреже). Одабир погрешних вредности параметара може довести до превеликог прилагођавања или до недовољног обучавања.

Након што је одабран одговарајући алгоритам, изграђује се модел користећи податке за обуку. Модел се тренира прилагођавањем његових параметара како би се минимализовала грешка између његових предвиђања и стварних исхода у подацима за обуку. Овај процес је познат као прилагођавање, односно „фитовање“ модела илити учење модела (енгл. *model learning*).

2.2.5. Тренирање и тестирање скупова података

Да би се оценила ефикасност модела истраживања података, неопходан је скуп података који се може користити за мерење његове прецизности и способности генерализације. Само правилан одабир хеуристике које користе алгоритми машинског учења може довести до тачних закључака. За алгоритам се каже да има пристрасност (енгл. *bias*) уколико су закључци систематски погрешни, тј. уколико су погрешни на неки предвидљив начин. Стога је важно осигурати да скупови за обуку модела буду репрезентативни у односу на расподелу података и да не уводе било какве предрасуде које би могле утицати на резултате евалуације.

Обрасци које модел открије нису нужно ни робусни, ни валидни. Пронађени обрасци могу бити карактеристични само за податке коришћене за обуку модела, а који не постоје у независним изворима података. Овај феномен постоји у различитим облицима у различитим задацима истраживања података: у моделирању је познат као „прекомерно прилагођавање“ (енгл.

overfitting), док се код претраживања образаца и статистичком тестирању појављује у виду „лажно позитивних резултата“. За све ове случаје је заједничко то што се над анализираним подацима добијају веома добри резултати, а лоши када се модел примени над непознатим подацима.

Постоје различити приступи којима се гарантује валидност и робусност резултата. Један приступ је да се подаци поделе у скуп за обуку и скуп за тестирање. Над скупом за обуку се користи одабрана метода истраживања података, док се скуп за тестирање користи за оцену његових перформанси над подацима који су за модел до тада невиђени. Неки типови података дозвољавају коришћење метода рандомизације, где се у обуци користе подаци у целини, али се обучени метод потом примењује над великим бројем рандомизованих верзија оригиналних података.

Имплементација скупова података за обуку и тестирање је кључна за мерење перформанси алгоритама и модела истраживања података. Скуп података за обуку, илити тренинг скуп, је у основи подскуп оригиналног скупа података који се користи за тренирање машинских модела учења. Овај скуп олакшава учење веза и шаблона унутар скупа података чиме се омогућава предвиђање исхода за нове инстанце података. Математички изражено, дати скуп података D се дели на два подскупа: *D_train* и *D_test*, где је *D_train* скуп података за обуку, а *D_test* скуп података за тестирање. Ова подела се обично врши насумично, али се притом води рачуна да се очува расподела података која постоји код оригиналног скупа.

С друге стране, скуп података за тестирање се састоји од остатка података, оних који нису укључени у скуп података за обуку. Овај подскуп се користи за процену перформанси модела, односно да се тестира способност модела да генерализује своје учење на до тада невиђене податке. Иако је *D_test* обично мањи по величини у односу *на D_train*, он служи као кључна мера реалних перформанси модела.

Постоје бројне праксе које се користе за поделу података на ове скупове података. Процес дељења података на скупове за обуку и тестирање може се

извести помоћу различитих техника, као што су случајно узорковање, стратификовано узорковање или дељење засновано на времену. Један конвенционалан приступ је метода *Holdout*, где се оригинални скуп података дели на два дисјунктна подскупа. Пропорција је често одређена величином и природом података, али је уобичајена пракса коришћење 70-80% података за тренирање, а остатка за тестирање. Овај случајни узорак осигурава очување расподеле података, која је математички изражена као:

$$D\_train \cup D\_test = D$$
$$D\_train \cap D\_test = \emptyset \qquad (2.1)$$

У контексту регресионих проблема, а нарочито нелинеарне регресије, постоје специфичне праксе. Нелинеарни регресиони проблеми укључују предвиђање континуалне одредишне променљиве на основу једне или више предикторских променљивих, при чему је веза између предиктора и исхода нелинеарна. Једна од кључних ствари је осигуравање прикладности модела. Пошто су нелинеарне везе сложене и разнолике, потребно је темељно разумевање података и основних процеса да би се одредио прикладан модел. Штавише, због сложености модела, можда ће бити потребан већи скуп података да би се „ухватиле“ све нијансе које постоје у односима међу подацима. Нелинеарна регресија може често довести до претераног прилагођавања јер модели покушавају да што боље фитују податке за обуку, притом се прилагођавајући шуму и аномалијама у подацима, што може негативно утицати на перформансе када се модел примени на невиђеним подацима. Стога су робусне технике унакрсне валидације неопходне за процену перформанси модела.

2.2.6. Оптимизација

Оптимизација има за циљ да финим подешавањем параметара и поставки побољша перформансе и тачност алгоритама и модела истраживања

података. Још један циљ оптимизације јесте и смањивање времена извршења. Технике оптимизације могу се примењивати у току различитих стадијума процеса обраде података, нпр. током одабира карактеристика, обуке модела и процене модела.

Постоји неколико техника оптимизације које се могу користити у процесу истраживања података, укључујући:

1. **Мрежна претрага** (енгл. *grid search*): Мрежна претрага је једноставна и широко коришћена техника оптимизације која укључује исцрпну претрагу кроз предефинисани скуп вредности параметара да би се нашла најбоља комбинација њихових вредности. Простор за претрагу дефинисан је као мрежа, при чему свака димензија представља различит параметар, а свака тачка у мрежи представља одређену комбинацију вредности параметара. За сваку комбинацију параметара је потребно оценити перформансе модела, те се на крају процеса добиј најбоља комбинација параметара (у складу с одабраном метриком перформанси).

2. **Случајна претрага** (енгл. *random search*): Случајна претрага је алтернатива мрежној претрази, код ње се узима случајно узорковање вредности параметара из предефинисаног скупа за претрагу. Овај приступ може бити ефикаснији од мрежне претраге јер не захтева оцену свих могућих комбинација вредности параметара. Уместо тога, он се ослања на претпоставку да се најбоља комбинација вредности параметара може пронаћи узорковањем довољног броја случајних тачака у простору за претрагу.

3. **Бајесова оптимизација**: Бајесова оптимизација је напреднија техника оптимизације која подразумева изградњу пробабилистичког модела циљне функције. Тај модел се потом користи за потрагу за најбољом комбинацијом вредности параметара. Модел се итеративно ажурира на основу перформанси процењених комбинација, а претрага се врши вођена проналажењем равнотежног односа између истраживања (узорковање тачака у неистраженим регионима простора за претрагу) и експлоатације (узорковање тачака с високим предвиђеним перформансама).

4. **Генетски алгоритми**: Генетски алгоритми су врста техника оптимизације који су инспирисани процесом природне селекције. Они подразумевају одржавање популације кандидата за решења, а који су представљени у виду хромозома или низова вредности параметара. Популација еволуира током времена пролазећи кроз процес селекције, укрштања и мутације, при чему најприлагођеније индивидуе имају и највећу вероватноћу да ће се размножавати и преносити своје гене на следећу генерацију. Овај процес се наставља све док се не достигне предефинисани критеријум за заустављање, као што је максималан број генерација или циљни ниво перформанси.

5. **Оптимизација заснована на градијенту**: Технике оптимизације засноване на градијенту, као што су градијентни спуст (енгл. *gradient descent*) или стохастички градијентни спуст, укључују итеративно ажурирање вредности параметара на основу градијента циљне функције у односу на параметре. Ове технике су посебно прикладне за оптимизацију циљних функција које се могу оценити и прилагодити, као што су оне на које се наилази при тренирању неуронских мрежа.

2.2.7. Евалуација модела

Последњи корак у процесу учења јесте процена, тј. евалуација, његовог успеха и мерење постигнутог учинка.

Информације добијене у фази евалуације се могу искористити и за доношење одлуке о додатној обуци, уколико се оцени да је то потребно. Фаза евалуације омогућава процену перформансе модела и алгоритама, идентификовање потенцијалних проблема и доношење информисаних одлука о одабиру и прилагођавању модела. Евалуација модела почива на тачној и смисленој примени ових евалуационих метрика. Постоје бројне метрике, свака са својом засебном дефиницијом и применом, а њихове области примене зависе од типа проблема и циљева истраживања. Овај одељак пружа преглед кључних концепата и техника укључених у процес евалуације, укључујући ту обуку и

тестирање, метрике евалуације, проблеми превелике и недовољне подешености, избор модела и прилагођавање модела.

2.2.7.1. Метрике евалуације

Евалуационе метрике играју кључну улогу у развоју и оптимизацији модела истраживања података. Њихова примарна функција је да објективно процене перформансе модела у односу на његов специфични задатак, омогућавајући на тај начин доношење информисаних одлука о избору и оптимизацији модела. Ефикасна употреба ових метрика захтева разумевање њихових карактеристика, њихове релевантности за одређени домен проблема и њихових одговарајућих снага и ограничења. Избор одговарајуће метрике утиче на ток истраживања података, на квалитет резултујућег модела и, на крају, на увиде изведене из података. У зависности од конкретног проблема, постоји више метрика евалуације које се могу користити за процену перформанси модела истраживања података.

Што се тиче задатака регресије, метрике евалуације служе за мерење неслагања између предвиђања модела и стварних вредности. Ове метрике наглашавају различите аспекте грешака у предвиђању: неке су осетљиве на веће грешке због операције степеновања, док друге пружају директну интерпретацију просечне грешке. У већини случајева се базирају на поређењу процењених вредности модела и стварних вредности посматране функције. Избор метрике регресије треба да буде усклађен са циљевима задатка регресије и природом дистрибуције грешке. Усвојићемо следеће ознаке које ће бити коришћене у наведеним метрикама:

N — укупан број инстанци,

y_i — вредност i-те инстанце одређене математичким моделом,

а $\widehat{y_i}$ — вредност i-те инстанце добијене мерењем.

Најважније метрике задатака регресије [13], [14] су:

- **Средња апсолутна грешка** (енгл. *Mean Absolute Error* (MAE)) је метрика евалуације која се обично користи код проблема регресије. Она рачуна

средњу апсолутну разлику између предвиђених и стварних вредности, где нижа вредност указује на бољи учинак. MAE је нарочито користан у случајевима када велике грешке нису знатно штетније од малих. Математички се изражава као:

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}|\mathrm{y_i} - \widehat{y_\mathrm{i}}| \qquad (2.2)$$

- **Средње квадратна грешка** (енгл. *Mean Square Error* (MSE)) је метрика која рачуна средњу вредност квадрата разлика између предвиђених и стварних вредности. Мања MSE указује на боље подударање. Математички се изражава као:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \qquad (2.3)$$

- **Корен средње квадратне грешке** (енгл. *Root Mean Squared Error* (RMSE)) се често користи код проблема регресије. Рачуна се као квадратни корен средње квадратне разлике између предвиђених и стварних вредности, где нижа вредност указује на бољи учинак. У односу на MAE, ова метрика више кажњава велике грешке. Математички се изражава као:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2} \qquad (2.4)$$

- **Резидуална сума квадрата** (енгл. *Residual Sum of Squares* (RSS)) показује укупну варијансу у резидуалима, односно укупну разлику између стварних и предвиђених вредности. Модел са мањом RSS обично указује на боље фитовање података. Математички се дефинише као:

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \qquad (2.5)$$

- **Релативна апсолутна грешка** (енгл. *Relative Absolute Error* (RAE)): Ово је метрика која показује укупну апсолутну грешку у односу на укупну апсолутну грешку коју производи једноставан (наиван) модел (ово је обично средња вредност стварних вредности). RAE нема димензије и обично се изражава у виду процента. Уколико је модел исправан, производи резултате који су бољи од резултата наивног модела. Математички израз гласи:

$$RAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} |y_i - \bar{y}|} \qquad (2.6)$$

- **Коефицијент одређености** (енгл. *R-squared*): R^2 мери пропорцију варијансе зависних променљивих које се могу предвидети помоћу независних променљивих. Вредност 1 указује на савршено подударање, а 0 указује на то да модел нема могућности да објасни било какву варијацију излазне променљиве. Математичка формула R^2 гласи:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \qquad (2.7)$$

- **Прилагођен R^2** (енгл. *Adjusted R-squared*): Ово је модификација коефицијента одређености, од њега се разликује по томе што је прилагођен броју предиктора у моделу. За разлику од R^2, прилагођен R^2 се увећава само уколико нови фактор унапређује модел више него што би то учинила насумична варијација. Математичка формула је:

$$\overline{R^2} = 1 - (1 - R^2) \frac{N - 1}{N - p - 1} \qquad (2.8)$$

где је N број опсервација, а p је број предиктора.

- **Средња квадратна логаритамска грешка** (енгл. *Mean Squared Logarithmic Error* (MSLE)): MSLE је метрика која је нарочито корисна

када циљна променљива има широк опсег могућих вредности. Она мери средњу вредност квадрата логаритамских разлика између стварних и процењених вредности. MSLE је нарочито прикладан када се има експоненцијалан раст јер даје мање на снази екстремним разликама. Математички облик гласи:

$$MSLE = \frac{1}{N}\sum_{i=1}^{N}(\log(1 + y_i) - \log(1 + \hat{y_i}))^2 \qquad (2.9)$$

- **Медијална апсолутна девијација** (енгл. *Median Absolute Deviation* (MAD)): MAD је робусна метрика која се користи да се измери дисперзија у оквиру скупа података. Израчунава се као медијана апсолутних девијација у односу на медијану података. За разлику од стандардне девијације, MAD није осетљив на есктремне вредности у узорку, те је стога **поузданији код расподела сигнала.** Формула по којој се рачуна:

$$MAD = \text{median}(|y_i - \text{median}(y)|) \qquad (2.10)$$

- **Акаикеов информациони критеријум** (енгл. *Akaike Information Criterion* (AIC)) је један од најпознатијих критеријума одабира модела који пореди различите моделе и одабира онај који минимализује губитак информација. Тачније, AIC описује тачност фитовања датог статистичког модела и служи за процену његовог квалитета. AIC не даје одговор да ли је модел добар у апсолутном смислу, већ ће између више лоших модела само указати на то који је најмање лош. Да би се спречила преобученост модела, AIC пенализује већи број параметара. Као најбољи модел се одабира онај који има најмању вредност AIC критеријума. Формула по којој се рачуна:

$$AIC = 2k - 2\,ln(\hat{L}) \qquad (2.11)$$

где k представља број параметара модела, а L представља максималну вредност функције веродостојности модела.

- **Бајесов информациони критеријум** (енгл. *Bayesian Information Criterion* (BIC)) је сличан AIC, али се док AIC процењује квалитет модела, BIC се може искористити за одабир модела који има највећу вероватноћу да буде одговарајући за дати скуп података [15]. Још једна разлика између њих је у погледу кажњавања броја параметара. AIC кажњава само сложеност модела, док BIC узима у обзир и број узорака који се посматрају. Последично, BIC више кажњава комплексне моделе, те је нарочито погодан код модела код којих не сме доћи до преобучености. Као најбољи модел се одабира онај који има најмању вредност AIC критеријума. Формула по којој се рачуна:

$$BIC = k\,ln(N) - 2\,ln(\hat{L}) \qquad (2.12)$$

  где k представља број параметара модела, L представља максималну вредност функције веродостојности модела, а N број узорака.

Поред наведених метрика, постоји и мноштво других метрика евалуације дизајнираних за више специјализоване задатке, као што су рангирање, груписање и предвиђање временских серија. Ове метрике су дизајниране да обухвате нијансе и сложеност ових специјализованих задатака. Као и код сваке евалуације, избор и тумачење ових специјализованих метрика захтевају јасно разумевање циљева и ограничења специфичних за задатак. Коришћење ових напредних метрика наглашава сложеност и разноликост области машинског учења, као и важност пажљивог одабира метрика у потрази за проницљивим и ефикасним истраживањем података.

2.2.7.2. Избор Модела

Избор модела подразумева издвајање модела са најбољим перформансама из скупа потенцијалних кандидата, при чему су сви модели примењени на скупу тестних података. Главни циљ овог процеса је да се изабере модел који обезбеђује најбоље могуће предвиђање или процену у складу са ограничењима проблема, тако да се пронађе равнотежа између сложености модела и његових перформанси. Ово се може урадити коришћењем

различитих техника, као што је Бајесов избор модела, унакрсна валидација, или бутстраповање.

Унакрсна валидација (енгл. *cross-validation*) је једноставна и свестрана техника која се често користи када се ради са малим скуповима података. Она функционише тако што се скуп података подели на неколико подскупова. Ови подскупови су дисјунктни и имају двоструку улогу, наизменично служе и као скупови за обуку и за валидацију. Један подскуп се издваја за процену перформанси модела кандидата, док преостали подскупови служе за обуку. Ово се понавља за све подскупове, а перформансе (обично изражене као средња квадратна грешка или логаритамски губитак) се израчунавају као просек свих добијених перформанси. Модел који даје најнижи просек се обично сматра супериорнијим. На пример, у *k-Fold* унакрсној валидацији, оригинални скуп података се дели на 'k' једнаких подскупова. Затим се модел тренира на 'k-1' подскупова, а остали подскуп се користи за тестирање. Овај процес се понавља k пута, са сваким подскупом тачно једном искоришћеним за тестирање. Перформансе крајњег модела су просек перформанси свих k модела. Математички, ако је дат скуп података D подељен на k подскупова S_1, S_2, ..., S_k, тада се перформансе P модела M рачунају као:

$$P = (1/k) * \sum P(M_{train}(S - S_i), S_i) \qquad (2.13)$$

Где је S_i тестни скуп у i-тој итерацији, $S - S_i$ представља преостале подскупове коришћене за тренирање, а $P(M_{train}(S - S_i), S_i)$ је перформанса модела трениранога на $S - S_i$ и тестираног на S_i.

Бутстраповање (eng. *bootstraping*) је метода генерисања нових узорака која пружа оцену тачности (пристрасност, стандардна грешка, интервал поузданости) процењених параметара. За разлику од традиционалних метода, код бутстреповања нема претпоставки о расподелама грешака или облику података, па је погодан за рад са комплексним скуповима података. Бутстраповање врши генерисање нових реплика оригиналног скупа података методом насумичног узорковања са заменом. Сваки поново узорковани скуп

података се користи за процену модела, а резултујући скуп процењених параметара се користи за формирање расподеле бутстраповања.

Бајесов избор модела је пробабилистички приступ који надограђује једноставно поређење фитовања тако да се укључи и кажњавање сложености. Ова метода примењује Бајесову теорему за израчунавање постериорних вероватноћа сваког модела према подацима. Модел са највећом постериорном вероватноћом бива одабран. Предност овог приступа јесте то што спречава одабир превише комплексних модела који имају лошу способност генерализације.

Код регресионих проблема, а нарочито код нелинеарне регресије, поступак одабира модела је још сложенији. Код НР се посматрани подаци моделују функцијом која је нелинеарна комбинација параметара модела и зависи од једне или више независних променљивих. Подаци се уклапају методом узастопних апроксимација. Избор модела у овом случају захтева пажљивији приступ јер се у раду са нелинеарним моделима мора водити рачуна о стварима као што су: прекомерно прилагођавање, локални минимуми и сложене интеракције параметара. Стога, када се ради о нелинеарној регресији, у процес одабира модела није довољно узети у обзир само учинак на тестном скупу података. Овде се морају укључити додатни критеријуми, као што су кажњавање сложености и моћ предвиђања ван узорака. Један од најчешћих приступа избора модела код НР јесте коришћењем информационих критеријума, нарочито AIC и BIC критеријума. Ове метрике одабирају моделе који могу пронаћи баланс између фитовања и сложености модела.

Још један ефикасан приступ у нелинеарној регресији је употреба техника регуларизације, као што су гребен (енгл. *ridge*) регресија или ласо (енгл. *lasso*) регресија, које могу да се носе са димензионалностима и колинеарношћу међу предикторима. Ове методе проширују функцију губитка кажњавањем, при чему Гребен додаје казну која је сума квадрата коефицијената (Л2 регуларизација), док Ласо која је једнака апсолутној вредности коефицијената (Л1 регуларизација). Увођење ових казни смањује сложеност модела тиме што

кажњава високе вредности коефицијената, а награђује вредности блиске нули. Ово олакшава одабир карактеристика и ублажава прекомерно прилагођавање.

Стога, успешан избор одговарајућег модела у нелинеарној регресији захтева сложену фузију ових метода. Узимајући у обзир не само фитовање модела, већ и његову сложеност, стабилност и интерпретабилност, може се обезбедити свеобухватније разумевање нелинеарних односа у подацима и робусне предиктивне перформансе.

2.2.7.3. Прилагођавање Модела

Прилагођавање модела укључује прилагођавање параметара и поставки модела да би се побољшале његове перформансе на скупу за тестирање. Постоје бројне методе којима се ово може извршити, као што су мрежна претрага, случајна претрага, Бајесова оптимизација или генетски алгоритми, како је објашњено у одељку о Оптимизацији.

Ово је итеративан процес код кога се пажљивим оцењивањем перформанси модела и алгоритама пружају вредни увиди у предности и мане различитих модела и алгоритама. Пролазећи кроз процес избора и прилагођавања модела, у крајњој линији се долази ка најефикаснијем решењу у датом случају.

2.3. Валидација и тестирање алгоритама

Валидација и тестирање алгоритама је важан аспект анализе сигнала јер се овим осигурава прецизност и поузданост резултата. У наставку ће бити образложени критеријуми евалуације успешности алгоритама, као и метода за тестирање великог броја генерисаних сигнала. Ово пружа свеобухватно разумевање перформанси алгоритма и њихову применљивост на проблеме у реалним процесима.

Валидација модела омогућава квантификацију перформанси алгоритма у погледу њихове способности да тачно изврше задатак препознавања. У ову

сврху се могу користити различите метрике, укључујући тестове адекватности, процене грешке, као и методе за избор модела. Упоређивањем перформанси различитих алгоритама користећи ове метрике, могуће је одредити најефикаснији приступ.

Поред оцене модела, важно је разматрати и робусност и општу применљивост алгоритама. Ово се може постићи коришћењем техника унакрсне валидације, које укључују поделу података на скупове за обуку и за тестирање, као и оцену перформанси алгоритама применом над оба скупа. Проценом перформанси алгоритама над непознатим подацима се избегава његово претерано прилагођавање подацима за обуку, тј. унакрсна валидација омогућава проверу да ли је алгоритам погодан за примену и у новим ситуацијама.

Још један важан аспект валидације алгоритма је тестирање великог броја генерисаних сигнала. Ово се може постићи коришћењем студија симулација, које укључују генерисање псеудо-случајних сигнала са познатим расподелама и параметрима и оцену перформанси алгоритама у препознавању ових дистрибуција и процени њихових параметара. Симулације пружају контролисано окружење за тестирање алгоритама, омогућавајући идентификацију потенцијалних слабости и могућности за побољшање.

2.4. Препознавање дистрибуција и њихових параметара

Препознавање дистрибуција и процена вредности њихових параметара је суштински аспект наше анализе улазних сигнала, њоме се омогућава идентификација основних статистичких модела који одговарају подацима. За ову сврху могу се употребити различите методе машинског учења, укључујући линеарну и нелинеарну регресију, као и неуронске мреже.

Расподеле вероватноће се често користе за моделовање универијантних података. Ово се може урадити у две фазе:

1. Проналажење расподеле која најбоље фитује податке.

2. Процена параметара (нпр. параметре локације, облика и скале) за одабрану расподелу вероватноће.

Постоји неколико различитих приступа за процену параметара расподела вероватноће.

2.4.1. Метода момената

Метода момената користи **узорке момената** за процену параметара. Метода момената је погодна за примену услед своје једноставности. Међутим, оно што отежава њихову употребу јесте то што нису увек доступне, а такође им недостају и неке од пожељних особина које одликују методе процене максималне вероватноће и процене најмањих квадрата.

2.4.2. Процена максималне вероватноће

Полазна тачка процене максималне вероватноће јесте математички израз који се назива функција вероватноће података. За дати модел вероватноће функција вероватноће представља вероватноћу проналажења скупа података. У овом изразу постоје непознати параметри. Процена максималне вероватноће (MLE) се односи на скуп параметара који максимизирају вероватноћу узорка [16].

Ова метода има пожељна математичка својства и својства оптималности метода максималне вероватноће. Када се величина узорка повећа, методе максималне вероватноће постају непристрасни естиматори са минималном варијансом. Методе макималне вероватноће имају приближно нормалне расподеле и приближне варијансе узорка, што се може искористити за генерисање граница поузданости параметара и тестова статистичких хипотеза.

Један од недостатака ове методе јесте што формуле вероватноће морају бити тачно дизајниране за дату расподелу и проблем процене параметара. Математички апарат који стоји иза овога је обично компликован, посебно ако

су потребни интервали поверења за параметре. Поред тога, није једноставно извести нумеричку процену вредности. Иако у неким случајевима постоје једноставне формуле максималне вероватноће, процене максималне вероватноће је обично најлакше добити коришћењем висококвалитетног статистичког софтвера. Још једна од мана процене максималне вероватноће јесте то што она над малим узорцима може бити веома пристрасна. Коначно, процена максималне вероватноће може бити подложна утицају одабраних почетних вредности [16].

2.4.3. Најмањи квадрати

Метода нелинеарних најмањих квадрата може бити замена за методу процене максималне вероватноће. Једна од предности овог приступа је то што већина статистичких софтверских пакета подржава ову методу. Ова метода се може користити чешће у поређењу са проценом максималне вероватноће. Уколико одабрани софтвер подржава нелинеарно фитовање и омогућава дефинисање одабране расподелу вероватноће, тада се помоћу методе најмањих квадрата могу генерисати процењене вредности за дату расподелу. Ово омогућава добијање разумне процене расподеле чак и ако софтвер не подржава процену максималне вероватноће.

С друге стране, једна од мана овог приступа је то што се метода најмањих квадрата не може лако применити на цензурисане податке. Такође, ова метода је јако подложна утицају одабраних почетних вредности [16].

2.5. Алгоритми истраживања података

Над пречишћеним и трансформисаним подацима потребно је и применити одговарајући алгоритам истраживања података. Сам одабир алгоритма зависи од циља који је постављен, при чему обично постоји више

одговарајућих решења која у већој или мањој мери могу предвидети зависне променљиве. Постоји много метода које се, са различитим циљевима, користе у пракси, а могу се грубо поделити на две главне категорије: методе које покушавају да моделују податке (тј. да науче глобалну структуру података) и методе које покушавају да пронађу обрасце који су корисни (тј. да науче локалне структуре из података). Са рачунске тачке гледишта, проблеми истраживања података су често НП-комплетни, што значи да немају оптималан алгоритам решења и да су им потребна апроксимирана решења.

Основна подела најважнијих техника истраживања података јесте на надгледано учење (енгл. *supervised learning*), ненадгледано учење (енгл. un*supervised learning*), полунадгледано учење и учење са подршком (енгл. *reinforcement learning*).

Алгоритми машинског учења су подељени у категорије према својој намени.

Предиктивни модел се користи за задатке који укључују, као што назив имплицира, предвиђање једне вредности користећи друге вредности у скупу података. Алгоритам учења покушава да открије и моделира однос између циљног обележја (обележја које се предвиђа) и других обележја.

Пошто се предиктивним моделима дају јасна упутства о томе шта треба да науче и како треба да то науче, процес обуке предиктивног модела је познат као надгледано учење. Под надгледањем се не мисли на људско учешће, већ на чињеницу да циљне вредности обезбеђују начин да рачунар сазна колико је добро научио жељени задатак. Формалније речено, за дати скуп података, алгоритам за учење под надзором покушава да оптимизује функцију (модел) да пронађе комбинацију вредности карактеристика које резултирају циљним излазом.

Класификација је још један од често коришћених алгоритама надгледаног машинског учења где се предвиђа којој категорији припада неки пример. Класификација потпада под две категорије. Прво, ако је доступно знање о груписању ентитета, као што је нпр. пол, онда се класификација може применити да би се пронашле ствари које групе чине различитим. Друго,

након што се такве теме идентификују, класификација може предвидети груписање нових ентитета. Примери класификације су предвиђања да ли је е-пошта непожељна, да ли пацијент има рак, да ли ће клијент исплатити кредит. У класификацији, циљна карактеристика коју треба предвидети је категоричка карактеристика позната као класа, која је подељена у категорије које се називају нивои. Класа може имати два или више нивоа, а нивои могу или не морају имати редослед. Класификација се толико користи у машинском учењу да постоји много типова класификационих алгоритама, са предностима и слабостима погодним за различите типове улазних података. Учење под надзором се може користити за предвиђање нумеричких података, као што су приход, лабораторијске вредности, резултати тестова или број предмета. Да би се предвиделе такве нумеричке вредности, уобичајени облик нумеричког предвиђања фитује моделе линеарне регресије на улазне податке. Иако регресија није једина метода за нумеричко предвиђање, она је далеко најчешће коришћена. Регресијске методе се користе за предвиђање у разним областима јер могу тачно да квантификују повезаност између улаза и циља, укључујући ту и величину и неизвесност односа. Уобичајене технике класификације су дискриминантна анализа, стабла одлучивања, методе најближег суседа, неуронске мреже и машине векторе подршке.

С обзиром на то да је лако претворити бројеве у категорије (на пример, узраста од 13 до 19 година су тинејџери) и категорије у бројеве (на пример, доделити 1 свим мушкарцима и 0 свим женама), граница између класификационих модела и модела нумеричког предвиђања није увек јасно дефинисана.

Дескриптивни модел се користи за задатке који би имали користи од увида стеченог сумаризацијом података на неке нове и занимљиве начине. За разлику од предиктивних модела који предвиђају неки циљ, у дескриптивном моделу ниједна појединачна карактеристика није важнија од било које друге. А пошто не постоји циљ за учење, процес обуке дескриптивног модела назива се учење без надзора. Они се прилично редовно користе за рударење података. На пример, задатак дескриптивног моделирања који се зове откривање образаца користи се за идентификацију корисних асоцијација унутар

података. Пример откривања образаца јесте анализа тржишне корпе, где трговци анализом трансакција клијената покушавају да идентификују артикле који се често купују заједно, тако да се те информације могу користити за прецизирање маркетиншких тактика. Даље, на пример, може се користити за откривање образаца лажног понашања, откривање генетских дефеката или идентификацију жаришта за криминалне активности.

Задатак дескриптивног моделирања дељења скупа података у хомогене групе назива се груписање (кластеровање). Груписање покушава да у подацима пронађе групе ентитета сличног или сличног по понашању. Уопштено говорећи, груписање је слепо за познате структуре података у смислу да не користи никакво знање изван података за додељивање ентитета кластерима. Ово се понекад користи за сегментациону анализу која идентификује групе појединаца са сличним понашањем или демографским информацијама како би их циљали рекламним кампањама на основу њихових заједничких карактеристика. Овим приступом, машина идентификује кластере, али је потребна људска интервенција да би се протумачили. На пример, с обзиром на пет кластера купаца у продавници, маркетиншки тим ће морати да разуме разлике међу групама како би направио промоцију која најбоље одговара свакој групи. Постоји велики избор метода груписања, али међу најпознатијима су *k-means* кластерисање и хијерархијско груписање.

На крају, постоји класа алгоритама за машинско учење познатих као мета-ученици која није везана за одређени задатак учења, већ је фокусирана на учење о ефикаснијем учењу. Алгоритам за мета-учење користи резултат претходног учења као основу за додатно учење. Ово обухвата алгоритме за учење који уче да раде заједно у тимовима који се називају ансамбли, као и алгоритме за које се чини да се временом развијају у процесу који се зове учење са подстицајем (енгл. *reinforcement learning*). Мета-учење може бити корисно за веома изазовне проблеме или када перформансе предиктивног алгоритма треба да буду што прецизније.

Следећа табела наводи неке од општих типова алгоритама машинског учења.

Табела 1. Општи типови алгоритама машинског учења и њихови задаци [17]

Модел	**Задатак учења**
Надгледани алгоритми учења	
k-најближих суседа	класификација
Наивни Бајес	класификација
Стабла одлучивања	класификација
Правила за учење класификације	класификација
Линеарна регресија	нумеричка предикција
Регресијска стабла	нумеричка предикција
Модел стабла	нумеричка предикција
Неуронске мреже	класификација, нумеричка предикција
Машине за подршку векторима	класификација, нумеричка предикција
Алгоритми ненадгледаног учења	
Асоцијативна правила	откривање образаца
k-меанс кластеровање	кластеровање
Мета-алгоритми учења	
Агрегација	класификација, нумеричка предикција
Појачавање	класификација, нумеричка предикција
Насумичне шуме	класификација, нумеричка предикција

За успешну примену машинског учења над неким реалним проблемом неопходно је утврдити којем типу учења тај проблем припада: класификацији, нумеричком предвиђању, откривању шаблона или груписању. Тип учења ће одредити и избор алгоритма.

2.6. Регресиона анализа

Регресија обухвата веома велику породицу метода. Општа идеја регресије је, просто речено, да се пронађе функција предикторских променљивих која одговара одзиву са најмањом грешком. Дакле, регресија се може користити за, на пример, проучавање одговора организма на дозу лека. Опсег метода варира од класичне линеарне регресије, преко генерализованих линеарних модела, па до генерализованих адитивних мешовитих модела (енгл. *additive mixed models*) и модела мешавине (енгл. *mixture models*).

Као што је речено изнад, регресија је изграђивање функције од независних променљивих (познатих и као „предиктори“) да би се предвидела зависна променљива (звана још и „одзив“). Регресиона анализа је статистичка метода за истраживање односа између варијабли, која укључује низ техника за моделирање и анализу неколико варијабли. Фокус је на односу између зависне променљиве и једне или више независних променљивих [18]. Регресиона анализа може осликати како се типична вредност зависне променљива мења када се неке од независних променљивих мењају, док се друге независне варијабле држе фиксним.

Генерално, постоје две врсте регресионе анализе које се разликују према томе да ли подаци апроксимирају линеарну или нелинеарну функцију, и то су: линеарна регресија и нелинеарна регресија. Један веома општи облик регресионог модела је:

$$Y = f(x_1, x_2, \cdots, x_n) + \varepsilon \qquad (2.14)$$

где је f нека непозната функција а ε грешка. Да би се извршила регресиона анализа, облик функције f мора бити назначен. Пошто обично немамо довољно података да покушамо директно да проценимо f, морамо претпоставити да оно има неки ограничени облик као у линеарном моделу:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \qquad (2.15)$$

где су $\beta_i = 0, 1, \ldots, n$ непознати параметри, а β_0 се назива термин пресека. У нелинеарним моделима, регресија покушава да опише однос између варијабли из математичке теорије са неким посебним карактеристикама.

Након што се одреди однос између променљивих, проблем се своди на процену параметара фитовањем доступних информација о овим променљивима. У свим случајевима, циљ процене јесте функција независних варијабли, такозвана функција регресије [19]. Процена најмањих квадрата је једна од широко коришћених метода. Такође, постоје различити стандарди за дефинисање добрих проценитеља и више метода за проналажење непристрасног проценитеља [19]. У регресионој анализи је од интереса и да се

окарактерише варијација зависне променљиве око функције регресије, а која се може описати неком расподелом вероватноће.

Регресиона анализа се широко користи за предвиђање. Такође се користи и да се разуме које од независних променљивих су повезане са зависном променљивом и да се истраже облици њихових односа [18].

2.6.1. Нелинеарна регресија

Под стандардном нелинеарном регресијом подразумевамо моделе који имају следећи облик:

$$y = f(\theta, x) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2) \qquad (2.16)$$

где је y одзив (зависна променљива), x је вектор састављен од p независних променљивих, θ је вектор састављен од k параметара модела, f је нека позната регресиона функција, а ϵ је резидуална грешка (за коју се претпоставља да има нормалну расподелу и да је центрирана око нуле са непознатом варијансом (σ^2)). Претпостављамо да су услови резидуалне грешке међусобно независни, као што се обично и претпоставља за стандардну НР анализу [20]. Параметри модела карактеришу однос између x и y преко функције f. Процена параметара НР модела је заснована на итеративној процедури која укључује линеарну апроксимацију, што доводи до решавања проблема најмањих квадрата у сваком кораку.

Примери неких од најчешће коришћених модела нелинеарне регресије су:

- Полиномна регресија — врста нелинеарне регресије која укључује додавање виших редова независних променљивих у линеарни модел. Ова регресија фитује закривљене или таласасте податке него права линија, али такође може да доведе до претераног или недовољног прилагођавања уколико је степен полинома превисок или пренизак.

- Експоненцијална регресија — врста нелинеарне регресије која укључује фитовање података експоненцијалном функцијом. Експоненцијална функција има облик $y = a * b^x$. Добро моделује податке који се брзо повећавају или смањују на почетку, а затим се стабилизују. Такође може да обради податке који имају константан процентуални пораст или мултипликативни однос.
- Логистичка регресија — врста нелинеарне регресије која укључује уклапање логистичке функције у податке. Логистичка функција има облик $y = 1 / (1 + e - (a + b * x))$, где су a и b константе, а e је основа природног логаритма. Логистичка регресија може да моделује податке који имају бинарни исход, као што су успех или неуспех, да или не, или 0 или 1. Такође може да обради податке који имају S-облик криве или ограничен опсег.
- Степена регресија — врста нелинеарне регресије која укључује уклапање степене функције у податке. Степена функција има облик $y = a * x^b$. Степена регресија може да моделира податке који имају промењљиву стопу промене или нелинеарни однос.

Нелинеарна регресија има широку примену у разним дисциплинама [21]. Као и код линеарне регресије, нелинеарна регресија пружа процене параметара коришћењем најмањих квадрата као критеријума одабира. Но ипак, за разлику од линеарне регресије, нема експлицитног математичког решења које је универзално, већ се морају примењивати одређени алгоритми за решавање проблема минимализације, укључујући ту и нумеричке апроксимације. Нажалост, нелинеарне везе између параметара чине да минимализација, а ни оптимизација у општем случају, није увек једноставна за такве моделе [22]. Последично, добијање корисних фитовања модела нелинеарне регресије може захтевати пажљиво одређивање детаља модела и евалуацију резултујућег излаза, а можда и коришћење статистичких модела који нису превише зависне од претпоставки модела.

Процена параметара нелинеарних модела се обично изводи помоћу неке варијанте критеријума најмањих квадрата. Овај процес се врши итеративно

све док се не добију (у идеалном случају) оптималне процењене вредности. Стога је неопходна *a priori* идеја о очекиваним вредностима параметара, знање о могућностима валидације предложеног модела (зарад провере кључних претпоставки), и разумевање каква је могућност валидације почетне хипотезе фитованог модела и утицај тога на прецизност процене параметара.

2.6.1.1. Нелинеарни најмањи квадрати

Уколико имамо n опсервација, једначина (2.16) се може записати на следећи начин:

$$Y_i = f(x_i, \theta) + \epsilon_i \qquad (2.17)$$

где је $i = 1,2, \dots, n$. Сума квадрата грешака нелинеарног модела се дефинише као:

$$S(\theta) = \sum_{i=1}^{n} \left(y_i - f(\mathrm{x}, \theta)\right)^2 \qquad (2.18)$$

С обзиром на то да су y_i и x везани за опсервације, сума квадрата је функција од θ. Означимо са $\hat{\theta}$ вредност θ која минимизује $S(\theta)$. Да бисмо пронашли $\hat{\theta}$, прво се морају наћи изводи од S(θ) у односу на θ. Затим се сви парцијални извода решавају у односу на θ, а параметри θ_i се замењују са $\hat{\theta}_i$. Функције које треба решити су нелинеарне у процењеним параметрима и често је тешко решити их, чак и у најједноставнијим случајевима. Стога се за процену параметара нелинеарних модела често користе итеративне нумеричке методе.

Неки од најчешће коришћених алгоритама за нелинеарно фитовање методом најмањих квадрата укључују:

- Њутнову методу — ово је класична метода заснована на методи градијентног спуста, али која може бити рачунарски захтевна и у великој мери зависна од добрих почетних вредности.

- Гаус-Њутнов алгоритам — ово је модификација Њутнове методе која даје добру апроксимацију решења до којег би Њутнова метода требало да стигне, али није загарантовано да ће конвергирати.
- Алгоритам Левенберг-Маркварт — метода која може да се избори са рачунарским потешкоћама које се јављају код других метода, али може да захтева проналажење оптималних вредности параметра.

2.6.1.2. Алгоритам Левенберг-Маркварт

Алгоритам Левенберг-Маркварт је итеративна метода која се користи за решавање проблема нелинеарних квадрата проналажењем минимума мултиваријантнтне функције. Опште је прихваћен и нашао је примену у многим пољима. Развили су га 1960-их година, независно један од другог, Кенет Левенберг и Доналд, овај алгоритам је комбинација две минимизационе методе, градијентни спуст и Гаус-Њутнова метода. Када је тренутно решење далеко од тачног, користи се метода најбржег спуста која осигурава конвергенцију погодним одабиром величине корака, иако се тако до конвергенције стиже споро. Гаус-Њутнова метода може дивергирати, али брзо конвергује у близини локалног или глобалног минимума [23].

Проблем који се решава овим алгоритмом се може изразити на следећи начин. Нека је модел који треба фитовати са подацима изражен као:

$$E(y) = f(x_1, x_2, \ldots, x_m;\ \beta_1, \beta_2, \ldots, \beta_k,) = f(\mathrm{x}, \beta) \qquad (2.19)$$

где су $x_1, x_2, \ldots, x_m$, независне променљиве, $\beta_1, \beta_2, \ldots, \beta_k$ су вредности k параметара, а $E(y)$ је очекивана вредност зависне променљиве y. Нека су подаци означени као:

$$(Y_i, X_{1i}, X_{2i}, \ldots, X_{mi}), \quad i = 1, 2, \ldots, n \qquad (2.20)$$

Потребно је израчунати вредности параметара које ће минимизовати израз:

$$\phi = \sum_{i=1}^{n} \left[Y_i - \hat{Y}_i\right]^2 \& = \; || \mathbf{Y} - \hat{\mathbf{Y}} ||^2 \qquad (2.21)$$

где $\hat{Y}_i$ представља вредност *y* предвиђену помоћу (2.19) за *i*-ту тачку у оквиру података.

Када су β параметри функције f линеарни, тада су контуре константе ϕ образују елипсоид, а када је f нелинеарно, тада су контуре изобличене у складу са степеном нелинеарности. Међутим, чак и код нелинеарних модела се контуре могу сматрати елиптичним у непосредној близини минимума ϕ.

Да би се одредиле почетне вредности параметара, функција се може минимизовати тако што се Гаус-Њутновом методом развије f у Тејлоров ред, тако се добија следећа једначина:

$$\langle Y(\mathbf{x}_i, \beta + \Delta\beta) \rangle = f(\mathrm{x}_i, \beta) + \sum_{j=1}^{k} \left(\frac{\partial f_i}{\partial \beta_j}\right) \Delta\beta_j \qquad (2.22)$$

или:

$$\langle \mathrm{Y} \rangle = \mathrm{f}_0 + P\Delta\beta \qquad (2.23)$$

Заграде <> у једначини (2.22) се користе да би се увела разлика у односу на предвиђања заснованих над стварним нелинеарним моделом. Вредност ϕ предвиђена на основу (2.22) се може записати као:

$$\langle \Phi \rangle = \sum_{i=1}^{n} [Y_i - \langle Y_i \rangle]^2 \qquad (2.24)$$

С обзиром на то да је Δβ приближно линеарно у (2.22), може се пронаћи стандардном методом најмањих квадрата постављањем $\frac{\partial \phi}{\partial \beta_j} = 0$ за све вредности j. Тако се Δβ може пронаћи решавањем следеће једначине:

$$A_{\Delta\beta} = \mathrm{g}, \qquad (2.25)$$

где

$$A^{[k\times k]} = P^T P, \qquad (2.26)$$

$$P^{[n\times k]} = \left(\frac{\partial f_j}{\partial b_j}\right), \quad i = 1,2,\dots,n; j = 1,2,\dots,k, \qquad (2.27)$$

$$\mathrm{g}^{[k\times 1]} = \left(\sum_{i=1}^{n} (Y_i - f_i)\frac{\partial f_i}{\partial b_j}\right), j = 1,2,\cdots,k, = P^T(\mathrm{Y} - \mathrm{f}_0). \qquad (2.28)$$

У пракси се показало корисним да се вредност β коригује само за део Δβ, у супротном екстраполација може бити ван региона где ф може бити адекватно представљена изразом (2.22), те може доћи до неуспеха конвергенције.

Метода градијента, с друге стране, једноставно праве одабира корак од тренутне вредности у правцу негативног градијента функције ф. Тако да:

$$\boldsymbol{\delta}_g = -\left(\frac{\partial \Phi}{\partial \beta_1}, \frac{\partial \Phi}{\partial \beta_2}, \cdots, \frac{\partial \Phi}{\partial \beta_k}\right)^T \qquad (2.29)$$

Методе најстрмијег спуста могу довести до веома споре конвергенције. Са овим методама, као и са методама Тејлоровог реда, неопходно је пажљиво контролисати величину корака једном када је правац вектора корекције утврђен. Чак и тада, спора конвергенција је пре правило него изузетак.

С обзиром на то да перформансе ових алгоритама умногоме зависе од исправног одабира почетних вредности параметара, Левенберг и Маркварт су предложили модификацију методе изложене изнад:

$$(A + \lambda I)\Delta\beta = \mathrm{g} \qquad (2.30)$$

где су λ Лагранжов мултипликатор, а I је јединична матрица.

Решење једначине (2.25) је инваријантно на линеарну трансформацију параметарског простора, за разлику од градијентних метода. Скалирање параметарског простора користи стандардне девијације извода $\frac{\partial f_i}{\partial b_j}$, над тачкама $i = 1, 2, \ldots, n$, трансформишући А матрицу у матрицу корелационих коефицијената међу $\frac{\partial f_i}{\partial b_j}$, чиме се побољшава нумеричка стабилност у линеарним проблемима најмањих квадрата.

Стога, дефинишемо скалирану матрицу A^* и скалирани вектор g^*:

$$A^* = \left(a^*_{j'}\right) = \left(\frac{a_{jj'}}{\sqrt{a_{jj}}\sqrt{a_{j'j'}}}\right) \qquad (2.31)$$

$$\mathrm{g}^* = \left(g^*_j\right) = \left(\frac{g_j}{\sqrt{a_{jj}}}\right) \qquad (2.32)$$

Решавајући за Тејлоров ред користећи једначину:

$$A^*\delta^*_t = \mathrm{g}^* \qquad (2.33)$$

важи да је:

$$\beta_j = \beta^*_j / \sqrt{a_{jj}} \qquad (2.34)$$

Из свега претходног је јасан изглед алгоритма, у r-тој итерацији се добија једначина:

$$\left(A^{*(r)} + \lambda^{(r)} I\right)\Delta\beta^{*(r)} = \mathrm{g}^{*(r)} \qquad (2.35)$$

Решавањем ове једначине за $\mathbf{g}^{*(r)}$ и коришћењем једначине (2.34) да се добије $\Delta\beta^{(r)}$. Коришћењем вектора:

$$\beta^{(r+1)} = \beta^{(r)} + \Delta\beta^{(r)} \qquad (2.36)$$

добија се нова сума квадрата $\Phi^{(r+1)}$. Од кључног значаја је одабир вредности $\lambda^{(r)}$ тако да важи:

$$\Phi^{(r+1)} < \Phi^{(r)} \qquad (2.37)$$

Јасно је да увек постоји $\lambda^{(r)}$ довољно велико да важи (2.37), осим ако је $\beta^{(r)}$ већ на минимуму Ф. Итеративним понављањем и корекцијом вредности $\lambda^{(r)}$ тако да се задовољити (2.37) долази се до брзе конвергенције алгоритма ка вредностима најмањих квадрата.

У свакој итерацији се жели минимизовати Ф у максималној околини где се апроксимира нелинеарна функција. Приликом одабира вредности $\lambda^{(r)}$ треба водити рачуна да метода Тејлоровог реда може конвергирати. Ово је нарочито важно у каснијим фазама процедуре конвергенције, када се траже вредности у непосредној близини минимума. Велике вредности $\lambda^{(r)}$ би стога требало користити само онда када је неопходно задовољити услов (2.37). Иако је тачно да $\Phi^{(r+1)}$ као функција од λ има минимум, она није погодан избор јер обично захтева знатно већу вредност λ него што је неопходно да се задовољи (2.37).

Маркварт [23] дефинише кораке алгоритма на следећи начин:

Нека је $\nu > 1$, и нека $\lambda^{(r-1)}$ означава вредност λ у претходној итерацији, иницијално можемо поставити $\lambda^{(0)} = 10^{-2}$. Рачунање $\Phi\left(\lambda^{(r-1)}\right)$ и$\Phi\left(\lambda^{(r-1)}/\nu\right)$ се врши у три корака:

1. Ако је $\Phi\left(\lambda^{(r-1)}/\nu\right) \leqq \Phi^{(r)}$, тада је $\lambda^{(r)} = \lambda^{(r-1)}/\nu$.
2. Ако је $\Phi\left(\lambda^{(r-1)}/\nu\right) > \Phi^{(r)}$ и $\Phi\left(\lambda^{(r-1)}\right) \leqq \Phi^{(r)}$, тада је $\lambda^{(r)} = \lambda^{(r-1)}$.
3. Ако је $\Phi\left(\lambda^{(r-1)}/\nu\right) > \Phi^{(r)}$ и $\Phi\left(\lambda^{(r-1)}\right) > \Phi^{(r)}$, сукцесивно множити λ са ν док за неко најмање w не буде $\Phi\left(\lambda^{(r-1)}\nu^{w}\right) \leqq \Phi^{(r)}$. Тада је $\lambda^{(r)} = \lambda^{(r-1)}\nu^{w}$.

Итерација је конвергирала када $\frac{|\Delta\beta^{(r)}|}{\tau+\left|\beta_j^{(r)}\right|} < \epsilon$, за свако j, за довољно мало $\epsilon > 0$ и за неко одговарајуће τ. Одабир ν се врши произвољно.

Описани алгоритам, попут метода градијента, има способност да конвергира од почетног нагађања које може бити ван региона конвергенције других метода. Истовремено, алгоритам, попут методе Тејлоровог реда, има способност брзог приближавања конвергентним вредностима када се нађе у њиховој околини. Дакле, алгоритам комбинује најбоље одлике својих претходника, избегавајући њихова ограничења.

2.7. Преглед алата за истраживање података коришћених у дисертацији

За предобраду и анализу података се обично користе скриптни језици (нпр. *Perl*, *Python*), интерактивна окружења за анализу података (нпр. *R Studio* и *Matlab*) или графички алати (нпр. *Weka*), као и неки специјализовани пакети за истраживање података. Према резултатима истраживања Kaggle[1], највеће заједнице посвећене машинском учењу и истраживању података, у 2024. години међу најпопуларније језике и алате истраживања података спадају Python, R, MATLAB, RapidMiner, Orange, Weka, TensorFlow и SAS VDMML.

У истраживању које је спроведено у оквиру ове докторске дисертације, аутор је користио алате са бесплатном лиценцом и отвореног кода: програмски језик R и радно окружење RStudio.

R [24] је бесплатно софтверско окружење за статистичка израчунавања и графику. Ово је интерпретирани језик који подржава и функционално и објектно-оријентисано програмирање. Развијен је од стране Роса Ихака и Роберта Џентлмена на Универзитету у Окланду на Новом Зеланду, а од 1997. године се развија од стране *R Core Development* тима. Доступан је у виду отвореног кода, лако је проширив помоћу бројних пакета и додатака који су

[1] www.kaggle.com

доступни у оквиру разних репозиторијума пакета, од којих је најважнији CRAN (Comprehensive R Archive Network) [25].

Већина функција у језику R су написане управо овим језиком, али могуће је повезати и функције које су написане у другим језицима, нпр. C, C++. Уз помоћ додатака, R се може повезати и са другим језицима и са разним системима за управљање базама података.

Иако је синтаксно комплекснији од Пајтона (који је водећи језик на пољу истраживања података) и има стрмију криву учења, језик R је направљен управо за истраживање података и подржава изузетно комплексне статистичке анализе. R садржи имплементације, понекад и у неколико верзија, свих често коришћених алгоритама истраживања података, као што су: класификација, регресиона стабла, кластер анализа, неуралне мреже, итд. Ови алгоритми су доступни у облику отвореног кода, а детаљна документација о овим R пакетима се може пронаћи страници CRAN.

RStudio је интегрисано развојно окружење (IDE) развијено од стране компаније Posit, отвореног је кода и бесплатан за некомерцијалну употребу, са потпуном подршком за R језик и компатибилношћу са свим R пакетима и алатима. Омогућава ефикасан рад са R-ом кроз низ функционалности које олакшавају писање, извршавање и дебаговање R кода, управљање пројектима, увоз и претраживање података, вршење анализа и креирање извештаја. Модуларног је дизајна, корисници могу истовремено прегледати више аспеката пројекта, на располагању су им уређивач кода, конзола за R, панели за приказ радних простора, историје, излаза, пакета, графикона и помоћних докумената, омогућавајући корисницима да прегледају више аспеката пројекта истовремено. RStudio такође подржава алате за рад у тиму, као што су контроле верзија, интеграције са системима за груписани рад и опцију за покретање R сесија на удаљеним серверима или кластерима. Нарочито је популаран међу истраживачима, аналитичарима података и статистичарима за ефикасно истраживање, манипулацију и визуелизацију података користећи R.

Поред наведеног, аутор је користио и Oracle Machine Learning for R (OML4R) за приказ како би се предложени алгоритам могао применити над великим скуповима података, где је скалабилност од суштинске важности. OML4R омогућава покретање R скрипти и коришћење R функција за машинско учење директно на Oracle 19c бази података, чиме се могу користити могућности које Оракл база података нуди за паралелну обраду података и машинско учење.

3. ПРОЦЕС ПРЕПОЗНАВАЊА РАСПОДЕЛА ВЕРОВАТНОЋЕ У КАНАЛИМА ФЕДИНГА И ЊИХОВИХ ПАРАМЕТАРА

Непрекидан развој различитих сервиса бежичних телекомуникационих система узрокује потребу за унапређењем њихових перформанси. Приликом проучавања перформанси бежичних телекомуникационих система највише пажње се посвећује обезбеђивању великих брзина преноса, великог капацитета канала и што већег домета везе са што мањом вероватноћом грешке. Простирање корисног сигнала кроз атмосферу прате разни нежељени ефекти. Најзначајнија сметња која се јавља приликом преноса дигитално модулисаних сигнала јесте појава промене нивоа корисног сигнала у времену - тј. фединг.

3.1. Фединг и моделовање пропагационих окружења

Пренос сигнала до пријемника је комплексан процес приликом кога долази до разних варијација у јачини сигнала. Ове варијације су познате као фединг (енгл. *fading*) и представљају један од основних разлога умањивања перформанси у телекомуникационим системима. Уколико се овом феномену приступи са аспекта проналажења тачне математичке карактеризације, комплексност тог приступа доводи у питање његову практичност. Ипак, уложени су значајни напори да се пронађу једноставни и тачни статистички модели који описују различите типове пропагационих окружења.

Карактеристике окружења кроз који се врши пренос сигнала утиче на јављање временских флуктуација анвелопе и фазе пренесеног сигнала. Ове сметње је јављају под утицајем одбијања радио сигнала од земљине површине или од

атмосферског омотача, што доводи до промене у поларизацији таласа, а последично до промена у јачини примљеног сигнала.

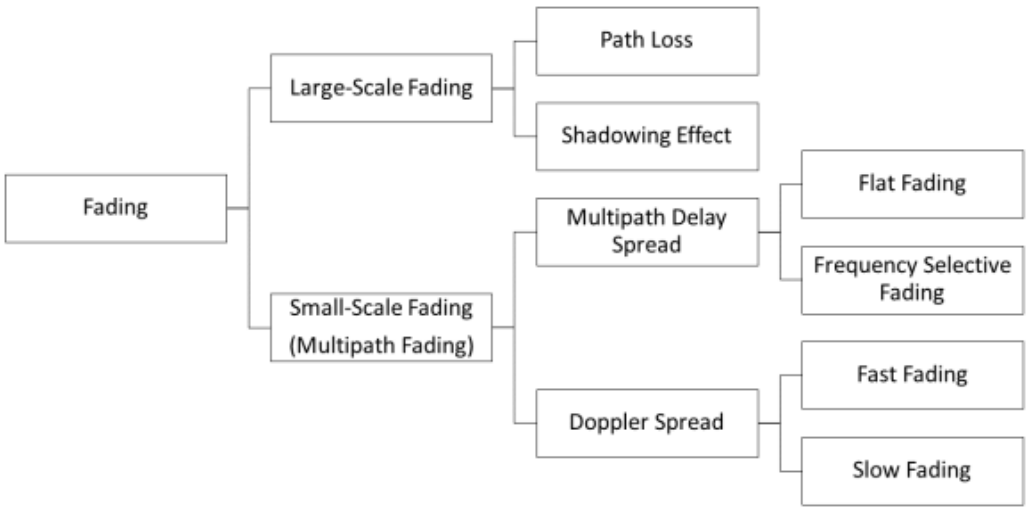
Брзи феднг (енгл. s*hort fading*) (познат и као кратки феднг, *multipath* губици) описује сложене феномене пропагације сигнала који су изазвани одбијањима посматраног сигнала од различитих објеката, а на такав начин да се на пријему случајно закашњене, рефлектоване и компоненте расејања сигнала комбинују на конструктиван, или, чешће, на деструктиван начин. Овај феномен узрокује да је временско трајање симбола знатно увећано, тако да функција корелације опада испод оптималних вредности [26], [27]. Поред овог феномена, приликом дизајнирања или развоја бежичног комуникационог система, мора се узети у обзир пропагација по више путања, позната као *multipath* пропагација. До овог типа пропагације долази јер се код бежичних комуникационих система често дешава да сигнал до пријемника не стиже само директним путем, већ приликом преноса стиже и сигнал одбијен од различитих „препрека“ које стоје између пријемника и базне станице (нпр. узвишења, грађевински објекти, итд). *Multipath* губици узрокују краткорочне варијације сигнала, а начин на који се мења анвелопа сигнала приликом излагања његовом утицају је моделиран помоћу неколико модела. У овом поглављу ће посебна пажња бити посвећена специфичним моделима који имају широку могућност примене. Спори феднг (енгл. *slow fading*) (односно дугорочни губици, засенчење) често настају када се неки топографски елементи, као што су дрвеће или високе зграде, налазе између предајника и пријемника, тако да дугорочне варијације сигнала узрокују да се импулсни одзив канала споро мења током трајања симбола.

Достизање потребног квалитета услуге (енгл. *quality of service* - QoS) и степена услуге (енгл. *grade of service* - GoS) у складу са захтевима спектралне ефикасности један је од основних циљева дизајнирања бежичних система. Једно од кључних питања у овом процесу је анализа својстава и утицаја било које врсте интерференције на бежичну комуникацију. Појмом интерференција истог канала (енгл. *co-channel interference* - CCI) се означава сигнал који има исту фреквенцију носиоца као и жељени информациони сигнал. Сви бежични

системи теже да одржавају доступан спектар поновним коришћењем додељених фреквенцијских канала у областима које су географски блиске једна другој. Услед честе поновне употребе фреквенције долази до интерференције сигнала који стижу са различитих канала и из различитих извора, али који раде на истој фреквенцији.

3.1.1. Multipath фединг

Фединг настаје под утицајем различитих фактора, као што су време, путања, радио-фреквенција и позиције пријемника и предајника. Када се говори о фиксној ситуацији (где се пријемник и предајник не крећу), у том случају до појаве фединга доводе и атмосферски услови, као што су киша или грмљавина. С друге стране, приликом кретања пријемника/предајника долази до наилажења на препреке (које се мењају током времена) под чијим утицајем такође долази до фединга. У бежичном комуникационом систему, *multipath* фединг се на основу различитих особина канала и позиције предајника и пријемника категорише у два типа, како је и приказано на Слика 1.



Слика 1. Типови фединга [28]

Као што је речено раније, приликом бежичног преноса долази до варијација фазе и анвелопе сигнала. Утицај фазних варијација се често може занемарити, нарочито када се у преносу користе некохерентни модулациони формати, што и јесте најчешћи случај у пракси. Међутим, посебна пажња се мора посветити временској варијацији анвелопе жељеног сигнала.

3.2. Модели феднига

Постоје бројни статистички модели који су развијени у сврху моделирању канала кроз које се преносе радио таласи између пријемника и предајника и проучавања фактора који на тај пренос утичу. Постоји више расподела вероватноће које се могу употребити у сврху описивања флуктуација анвелопе сигнала у бежичним каналима. У овој дисертацији су коришћене расподеле које се најчешће јављају код краткотрајног феднига у бежичним каналима: Гама, Рејлијева, Рајсова, Накагамијева и Вејбулова расподела [29].

3.2.1. Гама модел феднига

Гама расподела се користи у бежичним комуникацијама за моделовање снаге у каналима феднига [29]. Вероватноћа се одликује добром математичком трактабилношћу и добро фитује податке у широком опсегу пропагационих услова [26], [30].

Функција густине вероватноће у каналу са Гама федингом дата је помоћу [26]:

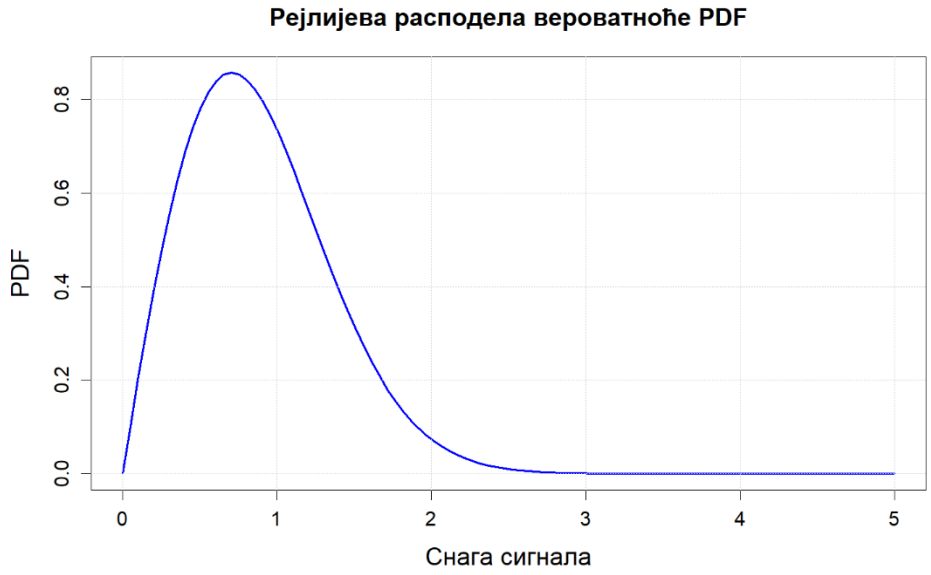
$$f_R(r) = \frac{r^{c-1}}{\Gamma(c)\Omega^c} e^{-\frac{r}{\Omega}}, r > 0, \qquad (3.1)$$

овде $\Omega = 2\sigma^2$ представља параметар скале, тј. средњу вредност снаге сигнала, а која је дефинисана још и као $\Omega = E(r^2)$, где E означава математичко очекивање статистичког процеса. Параметар c је параметар облика. Када

параметар $c = 1$ гама расподела постаје експоненцијална расподела вероватноће.

3.2.2. Рејлијев модел фединга

Рејлијев модел фединга је уједно и врло једноставан и прецизан, те је и један од најчешће коришћених модела за израчунавање утицаја окружења на простирање радио сигнала. Овај модел се користи за описивање простирања сигнала у урбаним срединама, али и у приградским подручјима, тамо где постоји велики број уређаја, а нема линија оптичке видљивости (енгл. *line of sight* — LOS).



Слика 2. Рејлијева расподела вероватноће

Када сигнал прође кроз комуникациони канал где не постоји линија видљивости комуникације између предајника и пријемника, тада долази до *multipath* простирања сигнала. Антена мобилне станице у тим случајевима не прима сигнал преко линије видљивости, већ се, услед *multipath* простирања, прима неколико рефлексија сигнала, дифракционих образаца и расејаних

(енгл. *scattered*) таласа. Тада фазе постају насумичне, а снага примљеног сигнала постаје случајна променљива, као што је приказано на Слика 2.

Прикладност коришћења Рејлијеве расподеле вероватноће је потврђена мерењима анвелопе примљеног сигнала [27] у областима где је линија видљивости блокирана различитим препрекама.

Рејлијев сигнал случајне амплитуде r може се формирати из фазних и квадратурних компоненти (x_1 и x_2). Ове компоненте имају средњу вредност једнаку нули, статистички су независне и имају нормалну расподелу, свака са варијансом σ^2. Израз:

$$r = \sqrt{x_1^2 + x_2^2}, \qquad (3.2)$$

представља Рејлијеву случајну променљиву. Њена функција густине вероватноће (енгл. *probability density function* - PDF) еквивалентне амплитуде r се даље представља као [26]:

$$f_R(r) = \frac{2r}{\Omega} exp\left(-\frac{r^2}{\Omega}\right), \qquad (3.3)$$

овде $\Omega = E(r^2) = 2\sigma^2$ представља параметар скале, односно средњу вредност снаге сигнала.

3.2.3. Рајсов модел фединга

У урбаним срединама није једноставно успоставити LOS између предајне станице и пријемника, те се услед *multipath* ефекта мобилни сигнал се састоји од неколико копија оригиналног сигнала. Додатно, чак и када постоји неки отворени простор, сигнал који се простире директно може бити ослабљен пре него што стигне до пријемника. Рајсов феднинг се користи да опише услове преноса сигнала у тзв. субурбаним срединама, где често постоји доминанатан сигнал са робусном линијом виљивости, уз вишеструке слабије *multipath*

компоненте [31]. У складу са тим, модел Рајсовог фединга се може добити из Рејлијевог модела фединга увођењем једне јаке директне LOS компоненте међу бројним случајно рефлектованим компонентама сигнала (које се обично сматрају слабијим). Наиме, Рајсов сигнал случајне амплитуде r може се формирати као:

$$r = \sqrt{(x_1 + A)^2 + x_2^2}, \qquad (3.4)$$

са функцијом густине вероватноће еквивалентне амплитуде израженом као [32]:

$$f_R(r) = \frac{r}{\sigma^2} exp\left(-\frac{r^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{rA}{\sigma^2}\right), r \geq 0, \qquad (3.5)$$

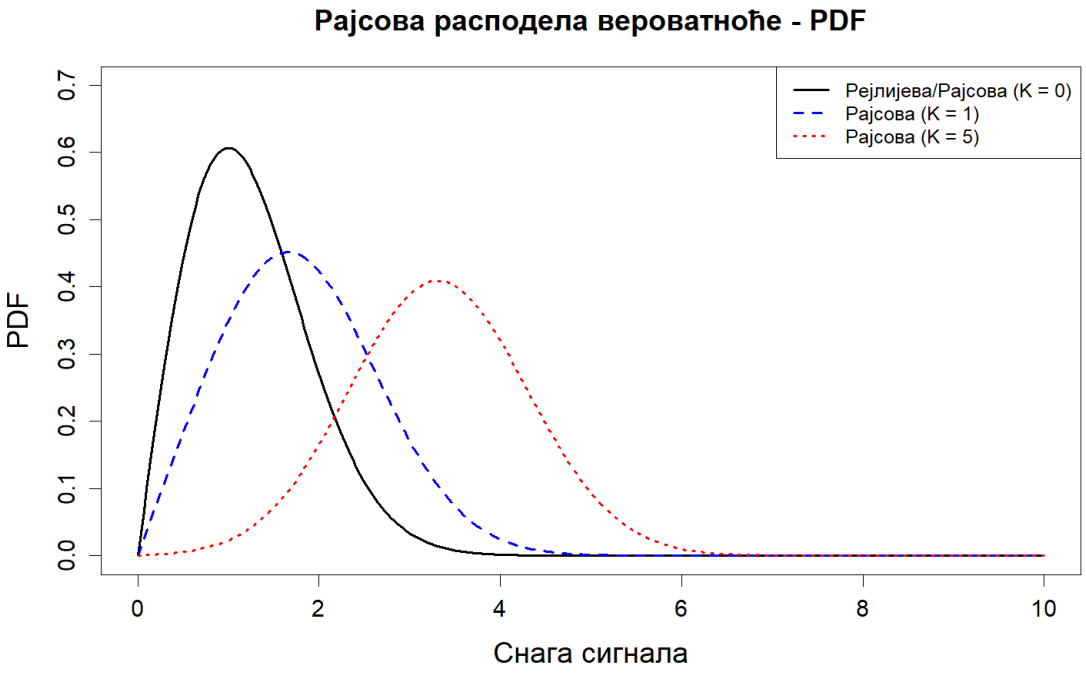
овде A представља средњу снагу сигнала компоненте, док $2\sigma^2$ представља средњу вредност снаге сигнала у расејаним компонентама. $I_o(x)$ представља модификовану Беселову функцију нултог реда прве врсте [6, Једн. 8.402].

Увођењем трансформације, $K = A^2/(2\sigma^2) = A^2/(\Omega)$, где је Ω средња снага анвелопе, Рајсов модел расподеле се може представити у другом облику — помоћу Рајсовог K фединг параметра. Овај параметар представља однос снаге сигнала у доминантној компоненти жељеног сигнала и просечне снаге фединга примљеног преко недоминантних путева:

$$f_R(r) = \frac{2(1 + K)r}{e^K \Omega} \; exp\left(-\frac{(1 + K)r^2}{\Omega}\right) \times \; I_0\left[2\sqrt{\frac{K(1 + K)r^2}{\Omega}}\right], r \geq 0. \qquad (3.6)$$

Када је K = 0, нема LOS или спекуларне компоненте, те модел постаје Рејлијев модел фединга. Када K расте, оштрина утицаја Рајсовог фединга опада, а перформансе система расту. Када Рајсов фактор тежи бесконачности, тада не постоји компонента расејања и фединга, као што је приказано на Слика 3. Скала вредности Рајсовог К фактора је линеарна [34].

Овај модел фединга се примењује за моделирање феднинг канала у фреквенцијском домену [35], у мобилним сателитским комуникацијама [27], [36].



Слика 3. Рајсова расподела вероватноће

3.2.4. Накагамијев модел фединга

Накагами-*m* расподела вероватноће укључује као специјалне случајеве Рејлијеву и Рајсову расподелу [26]. У раду [37] Минору Накагами је утврдио да Рејлијева расподела не може успешно предвидети понашање канала на великим удаљеностима и високим фреквенцијама. У истом раду је предложено, а у радовима различитих истраживача потврђено, описивање експерименталних података коришћењем параметарске функције густине засноване на Гама расподели. Осим тога, Накагамијева расподела је флексибилнија и обично боље фитује PDF амплитуде примљеног сигнала у урбаним срединама у односу на Рајсову и Рејлијеву расподелу [38]. Ова

расподела добро фитује податке и у спољашњим и у унутрашњим срединама [39], а показало се и да такође добро фитује сателитске сигнале [40].

Накагамијева расподела не претпоставља постојање линију видљивости између предајника и пријемника, као што је случај код Рајсове расподеле. Ова расподела користи параметарску функцију густине засноване на Гама расподели за карактеризацију експерименталних података и апроксимације расподеле [36].

Уколико је дат произвољан број компоненти у фази и у квадратури *multipath* феднига, које се моделују случајним променљивама x_{1i} и x_{2i}, i=1, …, m и које имају средњу вредност једнаку нули, следе Гаусову расподелу и имају једнаке варијансе. Тада квадратни корен суме ових променљивих:

$$r = \sqrt{\sum_{i=1}^{m}(x_{1i}^2 + x_{2i}^2)}, \qquad (3.7)$$

представља насумични процес са Накагами-m расподелом. Функција густине вероватноће Накагамијеве расподеле је [37]:

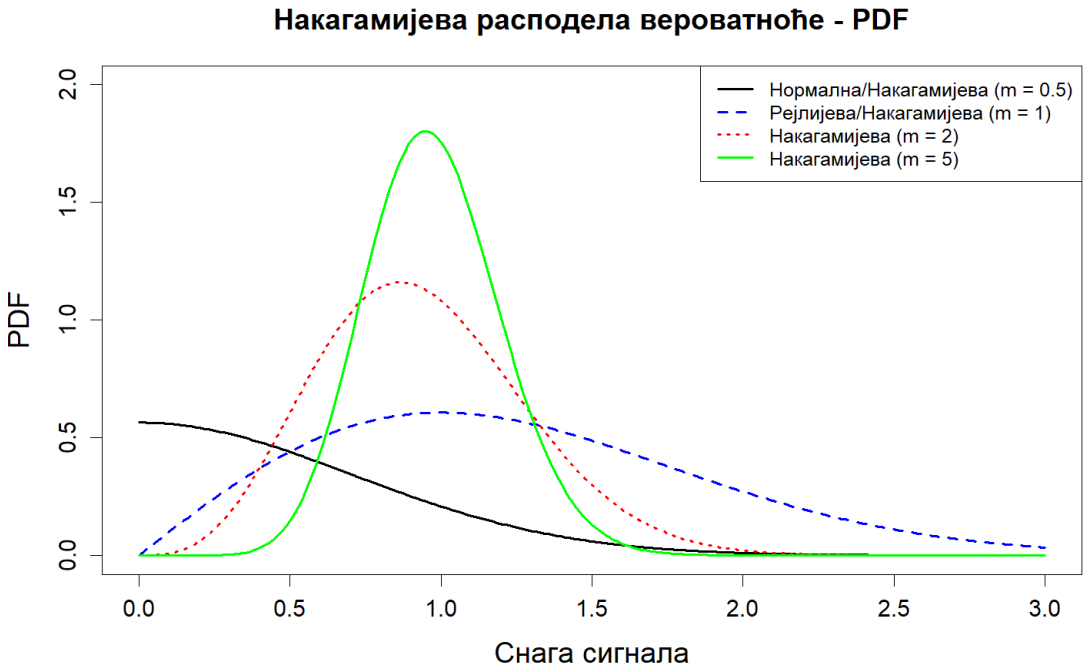
$$f_R(r) = \frac{2m^m r^{2m-1}}{\Gamma(m)\Omega^m} exp\left(-\frac{mr^2}{\Omega}\right), \quad r \geq 0, \qquad (3.8)$$

где $\Gamma(m)$ представља Гама функцију, параметар $\Omega = E(r^2)$ је средња снага сигнала, односно параметар скале, и мора бити већи од 0, а параметар m је Накагамијев параметар, односно параметар облика, и означава инверзну нормализовану варијансу r^2, он мора задовољити услов $\geq 1/2$, а од њега зависи оштрина феднига [7, стр. 46]. Параметар m показује степен феднига сигнала услед расејања и процеса *multipath* интерференције; када вредност параметра m расте, оштрина феднига опада. Када параметар m тежи бесконачности, Накагами канал постаје канал без феднига [41].

Када је параметар m једнак 1, Накагамијева расподела постаје Рејлијева расподела. Када је $m = \frac{1}{2}$, Накагами расподела прелази у Нормалну расподелу. За $m > 1$, Накагамијева расподела се приближава Рајсовој расподели, у том случају важи следећа зависност између параметра m и Рајсовог K фактора:

$$m = \frac{(K+1)^2}{2K+1}, m > 1. \qquad (3.9)$$

Због претходно наведеног, Рејлијева и Рајсова расподела се могу посматрати као специјални случајеви Накагамијеве расподеле, као што је показано на Слика 4.



Слика 4. Накагамијева расподела вероватноће

3.2.5. Вејбулов модел фединга

Да би се обухватили још неки постојећи феномени преноса, треба обрадити нелинеарна својства окружења, али ово је још увек отворен проблем у оквиру научних истраживања [42]. Вејбулов модел фединга произлази из Рејлијевог модела фединга као његов општији случај, а предложен је услед наведене нелинеарности окружења кроз које сигнал пролази.

Вејбулов модел фединга се примењује код сигнала сачињених од више *multipath* компоненти у оквиру нехогомене средине, при чему је анвелопа сигнала нелинеарна функција суме тих компоненти. Вејбулов фединг се добија као нелинеарна функција збира две случајне промењљиве са средњом вредношћу једнаком нули и нормалном расподелом насумичних промењљивих у фази и квадратури x_1 и x_2 и са једнаким варијансама:

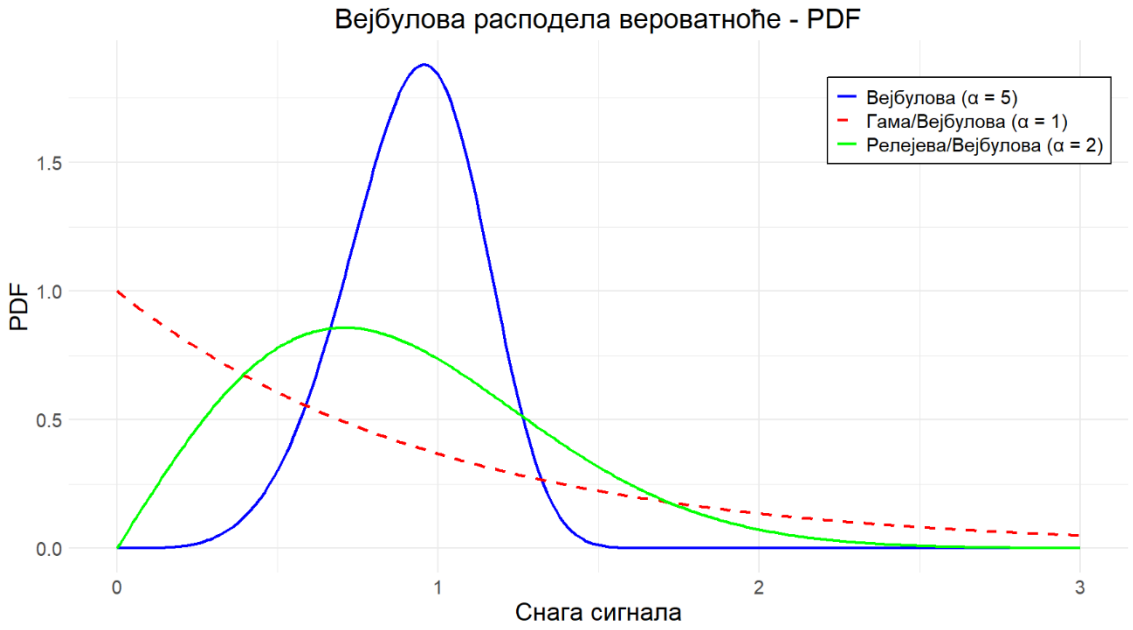
$$r = \sqrt[\alpha]{x_1^2 + x_2^2}, \qquad (3.10)$$

и са PDF израженом као [43]:

$$f_R(r) = \frac{\alpha r^{\alpha-1}}{\Omega} exp\left(-\frac{r^\alpha}{\Omega}\right), r \geq 0. \qquad (3.11)$$

У изразу (3.11) α означава параметар нелинеарности околине, док $\Omega = E(r^\alpha)$ представља просечну снагу сигнала. Параметар α такође описује и степен фединга (при чему је $\alpha \geq 0$), будући да се у областима с вишим вредностима параметра α степен фединга смањује [44]. Као што се може видети из (3.11), када је параметар $\alpha = 2$, Вејбулова расподела се своди на Рејлијеву. У случају када је α = 1, Вејбулова расподела прелази у експоненцијалну расподелу, као што је показано на Слика 5. Када α опада, расте оштрина Вејбуловог фединга, када α тежи бесконачности, Вејбулов модел прелази у модел без фединга.

Модел Вејбуловог фединга је нашао је примену у моделирању различитих сценарија код којих је потребна обрада у реалном времену. Вејбулов модел фединга се може користити и за моделирање фединга у спољашњем простору у градским срединама [44], али и унутрашњим срединама [45].



Слика 5. Вејбулова расподела вероватноће

3.2.6. Други модели фединга

У овој докторској дисертацији посматрани су одређени аспекти пропагације сигнала у бежичним комуникационим системима, међутим, постоји још модели краткорочног фединга који нису анализирани у оквиру истраживања представљеног у овом раду. У наставку је дат кратак опис тих модела:

- Хојтов модел федина - произилази из Рејлијевог модела фединга, као његов општији случај (такође је познат и као Накагами-q модел). Користи се за моделирање процеса који се обично јављају на сателитским везама које су подложне јакој јоносферској скинтилацији [35]. Хојтов модел има за граничне услове једнострани Гаусов фединг када је $q = 0$, и Рејлијев фединг када је $q = 1$ (односно, када су вредности варијанси једнаке).
- α-μ модел фединга такође укључује пропагацију кроз кластере *multipath* таласа, а има нелинеарне аспекте пропагације . Овај модел фединга подразумева сигнал са следећим својствима: 1) случајне фазе расејаних таласа унутар једног кластера имају слично временско

кашњење, док су временска кашњења различитих кластера релативно велика; 2) расејани таласи имају идентичну снагу унутар кластера *multipath* таласа [46]. Овај модел укључује Вејбулов и Накагами-m модел као специјалне случајеве.

- к-μ модел фединга укључује пропагацију кроз кластере вишеструких таласа, а изван нелинеарних аспекта пропагације [47]. Модел фединга са к-μ расподелом вероватноће је прилагођен случајевима примене са линијом директне видљивости будући да се претпоставља да свака група *multipath* таласа има доминанту компоненту. Расподела к-μ је општи физички модел фединга који може бити сведен на Рајсов и Накагами-*m* модел фединга (а самим тиме и на Рејлијев модел) у одређеним специјалним случајевима.
- η-μ модел фединга укључује пропагацију кроз кластере вишеструких таласа без нелинеарних и LOS аспеката. Дати модел укључује друге моделе фединга као специјалне случајеве, и то: Хојтов, Рејлијев и Накагами-*m* модел .
- Као што се може видети из претходно написаног, за неке моделе фединга се претпоставља резултујуће хомегено дифузно поље расејања које настаје од насумично распоређених извора расејања. Међутим, површи често имају међусобну просторну корелацију и карактеришу нелинеарну околину. Истражујући чињеницу да резултујућа анвелопа може бити нелинеарна функција збира *multipath* компоненти. α-η-μ и α-к-μ модели фединга представљају напредне верзије са три параметра: α (нелинеарност околине), μ (број *multipath* кластера у окружењу) и η (односом снаге расејаног таласа између компоненти), односно к (однос између доминантне компоненте у фази и доминантне компоненте у квадратури) [48]. Ови су општи модели и могу се, у зависности од вредности параметара, сводити на једноставније модела расподела. Модел α-η-μ укључује као специјалне случајеве и друге расподеле кратког фединга, као што су Рејлијев, Хојтов, Накагами-*m*, η-μ и

Вајбулов модел. Такође, α-κ-μ модел фединга укључује као специјалне случајеве κ-μ, Накагами-*m*, Рајсов, Вејбулов и Рејлијев модел расподеле.

3.3. Диверзити технологија и типови комбинера

Као што је објашњено раније, услед слабљења пријемног сигнала приликом преноса преко канала са федингом, потребно је да се на пријемној страни комбинују пристигле компоненте. Додатно, услед засенчења, сигнал губи снагу, чиме се додатно нарушавају перформансе комуникационих система. Диверзити технологија је неопходан део многих модерних система за бежичну комуникацију јер пружа ефикасан начин за превазилажење ових сметњи. Овај концепт је средином двадесетог века настао као начин умањивања ефеката *multipath* фединга, а и даље представља један од најбољих начина за побољшање радио веза код многих актуелних бежичних технологија [27]. Диверзити пријем се заснива на комбиновању две или више копија (путања) истог сигнала у циљу повећања укупног односа сигнал-шум (енгл. *signal-to-noise ratio* (SNR)).

Диверзити технике су кључне у смањењу утицаја фединга како би се добиле што боље перформансе система. Приликом пријема сигнала веома је битно утврдити о којој врсти фединга се ради. На основу типа фединга диверзити систем треба да примени даљи алгоритам избора ''најбољег'' сигнала са једне од више антена и даље обраде. У овој докторској дисертацији фокус је на процени типа фединга на диверзити антенама. Мотивација за истраживање је развити што ефикаснију процену типа фединга на диверзити пријему, тј. што ефикаснију предобраду сигнала како би у даљем процесирању добили што квалитетнији сигнал. Процена типа фединга је кључна у даљој обради примљеног сигнала, посебно имајући у виду да постоји велики број модела фединга у бежичним комуникационим системима. Ефикасна процена типа фединга убрзава даљу обраду примљених сигнала и тиме се повећавају перформасе комуникационих система.

3.3.1. Концепт диверзитија

Комбиновање диверзитија заправо подразумева редундантни пријем једног информацијског сигнала преко два или више феднинг канала. Овако примљене реплике сигнала се потом на пријемнику комбинују са циљем повећавања укупног примљеног SNR. Ово је оправдано тиме што постоји мања вероватноћа да на свим диверзити каналима дође до истовременог јављања феднинга, а самим тим је и мања вероватноћа јављања грешака и прекида сигнала.

Ове вишеструке реплике сигнала се издвајају на основу различитих радио путева. Неке од најпознатијих диверзити техника којима се постиже поменуто издвајање реплика сигнала

- Просторни диверзити — коришћењем више пријемних антена (ово је антенски или локацијски диверзитет) тако да сигнали буду некорелисани.
- Диверзити фреквенције — коришћењем неколико „размакнутих“ фреквенцијских канала за пренос сигнала. Код овог типа диверзитија се користи чињеница да сигнали различитих фреквенција другачије подлежу утицају пропагационе средине.
- Временски — код овог типа диверзитија се поништава утицај краткотрајног феднинга коришћењем чињенице да феднинг не утиче подједнако на сигнале са различитим временским оквирима. Овде се користи чињеница да сигнали путују различитим путањама, те самим тиме имају различита кашњења.
- Поларизациони —радио таласи могу бити поларизовани вертикално или хоризонтално, а рефлексија таласа од препрека на путу може променити њихову поларизацију. Коришћењем антена са различитим поларизацијама, могуће је добити два одвојена канала.

Сви наведени системи на улазу обједињују више сигнала са циљем да се на излазу добије што исправнији и снажнији сигнал. Обично се ово обједињавање

сигнала врши на два начина, први је мерењем улазних сигнала и одабиром најснажнијег, други је сабирањем сигнала. Код првог приступа треба водити рачуна о томе да сигнали имају различита кашњења, те се не може тренутно прећи са једног сигнала на други. Код сабирања сигнала је неопходно избећи интерференцију усаглашавањем времена (тј. баферовањем) приспећа сигнала.

Када је у питању просторно издвајање реплика сигнала, диверзити технике се могу поделити на две врсте:

1. микродиверзити,
2. макродиверзити.

Макродиверзити се користи за спречавање дуготрајног фединга (енгл. *large scale fading* - LSF), а примењује се и на страни пријемника и на страни предајника. Код ове технике се врши прикупљање сигнала помоћу антена које су међусобно просторно удаљене. Дуготрајни фединг настаје услед физички великих препрека услед чега сигнал слаби на дужи временски период. Могуће је да већина компоненти сигнала буде ослабљена и да се исправни сигнал може добити искључиво постављањем пријемних антена на удаљене локације, где се врши сабирање и одабирање компоненти сигнала. Ова техника диверзитија се ретко среће у пракси услед тешкоће имплементације и високе цене.

Микродиверзити техника се користи за спречавање краткотрајног фединга (енгл. small scale fading - SSF), и ради тако што се прикупљање сигнала врши врши помоћу једне или више антена смештених на једној локацији. Овај тип фединга карактеришу краткотрајни интервали високог нивоа деградације сигнала, а статистички гледано је мало вероватно да више сигнала буду истовремено драстично ослабљени.

Перформансе система диверзити техника исказују се кроз статистичке карактеристике првог реда, које обухватају: густину расподеле вероватноће сигнала, вероватноћу отказа система, средњу вероватноћу грешке по биту (ABER - Average Bit Error Rate), кумулативну расподелу вероватноће сигнала,

као и карактеристичну функцију и моменте сигнала (средњу вредност, средњи квадрат, средњи кубик и варијансу). Моменти омогућавају процену природе и нивоа фединга, који се дефинише као однос варијансе сигнала на излазу и квадрата средње вредности. За одређивање перформанси другог реда, неопходно је утврдити здружену расподелу вероватноће сигнала и његовог првог извода на излазу из диверзити пријемника. Статистичке карактеристике другог реда укључују средњи број преласка кроз нулу и средње трајање отказа система.

3.4. Процена расподела вероватноће у каналима са федингом - преглед досадашњих истраживања

За процену параметара фединга је коришћено много различитих метода.

Истраживачи су предлагали коришћење дубоког учења, тачније конволуционих неуронских мрежа (енг. *convolutional neural network* (CNN)) за процену Рајсовог и Накагамијевог фединга [49] у сигналима таласног облика [50] као и за препознавање фединг канала са Рејлијевим федингом [51].

У раду [52] је коришћена статистика тренутне фреквенције (енг. *instantaneous frequency* (IF)) сигнала примљеног на базној станици, у раду [53] аутори користе естиматор Рајсовог фактора који је заснован на анвелопи немодулисаног примљеног сигнала, аутори користе израз К фактора [54] да процене његову вредност.

Неки аутори су користили естиматор са оптималним диверзити пријемником за процену Рајсовог [55], Накагамијевог [56] и Рејлијевог [57] фединга помоћу линеарне процене канала преко најмањих квадрата (енг. least-squares (LS)) и најмањом просечном квадратном грешком (енгл. *minimum mean square error* (MMSE)).

У радовима [58], [59], [60], [61] је вршена процена Накагами-m параметра коришћењем методе момената вишег реда. Аутори у [58] показују да коришћење асимптотске варијансе у естиматорима који су заснованим на генерализованој методи момената, при чему се добијају резултати који су блиски резултатима добијеним помоћу приступа заснованом на методи процене максималне вероватноће. У радовима [61] и [62] су коришћени естиматори засновани на методи момената, с тим што су у раду [61] аутори користили тзв. *лукап* табелу коју су претходно добили израчунавањем и смештањем вредности инверзне функције, а у раду [62] је описан проблем процене параметара фединга код *multipath u* спорог фединга са флуктуацијама анвелопе и шумом.

У раду [63] је коришћена метода процене максималне вероватноће, али је израчунавање параметара расподела вршено без рачунања момената вишег реда. У раду [64] су за процењивање ових параметара коришћене и метода момената и процена максималне вероватноће.

Развој дубоког учења је довео до његове примене у различитим комуникационим системима. Класификација модулација сигнала у различитим окружењима и на различитим каналима применом дубоког учења је покушана у следећим радовима. У раду [65] аутори су искористили методу дубоког учења засновану на дводимензионалној спектралној корелационој функцији (енгл. *2D-spectral correlation function*) која за класификацију модулације користи комбинацију техника које се иначе користе за препознавања слика. У раду [66] је извршена класификација пет модулационих формата фединга коришћењем дубоке неуронске мреже. Модулациони формати у овом раду су генерисани са адитивним белим гаусовски шумом. У раду [67] аутори предлажу аутоматску методу класификације модулације засновану на предобради таласића уклањањем шума коришћењем унапређеног дизајна конволуционе неуронске мреже (енгл. *convolutional neural network* (CNN)) датог у раду [68]. У раду [67] аутори су извршили експерименте са применом класификатора заснованих на CNN над различитим фазним померајима, Рејлијевим федингом и Доплеровим

ефектом. Осим наведеног, у радовима [69], [70] је примењено дубоко учење и за идентификацију бежичних система у глобалном систем мобилних комуникација (енгл. *Global System for Mobile Communications* (GSM)), универзалног мобилног телефонског система (енгл. *Universal Mobile Telecommunications Service* (UMTS)), и дугорочна еволуција (енгл. *Long-Term Evolution* (LTE)). У радовима [50], [71] аутори користе конволуциону неуронску мрежу за процену параметара фединга директно из модулисаних сигнала; при чему су у раду [50] дати резултате са различитим вредностима SNR и величинама скупова података.

4. ГЕНЕРИСАЊЕ И ОБРАДА СИГНАЛА

Кључни допринос овог рада јесте да се применом метода истраживања података покаже да је могуће успешно утврдити расподелу и параметре анализираног сигнала у реалним радним условима. На примеру пријемника базне станице мобилне телефоније, након идентификације расподела сигнала примљених на вишеструким антенама, логика пријемника треба да одабере сигнал најбољег квалитета. То се може учинити тако што ће проценити и упоредити расподеле и параметре расподела добијених сигнала са сваке антене појединачно, а потом донети одлуку који од тих сигнала омогућава оптимално пружање услуге.

С обзиром на то да приступ подацима стварних сигнала на улазу пријемника базне станице није могућ, у оквиру ове дисертације коришћени су насумично генерисани узорци података сигнала у складу са најчешћим дистрибуцијама омотача сигнала и вредностима њихових параметара пронађеним у литератури [27], [34].

4.1. Генерисање псеудо-случајних сигнала федингa

Генерисање псеудослучајних сигнала је тема од великог значаја у савременом свету науке и технологије, посебно у контексту криптографије, статистике, симулација и многих других области.

Псеудослучајни сигнали се карактеришу као излазни сигнали из система који наизглед производе насумичне резултате, али у ствари, генеришу сигнале који су детерминисани унапред задатим алгоритмом или процесом [72]. Иако ови сигнали могу изгледати насумични на први поглед, они нису истински случајни, јер се могу репродуковати ако се почне са истим почетним условима, попут семена генератора.

У програмском језику R, генерисање псеудо-случајних сигнала је олакшано широким скупом функција написаних у ту сврху. Поред уграђених функција, језик R пружа широк скуп различитих пакета за генерисање псеудо-случајних сигнала комплекснијих расподела вероватноћа. Ови пакети пружају додатне функције и алате за генерисање и манипулисање псеудо-случајним сигналима, чиме се истраживачима омогућује да их прилагоде својим потребама.

Да би се могла измерити ефикасност алгоритма за препознавање модела расподела вероватноће, неопходно је извршити симулацију пријема одмерака сигнала на улазима пријемника. У овом циљу је генерисан скуп тестних података који је довољно велики да се исправно процени прецизност препознавања. Генерисање насумичних вредности узорака за изабране расподеле вероватноће и њихове задате вредности параметара је извршено коришћењем језика R и одговарајућих пакета.

Вредности параметара расподела вероватноће генерисаних сигнала су одабране тако да одговарају случајевима који се могу срести у стварности, у складу са опсегом референтних вредности датим у [26]. У складу са наведеном литературом, одабрани су одговарајући опсези вредности параметара пет расподела вероватноће које су коришћене у овој докторској дисертацији. Код свих пет расподела је вредност параметра Ω варирана од 1 до 2,5 у корацима од 0,5. Резултати објављени у [21] су показали да резултати препознавања не зависе од вредности параметра скале. Поред овога, посматране су следеће вредности параметара који се користе у четири расподеле вероватноће:

- Гама расподела: пет вредности параметра c, од 1 до 3 у корацима од 0,5;
- Накагами-м расподела, седам вредности параметра m, почевши од 0,5, и од 1 до 5 у корацима од 1, и 5,5;
- Рајсова расподела: шест вредности параметра K, од 0 до 5, у корацима од 1;
- Вејбулова расподела: шест вредности параметра c, од 0,5 до 3 у корацима од 0,5.

Поред фединга, који представља најзначајнију врсту сметње у бежичним комуникационим системима, често се јављају и друге врсте сметњи. У реалним условима, сигнал је често контаминиран нежељеним шумом. До шума долази из различитих разлога, било због термичких карактеристика, шума који уводе електронске компоненте пријемника и предајника, и слично. Адитивни бели Гаусовог шум се често користи за симулирање позадинског шума у каналу [73]. Математички израз примљеног зашумљеног сигнала је:

$$r(t) = s(t) + n(t) \qquad (4.1)$$

где је s(t) пренети сигнал, а n(t) је шум.

С обзиром на наведено, поред анализирања самог сигнала генерисаног у складу са посматраним расподелама, увели смо и два додатна нивоа додатог шума, ниво сигнал-шум односа (SNR) од 25dB и SNR од 20dB. Треба напоменути да је тестирање извршено без увођења додатних технике обраде зашумљених сигнала, већ је алгоритам примењен директно над зашумљеним сигналима. За веће нивое шума, односно ниже SNR вредности, било би неопходно имплементирати напредније технике за филтрирање шума и издвајање сигнала. Основни циљ је био да се анализира робусност тестираних метода у присуству различитих нивоа шума и процени њихова применљивост у условима када је однос сигнала и шума деградиран.

4.2. Предобрада сигнала

Предобрада сигнала је важан корак у анализи псеудо-случајних сигнала јер је корак који претходи препознавању расподела и њихових параметара. Технике предобраде имају за циљ да побољшају квалитет сигнала, да уклоне шумове, и нормализују податке. Ово је неопходан корак за добијање квалитетних података који се могу даље обрадити и анализирати.

Једна од карактеристика краткотрајног сигнала која се може релативно брзо израчунати на основу вредности његових одмерака је нормализована дискретна кумулативна расподела сигнала (НДКР) са одређеним бројем опсега вредности (енгл. *bin*) за предобраду сигнала. НДКР представља непараметарско процењивање кумулативне функције расподеле вероватноће, која обезбеђује компактну репрезентацију дистрибуције сигнала. Дискретизација НДКР омогућава смањивање димензионалности података и олакшава ефикаснију анализу.

Нормализација је још један од суштинских корака предобраде. Циљ нормализације јесте да се стандардизује скала сигнала, чиме се осигурава да су вредности у оквиру уједначеног опсега. Ово је нарочито важно када се пореде сигнали различитих величина или јединица мере, јер у супротном не би било могућности за смислено поређење.

Процес израчунавања НДКР почиње одређивањем опсега могућих вредности (од минималне до максималне) које одмерци тог сигнала имају, а тај опсег се потом подели у K једнаких интервала (под-опсега, $k = 1, \dots, K$). Након тога се за сваки интервал одреди број одмерака Yk који припадају сигналу и чија је вредност мања или једнака горњој граници Xk тог под-опсега. На крају процеса, НДКР се нормализује дељењем са укупним бројем узорака, чиме се добија нормализовани НДКР и тада се карактеристика $Y(X)$ креће у границама $(0, 1)$.

Сами бинови и њима одговарајуће вредности представљају вредности нормализованог НДКР за x и y осе, респективно. На крају овог процеса су добијене координате тачака које ће послужити за проверу подударања модела расподела вероватноћа, тзв. „фитовање“ (енгл. *fitting*). Овај корак предобраде упрошћава представљање података и омогућава ефикасније фитовање кривих и процену параметара.

Зарад бољег илустровања овог израчунавања НДКР из вредности узорака улазног сигнала, генерисали смо двадесет одмерака сигнала који прате расподелу вероватноће одабрану у генератору случајних сигнала (у овом

случају — Гама расподела са параметрима $\Omega = 1$ и $c = 2$) , који су приказани на Слика 6.

Signal values: 0.227531, 1.80523, 0.165631, 1.24947, 5.08724, 3.10942, 1.14969, 1.28665, 3.66837, 9.47541, 3.14633, 2.65018, 1.58915, 3.29589, 0.33547, 1.06296, 1.70372, 1.23008, 0.766655, 1.8028

Bin #	Bin MaxV	dCDF	normalized dCDF
1	0	0	0
2	1.15	6	0.3
3	2.3	13	0.65
4	3.45	17	0.85
5	4.6	18	0.9
6	5.75	19	0.95
7	6.9	19	0.95
8	8.05	19	0.95
9	9.2	19	0.95
10	10.35	20	1

x = (0, 1.15, 2.3, 3.45, 4.6, 5.75, 6.9, 8.05, 9.2, 10.35)

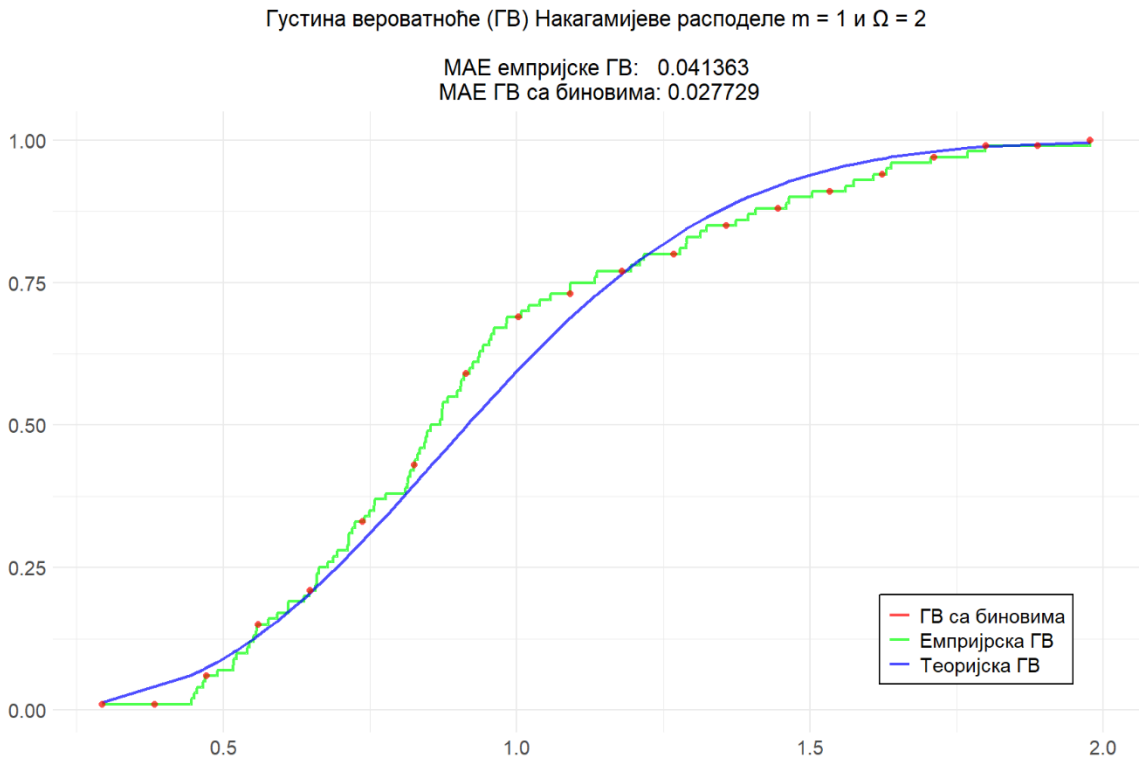
y = (0, 0.3, 0.65, 0.85, 0.9, 0.95, 0.95, 0.95, 0.95, 1)

Слика 6. Пример израчунавања НДКР одмерака сигнала [21]

Као што се на Слици 6. може видети, генерисане вредности узорка у овом примеру су у опсегу од 0,165 до 9,475. Овај опсег вредности се затим дели на неколико под-опсега (бинова) (колона *Bin MaxV* на поменутој слици). НДКР вредност сваког под-опсега се рачуна као број вредности узорка који су мањи или једнаки његовој максималној вредности (тј. број вредности узорка који су мањи или једнаки максималној вредности четвртог под-опсега у овом примеру је 17, као што је приказано у колони *dCDF* (тј. НДКР)). На крају процеса, НДКР се нормализује (дели) укупним бројем узорака (колона *normalized dCDF* (тј. НДКР)). Избором вредности из колона *Bin MaxV* и НДКР за x и y осе, добијају се координате тачака података за фитовање криве у наредним корацима примене НР.

4.3. Процес процене вредности параметара расподеле вероватноће

Процес процене вредности параметара расподеле вероватноће се заснива на проналажењу модела који најбоље одговара тачкама података сигнала. С обзиром на то да регресија проналази најбоље вредности параметара минимизирањем грешке која постоји између стварних вредности података у односу на израчунате вредности из математичких једначина (референтне вредности), ове параметре не рачунамо директно из емпиријске густине расподеле сигнала, већ помоћу НДКР сигнала [74].



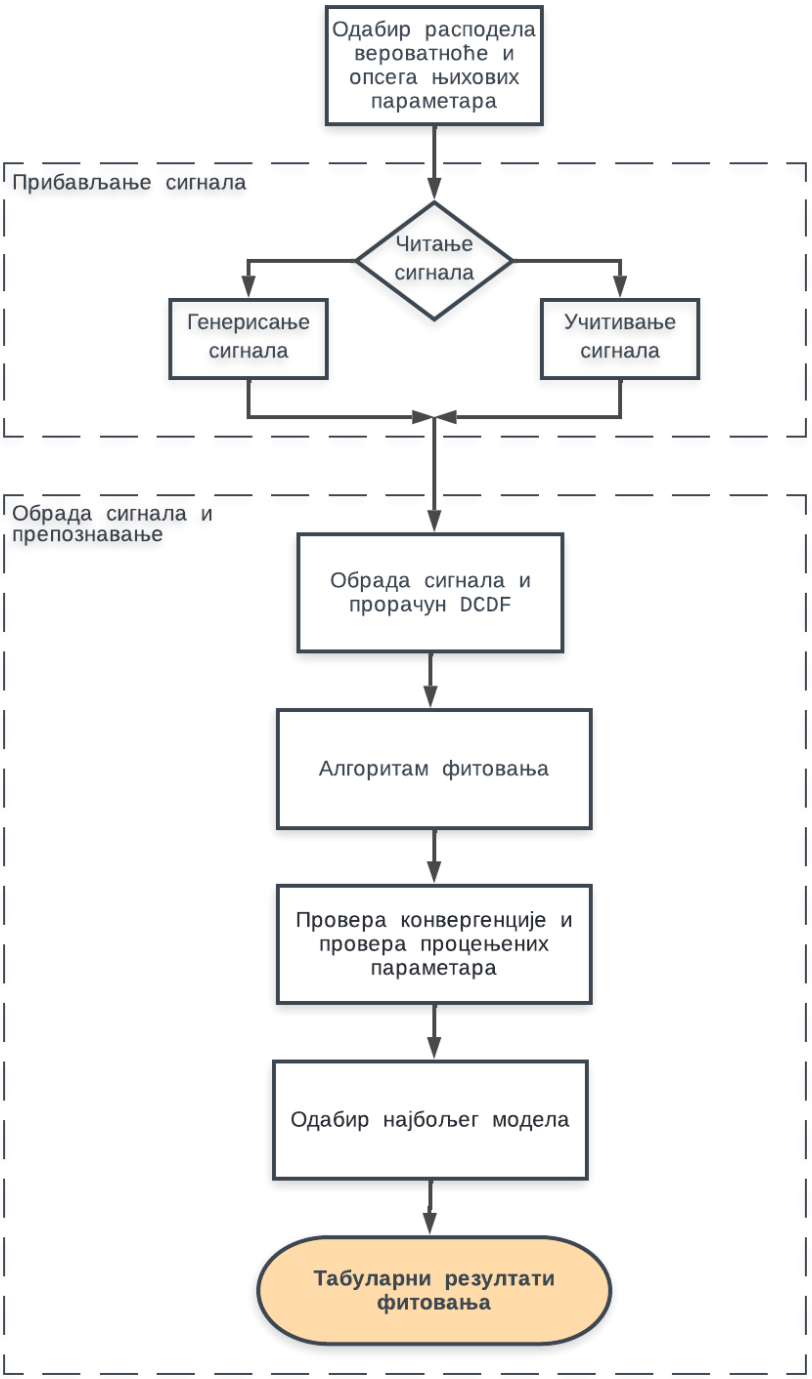
Слика 7. Поређење грешака емпиријске густине расподеле и густине расподеле по биновима

Како би се графички представила логика која стоји иза коришћења овог приступа, на Слика 7. Поређење грешака емпиријске густине расподеле и

густине расподеле по биновима су представљене три густине расподеле за Накагамијеву расподелу (са параметрима $m = 1$ и $\Omega = 2$). Плавом бојом је представљена крива идеалне расподеле, зеленом бојом емпиријска густина расподеле[75] за генерисани насумични сигнал од 100 одмерака, а црвеном бојом је представљена вредност густине вероватноће за исти сигнал, али сада за задати број бинова (у овом случају 50). Са слике се може уочити да је MAE грешка за трећину мања код густине вероватноће изложене по биновима.

У телекомуникационим системима се посебна пажња посвећује томе да перформансе буду што је могуће ближе раду у реалном времену и мора се водити рачуна да се не унесу значајна кашњења. Имајући то у виду, НДКР примљеног сигнала треба израчунати из разумно малог броја тачака (Ns), али водећи рачуна о томе да краћа дужина сигнала не узрокује значајно смањење тачности мерења. У раду [21] је коришћена величина узорка $Ns = 1000$, јер се показало да је то довољно за поуздану процену параметара дистрибуције вероватноће [76]. Други изазов је одредити одговарајући број бинова за НДКР; недовољан бинова може учинити да се жртвује превише информација, а превелики бројеви бинова могу изобличити криву да би се прилагодили подацима. У раду [21] коришћена је вредност $Nb = 50$.

Како бисмо испитали зависност перформанси од дужине сигнала и утврдили оптимални број тачака НДКР у циљу брзог и тачног препознавање, генерисани су сигнали са три различите дужине: 100, 300 и 1000 одмерака. Да би се одредиле густине расподеле генерисаних сигнала, вредности одмерака у тим сигналима су подељене на униформне под-опсеге – бинове. Одабране су четири величине броја бинова, изражене као проценат броја одмерака посматраног сигнала: 5%, 10%, 25% и 100%. Сходно томе, за сигнал од 100 одмерака, број генерисаних под-опсега, односно бинова је 5, 10, 25 и 100 респективно за 5%, 10%, 25% и 100%, за сигнал од 300 одмерака, број бинова је 15, 30, 75 и 300, а за сигнал од 1000 одмерака 1000, број бинова је 50, 100, 250 и 1000.



Слика 8 Дијаграм алгоритма

На Слика 8 је представљен блок дијаграм коришћеног алгоритма. За препознавање расподела вероватноће сигнала и њихових параметара коришћене су методе нелинеарне регресије. Улазни подаци су x и y вредности

НДКР сигнала, а математички модели (тј. једначине) најчешћих кумулативних расподела вероватноће са њиховим параметрима (тј. Ω, α, c, K, m из раније наведених једначина) као променљиве. Као резултат фитовања, алгоритам даје на излазу успешно фитоване моделе и процењене вредности њихових параметара заједно са критеријумима информација за избор модела. Процес проналажења најбољег уклапања НДКР улазних података са раније дефинисаним моделима расподеле вероватноће састоји се од неколико корака: (1) постављање почетних вредности параметара које треба препознати и њихових могућих опсега, (2) прилагођавање модела, (3) провера конвергенције и ограничења вредности параметара и (4) избор најбољег модела. Псеудокод корака наведеног алгоритма је представљен на Слика 9.

```
Input: ulazni signal, modeli raspodela verovatnoće
// Faza pripreme:
    Pronađi maksimalnu vrednost odmerka
    veličina_bina = maks_odmerak / broj_binova
    for i = 1 to broj_binova do
        binovi[i] = i * veličina_bina
    end for
    for i = 1 to broj_odmeraka do
        for j = 1 to broj_binova do
            if odmerak[i] <= binovi[j] then
                DKR[j] += 1
            end if
        end for
    end for
    for i = 1 to broj_binova do
        NDKR[i]=NDKR[k]/broj_odmeraka
    end for
// Faza prepoznavanja:
    for i = 1 to broj_modela_raspodela do
        Fituj model model[i] sa NDKR
        if fitovanje konvergira then
            if parametri su u okviru opsega then
                Dodaj model rezultatima
            end if
        end if
    end for
    for model = 1 to broj_fitovanih_modela do
        Pronađi model sa najboljim metrikama
    end for
Output: najbolji model
```

Слика 9. Псеудокод примене НР у препознавању статистичких модела

Псеудокод са Слике 9. је у раду [21] имплементиран у облику R скрипте која имплементира кораке горе наведене методологије. За потребе ове дисертације, наведена скрипта је знатно проширена и додат јој је графички кориснички интерфејс.

Први корак у обради јесте одабир модела расподела вероватноће и граничних опсега њихових параметара. Развијени алат може генерисати скуп сигнала или их учитати из улазних датотека, након чега следи обрада сигнала и прорачун НДКР. Затим следи фитовање одабраних модела и НДКР, врши се провера да ли су параметри успешно фитованих модела унутар референтних вредности наведених у литератури [26]. Предност овог приступа је проточна обрада, односно док се на улазу врши прикупљање (до жељеног броја одмерака) и обрада улазног сигнала у реалном времену, може се истовремено вршити евалуација модела расподела вероватноће. Систем одбацује моделе који нису успешно фитовани и моделе чији су параметри ван очекиваних опсега, а затим одабира најбољи међу преосталим моделима. Одабира се модел са најбољим вредностима одабраних метрика (*AIC*, *BIC*, *RMSE*, *MAE*, *R squared*, *adjusted R squared*).

5. ПРИМЕНА НЕЛИНЕАРНЕ РЕГРЕСИЈЕ ЗА ПРЕПОЗНАВАЊЕ КАНАЛА ФЕДИНГА И АНАЛИЗА РЕЗУЛТАТА

Да би се добили непристрасни резултати у процени (препознавању) расподела вероватноће и вредности њихових параметара на основу скупа сигнала, за сваки од три нивоа шума у сигналу, и за сваку од 96 комбинација параметара посматраних расподела вероватноће, и за сваку од 12 комбинација бинова и одмерака, генерисано је по сто сигнала који су препознавани. Тако је у овом експерименту коришћено укупно 345.600 различитих сигнала. Над сваким од тих сигнала извршено је фитовање кривих и извршен одабир најбољег процењеног модела. У поглављима која следе, дати су резултати спроведене анализе.

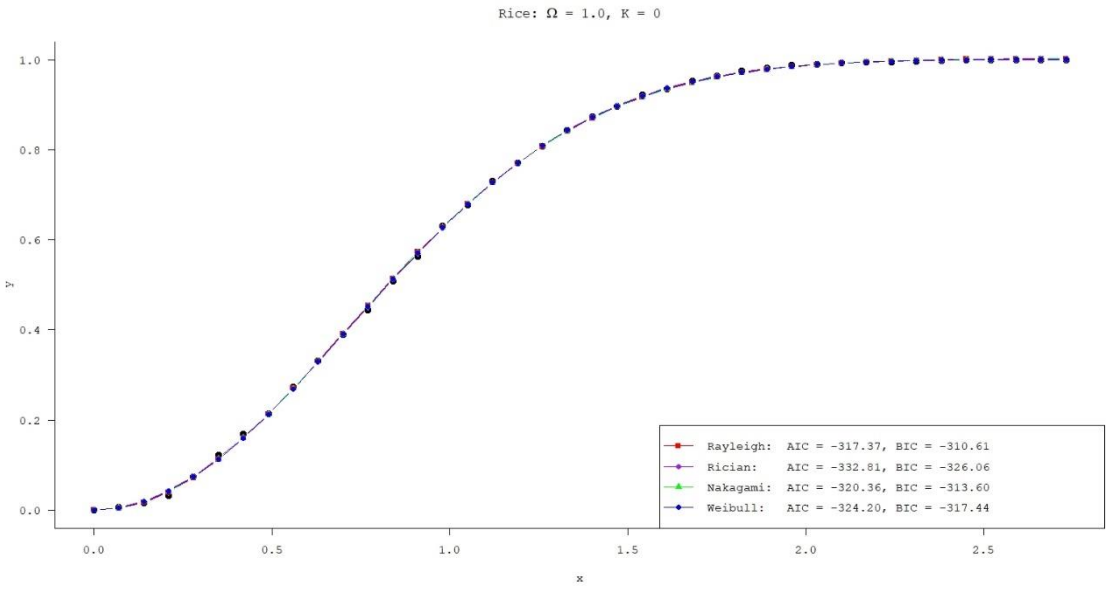
Изложени резултати представљају наставак истраживања датог у раду [21], допуњени понављањем експеримената препознавања расподела са различитим дужинама сигнала и израчунавањем НДКР сигнала са различитим бројем тачака, како би се установила оптимална дужина сигнала и број тачака НДКР за брзо и тачно препознавање.

5.1. Препознавање специјалних случајева расподела вероватноће

Фитовање кривих у статистичкој анализи постаје тешко када се бави специјалним случајевима где две или више кривих конвергирају у исти облик. Ово може сакрити неке суптилне разлике између расподела вероватноће, што отежава утврђивање модела који најбоље фитује податке. Када се облици ових кривих преклапају, традиционалне методе фитовања криве као што су MLE или најмањи квадрати могу дати двосмислене резултате, јер се ове технике у

великој мери ослањају на разликовање облика кривих како би идентификовале најбоље одговарајуће параметре. Ова инхерентна флексибилност која постоји међу расподелама је важан фактор у моделовању стохастичке природе феднинга сигнала.

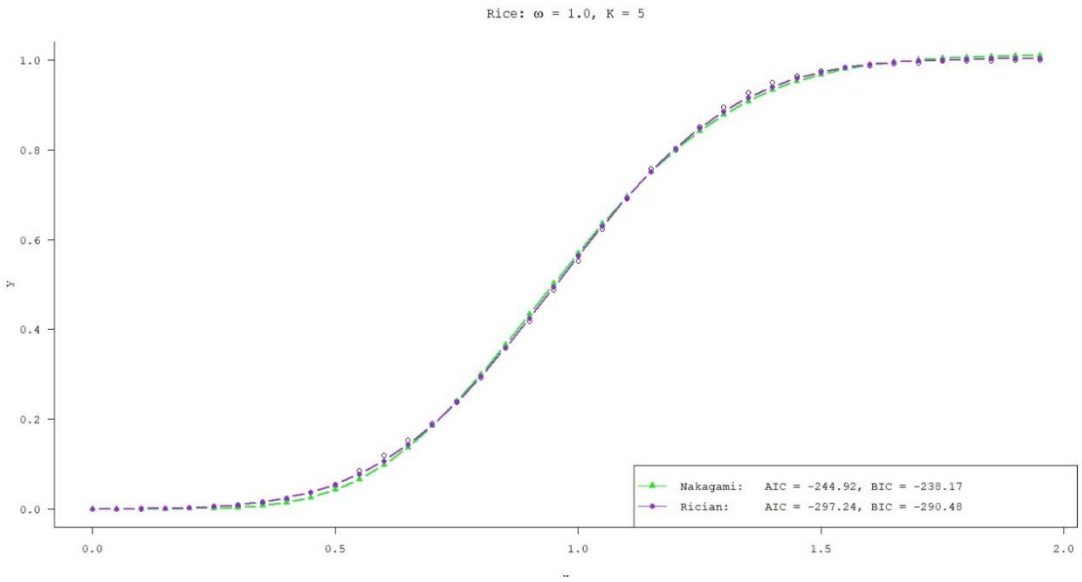
Када је реч о фитовању CDF и PDF расподела, када се вредности параметара приближавају специјалним случајевима, облици кривих почињу да личе (а у самим специјалним случајевима имамо потпуно преклапање) једни на друге. Приликом процеса фитовања кривих, посебна пажња се мора посветити разликовању суптилних нијанси у облицима кривих расподеле. Такве тешкоће су посебно изражене у практичним применама попут бежичних комуникација, где је прецизно моделирање понашања феднинга сигнала кључно за оптимизацију перформанси и поузданости система.



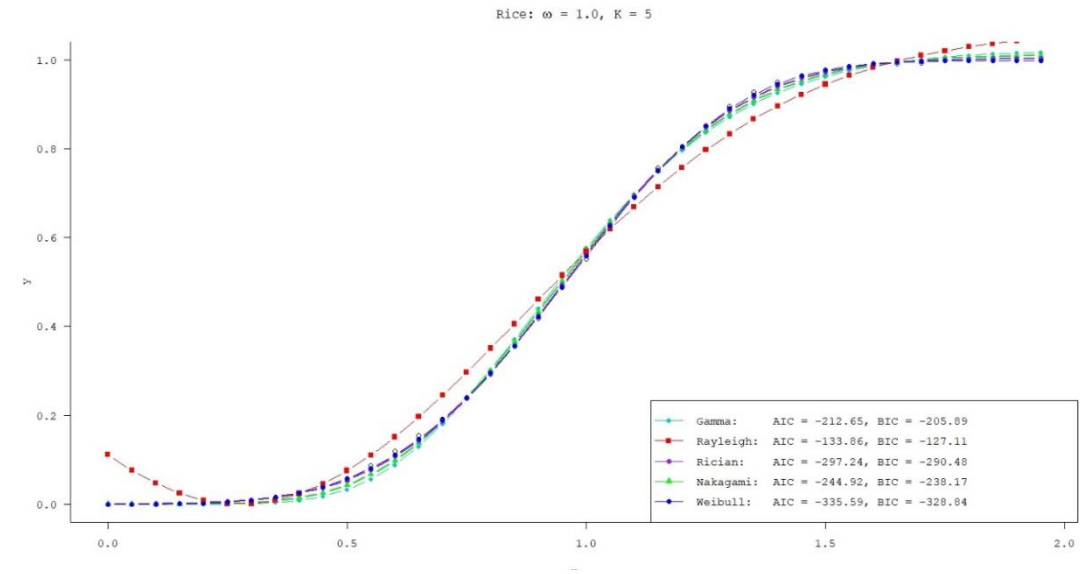
Слика 10. Фитовање криве за Рајсову расподелу за вредност параметра K=0

На Слика 10 је дат репрезентативан пример фитовања кривих расподела вероватноће за случај када имамо преклапање више модела. На слици је дато фитовање сигнала генерисаног по Рајсовој расподели са вредношћу параметра $K = 0$ и $\Omega = 1$, а то је специјални случај када се Рајсова расподела

своди на Рејлијеву расподелу, а самим тиме се преклапа и са Накагамијевом расподелом за вредност параметра $m = 1$, Вејбуловом расподелом за вредност параметра $\alpha = 2$. На слици се може се приметити да су вредности AIC и BIC параметара блиске, што је и очекивано будући да се заправо ради о једној истој расподели.



а) ограничено фитовање кривих



б) неограничено фитовање кривих

Слика 11. Ограничено и неограничено фитовање кривих

На Слика 11. Ограничено и неограничено фитовање кривихје приказан рад алгоритма код фитовања сигнала који следи Рајсову расподелу вероватноће са вредностима параметара $\Omega = 1$ и $К = 5$. С обзиром на то да ће за дати сигнал алгоритам утврдити да су само вредности параметара Рајсове и Накагамијеве расподеле вероватноће у оквиру очекиваних вредности — врши се фитовање ове две расподеле и алгоритам успешно утврђује да улазни сигнал одговара Рајсовој расподели (а). На слици (б) је дат и случај фитовања кривих када се не врши провера да ли су вредности израчунатих параметара у оквиру очекиваних вредности, у датом случају се за криву Вебјулове расподеле добија најбоље преклапање упркос томе што параметар $\alpha = 3{,}8$ излази ван референтних вредности [26].

Када је параметар облика код Гама (параметар с) и Вејбулове расподеле (параметар α) једнак јединици, тада се добија специјалан случај ове две расподеле у коме оне постају Експоненцијална расподела вероватноће [77], што значи да криве њихове CDF добијају исти облик. Услед наведеног, постоји подједнака вероватноћа да ли ће алгоритам „препознати“ сигнал као гама или као Вејбулову расподелу.

Табела 2 Резултати препознавања Гама расподеле са вредношћу параметра облика с = 1

бр. одмерака	**бр. бинова**	**Гама**	**Вејбулова**	**Накагамијева**
100	5	180	219	1
100	10	213	165	22
100	25	177	220	3
100	100	181	219	0
300	15	204	185	11
300	30	193	203	4
300	75	175	225	0
300	300	188	210	2
1000	50	189	208	3
1000	100	169	231	0
1000	250	181	215	4
1000	1000	176	221	3

у

Табела 2 су дати резултати препознавања расподеле сигнала који су генерисани да буду у складу са гама расподелом где је вредност параметра $c = 1$. Резултати су изложени по броју одмерака и броју бинова, при чему је у сваком реду укупан број експеримената једнак 400 (јер су груписани за четири различите вредности параметра Ω). Резултати препознавања су показали да су ти одмерци препознати да припадају гама расподели у 46,4%, Вејбулова расподела је препозната у 52,3% случајева, док је 1,1% случајева препозната Накагамијева расподела (при чему број фитовања Накагамијеве расподеле вероватноће опада са порастом броја одмерака сигнала).

На Слика 12. Поређење процењених вредности параметара облика модела гама и Вејбулове расподеле дат је програмски код којим се рачуна средња вредност процењених параметара облика код свих сигнала генерисаних у складу са гама расподелом са вредношћу параметра $c = 1$. Може се видети да фитовани модели одговарају специјалном случају који је описан изнад.

```
> # izdvajamo signale generisane sa parametrom c = 1
> tmp <- gamma_SNRfree %>% filter(gamma_SNRfree$param1==1)
> mean(as.numeric(tmp$weibull_shape), na.rm = TRUE)
[1] 1.002186
> mean(as.numeric(tmp$gamma_shape), na.rm = TRUE)
[1] 1.006544
> # SNR = 25dB
> # izdvajamo signale generisane sa parametrom c = 1
> tmp <- gamma_SNR25dB %>% filter(gamma_SNR25dB$param1==1)
> mean(as.numeric(tmp$weibull_shape), na.rm = TRUE)
[1] 0.9924241
> mean(as.numeric(tmp$gamma_shape), na.rm = TRUE)
[1] 1.008099
> # SNR = 20dB
> # izdvajamo signale generisane sa parametrom c = 1
> tmp <- gamma_SNR20dB %>% filter(gamma_SNR20dB$param1==1)
> mean(as.numeric(tmp$weibull_shape), na.rm = TRUE)
[1] 0.9939317
> mean(as.numeric(tmp$gamma_shape), na.rm = TRUE)
[1] 1.009051
```

Слика 12. Поређење процењених вредности параметара облика модела гама и Вејбулове расподеле

Резултати аналогни наведеним, добијени су и у осталим специјалним случајевима разматраним у поглављу 3.2, и у складу су и са резултатима које смо претходно публиковали у [21]. Стога, с обзиром на конзистентност резултата са изложеном теоријском поставком, у наредним поглављима нећемо понављати детаљно елаборирање ових случајева.

5.2. Перформансе критеријума евалуације модела за препознавање сигнала

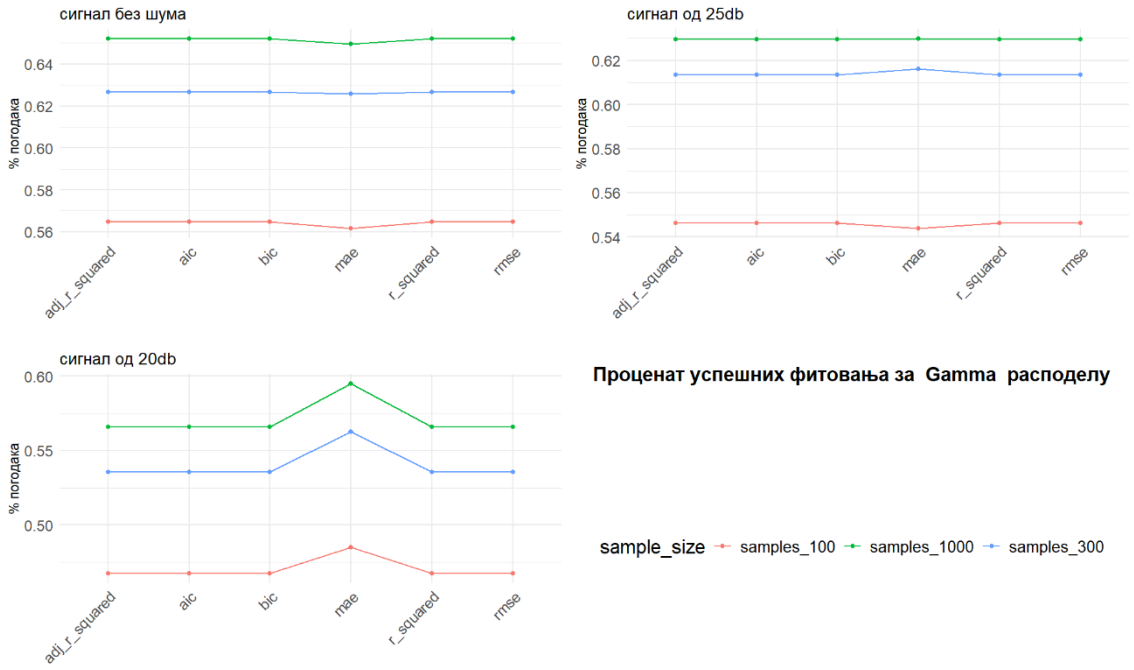
Статистичко моделовање улазних параметара је основни аспект препознавања дистрибуција јер омогућава квантификовање веза између посматраних података и посматраних модела. Различити тестови адекватности (енгл. *goodness-of-fit*) и методе избора модела се могу користити

за процену квалитета фитованих модела и одређивање најпогоднije расподеле вероватноће за дате податке.

У циљу утврђивања најадекватније метрике за процену успешности препознавања сигнала одабраних расподела вероватноће, спроведена је свеобухватна упоредна анализа фитовања коришћењем неколико статистичких метрика. Конкретно, упоређени су резултати добијени фитовањем сигнала применом критеријума AIC, BIC, MAE, RMSE, R^2 и прилагођеног R^2. Поређење броја успешних препознавања наведених метрика је извршено за три различите дужине улазног сигнала, са 100, 300 и 1000 одмерака. Резултати су представљени у виду графикона на којима су на x оси дате различити критеријуми одабира модела, док је на у оси дат проценат успешних фитовања одабраног скупа сигнала. На сваком графикону су црвеном бојом дати резултати за 100 одмерака, плавом за 300 одмерака, а зеленом за 1000 одмерака.

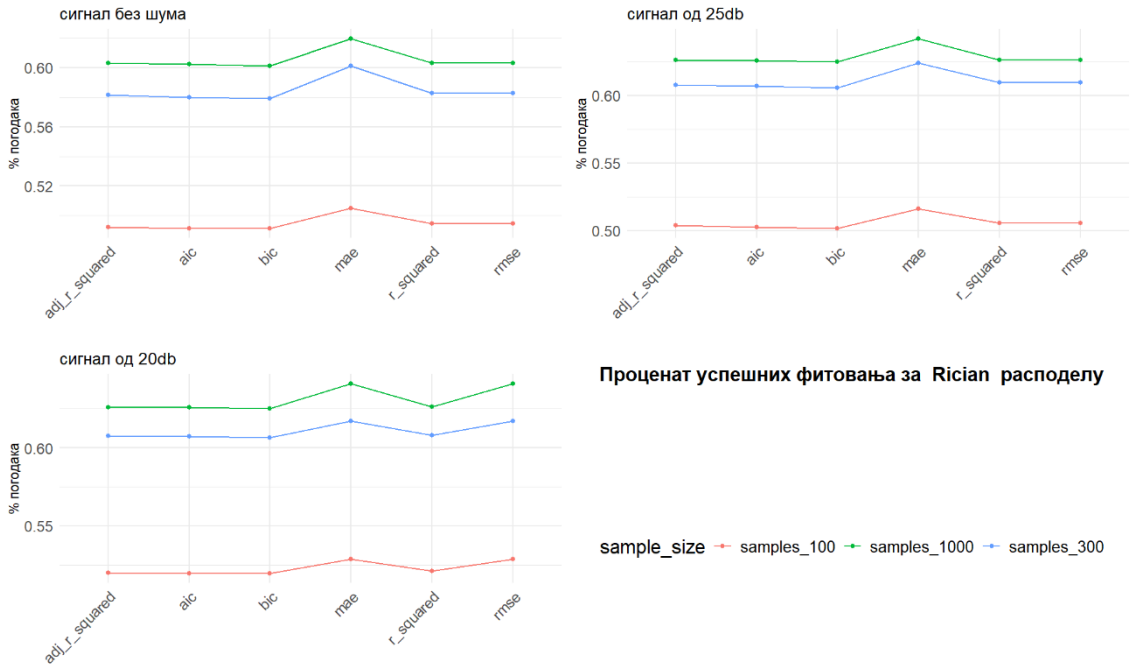
Разлика у броју исправних фитовања очигледно зависи од карактеристика података, комплексности модела и природе оптимизационог алгоритма. С обзиром на то да су у посматрано случају подаци добро дефинисани, а коришћени модели сличне комплексности, одабрани критеријуми дају резултате који су у великој мери слични. Међутим, неколико запажања се могу извући на основу упоредне анализе резултата.

На Слика 13, која се односи на гама расподелу вероватноће, примећује се знатна униформност перформанси коришћених метрика код сигнала без шума и сигнала са шумом од 25dB, уз благе осцилације. Ова чињеница указује на релативно сличне способности различитих метрика да прецизно моделују гама расподелу у таквим условима. Међутим, када је у питању сигнал са вишим нивоом шума од 20dB, уочава се јасна предност MAE критеријума у односу на остале, и то за све три испитиване дужине сигнала.



Слика 13. Графички приказ упоредне анализе фитовања сигнала гама расподеле помоћу различитих метрика

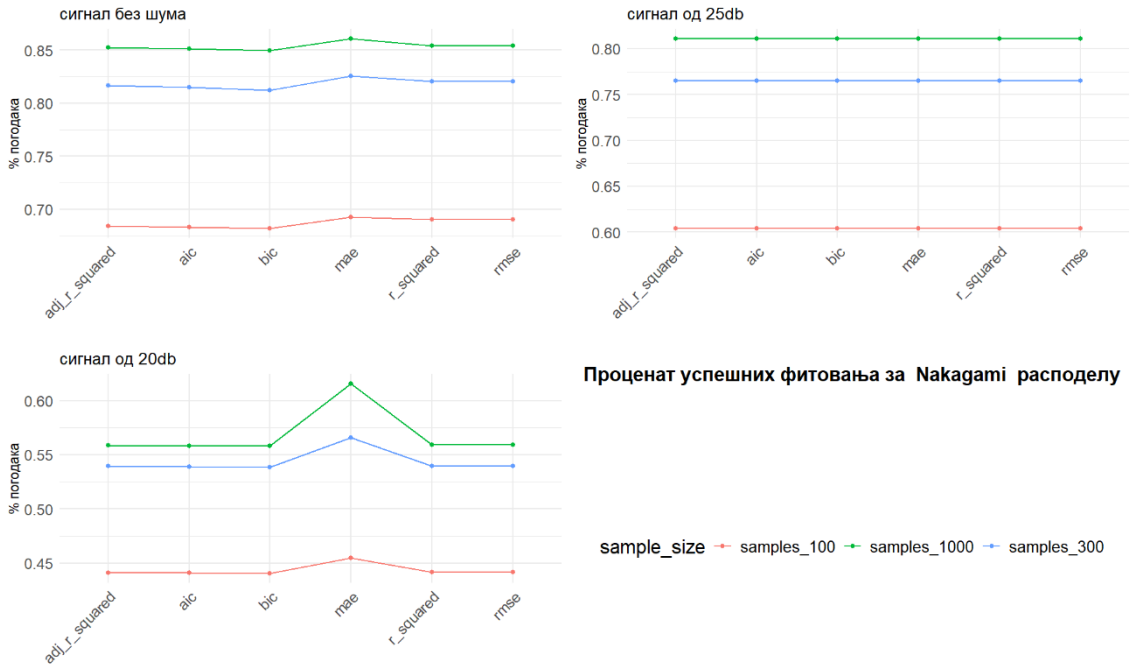
Слична тенденција је приметна и код Рајсове расподеле, представљене на Слика 14. Овде се јасно уочава да метрика MAE константно надмашује остале метрике, дајући боље резултате за све три дужине сигнала, без обзира на ниво присутног шума. С друге стране, са појавом шума у сигналу расте проценат успешног препознавања помоћу метрике RMSE.



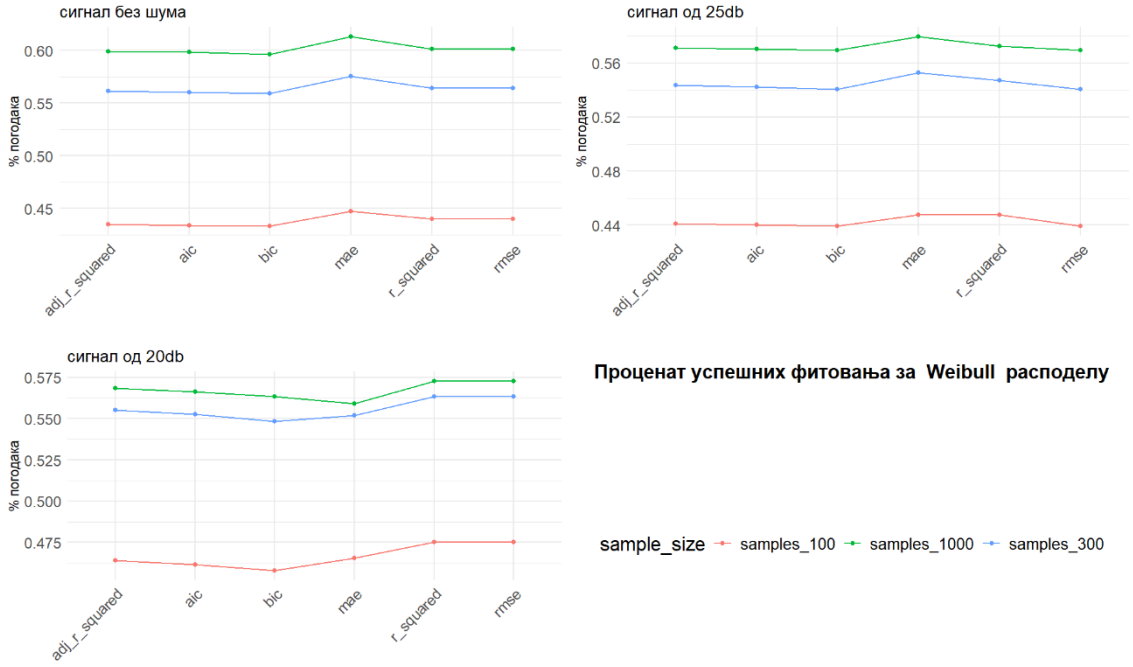
Слика 14. Графички приказ упоредне анализе фитовања сигнала Рајсове расподеле помоћу различитих метрика

Слика 15, која приказује резултате за Накагамијеву расподелу, открива прилично уједначене перформансе различитих критеријума код сигнала без шума и са шумом од 25dB. Међутим, код сигнала са 20dB шумом, за све три посматране дужине сигнала се може приметити значајан скок у проценту препознавања када се користи MAE метрика.

Случај Вејбулове расподеле представља изузетак од уоченог тренда, на Слика 16 се уочава да са повећањем нивоа шума предност MAE критеријума опада, а код сигнала од 20dB шума најбоље перформансе остварују критеријуми R^2 и RMSE. Ово одступање у сличној мери постоји код све три посматране дужине сигнала.



Слика 15. Графички приказ упоредне анализе фитовања сигнала Накагами расподеле помоћу различитих метрика



Слика 16. Графички приказ упоредне анализе фитовања сигнала Вејбулове расподеле помоћу различитих метрика

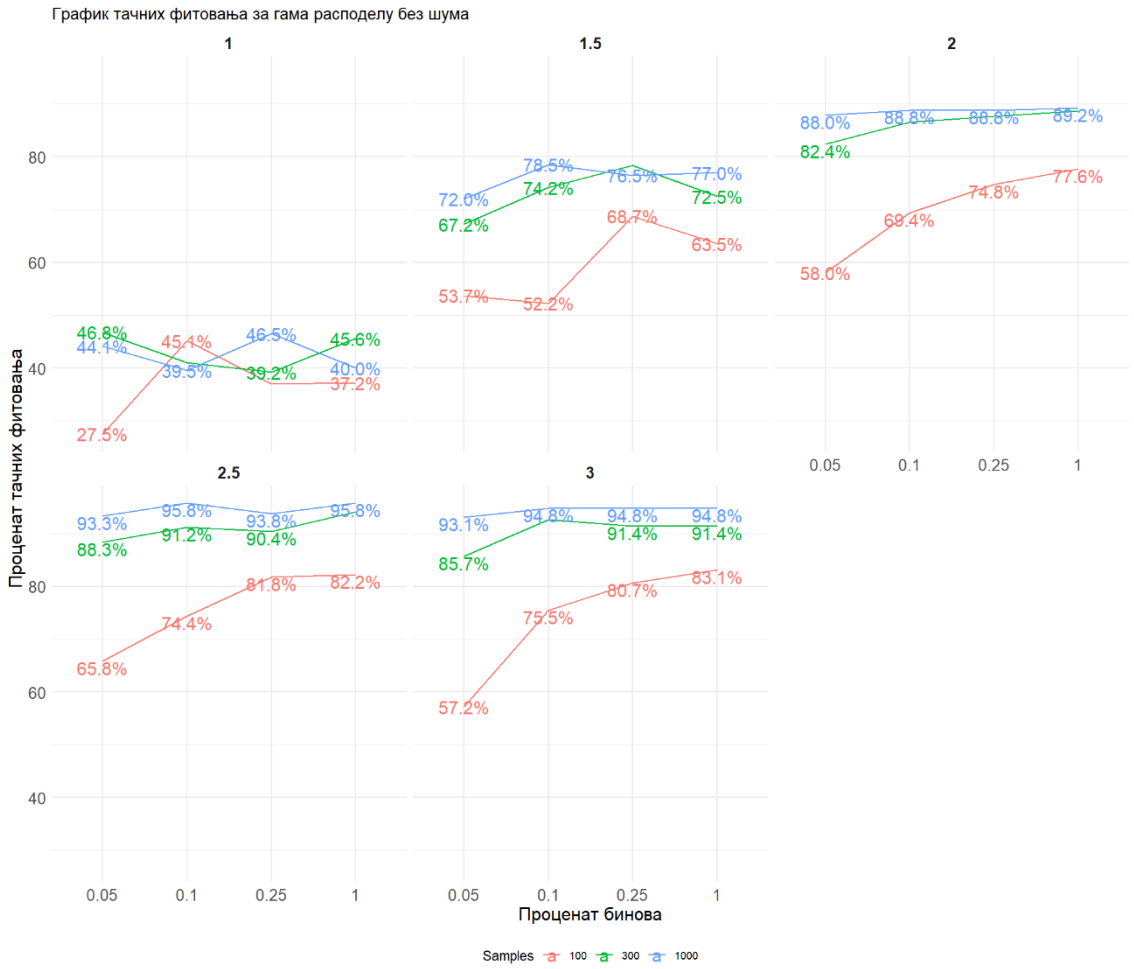
Доследни резултати фитовања коришћењем одабраних метрика, а нарочито у сценаријима са зашумљеним сигналима, показују њихову робусност, отпорност на шум и способност да пруже прецизне процене параметара различитих расподела вероватноће, што је од кључног значаја за поуздану карактеризацију и моделовање сигнала у реалним, често шумовитим условима. Ова конзистентност перформанси у различитим сценаријима омогућава примену униформног приступа за све расподеле, олакшавајући интерпретацију и упоредивост резултата.

5.3. Резултати препознавања расподела сигнала у бежичним каналима

5.3.1. Резултати препознавања гама расподеле

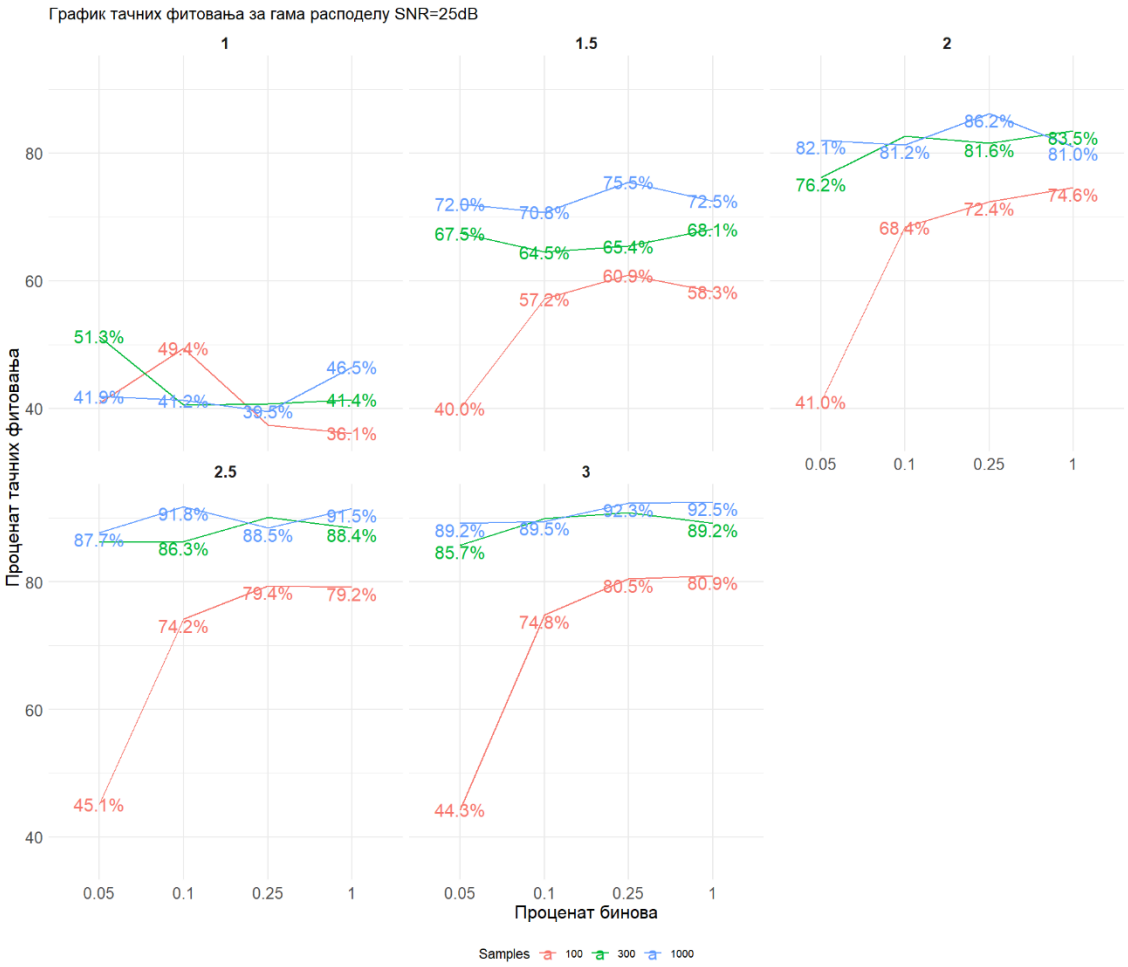
У наставку су изложени резултати препознавања случајних сигнала генерисаних по гама расподели вероватноће. Добијени резултати су приказани графички, груписани по вредности параметра облика c (у опсегу од 1 до 3, са кораком од 0,5). С обзиром на то да је у раду [21] показано, а у истраживању које је представљено овде и потврђено, параметар Ω (тј. параметар скале) нема јасног утицаја на резултате препознавања.

Као што је наведено раније, посматране сигнале смо поделили на униформне опсеге истих дужина како бисмо на основу њих израчунали НДКР. За све три посматране дужине сигнала одабрали смо број бинова који представља 5%, 10%, 25% и 100% од укупног броја одмерака сигнала. На сликама су показани резултати успешног препознавања густине вероватноћа и њихових параметара из сигнала у зависности од броја подопсега (бинова) на које су подељене вредности одмерака сигнала. Дакле, на x оси је дат нормализовани број бинова (изражен као опсег од 0,05 до 1), док је на y оси дат проценат тачних препознавања расподеле по којој је сигнал генерисан.



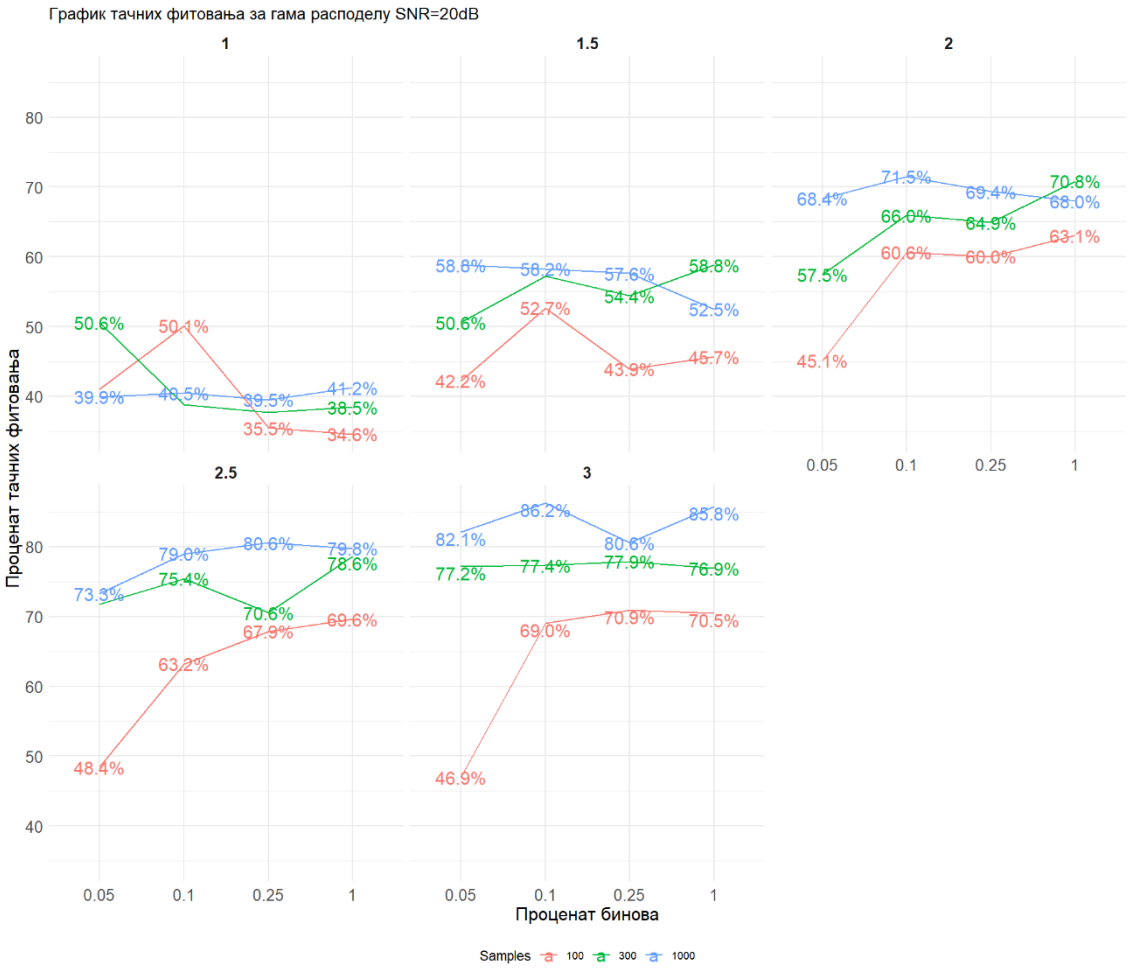
Слика 17 Графици тачности фитовања сигнала генерисаних по гама расподели без шума, груписани по вредностима параметра c

На Слика 17 су дати графици тачности фитовања сигнала генерисаних у складу са гама расподелом без икаквих унетих сметњи. На основу добијених резултата, могу се извести следећи закључци. Када је параметар облика код Гама (параметар c) и Вејбулове расподеле (параметар α) једнак јединици, тада се добија специјалан случај ове две расподеле у коме оне постају Експоненцијална расподела вероватноће [77], што значи да криве њихове CDF добијају исти облик. Услед наведеног, постоји подједнака вероватноћа да ли ће алгоритам „препознати“ сигнал као Гама или као Вејбулову расподелу. Донекле слабије перформансе се уочавају и код вредности $c = 1{,}5$, док са повећањем вредности параметра c расте и проценат успешности.



Слика 18. Графици тачности фитовања сигнала генерисаних по гама расподели са 25dB односом сигнал/шум, груписани по вредностима параметра c

С обзиром на то да се ради о незашумљеном сигналу, већ са 100 одмерака се добијају добри резултати. Даље, подизањем броја одмерака са 300 на 1000 побољшање перформанси је свега неколико процената, указујући на постојање границе после које додатни одмерци сигнала не доприносе битно бољој обради. Што се тиче утицаја параметра c, примећује се благ пораст процента препознавања са повећањем вредности овог параметра. Највећи укупни постигнути проценат успешног препознавања сигнала од 74% остварен је при комбинацији 1000 одмерака и 250 бинова.



Слика 19. Графици тачности фитовања сигнала генерисаних по гама расподели са 20dB односом сигнал/шум, груписани по вредностима параметра c

На Слика 18 су дати графици тачности фитовања сигнала генерисаних са SNR = 25dB, а на Слика 19 графици тачности фитовања сигнала генерисаних са SNR = 20dB. Са порастом шума долази до опадања перформанси алгоритма, а уочено слабљење је линеарно за цео опсег вредности параметра c. Међутим, наши експерименти су показали да се на већим SNR вредностима од посматраних, долази до додатне деградације резултата код вредности параметара c већих или једнаких 2.

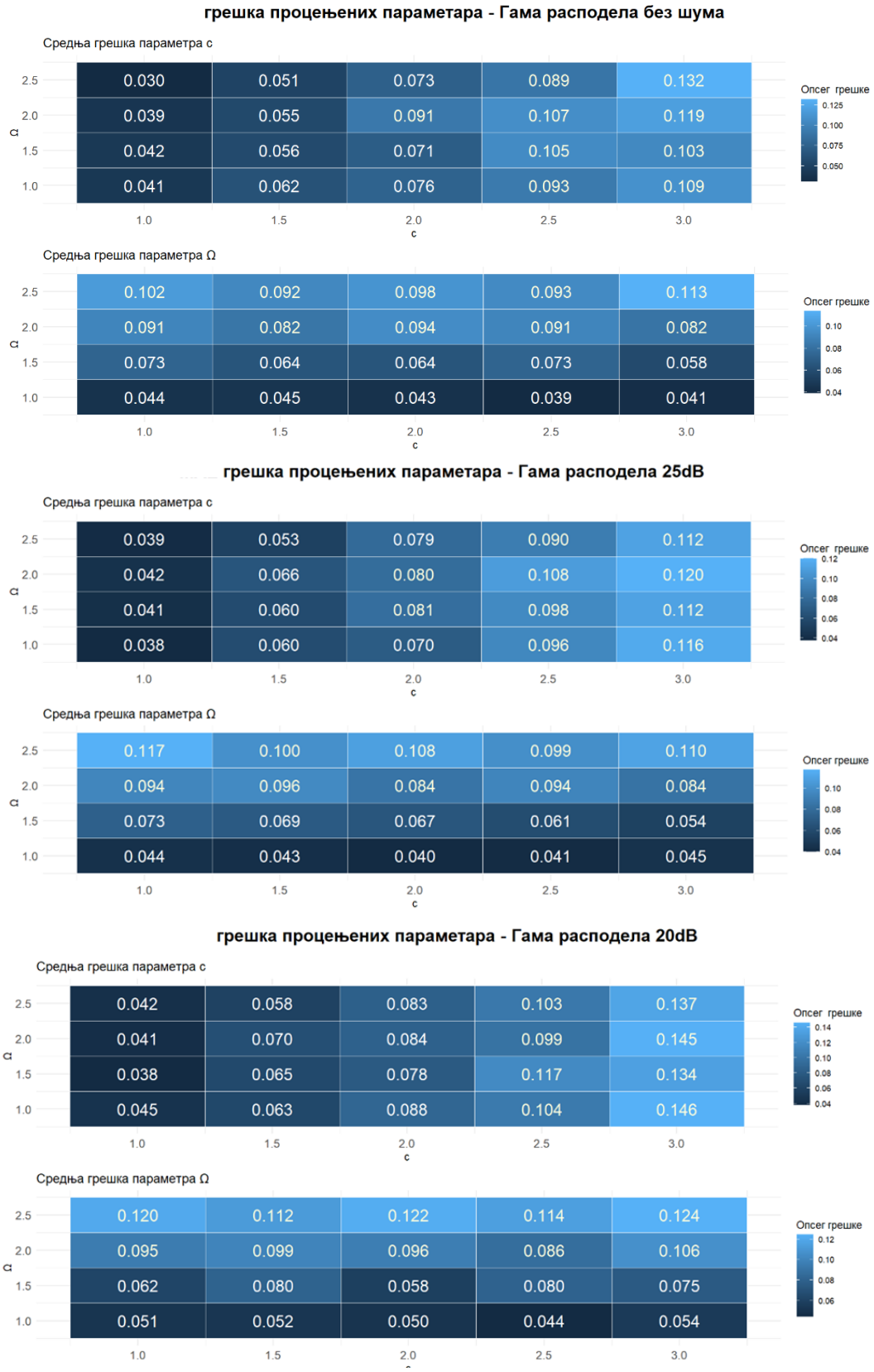
У Табела 3 су представљени процентуални резултати успешности фитовања гама модела. Резултати су дати груписано за све опсеге параметара, како би се

проучио утицај дужине сигнала и броја бинова. Из представљених резултат се уочава да се са повећањем броја бинова повећава и тачност фитовања, осим код дужине 1000 одмерака, где се може приметити да НДКР од 10% бинова даје у благој мери боље резултате у односу на друге вредности процента бинова.

Табела 3. Укупни резултати препознавања сигнала гама расподеле вероватноће

шум	бр. одмерака	број бинова			
		5%	10%	25%	100%
без сметњи	100	52.4	63.3	68.6	68.7
	300	74.1	77.1	77.4	78.4
	1000	78.1	79.4	80	79.4
25dB	100	42.2	64.8	66.1	65.8
	300	73.4	72.8	73.7	74.1
	1000	74.6	74.9	76.4	76.8
20dB	100	44.7	59.1	55.6	56.7
	300	61.5	63.0	61.1	64.7
	1000	64.5	67.1	65.5	65.4
укупно	100	46.4	62.4	63.4	63.7
	300	69.7	71.0	70.7	72.4
	1000	72.4	73.8	74.0	73.9

На Слика 20 је дата топлотна мапа грешака процењених вредности параметара гама расподеле у сва три посматрана случаја. Може се уочити да да са порастом вредности параметра облика и параметра скале опада тачност процене параметра, с тим што су ова два пораста међусобно независна. Додатно, посматрани нивои шума не утичу на смањење тачности процене параметара.



Слика 20. Топлотна мапа грешака процењених вредности параметара гама расподеле код сигнала без шума (а), са SNR=25dB (б), и са SNR=20dB (в)

5.3.2. Резултати препознавања Рејлијеве расподеле

Као што је наведено раније, Рејлијева расподела вероватноће представља специјалан случај неколико других расподела:

- За $K = 0$, формула (3.6) се своди на формулу (3.3), тј. добија се специјални случај када се Рајсова расподела своди на Рејлијеву расподелу [78].
- За $m = 1$ формула (3.8 се своди на формулу (3.3), тј. добија се специјални случај када се Накагами-м дистрибуција своди на Рејлијеву расподелу [41].
- За $\alpha=2$ формула (3.11) се своди на формулу (3.3), тј. добија се специјални случај када се Вејбулова расподела своди на Рејлијеву расподелу [79].

Услед наведених односа између расподела, сигнали генерисани по једној расподели вероватноћа истовремено одговарају и другим расподелама. Чињеница да је Рејлијева расподела обухваћена другим расподелама вероватноће као специјалан случај, у складу је са резултатима које смо представили у раду [21].

У Табела 4 [21]. су изложени резултати препознавања расподеле скупа сигнала који су генерисани по Рејлијевој расподели вероватноће, при чему се може видети да су резултати препознавања подељени између ње и расподела са којима дели наведене специјалне случајеве

Табела 4. Резултати препознавања сигнала Рејлијеве расподеле вероватноће [21]

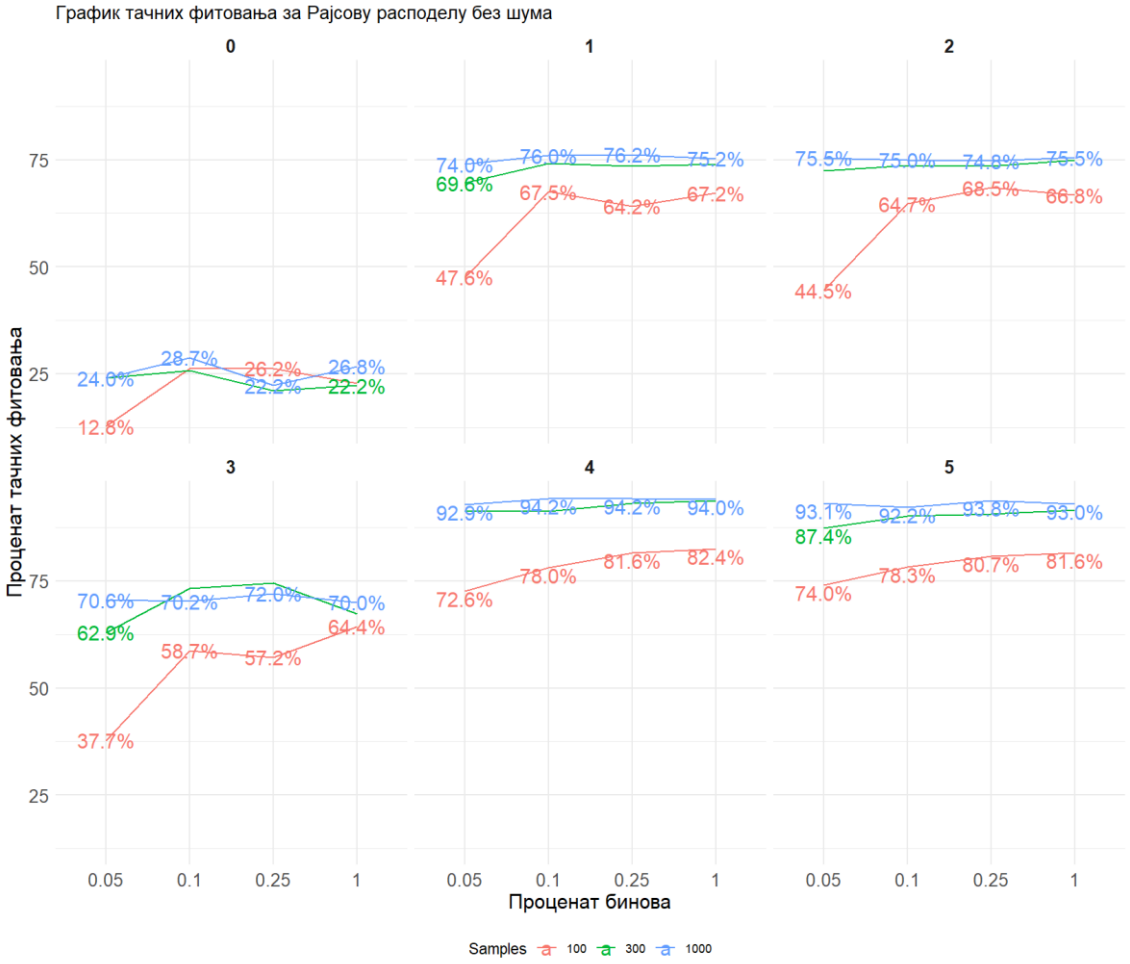
	Rayleigh	**Rice**	**Nakagami**	**Weibull**	**TOTAL**
$1.0 \leq \Omega \leq 1.3$	28.5	19	22	30.5	100
$1.4 \leq \Omega \leq 1.7$	28.75	15.75	24	31.5	100
$1.8 \leq \Omega \leq 2.1$	33.5	21.75	19.25	25.5	100
$2.2 \leq \Omega \leq 2.5$	32.5	22.25	18.25	27	100
AVERAGE	30.81	19.69	20.88	28.63	100

Слика 21. Топлотна мапа грешака процењених вредности параметара Рејлијеве расподеле код сигнала без шума, са SNR=25dB, и са SNR=20dB

На Слика 21. Топлотна мапа грешака процењених вредности параметара Рејлијеве расподелеје дата топлотна мапа грешака процењених вредности параметра Рејлијеве расподеле вероватноће. Слично резултатима датим у претходном поглављу, уочава се зависност грешке од вредности параметра Ω.

5.3.3. Резултати препознавања Рајсове расподеле

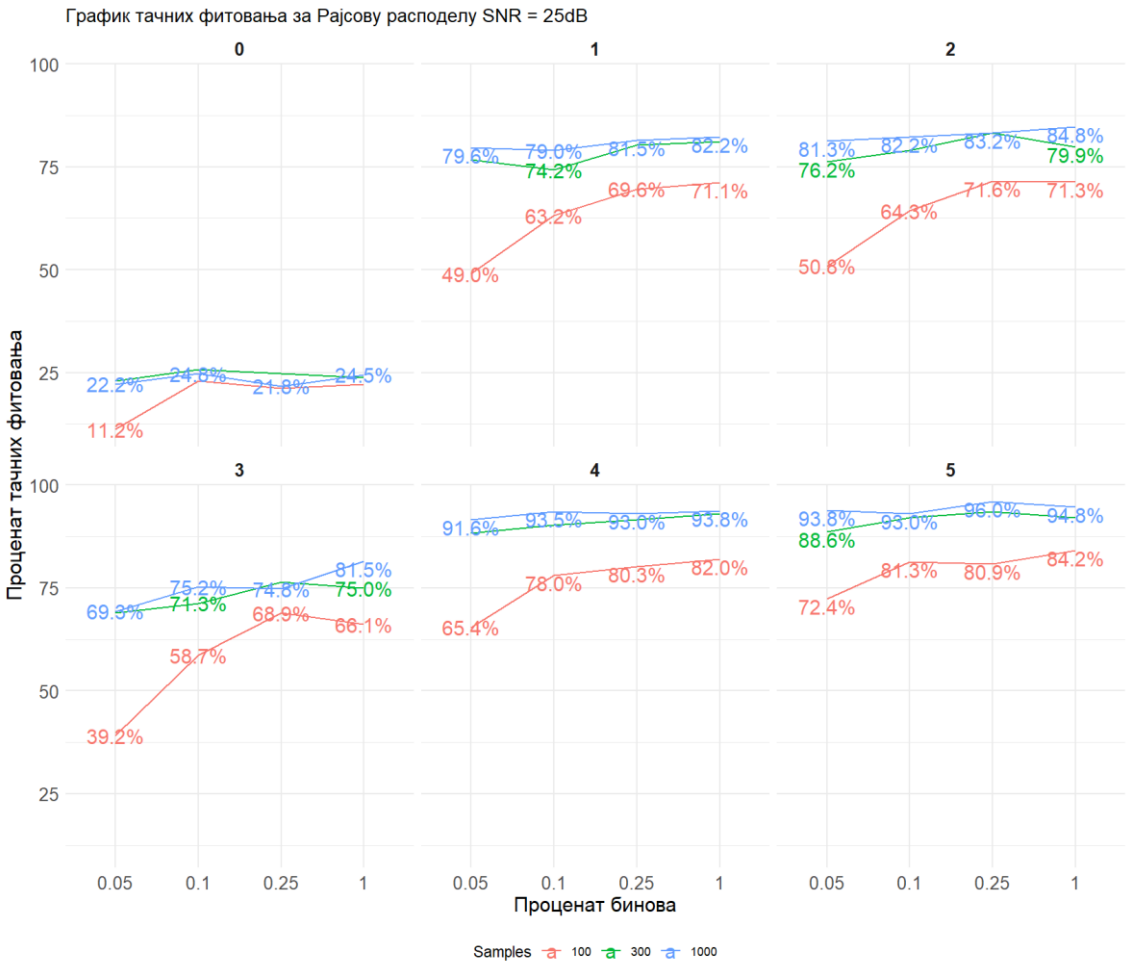
У наставку су изложени резултати препознавања скупа сигнала генерисаних по Рајсовој расподели вероватноће. Добијени резултати су приказани графички, груписани по вредности параметра облика K.



Слика 22 Графици тачности фитовања сигнала генерисаних по Рајсовој расподели без сметњи, груписани по вредностима параметра К

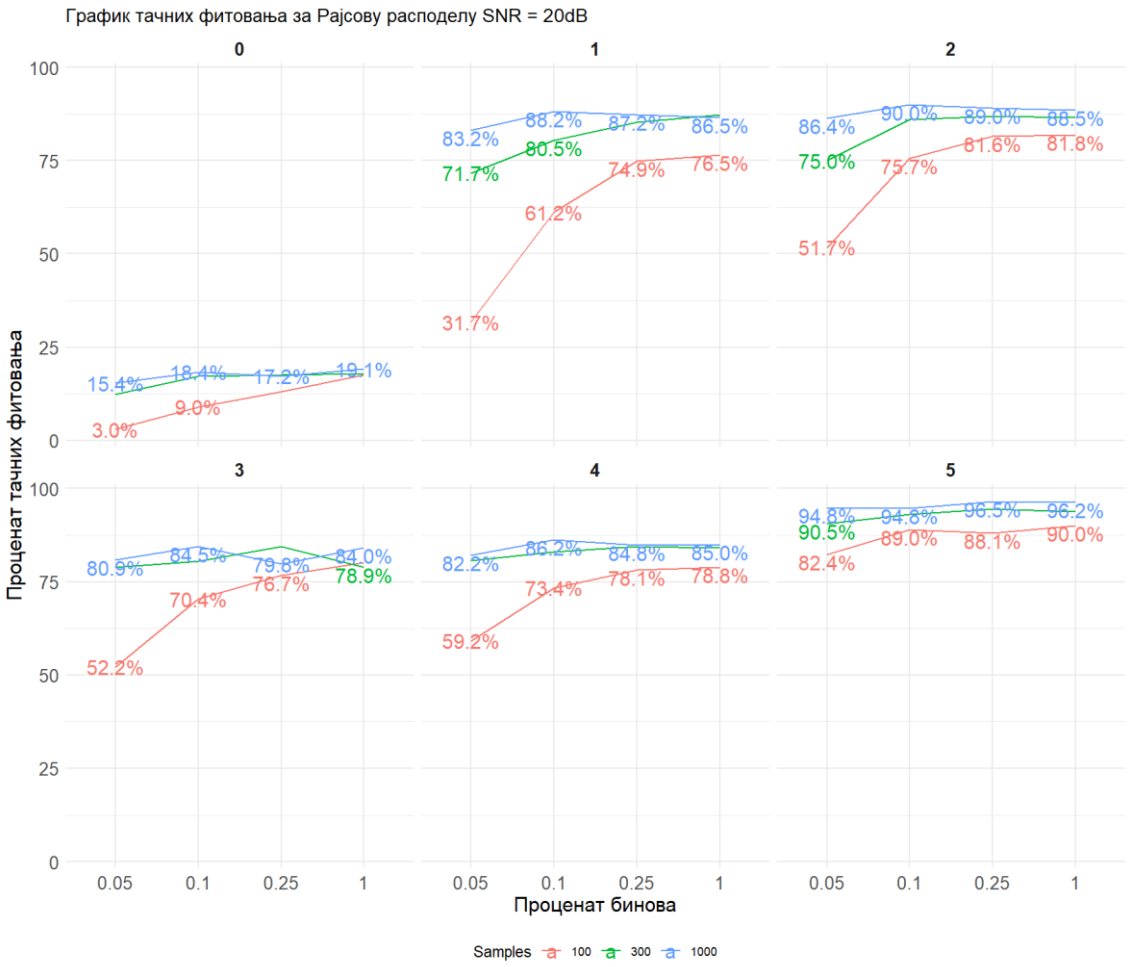
На Слика 22 су дати резултати препознавања сигнала генерисаних по Рајсовој расподели вероватноће без присуства шума. Резултати су изложени за различите вредности фактора K. Из добијених резултата се види да се при већим вредностима параметра К добијају бољи резултати препознавања. Једини случај када су добијени лоши резултати јесте за случај када је $K = 0$, односно за специјални случај[2] када Рајсова расподела постаје Рејлијева.

[2] Овај специјални случај се поклапа са специјалним случајевима када се и Накагамијева расподела (за $m = 1$) и Вејбулова расподела (за $\alpha = 2$) своде на Рејлијеву расподелу.



Слика 23 Графици тачности фитовања сигнала генерисаних по Pajсовој расподели са 25dB односом сигнал/шум, груписани по вредностима параметра K

На Слика 23 су дати графици тачности фитовања за сигнала генерисаних са SNR = 25dB, а на Слика 24 графици тачности фитовања сигнала генерисаних са SNR = 20dB. Оно што делује контра-интуитивно јесте повећање процента препознавања за повећањем нивоа SNR, као и то да се истовремено повећава проценат препознавања за краће дужине сигнала. Међутим, на већим SNR вредностима од посматраних, долази до деградације резултата.



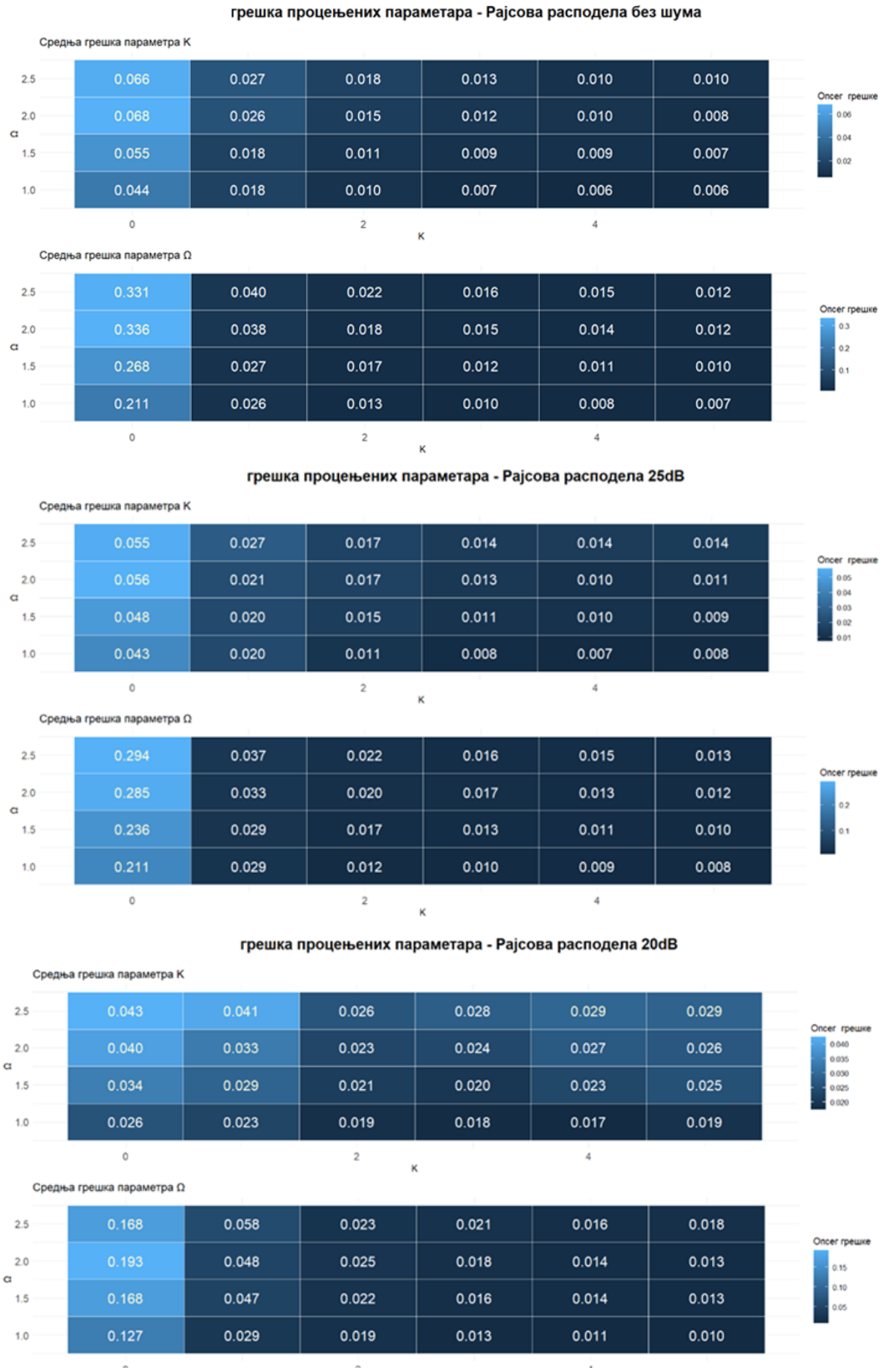
Слика 24 Графици тачности фитовања сигнала генерисаних по Рајсовој расподели са 20dB односом сигнал/шум, груписани по вредностима параметра K

У Табели 5 су представљени процентуални резултати успешности фитовања модела Рајсове расподеле вероватноће. Из представљених резултата се уочава да повећање броја одмерака са 300 на 1000 доводи до пораста процента препознавања који је нижи него код других посматраних модела расподела.

Табела 5. Укупни резултати препознавања сигнала Рајсове расподеле вероватноће

шум	бр. одмерака	проценат бинова			
		0.05	**0.1**	**0.25**	**1**
без сметњи	100	47.80	61.63	62.50	63.69
	300	67.73	71.25	70.97	70.60
	1000	73.04	72.75	72.21	72.42
25dB	100	47.66	60.88	64.96	65.69
	300	70.10	71.98	74.93	74.07
	1000	74.42	74.63	75.04	76.92
20dB	100	42.37	58.44	63.55	65.21
	300	67.45	72.39	74.56	73.70
	1000	73.67	76.00	74.79	75.50
укупно	100	45.94	60.32	63.67	64.86
	300	68.43	71.87	73.49	72.79
	1000	73.71	74.46	74.01	74.95

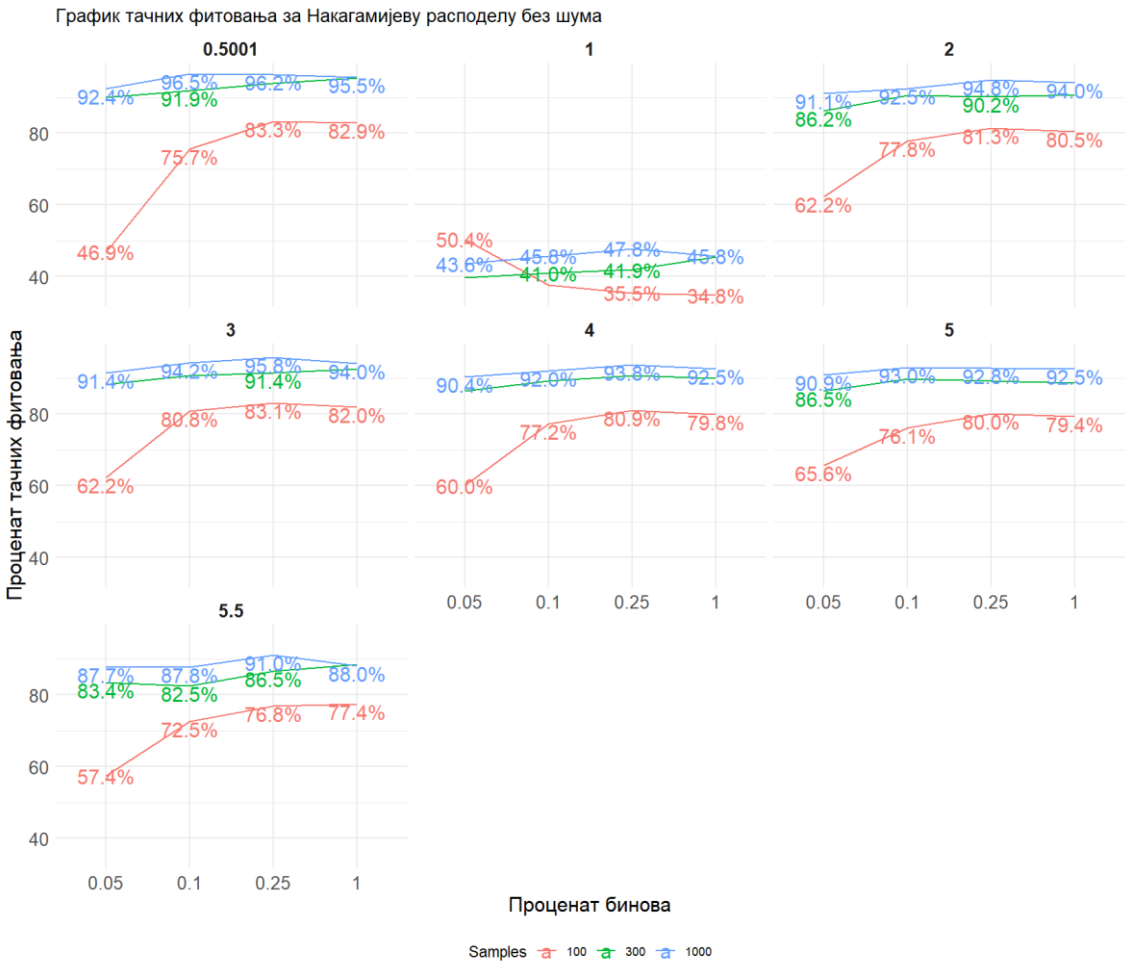
На Слици 25. је дата топлотна мапа грешака процењених вредности параметара Рајсове расподеле за сва три нивоа SNR. Уочава се да су грешке процене параметра K веће на нижим вредностима, док су константе када је $K \geq 2$. Додатано, уочава се тренд да са повећањем нивао шума у сигналу долази до истовременог повећања тачности процене параметра Ω, док тачност процене параметра K опада.



Слика 25. Топлотна мапа грешака процењених вредности параметара Рајсове расподеле код сигнала без шума (а), са SNR=25dB (б) и са SNR=20dB (в)

5.3.4. Резултати препознавања Накагамијеве расподеле

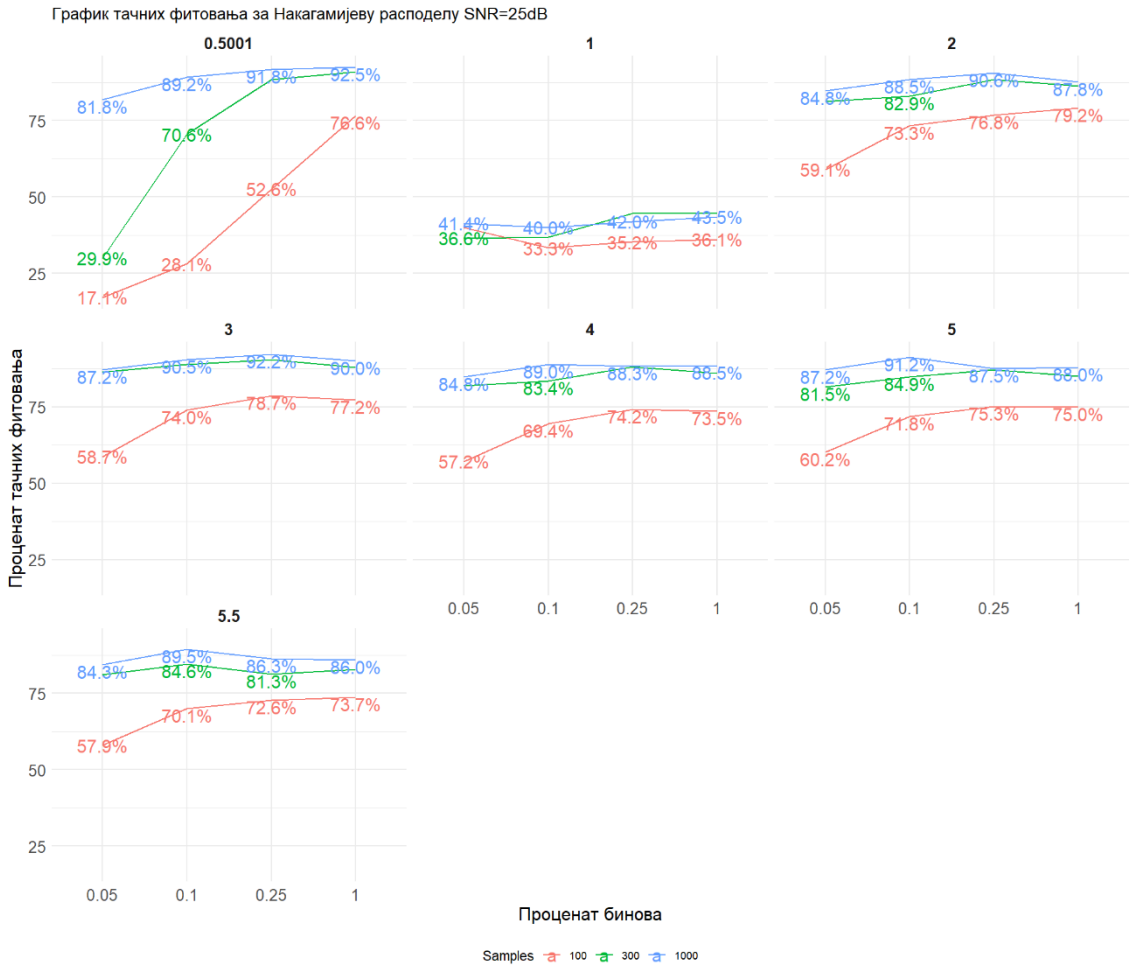
У овом поглављу су дати резултати препознавања сигнала генерисаног по Накагамијевој расподели вероватноће. Добијени резултати су приказани графички, груписани по вредности параметра облика m.



Слика 26 Графици тачности фитовања сигнала генерисаних по Накагамијевој расподели без сметњи, груписани по вредностима параметра m

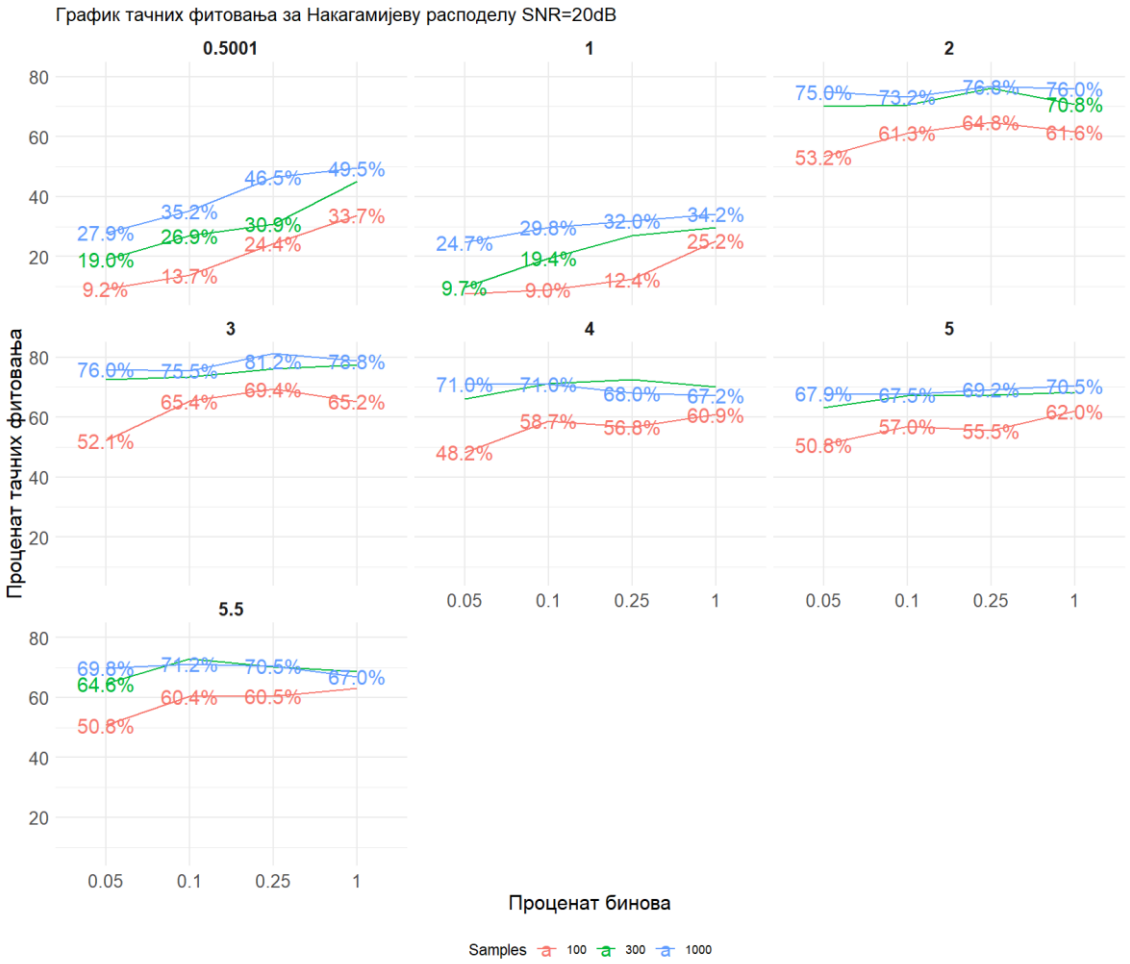
На Слика 26 су приказани резултати препознавања сигнала генерисаних по Накагамијевој расподели. На датој слици се примећује висок проценат (око 90%) успешности препознавања сигнала Накагамијеве расподеле, осим у случају када је параметар $m = 1$. Као што је речено раније, ово је у складу са

чињеницом да за ту вредност параметра Накагамијева расподела постаје један од специјалних случајева Рејлијеве расподеле вероватноће (као што је показано на Слика 4).



Слика 27 Графици тачности фитовања сигнала генерисаних по Накагамијевој расподели са 25dB односом сигнал/шум, груписани по вредностима параметра m

На Слика 27 су дати графици тачности фитовања сигнала генерисаних са SNR = 25dB, а на Слика 28 графици тачности фитовања сигнала генерисаних са SNR = 20dB. Шум од 25dB нема значајнијег утицаја на резултате препознавања код виших вредности параметра m, али се уочава пад за вредности $m = 0.5$ и $m = 1$. Ово слабљење перформанси је нарочито уочљиво код сигнала са SNR = 20dB.



Слика 28 Графици тачности фитовања сигнала генерисаних по Накагамијевој расподели са 20dB односом сигнал/шум, груписани по вредностима параметра m

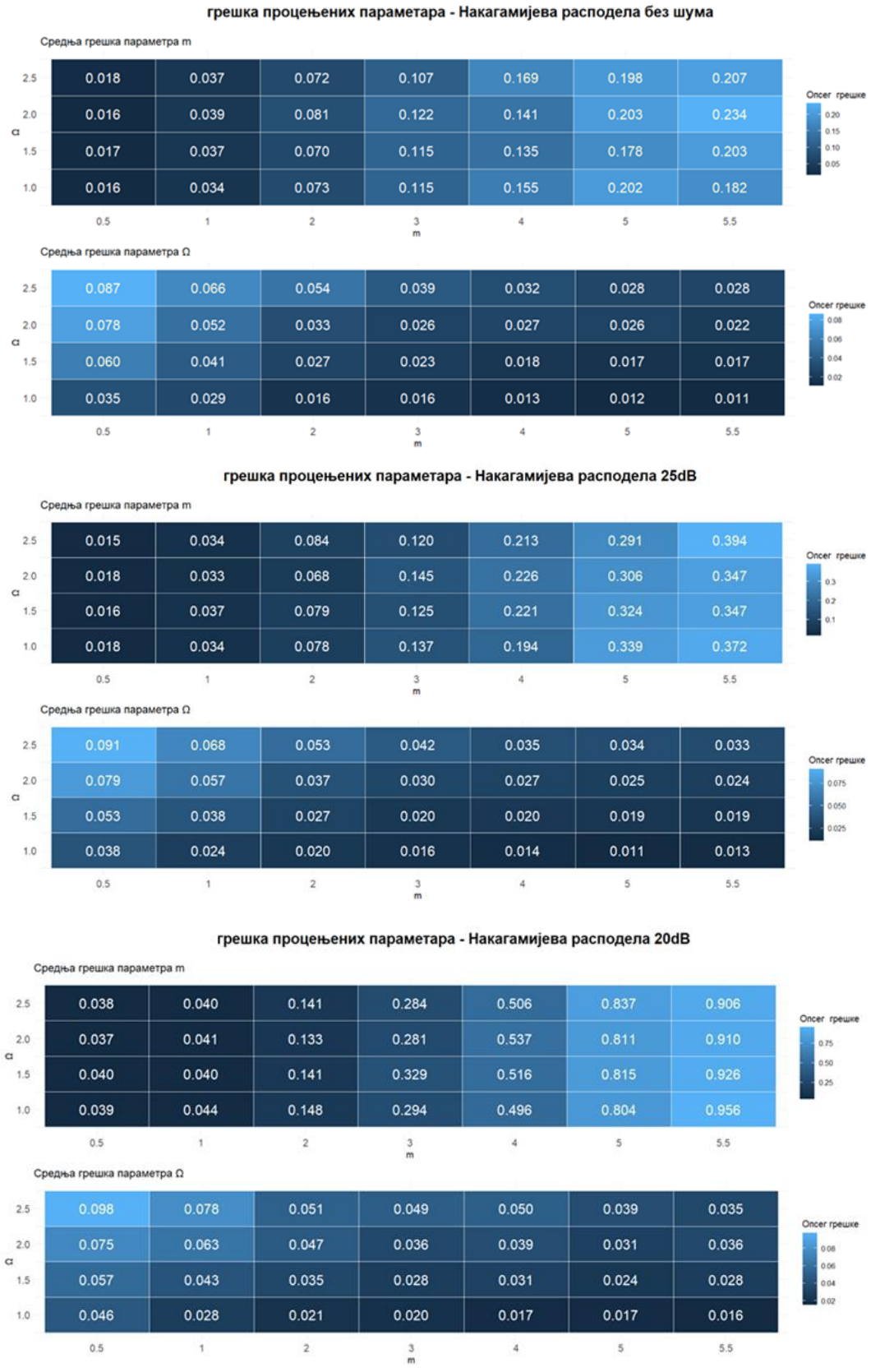
У Табела 6. Укупни резултати препознавања сигнала Накагамијеве расподеле вероватноће су представљени процентуални резултати успешности фитовања модела Накагамијеве расподеле. Најбољи резултати препознавања се добијају за дужину сигнала од 1000 одмерака и 25% бинова. Ова предност у односу на остале процентуалне вредности бинова, и то код сва три нивоа зашумљености. Иако ова предност коришћења 25% бинова опада са порастом нивоа шума у сигналу, али остаје присутна. Ови резултати указују на постојање оптималног односа између дужине сигнала и резолуције његовог

представљања путем броја бинова, специфичног за одређену расподелу и ниво шума.

Табела 6. Укупни резултати препознавања сигнала Накагамијеве расподеле вероватноће

шум	бр. одмерака	проценат бинова			
		0.05	**0.1**	**0.25**	**1**
без сметњи	100	57.8	71.1	74.4	73.8
	300	80.1	82.2	83.4	84.5
	1000	84.8	86	87.4	86
25dB	100	50.1	60	66.5	70.2
	300	68.4	76	81.2	80.5
	1000	78.8	82.6	82.7	82.3
20dB	100	37.5	44.4	47	53.1
	300	49.5	55.4	60.1	61.4
	1000	58.9	60.5	63.5	63.3
укупно	100	48.5	58.5	62.6	65.7
	300	66	71.2	74.9	75.5
	1000	74.2	76.4	77.9	77.2

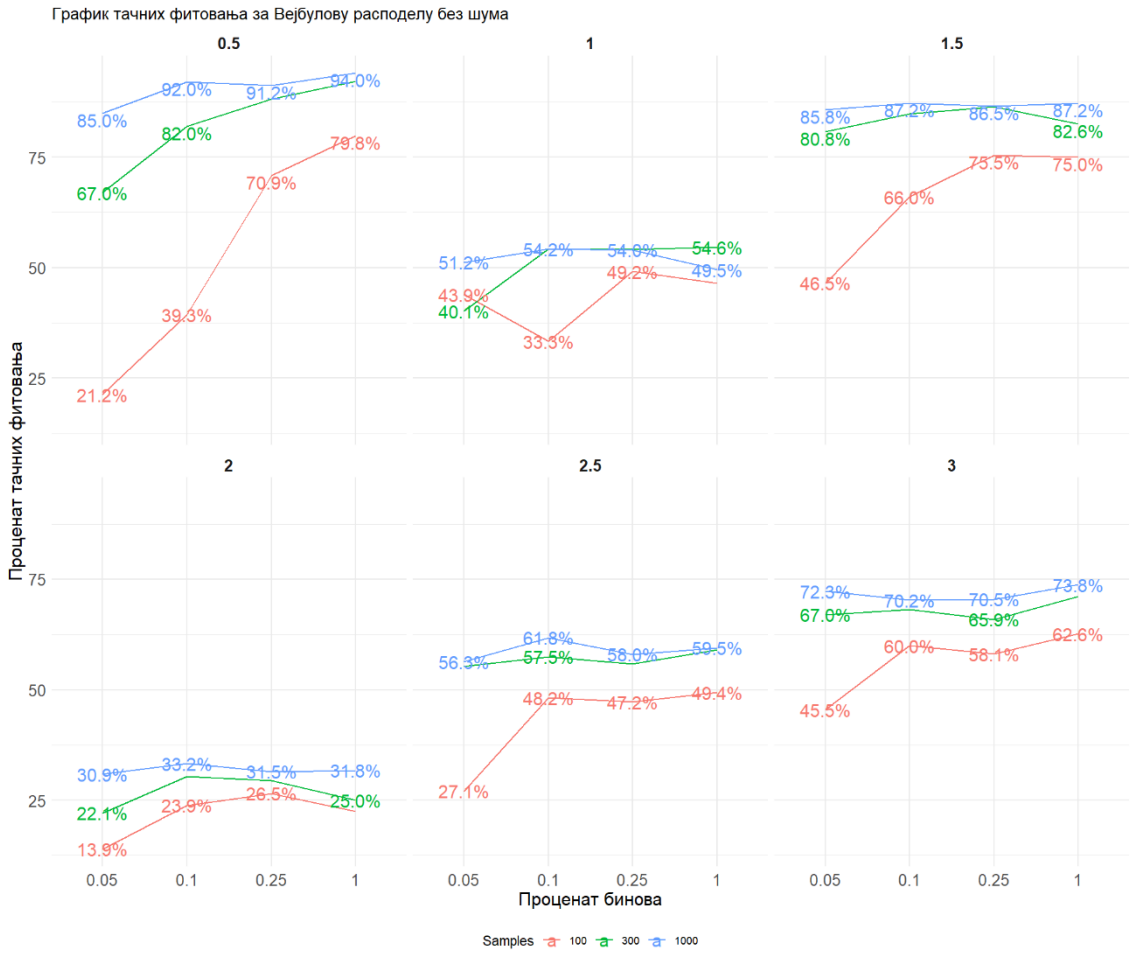
На Слици 29. је дата топлотна мапа грешака процењених вредности параметара гама расподеле у сва три посматрана случаја. И у овом случају се уочава Резултати показују да са порастом вредности параметра облика и параметра скале опада тачност процене параметра, међутим, може се уочити да грешка процена параметра Ω опада са повећањем вредности параметра m. Додатно, може се приметити да грешка процењене вредности параметра m код сигнала без шума расте линеарно, овај раст постаје експоненцијалан са повећањем нивоа SNR.



Слика 29. Топлотна мапа грешака процењених вредности параметара Накагамијеве расподеле код сигнала без шума (а), са SNR=25dB (б) и са SNR=20dB (в)

5.3.5. Резултати препознавања Вејбулове расподеле

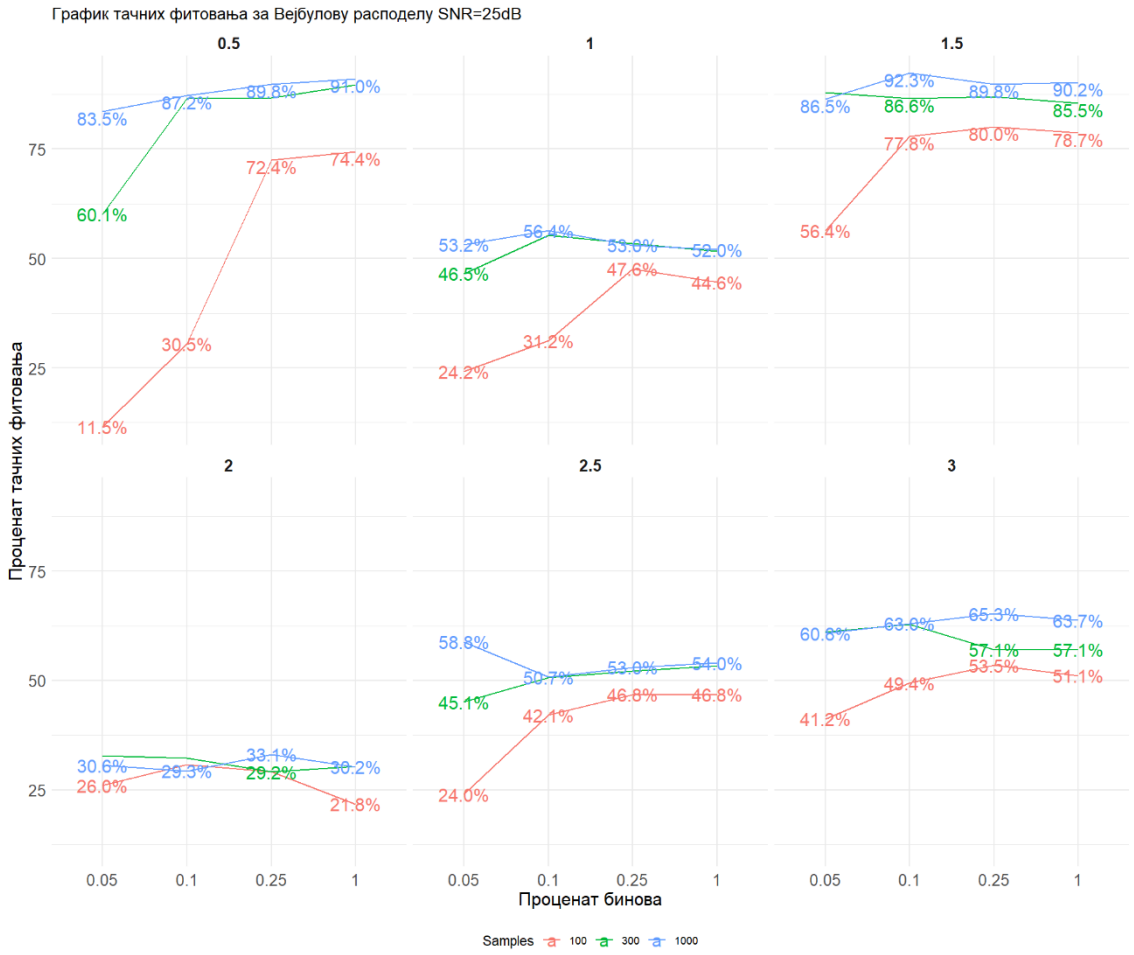
У овом поглављу су дати резултати препознавања скупа сигнала генерисаних по Вејбуловој расподели вероватноће. С обзиром на то да нема видног утицаја параметра Ω [21], на графицима је представљена зависност тачности препознавања од вредности параметра α.



Слика 30 Графици тачности фитовања сигнала генерисаних по Вејбуловој расподели без сметњи, груписани по вредностима параметра α

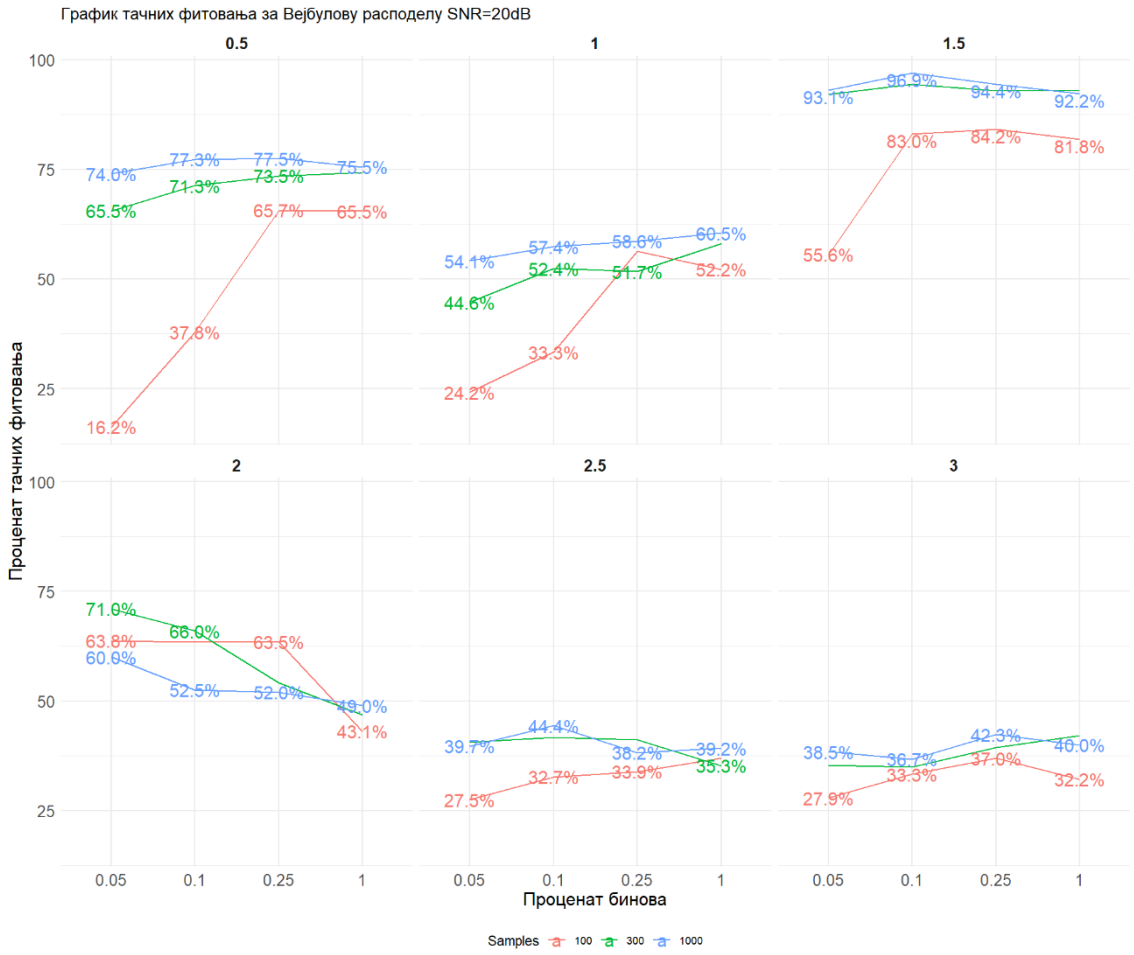
Слика 30 приказује резултате препознавања сигнала генерисаних по Вејбуловој расподели без присуства шума. Може се уочити да са порастом

вредности параметра α долази до опадања процента препознавања случајних сигнала генерисаних по Вејбуловој расподели. Значајан пад процента успешних фитовања постоји у случајевима када је $\alpha = 1$ и $\alpha = 2$. Као што је раније објашњено, ово су специјални случајеви Вејбулове расподеле када се она своди на експоненцијалну расподелу (за $\alpha = 1$) и када се своди на Рејлијеву расподелу (за $\alpha = 2$). Разлика у процентима у ова два случаја произилази из чињенице да у првом случају имамо преклапање две расподеле, док у другом имамо преклапање четири расподеле вероватноће.



Слика 31 Графици тачности фитовања сигнала генерисаних по Вејбуловој расподели са 25dB односом сигнал/шум, груписани по вредностима параметра α

На Слика 31 су приказани резултати препознавања сигнала генерисаних по Вејбуловој расподели за SNR=25dB, а на Слика 32 су приказани резултати за скуп сигнала са SNR=20dB. Са повећањем нивоа шума долази до наглог опадања успешности у случајевима са вишим вредностима параметра α. Код опсега вредности који обухватају специјалне случајеве, приметан је пораст броја препознавања модела Вејбулове расподеле. Сличан феномен је уочен и код других расподела вероватноће, те у специјалним случајевима расподеле добијамо боље препознавање модела Вејбулове расподеле.



Слика 32 Графици тачности фитовања сигнала генерисаних по Вејбуловој расподели са 20dB односом сигнал/шум, груписани по вредностима параметра α

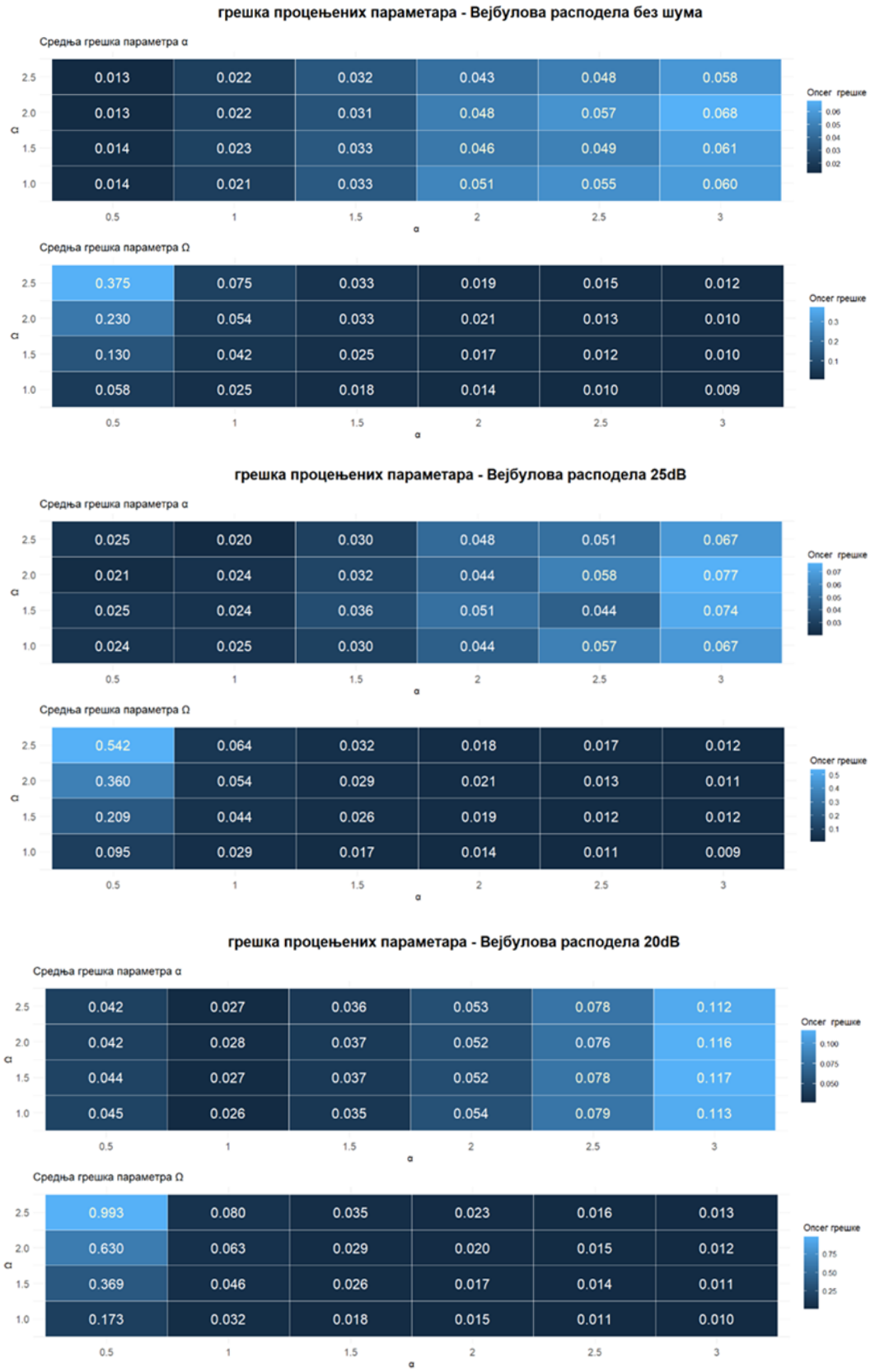
У Табела 7. Укупни резултати препознавања сигнала Вејбулове расподеле вероватноће су приказани резултати препознавања за све нивое шумова, као

и укупне перформансе система. Иако су резултати представљени груписано за све вредности параметара, што ствара утисак да не долази до веће деградације перформанси, као што је наведено, увођењем шума у сигнал долази до већих осцилација перформанси за различите вредности параметра α. Занимљиво је да сигнал од 300 одмерака даје приближно једнаке перформансе као сигнал од 1000 одмерака. Мада је разлика незнатна, најбољи резултати се постижу за вредности бинова 10% и 25%.

Табела 7. Укупни резултати препознавања сигнала Вејбулове расподеле вероватноће

шум	**бр. одмерака**	**проценат бинова**			
		0.05	**0.1**	**0.25**	**1**
без сметњи	100	33.0	45.1	54.6	56
	300	55.4	62.8	63.3	64.1
	1000	63.6	66.5	65.3	66
25dB	100	30.5	43.6	54.9	52.9
	300	55.6	62.4	60.9	61.3
	1000	62.2	63.2	64	63.5
20dB	100	35.8	47.3	56.8	51.9
	300	58.2	60.1	58.8	58.2
	1000	59.9	60.9	60.5	59.4
укупно	100	33.1	45.3	55.4	53.6
	300	56.4	61.8	61	61.2
	1000	61.9	63.5	63.3	63

На Слици 33. је приказана топлотна мапа грешака препознавања параметра Вејбулове расподеле. Са увођењем шума, грешке препознавања параметра α расту на крајевима посматраног опсега вредности, док на грешке препознавања у опсегу специјалних случајева увођење шума нема утицаја - ова робусност модела у тим случајевима објашњава добре перформансе за наведене случајеве. Грешка параметра омега је практично константна на целом опсегу, осим код $\alpha = 0{,}5$ где се ове грешке код $SNR = 20dB$ утростручују. Такође, уочава се да грешка расте са порастом вредности параметра размере.



Слика 33. Топлотна мапа грешака процењених вредности параметара Вејбулове расподеле код сигнала без шума (а), са SNR=25dB (б) и са SNR=20dB (в)

6. ПРИМЕНА НЕЛИНЕАРНЕ РЕГРЕСИЈЕ ЗА АНАЛИЗИРАЊЕ ИНТЕРНЕТ САОБРАЋАЈА

Огромне размере и комплексност интернета има несагледив утицај на свет данашњице. Како се ова огромна мрежа шири, разумевање њених замршених динамика постаје све важније. Под ширењем не означавамо овде само физичку инфраструктуру, већ и обимне и разноврсне податке који се генеришу. Да би се из ових огромних количина података које генерише интернет саобраћај извукли смислени увиди, уочили трендови и предвидели обрасци, неопходно је применити неке од техника истраживања података. Применом техника машинског учења, могуће је открити скривене корелације, идентификовати аномалије и екстраховати вредне увиде из сирових података интернет саобраћаја. Ови увиди могу бити од непроцењиве вредности у оптимизацији мрежних ресурса, побољшању квалитета услуга, откривању безбедносних претњи и развоју напредних апликација прилагођених корисничким потребама.

Расподеле вероватноће могу бити моћан алат за представљање стохастичке природе интернет саобраћаја и образаца коришћења, пружајући основу за предиктивну аналитику и процесе доношења одлука. С обзиром на наведено, ово поглавље представља примену нелинеарне регресије на податке о интернет саобраћају као студију случаја.

Као што је речено, развој технологија и мрежних услуга на интернету којем сведочимо указује на потребу за описивањем промена у интернет саобраћају. Информације добијене на овај начин се могу искористити на много начина, укључујући следеће:

- Предвиђање загушења мреже [80]: Разумевањем расподеле вероватноће дужине пакета може помоћи администраторима мреже да предвиде загушење и предузму одговарајуће мере да спрече појаву уских грла.

- Оптимизација мрежних перформанси [81]: Анализа карактеристика мрежног саобраћаја може помоћи у оптимизацији перформанси мреже тако што омогућује доношење исправних одлука о прилагођавању параметара као што су величине бафера и брзине преноса.
- Откривање аномалија и безбедност мреже [82]: Поређење саобраћаја у реалном времену са очекиваним обрасцима заснованим на историјским подацима може помоћи у откривању аномалија или потенцијалних напада на мрежу, чиме се побољшава њена безбедност.

DDoS напади (енгл. *Distributed Denial of Service*) је један од најзаступљенијих напада преко интернета. Основни циљ оваквог напада јесте да исцрпи ресурсе сервера и омете мрежне услуге. У основи, ова техника напада има две категорије: напад за преузимање протока и напад за преузимање ресурса, а ове се своде на то да нападачи (тзв. ботнети) преплаве Веб сајт интернет саобраћајем тако да се онемогући његово нормално коришћење за нормалне кориснике. Истраживање података нуди моћне алате за борбу против ове претње. Анализом великих количина података о мрежном саобраћају, системским логовима и другим изворима, могуће је идентификовати обрасце и аномалије који указују на ДДОС нападе. Ово омогућава организацијама да ефикасније реагују и заштите своје системе [83].

Традиционални приступи откривању DDoS напада укључују статистичке методе као што је мерење ентропије мрежног саобраћаја [84], или коришћење Бајесове класификације. Међутим, ове методе могу бити ограничене у њиховој способности да уоче сложене обрасце и аномалије у мрежном саобраћају, нарочито када су напади маскирани или користе нове технике.

Нелинеарна регресија се јавља као моћан алтернативан приступ за откривање DDoS напада. Ова техника може моделовати нелинеарне односе између различитих карактеристика мрежног саобраћаја, омогућавајући боље уочавање суптилних аномалија које могу указивати на напад.

Једна од предности нелинеарне регресије је њена флексибилност. Различити модели нелинеарне регресије могу се користити у зависности од врсте података и карактеристика мрежног саобраћаја. Ово омогућава прилагођавање приступа откривању специфичним потребама сваке организације. Друга предност је отпорност на нове технике напада. Нелинеарни модели регресије су у стању да уоче сложене обрасце у подацима, чак и када су напади маскирани или користе нове технике. Ово их чини ефикаснијим у откривању нових и еволуирајућих претњи.

Неке од најважнијих особина за одређивање да ли пакет припада нормалном саобраћају или саобраћају напада јесу проток пакета и проток бајтова. У раду [85] се наводи да се са порастом вредности ова два параметра истовремено расте и вероватноћа да пакет припада саобраћају напада. У раду [86] се за описивање саобраћаја користе број пакета по току, количина података по току, као и проток по јединици времена. У раду [87] се врши испитивање информативног добитка различитих особина саобраћаја UDP и TCP саобраћаја, као што су број јединствених извора, број SYN пакета у TCP саобраћају, број TCP токова и број UDP токова. У раду [88] се врши моделовање различитих особина IP адреса у ову сврху, као што су проток података, одредишни и изворни портови, и типови протокола.

6.1. Моделовање интернет саобраћаја техникама нелинеарне регресије

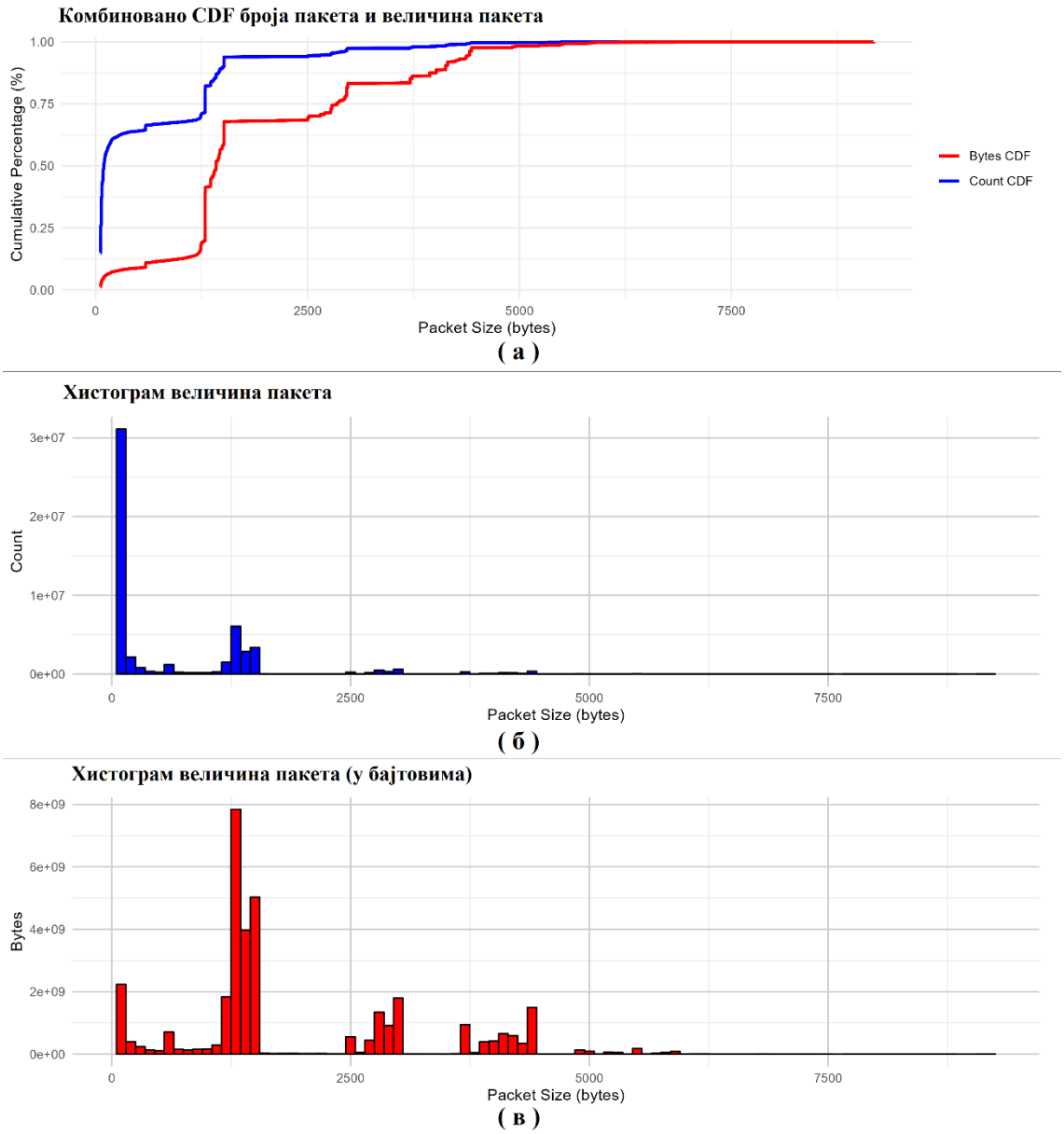
У овом поглављу дајемо студију случаја моделовања интернет саобраћаја помоћу НДКР за различите моделе расподела вероватноћа који у датом случају најбоље описују интернет саобраћај. У наставку приказујемо анализу прикладности осам модела кумулативне расподеле вероватноће за фитовање коју смо објавили у раду [89]. На основу литературе [81], [82], [90], одабране су следеће расподеле вероватноће за фитовање датог скупа података: Бета, Експоненцијална, Гама, Генерализована екстремна вредност (ГЕВ), Лог-

нормална, Накагамијева, Парето и Вејбулова. Из скупа пакета је на основу података о величинама Етернет пакета израчуната њихова дискретна густина расподеле вероватноће, а након тога су фитоване задате расподеле вероватноће. У сврху илустровања, као критеријум за одабир модела расподеле који најбоље одговара датим подацима искоришћено је неколико критеријума истовремено (AIC, BIC, RSS, RMSE, R-square, Adjusted-R-squared).

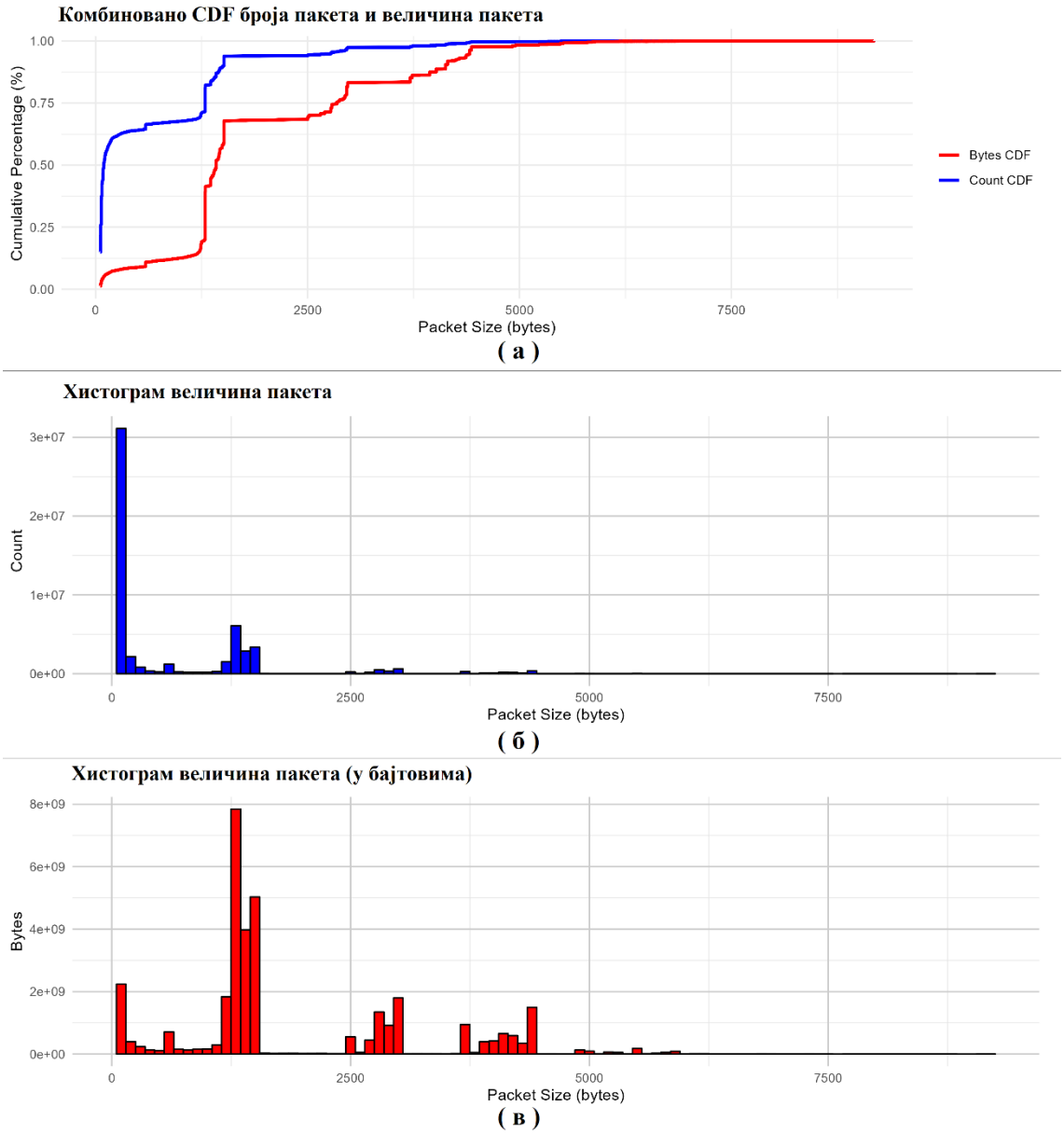
За потребе ове анализе је коришћена архива интернет саобраћаја саобраћаја MAWI радне групе [91] (енгл. *Measurement and Analysis on the WIDE Internet*), која представља заједнички пројекат јапанских компанија за истраживања интернет мрежа и академских институција. Скуп података који је коришћен у овој студији се састоји од интернет саобраћаја прикупљеног у оквиру пројекта пројекта "Дан у животу Интернета" [92] (енгл. *Day in the Life of the Internet*). Анализирана је датотека (енгл. *dump file*) 202304301400.pcap [93], која садржи интернет саобраћај прикупљен 30. априла 2023. године између 14.00 и 14:15 часова. Величина датотеке је 4GB и састоји се од 53.857.184 пакета.

Анализа овако великог скупа података може бити непрактична, те је први корак у анализи трансформација у облик који је погодан за анализу. Први корак предобраде јесте извршено филтрирање пакета Етернет протокола, при чему је уведено ограничење величине пакета у посматраном скупу података на максимално 1500 бајтова — што одговара величини максималне јединице за пренос (енгл. *Maximum Transmission Unit (MTU)*). MTU је највећа јединица података која се може пренети у једном Етернет оквиру мрежног слоја. Из прикупљених пакета су издвојене дужине свих Етернет оквира. Добијена датотека је веома велика (~200MB), али значајна количина података у њему је редундантна, с обзиром на то да има неколико десетина милиона редова, али само 1500 могућих дужина. Следећи корак обраде јесте сумирање добијених података као учесталости појављивања дужине пакетних података.

На



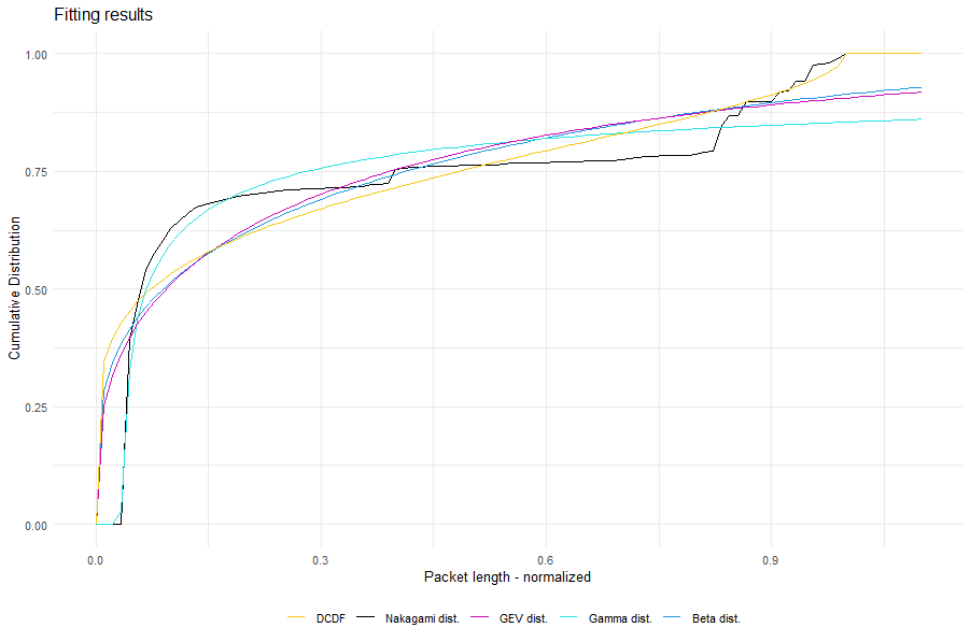
Слика 34 Комбинована CDF броја пакета и бајтова у анализи MAWI скупа податакаје приказана кумулативна расподела вероватноће дужине пакета и бајтова у оригиналном скупу података (а), са фокусом на пакете до 1500 бајтова (Етернет пакети), као и хистограмски приказ величина пакета (б) и количине бајтова (в) пренетих тим пакетима. Плавом бојом су дати подаци о броју пакета, а црвеном бојом подаци о броју пренетих бајтова.



Слика 34 Комбинована CDF броја пакета и бајтова у анализи MAWI скупа података

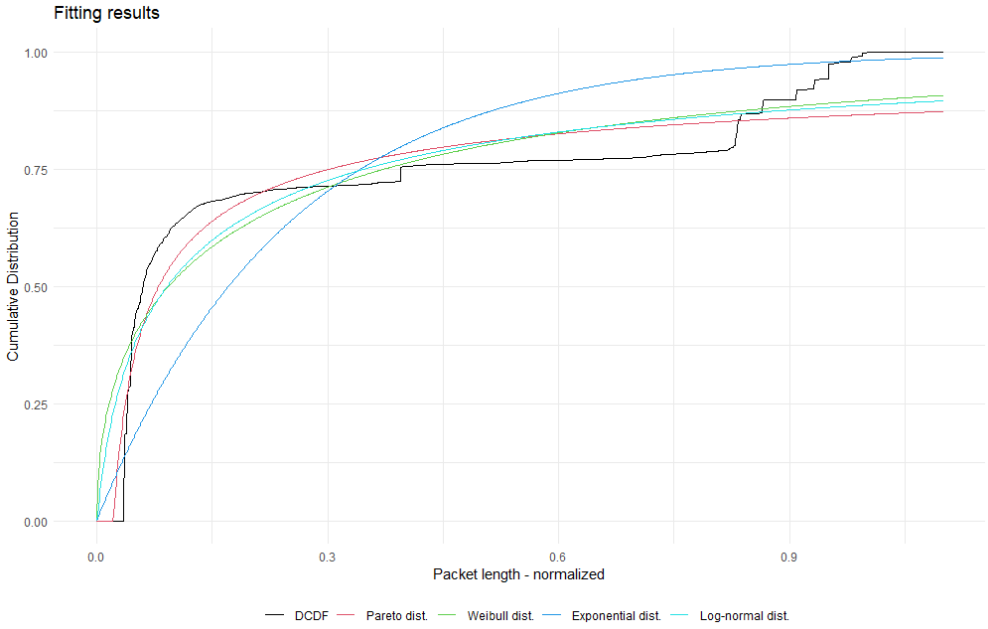
На основу изложеног се може уочити присуство веома великих и веома малих пакета, а такви пакети узрокују најстрмије делове кривих. Ово је у складу са Паретовим правилом (такође познатим и као „80—20 правило“) које каже да код многих процеса постоји појава да приближно 20% узрока носи око 80% (позитивних или негативних) последица Ово је типична особина интернет саобраћаја, и назива се Интернет микс (енгл. *Internet Mix* (IMIX)) [94], [95]. IMIX описује шаблоне стварног интернет саобраћаја и заснован је на статистичком

узорку са интернет рутера. IMIX профили постоје за IPv4, TCP, VPN (IPsec) и IPv6 саобраћај, с различитим величинама оквира заснованим на ограничењима MTU.



Слика 35 Фитовање Накагамијеве, GEV, Гама и Бета расподеле [89]

Да би се обезбедила боља прегледност и читљивост дијаграма, резултати уклапања анализе су представљени на две одвојене слике. Слика 35 приказује резултате фитовања ГЕВ, Накагами, Гама и Бета дистрибуција, док Слика 36 приказује фитовања Парето, Вејбулове, Експоненцијалне и Логнормалне расподеле. Ове две слике омогућавају визуелну процену поклапања наведених осам расподела са датим подацима.



Слика 36 Фитовање Парето, Вејбулове, Експоненцијалне и Лог-нормалне расподеле [89]

Поред визуелног приказа, у наставку су дати табеларни прикази резултата анализе. Табела 8 приказује процењене вредности параметара расподела вероватноћа. Табела 9 приказује вредности критеријума метода за одабир модела и тестова адекватности. Код прва четири критеријума ниже вредности указују на боље фитовања, док се код последња два на то указује већим вредностима. Из добијених резултата јесте јасно да су најбољи резултати добијени за GEV и Парето расподеле.

Табела 8 Процењене вредности параметара фитованих расподела вероватноће [89]

Расподела	**Параметар 1**	**Параметар 2**	**Параметар 3**
Парето	облик = 0.0215	скала = 0.5265	
Вејбулова	облик = 0.4997	скала 0.1927	
Експоненцијална	rate = 4.057		
Лог-нормална	ср. вред.= 0.023	стд. дев. = 1.966	
Накагамијева	облик = 0.2914	скала = 0.1352	
GEV	облик = 2.23463	скала = 0.03438	локација = 0.049
Гама	облик = 0.3319	скала = 1.0104	
Бета	облик 1 = 0.1941	облик 2 = 0.5430	

Табела 9 Критеријуми за одабир фитованих модела [89]

Расподела	AIC	BIC	RSS	RMSE	R^2	Adj. R^2
Парето	-242.7	-234.9	0.48	0.069	0.87	0.867
Вејбулова	-212.54	-204.73	0.65	0.081	0.824	0.82
Експоненцијална	-111.8	-106.6	1.83	0.135	0.51	0.5
Лог-нормална	-221.9	-214	0.59	0.077	0.84	0.836
Накагамијева	-203	-195	0.72	0.085	0.807	0.803
GEV	-250.7	-240.2	0.44	0.066	0.882	0.878
Гама	-206.4	-198.6	0.69	0.083	0.813	0.809
Beta	-205.9	-198.0	0.7	0.085	0.811	0.808

Добијени резултати су у складу са IMIX карактеристиком интернет саобраћаја. Као што је већ споменуто, Парето расподела и GEV расподела се могу користи за моделовање екстремних случајева, као што су највеће или најмање вредности у скупу података. Један од разлога зашто ове две расподела добро осликавају кумулативну расподелу вероватноће величина интернет пакета јесте њихова особина „тешког репа“ [96] (тј. реп расподеле вероватноће неке случајне величине одређен је вероватноћом да случајна величина узима вредности веће од неког задатог прага). Ово значи да постоји релативно велика вероватноћа да у скупу података постоји већи број малих и великих величина пакета, што се може сматрати екстремним догађајима.

6.2. Примена нелинеарне регресије за анализу великих скупова података у Oracle бази података

У овом поглављу дајемо студију случаја коришћења НДКР за моделовање великих скупова података похрањених у базама података. За потребе ове анализе извезли смо у базу података анотирани скуп података Bogazic¸i University DDoS dataset [97] (BOUN DDoS) генерисан на Универзитету Богазичи у Истамбулу. Наведени скуп података се састоји из два подскупа који

одговарају двама сценаријима напада трајања од по осам минута, у оба сценарија се користе насумично одабране IP адресе за извођење DDoS напада. У оба случаја након 80 секунди нормалног саобраћаја следи 20 секунди саобраћаја DDoS напада. Саобраћај напада је подељен на четири подгрупе, са 1000, 1500, 2000 и 2500 пакета у секунди. Пакети који су део DDoS напада се разликују од осталих пакета по одредишној IP адреси (она је код свих 10.50.199.86). Код напада преко TCP протокола, коришћен је одредишни порт 80.

Скупови података су изворно доступни у облику CSV датотеке [98], а подаци су груписани по следећим колонама:

- Време — почиње од нуле, инкремент је на нивоу једне микросекунде.
- Број фрејма — редни број пакета у скупу.
- Дужина фрејма — величина пакета у бајтовима.
- Изворна IP адреса пакета.
- Одредишна IP адреса пакета.
- Изворни порт — ово је одредишни TCP порт пакета, уколико се не ради о TCP пакету, поље је празно.
- Одредишни порт — TCP порт одредишта пакета, празно уколико се не ради о TCP пакету.
- SYN — уколико је TCP пакет и SYN флаг је постављен, тада има вредност 1, уколико флаг није постављен, има вредност 0, у остали случајевима је празно.
- ACK — уколико је TCP пакет и ACK флаг је постављен, тада има вредност 1, уколико флаг није постављен, има вредност 0, у остали случајевима је празно.
- RST — уколико је TCP пакет и RST флаг је постављен, тада има вредност 1, уколико флаг није постављен, има вредност 0, у остали случајевима је празно.
- TTL — време живота пакета.

- TCP протокол — уколико пакет припада транспортном слоју, има вредност „TCP“ или „UDP“, за друге протоколе има различите вредности.

BOUN DDOS скуп података о TCP саобраћају обухвата 9.335.605 редова, извезен је у OML4R Oracle 19c базу података са циљем симулације обраде великог обима података за демонстрирање капацитета и перформанси система у контексту анализе великих података. Имплементација OML4R омогућила је директну интеракцију са подацима унутар базе, што повећава ефикасност и минимизује време потребно за обраду комплексних упита, чиме је дата слика о могућностима примене нашег алгоритма у реалним оперативним условима.

Код препознавања DDoS напада, У раду [86] аутори су израчунали информациони добитак (енгл. *information gain* (IG)) колона BOUN DDOS скупа података тако што је вршена процена колико дате особине пружају информација о томе којој класи (напад или нормалан саобраћај) припада неки скуп пакета (Табела 10). IG је често коришћен критеријум за одабир атрибута и представља меру колико одређени атрибут доприноси укупној информацији у обучавајућем скупу података. Што је већа вредност информационог добитка, то та особина даје више информација о томе да ли подаци одговарају саобраћају напада или не.

Табела 10. Информациони добитак израчунат за различите векторе особина за BOUN DDOS скуп података [86]

Информациони добитак	**Назив карактеристике**
0.71	Јединствени извори
07	SYN пакети
0.7	Подаци по току
0.68	Просечни подаци
0.67	TCP ток
0.64	Пакети по току
0.67	UDP ток
0.63	Број свих хостова
0.55	Број токова

0.24	Број пакета
0.23	TCP пакети
0.22	ACK пакети
0.13	UDP пакети
0.03	Јединствене дестинације
0.01	RST пакети

Према истраживањима компаније Касперски[3], једне од водећих компанија у области сајбер безбедности, међу најзаступљеније типове напада у последњих неколико претходних година [99], [100], [101] спадају SYN преплављивање, UDP преплављивање и TCP преплављивање напади. SYN преплављивање и TCP преплављивање чине више од половине свих DDoS напада на интернету. Обе ове врсте напада користе ТЦП протокол, али постоје одређене разлике између њих. SYN преплављивање за напад користи TCP трофазно руковање, након пристиглог SYN захтева циљни сервер одговара SYN-ACK (синхронизација је потврђена) пакетом, обавештавајући учесника да је SYN примљен и да је спреман за комуникацију. Међутим одговор не долази, те сервер мора задржати ову полуотворену везу и доделити ресурсе за чекање изгубљеног ACK пакета, чиме се исцрпљују његови ресурси намењени новим везама. Насупрот томе, TCP преплављивање преоптерећује сервер великим бројем TCP пакета (али ови пакети могу бити било који део TCP везе), потенцијално укључујући потпуне или делимичне везе, тако засићујући капацитете сервера и мреже. Нападачи могу слати велику количину потпуних TCP захтева или насумичних TCP флагова како би преоптеретили сервер.

Управо с обзиром на распрострањеност ове две врсте напада, као и на њеихове велике вредности информационог добитка, одлучили смо да за потребе спроведене анализе искористимо "SYN пакети" и "TCP ток" карактеристике из Табела 10.

[3] https://www.kaspersky.com/

6.2.1. Oracle машинско учење и R

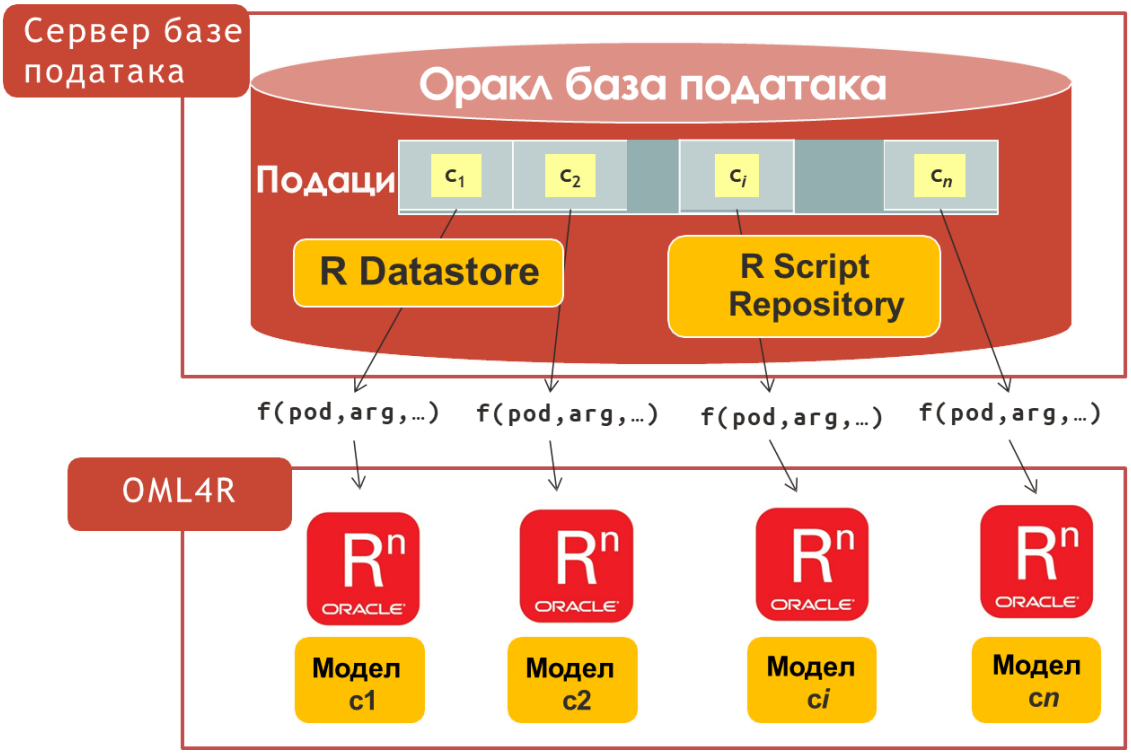
Програмски језик R има широку примену у области статистике и машинског учења. Подаци над којима се ради су често смештени у базе података, те је приликом рада потребно или програмски приступати подацима или их извозити у датотеке. Међутим, данас није реткост радити над скуповима података који су величине десетина терабајта, где постоје стотине милиона редова, при чему сваки од њих може имати на стотине променљивих. Традиционални приступ анализи оваквог скупа података би подразумевао сложену комбинацију R-а и SQL-а где би подаци би морали екстраховани из базе података помоћу SQL упита, а затим би морали бити увезени у R за даљу обраду и анализу. С обзиром на то да се ради са огромним количина података све то укључује значајне трошкове у смислу времена и ресурса. Поред тога, овај процес може укључивати додатне кораке, као што су чишћење и трансформација, који морају бити поновљени сваки пут када се подаци ажурирају или мењају. Када модел за анализу података развијен, може бити потребно и пренети резултате назад у базу података за даље коришћење и имплементацију.

Оракл машинско учење за R [102] (енгл. *Oracle Machine Learning for R* ((OML4R) решава изазове наведене изнад кроз интеграцију језика R са Оракл базом података. OML4R је моћна платформа за напредну аналитику која комбинује предности "екосистема" језика R и података којима се управља преко Оракл базе података[4]. Могућности језика R су проширене транспарентним приступом и директним управљањем базом података, као и применом алгоритама машинског учења директно у оквиру базе података.

Коришћењем OML4R могуће је делегирати захтевне рачунске задатке на базу података, чиме се елиминише потреба за писањем комплексних SQL упита, већ је могуће вршити екстраховање и трансформацију великих података

[4] Поред OML4R, постоје верзије и за SQL (OML4SQL) и пајтон (OML4Py).

коришћењем познате R синтаксе [103], као што је то показано на Слика 37. Изградња модела у OML4R OML4R значајно смањује потребу за преносом података, што је главно уско грло у традиционалним токовима аналитике.



Слика 37. Изградња модела у OML4R

Паралелни алгоритми машинског учења у бази су изложени кроз R API интерфејс, а могуће је складиштење корисничких R функција и других R објеката директно у бази података. Извршавање се врши у R окружењима покренутим и управљаним у оквиру базе користећи R, SQL, а у аутономним базама података и REST интерфејс. Овим се обезбеђује да се решења могу лако имплементирати, користећи инхерентне безбедносне функције Оракл базе и избегавајући ограничења обичних датотека.

За потребе извођења анализе представљене у овој докторској дисертацији, на виртуелној машини са оперативним системом Centos Linux 8 инсталирали смо Oracle 19c базу података, након чега смо инсталирали Oracle R Enterprise

Server, као и Oracle R Enterprise Client са одговарајућим "support" пакетима. У оба случаја је инсталирана верзија језика R-4.0.5. Главни циљ је био омогућити непосредан рад са R језиком директно на базама података, чиме се елиминише потреба за екстракцијом података и њиховим накнадним учитавањем из датотеке, што доприноси ефикасности и интегритету аналитичког процеса.

```
> # OML4R - ore.frame objekat
> str(BOUNTCPdf)
'data.frame': 9335605 obs. of  12 variables:
Formal class 'ore.frame' [package "OREbase"] with 14 slots
  ..@ .Data    : list()
  ..@ dataQry  : Named chr "( select /*+ no_merge(t) */
\"FRAME_NUMBER\" NAME001, \"TIME\" VAL001,\"FRAME_NUMBER\"
VAL002,\"FRAME_LENGTH\" "| __truncated__
  .. ..- attr(*, "names")= chr "55_2"
  ..@ dataObj  : chr "55_2"
  ..@ desc     :'data.frame': 12 obs. of  2 variables:
  .. ..$ name  : chr  "TIME" "FRAME_NUM" "FRAME_LENGTH" "SOURCE_IP"...
  .. ..$ Sclass: chr  "numeric" "numeric" "numeric" "factor" ...
  ..@ sqlName  : chr "\"FRAME_NUMBER\""
  ..@ sqlValue : chr  "\"TIME\"" "\"FRAME_NUMBER\"" "\"FRAME_LENGTH\""
"\"SOURCE_IP\"" ...
  ..@ sqlTable : chr "\"OML_USER\".\"BOUNTCP\""
  ..@ sqlExtra :List of 2
  .. ..$ rIdTab: chr ""
  .. ..$ rIdSeq: NULL
```

Слика 38. Коришћење прокси објеката у OML4R

За смештање података у оквиру OML4R се користе прокси објекти Објекат ore.frame представља релациони упит у оквиру Oracle 19c базе података, ово је еквивалент data.frame у Oracle R Enterprise. Обично ore.frame објекти представљају прокси за табеле у бази. Помоћу њих се могу додавати нове колоне или вршити друге промене над подацима. Свака тако начињена промена не утиче на основну табелу. Функција слоја транспарентности генерише SQL упит који ради над подацима изворне табеле, али се сама табела не мења.

У OML4R све R наредбе се у позадини преводе у SQL, што омогућава ефикасну обраду и анализу великих скупова података директно у бази података. На Слика 38. Коришћење прокси објеката у OML4Rје дат приказ једног ore.frame објекта. У односу на стандардни data.frame објекат, уочавају се следеће разлике:

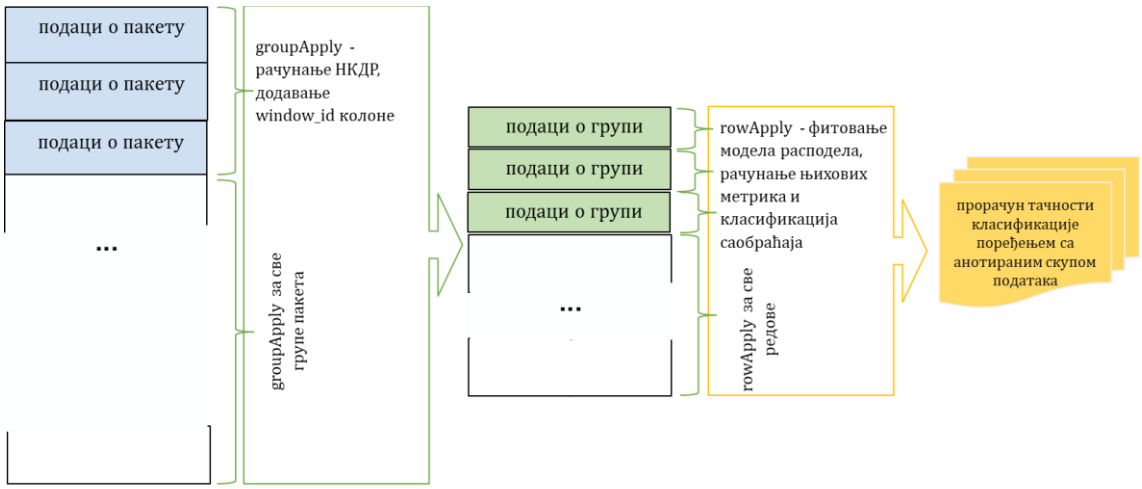
- Поље `dataQry` садржи SQL упит који се користи за дефинисање података који се добијају из базе. Упит укључује одабир специфичних колона из табеле.
- Поље `desc` описује мапирање између имена променљивих и њихових SQL типова података, што осигурава правилно тумачење података у R окружењу.
- Поља `sqlName`, `sqlValue`, и `sqlTable` користе се за чување метаподатака о SQL операцијама, указујући на идентификаторе и колоне којима се манипулише, као и на табелу из које се подаци добијају.

R enterprise омогућава директну примену многих алгоритама машинског учења, укључујући линеарне моделе, неуронске мреже и слично, али још увек нема интегрисану подршку за директно извршавање нелинеарних модела. Услед тога смо користили уграђено извршавање R скрипти (енгл. *Embedded R Execution*) које омогућава да се скрипте чувају и покрећу из базе података. Основна предност овог приступа над класичним извршавањем R скрипти јесте то што се израчунавање може вршити у бази на серверу, чиме се губи потреба за преносом великих количина података, а истовремено се за извршавање обраде може користити инфраструктура сервера.

Најважнији функционалности које омогућује Embedded R Execution API су:

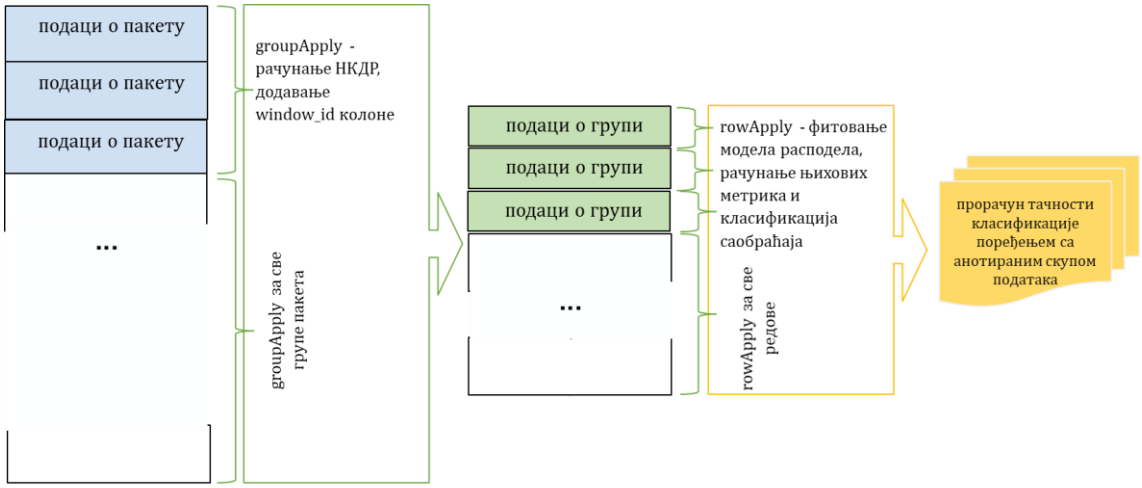
- ore.tableApply - смешта све редове у ore.frame објекат и извршава R функцију над њима;
- ore.groupApply - извршава R функцију над подацима груписаним по некој колони (ово може да се паралелизује на серверу);

- ore.rowApply - извршава R функцију над назначеним бројем редова (и ово може да се паралелизује на серверу)
- ore.rowApply - извршава функцију R прослеђујући одређени број редова улазног ore.frame објекта.
- ore.doEval - извршава R функцију без аутоматског преноса података.



Слика 39 Кораци извршене анализе у оквиру OML4R

На



Слика 39 Кораци извршене анализе у оквиру OML4Rје дат приказ одабраног поступка за процес анализе посматраног скупа података. Овај процес обухвата следеће фазе: из табеле која садржи пакете TCP саобраћаја се коришћењем функције groupApply функције врши груписање пакета по назначеном критеријуму. За параметар груписања узета је дужина

интервала који се анализира, на пример, груписање редова у интервале од по једне секунде. За сваку тако дефинисану групу, рачуна се нормализована дискретна кумулативна расподела вероватноће (псеудокод је дат на

```
Input: dataset paketa
// Faza pripreme:
    Filtriraj pakete TCP protokola sa postavljenim SYN flagom
    for i = 1 to broj_prozora do
       prozor = Time / veličina_bina
    end for
    Grupiši podatke po prozorima
// Faza izračunavanja
    Za svaku grupu:
        Izračunaj brojSYN u svakoj grupi
    // Normalizovanje podataka
    Pronađi min i max broj SYN paketa
    Za svaki unos u skupu:
        binovi = (brojSYN - min) / (max - min)
    Sortiraj podatke po binovima i izračunaj NDKR u njima
Output: izračunata NDKR
```

Слика 40. Псеудокод рачунања НДКР над скупом пакета.) Резултат ове операције је нови ore.frame објекат који за сваки ред садржи идентификатор временског прозора (window_id) и вредности НДКР (које представљају x,y координате). У следећем кораку анализе се над свим редовима новодобијеног објекта примењује функција rowApply, где се за сваки ред врши фитовање одабраног скупа модела расподела (псеудокод ове функције дат је на Слика 41,) ослањајући се на вредности метрика попут AIC, BIC, RMSE и других, за одабир најбољег модела.

```
Input: dataset paketa
// Faza pripreme:
    Filtriraj pakete TCP protokola sa postavljenim SYN flagom
    for i = 1 to broj_prozora do
       prozor = Time / veličina_bina
    end for
    Grupiši podatke po prozorima
// Faza izračunavanja
    Za svaku grupu:
        Izračunaj brojSYN u svakoj grupi
    // Normalizovanje podataka
    Pronađi min i max broj SYN paketa
    Za svaki unos u skupu:
        binovi = (brojSYN - min) / (max - min)
    Sortiraj podatke po binovima i izračunaj NDKR u njima
Output: izračunata NDKR
```

Слика 40. Псеудокод рачунања НДКР над скупом пакета

```
Input: NDKR paketa
// Faza pripreme:
    Učitaj potrebne biblioteke i kreiraj modele raspodela verovatnoće
// Faza prepoznavanja:
    for i = 1 to broj_modela_raspodela do
        Fituj model model[i] sa NDKR
        if fitovanje konvergira then
            if parametri su u okviru opsega then
                Dodaj model rezultatima
            end if
        end if
    end for
    for model = 1 to broj_fitovanih_modela do
        Izračunaj vrednosti metrika modela
    end for
Output: lista modela
```

Слика 41. Псеудокод фитовања модела

На Слици 42. је дат програмски код који се покреће у оквиру OML4R за примену корака описане анализе. Изложени кораци анализе су истоветни код моделовања других карактеристика посматраног скупа података.

```
# Dodajemo polje koje oznacava vremenski prozor kojem paketi pripadaju
BOUNTCP$window_id <- as.integer(BOUNTCP$TIME / 1)

# Pakete grupisemo po vremenskim prozorima i pozivamo funkciju koja
# racuna NDKR svakog pojedinacnog opsega paketa
rezultat <- ore.groupApply( BOUNTCP,
                            INDEX = TCP$window_id,
                            FUN.NAME = "izracunaj_syn_cdf",
                            window_size = .1,
                            ore.connect=TRUE)

# Prosledjujemo rezultatni skup u Orakl bazu podataka kao  ore.frame
ore.push(rezultat)

# Nad svim redovima rezultata vrsimo fitovanje modela NR
r1 <- ore.rowApply( rezultat,
                    FUN.NAME = "obrada_redova")
```

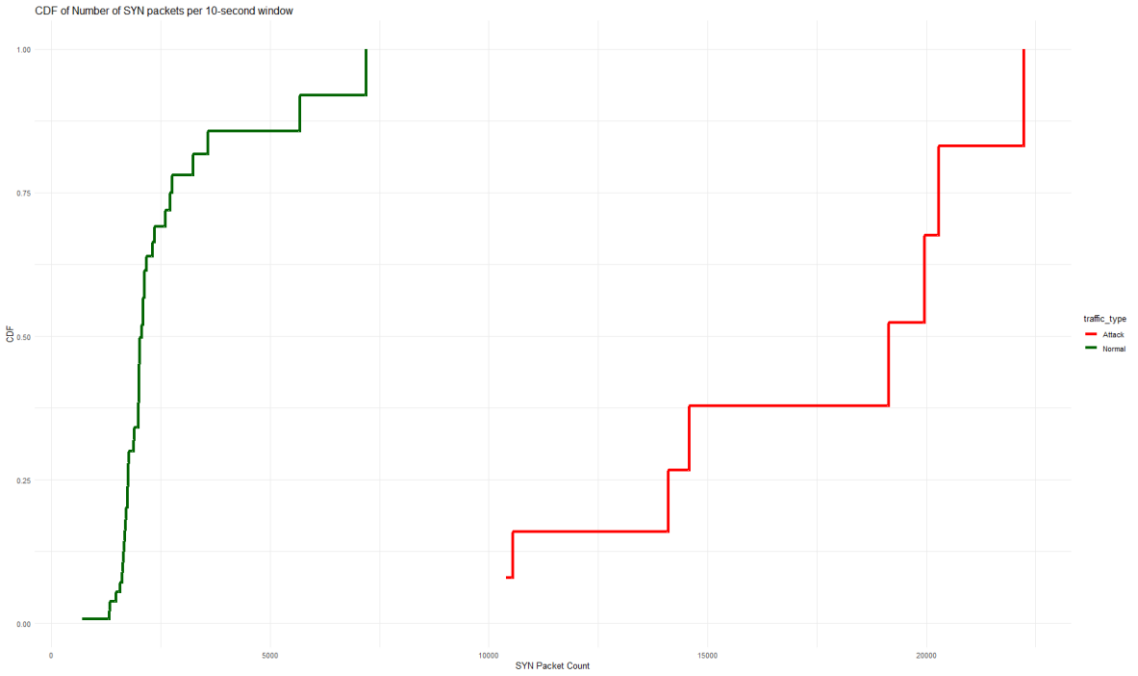
Слика 42. Програмски код за обраду података у OML4R

6.3. Препознавање DdoS напада препознавањем расподеле броја TCP пакета са ознаком SYN у одређеном временском интервалу

У овом одељку су представљени резултати коришћења нелинеарне регресије за анализу SYN TCP пакета у оквиру DDoS скупа података, са циљем разликовања нормалног и саобраћаја DDoS напада. С обзиром на висок ниво информационог добитка овог параметра, претпоставка је да учесталост и расподела броја SYN TCP пакета, праћених током различитих временских интервала, може статистички идентификовати и категорисати интернет саобраћај.

Користећи Oracle Machine Learning for R (OML4R), ова студија моделује интернет саобраћај фитујући број SYN пакета са различитим расподелама

вероватноће. Анализа истражује како промена дужине интервала утиче на прецизност и поузданост модела. На пример, Слика 43 илуструје расподеле густине SYN пакета за класе саобраћаја, јасно оцртавајући разлике између две врсте саобраћаја, чиме се валидира ефективност одабраног приступа моделовања.



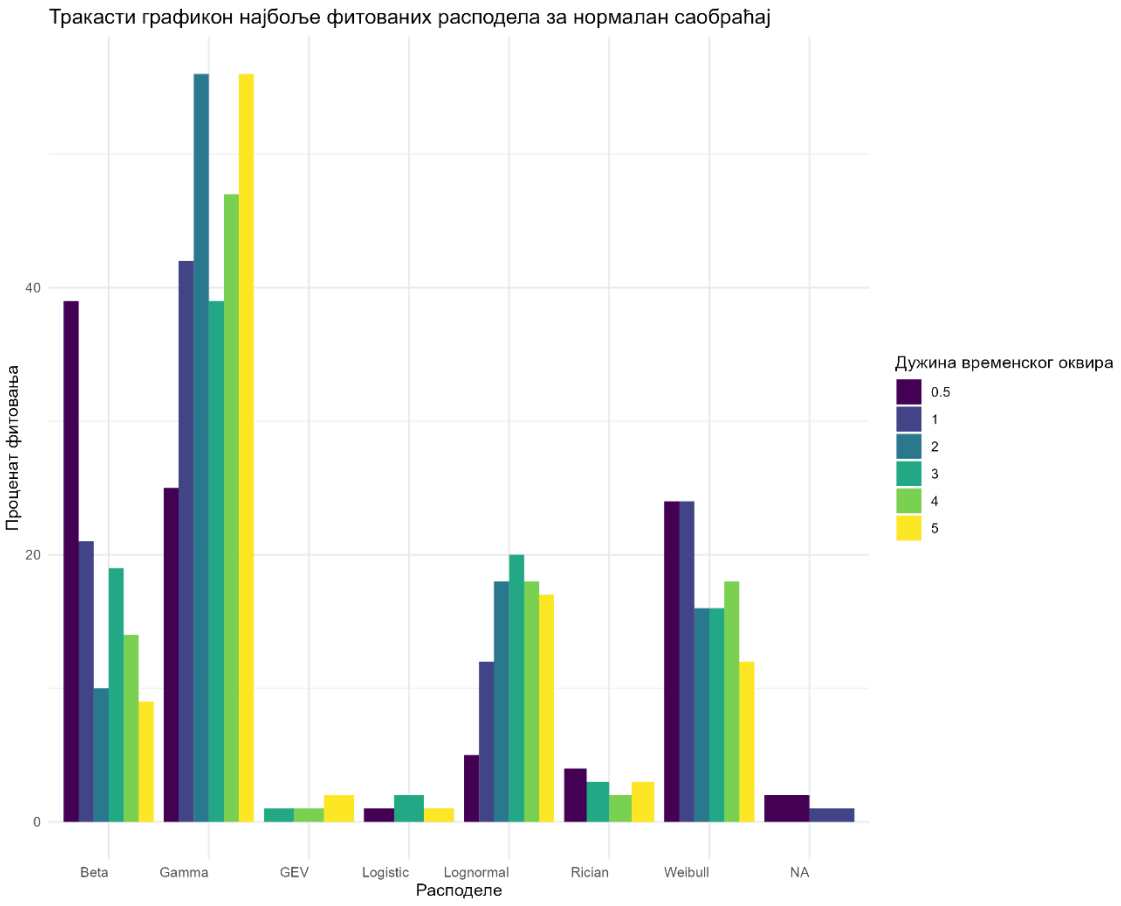
Слика 43. Густина расподеле за SYN пакете нормалног саобраћаја и саобраћаја DDoS напада

Код препознавања DDoS напада, рана детекција је од кључног значаја како би систем могао благовремено имплементирати механизме одбране. Стога смо спровели експерименте са циљем да утврдимо како перформансе модела препознавања базираних на расподели вероватноће варирају у зависности од дужине временских оквира над којима се посматра саобраћај. Прецизније, жеља је идентификовати обрасце који би омогућили поуздану класификацију саобраћаја као нормалног или нападачког након што се акумулира минимална количина података, односно за што краће временске оквире. Ово

је важно јер би детекција у раној фази напада омогућила брже реаговање система.

У том циљу, тестирали смо шест различитих дужина временских оквира: 0,5, 1, 2, 3, 4 и 5 секунди. Анализирајући перформансе модела за сваки од ових интервала, могли бисмо проценити компромис између брзине детекције и пероформанси. Краћи временски оквири би обезбедили бржу реакцију, али потенцијално на рачун смањене тачности услед недостатка података. Са друге стране, дужи интервали би акумулирали више информација, што би могло побољшати прецизност али успорити детекцију.

Овај приступ нам омогућава да нађемо компромис између времена потребног за детекцију и поузданости класификације, пружајући увид у оптималне параметре за дизајн ефикасног DDoS детекционог система који балансира ове две кључне компоненте. Резултати су представљени на одвојеним графиконима за нормални и нападачки саобраћај. На Слика 44. су дати резултати спроведене анализе за нормалан интернет саобраћај. Може се уочити да се на кратким временским оквирима, од 1 од 2 секунде, Бета расподела показује најбоље резултате фитовања и за нормалан и за саобраћај напада. Ово указује на то да Бета расподела ефикасно моделује брзе промене у количини SYN пакета који су типични за кратке временске периоде. Међутим, како се дужина временског интервала повећава, уочава се смањење успешности фитовања са Бета расподелом, што је праћено јасним разликовањем узорака саобраћаја.

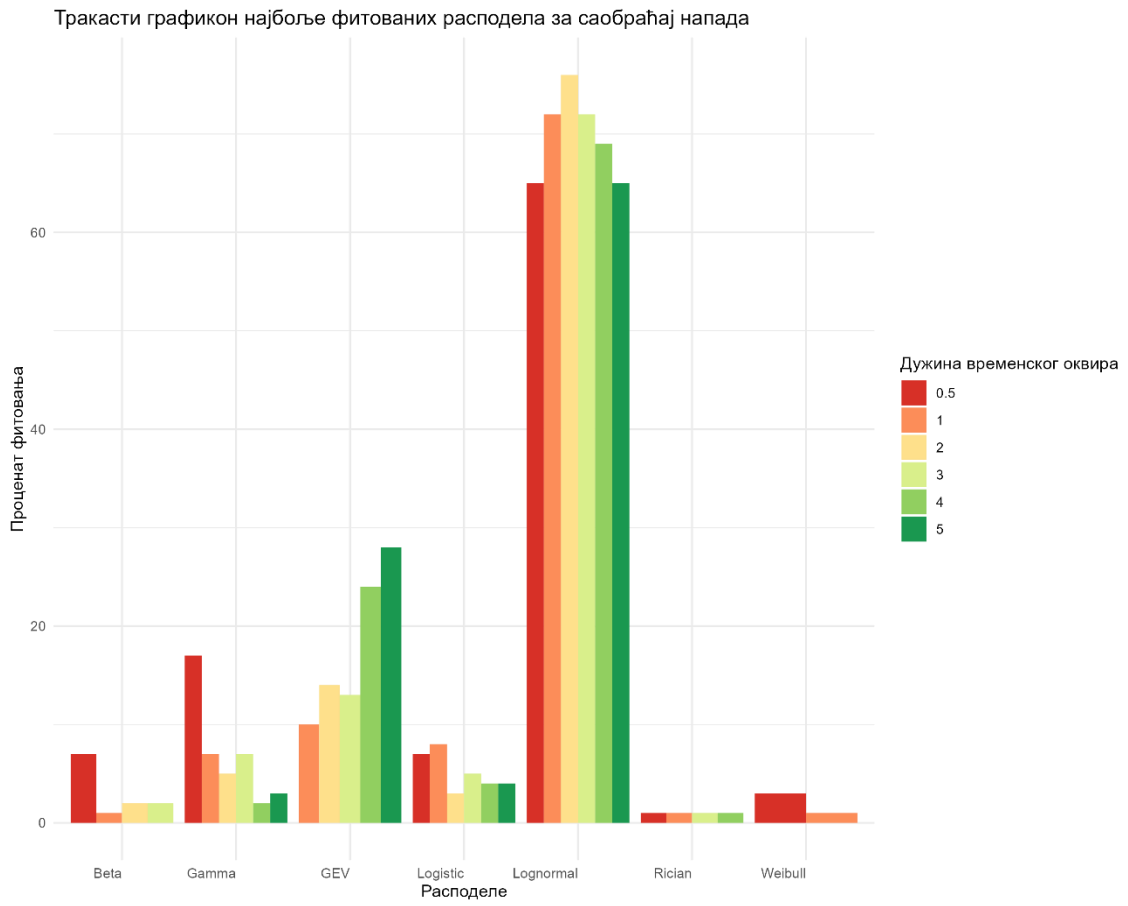


Слика 44 Графикон броја фитовања расподела по дужини временског оквира за нормалан саобраћај за SYN пакете

За нормални саобраћај, Гама расподела показује најбоље укупне резултате, чинећи око 40% свих фитовања. Ова расподела је посебно ефикасна за дуготрајне интервале, при чему непрекидно моделује уобичајене флуктуације у обиму саобраћаја. У краћим временским интервалима нормалног саобраћаја, иако преовладава Бета расподела, примећује се значајан пораст у броју успешних фитовања логнормалне расподеле како временски интервали постају дужи. Логнормална расподела боље описује дуготрајне и полако променљиве обрасце саобраћаја који су уобичајени за нормалне операције.

Осим ових, Вејбулова расподела показује константно добре резултате, чинећи скоро 20% свих успешних фитовања. Ова расподела је посебно корисна због своје флексибилности у моделовању различитих врста података, укључујући

и оне са снажно израженим реповима расподеле, што је типично за аномалне појаве у саобраћају.



Слика 45 Графикон броја фитовања расподела по дужини временског оквира за саобраћај напада за SYN пакете

На Слици 45 приказани су резултати анализе саобраћаја напада, где се издваја Логнормална расподела као најефикаснији модел у скоро 70% случајева анализе. Њена превласт у моделовању не варира у зависности од дужине временског интервала, што указује на њену стабилност и релевантност у контексту променљивих услова напада.

Занимљиво је да се код саобраћаја напада на краћим временским интервалима (0,5s) долази до израженог повећања у броју фитовања Бета и Гама расподела. Овај феномен се може објаснити краткоћом временског интервала који може

довести до преклапања карактеристика две различите врсте саобраћаја, чиме се ствара погрешна слика о преовлађујућем типу саобраћаја.

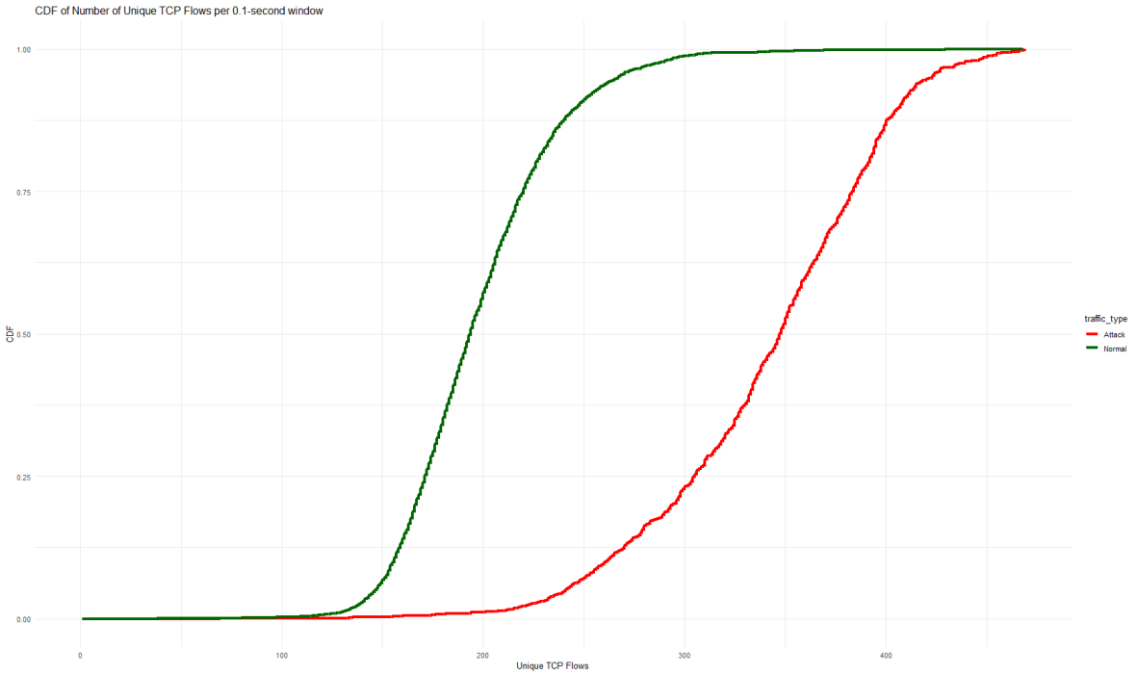
Осим тога, вредно је поменути и GEV (Generalized Extreme Value) расподелу која показује занимљив тренд у зависности од дужине временског интервала. Док за интервале до 3 секунде GEV расподела учествује у око 10% успешних фитовања, овај проценат расте са повећањем временског интервала, достижући до 30% за временске интервале дужине 5 секунди. Овакав раст подржава употребу GEV расподеле за дуготрајније анализе саобраћаја, нудећи поуздане моделе за идентификацију и карактеризацију екстремних варијација у обиму саобраћаја које су типичне за нападе.

6.3.1. Препознавање DdoS напада препознавањем расподеле броја TCP токова у одређеном временском интервалу

У овом одељку представљени су резултати примене нелинеарне регресије на анализу броја TCP токова у оквиру DDoS скупа података, ради разликовања између нормалног и саобраћаја напада. Ова анализа посматра број TCP токова током разних временских интервала, претпостављајући да ова метрика може обезбедити значајне информације за статистичко категорисање саобраћаја. Подаци о TCP токовима, као одраз активности на мрежи, могу указивати на нормалне или неприродно интензивне обрасце који су карактеристични за DDoS нападе.

Као што је речено у претходном поглављу, перформансе модела препознавања DDoS напада базираних на расподели вероватноће саобраћаја у великој мери зависе од дужине временских оквира у којима се посматрају пренети пакети. И у овом случају смо тестирали шест различитих дужина интервала: 0,5, 1, 2, 3, 4 и 5 секунди како бисмо утврдили оптималну дужину временског оквира у ком можемо добити добре перофрмансе.

Слика 46 показује расподеле густине броја TCP токова за обе класе саобраћаја, наглашавајући разлике у учесталости и облику расподела између нормалног и нападачког саобраћаја. Претпоставка је да ће се моделовањем интернет саобраћа моћи адекватно раздвојити два типа саобраћаја на основу анализираних параметара.

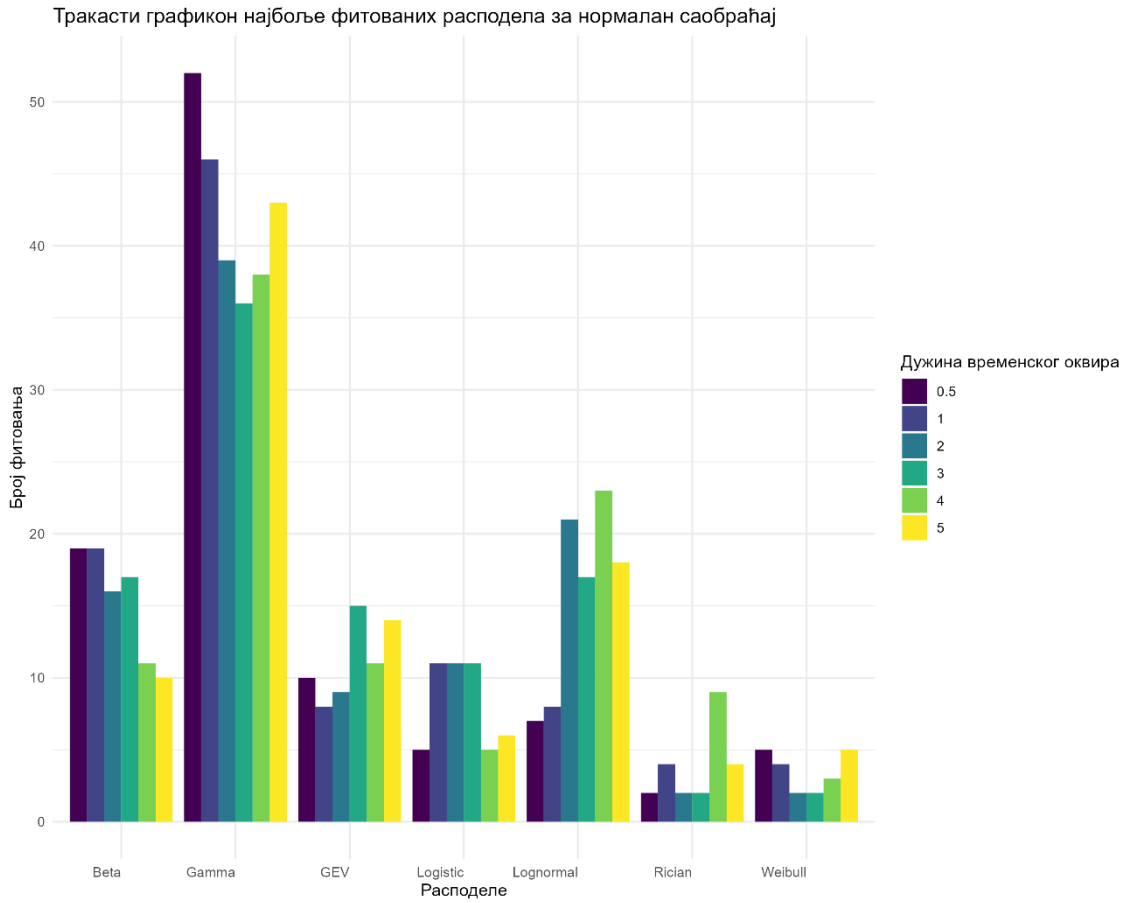


Слика 46. Густина расподеле за број TCP токова у оквиру нормалног саобраћаја и саобраћаја DDoS напада

На Слика 47 су представљени резултати анализе за нормалан интернет саобраћај. Овде се уочавају донекле другачији трендови у односу на резултате код SYN преплављивања, на краћим временским оквирима, од 0.5 до 1 секунде, Гама расподела показује најбоље резултате фитовања и за нормалан и за саобраћај напада, са преко 50% свих фитовања. С друге стране, Бета расподела показује слабљење броја фитовања са повећањем временског интервала, али и даље остварује скоро 20% успешних фитовања на кратким дужинама.

Такође се примећује значајно повећање броја успешних фитовања логнормалне расподеле како временски интервали постају дужи.

Логнормална расподела наставља боље да описује дуготрајне и полако променљиве обрасце саобраћаја, који су уобичајени за нормалне операције. Поред тога, примећује се повећање броја фитовања са GEV и логистичком расподелом, док се Вејбулова расподела показује са смањеним бројем успешних фитовања, што указује на промену у ефикасности ове расподеле у моделовању флуктуација у саобраћају.

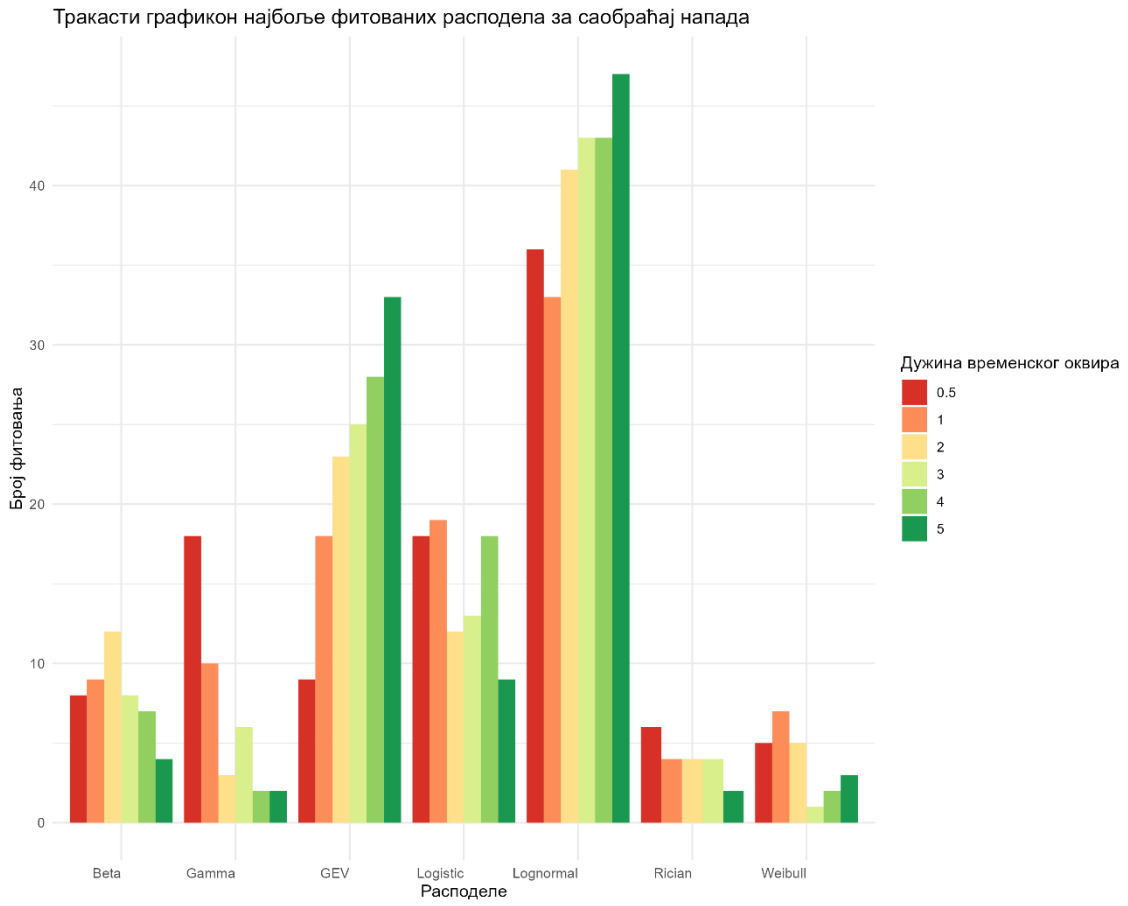


Слика 47. Графикон броја фитовања расподела по дужини временског оквира за нормалан саобраћај за TCP токове

На Слика 48 приказани су аналитички резултати специфични за саобраћај напада, где логнормална расподела и даље има највећи број успешних фитовања, мада је њен удео опао на око 50%, а чак и на 35% за краће временске интервале (од 0,5 до 1 секунде). Овај пад у удеу фитовања

логнормалне расподеле углавном иде на рачун GEV расподеле, која на краћим временским интервалима учествује са мање од 20% успешних фитовања, док на дужим интервалима њен удео расте до скоро 35%.

Гама расподела на нижим временским интервалима учествује са око 20% успешних фитовања, али тај проценат постаје занемарљив на дужим временским интервалима. Занимљиво је и велико повећање удела логистичке расподеле до скоро 20%, као и Бета расподеле, што указује на променљиву динамику у моделовању саобраћаја напада у различитим временским оквирима.



Слика 48. Графикон броја фитовања расподела по дужини временског оквира за саобраћај напада за TCP токове

На основу пружених текстова и визуелних података са графикона, можемо извести неколико закључака о корисности анализе расподела за детекцију DDoS напада и управљање интернет саобраћајем:

1. Логнормална расподела показује конзистентно високе перформансе у моделирању саобраћаја напада на различитим временским интервалима, иако њен удео у фитовањима опада са смањењем временског интервала. Ово указује на њену способност да стабилно моделира променљиве услове који карактеришу DDoS нападе.

2. Гама и GEV расподеле показују различите трендове у зависности од дужине временског интервала. Док Гама расподела има већи успех на краћим интервалима, GEV расподела показује повећање успешности са продужавањем временског интервала. Ова динамика може бити корисна за диференцијацију између нормалног и аномалног саобраћаја, посебно у контексту дуготрајнијих анализа.

3. Упоређивање модела расподела за нормалан и нападни саобраћај указује на то да различите расподеле имају различите предности у зависности од врсте саобраћаја. На пример, логистичка и Бета расподела показују значајан пораст у броју успешних фитовања у случају нападног саобраћаја, указујући на могућност њиховог коришћења као индикатора за специфичне типове напада.

Висок степен фитовања одређених расподела, као што су логнормална и GEV у случају дужих интервала, може помоћи у идентификацији и карактеризацији екстремних варијација у саобраћају које су типичне за DDoS нападе. Ова спознаја може бити интегрисана у алгоритме за детекцију који могу прецизније да предвиде и реагују на такве аномалије.

Основна предност нашег приступа лежи у способности да моделира саобраћај без потребе за априори знањем или посебним подешавањима. Ово омогућава анализу саобраћаја на основу стварних података, без унапред дефинисаних ограничења или претпоставки о природи саобраћаја. Као резултат, метода може адаптивно и динамично да одговори на променљиве услове саобраћаја и аномалије, што је кључно за ефикасно управљање и заштиту мрежних

ресурса у реалном времену. Ова карактеристика чини наш приступ изузетно вредним у контекстима где је брза и прецизна реакција на непредвиђене или малициозне промене у саобраћају од суштинске важности.

Иако се подаци из анализе могу користити за унапређење метода откривања DDoS напада, важно је истаћи да треба обратити пажњу на контекстуалне факторе и потенцијалне промене у обрасцима саобраћаја које могу утицати на избор и ефикасност различитих модела расподела.

7. СОФТВЕР ЗА РАД СА СИГНАЛИМА, ПРЕПОЗНАВАЊЕ СИГНАЛА И АНАЛИЗУ РЕЗУЛТАТА

Развијени софтвер[5][6] је представљен у раду [83] и пружа свеобухватно решење за фитовање расподела вероватноће, при чему допушта корисницима да лако идентификују одговарајућу расподелу вероватноће и процене њене параметре. Овај алат је развијен за Веб окружење са циљем да буде широко примењив и лак за употребу, те може послужити корисницима из различитих области и са различитим нивоима знања.

При развоју алата коришћен је *Shiny* [104] пакет који пружа Веб оквир за израду интерактивних Веб апликација користећи R. Креиран је од стране компаније Posit (некада зване RStudio) [105] и користи се за развој Веб аплета који имају интегрисану подршку за R. Овај софтверски оквир омогућује корисницима да креирају интерактивне Веб апликације и олакшано повезивање између R кода и корисничког интерфејса уз врло мало коришћење (па и знање) HTML, CSS, или JavaScript језика. Апликације написане на овај начин могу бити хостоване локално или на серверу, омогућавајући реално време интеракције са корисником, анализе података, визуализације, и сличног.

Најважнија промена у односу на алат представљен у раду [21] јесте то што је модул за генерисање сигнала сада имплементиран у језику R, те је успостављена униформност и искључена потреба за коришћењем MATLAB окружења. Овим се избегава непотребно мешање различитих програмских језика и софтверских платформи, тј. избегавају се и проблеми

[5] Апликација је доступна на хипер-вези: https://probdistid.shinyapps.io/ProbDistID/

[6] Изворни код је доступан на адреси: https://github.com/dragisa-miljkovic/probdistid

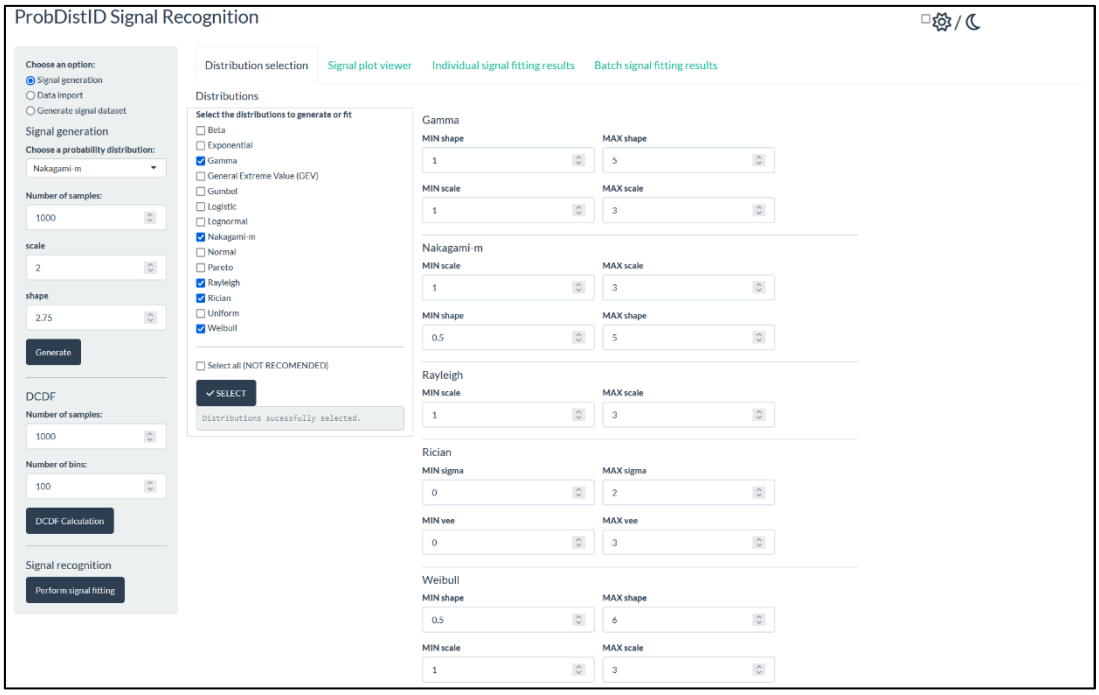
компатабилности и умањује комплексност алата, чиме се омогућава кохерентнији и ефикаснији рад на његовом даљем развијању.

С обзиром на то да није увек лако и практично имати реалне податке из стварног окружења, често постоји потреба за коришћењем синтетичких (вештачки генерисаних) података. Развијени софтверски алат омогућава корисницима да генеришу насумичне податке у сврху тестирања. Постоје два начина генерисања сигнала, појединачно - обично у ситуацијама када се алат користи у сврхе експериментисања, или серијски - у ситуацијама када корисник жели да изврши фитовање над великим бројем улазних сигнала.

Интуитивни графички кориснички интерфејс омогућава корисницима визуелизацију процеса фитовања и праћење корака обраде. Резултати фитовања се приказују у табеларном облику, где се исписују поруке о успешности фитовања, вредности процењених параметара, као и вредности AIC и BIC метода за одабир модела и MAE, RMSE, R-squared, Adjusted R-squared тестова за испитивање адекватности модела. Корисник може на основу вредности метрика вршити поређење резултата фитовања расподела вероватноће и на основу њих доносити одлуке о расподели која најбоље одговара улазним подацима.

7.1. Интерфејс алата и функционалности

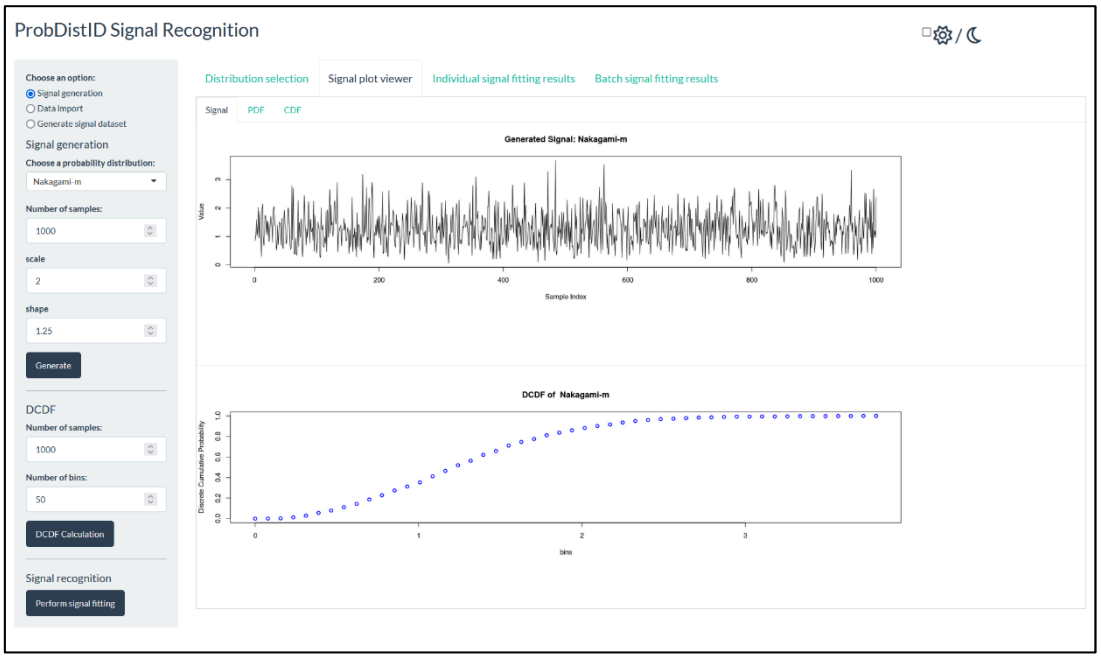
Графички кориснички интерфејс алата је замишљен тако да буде интуитиван и лак за коришћење. Сам прозор апликације је подељен на помоћни панел, са леве стране, и на главни прозор. У помоћном панелу се налазе опције везане за увоз/извоз података, генерисање сигнала, као и опција везаних за НДКР и препознавање сигнала. Главни прозор апликације се састоји из четири под-панела намењених за различите фазе/делове процеса препознавања.



Слика 49. Панел за одабир модела расподела вероватноће и подешавање њихових параметара

На Слика 49. Панел за одабир модела расподела вероватноће и подешавање њихових параметараје приказан панел служи за одабир расподела вероватноће и унос горњих и доњих граница њихових параметара. Алат тренутно подржава 14 расподела вероватноће: Бета, Вејбулова, Гама, Гумбелова, Експоненцијална, Логистичка, Логнормална, Нормална, Накагамијева, Општа расподела есктремних вредности (енгл. *General Extreme Value* (GEV)), Рејлијева, Рајсова, Пуасонова и Униформна расподела вероватноће. Ове расподеле су одабране јер заступљене у различитим дисциплинама, те имају широк опсег примене. Имплементација је извршена коришћењем R пакета pracma [106], VGAM [107] extraDistr [108]. У даљем развоју апликације, како се рад на унапређивању алата буде настављао, планирано је укључивање додатних расподела вероватноће.

На Слика 50 је приказан панел за плотовање сигнала на три начина: у виду одмерака, као PDF, или као CDF. Након подешавања НДКР, на истом панелу се приказује и њена кумулативна расподела вероватноће.



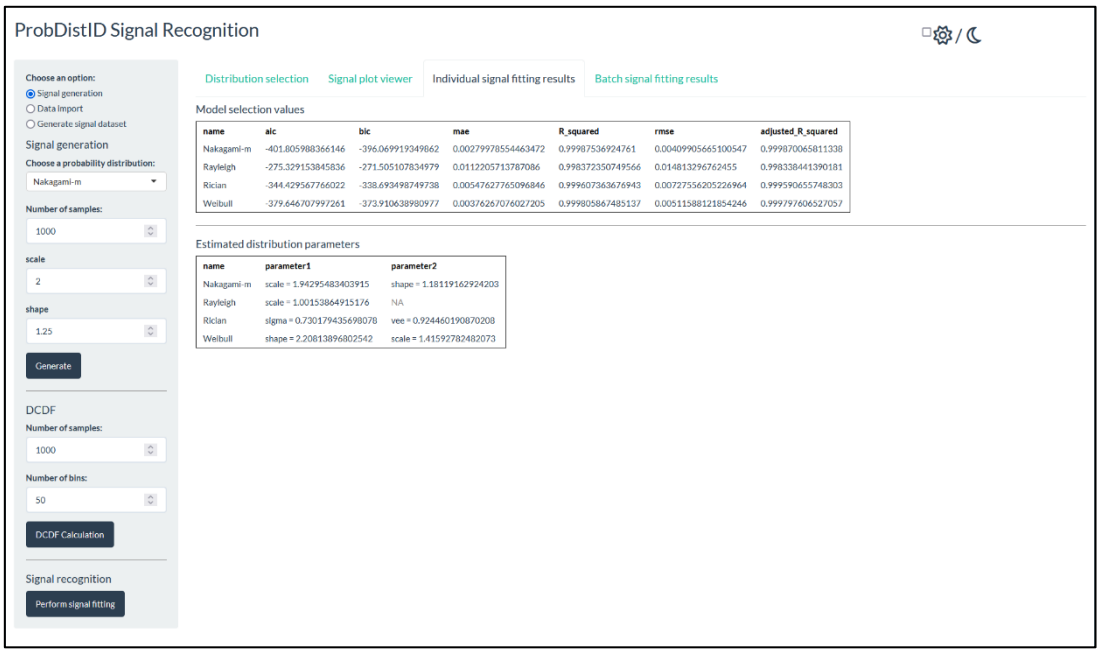
Слика 50. Панел за плотовање сигнала

У помоћном панелу корисник може генерисати или генерисати индивидуалне сигнале расподела са задатим вредностима параметара и бројем одмерака (и такви сигнали се могу прегледати и фитовати), или може генерисати датасет сигнала, при чему се назначује број одмерака, корак инкрементирања параметара сигнала (чији опсег је одабран раније у главном прозору) и број сигнала за сваку комбинацију параметара. Датасет је могуће преузети као датотеку у једном од четири формата: RData, CSV, JSON и XML.

На Слика 51 је дат приказ панела за приказивање табеларно приказаних резултата фитовања генерисаног сигнала или увезеног скупа сигнала. Поред овог панела се налази панел који служи за приказивање резултата групне обраде сигнала.

Апликација омогућава и увоз произвољног броја сигнала за фитовање. Поред наведених формата датотека које апликација подржава, за увозне датотеке је додата и могућност увода сигнала у текстуалном облику. Текстуелни формат

дозвољава унос сигнала без назначавања расподеле вероватноће и њених параметара, док се у осталим форматима подразумева да ти подаци постоје.



Слика 51. Панел за приказ резултата фитовања модела расподела

Код препознавања сигнала, у помоћном панелу треба подесити опције везане за подешавање НДКР, након чега се може покренути препознавање сигнала. Иако корисник може учитати скуп сигнала пре одабира расподеле вероватноће и њених параметара, систем захтева да се ове опције одаберу пре покретања процеса фитовања.

8. ЗАКЉУЧАК

У савременом добу, напредак технологије омогућио је мерење и анализу података у реалном времену. Одмеравањем таквих сигнала добијају се њихове дискретне вредности које се затим могу обрадити и анализирати применом различитих алгоритама. У многим практичним применама пресудно је утврдити сличност расподеле улазног сигнала са одређеном теоријски дефинисаном расподелом.

Предмет истраживања ове докторске дисертације јесте проналажење новог приступа за детекцију и препознавање расподеле анализираног сигнала, као и одређивање параметара те расподеле применом метода истраживања података. Конкретан пример примене јесте на препознавању расподела сигнала код диверзити пријемника у базним станицама мобилне телефоније. Циљ је био омогућити пријемнику избор сигнала са најбољим квалитетом на основу распознате расподеле анвелопе фединга и њених параметара. Други пример јесте анализа интернет саобраћаја где је покушано проналажење расподела вероватноће које најбоље моделују његове карактеристике.

Поред увода и закључка, који представљају прво и последње поглавље ове дисертације, она садржи још пет тематских целина.

У уводном делу, аутор описује предмет и циљеве истраживања, као и начин примене метода истраживања података за ефикасно препознавање и анализу расподела вероватноће које се јављају у реалним комуникационим системима.

Друго поглавље обухвата детаљан преглед метода истраживања података, укључујући њихов историјски развој, област примене, и различите фазе процеса истраживања података као што су сакупљање података, одабир и екстракција карактеристика, трансформација, редукција, моделовање и оптимизација процеса. Посебан акценат је на техникама за препознавање расподела и процену њихових параметара.

Тема трећег поглавља јесте описивање расподела вероватноће које се јављају у каналима фединга. Представљени су најчешћи модели фединга: гама, Рејлијев, Рајсов, Накагамијев и Вејбулов модел, и разматра се њихова примена у анализи различитих типова комбинера у диверзити системима.

У четвртом поглављу разматран је нови приступ процени модела бежичног сигнала применом алгоритама нелинеарне регресије. Такође, показани су поступци генерисања и предобраде сигнала са унапред познатим расподелама и параметрима, како би се омогућило квантификовање грешака препознавања.

У петом поглављу су представљени резултати примене нелинеарне регресије за препознавање канала фединга и анализа добијених резултата. Изложени су резултати препознавања различитих расподела у бежичним каналима. Показано је како препознавање расподеле и параметара расподеле, зависе од броја одмерака сигнала и броја тачака НДКР, чиме је испуњен други (практични) циљ дисертације.

Шесто поглавље проширује примену нелинеарне регресије на анализу интернет саобраћаја и препознавање DDoS напада, користећи велике скупове података и интеграцију са Oracle машинским учењем. У овом поглављу је показано како Оракл машинско учење за R решава проблеме анализе великих скупова података смештених у базама података. Показано је и на који начин се могу извршавати алгоритми машинског учења директно у бази података. Овај приступ елиминише потребу за комплексним SQL упитима и преносом великих количина података између клијента и сервера, што значајно убрзава процес анализе великих података.

У седмом поглављу је представљена и апликативни софтвер који је развијен током рада на овој дисертацији, чиме је испуњен трећи циљ истраживања. Развијени софтвер представља свеобухватно решење за генерисање псеудослучајних сигнала, као и фитовање расподела вероватноће и процењивање вредности њихових параметара.

Анализом и применом новог приступа за препознавање расподела вероватноће сигнала, остварени су следећи доприноси у докторској дисертацији:

- извршена је анализа резултата препознавања расподела вероватноће код канала са федингом над великим скупом генерисаних сигнала применом предложене методологије;
- показано је како предложени приступ омогућава процену расподела вероватноће сигнала и њихових параметара без априори познавања параметара или карактеристика самог сигнала;
- утврђено је да параметар скале нема утицаја на перформансе препознавања расподеле вероватноће сигнала;
- спроведена је упоредна анализа утицаја дужине сигнала и броја тачака нормализоване друге централне кумулативне расподеле на брзину и тачност препознавања сигнала;
- извршена је анализа моделовања различитих карактеристика интернет саобраћаја и указано на могућности примене за детекцију дистрибуираних напада путем преоптерећења сервиса;
- анализиран је велики скуп података похрањен у бази података и показане су могућности извршавања машинског учења директно у оквиру базе података, без преноса података између клијента и сервера.

Добијени резултати могу бити од помоћи пројектантима бежичних комуникационих система код дизајнирања система са оптималним вредностима параметара модела бежичних канала тако да се добијају што боље перформансе приликом пријема и обраде сигнала у бежичним каналима. Такође, предложени приступ би могао бити од практичног значаја у рачунарским мрежама где мрежни администратори могу искористити предложено решење како би добили јаснији увид у оптерећење система, безбедносне претње и карактеристике саобраћаја.

Правац будућих истраживања би могао бити усмерен на унапређивање датог алгоритма тако да се омогући прецизније откривање специјалних случајева у

којима се расподеле вероватноће своде једна на другу, а у циљу добијања једнозначних резултата. Постојећи алгоритам би требало побољшати додавањем функционалности за уклањање шума из сигнала. Такође, предложени приступ би се могао применити над сложенијим моделима статистичких расподела бежичних канала, као и у другим реалним процесима где је неопходно открити расподелу вероватноће.

9. ЛИТЕРАТУРА

[1] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[2] T. M. Mitchell, *Machine learning*, vol. 1, no. 9. McGraw-hill New York, 1997.

[3] N. Kunnathuvalappil Hariharan, "Applications of data mining in finance," *Naveen Kunnathuvalappil Hariharan2018 Appl. DATA Min. FINANCE Int. J. Innov. Eng. Res. Technol.*, vol. 5, no. 2, pp. 72–77, 2018.

[4] A. Sharma and P. K. Panigrahi, "A review of financial accounting fraud detection based on data mining techniques," *ArXiv Prepr. ArXiv13093944*, 2013.

[5] J. Yang *et al.*, "Brief introduction of medical database and data mining technology in big data era," *J. Evidence-Based Med.*, vol. 13, no. 1, pp. 57–69, 2020.

[6] P. Gundecha and H. Liu, "Mining social media: a brief introduction," *New Dir. Inform. Optim. Logist. Prod.*, pp. 1–17, 2012.

[7] S. Dua and X. Du, *Data mining and machine learning in cybersecurity*. CRC press, 2016.

[8] U. Shafique and H. Qaiser, "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)," *Int. J. Innov. Sci. Res.*, vol. 12, no. 1, pp. 217–222, 2014.

[9] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," *J. Data Warehous.*, vol. 5, no. 4, pp. 13–22, 2000.

[10] G. Piatetsky-Shapiro, "Knowledge discovery in real databases: A report on the IJCAI-89 Workshop," *AI Mag.*, vol. 11, no. 4, pp. 68–68, 1990.

[11] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.

[12] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*, vol. 72. Springer, 2015.

[13] P. Bruce, A. Bruce, and P. Gedeck, *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media, 2020.

[14] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th edition. Amsterdam: Morgan Kaufmann, 2016.

[15] A. Chakrabarti and J. K. Ghosh, "AIC, BIC and recent advances in model selection," *Philos. Stat.*, pp. 583–605, 2011.

[16] "NIST/SEMATECH e-Handbook of Statistical Methods." Accessed: Jun. 30, 2023. [Online]. Available: https://doi.org/10.18434/M32189

[17] B. Lantz, *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.

[18] N. R. Draper and H. Smith, *Applied Regression Analysis*. John Wiley & Sons, 1998.

[19] W. K. Härdle and L. Simar, *Applied multivariate statistical analysis*. Springer Nature, 2019.

[20] D. M. Bates and D. G. Watts, *Nonlinear regression analysis and its applications*. Wiley, 1988.

[21] D. Miljković, S. Ilić, D. Radosavljević, and S. Pitulić, "Application of nonlinear regression in recognizing distribution of signals in wireless channels," *Proc. Est. Acad. Sci.*, vol. 72, no. 2, pp. 105–114, 2023.

[22] J. C. Nash and R. Varadhan, "Unifying optimization algorithms to aid software system users: optimx for R," *J. Stat. Softw.*, vol. 43, pp. 1–14, 2011.

[23] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, 1963.

[24] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2023. [Online]. Available: https://www.R-project.org

[25] "The Comprehensive R Archive Network." Accessed: May 13, 2023. [Online]. Available: https://cran.r-project.org/

[26] S. Panic, M. Stefanovic, J. Anastasov, and P. Spalevic, *Fading and Interference Mitigation in Wireless Communications*. Boca Raton: CRC Press, 2015. doi: 10.1201/b16275.

[27] M. K. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels*. Wiley, 2005.

[28] "Fading basics | types of Fading in wireless communication." Accessed: Jun. 08, 2024. [Online]. Available: https://www.rfwireless-world.com/Articles/Fading-basics-and-types-of-fading-in-wireless-communication.html

[29] P. M. Shankar, *Fading and shadowing in wireless systems*. Springer, 2017.

[30] A. Abdi and M. Kaveh, "On the utility of gamma PDF in modeling shadow fading (slow fading)," presented at the 1999 IEEE 49th Vehicular Technology Conference (Cat. No. 99CH36363), IEEE, 1999, pp. 2308–2312.

[31] P. N. Murthy and R. Satyanarayana, "A Comparison of Rayleigh and Rician Fading Channels Under Frequency-Selective Fading.," *IUP J. Electr. Electron. Eng.*, vol. 3, no. 4, 2010.

[32] S. O. Rice, "Statistical properties of a sine wave plus random noise," *Bell Syst. Tech. J.*, vol. 27, no. 1, pp. 109–157, 1948.

[33] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.

[34] G. L. Stüber, *Principles of Mobile Communication*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-55615-4.

[35] K. Witrisal, Y.-H. Kim, and R. Prasad, "A new method to measure parameters of frequency-selective radio channels using power measurements," *IEEE Trans. Commun.*, vol. 49, no. 10, pp. 1788–1800, 2001.

[36] G. E. Corazza and F. Vatalaro, "A statistical model for land mobile satellite channels and its application to nongeostationary orbit systems," *IEEE Trans. Veh. Technol.*, vol. 43, no. 3, pp. 738–742, 1994.

[37] M. Nakagami, "The m-distribution—A general formula of intensity distribution of rapid fading," in *Statistical methods in radio wave propagation*, Elsevier, 1960, pp. 3–36.

[38] M. Patzold, U. Killat, F. Laue, and Y. Li, "On the statistical properties of deterministic simulation models for mobile fading channels," *IEEE Trans. Veh. Technol.*, vol. 47, no. 1, pp. 254–269, 1998.

[39] L. Rubio, J. Reig, and N. Cardona, "Evaluation of Nakagami fading behaviour based on measurements in urban scenarios," *AEU - Int. J. Electron. Commun.*, vol. 61, no. 2, pp. 135–138, Feb. 2007, doi: 10.1016/j.aeue.2006.03.004.

[40] E. Pajala, T. Isotalo, A. Lakhzouri, E. S. Lohan, and M. Renfors, "An improved simulation model for Nakagami-m fading channels for satellite positioning applications," presented at the 3rd Workshop on Position, Navigation and Communication, Hannover, Germany, 2006, pp. 81–89.

[41] H. A. Suraweera, R. H. Y. Louie, Y. Li, G. K. Karagiannidis, and B. Vucetic, "Two hop amplify-and-forward transmission in mixed rayleigh and rician fading channels," *IEEE Commun. Lett.*, vol. 13, no. 4, pp. 227–229, Apr. 2009, doi: 10.1109/LCOMM.2009.081943.

[42] K. G. Budden, "Radio waves in the ionosphere," *Radio Waves Ionos.*, 2009.

[43] A. Papoulis and S. Unnikrishna Pillai, *Probability, random variables and stochastic processes*. 2002.

[44] G. Tzeremes and C. Christodoulou, "Use of Weibull distribution for describing outdoor multipath fading," presented at the IEEE antennas and propagation society international symposium (IEEE Cat. No. 02CH37313), IEEE, 2002, pp. 232–235.

[45] H. Hashemi, "The indoor radio propagation channel," *Proc. IEEE*, vol. 81, no. 7, pp. 943–968, 1993.

[46] M. D. Yacoub, "The $\alpha $-$\mu $ Distribution: A Physical Fading Model for the Stacy Distribution," *IEEE Trans. Veh. Technol.*, vol. 56, no. 1, pp. 27–34, 2007.

[47] M. D. Yacoub, "The κ-μ distribution and the η-μ distribution," *IEEE Antennas Propag. Mag.*, vol. 49, no. 1, pp. 68–81, 2007.

[48] G. Fraidenraich and M. D. Yacoub, “The α-η-μ and α-κ-μ fading distributions,” presented at the 2006 IEEE ninth international symposium on spread spectrum techniques and applications, IEEE, 2006, pp. 16–20.

[49] M. Alymani, “Fading Channel Parameter Estimation Using Deep Learning,” PhD Thesis, Stevens Institute of Technology, USA, 2021.

[50] M. Alymani, M. H. Alhazmi, A. Almarhabi, H. Alhazmi, A. Samarkandi, and Y.-D. Yao, “Rician K-Factor Estimation Using Deep Learning,” in *2020 29th Wireless and Optical Communications Conference (WOCC)*, May 2020, pp. 1–4. doi: 10.1109/WOCC48579.2020.9114948.

[51] L. Ahrens, J. Ahrens, and H. D. Schotten, “Convolutional-Type Neural Networks for Fading Channel Forecasting,” *IEEE Access*, vol. 8, pp. 193075–193090, 2020, doi: 10.1109/ACCESS.2020.3032933.

[52] G. Azemi, B. Senadji, and B. Boashash, “Ricean K-factor estimation in mobile communication systems,” *IEEE Commun. Lett.*, vol. 8, no. 10, pp. 617–619, 2004.

[53] A. Doukas and G. Kalivas, “Rician K factor estimation for wireless communication systems,” presented at the 2006 International Conference on Wireless and Mobile Communications (ICWMC’06), IEEE, 2006, pp. 69–69.

[54] F. van der Wijk, A. Kegel, and R. Prasad, “Assessment of a pico-cellular system using propagation measurements at 1.9 GHz for indoor wireless communications,” *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 155–162, 1995.

[55] J. Wu and C. Xiao, “Optimal diversity combining based on linear estimation of Rician fading channels,” presented at the 2007 IEEE international conference on communications, IEEE, 2007, pp. 3999–4004.

[56] N. C. Beaulieu and A. M. Rabiei, “Linear Diversity Combining on Nakagami-0.5 Fading Channels,” *IEEE Trans. Commun.*, vol. 59, no. 10, pp. 2742–2752, Oct. 2011, doi: 10.1109/TCOMM.2011.080111.100373.

[57] R. Annavajjala and L. B. Milstein, “Performance analysis of linear diversity-combining schemes on Rayleigh fading channels with binary signaling and Gaussian weighting errors,” *IEEE Trans. Wirel. Commun.*, vol. 4, no. 5, pp. 2267–2278, 2005, doi: https://doi.org/10.1109/TWC.2005.853834.

[58] N. Wang, X. Song, and J. Cheng, “Generalized Method of Moments Estimation of the Nakagami-m Fading Parameter,” *IEEE Trans. Wirel. Commun.*, vol. 11, no. 9, pp. 3316–3325, Sep. 2012, doi: 10.1109/TWC.2012.071612.111838.

[59] A. Abdi and M. Kaveh, “Performance comparison of three different estimators for the Nakagami m parameter using Monte Carlo simulation,” *IEEE Commun. Lett.*, vol. 4, no. 4, pp. 119–121, 2000.

[60] J. Cheng and N. C. Beaulieu, “Maximum-likelihood based estimation of the Nakagami m parameter,” *IEEE Commun. Lett.*, vol. 5, no. 3, pp. 101–103, 2001.

[61] C. Tepedelenlioglu and P. Gao, “Estimators of the Nakagami-m parameter and performance analysis,” *IEEE Trans. Wirel. Commun.*, vol. 4, no. 2, pp. 519–527, 2005.

[62] N. Hadziahmetovic, M. Milisic, M. Ahic Dokic, and M. Hadzialic, “Estimation of nakagami distribution parameters based on signal samples corrupted with multiplicative and additive disturbances,” in *ELMAR 2007*, Sep. 2007, pp. 235–238. doi: 10.1109/ELMAR.2007.4418838.

[63] R. Mitra, A. K. Mishra, and T. Choubisa, “Maximum likelihood estimate of parameters of Nakagami-m distribution,” in *2012 International Conference on Communications, Devices and Intelligent Systems (CODIS)*, Dec. 2012, pp. 9–12. doi: 10.1109/CODIS.2012.6422123.

[64] M. Hadzialic, M. Milisic, N. Hadziahmetovic, and A. Sarajlic, “Moment-based and Maximum Likelihood-based Quotiential estimation of the Nakagami-m fading parameter,” in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, Apr. 2007, pp. 549–553. doi: 10.1109/VETECS.2007.124.

[65] G. J. Mendis, J. Wei, and A. Madanayake, “Deep learning-based automated modulation classification for cognitive radio,” in *2016 IEEE International Conference on Communication Systems (ICCS)*, Dec. 2016, pp. 1–6. doi: 10.1109/ICCS.2016.7833571.

[66] B. Kim, J. Kim, H. Chae, D. Yoon, and J. W. Choi, “Deep neural network-based automatic modulation classification technique,” in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2016, pp. 579–582. doi: 10.1109/ICTC.2016.7763537.

[67] W. Yongshi, G. Jie, L. Hao, L. Li, W. Zhigang, and W. Houjun, “CNN-based modulation classification in the complicated communication channel,” in *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, Oct. 2017, pp. 512–516. doi: 10.1109/ICEMI.2017.8265870.

[68] T. J. O’Shea, J. Corgan, and T. C. Clancy, “Convolutional Radio Modulation Recognition Networks,” in *Engineering Applications of Neural Networks*, C. Jayne and L. Iliadis, Eds., in Communications in Computer and Information Science. Cham: Springer International Publishing, 2016, pp. 213–226. doi: 10.1007/978-3-319-44188-7_16.

[69] H. Xia *et al.*, “Cellular Signal Identification Using Convolutional Neural Networks: AWGN and Rayleigh Fading Channels,” in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Nov. 2019, pp. 1–5. doi: 10.1109/DySPAN.2019.8935857.

[70] K. Alshathri, H. Xia, V. Lawrence, and Y.-D. Yao, “Cellular System Identification Using Deep Learning: GSM, UMTS and LTE,” in *2019 28th Wireless and Optical Communications Conference (WOCC)*, May 2019, pp. 1–4. doi: 10.1109/WOCC.2019.8770700.

[71] G. Lu, Q. Zhang, X. Zhang, F. Shen, and F. Qin, “CNN BASED RICIAN K FACTOR ESTIMATION FOR NON-STATIONARY INDUSTRIAL FADING CHANNEL,” in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov. 2018, pp. 594–598. doi: 10.1109/GlobalSIP.2018.8646650.

[72] H.-J. Zepernick and A. Finger, *Pseudo random signal processing: theory and application*. John Wiley & Sons, 2013.

[73] A. S. Babu and D. K. S. Rao, “Evaluation of BER for AWGN, Rayleigh and Rician fading channels under various modulation schemes,” *Int. J. Comput. Appl.*, vol. 26, no. 9, pp. 23–28, 2011.

[74] M. A. Stephens, “Introduction to Kolmogorov (1933) On the Empirical Determination of a Distribution,” in *Breakthroughs in Statistics: Methodology and Distribution*, S. Kotz and N. L. Johnson, Eds., New York, NY: Springer New York, 1992, pp. 93–105. doi: 10.1007/978-1-4612-4380-9_9.

[75] S. J. Bean, M. Heuser, and P. N. Somerville, “A Fortran program for estimating parameters in a cumulative distribution function,” UNIVERSITY OF CENTRAL FLORIDA ORLANDO DEPT OF MATHEMATICS AND STATISTICS, 1981.

[76] Y. Chen and N. C. Beaulieu, “Estimation of Ricean and Nakagami distribution parameters using noisy samples,” in *2004 IEEE International Conference on Communications (IEEE Cat. No.04CH37577)*, Jun. 2004, pp. 562-566 Vol.1. doi: 10.1109/ICC.2004.1312552.

[77] K. Kuter, “4.6: Weibull Distributions,” in *Math 345 - Probability*, Saint Mary’s College: LibreTexts, 2023. Accessed: Mar. 02, 2024. [Online]. Available: https://batch.libretexts.org/print/Letter/Finished/stats-3243/Full.pdf

[78] L. Bernadó *et al.*, “Multi-dimensional K-factor analysis for V2V radio channels in open sub-urban street crossings,” in *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Sep. 2010, pp. 58–63. doi: 10.1109/PIMRC.2010.5671774.

[79] J. Reig, “Multivariate Nakagami-m distribution with constant correlation model,” *AEU - Int. J. Electron. Commun.*, vol. 63, no. 1, pp. 46–51, Jan. 2009, doi: 10.1016/j.aeue.2007.10.009.

[80] Y. Yang, Y. Fan, and J. O. Royset, “Estimating probability distributions of travel demand on a congested network,” *Transp. Res. Part B Methodol.*, vol. 122, pp. 265–286, 2019.

[81] E. R. Castro, M. S. Alencar, and I. E. Fonseca, “Probability density functions of the packet length for computer networks with bimodal traffic,” *Int. J. Comput. Netw. Commun.*, vol. 5, no. 3, p. 17, 2013.

[82] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” presented at the Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, 2010, pp. 267–280.

[83] D. Miljković, S. Ilić, B. Jakšić, and D. Radosavljević, “Probdistid: A web-based tool for identifying and parameter estimation of probability distributions,” in *2nd International Conference “Conference on advances in science and technology,”* Herceg Novi, Montenegro, Jun. 2023, pp. 251–258. [Online]. Available: https://confcoast.com/img-publications/52/Zbornik%20radova%202023.pdf

[84] M. H. Bhuyan, D. Bhattacharyya, and J. K. Kalita, “An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection,” *Pattern*

Recognit. Lett., vol. 51, pp. 1–7, 2015, doi: https://doi.org/10.1016/j.patrec.2014.07.019.

[85] R. Sanjeetha, A. Raj, K. Saivenu, M. I. Ahmed, B. Sathvik, and A. Kanavalli, "Detection and mitigation of botnet based DDoS attacks using catboost machine learning algorithm in SDN environment," *Int. J. Adv. Technol. Eng. Explor.*, vol. 8, no. 76, p. 445, 2021.

[86] D. Erhan and E. Anarım, "Statistical properties of DDoS attacks," presented at the 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT), IEEE, 2019, pp. 1238–1242.

[87] N. Ahuja, G. Singal, D. Mukhopadhyay, and N. Kumar, "Automated DDOS attack detection in software defined networking," *J. Netw. Comput. Appl.*, vol. 187, p. 103108, 2021.

[88] X. Yang, B. Han, Z. Sun, and J. Huang, "Sdn-based ddos attack detection with cross-plane collaboration and lightweight flow monitoring," presented at the GLOBECOM 2017-2017 IEEE Global Communications Conference, IEEE, 2017, pp. 1–6.

[89] D. Miljković, S. Ilić, B. Jakšić, P. Milić, and S. Pitulić, "Modeling Internet Traffic Packet Length Using Probdistid: A Case Study," in *International Scientific Conference on Information Technology and Data Related Research*, 2023, pp. 172–177. doi: 10.15308/Sinteza-2023-172-177.

[90] K. M. Rezaul and A. Pakštas, "Web traffic analysis based on EDF statistics," *transformation*, vol. 9, no. 1, p. 14, 2006.

[91] "MAWI Working Group Traffic Archive." Accessed: May 06, 2023. [Online]. Available: http://mawi.wide.ad.jp/mawi/

[92] "A Day in the Life of the Internet (DITL)," CAIDA. Accessed: Feb. 26, 2024. [Online]. Available: https://www.caida.org/projects/ditl/

[93] "Traffic Trace Info." Accessed: May 06, 2023. [Online]. Available: http://mawi.wide.ad.jp/mawi/samplepoint-F/2023/202304301400.html

[94] W. John and S. Tafvelin, "Analysis of internet backbone traffic and header anomalies observed," presented at the Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007, pp. 111–116.

[95] K. Pentikousis and H. Badr, "Quantifying the deployment of TCP options-a comparative study," *IEEE Commun. Lett.*, vol. 8, no. 10, pp. 647–649, 2004.

[96] I. Ilić, "Ocenjivanje indeksa repa raspodele korišćenjem nekompletnih uzoraka," *Универзитет У Београду*, 2013.

[97] D. Erhan and E. Anarım, "Boğaziçi University distributed denial of service dataset," *Data Brief*, vol. 32, 2020, doi: https://doi.org/10.1016/j.dib.2020.106187.

[98] D. Erhan, "Boğaziçi University DDoS Dataset." IEEE, Oct. 09, 2019. doi: https://dx.doi.org/10.21227/45m9-9p82.

[99] O. Kupreev, A. Gutnikov, and Y. Shmelev, “Report on DDoS attacks in Q3 2022.” Accessed: Mar. 26, 2024. [Online]. Available: https://securelist.com/ddos-report-q3-2022/107860/

[100] A. Gutnikov, O. Kupreev, and Y. Shmelev, “Kaspersky DDoS report, Q1 2022.” Accessed: Mar. 26, 2024. [Online]. Available: https://securelist.com/ddos-attacks-in-q1-2022/106358/

[101] O. Kupreev, A. Gutnikov, and Y. Shmelev, “DDoS attacks in Q3 2021.” Accessed: Mar. 26, 2024. [Online]. Available: https://securelist.com/ddos-attacks-in-q3-2021/104796/

[102] D. McDermid, *Oracle Machine Learning for R Installation and Administration Guide, Release 1.5.1*, E97849-12th ed. Oracle. [Online]. Available: https://docs.oracle.com/en/database/oracle/machine-learning/oml4r/1.5.1/oread/oracle-machine-learning-r-installation-and-administration-guide.pdf

[103] A. Png and H. Helskyaho, “Oracle Machine Learning in Autonomous Database,” in *Extending Oracle Application Express with Oracle Cloud Features: A Guide to Enhancing APEX Web Applications with Cloud-Native and Machine Learning Technologies*, A. Png and H. Helskyaho, Eds., Berkeley, CA: Apress, 2022, pp. 139–191. doi: 10.1007/978-1-4842-8170-3_5.

[104] W. Chang *et al.*, *shiny: Web Application Framework for R*. 2022. [Online]. Available: https://CRAN.R-project.org/package=shiny

[105] “Posit,” Posit. Accessed: Feb. 10, 2023. [Online]. Available: https://www.posit.co/

[106] H. W. Borchers, *pracma: Practical Numerical Math Functions*. 2022. [Online]. Available: https://CRAN.R-project.org/package=pracma

[107] T. W. Yee, *VGAM: Vector Generalized Linear and Additive Models*. 2023. [Online]. Available: https://CRAN.R-project.org/package=VGAM

[108] T. Wolodzko, *extraDistr: Additional Univariate and Multivariate Distributions*. 2020. [Online]. Available: https://CRAN.R-project.org/package=extraDistr

10. СПИСАК ТАБЕЛА

11. СПИСАК СЛИКА