Utjecaj parametara sportske dinamike na učinkovitost prediktivnih modela u sportu

Šarčević, Ana

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Ana Šarčević

UTJECAJ PARAMETARA SPORTSKE DINAMIKE NA UČINKOVITOST PREDIKTIVNIH MODELA U SPORTU

DOKTORSKI RAD

Zagreb, 2023.

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Ana Šarčević

UTJECAJ PARAMETARA SPORTSKE DINAMIKE NA UČINKOVITOST PREDIKTIVNIH MODELA U SPORTU

DOKTORSKI RAD

Mentor: Izv. prof. dr. sc. Mihaela Vranić

Zagreb, 2023.

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Ana Šarčević

INFLUENCE OF SPORTS DYNAMICS PARAMETERS ON THE EFFICIENCY OF PREDICTIVE MODELS IN SPORTS

DOCTORAL THESIS

Supervisor: Associate professor Mihaela Vranić, PhD

Zagreb, 2023

Doktorski rad izrađen je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva, na Zavodu osnove elektrotehnike i električka mjerenja.

Mentor: izv. prof. dr. sc. Mihaela Vranić

Doktorski rad ima: 112 stranica

Doktorski rad br.: _____

O mentoru

Mihaela Vranić rođena je 12. rujna 1977. godine u Zagrebu. Nakon završene XV. gimnazije u Zagrebu, 1996. upisuje paralelno dva fakulteta Sveučilišta u Zagrebu, Fakultet elektrotehnike i računarstva (FER) te Ekonomski fakultet. Na FER-u, smjer Telekomunikacije i informatika, diplomirala je s izvrsnim uspjehom i s naglaskom na znanstveno-istraživački rad u lipnju 2001. godine. Iste godine zapošljava se na Fakultetu elektrotehnike i računarstva kao znanstveni novak. Na Ekonomskom fakultetu, smjer Organizacija i management, diplomirala je u rujnu 2002. godine. Na FER-u je magistrirala 2007. godine u polju elektrotehnike s radom pod nazivom: "Primjena poslovne inteligencije u akademskom okruženju" te je doktorirala 2011. godine u polju računarstva s radom naziva "Oblikovanje sažetih prikaza povezanosti elemenata u transakcijskim podacima".

Radila je kao istraživač na više znanstveno-istraživačkih i stručnih projekata. Na nekima od njih bila je voditeljica ili operativna voditeljica. Od svog zaposlenja radila je u nastavi na nizu predmeta preddiplomskih i diplomskih studija na FER-u. Od travnja 2015. godine do lipnja 2016. godine radila je u djelu radnog vremena i kao docent na Hrvatskom katoličkom sveučilištu (HKS-u). Na FER-u je od 1. 10. 2018. godine do 30. 9. 2022. godine u dva mandata obnašala dužnost prodekanice za nastavu.

Znanstveni interesi joj obuhvaćaju elektroničko poslovanje, baze podataka, skladištenje podataka, poslovnu inteligenciju i dubinsku analizu podataka. Autor je većeg broja znanstvenih radova iz područja svog znanstvenog interesa objavljenih na međunarodnim konferencijama te časopisima indeksiranim u vodećim bazama kao što su Current Contents, SCI Expanded, SCOPUS, EBSCO, INSPEC. Članica je programskih odbora i recenzent na više konferencija i časopisa u njezinom području interesa. Članica je Institute of Electrical and Electronics Engineers (IEEE) te Znanstvenog centra izvrsnosti za znanost o podacima i kooperativne sustave.

About the Supervisor

Mihaela Vranić was born on September 12, 1977, in Zagreb. After the end of the XV. gymnasium in Zagreb, in 1996 she enrolled in two faculties of the University of Zagreb, the Faculty of Electrical Engineering and Computing (FER) and the Faculty of Economics. In June 2001, she graduated from FER, majoring in Telecommunications and Informatics, with excellent grades and with an emphasis on science and research. In the same year, she was employed at the Faculty of Electrical Engineering and Computing as a researcher. She graduated from the Faculty of Economics, majoring in Organization and Management, in September 2002. At FER, she received her master science degree in 2007 in the field of electrical engineering with the thesis: "Application of business intelligence in an academic environment" and her doctorate in

2011 in the field of computing with the thesis titled "Designing concise representations of the connection of elements in transactional data".

She worked as a researcher on several scientific research and professional projects. At some of them, she was the head or operational manager. Since her employment, she has taught a number of undergraduate and graduate courses at FER. From April 2015 to June 2016, in part of working time she worked as an assistant professor at the Catholic University of Croatia. At FER, from October 1, 2018 to September 30, 2022, in two terms she held the position of vice dean for education.

Her scientific interests include electronic business, databases, data warehouses, business intelligence and data mining. She is the author of a large number of scientific papers in the field of her scientific interest published at international conferences and journals indexed in leading bases such as Current Contents, SCI Expanded, SCOPUS, EBSCO, INSPEC. She is a member of program committees and a reviewer at several conferences and journals in her field of interest. She is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the Centre of Research Excellence for Data Science and Advanced Cooperative Systems.

Sažetak

Dubinska analiza podataka danas se integrira u mnoštvo domena, od marketinga i maloprodaje, sve do zdravstvene i financijske domene, a svoju primjenu pronašla je i u sportskoj domeni. Menadžeri, momčadi, treneri i ostali subjekti u sportskoj domeni postaju sve spremniji na korištenje podataka za postizanje konkurentske prednosti, kako na terenu tako i izvan njega. Sportska analitika privlači pažnju i kladionica i opće populacije, a oni su uglavnom fokusirani na predviđanje ishoda sportskog događaja.

Priroda sporta uvelike utječe na prilagodbu naprednih tehnika koje će se koristiti za predviđanja. Ova doktorska disertacija fokusirana je na skupinu sportova čija se utakmica sastoji od niza nadigravanja i igra se dok se ne ostvari bodovni cilj, a na njih se referira pojmom *sportovi s bodovnim ograničenjem*. Među sportovima s bodovnim ograničenjem najpopularniji su tenis kao primjer individualnog takvog sporta i odbojka kao primjer timskog sporta iz te skupine sportova. Sportovi poput tenisa i odbojke u znanstvenoj se literaturi često opisuju vremensko diskretnim Markovljevim lancima. Ti modeli se baziraju na procjeni vjerojatnosti osvajanja poena na vlastitom servisu igrača ili momčadi, a oni se, skupa s pretpostavkom o jednolikoj i neovisnoj distribuciji poena, koriste za predviđanje ishoda svake razine odbojkaške utakmice odnosno teniskog meča.

Pretpostavka o jednolikoj i neovisnoj distribuciji poena sukobljava se s konceptima vrlo popularnima u psihologiji poput psihološkog zamaha ili psihološkog pritiska, a osim toga, onemogućava modeliranje umora i inkorporaciju dinamičkih parametara u prediktivne modele. Psihološki zamah, u sportu poznatiji pod pojmom vruća ruka (engl. *hot hand*) ili vrući niz (engl. *hot streak*), definiran je kao kratkoročna iznadprosječna izvedba igrača ili momčadi nakon što se dogodi jedan ili više uzastopnih motivirajućih događaja u igri. Istraživanja na temu postojanja fenomena psihološkog zamaha intenzivno se provode od kraja dvadesetog stoljeća, a suprotno čvrstoj intuiciji većine promatrača sportskih događaja, mišljenja znanstvenika i dalje su jako podijeljena.

Fokus doktorske disertacije se usmjerava na izgradnju prediktivnog modela odbojkaških utakmica, a model se može generalizirati i na ostale sportove koji koriste sličan sustav bodovanja što će biti i opisano. Doktorska disertacija uvodi dva parametra sportske dinamike u modele za predviđanje ishoda odbojkaških utakmica - parametar kratkoročnog zamaha i parametar dugoročnog zamaha. Kroz formulaciju kratkoročnog zamaha, disertacija opisuje kratkoročne iznadprosječne izvedbe igrača nakon određenog motivirajućeg događaja u utakmici. Dugoročni zamah, s druge strane, kombinira povijesne statistike momčadi i simulacijske podatke utakmica čiji se ishodi previđaju. Disertacija nudi matematičke modele za spomenuta dva parametra sportske dinamike koji se temelje na uvjetnoj vjerojatnosti i empirijskom Bayesovom pravilu ažuriranja te nudi hibridni prediktivni model koji objedinjuje navedene parametre. S obzirom

na to da se stručnjaci u domeni sporta često fokusiraju na analizu mogućih rezultatskih sekvenci utakmice, doktorska disertacija koristi Monte Carlo simulaciju koje je interpretabilna od strane domenskih stručnjaka.

Ključne riječi: predviđanje ishoda sportskih događaja, odbojka, jednolika i neovisna distribucija, Markovljevi lanci, Bayesova statistika, Monte Carlo simulacija, psihološki zamah, sportska dinamika.

Influence of sports dynamics parameters on the efficiency of predictive models in sports

Introduction

Soccer, tennis, and volleyball are popular sports played all over the world with players constantly looking for flaws in their opponents' performances. With the advancement of sports analytics, players can abandon traditional methods of analyzing opponents. Observing the opponent's game to research him is now being replaced by extensive data analysis to evaluate the opponent's performance. Coaches can gain insight into their players' performance, analyze player injuries, and study scenarios on the field that may arise under various circumstances. The results of the analyzes can then be used by players, teams and sports organizations to make informed decisions and gain a competitive advantage on and off the field. Aside from coaches, teams, sports organizations, and managers, sports analytics piques the interest of bookmakers and the general public, who are frequently interested in predicting the outcome of a sporting event. Betting on the outcomes of sporting events has grown dramatically in popularity as the Internet has advanced. The Internet offers a more dynamic and practical way of betting while also allowing you to bet on current sporting events.

The nature of each sport greatly influences the adaptation of advanced techniques to be used for prediction. Sports such as golf, cricket, tennis, and volleyball are easy to model with *discrete-time* stochastic processes such as discrete-time Markov chains because they generate sequences of observations that correspond to outcomes after a countable number of holes, balls, shots, and rallies [1]. On the other hand, sports such as football, hockey, athletics, swimming and basketball are often modeled by *continuous-time* stochastic processes such as continuous-time Markov chains since they generate sequences of observations corresponding to scored goals/baskets, changes of ownership or changes of leadership after an uncountable period of time [1]. This dissertation is focused on a group of sports whose match consists of a series of rallies and is played until the point goal is achieved, and they are referred to by the term *sports with a point limit*. Among the sports with a point limit, the most popular are tennis as

an example of an individual such sport and volleyball as an example of a team sport from that group of sports.

Tennis and volleyball are two sports that can be very simply described by a set of states and transitions between them. Markov chains of the first order are commonly used to model such sports, and very important parameters are the statistics of previous matches of the observed players or teams [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Such models rely on estimating the probability of winning points on own serve for both players or teams, which is then fed into mathematical equations based on the Markov chain to estimate the probability of a specific match outcome. The parameters mentioned above are derived from historical data by aggregating a set of previously played matches. Models developed to predict the outcome of tennis or volleyball matches are frequently based on the assumption that the probability of winning a point on one's own serve is independent and identically distributed [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17].

The aforementioned assumption conflict with the intuition about psychological momentum and pressure and make it impossible to model player/team fatigue or other dynamic parameters. Psychological momentum is defined as a brief above-average performance of players/teams after a certain motivating event in a match [18]. Although it is a very intuitive concept that is believed by a large number of sports experts and laymen, in scientific research, the belief in the existence of psychological momentum is still divided [19]. Aside from ignoring the aforementioned parameters, a significant disadvantage of existing models is the profiling of players based solely on historical data. In other words, the player's input statistics are not updated with the statistics of the current match, but remain constant throughout the match. Such models will fail to adequately represent situations in which a sporting event deviates significantly from historical data. So, for example, a strong team may be missing key players due to injuries, resulting in poorer results than what historical statistics show, and a model based solely on historical data may still favor the team more than it should.

The dissertation introduces two sports dynamics parameters: short-term and long-term momentum. Whether it is psychological momentum, pressure, fatigue, another factor, or the style of play itself, the goal of the dissertation is to observe patterns of behavior or reactions of the team at specific points during the match, to incorporate them into the model through the shor-term momentum formulation and to analyze their impact on the predictive model's effectiveness. The dissertation also incorporates the potential directions of the match into the model via the formulation of long-term momentum, and the formulation can be used for the previously mentioned problem of updating historical statistics with the statistics of the current match. Given that sports experts frequently focus on the analysis of possible match result sequences, the dissertation employs an interpretable simulation method to accomplish this. In more detail, the dissertation seeks to depart from the assumption of a identical and independent distribution of

points by incorporating sports dynamics parameters into the model and finally analyzing their effect on the prediction of the basic characteristics of a volleyball match. The goal of the research is to create and propose mathematical models for various parameters of sports dynamics, as well as to create and propose a hybrid predictive model that will unify the previously mentioned parameters. The emphasis will be on developing a predictive model of volleyball matches that can be generalized to other sports that use a similar scoring system, such as tennis.

Description of the field of study

The second chapter gives a brief overview of the field of study. The chapter starts with a description of volleyball's basic rules and scoring structure which are important because they are later incorporated into the model for predicting volleyball match outcomes. Volleyball is an Olympic ball sport in which two opposing teams of 6 players each play on a court with a net in the middle. The game consists of a series of rallies, and each team aims to score a point in the rally. A rally is a sequence of actions in the game from the moment the server played the ball to the moment the ball is out of play. The ball is out of play at the moment of a foul that was whistled by one of the referees, and if there is no foul, at the moment of the whistle. A completed rally is a sequence of actions in the game, the outcome of which is the awarding of points. If the serving team wins the rally, it gets the point and keeps serving. If the receiving team wins the rally, it receives the point and must serve next. A volleyball match is made up of sets. The set is won by the team that first wins 25 points with a lead of at least two points (except in the decisive fifth set). If the score is 24:24, the game continues until a two-point lead is established (26:24, 27:25, etc.). The team that wins three sets wins the match. If the game reaches 2:2, the fifth set is played to 15 points with at least a two-point lead. The first serve in the first set, as well as in the decisive (fifth) set, is performed by the team determined by lot [20].

The chapter then moves on to a description of the domain of betting on sporting events, defining the terminology required to understand the rest of the dissertation, highlighting and describing the problem of adjusting the odds offered by the bookmaker. Betting brings with it a certain risk. The odds represent the inverse risk ratios for each of the bets. Odds are most often presented in three forms, and the most popular is the decimal form. Odds are often much more adjusted to market trends than to the actual state of the match. Due to this behavior there is no way to calculate the real probability of the outcome of a sports event from the odds [12, 21]. This is why normalization is used. Normalization represents the process of converting the odds into the probability of realizing a particular bet. There are several types of normalization, and the dissertation focuses on Shin's normalization because it has been empirically proven to give the best results in the volleyball domain [21]. Finally, the dissertation focuses on the types of betting. The most popular type of betting in volleyball is betting on the winner of the match, and due to the higher odds compared to betting on the winner of the match, handicap betting and

betting on the total number of points that will be played in the match are becoming increasingly popular. This is why the dissertation focuses on handicap betting and betting on the total number of points in a match. Handicap betting in volleyball can be defined as betting on the difference in points or sets scored by two teams.

Probability theory fundamentals and Monte Carlo simulation

The third chapter describes the fundamental concepts of probability theory required to comprehend the dissertation. Random variables are defined and described, with a focus on random variables with identical and independent distribution. The following is a description of stochastic processes, with an emphasis on discrete-time Markov chains. The theoretical basis of conditional probability and Bayesian statistics is then described. Finally, the chapter describes the Monte Carlo simulation that underpins the model for predicting the outcome of volleyball matches. The entire chapter is filled with examples that aid in the acquisition of theoretical background. The most important terms are briefly described below:

- A collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent [22].
- A Markov chain is a probabilistic model consisting of a set of states and transitions between them. The transition to the next state depends solely on the current state [2, 12, 23].
- The conditional probability is a measure of the probability of an event occurring given that another event has already occurred [24].
- Proponents of the Bayesian approach believe that probability is subjective, that it is a degree of belief that changes with the arrival of new information [25].
- The Monte Carlo method (or simulation) is a probabilistic numerical technique used to estimate the outcome of a stochastic process. It is a method of simulating events that are difficult to model directly, and it is used to obtain numerical solutions to problems that are extremely difficult to solve analytically [26].

Related work

The fourth chapter provides an overview of existing research in the domain of modeling and analysis of tennis and volleyball matches. The chapter begins by describing the scientific literature on predicting the outcome of tennis and volleyball matches. The focus is on point-based models. Point-based models are based on the estimation of the probability of winning points on the player's or team's own serve, which are, together with the assumption of the identical and independent point distribution, used to predict the outcome of games, sets and matches in the case of predicting the outcome of a tennis events or set and matches in the case of predicting the outcome of a volleyball events [27, 28, 29]. The subsection that follows focuses on the calculation of the input parameters used in these models - the probabilities of

winning points on the player's or team's own serve. Here, the common opponent method stands out among the more popular methods [30]. Finally, the issue of psychological momentum in sports is discussed. It is an extremely important topic in sports.

Volleyball match modeling using Markov chains

Chapter 5 describes a Markov chain approach of modeling volleyball matches. It is an approach that has previously been described in the scientific literature [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. However, due to the importance of the chapter, its length, and the fluidity of the text, it has been highlighted separately. Volleyball matches are most often modeled through a hierarchical Markov chain, one models a volleyball set, and the other models a volleyball match. The states of that Markov chain are all possible results in the set or match, and the transitions are the probabilities of winning points on the team's own serve, or the probabilities of winning the set - depending on which level of the volleyball match is modeled. They are independent and independently distributed. Based on Markov chains, equations for predicting the outcome of a volleyball set and match are developed.

Dataset

Chapter 6 discusses various datasets in the sports domain and emphasizes the importance of sports analytics. The chapter describes the collection, cleaning, and preprocessing of the volleyball data set used in this dissertation.

Sports data contains information about sports events, players and all other participants in the sports sector, and provides insight into the performance of a team or individual. Datasets in the sports domain can have varying levels of detail. The most powerful datasets in the sports domain are known as play-by-play (PBP) datasets. This type of data provides information on every detail in every aspect of the game. PBP datasets, in other words, provide a transcript of the game in single event format [31, 32, 33, 34, 35]. In addition to detailed PBP datasets, there are sports datasets with a slightly lower level of detail.

The dataset gathered in this dissertation contains information on volleyball matches played between April 2016 and October 2017. The collected dataset lacks the level of detail found in the detailed PBP datasets described above. The dataset used in this dissertation, in particular, only contains information about changes in scores during matches, as well as the corresponding coefficients set by the betting house.

Sports dynamics parameters

The term *sports dynamics parameters* is self-explanatory, and it refers to those parameters that introduce a dynamic element into models used to describe sports events. Models based on an identical and independent distribution of points do not include sports dynamics parameters - the value of the probability of winning points on the team's own serve in such models remain constant throughout the match, regardless of the events in the match, the way of achieving a specific score, or the stage of the match. The match dynamics are not monitored in such mo-

dels. This dissertation introduces two parameters of sports dynamics - short-term momentum and long-term momentum parameter. In this dissertation, the term short-term momentum will be used to describe a player's or team's short-term above-average performance following a motivating event. Motivating events that are primarily considered are one or more consecutive points won on a team's own serve which lead to winning the next point in sequence. The term short-term momentums seems more appropriate than the term psychological momentum given that the dissertation does not consider the cause of such a phenomenon. Long-term momentum is a dynamic parameter that combines historical statistics and simulation data of matches whose outcome is predicted. In more detail, through the formulation of long-term momentum, the probabilities of winning points on the teams' own serve calculated from historical data (expectations) are gradually updated with the information of the score sequences of the simulation. This actually means that the probability of winning a point on team's own serve is slightly adjusted after each point played in the simulation. The defined parameters of sports dynamics are based on conditional probability and the empirical Bayes update rule.

Non-identical and dependent modeling of volleyball matches

Chapter 8 describes a hybrid formulation that combines short-term and long-term momentum parameters, as well as a Monte Carlo simulation method for predicting the outcome of a volleyball match. This model is referred to as *the evolving probability method*. The chapter also addresses the issue of profiling individual teams and proposes a method for profiling groups of similar matches. Finally, the generalization of the evolving probability method to other-point limit sports, as well as the optimization procedure, are described.

Results and discussion

In comparison to previously used models based on an identical and independent distribution of points, the evolving probability method proposed in the dissertation simulates the matches more precisely. When predicting the total number of points to be played in a match, the evolving probability method's error is on average 20% lower than the error of the approach based on an identical and independent point distribution. In the case of handicap prediction, the simulation error is on average 25% lower. The method's strength also lies in its interpretability, which allows it to be easily incorporated into various expert systems. The method is flexible and can be improved by defining additional parameters of sports dynamics, and it can be generalized to sports with similar scoring systems.

Conclusion

This dissertation challenges a common approach to modeling sports with point limits that is based on the assumption of independent and identical point distributions. It is demonstrated that such models can be improved by incorporating dynamics into the match models in the form of sport momentums. Formal mathematical models based on conditional probability and empirical Bayes estimation are proposed for implementing these momentums, which are combined

through a unifying hybrid approach based on Monte Carlo simulation. Finally, the method is applied to real-world volleyball data, demonstrating significant improvements over previous approaches in predicting match outcomes, particularly the handicap and total number of points scored in the match. The method can be integrated into an expert system to gain insight into player performance at various stages of a match or to study field scenarios that may arise under various conditions.

Sadržaj

Poglavlje 1

Uvod

1.1 Motivacija

Digitalizacijom i informatizacijom svijeta, s pojavom i razvojem interneta stvari (engl. *Internet of Things*, IoT), sve većom aktivnošću ljudi na društvenim mrežama i internetu općenito, stvara se i raste novo virtualno područje djelovanja čovječanstva koje zahtijeva spremanje, obradu i isporuku velike količine podataka. Veličina i broj dostupnih podatkovnih skupova drastično raste, a podatke prikupljaju različiti uređaji od mobitela i antena, preko kamera i mikrofona, sve do uređaja za identifikaciju radijske frekvencije (engl. *Radio-frequency identification*, RFID), bežičnih senzorskih mreža i drugih uređaja. Današnje društvo svakodnevno proizvodi ogromne količine podataka, a oni predstavljaju resurse iznimne važnosti. Sirovi, neobrađeni podatci često su beskorisni, nužna je njihova obrada, a često se važnim ispostavlja korištenje metoda za automatsko izvlačenje informacija iz njih.

Dubinska analiza podataka računalni je način obrade podataka u svrhu pronalaženja skrivenih i korisnih informacija, a uključuje statističke metode, metode strojnog učenja i sustave baza podataka [36]. Osnovu za dubinsku analizu podataka čini konačan skup podataka dobiven iz nekog procesa, a na temelju njega je moguće ekstrahirati korisne informacije, modelirati događaje od interesa i predviđati kretanja određenih varijabli [37]. Zadatci dubinske analize podataka dijele se u dvije kategorije: zadatci deskriptivne analize podataka i zadatci prediktivne analize podataka [38]. Zadatci deskriptivne analize podataka karakteriziraju svojstva podataka u ciljnom skupu podataka dok zadatci prediktivne analize podataka izvode indukciju na trenutnim podatcima kako bi se napravila predviđanja [38]. Drugim riječima, dok deskriptivna analiza podataka korisnicima pomaže razumjeti trenutne događaje, prediktivna analitika preuzima proaktivnu ulogu i analizom povijesnih podataka omogućuje korisnicima predvidjeti buduće događaje i ponašanja. S ciljem poboljšanja poslovanja, dubinska analiza podataka integrira se u mnoštvo domena od marketinga i maloprodaje, sve do zdravstvene i financijske domene, a svoju primjenu pronašla je i u domeni sporta gdje menadžeri i timovi postaju sve spremniji na

korištenje podataka za postizanje konkurentske prednosti.

Izvješće o globalnom sportskom tržištu iz 2022. godine iznosi da je globalno sportsko tržište jedno je od najvećih na svijetu te da je poraslo s 354,96 milijardi američkih dolara u 2021. godini na 496,52 milijarde američkih dolara u 2022. godini s prosječnom godišnjom stopom rasta (engl. *compound annual growth rate*, CAGR) od 39,9% [39]. Nogomet, tenis i odbojka sportovi su gledani diljem svijeta, a igrači neprestano pokušavaju pronaći rupu u izvedbi svojih protivnika. Razvojem sportske analitike igrači se mogu odmaknuti od konvencionalnih metoda analize protivnika. Istraživanja protivnika promatranjem njegove igre sada se mogu zamijeniti opsežnom analizom prikupljenih podataka iz kojih se može procijeniti izvedba protivnika. Osim toga, treneri mogu dobiti uvid u performanse svojih igrača [40, 41, 42], mogu analizirati ozljede igrača [43] ili pak proučiti scenarije na terenu koji mogu nastati pod različitim okolnostima [44]. Rezultate analiza tada mogu koristiti igrači, momčadi i sportske organizacije kako bi donosile informirane odluke te dobili konkurentsku prednost na terenu i izvan njega. Detaljnije, treneri i igrači mogu iskoristiti znanje kako bi poboljšali svoje performanse na terenu dok se sportske organizacije u osnovi bave poslovnom stranom sporta, a znanja koriste kako bi doveli do većeg rasta i profitabilnosti organizacije.

Osim trenera, momčadi, sportskih organizacija i menadžera, sportska analitika privlači pažnju i kladionica te opće populacije koji su najčešće fokusirani na predviđanje ishoda sportskog događaja. Napretkom interneta klađenje na ishode sportskih događaja zabilježilo je dramatičan porast popularnosti. Internet pruža dinamičniji i praktičniji način klađenja, a istovremeno nudi mogućnost klađenja na tekuće sportske događaje. Globalno tržište online kockanja* naraslo je sa 73,42 milijarde američkih dolara u 2021. godini na 81,08 milijardi američkih dolara u 2022. godini, a očekuje se da će se širiti po prosječnoj godišnjoj stopi rasta od 9,2% u periodu od 2022. godine do 2026. godine [45]. Europska udruga za igre na sreću i klađenje (engl. *European Gaming & Betting Association*, EGBA) tvrdi da europsko tržište kockanja na internetu raste za otprilike 10% po godini [46]. Prema Europskoj udruzi za igre na sreću i klađenje očekuje se da će se ekonomska veličina internetskog sektora Europske unije povećati sa 30,5 milijarde američkih dolara u 2020. godini na 52 milijarde američkih dolara u 2026. godini [46]. Sportska analitika dovela je sportsko klađenje na novu razinu. Razvijene su brojne tvrtke i web stranice kako bi se navijačima pružila najnovija saznanja koja su od iznimne važnosti za donošenje odluka.

Priroda svakog sporta uvelike utječe na prilagodbu naprednih tehnika koje će se koristiti za predviđanje. Sportove poput golfa, kriketa, tenisa i odbojke lako je modelirati vremensko diskretnim (engl. *discrete-time*) stohastičkim procesima poput vremensko diskretnih Markovljevih lanaca jer oni generiraju nizove opažanja koji odgovaraju rezultatima nakon prebrojivog

*Tržište online kockanja obuhvaća korištenje interneta za igranje casino igara, pokera i klađenje na ishode sportskih događaja.

broja rupa, lopti, udaraca i nadigravanja [1]. S druge strane, sportovi poput nogometa, hokeja, atletike, plivanja i košarke često se modeliraju vremensko kontinuiranim (engl. *continuous-time*) stohastičkim procesima poput vremensko kontinuiranih Markovljevih lanaca budući da oni generiraju nizove opažanja koji odgovaraju postignutim golovima/koševima, promjenama posjeda ili promjenama vodstva nakon neprebrojivog perioda vremena [1]. Ova doktorska disertacija fokusirana je na skupinu sportova čija se utakmica sastoji od niza nadigravanja i igra se dok se ne ostvari bodovni cilj, a na njih se referira pojmom *sportovi s bodovnim ograničenjem*. Među sportovima s bodovnim ograničenjem najpopularniji su tenis kao primjer individualnog takvog sporta i odbojka kao primjer timskog sporta iz te skupine sportova.

Tenis i odbojka dva su sporta koja se vrlo jednostavno mogu opisati skupom stanja i prijelazima između njih. Za modeliranje takvih sportova najčešće se koriste Markovljevi lanci prvog reda, a vrlo važne parametre predstavljaju statistike ranijih susreta promatranih igrača odnosno momčadi [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Detaljnije, većina modela za predviđanje ishoda teniskih mečeva ili odbojkaških utakmica koji se mogu pronaći u znanstvenoj literaturi, opisuju teniski meč odnosno odbojkašku utakmicu hijerarhijskim Markovljevim modelom. Takvi modeli oslanjaju se na procjenu vjerojatnosti osvajanja poena na vlastitom servisu za oba igrača odnosno momčadi, a one se naknadno unose u matematičke jednadžbe temeljene na Markovljevom lancu kako bi se procijenila vjerojatnost određenog ishoda utakmice. Navedeni parametri računaju se iz povijesnih podataka agregiranjem određenog broja prethodno odigranih utakmica. Izgrađeni modeli za predikciju ishoda teniskih mečeva ili odbojkaških utakmica najčešće se temelje na pretpostavci da je vjerojatnost osvajanja poena na vlastitom servisu neovisno i jednoliko distribuirana (engl. *independent and identically distributed*, IID) [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. Pod pojmom neovisno, misli se na to da vjerojatnost osvajanja poena na vlastitom servisu ne ovisi o ishodu prethodno odigranog poena. S druge strane, jednolika distribucija u ovom slučaju podrazumijeva da se svaki poen smatra jednakim, neovisno o tome radi li se o jako bitnom poenu u posljednjem setu utakmice ili se pak radi o manje važnom poenu tijekom utakmice [2].

Spomenuta pretpostavka o jednolikoj i neovisnoj distribuciji poena sukobljava se s intuicijom o psihološkom zamahu i pritisku te onemogućuju modeliranje umora igrača/momčadi ili drugih dinamičkih parametara. Psihološki zamah predstavlja kratku natprosječnu izvedbu igrača/momčadi nakon određenog motivirajućeg događaja u meču/utakmici [18], a psihološki pritisak predstavlja ispodprosječnu izvedbu igrača u odlučujućim trenutcima utakmice/meča [47]. Gotovo je nemoguće gledati utakmicu a da sportski komentator barem jednom tijekom utakmice nije spomenuo kako se određeni igrač „razigrao“, a promatra li se broj predaha (engl. *time-out*) koje treneri pozovu tijekom utakmica, nakon nekoliko uzastopno dobrih poteza protivničke momčadi, očito je da i treneri vjeruju u postojanje psihološkog zamaha. Od kraja dvadesetog stoljeća ispitivanje postojanja psihološkog zamaha u sportskoj domeni bitna je

tema znanstvenih radova. Istraživanja se provode na velikom broju sportova osobito u košarci [48, 49, 50, 51], bejzbolu [52], tenisu [53, 54, 55, 56, 57, 58, 59, 60] i odbojci [61, 62, 63]. Iako je riječ o vrlo intuitivnom konceptu u koji vjeruje velik broj sportskih stručnjaka i laika, u znanstvenim istraživanjima mišljena o postojanju psihološkog zamaha i dalje su podijeljena.

Osim zanemarivanja navedenih parametara, velika mana postojećih modela je i profiliranje igrača samo na temelju povijesnih podataka. Drugim riječima, ulazna statistika igrača (u opisanim modelima to su samo vjerojatnosti osvajanja poena na vlastitom servisu igrača/momčadi) ne ažurira se statistikom trenutne utakmice, već ostaje konstantna tijekom cijele utakmice. Takvi modeli neće na odgovarajući način predstavljati situacije kada sportski događaj snažno odstupa od povijesne statistike. Tako na primjer jakoj momčadi mogu nedostajati ključni igrači zbog ozljeda što dovodi do lošijih rezultata u odnosu na ono što pokazuje povijesna statistika, a model koji koristi samo povijesne podatke može i dalje favorizirati momčad jače nego što bi trebalo.

Doktorska disertacija uvodi dva parametra sportske dinamike, kratkoročni i dugoročni zamah. Neovisno o tome radi li se o psihološkom zamahu, pritisku, umoru, nekom drugom faktoru ili pak o samom stilu igre, ideja doktorske disertacije je, kroz formulaciju kratkoročnog zamaha, uočiti obrasce ponašanja odnosno reakcije momčadi u određenim trenutcima utakmice te analizirati njihov utjecaj na učinkovitost prediktivnih modela. Doktorska disertacija kroz formulaciju dugoročnog zamaha u model inkorporira potencijalne smjerove utakmice, a ista formulacija se može iskoristiti i za prethodno spomenuti problem ažuriranja povijesne statistike statistikom trenutne utakmice. S obzirom na to da se stručnjaci u domeni sporta često fokusiraju na analizu mogućih rezultatskih sekvenci utakmice, doktorska disertacija koristi interpretabilnu simulacijsku metodu kako bi se to moglo i izvesti. Detaljnije, ideja doktorske disertacije je odmaknuti se od pretpostavke o jednolikoj i neovisnoj distribuciji poena te u model uključiti parametre sportske dinamike i konačno analizirati njihov učinak na predikciju osnovnih karakteristika odbojkaške utakmice. Cilj istraživanja jest osmisliti i predložiti matematičke modele za različite parametre sportske dinamike te osmisliti i predložiti hibridni prediktivni model koji će objediniti navedene parametre. Fokus će se usmjeriti na izgradnju prediktivnog modela odbojkaških utakmica, a model će se moći generalizirati i na sportove koji koriste sličan sustav bodovanja, poput tenisa.

1.2 Doprinosi

Napomena: Zbog nedostatka potrebne terminologije ova doktorska disertacija uvodi pojam *sportovi s bodovnim ograničenjem*. Pod pojmom sportovi s bodovnim ograničenjem misli se na sportove čija se utakmica sastoji od niza nadigravanja i igra se dok se ne ostvari neki bodovni cilj. Primjeri takvih sportova su odbojka, odbojka na pijesku, tenis, stolni tenis, badminton, skvoš itd.

Znanstveni doprinos ovog rada sastoji se u sljedećem:

1. Formalno definiranje parametara sportske dinamike u sportovima s bodovnim ograničenjem koje se temelji na uvjetnoj vjerojatnosti i empirijskom Bayesovom pravilu ažuriranja.
2. Postupak oblikovanja modela za predviđanje razlike u broju osvojenih bodova suparnika i ukupnog broja dovršenih nadigravanja u utakmicama sportova s bodovnim ograničenjem koji uzima u obzir parametre sportske dinamike, a temelji se na simulaciji Monte Carlo.
3. Analiza utjecaja parametara sportske dinamike na u činkovitost modela za predviđanje razlike u broju osvojenih bodova suparnika i ukupnog broja dovršenih nadigravanja u utakmicama sportova s bodovnim ograničenjem.

1.3 Struktura rada

Struktura doktorske disertacije je sljedeća:

Napomena: Cijela doktorska disertacija primarno je fokusirana na modeliranje utakmica dvoranske odbojka, međutim, u poglavlju 8 opisana je generalizacija modela na utakmice drugih sportova s bodovnim ograničenjem.

Poglavlje 2 daje kratak opis domene od interesa. Poglavlje započinje opisom osnovnih pravila i strukture bodovanja u odbojci. Opis pravila i opis strukture bodovanja u odbojci bitni su jer se naknadno inkorporiraju u model za predviđanje ishoda odbojkaških utakmica. U poglavlju zatim slijedi opis domene klađenja na ishode sportskog događaja - definira se terminologija potrebna za razumijevanje ostatka doktorske disertacije, ističe se i opisuje problem prilagodbe koeficijenata koje nudi kladionica te se opisuje način rješavanja tog problema.

Poglavlje 3 opisuje osnovne koncepte teorije vjerojatnosti potrebne za razumijevanje doktorske disertacije. Definiraju se i opisuju slučajne varijable, a fokus se usmjerava na slučajne varijable s jednolikom i neovisnom razdiobom. Slijedi opis stohastičkih procesa s naglaskom na vremensko diskretne Markovljeve lance. U poglavlju je zatim opisana teorijska osnova uvjetne vjerojatnosni i Bayesove statistike. Poglavlje konačno opisuje Monte Carlo simulacijsku metodu na kojoj se bazira model za predviđanje ishoda odbojkaških utakmica. Cijelo poglavlje prožeto je primjerima koji olakšavaju usvajanje teorijske podloge.

Poglavlje 4 daje pregled znanstvenih istraživanja u domeni analize i modeliranja sportskih događaja s fokusom na analizu i modeliranje teniskih mečeva i odbojkaških utakmica. Na početku poglavlja dan je pregled postojećih istraživanja fokusiranih na predviđanje ishoda teniskih mečeva i odbojkaških utakmica. Zatim je dan pregled znanstvenih istraživanja koji se bave proračunom parametara koji ulaze u model za predviđanje ishoda sportskih događaja. Konačno je definiran pojam psihološkog zamaha te je opisana problematika psihološkog zamaha u sportskoj domeni.

Poglavlje 5 opisuje pristup modeliranja odbojkaških utakmica Markovljevim lancima. Radi se o pristupu otprije poznatom u znanstvenoj literaturi. Međutim, radi važnosti poglavlja, duljine poglavlja i fluidnosti teksta, poglavlje je izdvojeno zasebno.

Poglavlje 6 opisuje različite podatkovne skupove u sportskoj domeni i ističe važnost sportske analitike. Poglavlje zatim opisuje način prikupljanje, čišćenja i predobrade odbojkaškog podatkovnog skupa korištenog u ovoj doktorskoj disertaciji.

Poglavlje 7 definira pojam parametra sportske dinamike te uvodi, definira i opisuje dva takva parametra - kratkoročni zamah i dugoročni zamah. Definirani parametri sportske dinamike

temelje se na uvjetnoj vjerojatnosti i empirijskom Bayesovom pravilu ažuriranja. Poglavlje, u domeni odbojke, daje matematičke modele za ta dva parametra sportske dinamike.

Poglavlje 8 opisuje hibridnu formulaciju koja kombinira parametre kratkoročnog i dugoročnog zamaha te opisuje metodu za predviđanje ishoda odbojkaških utakmica temeljenu na Monte Carlo simulaciji. Doktorska disertacije se na taj model referira pojmom *Napredna metoda ažuriranja vjerojatnosti*. Poglavlje također opisuje problematiku profiliranja momčadi, a zbog ograničenosti podatkovnog skupa, u te svrhe predlaže i opisuje pristup profiliranja grupa sličnih utakmica. Poglavlje konačno opisuje generalizaciju napredne metode ažuriranja vjerojatnosti na druge sportove s bodovnim ograničenjem i opisuje postupak optimizacije napredne metode ažuriranja vjerojatnosti nad podatkovnim skupom iz domene odbojke.

Poglavlje 9 opisuje rezultate istraživanja.

Rad je zaključen s poglavljem 10 u kojem su dani zaključci istraživanja i smjernice za njegov nastavak.

Poglavlje 2

Opis domene od interesa

Fokus doktorske disertacije je na modeliranju utakmica dvoranske odbojke. Međutim, teorije i principi predloženi u doktorskoj disertaciji mogu se jednostavno primijeniti i na ostale sportove s bodovnim ograničenjem poput odbojke na pijesku, tenisa, stolnog tenisa, badmintona i slično (detaljnije u poglavlju 8.4). Poglavlje započinje opisom osnovnih pravila i strukture bodovanja u odbojci nakon čega se fokusira na domenu klađenja na ishode sportskih događaja. Tu se uvodi i definira terminologija potrebna za razumijevanje ostatka doktorske disertacija, a konačno se opisuje problem prilagodbe koeficijenata i predlaže se način njegovog rješavanja.

2.1 Osnovna pravila i struktura bodovanja u odbojci

Napomena: u nastavku doktorske disertacije se pod pojmom odbojka misli na dvoransku odbojku. Postoji još i odbojka na pijesku koja nije predmet istraživanja ove doktorske disertacije, ali kao što je već navedeno i na nju se mogu primijeniti razvijeni principi i metodologija.

Napomena: Ostatak poglavlja napisan je na temelju službenih odbojkaških pravila opisanih u [20].

Odbojka je olimpijski sport s loptom u kojem igraju dvije protivničke momčadi od po 6 igrača na terenu koji na sredini ima razapetu mrežu. Igra se sastoji od niza nadigravanja, a svaka momčad za cilj ima ostvariti bod u nadigravanju. Nadigravanje je slijed akcija u igri od trenutka kada je server odigrao loptu do trenutka kada je lopta izvan igre. Lopta je izvan igre u trenutku pogreške koju je zviždukom dosudio jedan od sudaca, a ako nema pogreške, u trenutku zvižduka. Dovršeno nadigravanje je slijed akcija u igri čiji je ishod dodjela poena.

Momčad može osvojiti poen na više načina:

1.kada postigne da lopta padne u protivni čko polje igrališta.

- •Lopta je „unutra“ ako u bilo kojem trenutku njenog dodira s podom, neki njen dio dodiruje pod igrališta uključujući i granične crte. Lopta upućena prema protivničkom polju igrališta mora prijeći iznad mreže unutar prostora za prijelaz. Prostor za prijelaz dio je vertikalne ravnine mreže ograničen: s donje strane gornjim rubom mreže; s obje strane antenama i njihovim zamišljenim produžetcima; s gornje strane stropom.

2.kada protivni čka momčad načini pogrešku.

- •Mom čad čini pogrešku kad izvede akciju u igri protivnu pravilima ili ih prekrši na bilo koji drugi način. Suci dosuđuju pogreške i određuju posljedice prema pravilima:
  - – ako su učinjene dvije ili više uzastopnih pogrešaka, uzima se u obzir samo prva.
  - – ako su dvije ili više pogrešaka istodobno učinili protivnički igrači, dosuđuje se obostrana pogreška i nadigravanje se ponavlja.

3.kada protivni čka momčad dobije kaznu.

- •Prvo odugovla čenje člana jedne momčadi na utakmici sankcionira se "opomenom zbog odugovlačenja". Drugo i sljedeća odugovlačenja bilo koje vrste od bilo kojeg člana iste momčadi u istoj utakmici smatraju se pogreškom i sankcioniraju "kaznom zbog odugovlačenja", što znači dodjeljivanje poena i prava serviranja protivniku.
- •Prvo neuljudno ponašanje na utakmici bilo kojeg člana momčadi kažnjava se dodjelom poena i prava serviranja protivniku.

Ako je momčad koja je servirala osvojila nadigravanje ona osvaja poen i nastavlja servirati. Ako je momčad koja je primala servis osvojila nadigravanje ona osvaja poen i mora servirati sljedeća. Odbojkaška utakmica sastoji se od setova. Set osvaja momčad koja je prva osvojila 25 poena (osim odlučujućeg, petog seta) s vodstvom od najmanje dva poena. U slučaju izjednačenog rezultata 24:24, igra se nastavlja dok se ne postigne vodstvo od dva poena (26:24, 27:25, itd.). Utakmicu dobiva momčad koja je pobjednik triju setova. U slučaju rezultata 2:2 u setovima, odlučujući peti set igra se do 15 poena s vodstvom od najmanje dva poena. Prvi servis u prvom setu, kao i u odlučujućem (petom) setu, izvodi ekipa određena ždrijebom.

Ovdje opisana pravila koristit će se u Monte Carlo simulaciji. Teorijska osnova za razumijevanje Monte Carlo simulacije dana je u poglavlju 3.5, a način inkorporiranja odbojkaških pravila u Monte Carlo simulaciju opisan je u poglavlju 8.2.

2.2 Domena klađenja na ishode sportskih događaja

Odbojka je jedan od najpopularnijih sportova na svijetu, a uživaju je milijuni gledatelja prateći brojne turnire kroz godinu. Radi se o naširoko igranom timskom sportu kojem je popularnost naglo porasla stjecanjem olimpijskog statusa 1957. godine. S porastom popularnosti tog sporta i klađenje na ishode odbojkaških utakmica uzelo je maha. Niz je odbojkaških prvenstava i natjecanja na koja se može kladiti, a industrija klađenja na odbojkaške utakmice broji milijunske iznose.

U industriji klađenja ističu se dvije uobičajene vrste oklada, klađenje prije početka sportskog događaja, takozvano *pre-mach* klađenje i klađenje za vrijeme sportskog događaja odnosno *in-play* klađenje. Pojmovi su samoopisni pa je jednostavno za zaključiti da je u klađenju prije početka sportskog događaja moguće postaviti okladu samo prije nego što je sportski događaj započeo dok klađenje za vrijeme sportskog događaja omogućuje postavljanje oklade tijekom sportskog događaja.

2.2.1 Koeficijenti u sportskim kladionicama

Za potrebe razumijevanja ostatka doktorske disertacije ključno je definirati pojam koeficijenta odnosno kvote. Klađenje sa sobom donosi određeni rizik. Koeficijenti predstavljaju inverzne omjere rizika za svaku od oklada. Koeficijenti se najčešće prikazuju u tri oblika. Dok se razlomački koeficijenti (poznati i kao Engleski koeficijenti) koriste samo u Velikoj Britaniji, a *moneyline* koeficijenti (poznati i kao Američki koeficijenti) u SAD-u, decimalni koeficijenti (poznati kao Europski koeficijenti) se najčešće koriste u ostatku svijeta [12]*.

Decimalni oblik koeficijenta, kako samo ime sugerira, je prikazan decimalnim brojem. Računanje potencijalnog dobitka koristeći ovaj oblik koeficijenta je iznimno lagan. Ponuđene decimalne koeficijente postavljene na ishod određenog događaja potrebno je pomnožiti s ulogom kladioničara (u ovom kontekstu pojam kladioničar opisuje osobu koja se kladi), a potom oduzeti od inicijalnog uloga [12]. Potrebno je napomenuti da pojam kladioničara ima dvojako značenje. Pod istim pojmom definira se onaj koji drži, vodi i upravlja kladionicom (engl. *bookmaker*). Isti pojam međutim definira i osobe koje se klade. U nastavku disertacije, sve dok se ne navede suprotno, pojam kladioničar koristit će se za opisivanje osobe koja vodi, drži upravlja kladionicom, a za osobu koja se kladi koristit će se pojam klijent.

Primjer računanja potencijalnog dobitka koristeći decimalni oblik koeficijenata: Zamislimo da utakmicu igraju momčad A protiv momčadi B, a kladionica je postavila koeficijent $1,50$ na pobjedu momčadi A i koeficijent $2,48$ na pobjedu momčadi B. Ako klijent postavi okladu od 100 novčanih jedinica na pobjedu momčadi A i ako se ona ostvari, klijent će zaraditi 50

*Osim ova tri oblika koeficijenata, postoje još Hong Kong-ški, Indonežanski i Malezijski tip koeficijenata, ali oni su manje poznati i popularni su uglavnom u državama čije ime i nose.

novčanih jedinica. Ako pak klijent uloži 100 novčanih jedinica na pobjedu momčadi *B* i ako se ona ostvari, klijent će osvojiti 148 novčanih jedinica.

Tko postavlja koeficijente i na koji način? U svijetu su zastupljene dvije vrste kladioničarskih kuća. *Betting exchanges* posluju na principu da se klijenti klade jedni protiv drugih i na temelju uplata klijenata se stvaraju koeficijenti. Takav oblik kladioničarskih kuća posluje bez rizika te bez obzira na ishod meča uzima proviziju od uplate svakog klijenta i pobjedniku oklade isplaćuje novac gubitnika [12]. Standardni kladioničari s druge strane samostalno postavljaju koeficijente za različite ishode mečeva na koje se u konačnici klijenti klade [12]. Algoritmi za računanje vjerojatnosti ishoda sportskog događaja i za postavljanje koeficijenata razlikuju se ovisno o kladioničarskoj kući, jako su složeni, a u obzir uzimaju različite varijable poput ozljede igrača, utjecaja publike, važnosti utakmice i slično. Idealno bi zbroj vjerojatnosti na svaki mogući ishod događaja trebao biti 1. U praksi to nije slučaj. Naime, koeficijenti u praksi ne predstavljaju pravu inverznu vjerojatnost ishoda jer je na "pravi" koeficijent dodan višak vjerojatnosti odnosno marža (engl. *bookmaker take* ili *bookmaker margin*). Marža kladionice definirana je kao razlika između koeficijenata koje se nude klijentu i stvarne vjerojatnosti ishoda. Ona osigurava kladioničarima konstantnu dobit. Što je manji koeficijent kladioničari smanjuju maržu kako bi ponudili što atraktivnije koeficijente, a nadaju se što manjem broju uplata na taj ishod. Svako tržište nije jednako zahtjevno. Hrvatsko tržište dopušta puno veću zaradu (nepovoljniji koeficijenti) od tržišta razvijenih zemalja gdje je konkurencija veća. Najmanja marža je upravo na oklade s dva moguća ishoda jer su takvi algoritmi najjednostavniji te je najlakše osigurati konstantnu dobit. Osim navedenoga, kladioničari svoje koeficijente konstantno prilagođavaju uplatama klijenata kako bi pazili na ujednačenost knjiga i izbjegli eventualne gubitke na kraju utakmice. Koeficijenti se često puno više prilagođavaju trendovima tržišta nego stvarnom stanju u utakmici [12]. Zbog ovakvog ponašanja kladioničara ne postoji niti jedan način da se iz koeficijenata izračuna stvarna vjerojatnost ishoda sportskog događaja. Upravo zbog toga koristi se normalizacija (više u poglavlju 2.3).

2.2.2 Tipovi klađenja

Konačno je još potrebno obraditi tipove sportskog klađenja. Bitno je napomenuti da postoji jako puno tipova klađenja na sportske događaje. U nastavku su objašnjeni oni tipovi potrebni za razumijevanje ostatka doktorske disertacije. Najpopularniji tip klađenja u odbojci je klađenje na pobjednika utakmice, a zbog većih kvota u usporedbi s klađenjem na pobjednika utakmice sve veću popularnost bilježi i hendikep klađenje i klađenje na ukupan broj poena koji će se odigrati u utakmici.

Hendikep je definiran kao prednost u golovima, koševima, poenima, postocima ili drugim veličinama koje su u vezi s događajem koji je predmet klađenja, a koja je unaprijed određena

i predstavljena klijentu u ponudi kladionice. Drugim riječima, hendikep klađenje u odbojci se može definirati kao klađenje na razliku postignutih poena ili setova dviju momčadi. Detaljnije, kladionica postavlja hipotetski broj poena ili setova, a da bi klijent osvojio okladu, favorit mora ostvariti veću prednost od postavljene.

Primjer hendikep klađenja: Pretpostavimo da utakmicu igraju momčad A protiv momčadi B, a za istu utakmicu postavljen je hendikep $-1,5$ setova za momčad A, odnosno $+1,5$ setova za momčad B. Ako se klijent kladi na momčad A, tada momčad A mora pobijediti momčad B s barem 2 seta razlike, odnosno momčad A mora pobijediti momčad B unatoč početnom "zaostatku" od $1,5$ seta. Dakle kako bi klijent osvojio okladu, momčad A mora pobijediti utakmicu rezultatima $3-1$ ili $3-0$. Isti principi vrijede i za hendikep oklade na razini poena.

Slično je i u klađenju na ukupan broj poena koji se očekuju u utakmici. Ovaj tip klađenja omogućuje klijentu da se kladi na to hoće li ukupni broj poena u utakmici biti veći ili manji od predefiniranog broja poena kojeg postavlja kladionica.

Primjer klađenja na ukupan broj poena u utakmici: Kladionica može ponuditi $182,5$ poena na određenu odbojkašku utakmicu. Klijent tada može postaviti okladu na viši broj poena ako smatra da će se u utakmici odigrati najmanje 183 poena ili pak može postaviti okladu na niži broj poena ako pretpostavlja da će se u utakmici odigrati manje od 182 poena.

2.3 Normalizacija

Kao što je spomenuto u poglavlju 2.2.1, kladionice neprestano prilagođavaju koeficijente ovisno o uplatama klijenata. Drugim riječima, one nude "nepoštene" koeficijente kako bi osigurale profit. Iz tog razloga se obrnuti koeficijenti ne mogu interpretirati kao vjerojatnosti već ih je u potrebno normalizirati [21].

Primjer prilagođavanja koeficijenata (primjer preuzet iz [12]): Kao i do sada, pretpostavimo da utakmicu igraju momčad A protiv momčadi B. Kladioničarska kuća na istu utakmicu nudi koeficijent $1,666$ na pobjedu momčadi A i $2,380$ na pobjedu momčadi B (stupac "Koeficijent" u tablici 2.1). Ako su uplate klijenata raspoređene kako je i očekivano ($\approx 60\%$ na pobjedu igrača $A \approx 42\%$ na pobjedu igrača B), tada koeficijenti ostaju stabilni.

Tablica 2.1: Raspodjela uplata igrača na određeni ishod utakmice

Momčad	Koeficijent	Vjerojatnost iz koeficijenata	Raspodjela uplata
A	1,666	60,02%	75%
B	2,380	42,02%	25%

Pretpostavimo sada da su uplate klijenata raspoređene kao u stupcu "Raspodjela uplata" u tablici 2.1. Pretpostavimo također da klijenti izvrše 100 uplata od 1 novčane jedinice prema toj raspodjeli. To znači da u slučaju pobjede momčadi A kladionice ostvaruju prihod od 100 novčanih jedinica, a moraju isplatiti $75 * 1,666 = 124,95$ novčanih jedinica. Radi se o izrazito nepovoljnom ishodu za kladionice koje su mogle izbjeći taj gubitak ispravljanjem koeficijenata. Upravo zbog ispravljanja koeficijenata nije moguće iz njih izračunati stvarnu vjerojatnost na ishod sportskog događaja.

Normalizacija predstavlja proces pretvorbe koeficijenata u vjerojatnosti ostvarivanja pojedine oklade [21, 64]. Nekoliko je metoda normalizacije koeficijenata. U nastavku je fokus na Shinovoj normalizaciji jer je empirijski dokazano da je Shinova normalizacija najpogodnija za transformaciju koeficijenata u vjerojatnosti kada su u pitanju ishodi odbojkaških utakmica [21]. Objašnjena je i osnovna normalizacija (engl. *basic normalization*) jer je potrebna za razumijevanje Shinove normalizacije.

2.3.1 Osnovna normalizacija

Napomena: Opis osnovne normalizacije dan u ostatku poglavlja napisan je na temelju izvora [12, 21, 65].

Pretpostavimo da su $k = (k_1, k_2, ...k_n)$ decimalni koeficijenti ponuđeni za utakmicu sa $n >=$ 2 mogućih ishoda. Neka je $k_i > 1$ za svaki $i = 1, ...n$. Inverzni koeficijenti $\pi = (\pi_1, \pi_2, ...\pi_n)$ tada su:

$$\pi_i = \frac{1}{k_i}. \tag{2.1}$$

Ti inverzni koeficijenti predstavljaju latentne varijable odnosa snaga. Oni ne mogu predstavljati stvarne vjerojatnosti jer je njihova suma veća od 1. Stvarna suma (engl. *booksum*) Π računa se po formuli:

$$\Pi = \sum_{i=1}^{n} \pi_i. \tag{2.2}$$

Višak vjerojatnosti ili marža tada je $\Pi - 1$. Kako bi se konačno standardizirali inverzni koeficijenti one se dijele sa stvarnom sumom Π što konačno daje formulu za osnovnu normalizaciju vjerojatnosti ishoda:

$$p_i = \frac{\pi_i}{\Pi}. \tag{2.3}$$

Osnovna normalizacija ne daje precizne vjerojatnosti ishoda za niske koeficijente. Razlog leži u činjenici da kladioničarske kuće, kako bi očuvale atraktivnost oklade, namjerno smanjuju vlastiti dobitak podižući niske koeficijente. To predstavlja najveću manu osnovne normalizacije [12, 21].

Primjer primjene osnovne normalizacije (primjer preuzet iz [12]): Neka je dan koeficijent $1,05$ na pobjedu momčadi *A*, odnosno koeficijent 9 na pobjedu momčadi *B*. Obrnute vrijednosti tih koeficijenata su vrijednosti $0,95$ i $0,11$, a stvarna suma Π tada iznosi $1,06$. Osnovnom normalizacijom ćemo dobiti da vjerojatnost na pobjedu momčadi *A* iznosi $89,6\%$, a vjerojatnost na pobjedu momčadi *B* iznosi $10,4\%$.

2.3.2 Shinova normalizacija

Napomena: Opis Shinove normalizacije dan u ostatku poglavlja napisan je na temelju izvora [12, 21, 65].

Shinova normalizacija odnosno Shinov model temeljen je na pretpostavci o prisustvu skupine insajdera (engl. *insiders*) sa savršenim poznavanjem ishoda sportskog događaja koji znatno utječu na tržište klađenja odnosno na ponuđene koeficijente. Kladioničari i neinformirani klijenti dijele vjerovanja $\mathbf{p} = (p_1, p_2, ..., p_n)$, dok insajderi znaju ishod događaja prije nego što je on započeo. U tom modelu kladioničarske kuće žele ograničiti izloženost insajderima, osobito u događajima s niskom vjerojatnošću realizacije, a visokom isplatom. To čine smanjivanjem koeficijenata koji su ponuđeni na te događaje. U Shinovom modelu kladioničarske kuće postavljaju koeficijente kako bi maksimizirali profit znajući unaprijed da moraju isplatiti insajdere.

Insajderi čine z udio populacije što predstavlja mjeru učestalosti trgovanja povlaštenim informacijama (engl. *insider trading*)*. Dakle, $z = 0$ ukazuje na odsustvo insajdera, a $z > 0$ označava odstupanje cijena od pravih vjerojatnosti zbog trgovanja povlaštenim informacijama.

Bez gubitka općenitosti možemo pretpostaviti da je ukupna vrijednost oklada 1, z dolazi od insajdera, a $1 - z$ od ostalih (neinformiranih) klijenata. Ako se ishod i realizira, količina uplata na i-ti ishod iznosi:

$$p_i(1-z)+z. \tag{2.4}$$

Ako kladioničar kotira s $k_i = 1/\pi_i$ za ishod i, tada on za događaj i snosi odgovornost (engl. *liability*)*:

$$\frac{1}{\pi_i}\left(p_i(1-z)+z\right). \tag{2.5}$$

Uz pretpostavku da kladioničar u svaki ishod vjeruje s vjerojatnošću $\mathbf{p}$ dobije se očekivana odgovornost kladioničara (*Napomena*: Očekivanje slučajne varijable definirano je kao zbroj umnožaka svake vrijednosti s pripadajućom vjerojatnošću):

$$\sum_{i=1}^{n} \frac{p_i}{\pi_i}\left(p_i(1-z)+z\right), \tag{2.6}$$

a ukupni profit koji kladioničar očekuje za taj događaj računa se formulom:

$$T(\pi) = 1 - \sum_{i=1}^{n} \frac{p_i}{\pi_i}\left(p_i(1-z)+z\right). \tag{2.7}$$

Kladionica postavlja ograničenje na stvarnu sumu Π takvu da vrijedi $0 <= \Pi < \Pi_{max}$, pri čemu je Π_{max} tržišno ograničenje stvarne sume koje je potrebno zadovoljiti kako bi kladionica ostala konkurentna. Uz navedeno Shin dolazi do rješenja (više detalja u [66]):

$$\pi_i = \sqrt{zp_i + (1-z)p_i^2} \sum_{j=1}^{n} \sqrt{zp_j + (1-z)p_j^2}. \tag{2.8}$$

U članku [67] je pokazano da se formula 2.8 može invertirati te se dobije izraz za takozvanu Shinovu vjerojatnost:

*Ključno obilježje trgovanja na temelju povlaštenih informacija je stjecanje nepravedne prednosti na temelju povlaštenih informacija na štetu trećih osoba koje ne raspolažu tim informacijama, što narušava integritet financijskih tržišta i povjerenje ulagatelja.

*Odgovornost u klađenju definira se kao novčani iznos potreban za pokriće ishoda oklade. Prilikom klađenja postoje dvije vrste odgovornosti - klijent koji se kladi na ishod događaja i kladioničar koji postavlja koeficijente i nada se gubitku klijenata. Klijent ne može postaviti okladu prije nego što uplati novac, a kladioničar unaprijed mora imati sredstva za isplatu u slučaju dobitka klijenta.

$$p_i = \frac{\sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\Pi}} - z}{2(1-z)}. \tag{2.9}$$

Konačno preostaje samo formula za izračunavanje udjela insajdera z. Uz korištenje uvjeta $\sum_{i=1}^{n} p_i = 1$ dobije se:

$$z = \frac{\sum_{i=1}^{n} \sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\Pi}} - 2}{n-2}. \tag{2.10}$$

Izraz 2.10 se može riješiti iteriranjem s fiksnom točkom počevši od $z_0 = 0$. Međutim u slučaju da događaj ima samo dva moguća ishoda ($n = 2$) kao što je to u slučaju predviđanja pobjednika odbojkaške utakmice moguće je koristiti izraz koji ima analitičko rješenje:

$$z = \frac{(\pi_+ - 1)\left(\pi_-^2 - \pi_+\right)}{\pi_+\left(\pi_-^2 - 1\right)}. \tag{2.11}$$

U formuli 2.11 vrijedi $\pi_+ = \pi_1 + \pi_2$ i $\pi_- = \pi_1 - \pi_2$. Dakle, π_+ predstavlja zbroj inverznih vrijednosti koeficijenata, a π_- predstavlja njihovu razliku.

Primjer primjene Shinove normalizacije (primjer preuzet iz [12]): Neka je dan koeficijent $k_1 = 1,05$ na pobjedu momčadi A odnosno koeficijent $k_2 = 9$ na pobjedu momčadi B. Obrnute vrijednosti tih koeficijenata su vrijednosti $\pi_1 = 1/k_1 = 0,95$ i $\pi_2 = 1/k_2 = 0,11$. S obzirom na to da događaj ima samo 2 ishoda koristi se formula 2.11 kako bi se izračunao udio insajdera z. On u ovom primjeru iznosi $z = 0.07267113$. Vjerojatnost na pobjedu momčadi A računa se koristeći formulaciju Shinove normalizacije (2.9) i iznosi $92,1\%$ dok vjerojatnost na pobjedu momčadi B izračunata istim pristupom iznosi $7,9\%$.

Poglavlje 3

Osnovni koncepti teorije vjerojatnosti i Monte Carlo simulacija

U ovom poglavlju dana je teoretska osnova potrebna za razumijevanje ostatka doktorske disertacije.

3.1 Slučajna varijabla

Slučajni eksperiment je proces kojim promatramo nešto nesigurno. Nakon što se eksperiment izvrši rezultat slučajnog eksperimenta je poznat. Rezultat slučajnog eksperimenta naziva se ishod, a skup svih mogućih ishoda naziva se prostor uzoraka (engl. *sample space*), a najčešće se označava slovom S [24]. Primjer slučajnog eksperimenta je bacanje novčića. U tom slučaju prostor uzoraka je S = pismo, glava. Slučajni eksperiment može se ponavljati više puta, a svako ponavljanje eksperimenta naziva se pokus. Drugim riječima, pokus je određena izvedba slučajnog eksperimenta [24]. U primjeru bacanja novčića svaki pokus rezultirat će pismom ili glavom. Bitno je napomenut da se prostor uzoraka definira na temelju slučajnog eksperimenta. U slučaju da bacamo novčić 3 puta i promatramo sekvence pisma (P)/glave (G) prostor uzoraka je: S = (G, G, G), (G, G, P), (G, P, G), (P, G, G), (G, P, P), (P, G, P), (P, P, G), (P, P, P). Definirajmo još pojam događaja (engl. *event*). Radi se o podskupu prostora uzoraka kojem pridjeljujemo određenu vjerojatnost, a najčešće se označava slovom E [24]. Primjerice, želimo procijeniti vjerojatnost da će kocka pasti na parni broj odnosno vjerojatnost događaja $E = \{2, 4, 6\}$. Ako rezultat slučajnog eksperimenta poprima vrijednost iz skupa E kaže se da se događaj realizirao [24].

U analizi slučajnog eksperimenta uobičajeno je fokusirati se na numeričke aspekte eksperimenta. Na primjer, u nogometnoj utakmici moguće je promatrati broj golova, udaraca na gol, udaraca iz kuta, prekršaja i slično. Promatramo li nogometnu utakmicu kao slučajni eksperiment svaki od ovih numeričkih rezultata daje informacije o ishodu slučajnog eksperimenta.

Upravo to su primjeri slučajnih varijabli. Drugim riječima, slučajna varijabla je varijabla čija je vrijednost određena slučajnim eksperimentom, a često se u literaturi označava slovom X. Formalnije, slučajna varijabla je funkcija koja dodjeljuje numeričku vrijednost svakom mogućem ishodu slučajnog eksperimenta [24].

Definicija 1. Slučajna varijabla X je funkcija iz prostora uzoraka u realni broj [24].

$$X : S \to \mathbb{R} \tag{3.1}$$

Primjer slučajne varijable (primjer preuzet iz [24]): Pretpostavimo da bacamo novčić 3 puta i analiziramo sekvence pismo/glava. Pretpostavimo da smo u eksperimentu zainteresirani za broj pojavljivanja glava. Moguće je definirati slučajnu varijablu X čija je vrijednost broj glava u sekvenci. Vrijednost slučajne varijable X u tom slučaju može biti 0, 1, 2 ili 3, ovisno o ishodu slučajnog eksperimenta. Tako će za ishod {G, G, G}, slučajna varijabla poprimiti vrijednost 3. Skup mogućih vrijednosti slučajne varijable X naziva se raspon slučajne varijable i označava s R_x [24]. U prethodnom primjeru $R_x = \{0, 1, 2, 3\}$.

Važno je razlikovati dvije vrste slučajnih varijabli, diskretne i kontinuirane slučajne varijable.

3.1.1 Diskretne slučajne varijable

Definicija 2. Slučajna varijabla X je diskretna ako je raspon slučajne varijable prebrojiv skup [24].

Definicija 3. Skup A je prebrojiv ako:

1. je skup konačan, $|A| < \infty$ ili
2. može se staviti u korespondenciju jedan na jedan s prirodnim brojevima $\mathbb{N}$. U tom se slučaju kaže da je skup prebrojivo beskonačan (engl. *countably infinite*) [24].

Za diskretnu slučajnu varijablu često nas zanima vjerojatnost da je $X = x_k$ odnosno da diskretna slučajna varijabla poprimi specifičnu vrijednost. Potrebno je imati na umu da je ovdje događaj $A = \{X = x_k\}$ definiran kao skup ishoda s u prostoru uzoraka S za koji je vrijednost X jednaka x_k, odnosno [24]:

$$A = \{\mathrm{s} \in \mathrm{S} \mid \mathrm{X(s)} = \mathrm{x_k}\}. \tag{3.2}$$

Vjerojatnosti događaja $\{X = x_k\}$ se formalno prikazuju funkcijom vjerojatnosti diskretne varijable (engl. *probability mass function*) [24].

Definicija 4. Neka je X diskretna slučajna varijabla s rasponom $R_x = \{x_1, x_2, x_3, \ldots\}$ (konačan ili prebrojivo beskonačan skup). Funkcija:

$$P(x_k) = P(X = x_k), \text{ za } k = 1,2,3\ldots, \tag{3.3}$$

se zove funkcija vjerojatnosti diskretne slučajne varijable X [24].
Drugim riječima, funkcija vjerojatnosti diskretne slučajne varijable daje vjerojatnosti mogućih vrijednosti diskretne slučajne varijable - svakoj vrijednosti slučajne varijable pridružuje određenu vjerojatnost. Za diskretne slučajne varijable funkcija vjerojatnosti se također naziva i raspodjela vjerojatnosti (engl. *probability distribution*) [68].
U slučaju diskretne slučajne varijable vrijedi [68]:

$$P(x_k) >= 0,$$

$$\sum_{k=1}^{k=n} P(x_k) = 1.$$

Funkcija vjerojatnosti slučajne varijable jedan je od načina opisivanja razdiobe slučajne varijable, međutim ona se ne može definirati za kontinuirane slučajne varijable (*napomena*: kontinuirane slučajne varijable definirane su u nastavku). Drugi način opisivanja razdiobe slučajne varijable je funkcija kumulativne razdiobe (engl. *cumulative distribution function*). Njezina prednost je da se može definirati kako za kontinuirane, tako i za diskretne slučajne varijable [24].
Definicija 5. Funkcija kumulativne razdiobe slučajne varijable X je definirana kao [68]:

$$F(x) = P(X \leq x), \text{ za sve x} \in \mathbb{R}. \tag{3.4}$$

Radi se o vjerojatnosti da slučajna varijabla poprima vrijednost unutar nekog intervala, u ovom slučaju vrijednost manju ili jednaku x [24]. Za funkciju kumulativne razdiobe u slučaju diskretne slučajne varijable vrijedi [68]:

$$F(x) = P(X \leq x) = \sum_{x_k <= x} P(x_k),$$

$$0 <= F(x) <= 1,$$

$$x <= y \rightarrow F(x) <= F(y).$$

3.1.2 Kontinuirane slučajne varijable

Kontinuirane slučajne varijable, za razliku od diskretnih slučajnih varijabli koje mogu poprimiti samo prebrojiv broj mogućih vrijednosti, poprimaju raspon vrijednosti u obliku intervala ili

unije intervala koji se ne preklapaju na realnom pravcu [24]. Formalna definicija kontinuiranih slučajnih varijabli dana je u nastavku.

Definicija 6. Slučajna varijabla X s funkcijom kumulativne razdiobe $F(x)$ je kontinuirana ako je $F(x)$ kontinuirana funkcija za sve $x \in \mathbb{R}$ [24].

Za kontinuirane slučajne varijable, stoga nema smisla promatrati vjerojatnost da varijabla poprimi određenu vrijednost jer ta vjerojatnost ima infinitezimalno malenu vrijednost, odnosno vrijednost približno jednaku nuli. Za opisivanje kontinuirane slučajne varijable koristi se funkcija kumulativne razdiobe i funkcija gustoće vjerojatnosti kontinuirane slučajne varijable (engl. *probability density function*) [24, 68].

Slično kao u slučaju diskretnih slučajnih varijabli, vrijednost kumulativne funkcije raspodjele, $F(x)$, kontinuirane varijable x u nekoj točki x_0, $F(x_0)$, predstavlja vjerojatnost da varijabla x poprimi vrijednost jednaku ili manju vrijednosti x_0 [68]. Funkcija kumulativne razdiobe kontinuirane slučajne varijable ima sljedeća svojstva [68]:

$$b >= a \rightarrow F(b) >= F(a),$$

$$\lim_{x \to a} F(x) = F(a),$$

$$\lim_{x \to -\inf} F(x) = 0,$$

$$\lim_{x \to +\inf} F(x) = 1.$$

Preostalo je još definirati funkciju gustoće vjerojatnosti kontinuirane slučajne varijable. Kako bi se dobio osjećaj u to što je funkcija gustoće vjerojatnosti kontinuirane slučajne varijable, najlakše ju je definirati na sljedeći način [24]:

$$f(x) = \lim_{\Delta \to 0} \frac{P(x < X <= x + \Delta)}{\Delta}. \tag{3.5}$$

Vrijedi i sljedeće - vjerojatnost da kontinuirana slučajna varijabla poprimi vrijednost unutar segmenta $a <= X <= b$ je dana sa sljedećim integralom [24]:

$$P(a <= X <= b) = \int_a^b f(x)dx = F(b) - F(a). \tag{3.6}$$

Konačno je iz 3.5 i 3.6 lako za zaključiti da vrijedi [24]:

$$f(x) = \lim_{\Delta \to 0} \frac{F(x + \Delta) - F(x)}{\Delta}.$$

To znači da je funkcija gustoće vjerojatnosti derivacija kumulativne funkcije razdiobe [68].

Definicija 7. Razmotrimo kontinuiranu slučajnu varijablu X s kontinuiranom funkcijom kumulativne razdiobe $F(x)$. Funkcija $f(x)$ definirana kao:

$$f(x) = \frac{dF(x)}{dx} = F'(x), \quad \text{ako je } F(x) \text{ diferencijabilna na } x, \tag{3.7}$$

naziva se funkcija gustoće vjerojatnosti kontinuirane slučajne varijable X [24].
Funkcija gustoće vjerojatnosti kontinuirane slučajne varijable, $f(x)$, ima sljedeća svojstva [68]:

$$f(x) >= 0,$$

$$\int_{-\inf}^{+\inf} f(x)dx = 1^*,$$

$$\int_{x_1}^{x_2} f(x)dx = P(x_1 < x < x_2).$$

Bitno je uočiti da funkcija gustoće vjerojatnosti nije vjerojatnost već je to funkcija koja opisuje relativnu vjerojatnost da kontinuirana slučajna varijabla poprimi određenu vrijednost, odnosno to je funkcija koja određuje vjerojatnost da varijabla poprimi određenu vrijednost unutar određenog intervala [68]. Moglo bi se reći da visina funkcije gustoće (y os) pri određenoj vrijednosti x sama po sebi nema stvarnu vrijednost, a matematički udžbenici često ne navode skalu na y osi (npr. [69]).

3.1.3 Jednolika i neovisna razdioba slučajne varijable

U teoriji vjerojatnosti i statistici kolekcija slučajnih varijabli je neovisno i jednoliko distribuirana (engl. *independent and identically distributed*, skraćeno i.i.d., iid, IID) ako svaka slučajna varijabla ima istu razdiobu vjerojatnosti kao i ostale (jednolika razdioba slučajne varijable) i sve su međusobno neovisne (neovisna razdioba slučajne varijable) [22]. Detaljnije, pojam jednolike razdiobe označava da nema ukupnih trendova – razdioba ne fluktuira i sve stavke u uzorku su uzete iz iste razdiobe vjerojatnosti. Pojam nezavisne razdiobe označava da su sve stavke uzorka neovisni događaji odnosno nisu povezane ni na koji način - poznavanje vrijednosti jedne varijable ne daje informaciju o vrijednosti druge i obrnuto [22]. Formalne definicije dane su u nastavku.
Definicija 8. Diskretne slučajne varijable X_1,...,X_n su nezavisne ako za sve (x_1,...x_n) $\in \mathbb{R}^n$ vrijedi [24]:

$$P(x_1, ...x_n) = P(x_1) * ... * P(x_n). \tag{3.8}$$

*ako područje vrijednosti kontinuirane slučajne varijable x nije čitav pravac, već samo određeni interval $[a,b]$ te vrijedi da je $f(x) = 0$ za svaku vrijednost varijable izvan tog intervala, onda možemo pisati $\int_a^b f(x)dx = 1$.

Definicija 9. Kontinuirane slučajne varijable X_1,...,X_n su nezavisne ako za sve (x_1,...x_n) $\in \mathbb{R}^n$ vrijedi [24]:

$$f(x_1, ...x_n) = f(x_1) * ... * f(x_n). \tag{3.9}$$

Definicija 10. Varijable su jednoliko distribuirane ako imaju iste funkcije kumulativne razdiobe [24].

Definicija 11. Kažemo da slučajna varijabla ima jednoliku i neovisnu distribuciju ako su zadovoljene prethodne definicije (definicija 8 i 10 u slučaju diskretne slučajne varijabli ili definicija 9 i 10 u slučaju kontinuirane slučajne varijable).

Primjer jednolike i neovisne razdiobe slučajne varijable: Pretpostavimo da bacamo novčić 100 puta i dobijemo glavu 53 puta i pismo 47 pismo. Želimo ponoviti bacanje 101. put. I 101. put moguće je dobiti glavu ili pismo s vjerojatnošću 0,5. Vjerojatnost dobivanja glave ili pisma ne ovisi niti o jednom od prethodnih ishoda. Drugim riječima, prošlost ne utječe na budućnost. Dakle ishod bacanja novčića je jednoliko i neovisno distribuiran događaj. Neovisno je distribuiran jer jedan ishod ne ovisi o drugom ishodu, a jednoliko jer svaki uzorak dolazi iz iste distribucije (nema promjene distribucije kada se baca novčić). *Napomena*: jednoliko distribuirano ne znači jednako vjerojatno, tj. nije potrebno da dvije slučajne varijable imaju vjerojatnosti 0,5 svaka ili da četiri slučajne varijable imaju vjerojatnosti 0,25 svaka.

3.2 Stohastički procesi

Pojam slučajne varijable neovisan je o vremenu. Međutim, mnogi procesi čiji je ishod neizvjestan, a koji se odvijaju u vremenu zahtijevaju da se koncept slučajne varijable poopći tako da uključuje i vremensku komponentu [70]. Upravo tako dolazi se do pojma stohastičkog procesa. Radi se o sustavu koji se razvija u vremenu dok prolazi kroz slučajne fluktuacije. Takav sustav može se definirati familijom slučajnih varijabli $\{X_t\}$, pri čemu X_t mjeri aspekt sustava od interesa u trenutku t [71]. Formalnije, neka slučajna varijabla X ovisi o parametru $t \in T \subset \mathbb{R}$. Za svako vrijeme $t \in T$ određena je slučajna varijabla koju ćemo označavati s X_t ili pak s $X(t)$. Nadalje, označimo sa S skup svih vrijednosti (prostor stanja) koje mogu poprimiti slučajne varijable X_t [70].

Definicija 12. Familija slučajnih varijabli

$$X = \{X_t, t \in T\}, \tag{3.10}$$

naziva se stohastički proces X sa skupom parametara T i skupom stanja S [70].

Stohastički proces možemo shvatiti kao funkciju dviju varijabli $X : T \times \Omega \to S$. Ovdje je S skup stanja, skup unutar kojeg proces poprima vrijednosti (npr. $S \in \mathbb{Z}$, $S \in \mathbb{R}$ ili $S \in \mathbb{C}$). Za izabrano vrijeme t i elementarni događaj $\omega \in \Omega$, $X(t, \omega)$ jest realizacija procesa [70].

Primjer stohastičkog procesa (primjer preuzet iz [71]): Naka je X_t broj kupaca koji čekaju u redu u trenutku t. Kako vrijeme prolazi kupci će dolaziti i odlaziti pa će se vrijednost od X_t mijenjati. U bilo kojem trenutku t, X_t poprima jednu od vrijednosti 0, 1, 2...; a t može biti bilo koja vrijednost u podskupu $(-\infty, +\infty)$.

3.2.1 Klasifikacija stohastičkih procesa

Svaka slučajna varijabla X_n poprima vrijednosti u prostoru stanja S koji predstavlja skup svih vrijednosti za proces. Kao i kod uobičajenih slučajnih varijabli prostor stanja S može biti diskretan ili kontinuiran. Ponovimo, diskretan prostor stanja označava skup različitih mogućih ishoda koji može biti konačan ili prebrojivo beskonačan. Kontinuirani prostor stanja označava neprebrojivo beskonačan kontinuum postupno varirajućih ishoda [23]. Nadalje, proces ima skup indeksa koji redaju slučajne varijable koje čine proces. Skup indeksa obično se tumači kao vremenska varijabla koja govori u kojim trenutcima se proces od interesa mjeri. Skup indeksa također može biti diskretan ili kontinuiran. Diskretno vrijeme označava proces uzorkovan u različitim točkama dok kontinuirano vrijeme označava proces koji se stalno prati tijekom vremena. Ovo daje četiri kategorije stohastičkih procesa [23]:

1.stohasti ̌cki procesi s diskretnim vremenom i diskretnim prostorom stanja,

2.stohasti ̌cki procesi s diskretnim vremenom i kontinuiranim prostorom stanja,

3.stohasti čki procesi s kontinuiranim vremenom i diskretnim prostorom stanja,
4.stohasti čki procesi s kontinuiranim vremenom i kontinuiranim prostorom stanja [23].

Ova doktorska disertacija fokusirana je na skupinu stohastičkih procesa koji zadovoljavaju Markrovljevo svojstvo (engl. *Markov property*) odnosno svojstvo odsustva pamćenja (engl. *memoryless property*) (detaljnije objašnjeno u poglavlju 3.2.2), a na takve stohastičke procese se referira pojmom Markovljevog procesa. Ovdje se znanstvena literatura razilazi oko terminologije. Dio znanstvene literature koristi pojam Markovljevog lanca kako bi opisali Markovljeve procese s diskretnim vremenom, neovisno o tome radi li se o stohastičkom procesu s diskretnim ili kontinuiranim skupom stanja [72]. S druge strane, dio znanstvene literature rezervira pojam Markovljevog lanca za Markovljeve procese s diskretnim skupom stanja, neovisno o tome radi li se o stohastičkom procesu s diskretnom ili kontinuiranom vremenskom komponentom [73, 74]. Ova doktorska disertacije koristi upravo tu terminologiju pa pojam Markovljevog lanca koristi za Markovljeve procese s diskretnim skupom stanja te razlikuje vremensko kontinuirane i vremensko diskretne Markovljeve lance. U ostatku disertacije fokus je na vremensko diskretnim Markovljevim lancima.

3.2.2 Vremensko diskretni Markovljevi lanci

Krucijalno svojstvo ovog tipa stohastičkog procesa je takozvano Markovljevo svojstvo, odnosno svojstvo odsustva pamćenja [23]. U teoriji vjerojatnosti stohastički proces ima Markovljevo svojstvo ako buduća raspodjela vjerojatnosti procesa, za dano trenutno stanje i sva prošla stanja, ovisi samo o trenutnom stanju i niti jednom drugom prethodnom stanju. Drugim riječima, prošlost i budućnost su uvjetno neovisni s obzirom na sadašnjost [23].
Definicija 13. Neka je $(X_n) = (X_0, X_1, X_2...)$ stohastički proces u diskretnom vremenu $n = 0, 1, 2, ...$ i s diskretnim skupom stanja S. (X_n) ima Markovljevo svojstvo ako za svako vrijeme n i za svako stanje $x_0, x_1, ..., x_n, x_{n+1} \in S$ vrijedi [23]:

$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n). \tag{3.11}$$

Markovljev lanac predstavlja vjerojatnosni model, a sastoji se od skupa stanja i prijelaza među njima. U svakom trenutku sustav može prijeći u neko novo stanje ili može ostati u istom stanju. Svako stanje modelirano takvim lancem je dosezljivo pa makar s jako malom vjerojatnošću. Promjene stanja nazivaju se tranzicijama ili prijelazima [2, 12]. Kako bi definirali Markovljev lanac potrebno je odrediti inicijalno stanje i vjerojatnosti prijelaza. Inicijalno stanje može se odrediti nasumično po određenoj razdiobi $\lambda_i = P(X_0 = i)$. Zbog Markovljevog svojstva, vjerojatnosti prijelaza moraju ovisiti samo o trenutnom stanju, a ne o cijeloj povijesti,

dakle $P(X_{n+1} = j|X_n = i)$. Markovljev proces sa svojstvom da vjerojatnosti prijelaza ostaju konstantne tijekom vremena (ne ovise o n) je vremensko homogen [23].

Definicija 14. Neka je λ_i razdioba vjerojatnosti na prostoru stanja S. Neka je p_{ij} (pri čemu je $i, j \in S$) takav da je $p_{ij} >= 0$ za sve i, j i neka je $\sum_j p_{ij} = 1$ za sve i. Neka je vremenski indeks $n = 0, 1, 2, ...$ Tada je diskretni Markovljev proces odnosno Markovljev lanac s homogenim vremenom (X_n) s inicijalnom razdiobom (λ_i) i vjerojatnostima prijelaza (p_{ij}) definiran kao [23]:

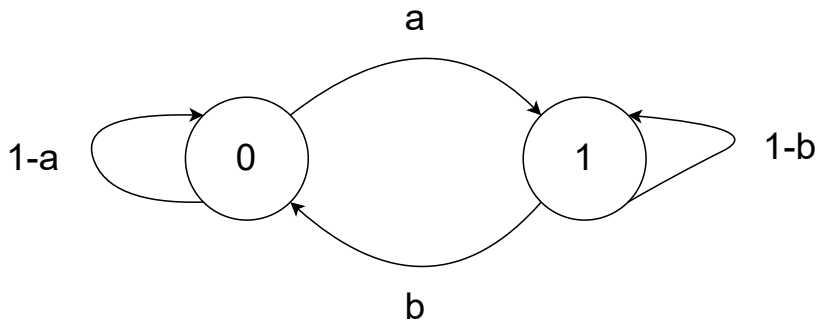
$$\begin{aligned} &P(X_0 = i) = \lambda_i, \\ &P(X_{n+1} = j|X_n = i, X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_{n+1} = j|X_n = i) = p_{ij}. \end{aligned} \tag{3.12}$$

Kada je prostor stanja konačan (a ponekad i kada nije) zgodno je zapisati vjerojatnosti tranzicije (p_{ij}) kroz matricu prijelaza P čiji je (i, j)-ti element p_{ij}. Tada je uvjet $\sum_j p_{ij} = 1$, uvjet da zbroj svakog retka matrice P jednak 1 [23]. Za prikaz Markovljevih lanaca najčešće se koriste i usmjereni grafovi i matrice prijelaza. Ako su poznate sve vjerojatnosti prijelaza među stanjima te sama topologija lanca moguće je jednoznačno odrediti vjerojatnost prijelaza između bilo koja dva stanja u lancu u određenom broju koraka [12].

Primjer vremensko diskretnog Markovljevog lanca s dva stanja (primjer preuzet iz [23]): Zamislimo vremensko diskretan Markovljev lanac sa skupom stanja S = 0, 1 i matricom prijelaza:

$$\mathrm{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

takvom da vrijedi $0 < a, b < 1$. *Napomena*: primijeti da je zbroj elemenata u istom retku jednak 1. Ovakav lanac može se prikazati i usmjerenim grafom (slika 3.1).



Slika 3.1: Usmjereni graf vremensko diskretnog Markovljevog lanca s dva stanja

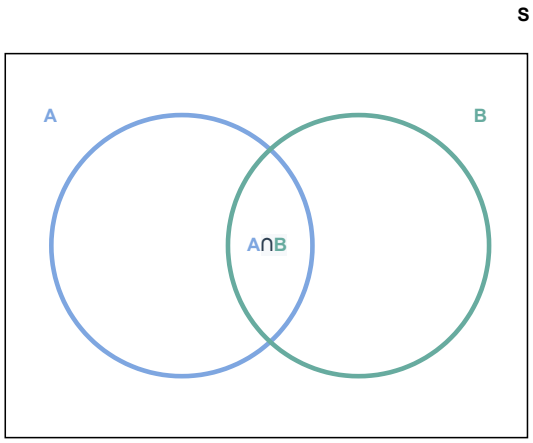
3.3 Uvjetna vjerojatnost

Kako ažurirati vjerojatnosti određenih događaja s pojavom novih informacija? Odgovor na to pitanje daje uvjetna vjerojatnost, a radi se o jednom od fundamentalnih koncepata teorije vjerojatnosti.

Definicija 15. Ako su A i B dva događaja u prostoru uzoraka S tada je uvjetna vjerojatnost od A uz B definirana kao [24]:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0. \tag{3.13}$$

Drugim riječima, ako je poznato da se dogodio događaj B, svaki ishod izvan B se odbacuje, tj. prostor uzoraka je reduciran samo na skup B. U tom slučaju jedini način na koji se događaj A može realizirati je kada ishod pripada skupu $A \cap B$. Navedeno je prikazano grafički na slici 3.2 [24].



Slika 3.2: Vennov dijagram za uvjetnu vjerojatnost

Vrijednost $P(A \cap B)$ dodatno se dijeli s $P(B)$ kako bi uvjetna vjerojatnost novog prostora uzoraka postala 1, odnosno kako bi se postiglo [24]:

$$P(B|B) = \frac{P(B \cap B)}{P(B)} = 1. \tag{3.14}$$

Primjer uvjetne vjerojatnosti (primjer preuzet iz [24]): Zamislimo da bacamo kocku. Neka je A događaj da je kocka pala na neparni broj odnosno $A = \{1,3,5\}$. Neka je B događaj da je ishod manji ili jednak 3, odnosno $B = \{1,2,3\}$. Uvjetna vjerojatnost $P(A|B)$ tada je $2/3$. Naime, ako je poznato da se dogodio događaj B ishod mora biti $\{1,2,3\}$. Kako bi se dogodio i događaj A ishod mora biti unutar skupa $A \cap B = \{1,3\}$. S obzirom na to da je svaki od tih ishoda jednako vjerojatan, tada je vjerojatnost jednaka $2/3$.

3.4 Bayesova statistika

3.4.1 Frekvencionistička i Bayesovska statistika

Dva su pristupa statističkoj analizi, frekvencionistički i Bayesovski. Osnovna razlika među njima leži u načinu na koji se interpretira pojam vjerojatnosti [25]. Kao što samo ime kaže frekvencionistička statistika, poznata i kao klasična, vjerojatnost tretira kao limes relativnih frekvencija. Iz frekvencionističke perspektive vjerojatnost predstavlja dugoročnu učestalost pojavljivanja određenog događaja [25]. S druge strane, zagovornici Bayesovskog pristupa smatraju da je vjerojatnost subjektivna odnosno da je ona stupanj uvjerenja koji se mijenja dolaskom novih informacija [25]. Dakle Bayesovski statističari vjerojatnost smatraju izrazom uvjerenja, a osnovna ideja Bayesovskog pristupa je da ažurira uvjerenja dolaskom novih informacija.
Primjer frekvencionističke i Bayesovske statistike: Razmotrimo problem bacanja pristranog novčića u kojem je potrebno pronaći vjerojatnost da će pristrani novčić pasti na glavu. Iz frekvencionističke perspektive vjerojatnost da će pristrani novčić pasti na glavu kroz neko vrijeme je dugoročna relativna učestalost pojavljivanja ishoda glava. Drugim riječima, kako se novčić više baca, broj dobivenih glava kao postotak ukupnog broja bacanja odgovara pravoj vjerojatnosti da će novčić pasti na glavu. Dakle, ovaj pristup ne uzima u obzir vlastita uvjerenja o nepristranosti drugih novčića. Prema Bayesovskom pristupu pojedinac može prije bacanja novčića vjerovati da je novčić nepristran, međutim kada se krene bacati, novčić krene učestalo padati na glavu. Kao rezultat, Bayesovski statističari modificiraju uvjerenje o nepristranosti novčića i uzimaju u obzir mogućnost da novčić možda ipak nije nepristran.

Usko vezan uz interpretaciju vjerojatnosti je pojam neizvjesnosti [25]. Pristaše klasičnog pristupa vjeruju kako je izvor neizvjesnosti isključivo realizacija slučajnih varijabli dok Bayesova statistika ide korak dalje i smatra tu slučajnost nedovoljnom. Stoga kroz parametre modela ugrađuju dodatnu neizvjesnost tako da i njih same promatra kao slučajne (za razliku od „klasičara“ koji ih uzimaju kao egzaktne) [25]. Nastavak poglavlja fokusiran je isključivo na Bayesovski pristup.

3.4.2 Bayesovsko zaključivanje

Kao što je prethodno napisano, zagovornici Bayesovskog pristupa vjerojatnost tretiraju kao stupanj uvjerenja te ju mijenjaju (“nadograđuju”) dolaskom novih informacija. Upravo to naziva se Bayesovsko zaključivanje. Navedeno se realizira koristeći Bayesov teorem koji predstavlja srž Bayesove statistike. Formalna definicija dana je u nastavku.
Definicija 16. S E označimo informacije koje imamo prije nego smo proučili podatke i pretpostavimo da subjektivnu vjerojatnost od E možemo izraziti kao $P(E)$. Bayesov teorem govori da, nakon što proučimo podatke D, vjerovanje u E se mijenja po formuli:

$$P(E|D) = \frac{P(D|E) * P(E)}{P(D)} = \frac{P(D|E) * P(E)}{\int P(E)P(D|E)dE}, \tag{3.15}$$

pri čemu je:
$P(D|E)$ uvjetna vjerojatnost danih podataka D uz uvjet da je vjerovanje u E istinito,
$P(D)$ marginalna gustoća podataka D, $P(D) > 0$ [25, 75].
Vjerojatnost događaja E prije analize podataka, $P(E)$, zovemo apriori vjerojatnost, a prilagođenu vjerojatnost, $P(E|D)$, aposteriori vjerojatnost.

U standardnim Bayesovskim metodama apriori distribucija fiksira se prije promatranja bilo kakvih podataka. Ovaj pristup je u suprotnosti s empirijskim Bayesovskim metodama. Naime, empirijske Bayesovske metode su metode za statističko zaključivanje u kojima se apriori distribucija vjerojatnosti procjenjuje iz podataka [75].

Konjugirani prior

U Bayesovoj teoriji vjerojatnosti, ako je aposteriori distribucija, $p(E|D)$, u istoj familiji vjerojatnosnih razdioba kao i apriori razdioba vjerojatnosti, $p(E)$, apriori i aposteriori distribucije se tada nazivaju konjugirane distribucije, a apriori distribucija se naziva konjugirani prior za funkciju vjerojatnosti, $p(D|E)$ [25].

Ponekad je izbor apriori distribucije vođen željom da se dobije analitički poželjna aposteriori distribucija. Ako se uoči da su podatci generirani određenom familijom distribucija, tada će nam izbor tzv. konjugirane apriori distribucije garantirati da je i aposteriori iz iste familije kao i apriori. Iako apriori i aposteriori imaju istu formu njihovi parametri se razlikuju - parametri aposteriori reflektiraju razmjenu između apriori i informacija dobivenih analizom podataka. To olakšava interpretaciju jer je moguće pratiti kako se mijenjaju parametri apriori distribucije nakon primjene Bayesovog pravila [25]. Osim toga, konjugirani prior je algebarska pogodnost koja daje izraz zatvorenog oblika za posterior. U suprotnom može biti potrebna numerička integracija [76]. Konjugirani priori korisni su jer svode Bayesovo ažuriranje na modificiranje parametara apriori distribucije (tzv. hiperparametra) umjesto izračunavanja integrala [76]. Tablica 3.1 prikazuje najčešće korištene konjugirane apriori distribucije. Tablica je preuzeta iz [25]. Potrebno je uočiti da je beta distribucija konjugirana apriori distribucija za binomnu razdiobu. Dokaz je dan u nastavku (dokaz *beta binomna konjugacija*). Prvo je potrebno definirati binomnu i beta razdiobu.

Definicija 16. Kažemo da slučajna varijabla ima binomnu distribuciju ili da je X binomna slučajna varijabla, što označavamo s $X \sim bin(n, \theta), \theta \in [0,1], n \in \mathbb{N}$, ako joj je funkcija gustoće dana s [25]:

$f(x \mid \theta)$	$\pi(\theta)$	$\pi(\theta \mid x)$
Normalna	Normalna	Normalna
$N\left(\theta, \sigma^2\right)$	$N\left(\mu, \tau^2\right)$	$N\left(\frac{\tau^2}{\tau^2+\sigma^2}X + \frac{\sigma^2}{\tau^2+\sigma^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$
Poissonova	Gamma	Gamma
$P(\theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha+X, \beta+1)$
Gamma	Gamma	Gamma
$\Gamma(v, \theta)$	$\Gamma(\alpha, \beta)$	$\Gamma(\alpha+v, \beta+X)$
Binomna	Beta	Beta
$B(n, \theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha+X, \beta-X+n)$

Tablica 3.1: Najčešće korištene konjugirane apriori razdiobe

$$P(X=x) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}, \quad x = 0,1,2,\ldots,n. \tag{3.16}$$

Radi preglednosti često se koristi i ova notacija:

$$P(X=x) =_n C_x \theta^x (1-\theta)^{n-x}, \quad x = 0,1,2,\ldots,n. \tag{3.17}$$

Binomna razdioba označava vjerojatnost određenog broja uspjeha x iz n pokušaja gdje je vjerojatnost uspjeha θ.

Definicija 17. Kažemo da slučajna varijabla X ima beta distribuciju s parametrima $\alpha > 0$ i $\beta > 0$, što označavamo s $X \sim \text{Beta}(\alpha, \beta)$, ako joj je funkcija gustoće dana s:

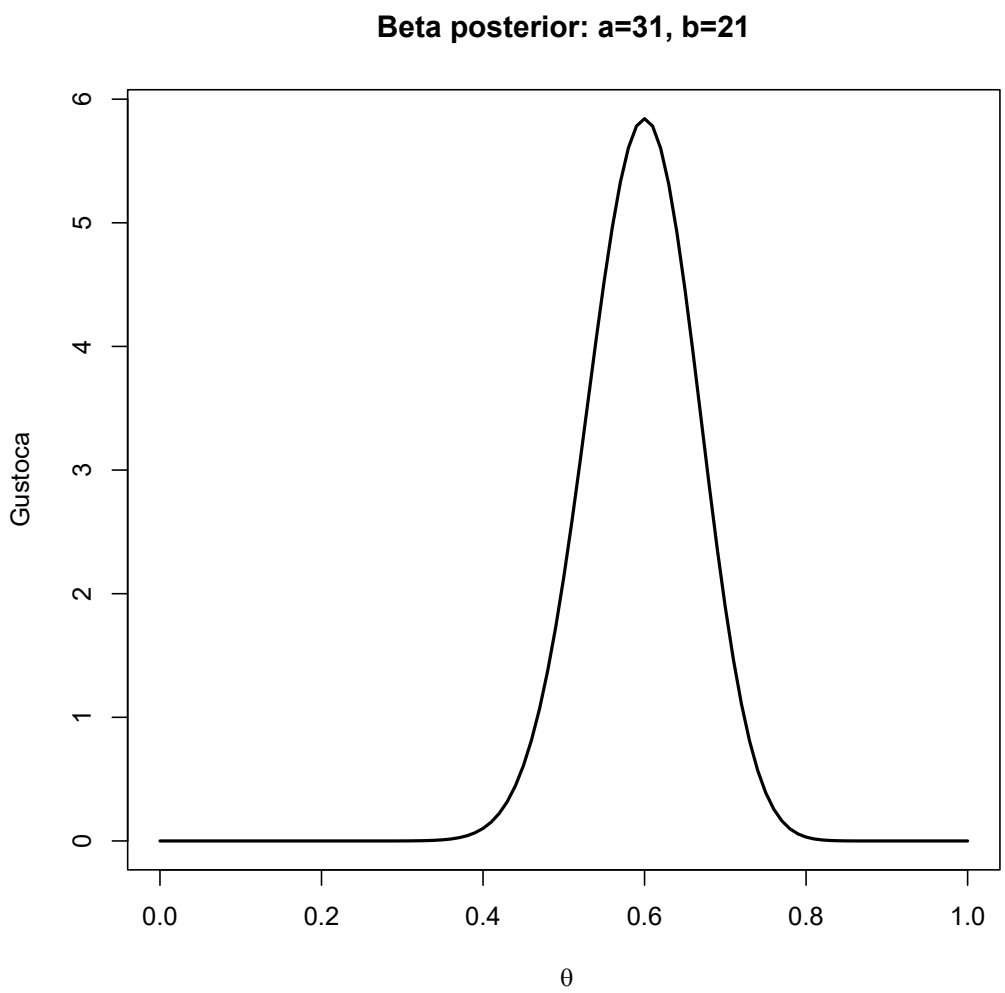
$$f(x \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1, \tag{3.18}$$

pri čemu je $B : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ beta funkcija definirana sa [25]:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx. \tag{3.19}$$

Beta distribucija naziva se i distribucijom vjerojatnosti jer je njezina domena ograničena između 0 i 1. Ona označava najvjerojatnije mogućnosti uspjeha nekog događaja. Ključna razlika između binomne i beta distribucije je u tome što je u beta distribuciji vjerojatnost x slučajna varijabla, a u binomnoj distribuciji, vjerojatnost θ je fiksni parametar. Parametri α i β kontroliraju oblik beta distribucije, a najjednostavnije ih je interpretirati tako da $\alpha - 1$ predstavlja broj uspjeha, a $\beta - 1$ broj neuspjeha u eksperimentu. Slično kao x i $n - x$ u binomnoj razdiobi.

Primjer beta razdiobe: Pretpostavimo da bacamo pristrani novčić 50 puta, a novčić padne 30 puta na glavu i 20 puta na pismo. Drugim riječima, eksperiment bilježi 30 uspjeha i 20 neuspjeha. Eksperiment se opisuje $Beta(31,21)$ distribucijom (slika 3.3). Na slici 3.3 je lako uočiti da je najvjerojatnija vjerojatnost da će ovaj novčić pasti na glavu $\theta = 0,6$ - to ima smisla s obzirom na to da je $30/50 = 0,6$. Na slici 3.3 se također da uočiti da je vjerojatnost da je parametar $\theta < 0,4$ i $\theta > 0,8$ gotovo jednaka nuli. U slučaju prikupljanja više podataka (npr. od 100 bacanja, novčić padne 60 puta na glavu), vrh beta distribucije bio bi uži jer bi se sigurnost u to da će pasti glava povećala. *Napomena*: potrebno je primijetiti da su vrijednosti na *y* osi veći od 1 jer je to funkcija gustoće vjerojatnosti (više detalja u poglavlju 3.1.2).



Slika 3.3: Graf funkcije gustoće Beta(31,21) razdiobe

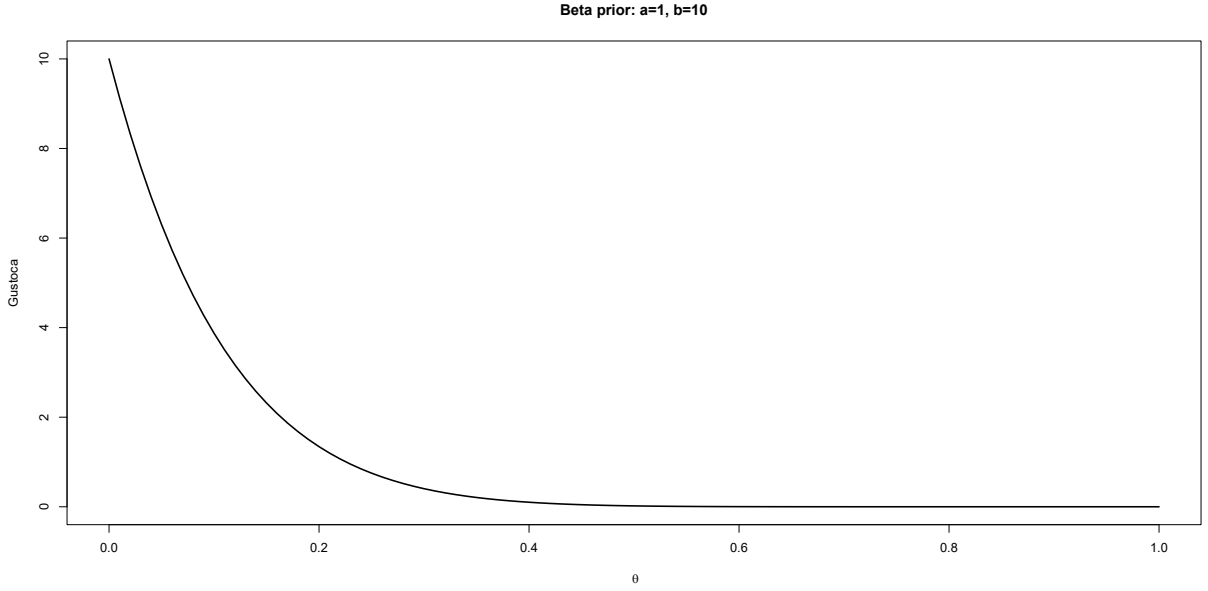
Kako parametar α postaje veći (više uspješnih događaja), vrh distribucije vjerojatnosti će se pomaknuti udesno. Povećanjem vrijednosti parametra β vrh distribucije vjerojatnosti se pomiče ulijevo (više neuspjeha). Vratimo se sad na dokaz beta binomnog konjugiranog modela.

Dokaz - beta binomna konjugacija (engl. beta binomial conjugacy): Pretpostavimo da izvodimo

n neovisnih Bernoullijevih pokusa s parametrom vjerojatnosti θ. Dobro je poznato da broj uspjeha x ima binomnu distribuciju. Uzimajući u obzir da je parametar vjerojatnosti θ nepoznat, imamo $x \mid \theta \sim \text{bin}(n,\theta)$. Funkcija gustoće vjerojatnosti dana je formulom 3.17. Pretpostavimo da je apriori distribuciju za parametar θ beta distribucija, tj. $\theta \sim \text{beta}(\alpha,\beta)$, s funkcijom gustoće vjerojatnosti $f(\theta|\alpha,\beta)$ (3.18). Ako uvrstimo funkcije gustoće vjerojatnosti u Bayesovu formulu 3.15 dobijemo [77]:

$$\begin{aligned} P(\theta|X) &= \frac{{}_nC_x\theta^x(1-\theta)^{n-x}\cdot\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 {}_nC_x\theta^x(1-\theta)^{n-x}\cdot\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &= \frac{\frac{{}_nC_x}{B(\alpha,\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\frac{{}_nC_x}{B(\alpha,\beta)}\int_0^1\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta} \\ &= \text{Beta}(x+\alpha,n-x+\beta). \end{aligned} \tag{3.20}$$

Primjer Bayesovog zaključivanja pomoću beta binomne konjugacije: Pretpostavimo da 100 učenika piše ispit iz matematike, a pitamo se koliko će učenika položiti ispit. Lako je za zaključiti da se radi o slučajnoj varijabli s binomnom razdiobom $X \sim bin(100,\theta)$ kojoj je vjerojatnost polaganja ispita θ nepoznata. Zanima nas razdioba vjerojatnosti polaganja ispita. Započinjemo sa apriori distribucijom odnosno s početnom pretpostavkom o parametru θ. Radi se o našem početnom očekivanju o parametru θ prije opservacije pravih podataka. Pretpostavimo da je apriori distribucija $Beta(1,10)$* (slika 3.4). To znači da je naše početno vjerovanje da je vjerojatnost polaganja ispita jako mala (uoči da je beta distribucija nakošena udesno).

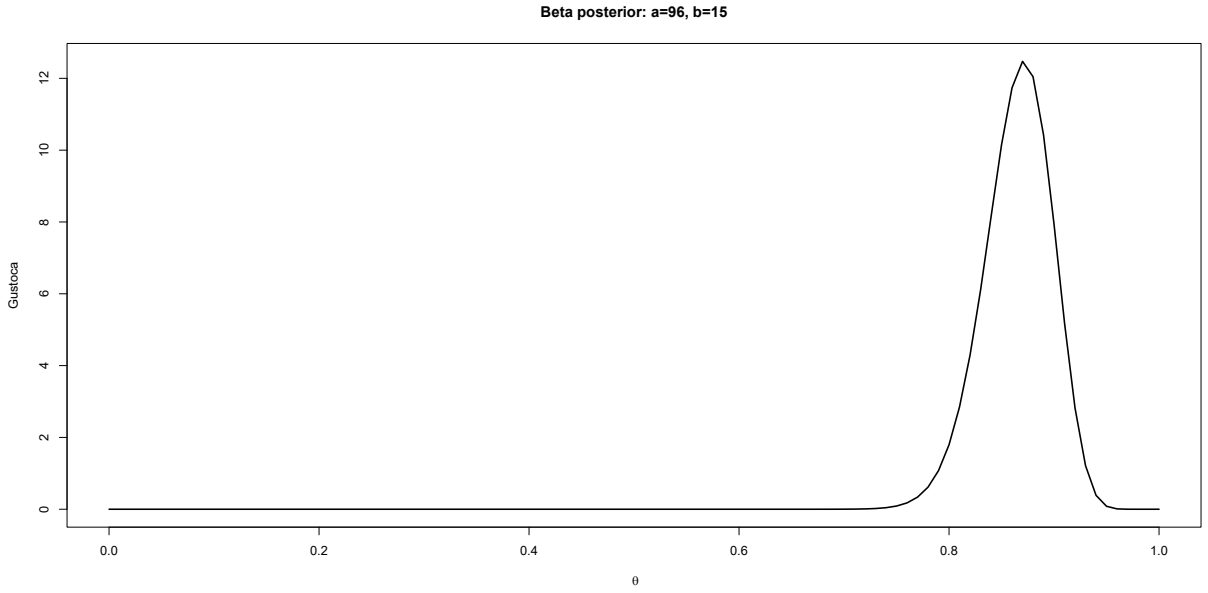


Slika 3.4: Graf funkcije gustoće Beta(1,10) razdiobe

Promotrimo sada stvarne podatke. Pretpostavimo da je ispit položilo $x = 95$ od $n = 100$

*konjugirani prior binomne razdiobe je beta razdioba.

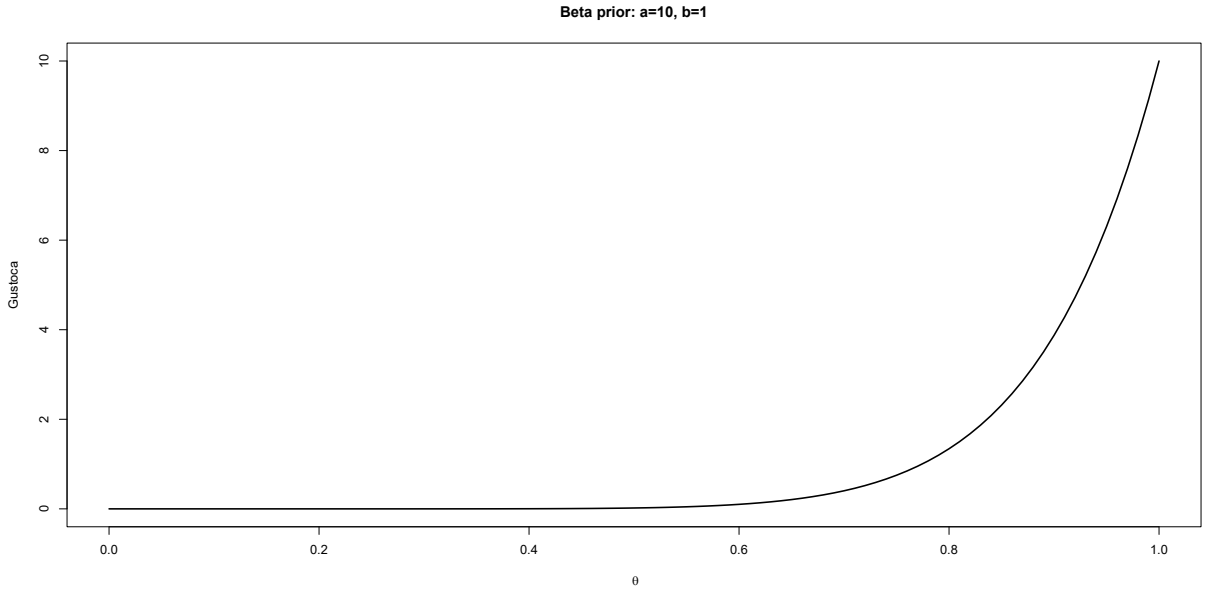
učenika. Primjeni li se Bayesov teorem, odnosno znanje o beta binomnom konjugiranom modelu, posterior tada ima distribuciju $Beta(x+\alpha, n-x+\beta)$, odnosno $Beta(96,15)$ (slika 3.5). Uočite da posterior puno više "vjeruje" stvarnim podatcima nego prior distribuciji, a to im više



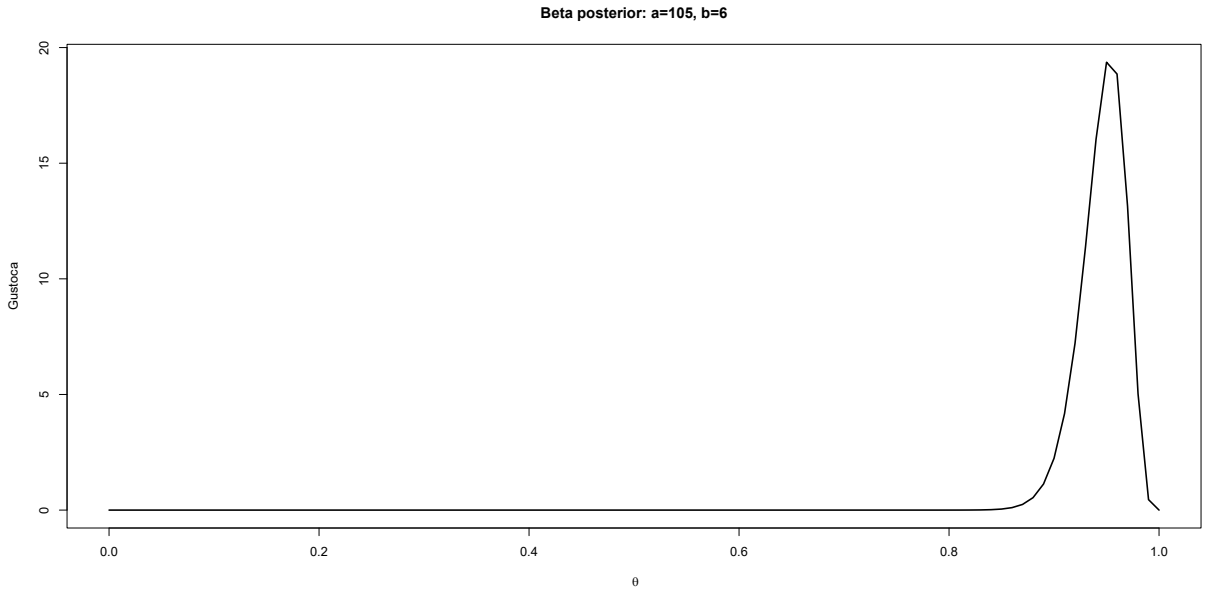
Slika 3.5: Graf funkcije gustoće Beta(96, 15) razdiobe

"vjeruje" što ima više podataka.

Zamislimo sada da su naša početna očekivanja za parametar θ bila puno veća, tj. vjerovali smo u početku da je vjerojatnost polaganja ispita velika te da se ravnala po distribuciji $Beta(10,1)$ (slika 3.6, uoči da je beta distribucija nakošena u lijevo). Ako je ispit položilo 95 od 100 učenika tada je posterior $Beta(105,6)$ (slika 3.7). Lako je za primijetiti da je posterior sada još više nakošen u lijevo.



Slika 3.6: Graf funkcije gustoće Beta(10,1) razdiobe



Slika 3.7: Graf funkcije gustoće Beta(105,6) razdiobe

3.5 Monte Carlo simulacija

Monte Carlo simulacija (ili metoda) je vjerojatnosna numerička tehnika koja se koristi kako bi se procijenio ishod stohastičkog procesa. Radi se o metodi simuliranja događaja koje je teško izravno modelirati, a koristi se za dobivanje numeričkih rješenja problema koje je jako teško riješiti analitički. Monte Carlo simulacija vrlo je intuitivan koncept, a vjerojatnost događaja interpretira kao dugoročni udio realiziranih događaja u ponovljenim pokusima [26]. Navedeno je teoretski opravdano jakim zakonom velikih brojeva.

Definicija 18. Razmotrimo uzastopna pojavljivanja pokusa slučajnog eksperimenta. Definirajmo sekvencu $X_1, X_2, ...$ pri čemu je:

$$X_k = \begin{cases} 1, & \text{ako se } A \text{ pojavi u } k\text{-tom pokusu} \\ 0, & \text{ako se } A \text{ ne pojavi u } k\text{-tom pokusu,} \end{cases} \tag{3.21}$$

za $k >= 0$. Tada je $(X_1 + X_2 + ...X_n)/n$ udio pojavljivanja događaja A u n pokusa. X_k je jednoliko distribuirana varijabla sa srednjom vrijednosti $E(X_k) = P(A)$. Prema jakom zakonu velikih brojeva vrijedi [26]:

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = P(A), \text{ sa vjerojatnošću 1.} \tag{3.22}$$

Za velike n, Monte Carlo procjena za $P(A)$ je [26]:

$$P(A) \approx \frac{X_1 + \cdots + X_n}{n}. \tag{3.23}$$

Monte Carlo simulacija koristi slučajno uzorkovanje i statističko modeliranje za procjenu matematičkih funkcija i oponašanje rada složenih sustava. Tri su koraka koja slijedi svaka Monte Carlo simulacija:

1.modeliranje sustava kao niz funkcija gusto će vjerojatnosti,

2.opetovano uzorkovanje iz funkcija gusto će vjerojatnosti,

3.izra čunavanje statistike od interesa [78].

Detaljnije, Monte Carlo simulacija zahtjeva poznavanje samo veze između ulaznih varijabli $(x_1, \ldots x_n)$ i izlaza (y):

$$y = f(x_1, \ldots x_n) \tag{3.24}$$

te poznavanje vjerojatnosne razdiobe ulaznih varijabli [79]. Metoda se bazira na brojnim ponavljanjima pokusa s računalno generiranim slučajnim brojevima i relevantnim matematičkim operacijama. U svakom pokusu ulaznim varijablama, $x_1, \ldots x_n$, dodjeljuje se nasumična vrijednost, ali takva da njihove razdiobe odgovaraju vjerojatnosnoj razdiobi te varijable. Izlazna

varijabla tada se računa koristeći formula 3.24. Izlazne varijable pokusa mogu se tada iskoristiti kako bi se konstruirao histogram izlaznih vrijednosti (y) koji predstavlja distribuciju izlazne varijable. Isti se može iskoristiti za računanje različitih statističkih veličina [79].

Primjer Monte Carlo simulacije: Razmotrimo jednostavan primjer bacanja dvaju kockica. Želimo odrediti vjerojatnost da će zbroj brojeva na kockicama biti 7. Postoji 36 mogućih kombinacija, a samo 6 od tih kombinacija zadovoljava prethodni uvjet. To znači da matematičku vjerojatnost da će zbroj brojeva na kockicama biti 7 izračunavamo kao $6/36$, odnosno 16,67 posto. Vjerojatnost je moguće aproksimirati bacanjem kockice n puta i bilježenjem ishoda bacanja. Pretpostavimo da smo to i učinili i dobili jednu od poželjnih kombinacija 17 od 100 puta. Monte Carlo simulacija je matematički prikaz ovog procesa. Omogućuje simuliranje fizičkog bacanja kockica. Svako bacanje kocke predstavlja jednu iteraciju u ukupnoj simulaciji, a s povećavanjem broja ponavljanja, rezultati simulacije postaju sve točniji. Za svaku iteraciju, varijabilni ulazi generiraju se nasumično. Rezultati simulacije daju statistički prikaz gore opisanog fizičkog eksperimenta.

Poglavlje 4

Vezana istraživanja

Slijedi poglavlje u kojem je dan pregled postojećih istraživanja u domeni modeliranja i analize teniskih mečeva i odbojkaških utakmica. Poglavlje započinje opisivanjem znanstvene literature koja je fokusirana na predviđanje ishoda sportskih događaja. Zatim slijedi potpoglavlje fokusirano na izračun ulaznih parametara koji se koriste u modelima za predviđanje sportskih događaja. Konačno je opisana problematika psihološkog zamaha u sportskoj domeni.

4.1 Prediktivno modeliranje i analiza sportskih događaja

Statističko modeliranje sportskih podataka posljednjih je godina postalo sve popularnije, a predložene su različite vrste modela, od onih za identificiranje ključnih karakteristika koje dovode igrače i momčadi do pobjede [80, 81], do modela za predviđanje ishoda sportskih događaja. Međutim, iako se radi o vrlo popularnom sportu, [82] ističe da je manjak znanstvenih istraživanja koja se bave analizom odbojkaških utakmica i modelima za predviđanje njihova ishoda. Tu pak, od sportova slične prirode, literaturom dominira prediktivno modeliranje i analiza teniskih mečeva.

Članak [29] daje detaljan pregled modela za predviđanje ishoda teniskih mečeva, a modele klasificira u tri kategorije, modele temeljene na poenima (engl. *point-based models*), modele parne usporedbe (*paired comparison models*) i modele strojnog učenja. Iste kategorije modela mogu se identificirati i u modeliranju odbojkaških utakmica, a u nastavku je fokus na prve dvije kategorije modela. Radi se o najcitiranijim metodama kada se istražuje literatura koja je fokusirana na predviđanje ishoda sportova s bodovnim ograničenjem osobito u slučaju tenisa. *Napomena*: Radi ograničenosti podatkovnog skupa korištenog u ovoj doktorskoj disertaciji (više u poglavlju 6), modeli strojnog učenja nisu fokus ove doktorske disertacije.

4.1.1 Modeli temeljeni na poenima

Modeli temeljeni na poenima baziraju se na procjeni vjerojatnosti osvajanja poena na vlastitom servisu igrača odnosno momčadi, a one se, skupa sa pretpostavkom o jednolikoj i neovisnoj distribuciji poena, koriste za predviđanje ishoda gema, seta i meča u slučaju predviđanja ishoda teniskog događaja odnosno seta i utakmice u slučaju predviđanja ishoda odbojkaškog događaja [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. Navedena pretpostavka o jednolikoj i neovisnoj distribuciji poena omogućuje korištenje vremensko diskretnih stohastičkih procesa poput Markovljevih lanaca za modeliranje ishoda takvih sportskih događaja. U ovoj kategoriji modela znanstvena literatura razdvaja dvije potkategorije, modele za predviđanje ishoda sportskog događaja prije nego što je događaj započeo i modele za predviđanje ishoda za vrijeme sportskog događaja [29]. Način modeliranja utakmica koristeći model Markovljevog lanca detaljno je opisan u poglavlju 5.

Već 1970-ih su predloženi prvi modeli za predviđanje ishoda gema i seta prije nego što je započeo teniski meč [3, 13] te modeli za procjenu ukupnog broja poena koji se očekuju u teniskom gemu ili setu [3]. Nedugo nakon toga objavljeno je još nekoliko znanstvenih članaka koji se bave prediktivnim modeliranjem teniskih mečeva prije nego što je meč započeo, a svi, koristeći drugačije pristupe, nude ekvivalentne hijerarhijske izraze za proračun vjerojatnosti osvajanja svake razine teniskog meča kao i izraze za izračun očekivanog broja poena [4, 14, 15, 16]. Modeli za predviđanje ishoda teniskog meča uživo, također su detaljno obrađeni u znanstvenoj literaturi. [17] se smatra prvim modelom za predviđanje ishoda teniskog meča uživo. Isti autor analizirao je vjerojatnosti osvajanja gema ovisno o rezultatu u gemu. Nakon toga, ova domena istraživanja postala je izrazio popularna. [5, 6] su ponudili rekurzivne formule za izračun vjerojatnosti osvajanja i trajanja svake razine teniskog meča uživo. Zbog računarske složenosti rekurzivnih izraza, [83] je predložio analitičko rješenje za izračun tih vjerojatnosti. Nakon toga ponuđeni su i rekurzivni izrazi za računanje vjerojatnosti na svaki mogući ishod gema, seta ili meča uživo [6, 8]. Konačno su ponuđene i kombinatoričke formule za rješavanje istog problema [9].

Iako se istraživanja u domeni tenisa provode još od 1970. godine, prekretnica u analizi odbojkaških utakmica dogodila se tek 2004. godine. Objavljen je znanstveni članak [84] u kojem su autori izveli izraze za računanje vjerojatnosti osvajanja odbojkaškog seta prije nego što je utakmica započela, a matematički izrazi su temeljeni na Legendrovim polinomima. Autori su također pokazali koliko je bitna informacija o tome tko prvi servira u odlučujućem setu u odbojkaškoj utakmici. Nedugo nakon toga, autori članka [10] prvi su dali izraze za računanje vjerojatnosti osvajanja i trajanja odbojkaškog seta i meča uživo koristeći Markovljeve lance i rekurzivni pristup. Lako je za uočiti da se radi o pristupu izrazito popularnom u modeliranju teniskih mečeva. Ferrante i Fonesca [11] preinačili su prethodno spomenuti članak s kombinatoričkog stajališta, a pri tome su napravili razliku između dva sustava bodovanja u odbojci. Izrazi

ponuđeni u tom članku mogu se koristiti za izračun vjerojatnosti za osvajanjem seta i utakmice prije nego što je utakmica započela. Konačno je [12] ponudio kombinatoričke izraze za izračun vjerojatni osvajanja seta i utakmice uživo. Svi ovi modeli temelje se na samo dva parametra, a radi se o vjerojatnostima osvajanja poena na vlastitom servisu igrača odnosno momčadi, što je dovelo do novog bitnog područja istraživanja - pronalazak optimalne metode za izračun istih.

4.1.2 Modeli temeljeni na parnoj usporedbi

Parna usporedba odnosi se na uspoređivanje entiteta u paru kako bi se odredio bolji entitet [28]. Doktorska disertacija ukratko opisuje dva često korištena modela u ovoj skupini modela, radi se o Bradley-Terry modelu i o ELO modelu.

Bradley-Terry model

Vrlo popularan model parne usporedbe je takozvani *Bradley-Terry* model, prezentiran 1952. godine u [85]. Postoji nekoliko varijacija Bradley-Terry modela [86], a intuicije radi, u nastavku je opisan standardni model. Opis Bradley-Terry modela napisan je na temelju [86, 87]

Pretpostavimo da se održava natjecanje u kojem sudjeluje K individua. Neka su i i j bilo koje dvije individue u tom skupu:

$$(i, j \in \{1,2,...,K\}; i \neq j).$$

Neka je svakoj individui dodijeljena realna pozitivna vrijednost koja predstavlja njezinu razinu sposobnosti ili vještine:

$$\{p_i, p_j, ..., p_K\} \in \mathbb{R}^+.$$

To znači da je prirodna logaritamska vrijednost razine sposobnosti ili vještine individue i dana kao:

$$\beta_i = ln(p_i), i = 1,2,...,,K.$$

Bradley-Terry model definira prirodnu logaritamsku vrijednost omjera šansi (engl. *log-odds*) da će individua i nadjačati individuu j, p_{ij}, kao:

$$logit(p_{ij}) = ln(\frac{p_{ij}}{1-p_{ij}}) = \beta_i - \beta_j, \tag{4.1}$$

a ishod modelira kao nezavisnu slučajnu varijabla s Bernoullijevom razdiobom, $Bernoulli(p_{ij})$. Izraz 4.1 svodi model na model logističke regresije. Pogodnost ovakve formulacije Bradley Terry modela je mogućnost proširenja modela kako bi se uključile značajke koje mogu pomoći

procijeniti i objasniti razliku između vještina individua.

Izvlačenjem vrijednosti p_{ij} iz formule 4.1 dobije se drugačija formulacija Bradley-Terry modela koja za dani par i i j procjenjuje da je parna usporedba $i > j$ istinita kao [29, 86]:

$$P(i > j) = p_{ij} = \frac{p_i}{p_i + p_j}. \tag{4.2}$$

Primjenjuje li se model u sportskoj domeni, tada parametri p_i i p_j predstavljaju sposobnosti igrača ili momčadi (p_i na primjer može biti procijenjen prema broju pobjeda igrača i u utakmicama određenog turnira), a ovisno o kontekstu, u formuli 4.2, usporedba $i > j$ može imati različita značenja. Tako primjerice može označavati da igrač i ima prednost u odnosu na igrača j ili pak da je igrač i bolje rangiran od igrača j ili da igrač i pobjeđuje igrača j [29]. Bitno je za naglasiti da kao i parametri vjerojatnosti osvajanja poena na vlastitom servisu u modelima temeljenim na bodovima, tako se i parametri koji opisuju sposobnosti igrača u standardnim Bradley Terry modelima izračunavaju iz povijesnih podataka i drže se konstantnima tijekom određenog fiksnog razdoblja [29, 86]. Konačno ostaje pitanje procjene parametara koji opisuju sposobnosti individue i, p_i. U nastavku je opisan algoritam izračuna parametra p_i u osnovnoj verziji Bradley Terry modela. Opis algoritma napisan je na temelju [88].

Pretpostavimo da promatramo parove između n igrača ili momčadi, a ideja je procijeniti parametre p_i, ..., p_n koristeći procjenu maksimalne vjerojatnosti (engl. *maximum likelihood estimation*). Uz pretpostavku da su ishodi različitih parova neovisni vrijedi:

$$\begin{aligned} ln(\mathbf{p}) &= \sum_{i}^{n}\sum_{j}^{n} ln(p_{ij})^{w_{ij}} \\ &= \sum_{i}^{n}\sum_{j}^{n} ln(\frac{p_i}{p_i + p_j})^{w_{ij}} \\ &= \sum_{i}^{n}\sum_{j}^{n} w_{ij} * ln(\frac{p_i}{p_i + p_j}) \\ &= \sum_{i}^{n}\sum_{j}^{n} [w_{ij} * ln(p_i) - w_{ij} * ln(p_i + p_j)], \end{aligned} \tag{4.3}$$

pri čemu je:
w_{ij} - broj koliko je puta i pobijedio j
$\mathbf{p}$ - vektor parametara $p_i, ..., p_n$.
Ako se sa W_i označi ukupan broj pobjeda igrača/momčadi i, polazeći od proizvoljnog vektora $\mathbf{p}$ algoritam iterativno provodi ažuriranje:

$$p_i' = W_i * (\sum_{j \neq i} \frac{w_{ij} + w_{ji}}{p_i + p_j})^{-}1 \tag{4.4}$$

Tablica 4.1: Rezultati međusobnih susreta momčadi

	M1	M2	M3	M4
M1	0	2	0	1
M2	3	0	5	0
M3	0	3	0	1
M4	4	0	3	0

za svaki i. Novoizračunati parametri, p'_i, tada se normaliziraju:

$$p_i = \frac{p'_i}{\sum_{j=1}^{n} p'_j}. \tag{4.5}$$

Opisani postupak procjene parametara p_i poboljšava prirodnu logaritamsku vrijednost vjerojatnosti (engl. *log-likelihood*) u svakoj iteraciji te konačno konvergira do jedinstvenog maksimuma.

Obilje je znanstvenih publikacija koji primjenjuju Bradley-Terry model kako u tenisu tako i u odbojci, a znanstvena literatura ponajviše je fokusirana na različite proračune parametara koji opisuju sposobnosti igrača odnosno momčadi, a u modele se sve češće uključuju vremenskovarijabilne parametre snaga [86, 89, 90, 91, 92].
Primjer primjene Bradley-Terry modela (primjer preuzet iz [93]): Pretpostavimo da su 4 momčadi međusobno odigrale ukupno 22 utakmice. Momčadi, njihovi protivnici i pobjede dane su tablično (tablica 4.1). Iz tablice se tako, između ostalog, da iščitati da je momčad $M1$ pobijedila momčad $M2$ 2 puta, a momčad $M2$ je pobijedila momčad $M1$ 3 puta.
Proizvoljno inicijaliziramo 4 elementa u vektoru parametra $\mathbf{p} = [p_1, p_2, p_3, p_4]$ - npr. dodijelimo vrijednosti 1 svakoj momčadi: $\mathbf{p} = [1,1,1,1]$. Proces procjene vrijednosti elemenata vektora $\mathbf{p}$ nastavlja se korištenjem formule 4.4. Uvrštavanjem odgovarajućih vrijednosti u formulu 4.4 proizlazi da je $p_1 = 0,6, p_2 = 1,231, p_3 = 0,667, p_4 = 1,556$. Slijedi normalizacija parametara (4.5). Normalizirane vrijednosti su $p_1 = 0,148, p_2 = 0,304, p_3 = 0,164, p_4 = 0,384$. Prvi korak iteracije ovime je završio. Kako bi se dobile što bolje procjene parametara, proces se ponavlja tako da se u formuli 4.4 sada koriste novoizračunate vrijednosti p'_i i p'_j umjesto p_i i p_j. Tada se ponovno provodi normalizacija. Nakon 20 koraka iteracija vrijednosti parametara u ovom primjeru su konvergirale u $\mathbf{p} = [0,139, 0,226, 0,143, 0,492]$. Konačno se može zaključiti da je momčad $M4$ najjača, momčad $M2$ je druga po snazi, a momčad $M1$ i momčad $M3$ su gotovo jednake snage, ali manje od momčadi $M2$ i $M4$. Bradley-Terryjev model omogućuje ekstrapolaciju odnosa između sve 4 momčadi, unatoč tome da neke momčadi nisu igrale jedna protiv druge.

ELO model

Osim Bradley-Terry modela, u skupinu modela parne usporedbe mogu se svrstati i ostali modeli rangiranja (engl. *ranking-based models*) [27]. Najpoznatiji modeli rangiranja u domeni tenisa su ATP rangiranje za igrače i WTA rangiranje za igračice. U domeni odbojke to je FIVB rangiranje. Međutim, kao alternativa navedenim svjetskim ljestvicama ističe se ELO rangiranje, nazvano prema svom kreatoru Aepadu Elu, koji ga je primarno stvorio u svrhe rangiranja šahista [94]. Za razliku od spomenutih svjetskih ljestvica koje dodjeljuju bodove igračima ovisno o fazi i razini turnira koje igrači dosegnu, broj ELO bodova koje igrači osvoje ili izgube određeni su ELO ocjenom njegovog protivnika. Nakon svake utakmice, pobjednička momčad uzima bodove poraženoj momčadi, a broj tih bodova određen je razlikom u rejtingu momčadi. Detaljnije, ELO rangiranje započinje dodjeljivanjem istih ELO bodova svakoj momčadi, a oni se zatim ažuriraju nakon svake odigrane utakmice. Kada igrač porazi igrača s više ELO bodova nego što ih sam ima, on od njega uzima puno više ELO bodova nego kada porazi igrača s manje ELO bodova od sebe. Ako pak igrač izgubi od igrača s manje ELO bodova, on gubi više ELO bodova nego ako izgubi od igrača s više ELO bodova. ELO bodovi se tada mogu koristiti za predviđanje pobjednika bilo koje igre s nultim zbrojem (engl. *zero-sum game*). Nulti zbroj je situacija u teoriji igara u kojoj je dobit jedne osobe jednak gubitku druge tako da je neto promjena bogatstva ili koristi jednaka nuli [95]. Igre poput šaha, tenisa ili odbojke, u kojima ima jedan pobjednik i jedan gubitnik, primjeri su igara s nultim zbrojem [95]. Opis ELO modela napisan je na temelju [29, 96, 97]

Formalno, pretpostavimo da je vrijeme diskretizirano u periode s indeksima $t = 1, ..., T$. Neka je θ_{it} procijenjena snaga momčadi i u trenutku t. Pretpostavimo da se tijekom perioda t momčad i natjecala protiv momčadi $j = 1, ..., J$ s procijenjenom snagom θ_{jt}. Neka su zadani bodovi momčadi i u trenutku t kao:

$$R_{it} = C + (\frac{400}{log10}) * \theta_{it}, \tag{4.6}$$

pri čemu je C proizvoljno odabrana konstanta (najčešće 1500).
Vjerojatnost da će igrač/momčad i pobijediti igrača/momčad j u trenutku t računa se po formuli (uoči da je prediktor razlika u bodovima igrača/momčadi):

$$We(R_{it}, R_{jt}) = \frac{1}{1 + 10^{-}(R_{it} - R_{jt})/400}. \tag{4.7}$$

Za razliku od parametara p_i i p_j u Bradley-Terry modelu i za razliku od parametara vjerojatnosti osvajanja poena na vlastitom servisu u modelima temeljenim na poenima, ELO bodovi igrača ažuriraju se nakon svake utakmice po formuli:

$$R_{i,t+1} = R_{it} + K * \sum_{j=1}^{J} (y_{ij} - We(R_{it}, R_{jt})), \tag{4.8}$$

gdje se vrijednost K može odabrati ili optimizirati kako bi odražavala vjerojatnu promjenu sposobnosti momčadi tijekom vremena, a parametar y_{ij} bilježi binarnu vrijednost ishoda događaja - parametar poprima vrijednost 1 ako je momčad i pobijedila momčad j, a 0 u suprotnom. Vrijednost K može se shvatiti kao veličina doprinosa rezultata trenutne utakmice na rejting prije utakmice. Tako visoke vrijednosti parametra K stavljaju veću težinu na rezultat trenutne utakmice, a niske vrijednosti parametra K stavljaju veći naglasak na rejting prije utakmice. U nekim implementacijama ELO modela, vrijednost K ovisi o rejtingu momčadi prije utakmice, a veće vrijednosti parametra K se postavljaju za momčadi sa slabijim rejtingom. Time se pretpostavlja da su slabije momčadi manje "stabilne" i veća je vjerojatnost da će mijenjati sposobnosti tijekom vremena.

Primjer primjene ELO modela: Pretpostavimo da utakmicu igra momčad A protiv momčadi B. Pretpostavimo također da momčad A ima 2600 ELO bodova, a momčad B ima 2300 ELO bodova. Očekivana vjerojatnost pobjede momčadi A nad momčadi B računa se po formuli 4.7 i iznosi $0,849$, a vjerojatnost pobjede momčadi B iznosi $0,151$ (također se računa koristeći izraz 4.7). Uz pretpostavku da je $K = 16$ i uz pretpostavku da je pobijedila momčad A novi broj bodova momčadi A tada iznosi 2602, a novi broj bodova momčadi B iznosi 2298 (4.8).

Izrazito je popularna implementacija ELO modela rangiranja koju su predložili novinari iz FiveThirtyEight.com. Oni su naime predložili prilagođeni K faktor u formuli 4.8. [27] je usporedio 11 dostupnih modela za predikciju ishoda teniskih mečeva, a upravo ELO model novinara pokazao se najpreciznijim. Izvorna ELO metoda koju su koristili novinari FiveThirtyEighta nije uzimala u potpunosti obzir trenutačnu formu svakog igrača i njihovu nedavnu izvedbu. Stoga su autori Angelini, Candila i De Angelis [96] predložili težinsku ELO metodu gdje se razmatra rezultat teniskog meča. Detaljnije, tzv. WELO uzima u obzir ne samo je li igrač pobijedio ili izgubio utakmicu, već i s kojim rezultatom je pobijedio ili izgubio utakmicu.

4.2 Izračun ulazne statistike

S obzirom na to da su vjerojatnosti osvajanja poena na vlastitom servisu momčadi bile jedini prediktori ishoda događaja u modelima temeljenim na poenima javila se skupina znanstvenih istraživanja koja se fokusirala na analizu i procjenu istih. Istraživanja u ovoj domeni započela su analizom utječe li odluka o tome tko prvi servira u setu na konačan rezultat utakmice. [10] tvrdi da momčad koja set započinje na prijemu ima prednost u osvajanju seta, a isto tumači tako da ona momčad koja prva prima servis ima i prva priliku za napad. Sličnu analizu autori su proveli i na razini utakmice, a tu su pokazali da na ishod utakmice ne utječe izbor želi li momčad započeti utakmicu na servisu ili na prijemu. Isto je i očekivano s obzirom na to da se na početku svakog seta momčadi na servisu izmjenjuju. Dođe li do petog seta, autor tvrdi da je u prednosti momčad koja ne započinje servirati u istom. Sličnu analizu proveli su i [84], a autori tvrde da momčad koja pobijedi u bacanju novčića u petom setu ima prednost ispravnim odabirom želi li započeti set na servisu ili ne. Momčad tako može ostvariti prednost odlukom da će započeti servirati u setu, ali samo ako je zbroj vjerojatnosti osvajanja poena na vlastitom servisu obje momčadi veća od 1. U suprotnom momčad može ostvariti prednost ako ipak odabere započeti set na prijemu. Oba istraživanja temelje se na pretpostavci da je vjerojatnost osvajanja poena na vlastitom servisu jednoliko i neovisno distribuirana.

Druga skupina znanstvenih radova fokusirala se pak na usrednjavanje povijesnih statističkih podataka ili na procjenu učinka protiv zajedničkih protivnika, a istraživanja su primarno provedena u domeni tenisa [98]. Prvi korak u procjeni vjerojatnosti osvajanja poena na vlastitom servisu je prikupljanje povijesnog podatkovnog skupa. Većina dostupnih podataka sadrži uprosječene vrijednosti po meču. Takvi podatci daju informacije o učinku igrača protiv prosječnog protivnika iz njegove prošlosti. To rješenje nije zadovoljavajuće jer je u obzir potrebno uzeti specifičnu snagu protivničkog igrača. [99] su predložili jednostavno, ali iznimno učinkovito rješenje za taj problem. Predloženo rješenje kombinira statistike igrača uzimajući u obzir prosječnu statistiku svih igrača kao i prosječnu statistiku turnira koja se pokazala kao izrazito značajan parameter skaliranja. Detaljnije, vjerojatnost osvajanja poena na vlastitom servisu igrača računa se kao ukupna vjerojatnost osvajanja poena na servisu za specifični turnir (ovime se u obzir uzima podloga terena) na koju se dodaje višak za koji igračeva vjerojatnost osvajanja poena na vlastitom servisu premašuje prosječnu (ovime se računa igračeva sposobnost servirana) te oduzima višak za koji protivnikova vjerojatnost primanja servisa premašuje prosječnu vrijednost (ovime se računa igračeva sposobnost vraćanja servisa). Ovo pokazuje da procjena točnih vjerojatnosti osvajanja poena na vlastitom servisu igrača nije toliko presudna kao mogućnost procjene razlike između vjerojatnosti osvajanja servisa između dva igrača. [100] su ponudili sličnu metodu za kombiniranje statistike igrača koristeći pretpostavku da vjerojatnost osvajanja poena na vlastitom servisu igrača i vjerojatnost vraćanja servisa igrača varira od meča

do meča i da se može modelirati kao slučajna varijabla s Gaussovom razdiobom. Predloženi model koristi 4 ulazna parametra za svakog igrača. Prva dva parametra su srednje vrijednosti osvajanja poena na vlastitom servisu igrača i srednje vrijednosni vraćanja servisa. Treći i četvrti parameter predstavljaju standardnu devijaciju oko tih srednjih vrijednosti koja mjeri igračevu dosljednost od utakmice do utakmice i od podloge do podloge. [30] su uveli metodu zajedničkih protivnika koja bi trebala dodatno poboljšati ulazne parameter ali zahtjeva posjedovanje veće količine podataka. Konkretno, ovi su autori koristili statističke podatke iz podskupa mečeva unutar predefiniranog razdoblja koje je uključivalo samo mečeve protiv protivnika s kojima su se susrela oba igrača koja se modeliraju. Cilj ovog pristupa je izračunati vjerojatnosti osvajanja poena na vlastitom i protivničkom servisu za svakog igrača uz eliminiranje pristranosti koja proizlazi iz različite kombinacije protivnika s kojima se svaki igrač susreo. Problem s ovako rigoroznim modelom je što se uvelike smanjuje skup podataka za učenje pa može doći do podnaučenosti. Autori iz tog razloga iznose i ideju o rekurzivnom pristupu ovoj strategiji, što znači da se u obzir uzmu i protivnici protivnika (ili protivnici protivnika protivnika, ovisno o dubini rekurzije). Takvo rješenje iskorištavalo bi veći udio dostupnog podatkovnog skupa pa bi se izbjegla podnaučenost. Međutim, u tom slučaju potrebno je paziti na dubinu rekurzije. Naime, korištenje ove strategije s predubokom rekurzijom postaje ekvivalentno zanemarivanju modela zajedničkih protivnika jer bi se skoro svi dostupni podaci uzeli u obzir. [98] je modificirala metodu zajedničkog protivnika tako da omogući da izračunate vjerojatnosti osvajanja poena na vlastitom servisu variraju od seta do seta. Time je djelomično odstupila od pretpostavke o jednolikoj i neovisnoj distribuciji poena. [101] je prezentirao Bayesov hijerarhijski model koji se koristi kako bi se procijenila vjerojatnost osvajanja poena na vlastitom servisu za danu podlogu, turnir i datum meča. Detaljnije, svakom igraču vještina serviranja i vraćanja servisa modelira se kao Gaussov nasumični hod tijekom vremena, a varira ovisno o podlozi i o turniru. [102] predlažu model stanja mogućih događaja koji dovode do osvajanja poena na servisu (npr. as, osvojeni poen na prvom servisu itd.) i kombiniraju vjerojatnosti tih stanja kako bi procijenili ukupnu vjerojatnosti osvajanja poena na servisu. U predloženom modelu procjene za vjerojatnosti svakog stanja su usrednjene za određeno razdoblje nedavne igre svakog igrača, a dodatno se prilagođavaju ovisno o snazi protivnika na način koji su predložili [99]. [103] su demonstrirali način kalibracije parametara vjerojatnosti osvajanja poena na vlastitom servisu prema vjerojatnostima pobjede izračunatih iz ELO rejtinga igrača. [2] je prezentirala načine kombiniranja ulazne statistike svakog meča ovisno o vremenu odigravanja povijesnih mečeva, o odstupanju od prosječne statistike igrača te o duljini meča odnosno broju odigranih poena. Konačno je [104] istaknuo dva najveća problema prethodnih metoda. Radi se o problemu ograničenog podatkovnog skupa i o problemu pristranosti podataka izazvane različitom kombinacijom protivnika (jedan od pristupa rješavanja tog problema je metoda zajedničkog protivnika, međutim ista rješava problem po cijenu smanjenja količine dostupnih podataka). U svrhe rje-

šavanja istih, autor nudi tri modifikacije postojećih metoda. Prva modificira Barnett Clarkovu metodu uvođenjem Effron Morrisovog procjenitelja. Druga u Barnett i Clarkovu metodu uvodi komponentu zajedničkog protivnika, a treća se temelji na računanju parametara vjerojatnosti osvajanja poena na vlastitom servisu iz vjerojatnosti osvajanja meča izračunate ELO metodom. Ta metoda pokazala je najbolje rezultate. Prednost tog pristupa je također što omogućuje procjenu parametara vjerojatnosti osvajanja poena na vlastitom servisu iz predviđanja koja nisu temeljena na poenima.

Razlog zašto se ovo poglavlje fokusiralo prvenstveno na domenu tenisa je ograničenost podatkovnog skupa korištenog u ovoj doktorskoj disertaciji (više u poglavlju 6). Naime, doktorska disertacija koristi podatkovni skup koji sadrži svaku izmjenu rezultata u odbojkaškoj utakmici, a pri tome ne bilježi niti akcije igrača niti postavu igrača niti rotaciju igrača. Cijela momčad promatra se kao jedan igrač što problem svodi na onaj kao u tenisu.

4.3 Psihološki zamah u sportu

Nekoliko znanstvenih istraživanja pokazuje da je područje ispitivanja kvalitete pretpostavke o jednolikoj i neovisnoj distribuciji poena u modeliranju sportskih događaja i dalje atraktivna tema znanstvenih istraživanja [1, 105]. Mnogi znanstvenici smatraju da ta pretpostavka nije dovoljno dobra zbog intuicije o psihološkom zamahu igrača. Psihološki zamah, u sportu poznatiji pod pojmom *vruća ruka* (engl. *hot hand*) ili *vrući niz* (engl. *hot streak*), definiran je kao kratkoročna iznadprosječna izvedba igrača ili momčadi nakon što se dogodi jedan ili više uzastopnih motivirajućih događaja u igri [18]. Koncept psihološkog zamaha vrlo je intuitivan, a u njega vjeruje velik udio sportskih stručnjaka i laika. Prema istraživanju [48], 91% obožavatelja košarke složilo se da je veća vjerojatnost da će igrač zabiti koš nakon što je prethodno zabio dva ili tri koša, nego kada je prethodno promašio dva ili tri koša. 68% ih smatra da, promatraju li se 2 uzastopna slobodna bacanja, igrač ima veću vjerojatnost da će zabiti drugi slobodni udarac nakon što je zabio prvi, a 84% obožavatelja smatra da je bitno dodati loptu igraču koji je prethodno zabio nekoliko koševa u nizu. Slična vjerovanja iskazali su i profesionalni košarkaši. Objavom upravo ovog znanstvenog članka započela su istraživanja na temu psihološkog zamaha. Lako je za zaključiti da i sportski treneri vjeruju u taj fenomen s obzirom na broj predaha (engl. *time-out*) koje pozovu nakon što protivnička momčad poveže nekoliko dobrih poteza, a svi koji prate sport složit će se da su barem jednom čuli kako sportski komentator ističe da se određeni igrač ili momčad "razigrala". Pri tome komentator upravo misli na fenomen psihološkog zamaha. Vrlo je teško osloboditi se dojma da će momčadi ili igrači nakon nekoliko uzastopno dobrih poteza nastaviti igrati na višoj razini od normalne, barem kratkoročno.

Ispitivanje postojanja fenomena psihološkog zamaha jako je bitna tema znanstvenih radova, a ključno pitanje tih znanstvenih radova je odstupa li opažena superiorna izvedba igrača od onoga što bi se moglo dogoditi slučajno [19]? Istraživanja se intenzivno provode od kraja dvadesetog stoljeća, a suprotno čvrstoj intuiciji većine promatrača sportskih događaja, mišljenja znanstvenika i dalje su jako podijeljena. Istraživanja u košarci [48, 49, 50, 51] i bejzbolu [52] dominiraju literaturom koja istražuje psihološki zamah, a iscrpno istraživanje provedeno je i u domeni tenisa [53, 54, 55, 56, 57, 58, 60]. Odbojka se smatra jednim od boljih „laboratorija“ za testiranje fenomena psihološkog zamaha jer mreža na sredini terena razdvaja protivničke igrače. Protivnici su manje sposobni koristiti strategije protiv razigranog igrača, za razliku od primjerice košarke gdje protivnici mogu pažljivo „čuvati“ vrućeg igrača [106]. Nekolicina je članaka upravo i fokusirana na ispitivanje postojanja psihološkog zamaha u odbojci [61, 62, 63].

[19] je opsežan članak koji izlistava i opisuje radove koji se bave istraživanjem postojanja psihološkog zamaha i daje pregled metodoloških nedostataka koji se tiču učinkovitosti testova korištenih u određenim istraživanjima. Autor je napravio opsežan pregled 24 najznačajnija znanstvena članka koja su se bavila ispitivanjem postojanja psihološkog zamaha u raznim spor-

tovima, od košarke, bejzbola, golfa, pikada, bilijara, tenisa i odbojke pa sve do kuglanja. Od 24 znanstvena članka, 13 članaka nije podupiralo pretpostavku da psihološki zamah postoji, dok je 11 znanstvenih članak podržavalo istu. Međutim, autor i dalje tvrdi da je znanstvena potpora za postojanje fenomena vruće ruke kontroverzna i prilično ograničena iz više razloga, poput korištenje nerealnog modela i upitnog podatkovnog skupa, postavljanja upitne definicije "vrućeg" i "hladnog" igrača i sličnih razloga. Autor međutim smatra da samo vjerovanje u postojanje fenomena psihološkog zamaha utječe na strategiju igre i klađenje [19]. Ovime se fokus s ispitivanja postojanja fenomena psihološkog zamaha prebacuje na analizu učinka vjerovanja u sam koncept. [63] su između ostalog, fokusirani na analizu učinka vjerovanja u fenomen vruće ruke na proces donošenja odluka u odbojci. Autori pokazuju da treneri mogu otkriti varijabilnost performansi svojih igrača i mogu ih iskoristiti za donošenje strateških odluka, a pokazuju da se igrači oslanjaju na vruće nizove kada odlučuju kome će dobaciti loptu. Slično istraživanje proveli su i [107]. [108] ističe važnost istraživanja utjecaja zablude o vrućoj ruci. Autor smatra da bi bilo zanimljivo analizirati utječe li iluzija da je igrač vruć na to da postane bolje. Mogu li lažna uvjerenja ikad dovesti do povoljnog ponašanja?

Ova doktorska disertacija je, potaknuta opisanim znanstvenim istraživanjima, između ostaloga fokusirana na analizu utjecaja kratkoročnog zamaha na modele za predviđanje sportskih događaja. Koji je uzrok igračeve iznadprosječne izvedbe? Može li uzrok biti igračeva kratkoročna neuroplastičnost odnosno sposobnost mozga da se prilagodi uvjetima u igri, ili se pak radi o boljem fokusu i mentalnoj pripremi? Dolazi li fenomena psihološkog zamaha zbog takozvanog pristupa istraživanja i odabira (engl. *explore and exploit*) koji se odnosi na kratko razdoblje istraživanja različitih pristupa rješavanja problema i odabira najboljeg od ponuđenih? Radi li se o specifičnom stilu igre igrača [109]? Iako je moguće spekulirati o tome je li igračeve kratkoročne iznadprosječne performanse odražavaju fizički učinak, psihološki ili kombinaciju, to nije moguće znati sa sigurnošću. Zato se u ostatku doktorske disertacije ne koristi pojam psihološkog zamaha već se koristi pojam kratkoročnog zamaha.

Za kraj je bitno napomenuti da istraživanja vezana uz temu psihološkog zamaha nisu zanimljiva samo u domeni sporta već se radi o vrlo zanimljivom području istraživanja i u ekonomiji [110] i u kognitivnoj znanosti [111]. Time istraživanja vezana uz temu psihološkog zamaha dodatno dobivaju na vrijednosti s obzirom na to da se teorije i ponašanja u domeni sporta mogu primijeniti i u realnim domena nesportskog karaktera.

Poglavlje 5

Modeliranje odbojkaških utakmica Markovljevim lancima

Modeli kojima se pokušava aproksimirati stanje unutar odbojkaške utakmice poznati su kao jedni od najsloženijih diskretnih stohastičkih modela [12]. Modeli odbojkaških utakmica najčešće se predstavljaju Markovljevim lancima, a bitna značajka u tim modelima je vjerojatnost osvajanja poena na vlastitom servisu momčadi. Naime, pravila i struktura bodovanja u odbojci su se kroz povijest značajno mijenjala, ali je vjerojatnost osvajanja poena na vlastitom servisu momčadi ostala glavni i presudni faktor u pokušajima da se izračuna vjerojatnost osvajanja seta ili meča [12].

Teorijska osnova za razumijevanje ovog poglavlja dana je u poglavlju 3.2.2. Međutim, bitno je napomenuti da je prilikom modeliranja odbojkaških utakmica jednostavnije na Markovljev lanac gledati kao na dinamički sustav gdje stanja zadovoljavaju rekurziju $X_n = f(X_{n-1}, Y_n)$, $n >= 1$, pri čemu su $Y_1, Y_2, ...$ jednoliko i neovisno distribuirane slučajne varijable, a f je deterministička funkcija. To znači da je novo stanje, X_n, funkcija prethodnog stanja, X_{n-1}, i slučajne varijable, Y_n [112]. Odbojkaške utakmice najčešće se modeliraju kroz hijerarhijski Markovljev lanac, jedan modelira odbojkaški set, a drugi odbojkašku utakmicu. Stanja tog Markovljevog lanca su svi mogući rezultati u setu odnosno utakmici, a tranzicije su vjerojatnosti osvajanja poena na vlastitom servisu momčadi odnosno vjerojatnosti osvajanja seta - ovisno koja razina odbojkaške utakmice se modelira. One su jednoliko i neovisno distribuirane [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Više o ovome dano je u nastavku.

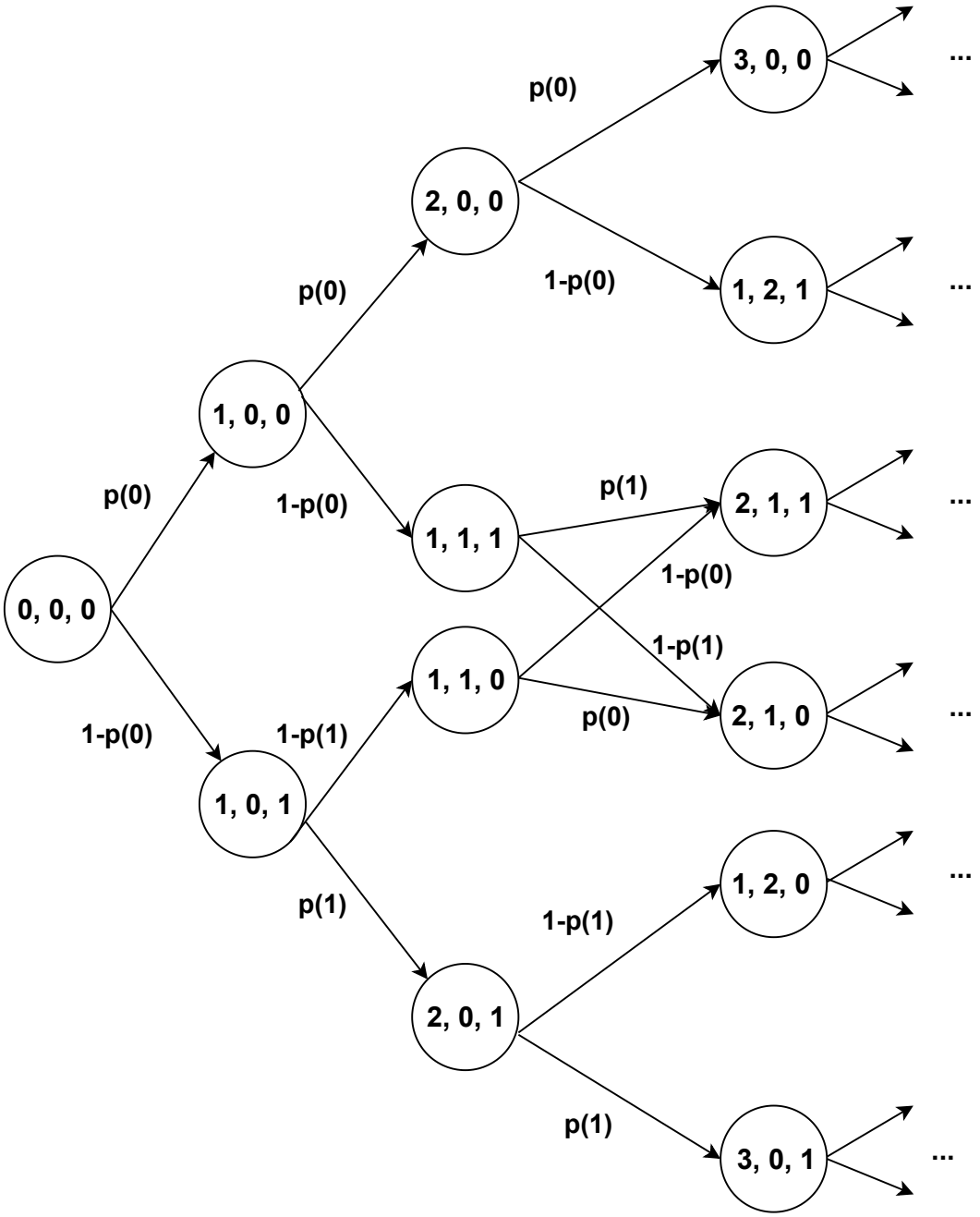
Za kraj je još potrebno prisjetiti se da u ovoj kategoriji modela znanstvena literatura razdvaja dvije potkategorije, *pre-match* modele i *in-play* modele, a u oba slučaja dva najpopularnija pristupa modeliranju sportskih događaja su rekurzivni i kombinatorički pristup. U nastavku doktorske disertacije detaljnije je opisan *pre-match* model izgrađen koristeći kombinatorički pristup. Radi se o učinkovitijem pristupu s aspekta vremena izvođenja - rekurzivni pristup naime inherentno sadrži računarsku složenost koja sprječava vremensku učinkovitost metoda

koje se oslanjaju na njih (dokaz u domeni tenisa dan je u radu [9]).

5.1 Markovljev lanac za odbojkaški set

Prilikom izgradnje bilo kojeg prediktivnog modela, kvaliteta rezultata najviše ovisi o kvaliteti ulaznih podataka. U analizi odbojkaških utakmica i teniskih mečeva vjerojatnost osvajanja poena na vlastitom servisu momčadi pokazala se kao jedna od najvažnijih značajki, a mnogi modeli u znanstvenoj literaturi grade prediktivne modele koristeći samo navedeni parametar (više u poglavljima 4.1.1 i 4.2). Ti modeli najčešće se temelje na vrlo jednostavnoj pretpostavci - vjerojatnosti osvajanja ili gubitka poena na vlastitom servisu momčadi ostaju konstantne tijekom cijele utakmice. Drugim riječima, vjerojatnost osvajanja/gubitka poena na vlastitom servisu momčadi ne ovisi o fazi utakmice niti o događajima koji su prethodili stjecanju iste. Pretpostavka o jednolikoj i neovisnoj distribuciji poena čini proces izgradnje modela relativno jednostavnim uz pretpostavku da analitičar ima pristup povijesnim podatcima suprotstavljenih momčadi. Iz njih analitičar može izračunati vjerojatnosti osvajanja poena na vlastitom servisu momčadi te ih zatim može koristiti kao vjerojatnosti tranzicija između stanja u Markovljevom lancu.

Kao i u [12], radi jednostavnosti u primjeru koji slijedi momčad domaćin označavat će se s indeksom 0, a gostujuća momčad s indeksom 1. Vjerojatnost osvajanja poena na vlastitom servisu domaće momčadi, stoga će se označavati s $p(0)$, a gostujuće s $p(1)$. Ograničenja na zadane vrijednosti su trivijalne, $0 < p(0) < 1$ i $0 < p(1) < 1$ [12]. Slika 5.1 prikazuje Markovljev lanac za prva tri poena u odbojkaškom setu. Stanja tog lanca prikazuju moguće rezultate u setu te momčad na servisu. Detaljnije, prvi broj označava broj osvojenih poena momčadi koja je na servisu, drugi broj označava broj osvojenih poena momčadi koja prima servis, a treći broj označava momčad na servisu. Ako je treći broj jednak 0 tada prvi broj označava broj osvojenih poena momčadi domaćina. Ako je pak treći broj jednak 1 tada prvi broj označava broj osvojenih poena gostujuće momčadi. Tako primjerice stanje $(1,0,0)$ označava da je trenutni rezultat $1:0$ što znači da momčad domaćin vodi jedan poen razlike i servira idući poen. Stanje $(2,0,1)$ označava rezultat $0:2$. To znači da momčad domaćin gubi 2 poena razlike od gostujuće momčadi, a gostujuća momčad servira idući poen. Inicijalno stanje prikazanog modela je stanje $(0,0,0)$ (bez gubitka općenitosti može se pretpostaviti da servira momčad domaćin). Uzmu li se u obzir prethodno opisana odbojkaška pravila iz tog stanja moguće je prijeći u stanje $(1,0,0)$ s vjerojatnošću $p(0)$, odnosno ako momčad domaćin osvoji poen na vlastitom servisu. U suprotnom model prelazi u stanje $(1,0,1)$ s vjerojatnošću $1-p(0)$. Dosegne li model stanje $(1,0,1)$ servis preuzima momčad gost. U tom slučaju model može prijeći u stanje $(1,1,0)$ ako momčad koja servira izgubi poen na vlastitom servisu. Vjerojatnost tog događaja je $1-p(1)$. Ako pak momčad osvoji poen na vlastitom servisu s vjerojatnošću $p(1)$ model prelazi u stanje



Slika 5.1: Usmjereni graf Markovljevog lanca za prva 3 poena u setu

$(2,0,1)$. Daljnji razvoj modela trivijalno je ekstrapolirati. *Napomena*: model na slici 5.1 je moguće poopćiti korištenjem notacije s i $1-s$ umjesto brojeva 0 i 1 na mjestima koji označavaju koja momčad je trenutno na servisu te korištenjem notacije $p(s)$ i $p(1-s)$ za označavanje vjerojatnosti osvajanja poena na vlastitom servisu momčadi domaćina i gostujuće momčadi. Takva notacija koristit će se u ostatku poglavlja.

Koristeći Markovljev lanac za odbojkaški set moguće je izvesti formule za predviđanje njegova ishoda. Definirajmo konačan skup stanja S kojim se pomoću Markovljevog lanca opisuje razvoj odbojkaškog seta [11, 12]:

$$S := \{(i,j,s) : i \in \{0,1,...,24,A,W\}, j \in \{0,1,...,24\}, s \in \{0,1\}\}.$$

Parametar i predstavlja broj osvojenih poena momčadi koja je trenutno na servisu. Vrijednost A parametra i označava prednost momčadi koja je trenutno na servisu - obuhvaća sve rezultate iznad 24, a momčad servira za pobjedu. Vrijednost W parametra i označava pobjedu momčadi koja je trenutno na servisu. Parameter j predstavlja broj osvojenih poena momčadi koja prima servis, a parametar s je zastavica koja označava koja je momčad trenutno na servisu. Završna stanja modela su $(W,0,s),(W,1,s),...,(W,24,s)$. Tranzicije između stanja definirane su na sljedeći način [11, 12]:

•Ako niti jedna mom čad nije osvojila preko 23 poena.

$(i,j,s) \to (i+1,j,s)$ s vjerojatnošću prijelaza $p(s)$

$(i,j,s) \to (j+1,i,1-s)$ s vjerojatnošću prijelaza $1-p(s)$

•Ako je jedna od mom čadi osvojila preko 23 poena.

$(23,24,s) \to (24,24,s)$ s vjerojatnošću prijelaza $p(s)$

$(23,24,s) \to (W,23,1-s)$ s vjerojatnošću prijelaza $1-p(s)$

$(24,23,s) \to (W,23,s)$ s vjerojatnošću prijelaza $p(s)$

$(24,23,s) \to (24,24,1-s)$ s vjerojatnošću prijelaza $1-p(s)$

$(24,24,s) \to (A,24,s)$ s vjerojatnošću prijelaza $p(s)$

$(24,24,s) \to (A,24,1-s)$ s vjerojatnošću prijelaza $1-p(s)$

$(A,24,s) \to (W,24,s)$ s vjerojatnošću prijelaza $p(s)$

$(A,24,s) \to (24,24,1-s)$ s vjerojatnošću prijelaza $1-p(s)$

Kako bi izračunali vjerojatnost da momčad koja započinje servirati u setu osvoji taj set, potrebno je izračunati vjerojatnosti da Markovljev lanac iz početnog stanja $(0,0,s)$, dosegne jedan od završnih stanja modela ($(W,0,s)$, $(W,1,s)$, ..., $(W,24,s)$). Jedan od mogućih pristupa svodi se na prolazak kroz sva stanja Markovljevog lanca koja vode do završnog stanja. Taj pristup je poprilično računarski zahtjevan jer se odbojkaški set opisuje velikom matricom prijelaza dimenzija $1265x1265$ [12]. Alternativni pristup obuhvaća izravan kombinatorički izračun vjerojatnosti da momčad koja započne servirati osvoji set, a dan je izrazom 5.1 [11, 12]:

$$p(W,s) = \sum_{l=0}^{23} p(W,l,s) + p(24,24,s) * p(A,s) + p(24,24,1-s) * (1 - p(A,1-s)), \quad (5.1)$$

pri čemu je:
$p(W,l,s)$ vjerojatnost da momčad s osvoji set uz l osvojenih poena momčadi $1-s$,
$p(24,24,s)$ vjerojatnost dostizanja stanja $(24,24,s)$ pri čemu je momčad s iduća na servisu,
$p(A,s)$ vjerojatnost da momčad s osvoji set s vodstvom od 2 poena ako je došlo do rezultata $24:24$, a na servisu je momčad s.
Svaki pribrojnik prethodne formule raspisuje se na sljedeći način [11, 12]:

$$p(W,0,s) = p_s^{25}, \quad (5.2)$$

$$p(W,l,s) = \sum_{k=1}^{l} \begin{pmatrix} l-1 \\ l-k \end{pmatrix} \begin{pmatrix} 25 \\ 25-k \end{pmatrix} p(s)^{25-k} p(1-s)^{l-k} (1-p(s))^k (1-p(1-s))^k, \quad (5.3)$$

$$p(24,24,s) = \sum_{k=1}^{24} \begin{pmatrix} 23 \\ 24-k \end{pmatrix} \begin{pmatrix} 24 \\ 24-k \end{pmatrix} p(s)^{24-k} p(1-s)^{24-k} (1-p(s))^k (1-p(1-s))^k, \quad (5.4)$$

$$p(24,24,1-s) = \sum_{k=1}^{24} \begin{pmatrix} 23 \\ 24-k \end{pmatrix} \begin{pmatrix} 24 \\ 25-k \end{pmatrix} p(s)^{25-k} p(1-s)^{24-k} (1-p(s))^k (1-p(1-s))^{k-1}, \quad (5.5)$$

$$p(A,s) = \frac{p(s)^2}{p(s)^2 * (1-p(1-s)) + p(1-s)^2 * (1-p(s)) + p(s) * p(1-s)}. \quad (5.6)$$

Primjer računanja vjerojatnosti osvajanja seta: U svrhe razumijevanja formula 5.2, 5.3, 5.4 i 5.5 doktorska disertacija nudi primjer računanja vjerojatnosti da momčad domaćin osvoji set s rezultatom $25:2$ uz pretpostavku da momčad domaćin započinje servirati u setu (iako se radi o izrazito nerealnom ishodu, primjer primarno ima svrhu demonstracije formula). Do osvajanja seta rezultatom $25:2$ moguće je doći na više načina. Momčad domaćin može osvojiti 24 poena na vlastitom servisu s vjerojatnošću $p(s)^{24}$. U tom slučaju momčad domaćin također

mora izgubiti jedan poen na vlastitom servisu s vjerojatnošću $1-p(s)$, zatim momčad gost mora osvojiti jedan poen na vlastitom servisu s vjerojatnošću $p(1-s)$, a potom izgubiti poen na vlastitom servisu s vjerojatnošću $1-p(1-s)$ (upravo tim redoslijedom). Ovaj slučaj može se dogoditi na 25 različitih načina jer je svejedno koji od 25 mogućih poena u setu će momčad domaćin izgubiti. Primijeni li se odgovarajuća formula (5.3) prvi pribrojnik sume je upravo $25 * p(s)^{24} * (1-p(s)) * p(1-s) * (1-p(1-s))$ čime je pokriven opisan slučaj. Drugi slučaj koji se može dogoditi je da momčad domaćin osvoji 23 poena na vlastitom servisu s vjerojatnošću $p(s)^{23}$. Da bi se u tom slučaju postigao rezultat 25 : 2 momčad domaćin mora izgubiti 2 poena na vlastitom servisu s vjerojatnošću $(1-p(s))^2$, a momčad gost također mora izgubiti 2 poena na vlastitom servisu s vjerojatnošću $(1-p(1-s))^2$. Ovaj slučaj može se realizirati na 300 načina - od 25 poena u odbojkaškom setu potrebno je izabrati 2 poena koja će momčad domaćin izgubiti, a isto je moguće na ${}_{25}C_2$. Ovaj slučaj pokriven je drugim pribrojnikom sume 5.3, $300 * p(s)^{23} * (1-p(s))^2 * (1-p(1-s))^2$.

Konačno je još potrebno izvesti formulu 5.6, a izvod je dan u nastavku.

Izvod formule 5.6. Nacrtajmo segment Markovljevog lanca za odbojkaški set koji modelira slučaj kada set dosegne rezultat 24 : 24 (slika 5.2). Model se sastoji od 6 stanja: $(24,24,s)$, $(24,24,1-s)$, $(A,24,s)$, $(A,24,1-s)$, $(W,24,s)$, $(W,24,1-s)$. Označimo sa w_s vjerojatnost da momčad s osvoji set s vodstvom od 2 poena u slučaju da je došlo do rezultata 24 : 24 uz pretpostavku da je početno stanje $(24,24,s)$, a sa w_{1-s} vjerojatnost da momčad s osvoji set s vodstvom od 2 poena u slučaju da je došlo do rezultata 24 : 24 uz pretpostavku da je početno stanje $(24,24,1-s)$. Uzmu li se u obzir svi putevi Markovljevog lanca koji iz stanja $(24,24,s)$ ili iz stanja $(24,24,1-s)$ vode u stanje $(W,24,s)$ dobiju se izrazi [84]:

$$w_s = p(s)^2 + (1-p(s)) * (1-p(1-s)) * w_s + p(s) * (1-p(s)) * w_{1-s}, \tag{5.7}$$

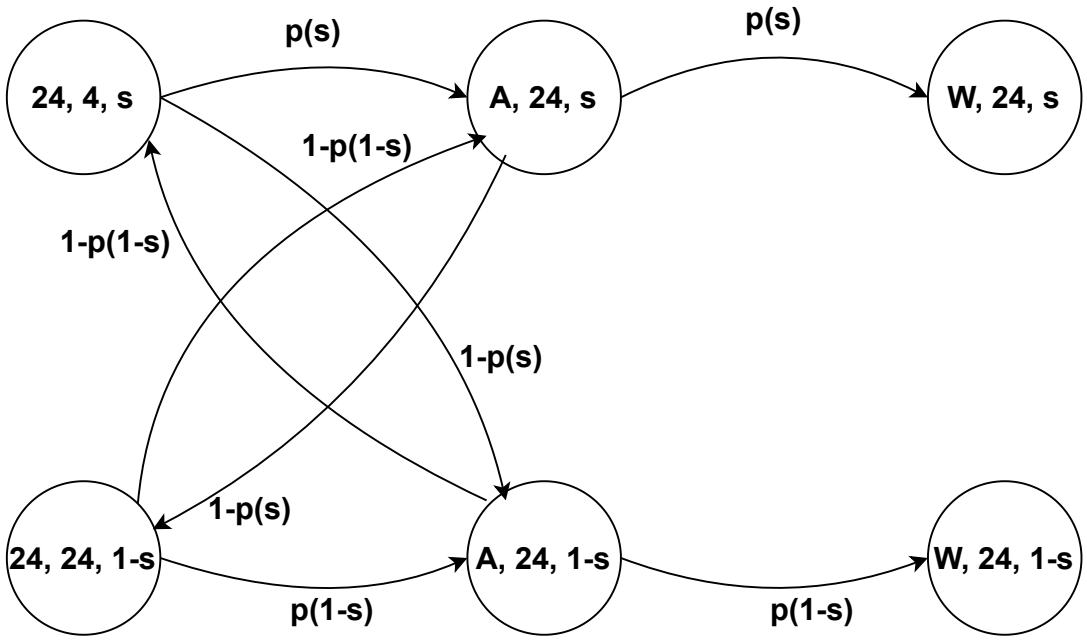
$$w_{1-s} = (1-p(1-s)) * p(s) + p(1-s) * (1-p(1-s)) * w_s + (1-p(1-s)) * (1-p(s)) * w_{1-s}. \tag{5.8}$$

Rješavanjem sustava jednadžbi 5.7 i 5.8 dobiju se izrazi [84]:

$$w_s = \frac{p(s)^2}{p(s)^2 + p(1-s)^2 + p(s) * p(1-s) * (1-p(s)-p(1-s))}, \tag{5.9}$$

$$w_{1-s} = \frac{p(s) * (1-p(1-s)) * (p(s)+p(1-s))}{p(s)^2 + p(1-s)^2 + p(s) * p(1-s) * (1-p(s)-p(1-s))}. \tag{5.10}$$

Lako je za uočiti da je izraz 5.9 ekvivalentan izrazu 5.6, dok je izraz 5.10 jednak izrazu $(1-p(A,1-s))$ u formuli 5.1



Slika 5.2: Segment usmjerenog grafa Markovljevog lanca za modeliranje slučaja izjednačenog rezultata

5.2 Markovljev lanac za odbojkašku utakmicu

Slične kombinatoričke izraze kao i za odbojkaški set moguće je ispisati i za odbojkašku utakmicu. Vjerojatnost da će momčad koja započne servirati u odbojkaškoj utakmici i osvojiti tu utakmicu dana je izrazom [11, 12]:

$$p(sW(3,0)) + p(sW(3,1)) + p(sW(3,2)), \tag{5.11}$$

pri čemu je:
$p(sW(3,0))$ vjerojatnost da momčad s osvoji meč sa $3:0$ u setovima,
$p(sW(3,1))$ vjerojatnost da momčad s osvoji meč sa $3:1$ u setovima,
$p(sW(3,2))$ vjerojatnost da momčad s osvoji meč sa $3:2$ u setovima.
Svaki od pribrojnika prethodne formula računa se sljedećim izrazima [11, 12]:

$$p(sW(3,0)) = p(W,s)^2 * (1 - p(W,1-s)), \tag{5.12}$$

$$p(sW(3,1)) = 2*(1-p(W,s))*p(W,s)*(1-p(W,1-s))^2 + p(W,s)^2 * p(W,1-s)*(1-p(W,1-s)), \tag{5.13}$$

$$p(sW(3,2)) = \Big[p(W,s)^2 * p(W,1-s)^2 + (1-p(W,s))^2 * (1-p(W,1-s))^2 + 4 * p(W,s) * p(W,1-s) * (1-p(W,s)) * (1-p(W,1-s))] * pp_{(T,s)}. \tag{5.14}$$

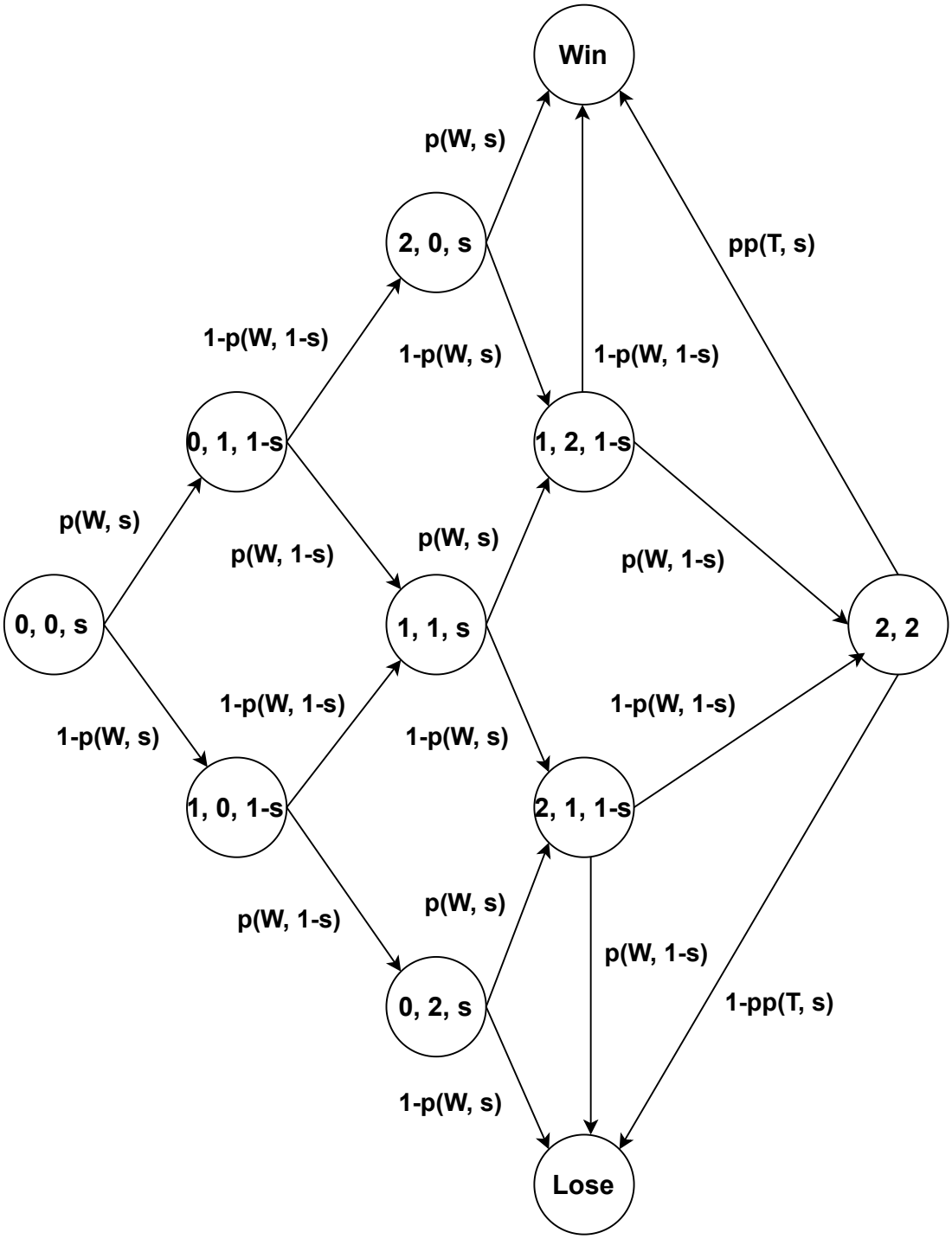
Parametar $pp_{(T,s)}$ u 5.14 predstavlja vjerojatnost osvajanja 5. seta. S obzirom na to da se baca novčić kako bi se odlučilo koja momčad započinje servirati u 5. setu, vjerojatnost da će momčad s osvojiti 5. set računa se izrazom [11, 12]:

$$pp_{(T,s)} = \frac{1}{2} * p_{(T,s)} + \frac{1}{2} * \left(1 - p_{(T,1-s)}\right). \tag{5.15}$$

Parametar $p_{(T,s)}$ u 5.15 predstavlja vjerojatnost osvajanja 5. seta ako 5. set započne servirati momčad s. Računa se na isti način kao i vjerojatnost osvajanja običnog seta (više u poglavlju 5.1) uz uvjet da se 5. set igra do 15 poena, a ne do 25 poena.

Prethodne formule lako je izvesti promatra li se Markovljev lanac za odbojkašku utakmicu prikazan slikom 5.3. Markovljevim lancem su prikazani svi putevi i odgovarajuće vjerojatnosti kojima momčad s može doći do pobjede.

Slični kombinatorički izrazi mogu se pronaći i za modeliranje ishoda teniskih mečeva [9].



Slika 5.3: Usmjereni graf Markovljevog lanca za odbojkašku utakmicu

Poglavlje 6

Podatkovni skup

6.1 Podatkovni skupovi u sportskoj domeni

Podaci igraju važnu ulogu gotovo u svakoj industriji. Iako se u domeni sporta još uvijek često čuje rečenica "utakmice se ne igraju na papiru", teško je zanemariti očite prednosti koje donosi sportska analitika [113]. Globalna sportska industrija evoluirala je tijekom vremena, a menadžeri i momčadi postaju sve spremniji na korištenje podataka za postizanje konkurentske prednosti na terenu i izvan njega. Analiza sportskih podataka pomaže sportskim subjektima da procjene izvedbu svojih igrača, regrutiraju nove igrače i poboljšaju izvedbu svoje momčadi. Analizom podataka može se procijeniti i izvedba protivnika, a treneri to mogu iskoristiti kako bi oblikovali najbolju taktiku protiv njega. Tehnologija je toliko napredovala da je moguće pratiti igrača u stvarnom vremenu tijekom utakmice. To se može iskoristiti za praćenje kretanja i brzine igrača. Praćena statistika sama po sebi ima velik značaj, ali način na koji se podaci mogu integrirati s partnerskim sustavima otključava novi potencijal [113]. Klubovi mogu koristiti prikupljene podatke kako bi procijenili igračevu sposobnost postizanja zgoditaka s određene pozicije ili kako bi analizirali kako igrač surađuje s drugim članovima momčadi. Prednosti sportske analitike nadilaze samu utakmicu. Timovi fizioterapeuta na temelju povijesne analize mogu koristiti metriku za procjenu je li igrač potpuno spreman za utakmicu ili se još uvijek muči s ozljedom [113]. Korištenje podataka sportskim subjektima pomaže u povećanju prihoda, smanjenju troškova poslovanja i jamči visoke povrate ulaganja. Podatci pomažu sportskim tvrtkama oblikovati marketinške strategije, povećati bazu obožavatelja i poboljšati prodaju robe. Koriste se za prikupljanje sponzorstava, provođenje simulacija i predviđanja [113]. Očekuje se da će globalno tržište sportske analitike postići ukupnu godišnju stopu rasta od 27,3% između 2021. godine i 2030. godine. Veličina tržišta u 2021. godini procijenjena je na 2,45 milijardi američkih dolara, a predviđa se da će doseći gotovo 22 milijarde američkih dolara do 2030. godine [114].

Sportski podaci sadrže informacije o sportskim događajima, igračima i svim ostalim sudi-

onicima u sportskom sektoru, a pružaju uvid o učinku momčadi ili pojedinca. Podatkovni skupovi u domeni sporta mogu sadržavati različite razine detalja. Takozvani *play-by-play* (PBP) podatkovni skupovi najmoćniji su podatkovni skupovi u sportskoj domeni. Takvi podaci pružaju informacije o svakom detalju u svim aspektima igre, od svakog bacanja u bejzbolu, svake akcije, lokacije pucanja i prekida u košarci do svakog dodira s loptom u nogometu i odbojci. Drugim riječima, PBP podatkovni skupovi daju transkript igre u formatu pojedinačnih događaja. Takvi podatci, pružajući dubinu koja se ne može naći u drugim izvorima podataka, lako pobuđuju kreativnost sportskih analitičara [31, 32, 33, 34, 35]. Osim detaljnih PBP podatkovnih skupova, postoje i sportski podatkovni skupovi s nešto manjom razinom detalja. Imperativ je međutim, da svaki podatkovni skup u domeni sporta uključuje povijesne podatke ili ključne statistike. To znači da se skup podataka koji ne može pružiti informacije iz povijesnog razdoblja igrača ili momčadi smatra se nepotpunim. Takvi se podatkovni skupovi naime, ne mogu se koristiti kako za procjenu performansi igrača ili momčadi, tako niti za analizu ili predviđanje budućih događaja.

Prikupljanje sportskih podataka može se obaviti pomoću cijelog niza metoda, a to uvelike ovisi o vrsti sporta koji se analizira i o količini podataka koju treba prikupiti. Statističke tvrtke (engl. *stats companies*) jedan su od glavnih izvora prikupljanja sportskih podataka. One vode evidenciju o golovima, prekršajima, rezultatima, podacima momčadi, pojedinačnim nastupima i biografijama igrača. Analitičari se često angažiraju kako bi ručno zabilježili i digitalno pohranili te podatke. Složeniji podaci bilježe se pomoću programske podrške za bilježenje svih događaja u utakmici, koji se pohranjuju u bazu podataka.

6.2 Opis i predobrada podatkovnog skupa u domeni odbojke

Podatkovni skup prikupljen u svrhu izrade ove doktorske disertacije sadrži podatke o odbojkaškim utakmicama koje su se održavale u vremenskom periodu od travnja 2016. godine sve do listopada 2017. godine [116]. On ne sadrži razinu detalja kao i gore opisani, detaljni PBP podatkovni skupovi. Podatkovni skup korišten u ovoj doktorskoj disertaciji sadrži samo informacije o izmjenama rezultata tijekom utakmica i sadrži odgovarajuće koeficijente koje je postavila kladioničarska kuća. Detaljnije, za svaku utakmicu podatkovni skup bilježi izmjene rezultata, očekivane vrijednosti i koeficijente za tri najčešća tipa klađenja – klađenje na pobjednika utakmice, klađenje na očekivani broj poena i klađenje na očekivani hendikep (više detalja u poglavlju 2.2). S obzirom na to da podatkovni skup sadrži takozvane *in-play* izmjene rezultata u utakmicama, za takav podatkovni skup, od sada pa nadalje, u ovoj doktorskoj disertaciji koristi će se termin *in-play* (IP) podatkovni skup. Korišteni podatkovni skup inicijalno je sadržavao informacije o 14153 odbojkaške utakmice, objavljen je javno [117], a atributi podatkovnog skupa zajedno s njihovim opisom prikazani su u tablici (tablica 6.1 [116]). Potrebno je napomenuti da su atributi *PrematchCoef1* i *PrematchCoef2* prikazani u tablici 6.1 izvedeni iz podatkovnog skupa na način opisan u nastavku poglavlja. Radi se o tzv. *pre-match* koeficijentima na ishod utakmice, odnosno o koeficijentima koje je kladioničarska kuća postavila na ishod utakmice prije nego što je utakmica počela.

In-play podatkovni skup poprilično ograničava sportskog analitičara, međutim uz određenu razinu kreativnosti, analiza takvih podataka omogućuje ekstrahiranje korisnih znanja i poboljšanje modela za predviđanje ishoda sportskih događaja. To je demonstrirano u ostatku ove doktorske disertacije. Demonstraciji rezultata prethodi opis procesa čišćenja, predobrade korištenog IP podatkovnog skupa te realizacija zamišljenog modela. Tri ključna koraka vezana uz problem pripreme podatkovnog skupa opisana su u nastavku [116]:

- **Nedostajuće in-play vrijednosti**.

  U podatkovnom skupu nedostajale su informacije za neke izmjene rezultata u utakmicama. Informacije o izmjenama rezultata ključne su za izgradnju modela predloženog u ovoj doktorskoj disertaciji. Iz tog razloga, određeni dio podataka morao je biti eliminiran. Kako bi se spriječila veća redukcija podatkovnog skupa eliminirani su podatci o onim utakmicama za koje nije zabilježen konačan rezultat utakmice. Podatak o konačnom rezultatu utakmice bitan je radi evaluacije modela. Kako bi se zadržala i statistička relevantnost analiza, eliminirane su i utakmice za koje je iz nekog razloga zabilježeno manje od 75 izmjena rezultata. Naime, analizom podatkovnog skupa korištenog u ovoj doktorskoj disertaciji zaključeno je da se u prosječnoj odbojkaškoj utakmici sa snažnim favoritom igra prosječno 134 poena, dok se u utakmici, u kojoj su suprotstavljene momčadi sličnog omjera snaga, igraju u prosjeku 182 poena. Podatci o utakmicama za koje je

Tablica 6.1: Opis atributa in-play odbojkaškog podatkovnog skupa

Naziv atributa	Opis atributa
SportEventID	Jedinstveni identifikator utakmice.
Score	Rezultat utakmice. Simbolom * bilježi se momčad na servisu. Tako npr. rezultat $1:1*23:20$ označava da je trenutni rezultat 1 : 1 u setovima. Igra se treći set u kojem je trenutni rezultat 23 : 20, a na servisu je momčad domaćin.
PrematchCoef1/ PrematchCoef2	Koeficijenti na pobjednika utakmice postavljeni prije nego što je utakmica započela. Koeficijenti se bilježe za obje momčadi.
CoefMatchWin1/ CoefMatchWin2	Koeficijenti na pobjednika utakmice postavljeni na određenom rezultatu utakmice. Koeficijenti se bilježe za obje momčadi.
Handicap	Aproksimacija razlike u osvojenim bodovima momčadi.
HandicapUnder	Koeficijent da će hendikep biti manji od aproksimirane vrijednosti.
HandicapOver	Koeficijent da će hendikep biti veći od aproksimirane vrijednosti.
TotalPoints	Aproksimacija ukupnog broja poena u utakmici.
TotalPointsUnder	Koeficijenti da će se odigrati manji broj poena od aproksimirane vrijednosti.
TotalPointsOver	Koeficijenti da će se odigrati veći broj poena od aproksimirane vrijednosti.

zabilježen konačan rezultat utakmice i za koje je zabilježeno barem 75 izmjena rezultata zadržane su u podatkovnom skupu.

- **Nedostajući pre-match koeficijenti**.
  Za većinu utakmica u podatkovnom skupu nije zabilježena vrijednost koeficijenta na ishod sportskog događaja kojeg je kladioničarska kuća postavila prije nego što je utakmica započela, tzv *pre-match* koeficijenta. Kako bi se izbjegla dodatna redukcija podatkovnog skupa, a time i gubitak vrijednih informacija, korišteno je zaobilazno rješenje za neizravni izračun podataka koji nedostaju. Detaljnije, kao pre-match koeficijenti u tom slučaju uzeti su koeficijenti postavljeni na rezultatima 0 : 0 1 : 1, 0 : 0 2 : 2 (značenje ovakvog zapisa rezultata objašnjeno je u tablici 6.1 u stupcu pod nazivom *Score*). Ideja je podržana hipotezom da se pre-match koeficijenti ne razlikuju puno od koeficijenata postavljenih na prethodno navedenim rezultatima s obzirom na to da je utakmica tek počela, a niti jedna od momčadi još uvijek nije mogla pokazati bolje performanse. Postavljenu hipotezu podržali su domenski eksperti, a eksperimentalno je utvrđeno da je srednja vrijednost relativnih promjena koeficijenata postavljenih na rezultatima 0 : 0 1 : 1, 0 : 0 2 : 2 tek 0.17%.
- **Pozadinske distribucije**.

Podatkovni skup korišten u ovoj doktorskoj disertaciji inicijalno je sadržavao četiri široke kategorije utakmica, odnosno četiri različite pozadinske distribucije poena. Detaljnije, podatkovni skup sadržavao je zapise o regularnoj i mladoj odbojkaškoj sekciji u muškoj i ženskoj konkurenciji. S obzirom na to da je u znanstvenoj literaturi pokazano da postoje primjetne razlike u načinu igre navedenih kategorija [118], disertacija polazi od pretpostavke da su specifičnosti svake kategorije dovoljno različite da njihovo zajedničko modeliranje nije dobar pristup. Kako bi se osigurao podatkovni skup s utakmicama koje imaju jednaku pozadinsku distribuciju poena i kako bi se osiguralo da simulacije aproksimiraju događaje iz stvarnog svijeta, odabrana je samo jedna od prethodno navedenih kategorija. Raspodjela broja utakmica po kategoriji u korištenom podatkovnom skupu je sljedeća:

- Mlada sekcija ženske konkurencije $\rightarrow$ 399 utakmica;
- Regularna sekcija ženske konkurencije $\rightarrow$ 2964 utakmice;
- Mlada sekcija muške konkurencije $\rightarrow$ 421 utakmice;
- Regularna sekcija muške konkurencije $\rightarrow$ 4704 utakmice.

Uzimajući u obzir ukupan broj odigranih utakmica po skupini u korištenom podatkovnom skupu, odabrana je kategorija regularne sekcije u muškoj konkurenciji. Utakmice navedene kategorije čine 55% utakmica inicijalnog podatkovnog skupa, što konačno rezultira podatkovnim skupom od 4704 utakmice.

Radi boljeg uvida u podatkovni skup koji se koristi u ovoj doktorskoj disertaciji, na slici 6.1 je dan primjer ispisa manjeg uzorka podataka. Zbog fluidnosti teksta, eksploratorna analiza korištenog podatkovnog skupa proteže se kroz poglavlja dana u ostatku doktorske disertacije i nije izdvojena kao zasebno poglavlje.

```
$ SportEventID     <dbl> 4920346, 4920346, 4920346, 4920346, 4920346, 4920346, 49203...
$ Team1            <fctr> Arda, Arda, Arda, Arda, Arda, Arda, Arda, Arda, Arda, Arda...
$ Team2            <fctr> CSKA Sofia, CSKA Sofia, CSKA Sofia, CSKA Sofia, CSKA Sofia...
$ Score            <fctr> 0:0  (1:1), 0:0  (2:1), 0:0  (3:1), 0:0  (4:1), 0:0  (4:2)...
$ PrematchCoef1    <dbl> 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 21,...
$ PrematchCoef2    <dbl> 1.005, 1.005, 1.005, 1.005, 1.005, 1.005, 1.005, 1.005, 1.0...
$ CoefMatchWin1    <fctr> 21.00, 19.50, 15.50, 11.50, 12.50, 13.00, 11.50, 11.00, 9....
$ CoefMatchWin2    <dbl> 1.005, 1.008, 1.012, 1.035, 1.030, 1.025, 1.035, 1.040, 1.0...
$ Handicap         <fctr> (+23.5), (+22.5), (+22.5), (+20.5), (+20.5), (+20.5), (+20...
$ HandicapUnder    <fctr> 1.93, 1.96, 1.82, 1.89, 1.90, 1.96, 1.84, 1.89, 1.90, 1.98...
$ HandicapOver     <fctr> 1.87, 1.84, 1.98, 1.909, 1.90, 1.84, 1.96, 1.909, 1.90, 1....
$ TotalPoints      <fctr> (128.5), (129.5), (130.5), (133.5), (133.5), (132.5), (134...
$ TotalPointsUnder <fctr> 1.86, 1.89, 1.94, 1.93, 1.89, 1.89, 1.87, 1.909, 1.90, 1.8...
$ TotalPointsOver  <fctr> 1.94, 1.909, 1.86, 1.87, 1.909, 1.909, 1.93, 1.89, 1.90, 1...
```

Slika 6.1: Uzorak korištenog podatkovnog skupa

Poglavlje 7

Parametri sportske dinamike

Pojam *parametar sportske dinamike* samoopisan je, a odnosi se na one parametre koji u modele za opisivanje sportskih događaja uvode dinamički element. Modeli temeljeni na jednolikoj i neovisnoj distribuciji poena ne inkorporiraju parametre sportske dinamike - vrijednosti vjerojatnosti osvajanja poena na vlastitom servisu momčadi u takvim modelima ostaju konstantnima tijekom cijele utakmice, neovisno o događajima u utakmici, načinu postizanja određenog rezultata ili o fazi utakmice. U takvim modelima dinamika utakmice se ne prati.
Ova doktorska disertacija polazi od tri hipoteze:

1. Jednolika i neovisna distribucija poena nije dovoljno dobra pretpostavka u modeliranju utakmica sportova s bodovnim ograničenjem.
2. Uklju čivanjem parametara sportske dinamike u modele sportova s bodovnim ograničenjem moguće je preciznije simulirati realni slijed utakmica te predvidjeti određene karakteristike utakmica.
3. Mogu će je verificirati razvijene modele korištenjem podatkovnih skupova stvarnih utakmica sportova s bodovnim ograničenjem.

Ova doktorska disertacija uvodi dva parametra sportske dinamike, tzv. kratkoročni zamah i dugoročni zamah. Ovi parametri sportske dinamike detaljnije su opisani u nastavku doktorske disertacije, a osim u modelima za predviđanje ishoda sportskih događaja mogu se koristiti i za analiziranje potencijalnih rezultatskih sekvenci utakmice.

7.1 Kratkoročni zamah

Pojam *kratkoročnog zamaha* u ovoj doktorskoj disertaciji koristit će se za opisivanje kratkoročne iznadprosječne izvedbe igrača odnosno momčadi nakon određenog motivirajućeg događaja [116]. Taj pojam čini se prikladnijim s obzirom na to da doktorska disertacija ne razmatra uzrok takvog fenomena. Za razliku od znanstvenih istraživanja obrađenih u poglavlju 4.3, ideja ove doktorske disertacije nije testirati postojanje fenomena psihološkog zamaha već izgraditi

vjerojatnosni model koji implementira kratkoročni zamah kao povećanje vjerojatnosti osvajanja poena na vlastitom servisu nakon motivirajućeg događaja i konačno demonstrirati da takav model ima bolje performanse u odnosu na modele koji ne implementiraju taj fenomen. To su modeli temeljeni na jednolikoj i neovisnoj distribuciji poena.

Motivirajući događaji koje ova doktorska disertacija uzima u obzir su osvajanje jednog ili više poena u nizu u utakmicama sportova s bodovnim ograničenjem [116]. U ovom trenutku potrebno je prvo eksploratornom analizom korištenog podatkovnog skupa racionalizirati potrebu za implementacijom kratkoročnog zamaha te odgovoriti na pitanje koliko je osvojenih poena u nizu, u utakmicama sportova s bodovnim ograničenjem, potrebno uzeti u razmatranje prilikom implementacije kratkoročnog zamaha u modele za predviđanje ishoda takvih sportskih događaja. U tu svrhu testirane su tri hipoteze [116]:

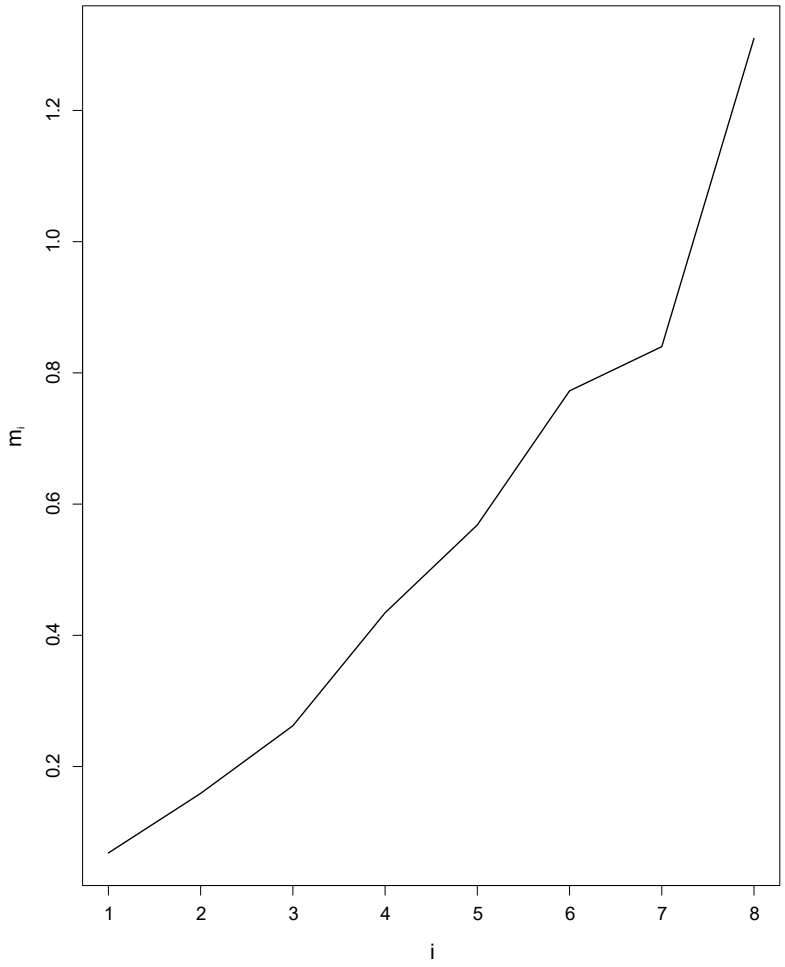
1. Ako momčad osvoji poen, pojavljuje se kratkoročni zamah koji se može opisati kao povećanje vjerojatnosti osvajanja idućeg poena na vlastitom servisu momčadi.
2. U činak kratkoročnog zamaha je kumulativan što znači da osvajanje više poena u nizu rezultira većom vjerojatnošću osvajanja idućeg poena.
3. Osvajanje više od tri poena u nizu događa se rijetko u odbojkaškim utakmicama.

Definirajmo za početak tri pojma [116]:

- *Red zamaha* (i) je definiran kao broj uzastopno osvojenih poena momčadi.
- *Broj pojavljivanja zamaha* (N_i) definiran je kao broj koji opisuje koliko puta se u utakmici (po momčadi) pojavljuje zamah i-tog reda.
- *Vrijednost zamaha* (m_i) je broj koji pokazuje koliko se poveća vjerojatnost osvajanja idućeg poena na vlastitom servisu momčadi ako momčad prethodno osvoji i poena u nizu.

Slika 7.1 [116] prikazuje ovisnost vrijednosti i reda zamaha za prosječnu momčad u korištenom podatkovnom skupu. Vrijednost m_2 tako prikazuje koliko će se (u postotku) povećati vjerojatnost osvajanja trećeg poena u nizu u odnosu na prosječnu vjerojatnost osvajana poena na vlastitom servisu momčadi ako je momčad prethodno osvojila 2 poena na vlastitom servisu [116]. *Napomena*: U ovom trenutku potrebno je obratiti pažnju na trend krivulje, a konkretne vrijednosti bit će analizirane u nastavku disertacije. Potrebno je uočiti da slika 7.1 potvrđuje dvije od tri postavljene hipoteze. Efekt kratkoročnog zamaha postoji (1. hipoteza) i kumulativan je (2. hipoteza).

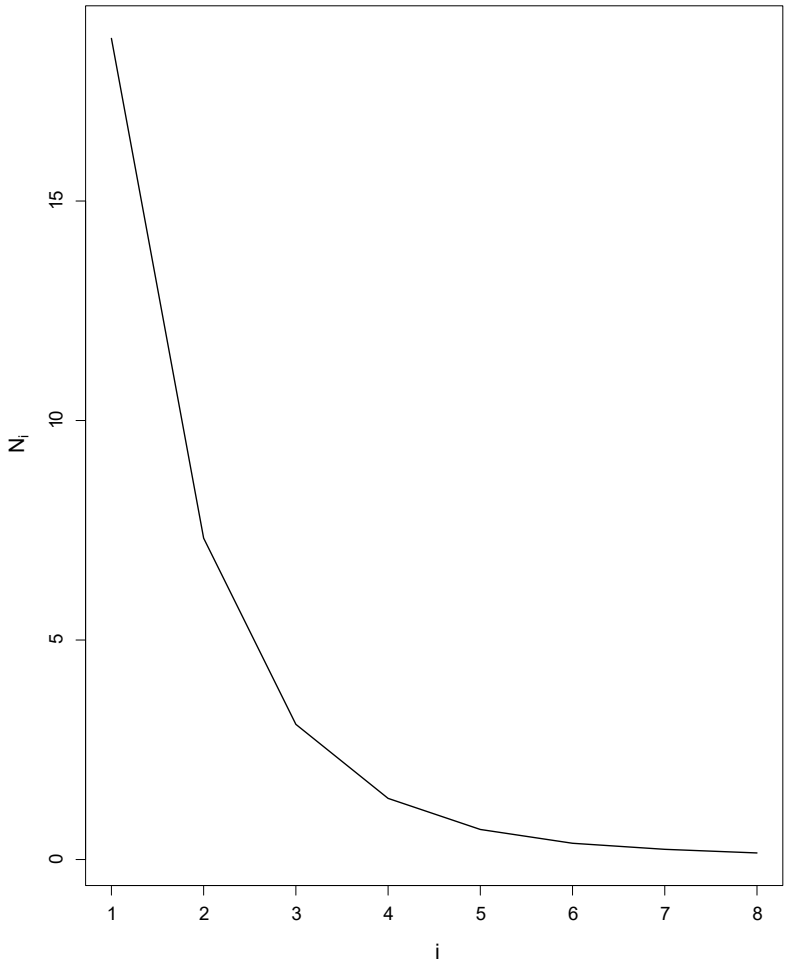
Slika 7.2 [116] prikazuje broj pojavljivanja kratkoročnog zamaha (N_i) u ovisnosti o redu zamaha (i) u prosječnoj odbojkaškoj utakmici za prosječnu odbojkašku momčad. N_2, tako, predstavlja broj koji ukazuje koliko puta je prosječna momčad osvojila 3 poena zaredom [116]. *Napomena*: obratite pažnju na trend krivulje. Iz slike 7.2 je evidentno da se kratkoročni zamasi reda većeg od 3 pojavljuju jako rijetko tijekom utakmica opisanih u korištenom podatkovnom skupu. Time je potvrđena i treća postavljena hipoteza. To znači da niski broj opažanja kratko-



Slika 7.1: Graf ovisnosti vrijednosti zamaha o redu zamaha

ročnih zamaha reda višeg od tri neće dati statistički relevantne rezultate koje bi smo mogli dalje iskoristiti. Zato se ova doktorska disertacija fokusira samo na implementaciju kratkoročnog zamaha prvog, drugog i trećeg reda. Kratkoročni zamah prvog reda u disertaciji je definiran kao uvjetna vjerojatnost (više detalja u poglavlju 3.3) osvajanja drugog poena u nizu na vlastitom servisu momčadi nakon što je momčad prethodno osvojila jedan poen. Slično tome, kratkoročni zamah drugog reda definiran je kao uvjetna vjerojatnost osvajanja trećeg poena u nizu na vlastitom servisu nakon što je momčad prethodno osvojila dva poena zaredom. Konačno, kratkoročni zamah trećeg reda definiran je kao vjerojatnost osvajanja četvrtog poena u nizu na vlastitom servisu nakon tri uzastopna poentiranja momčadi na vlastitom servisu [116].

Budući da podatci iz stvarnog svijeta pružaju dovoljno dokaza za racionalizaciju uvođenja kratkoročnog zamaha, pristupa se njegovom matematičkom modeliranju. Kada momčad koja je na servisu osvoji i poena zaredom, vjerojatnost osvajanja idućeg poena prilagođava se mijenjanjem vjerojatnosti p_{hist}/q_{hist} (ovisno koja je momčad na servisu) novom vjerojatnošću p_{stm}/q_{stm} (slova *stm* označavaju kratkoročni zamah (engl. *short-term momentum*)) prema formulama 7.1



Slika 7.2: Graf ovisnosti broja pojavljivanja zamaha o redu zamaha

i 7.2 [116]:

$$p_{stm} = p_{\text{hist}} * \left(1 + m_i^p\right), \tag{7.1}$$

$$q_{\text{stm}} = q_{\text{hist}} * \left(1 + m_i^q\right). \tag{7.2}$$

U formulama 7.1 i 7.2 je:

p_{hist}/q_{hist} - vjerojatnost osvajanja poena nakon tzv. *breaka* za momčad *A*/*B*, ovisno o tome koja je momčad na servisu. Po definiciji poen nakon *breaka* je prvi poen u nizu i kao takav ne uključuje parametar kratkoročnog zamaha na način na koji je on definiran u ovoj doktorskoj disertaciji.

m_i^p/m_i^q - vrijednost zamaha *i*-tog reda za momčad *A*/*B*, ovisno o tome koja je momčad na servisu. Indeks i označava red zamaha, a poprima vrijednosti u intervalu $i \in [1,3]$. Eksponenti p, q ne označavaju potenciju već služe za razlikovanje momčadi A i B.

Parametri formula 7.1 i 7.2 računaju se iz povijesnih podataka prije početka utakmice, a u doktorskoj disertaciji držat će se konstantnima tijekom cijele utakmice. To znači da doktorska disertacije ne razmatra potencijalnu važnost pojedinog poena i kako ta važnost utječe na kratkoročni zamah - vrijednosti parametra kratkoročnog zamaha tako pri kraju utakmice mogu biti manji radi umora igrača ili većeg psihološkog pritiska. Uz dovoljno povijesnih podataka nadogradnja ponuđenog pristupa trebala bi biti prilično jednostavna, a to je ostavljeno za buduća istraživanja.

Konačno je bitno napomenuti da je prilikom modeliranja kratkoročnog zamaha u obzir potrebno uzeti specifičnost pojedinog sporta. U slučaju odbojke, moguće je analizirati, ovako definiran kratkoročni zamah nakon svakog servisa momčadi. To proizlazi iz činjenice da pravila odbojke nalažu rotiranje servera nakon svakog izgubljenog servisa. Osim toga, u odbojci se jako često mijenja postava momčadi na terenu. Kratkoročni zamah događa se kada je ili dobar server na servisu ili je na terenu jako dobra postava koja osvaja poene. Drugi sportovi poput tenisa mogu imati potpuno drugačiju dinamiku meča (više u poglavlju 8.4).

7.2 Dugoročni zamah

Doktorska disertacija polazi od hipoteze da je u procesu simuliranja sportskih događaja u obzir potrebno uzeti rezultatske sekvence utakmice. Drugim riječima, disertacija pretpostavlja da vjerojatnosti osvajanja poena na vlastitom servisu igrača koje su izračunate iz povijesnih podataka nisu konstantne i ovise o tijeku utakmice. Stanja modela nisu neovisna, a način na koji je model dosegnuo određeno stanje je bitan. Hipoteza je testirana u poglavlju 9, a u ostatku ovog poglavlja predložen je parametar kojim se ostvaruje opisano.

Detaljnije, u disertaciji je predložen dugoročni zamah, dinamički parametar koji kombinira povijesnu statistiku momčadi i simulacijske podatke utakmice čiji se ishod predviđa. Detaljnije, kroz formulaciju dugoročnog zamaha (7.3, 7.4), vjerojatnosti osvajanja poena na vlastitom servisu igrača izračunate iz povijesnih podataka (očekivanja) postupno se ažuriraju s informacijama rezultatskih sekvenci simulacije [116]. To zapravo znači da se vjerojatnost osvajanja poena na vlastitom servisu neznatno prilagođava nakon svakog odigranog poena u simulaciji – ako je momčad napravila nekoliko jako dobrih poteza u utakmici čiji se ishod simulira, kroz formulaciju dugoročnog zamaha vjerojatnost osvajanja poena povećat će se u odnosu na povijesnu vjerojatnost osvajanja poena na vlastitom servisu, a time će se bolje opisati trenutna događanja u simulaciji. Nove vjerojatnosti osvajanja poena na vlastitom servisu se označavaju sa p_{ltm}/q_{ltm} (*ltm* označava dugoročni zamah, engl. *long-term momentum*) te se računaju po sljedećim formulama [116]:

$$p_{ltm} = \frac{totalPts_{current}}{\lambda} * p_{current} + \left(1 - \frac{totalPts_{current}}{\lambda}\right) * p_{hist}, \tag{7.3}$$

$$q_{ltm} = \frac{totalPts_{current}}{\lambda} * q_{current} + \left(1 - \frac{totalPts_{current}}{\lambda}\right) * q_{hist}, \tag{7.4}$$

pri čemu je:

$p_{current}$ i $q_{current}$ - vjerojatnosti osvajanja poena na vlastitom servisu koristeći samo informacije iz trenutno simulirane utakmice, odnosno omjer osvojenih poena na vlastitom servisu momčadi i ukupnog broja servisa momčadi,

$totalPts_{current}$ - trenutno odsimuliran broj poena,

p_{hist} i q_{hist} - povijesne vjerojatnost osvajanja poena na vlastitom servisu za momčadi A i B (kao i u izrazima 7.1 i 7.2),

λ - težina utjecaja dugoročnog zamaha na vjerojatnost osvajanja poena.

Formulacije 7.3, 7.4 temelje se na empirijskom Bayesovom pravilu ažuriranja.

Dokaz. Izraze za dugoročni zamah, 7.3, 7.4 moguće je opisati kao empirijski Bayesov procjenitelj jer su usko povezane s posebnim slučajem beta binomnog konjugiranog modela u Bayesovom zaključivanju (poglavlje 3.4.2). Detaljnije, pravilo ažuriranja vjerojatnosti osvajanja poena

na servisu igrača slična je posteriornoj procjeni srednje vrijednosti binomne slučajne varijable nakon n pokusa koja ima beta apriornu distribuciju $Beta(\alpha,\beta)$ sa $M=\alpha+\beta$. Drugim riječima, vjerojatnost osvajanja poena na vlastitom servisu igrača ravna se po binomnoj razdiobi, $x|\theta \sim bin(n,\theta)$ - binomna distribucija modelira broj uspjeha (u ovom slučaju, osvajanje poena) u fiksnom broju neovisnih pokusa (u ovom slučaju, broj servisa). Konjugirani prior binomne razdiobe je beta razdioba, $Beta(\alpha,\beta)$ (tablica 3.1). Posterior je u tom slučaju također beta razdioba (dokaz 3.20), međutim s drugačijim parametrima, $Beta(x+\alpha, n-x+\beta)$. Radi praktičnosti, u posteriornoj binomnoj razdiobi zamijenimo α i β s parametrima μ i M, pri čemu je $\mu=(\alpha)/(\alpha+\beta)$ što predstavlja srednju vrijednost prior distribucije, a $M=\alpha+\beta$. Konačna posterior beta distribucija tada je $Beta(x+M*\mu, n-x+M*(1-\mu))$. Srednja vrijednost posterior distribucije može se napisati kao [77]:

$$\frac{M*\mu+x}{M+n} = \frac{M}{M+n}*\mu + \frac{n}{M+n}*\frac{x}{n}. \tag{7.5}$$

Formula 7.5 opisuje osnovni koncept Bayesove statistike - vjerojatnost se tretira kao uvjerenje koje se mijenja (ažurira, nadograđuje) dolaskom novih informacija. Drugim riječima apriori uvjerenje o parametru (μ) ažurira se s novim informacijama (x/n) kako eksperiment napreduje. Težine vrednovanja apriornog uvjerenje i novih informacija opisane su formulom 7.5. Prilagođeni oblik formule 7.5 poprimaju formulacije 7.3, 7.4.

U formulama 7.3, 7.4 odabir vrijednosti parametra λ omogućuje kontroliranje koliko će snažno dugoročni zamah utjecati na vjerojatnost osvajanja poena na vlastitom servisu [116]. Dugoročni zamah mora se ponašati na takav način da povijesna statistika uvijek ima najveći utjecaj na početku utakmice, a utjecaj povijesne statistike se postupno smanjuje kako simulacija napreduje, izrazito u slučaju kada trenutno simulirani događaji jako odstupaju od povijesnih podataka. Odabirom većih vrijednosti parametra λ povijesna statistika ima veći utjecaj, dok niže vrijednosti parametra λ stavljaju veći naglasak na ono što se simulira u utakmici [116]. U ovom trenutku se postavlja logično pitanje - kako odabrati vrijednost parametra λ, odnosno kako na temelju apriornih informacija o utakmici odrediti najbolju vrijednost parametra. Zbog same prirode sporta, razlika u snagama momčadi izravno utječe na broj poena koji će se odigrati u utakmici. Jednostavan, ali učinkovit pristup je odabrati vrijednost parametra λ na temelju grube procjene očekivanog broja poena koji bi se trebao odigrati u utakmici ($totalPts_{exp}$, engl. *total points expected*). Naime, ako utakmicu igraju dvije momčadi sličnih snaga i sposobnosti, očekuje se da će se odigrati puno više poena u usporedbi s utakmicom u kojoj su suprotstavljene dvije izrazito neuravnotežene momčadi sa snažnim favoritom. U duljim utakmicama, dugoročni zamah postaje sve izraženiji što se može modelirati većom vrijednošću parametra λ kako u ranoj fazi utakmice ne bi preveliki naglasak bio na simulacijskim podacima [116]. Povijesna statistika za procjenu očekivanog broja poena koji će se odigrati u utakmici može se izračunati, bilo upotrebom podataka o susretima istih momčadi (ako se dogodio dovoljno puta), ili prvo

grupiranjem momčadi na temelju snage momčadi i izračunavanjem prosječnog broja poena između momčadi definiranih skupina – to je pristup koji će se koristiti u disertaciji i bit će detaljnije argumentiran i opisan u nastavku (poglavlje 8.3). Ukratko ako analitičar ima dovoljno dostupnih povijesnih podataka, tada on može modelirati parametar λ kao linearnu funkciju procijenjenog ukupnog broja poena koji će biti odigran u utakmici.

Nove formule s modificiranim parametrom λ prikazana su u izrazima 7.6 i 7.7 [116]:

$$p_{ltm} = \frac{totalPts_{current}}{k * totalPts_{exp} + l} * p_{current} + \left(1 - \frac{totalPts_{current}}{k * totalPts_{exp} + l}\right) * p_{hist}, \tag{7.6}$$

$$q_{ltm} = \frac{totalPts_{current}}{k * totalPts_{exp} + l} * q_{current} + \left(1 - \frac{totalPts_{current}}{k * totalPts_{exp} + l}\right) * q_{hist}. \tag{7.7}$$

U formulama 7.6 i 7.7 su dodani parametri k, l i $totalPts_{exp}$, koji se računa iz povijesnih podataka. U disertaciji se parametri k i l tretiraju kao univerzalni parametri za sve utakmice – profinjeniji pristup s uvjetnim vrijednostima parametara k i l (koji bi ovisili o individualnim svojstvima utakmica) ostavljen je za buduća istraživanja.

Pristup predložen u ovom poglavlju primjenjiv je i na druge sportove s bodovnim ograničenjem, a više detalja o tome dano je u poglavlju 8.4.

Za kraj je bitno napomenuti da pretpostavka o jednolikoj i neovisnoj distribuciji poena neće adekvatno opisati situacije kada događaji na terenu snažno odstupaju od događaja opisanih povijesnim podatcima. Odstupanja se mogu dogoditi iz više razloga. Jakoj momčadi, primjerice, zbog ozljede može nedostajati ključni igrač u utakmici. To može rezultirati lošijim performansama momčadi u toj utakmici u odnosu na prosječne performanse te momčadi koje se izračunavaju iz povijesnih podataka. U tom slučaju modeli koji se temelje samo na povijesnim podacima nastavljaju favorizirati tu momčad više nego što bi trebalo. Zato je važno, uz povijesne podatke u obzir uzeti i dostupne podatke o utakmici čiji se ishod predviđa, a u te svrhe može se koristiti blago modificirana formulacija dugoročnog zamaha koja bi umjesto rezultata simulacijskih sekvenci pratila utakmicu uživo i ažurirala povijesne vjerojatnosti osvajanja poena na vlastitom servisu igrača sa statistikom trenutne utakmice. Ovo je ostavljeno za buduća istraživanja.

Poglavlje 8

Nejednoliko i ovisno prediktivno modeliranje odbojkaških utakmica

8.1 Hibridna formula

Nakon što su predstavljeni koncepti kratkoročnog i dugoročnog zamaha, u nastavku je predstavljen pristup koji kombinira ova dva koncepta kroz objedinjujuću hibridnu formulaciju. Hibridna formula ima za cilj ažurirati povijesne vjerojatnosti osvajanja poena na vlastitom servisu momčadi s informacijama iz simulacijskih sekvenci utakmice čiji se ishod predviđa, a pri tome istovremeno ažurirati novoizračunate vjerojatnosti osvajanja poena na vlastitom servisu ovisno o pojavi događaja koji uzrokuje kratkoročnu promjenu performansi momčadi [116]. Kako bi se implementirala kombinirana formula za modeliranje nejednolike i ovisne razdiobe, formule 7.1 i 7.6 se kombiniraju, tretirajući p_{hist} iz formule 7.1 kao p_{ltm} u formuli 7.6. Slično vrijedi i za gostujuću momčad gdje se q_{hist} iz formule 7.2 tretira kao q_{ltm} u formuli 7.7 [116]:

$$p_{ltm} = [\frac{totalPts_{current}}{k * totalPts_{exp} + l} * p_{current} + \left(1 - \frac{totalPts_{current}}{k * totalPts_{exp} + l}\right) * p_{hist}] * (1 + m_i^p), \quad (8.1)$$

$$q_{ltm} = [\frac{totalPts_{current}}{k * totalPts_{exp} + l} * q_{current} + \left(1 - \frac{totalPts_{current}}{k * totalPts_{exp} + l}\right) * q_{hist}] * (1 + m_i^q). \quad (8.2)$$

Parametri formula 8.1 i 8.2 objašnjeni su u poglavlju 7.

Slika 8.1 [116] prikazuje stablo utakmice koje inkorporira predloženu hibridnu formulaciju. Tu se, na prijelazima stabla, prikazuje kako se mijenjaju vjerojatnosti osvajanja poena na vlastitom

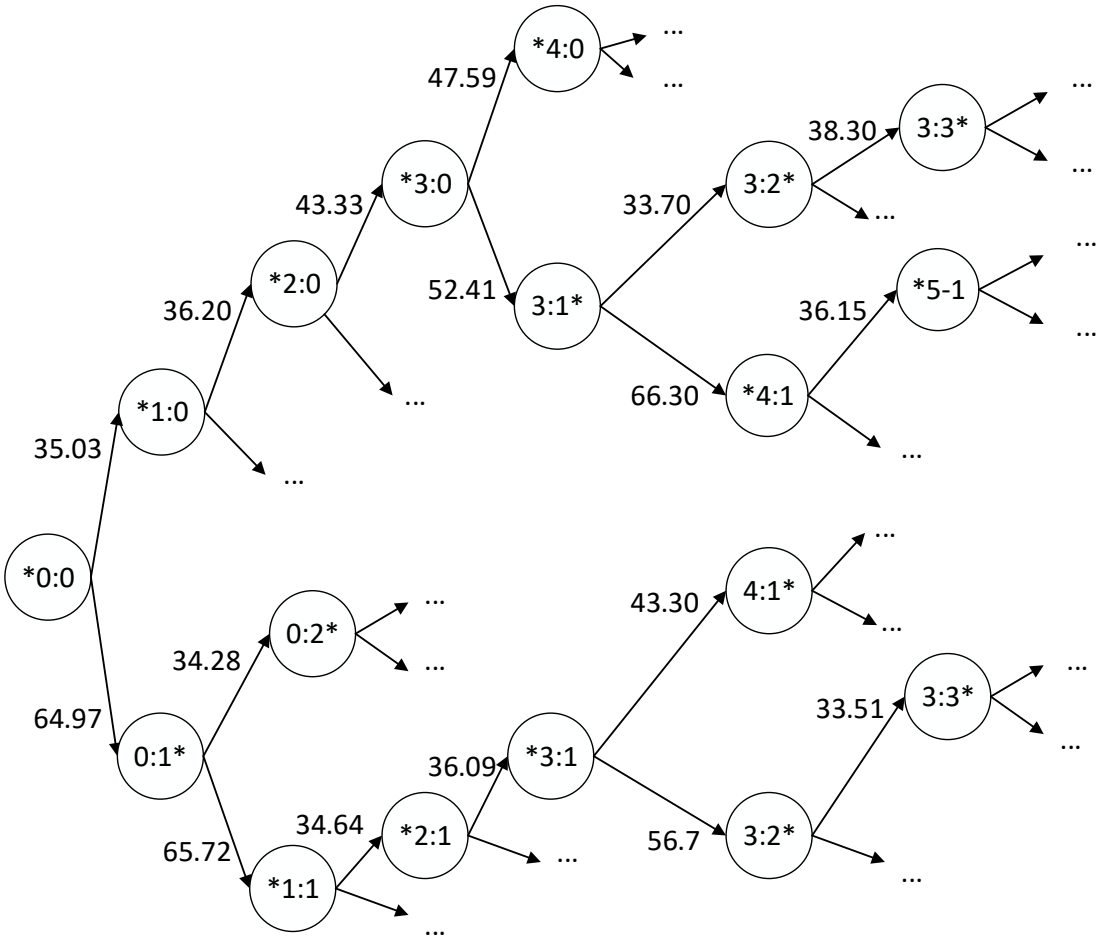
servisu obje momčadi kroz cijelu utakmicu. Kako bi se to demonstriralo, odabrana je utakmica sa sljedećim karakteristikama [116]:

- $p_{hist} = 35,03\%$,
- $q_{hist} = 34,47\%$,
- $m_1^p = 2,28\%$,
- $m_2^p = 21,77\%$,
- $m_3^p = 31,75\%$,
- $m_1^q = 5,51\%$,
- $m_2^q = 13,97\%$,
- $m_3^q = 34,58\%$
- $totalPts_{exp} = 179$

Radi demonstracije primjera, kao parametri linearne funkcije, odabrane su vrijednosti $k = 1$ i $l = 0$ su.

U ovom pristupu, nakon svakog nadigravanja, ulazni parametri formula 8.1 i 8.2 koji ne poprimaju konstantne vrijednosti se ažuriraju. To su: $totalPts_{current}$, $p_{current}$ i $q_{current}$. Nakon ažuriranja vrijednosti tih parametara, formule 8.1 i 8.2 se koriste za izračun vjerojatnosti osvajanja poena na vlastitom servisu momčadi. Svako stanje ima točno dvije isključene vjerojatnosti prijelaza, tako da je moguće opisati odbojkašku utakmicu kao binarno stablo gdje svaki list predstavlja nejedinstveni rezultat utakmice, a svi čvorovi s istim brojem ukupnih bodova su na istoj dubini stabla.

Razlika pristupa predloženog u ovoj doktorskoj disertaciji i pristupa temeljenog na jednolikoj i neovisnoj distribuciji poena je u tome što u modelu predloženom u disertaciji postoji više različitih stanja s istim rezultatom i istom momčadi na servisu. To masovno proširuje prostor stanja modela i povećava njegovu složenost. Pronalaženje analitičkog rješenja za ovakav problem vrlo je zahtjevan zadatak. Zbog transformacije konstantnih prijelaznih vjerojatnosti iz modela temeljenog na jednolikoj i neovisnoj distribuciji poena u evoluirajuće vjerojatnosti, nije moguće koristiti prijelazne matrice i kombinatorički pristup, a nameće se potreba za korištenjem Monte Carlo simulacijske metode.



Slika 8.1: Stablo odbojkaškog seta temeljeno na nejednolikoj i ovisnoj distribuciji

8.2 Monte Carlo simulacija odbojkaških utakmica

8.2.1 Opis osnovne Monte Carlo simulacije odbojkaških utakmica

Polazna točka za Monte Carlo simulaciju odbojkaških utakmica je generator slučajnih brojeva koji može uzorkovati vrijednosti iz uniformne distribucije na intervalu $[0,1]$. Pretpostavimo da parametar p predstavlja vjerojatnost osvajanja poena na servisu momčadi A, a parametar q predstavlja vjerojatnost osvajanja poena na servisu momčadi B. Kada servira momčad A, za svaki poen se uzorkuje jedna vrijednost iz jediničnog intervala. Ako je uzorkovana vrijednost u rasponu $[0,p]$, momčad A osvaja taj poen. U suprotnom momčad A gubi poen, a započinje servirati momčad B. Slično, kada momčad B servira, za svaki poen se uzorkuje vrijednost na jediničnom intervalu. Ako je uzorkovana vrijednost u rasponu $[0,q]$, momčad B osvaja taj poen, u suprotnom momčad B gubi poen. Simulacija poen po poen odvija se na ovaj način prema pravilima bodovanja u odbojci [119].

8.2.2 Napredna metoda ažuriranja vjerojatnosti

Ova doktorska disertacija ima za cilj pokazati da je inkorporiranjem dinamičkih parametara, kroz formulacije kratkoročnog i dugoročnog zamaha u prethodno opisanu Monte Carlo simulacijsku metodu, moguće preciznije simulirati stvarni tijek utakmice. To se dalje može koristiti za bolju procjenu ishoda utakmice, bilo da se procjenjuje pobjednik utakmice, hendikep ili ukupan broj poena koji će se odigrati u utakmici. Predviđanja bi trebala biti preciznija u usporedbi s predviđanjima koja su dobivena modelima temeljenim na pretpostavci o jednolikoj i neovisnoj distribuciji poena. Budući da je izrada preciznijih predviđanja jedan od ciljeva dubinske analize sportskih podataka, naglasak će biti na predviđanju hendikepa i ukupnog broja poena koji se očekuje u utakmici i u potpunosti su dovoljni za opisivanje odbojkaške utakmice [116]. Za predviđanje ovih vrijednosti, a na temelju metodologije predstavljene u prethodnim poglavljima, disertacija predlaže pristup koji prolazi kroz stablo utakmice i ažurira prijelazne vjerojatnosti osvajanja poena na vlastitom servisu nakon svakog odigranog poena pomoću jednadžbi 8.1 i 8.2. Monte Carlo simulacija koja uključuje predloženu hibridnu formulaciju nazvana je naprednom metodom ažuriranja vjerojatnosti [116]. Detaljnije, metoda identificira događaje koji uzrokuju kratkoročni zamah igrača i u tim trenutcima adekvatno prilagođava vjerojatnosti osvajanja poena na vlastitom servisu - ako momčad osvoji poen na vlastitom servisu, vjerojatnost osvajanja idućeg poena na vlastitom servisu se povećava. Vjerojatnost se kumulativno povećava sve do trenutka kada momčad osvoji četiri poena na vlastitom servisu. Nakon tog događaja u utakmici, vjerojatnosti osvajanja poena na idućim servisima (petom pa na dalje) ostaju jednake vrijednosti vjerojatnosti osvajanja poena na vlastitom servisu nakon tri uzastopna osvojena poena momčadi. Kada momčad izgubi poen na vlastitom servisu, kratkoročni zamah se poništava.

Istovremeno navedenome, nakon svakog odigranog poena u utakmici, metoda, kroz formulaciju dugoročnog zamaha, prilagođava povijesne vjerojatnosti novim statistikama.

Metodom se konačno mogu predviđati ishodi odbojkaških utakmica na sljedeći način [116]:

1. Monte Carlo simulacijom, koja uključuje hibridnu formulaciju za inkorporiranje kratkoročnog i dugoročnog zamaha, generiraj jedno stablo utakmice koje predstavlja jedan mogući tijek utakmice.
2. Iz generiranog stabla utakmice izračunaj hendikep i ukupni broj odigranih poena.
3. Ponovi korake 1 i 2 više puta (u doktorskoj disertaciji koraci su ponovljeni 10^6 puta).
4. U slučaju predviđanja ishoda pojedinačne utakmice, izračunaj odgovarajuću mjeru centralne tendencije (ovisno o obliku rezultantne distribucije) hendikepa i ukupnog broja poena. Ta mjera predstavlja konačno predviđanje.

Metoda je formalizirana kroz pseudokod u nastavku [116]. *Napomena*: u pseudokodu su uvedene nove varijable čije značenje, do sada, nije objašnjeno u disertaciji. U nastavku je dan njihov opis:

- $ownWon_p$/$ownWon_q$ - broj osvojenih poena na vlastitom servisu momčadi domaćina/gosta u iteraciji simulacije,
- $ownServed_p$/$ownServed_q$ - broj servisa momčadi domaćina/gosta u iteraciji simulacije,
- *serve* - zastavica koja označava koja momčad je na servisu u iteraciji simulacije (poprima vrijednost 1 kada servira momčad domaćin, u suprotnom poprima vrijednost 0),
- $points_p/points_q$ - ukupan broj osvojenih poena momčadi domaćina/gosta u iteraciji simulacije,
- $momentum_{counter}$ - red kratkoročnog zamaha u iteraciji simulacije,
- x - vektor slučajnih brojeva iz uniformne razdiobe,
- $start_p/start_q$ - zastavica koja označava je li momčad domaćin/gost servirala barem jedan poen (poprima vrijednost 0 ako momčad još nije servirala niti jedan poen, u suprotnom poprima vrijednost 1).

Algorithm 1 Napredna metoda ažuriranja vjerojatnosti

Input: $p_{hist}, q_{hist}, k, l, totalPts_{exp}, m_1^p, m_1^q, m_2^p, m_2^q, m_3^p, m_3^q$
Output: $totalPts_{pred}, handicap_{pred}$

Initialize: $ownWon_p = 0, ownWon_q = 0, ownServed_p = 0, ownServed_q = 0, totalPts_{current} = 0, serve = 1, points_p = 0, points_q = 0, momentumCounter = 0, totalPts_{pred} = 0, handicap_{pred} = 0, x = runif(400), start_p = 0, start_q = 0$
top:

if $start_p == 0$ **then**
    $p_{current} = 0$
else if $start_p == 1$ **then**
    $p_{current} = ownWon_p / ownServed_p$
end if
if $start_q == 0$ **then**
    $q_{current} = 0$
else if $start_q == 1$ **then**
    $q_{current} = ownWon_q / ownServed_q$
end if
if $(serve == 1)$ **then**
    p_{ltm} = equation 7.6
    $p_{new} = p_{ltm}$
    $ownServed_p += 1$
    $start_p = 1$
    if $(momentumCounter == 1)$ **then**
        $p_{new} = (1 + m_1^p) * p_{ltm}$
    else if $(momentumCounter == 2)$ **then**
        $p_{new} = (1 + m_2^p) * p_{ltm}$
    else if $(momentumCounter >= 3)$ **then**
        $p_{new} = (1 + m_3^p) * p_{ltm}$
    end if
    if $(x[i] < p_{new})$ **then**
        $ownWon_p += 1$
        $points_p += 1$
        $momentumCounter += 1$
    else
        $momentumCounter = 0$
        $points_q += 1$
        $serve = 0$
    end if
    $totalPts_{current} += 1$
else if $(serve == 0)$ **then**
    q_{ltm} = equation 7.7
    $q_{new} = q_{ltm}$
    $ownServed_q += 1$
    $start_q = 1$

Algorithm 1 Napredna metoda ažuriranja vjerojatnosti (nastavak)

41: **if** $(momentumCounter == 1)$ **then**
42: $q_{new} = (1 + m_1^q) * q_{ltm}$
43: **else if** $(momentumCounter == 2)$ **then**
44: $q_{new} = (1 + m_2^q) * q_{ltm}$
45: **else if** $(momentumCounret >= 3)$ **then**
46: $q_{new} = (1 + m_3^q) * q_{ltm}$
47: **end if**
48: **if** $(x[i] < q_{new})$ **then**
49: $ownWon_q += 1$
50: $points_q += 1$
51: $momentumCounter += 1$
52: **else**
53: $momentumCounter == 0$
54: $points_p += 1$
55: $serve = 1$
56: **end if**
57: $totalPts_{current} += 1$
58: **end if**
59: **if** $(SetIsOver)$ **then** ▷ provjeri je li set zavrsio
60: $totalPts_{pred}$ = $totalPts_{pred}$ + $points_p$ + $points_q$
61: $handicap_{pred}$ = $handicap_{pred}$ + $points_p$ - $points_q$
62: $points_p = 0$
63: $points_q = 0$
64: **end if**
65: **if** $(MatchIsNotOver)$ **then** ▷ provjeri je li utakmica zavrsila
66: **goto** *top*.
67: **else**
68: **return** $totalPts_{pred}$, $handicap_{pred}$
69: **end if**

Kako bi simulirali utakmicu koristeći naprednu metodu ažuriranja vjerojatnosti potrebno je opisati utakmicu s numeričkim parametrima koji predstavljaju profile izvedbe momčadi. Profili se grade iz povijesnih podataka. Više o gradnji profila dano je u poglavlju 8.3.

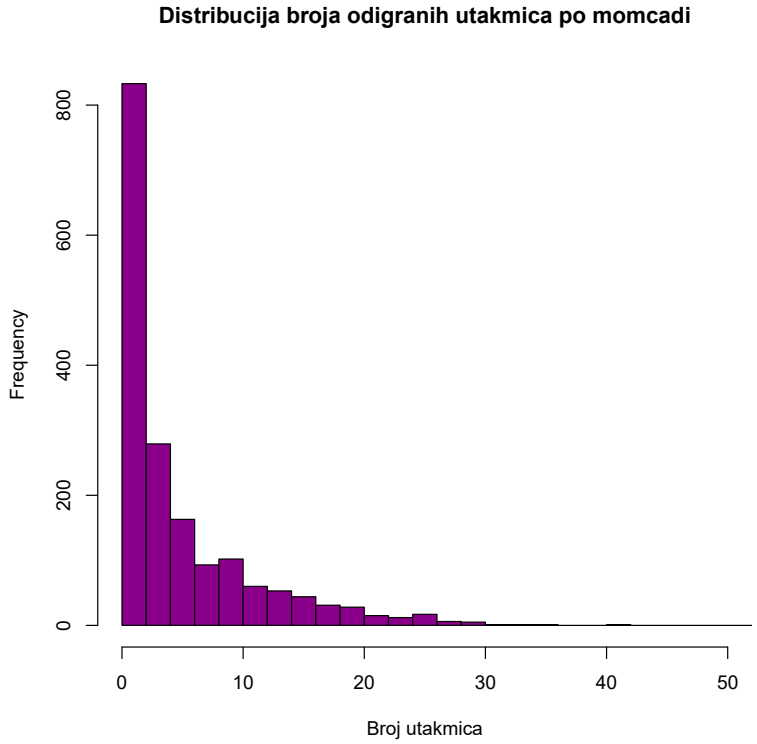
8.3 Profiliranje momčadi

Poznato je da na kvalitetu izlaza prediktivnog modela ponajviše utječe kvaliteta i količina ulaznih podataka. U analizi odbojkaških utakmica, vrlo važnim prediktorom pokazala se vjerojatnost osvajanja poena na vlastitom servisu igrača odnosno momčadi, a u znanstvenoj literaturi ističe se kompleksno područje istraživanja fokusirano upravo na izračun tog parametara (više detalja u poglavlju 4.2). Ova disertacija, osim parametra vjerojatnosti osvajanja poena na vlastitom servisu momčadi, u prediktivni model uključuje i parametre sportske dinamike.

Kada se modeli za predviđanje ishoda sportskih događaja implementiraju u sustave klađenja, radi preciznosti modela, potrebno je periodički ponavljati analize na razini igrača, momčadi ili turnira. Jednostavni sustavi rangiranja i prosječne statistike u većini slučajeva nisu dovoljne, a njihove rezultate često krivo interpretiraju čak i neke od vodećih kladionica [116].

Zamislimo na trenutak idealan scenarij u kojem posjedujemo dovoljnu količinu povijesnih odbojkaških utakmica koje se mogu upotrijebiti za konstrukciju profila performansi svake momčadi. Ti bi se podatci tada mogli iskoristiti kako bi se, za svaku momčad, izračunale vjerojatnosti osvajanja poena na vlastitom servisu s obzirom na snagu momčadi protivnika. Povijesni podaci bi se također mogli iskoristiti kako bi se za svaku momčad, ovisno o snazi protivničke momčadi, izračunali parametri kratkoročnog zamaha, a mogli bi se iskoristiti i za grubu procjenu očekivanog broja poena u utakmici koja je potrebna za izračun dugoročnog zamaha. Tako izračunati parametri bi se tada mogli upotrijebiti u predloženoj naprednoj metodi ažuriranja vjerojatnosti za procjenu hendikepa i ukupnog broja poena.

Idealan scenarij u stvarnom svijetu ne postoji, a heterogenost prikupljenog podatkovnog skupa ovoj doktorskoj disertaciji ne dozvoljava korištenje opisanog pristupa - uz korišteni podatkovni skup izgradnja individualnih profila momčadi nije moguća. Korišteni podatkovni skup, naime, ne sadrži dovoljno povijesnih utakmica za veliki dio momčadi sadržanih u njemu (slika 8.2), a prikupljanje većeg podatkovnog skupa za izradu ove doktorske disertacije nije bilo moguće. Iz navedenih razloga doktorska disertacija predlaže pristup koji kompenzira problem relativno malog broja opservacija u koraku profiliranja momčadi. Pristup se odmiče od ideje individualnog profiliranja momčadi, a predlaže konstrukciju profila grupe. Pri tome predlaže da grupe ne agregiraju performanse sličnih momčadi već svojstva utakmica u kojima su suprotstavljene momčadi sličnih omjera snaga [116]. Tako primjerice, jedna od grupa sadrži sve utakmice sa snažnim favoritom, a druga pak sadrži utakmice u kojima igraju momčadi sličnih snaga. Ovim pristupom korišteni podatkovni skup upotrijebljen je učinkovitije, a pojedinačne momčadi zastupljene su u više grupa ovisno o performansama svog protivnika. To znači da se iste momčadi opisuju s drugačijim profilima koji variraju ovisno o performansama momčadi protivnika [116]. Tako je snaga momčadi protivnika ugrađena u profil grupe te o toj snazi više nije potrebno posebno voditi računa. U svrhu realizacije opisanog, disertacija predlaže dva mo-



Slika 8.2: Graf razdiobe odigranih utakmica po momčadi

dula - prvi modul dijeli utakmice u grupe utakmica u kojima su suprotstavljene momčadi sličnih omjera snaga i gradi profile tih grupa na temelju povijesnih podataka, a drugi modul služi za validaciju modela.

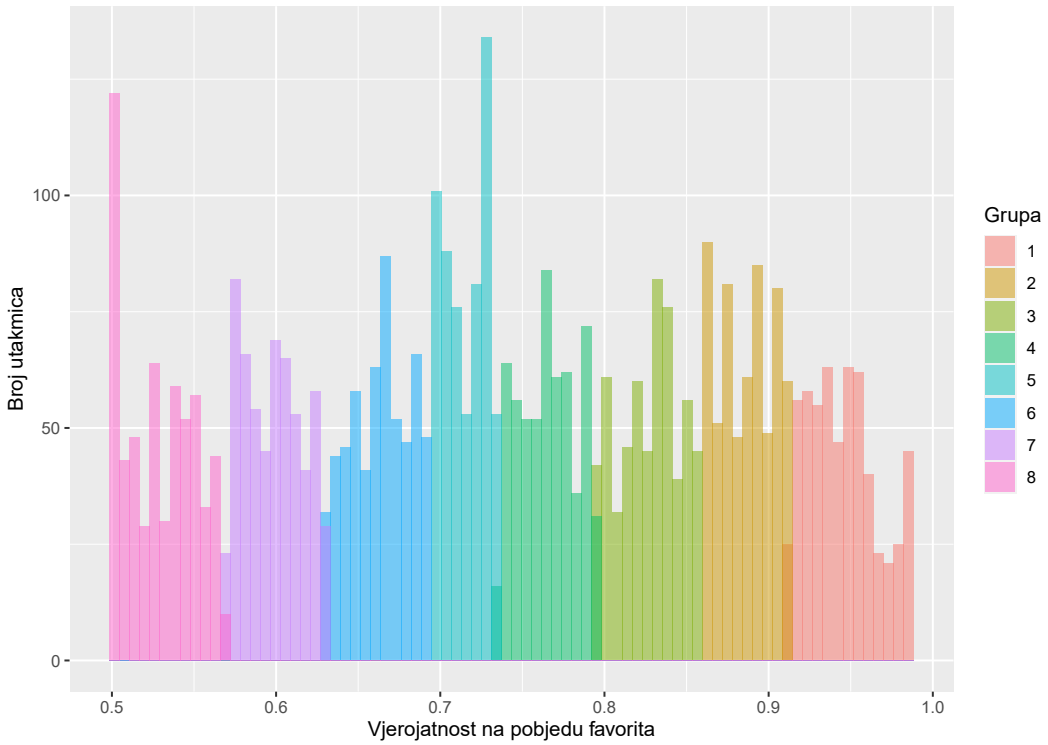
Modul 1. Kao što je rečeno, prvi modul služi za grupiranje utakmica i za izgradnju profila grupa utakmica. U te svrhe modul koristi podatkovni skupa za treniranje (70% cijelog podatkovnog skupa).

- **Grupiranje utakmica**. Shinovom normalizacijom pre-match koeficijenata, za obje momčadi se računaju vjerojatnosti osvajanja utakmice (više detalja u poglavlju 2.3.2). Izračunate vjerojatnosti tada se koriste kako bi se odredila momčad domaćin i momčad gost * (u doktorskoj disertaciji kao momčad domaćin odabrana je momčad s boljim performansama). Nakon toga, podatkovni skup se sortira prema izračunatoj pre-match vjerojatnosti osvajanja utakmice momčadi domaćina. Sortirani podatkovni skup konačno se "reže" na manje dijelove sličnih veličina. Drugim riječima, nad sortiranim skupom podataka definiraju se grupe utakmica s ograničenjem da svaka grupa ima podjednaki broj utakmica i nije preširoka [116]. Kako odrediti optimalan broj grupa? Broj grupa ne smije biti niti

*U doktorskoj disertaciji vjerojatnost osvajanja poena na vlastitom servisu za momčad domaćina i momčad gosta označava se s oznakama p i q, tim redoslijedom. Ako podatci ne sadrže informaciju o tome koja momčad je momčad domaćin, a koja momčad je momčad gost ili pak takva definicija nije prihvatljiva u određenom sportu, tada se kao momčad domaćin definira momčad s boljim povijesnim performansama. U slučaju da suprotstavljene momčadi imaju iste povijesne performanse, momčad domaćin odabire se nasumično. Ovaj pristup nužan je kako bi se održala dosljedna notacija u svim utakmicama podatkovnog skupa.

premalen niti prevelik (osim ako to dopušta veličina podatkovnog skupa). Malen broj grupa rezultirao bi nekim oblikom nakošene distribucije. Svaka od tih grupa sadržavala bi veliki broj utakmica, a razlika između omjera snaga momčadi na krajevima tih grupa bila bi prevelika. Uprosječe li se performanse momčadi takvih grupa i iskoriste li se za simuliranje tih utakmica, ne bi bilo moguće kvalitetno odsimulirati utakmice na rubovima grupa. Veći broj grupa bi pak rezultirao preciznijim simulacijama i boljim podudaranjem distribucija, međutim u tom slučaju podatkovni skup bi trebao biti puno veći [116]. U doktorskoj disertaciji odabrano je 8 grupa (slika 8.3). Broj grupa odabran je kao kompromis između veličine korištenog podatkovnog skupa i sličnosti utakmica unutar određene grupe [116].

- **Konstrukcija profila grupa**. Sve utakmice jedne grupe koriste se za izgradnju profila te grupe. Profili pojedinačnih grupa prikazani su tablicom 8.1 [116]. Profil za svaku grupu utakmica sastoji se od parametara vjerojatnosti osvajanja poena na vlastitom servisu za momčad domaćina (p) i momčad gosta (q), parametara kratkoročnog zamaha prvog (m_1^p), drugog (m_2^p) i trećeg (m_3^p) reda za momčad domaćina, te parametara kratkoročnog zamaha prvog (m_1^q), drugog (m_2^q) i trećeg (m_3^q) reda za momčad gosta. Profil također sadrži i grubu procjenu očekivanog broja poena ($totalPts_{exp}$) koja je potreba za izračun dugoročnog zamaha. Posljednja dva stupca ($p_{winfrom}$, p_{winto}) u tablici 8.1 predstavljaju granične vrijednosti vjerojatnosti osvajanja utakmice od strane favorita za određenu grupu utakmica (gornja granica je uključena) [116]. Tako primjerice prva grupa sadrži utakmice u kojima je vjerojatnost na pobjedu favorita veća od $91,266\%$, a posljednja grupa utakmica sadrži utakmice u kojima je vjerojatnost na pobjedu favorita između $50,000\%$ i $56,613\%$ (uključivo). Te vjerojatnosti su, kako je opisano prethodno, izračunate iz pre-match koeficijenata koristeći Shinovu normalizaciju, a služe kako bi se dobio osjećaj o omjeru snaga suprotstavljenih momčadi u toj grupi utakmica te kako bi se utakmica za koju se predviđa ishod mogla smjestiti u odgovarajuću grupu (više o tome u nastavku).



Slika 8.3: Podjela utakmica u grupe utakmica sličnih omjera snaga

Tablica 8.1: Profili grupa odbojkaških utakmica

Grupa	p (%)	q (%)	m_1^p (%)	m_2^p (%)	m_3^p (%)	m_1^q (%)	m_2^q (%)	m_3^q (%)	$totalPts_{exp}$	$p_{winfrom}$ (%)	p_{winto} (%)
1	41.55	28.55	2,95	11,73	13,07	15,17	24,64	43,84	134,0	91,266	100,000
2	39.34	29.38	6,23	11,45	9,40	9,61	23,40	28,48	139,5	85,934	91,266
3	38.63	30.93	3,02	7,50	7,09	8,07	21,03	31,57	162,0	79,619	85,934
4	37.27	32.20	4,74	15,47	23,58	9,23	19,56	39,01	174,0	73,674	79,619
5	36.34	32.58	4,44	12,05	27,68	11,59	17,89	28,71	181,0	69,321	73,674
6	36.67	33.14	5,49	9,43	24,13	8,26	17,57	31,13	180,0	62,995	69,321
7	34.91	33.83	4,58	19,86	33,46	6,40	12,42	33,34	182,0	56,613	62,995
8	35.41	34.35	1,44	16,37	32,74	5,33	16,30	30,20	180,0	50,000	56,613

Modul 2. Drugi modul služi za testiranje napredne metode ažuriranja vjerojatnosti i koristi podatkovni skup za testiranje (30%). Zbog opisanog ograničenja korištenog podatkovnog skupa, predviđanje ishoda pojedinačnih utakmica zamjenjuje se predviđanjem distribucija ishoda sportskih događaja određene grupe utakmica. Detaljnije, i utakmice testnog skupa podataka, na temelju pre-match koeficijenata, dijele se u 8 grupa definiranih stupcima $p_{winfrom}$ i p_{winto} u tablici 8.1. Svaka grupa utakmica testnog skupa podataka ima odgovarajuću distribuciju ishoda sportskog događaja od interesa. Izračunati parametri profila grupa (modul 1) koriste se kao aproksimativne vrijednosti za simuliranje utakmica odgovarajuće grupe testnog skupa podataka. Parametri odgovarajućeg profila koriste se kao ulaz u predloženoj naprednoj metodi ažuriranja vjerojatnosti.

Konačno je potrebno napomenuti da je svaki igrač, a tako i momčad, jedinstven. To znači da igrači i momčadi usporedivih performansi mogu reagirati potpuno drugačije u sličnim okolnostima. Zato je potrebno analizirati zamahe i reakcije za svaku momčad i igrača zasebno. Ovakav pristup se najčešće primjenjuje u velikim kladionicama, ali kao nužnost postavlja potrebu za velikom skupom podataka koji sadrži dovoljno informacija o povijesnim susretima svakog igrača i svake momčadi. Kada bi veličina podatkovnog skupa dozvolila profiliranje na razini igrača, proces izgradnje modela za predviđanje ishoda odbojkaških utakmica postao bi kompleksniji, a u obzir bi trebalo uzeti postave i rotacije obiju momčadi. Pristup predstavljen u doktorskoj disertaciji predlaže jedan od mogućih načina rješavanja problema na realnom skupu podataka. Opisani pristup predstavlja učinkovit način iskorištavanja ograničenog podatkovnog skupa [116].

8.4 Generalizacija napredne metode ažuriranja vjerojatnosti na druge sportove s bodovnim ograničenjem

Iako je fokus doktorske disertacije primarno na modeliranju i analizi utakmica dvoranske odbojke, teorije i principi razvijeni u ovoj doktorskoj disertaciji mogu se primijeniti i na druge sportove s bodovnim ograničenjem poput odbojke na pijesku, tenisa, stolnog tenisa, badmintona i sličnih sportova. U procesu prilagodbe napredne metode ažuriranja vjerojatnosti na druge sportove s bodovnim ograničenjem, ističu se 3 bitna koraka:

1. **Identifikacija parametara sportske dinamike i matematičko modeliranje identificiranih parametara**. U različitim sportovima s bodovnim ograničenjem moguće je identificirati drugačije parametre sportske dinamike, a isti parametri sportske dinamike, u različitim sportovima s bodovnim ograničenjem mogu imati drugačiji utjecaj na dinamiku utakmice te na točnost i preciznost predviđanja. U procesu identifikacije parametara sportske dinamike potrebno je voditi računa o prirodi i dinamici sporta te pravilima i strukturi bodovanja u sportu. Doktorska disertacija, u ovom trenutku, razmatra samo parametre kratkoročnog i dugoročnog zamaha. Matematički modeli parametara kratkoročnog i dugoročnog zamaha, u domeni odbojke, detaljno su objašnjeni u poglavljima 7.1 i 7.2. Ponuđene formulacije mogu se primijeniti i na druge sportove s bodovnim ograničenjem. Tu je međutim, pri definiranju i izračunu vrijednosti zamaha u formulama 7.1 i 7.2 kao i pri izračunu težina u formulama 7.3 i 7.4, u obzir potrebno uzeti specifičnosti određenog sporta koji se modelira. Fokusirajmo se na trenutak na domenu odbojke i tenisa *.

   Rezimirajmo saznanja o parametru kratkoročnog zamaha u domeni dvoranske odbojke - uzmu li se u obzir pravila i struktura bodovanja u domeni dvoranske odbojke detaljno opisana u poglavlju 2.1 kao i znanje ekstrahirano eksploratornom analizom korištenog podatkovnog skupa, lako je za uočiti da se u domeni dvoranske odbojke kratkoročni zamah događa i kumulativan je (više detalja u poglavlju 7.1). Kratkoročni zamah najčešće se događa kada je na servisu dobar server ili pak kada je na terenu dobra postava. Jednom kada momčad izgubi poen na vlastitom servisu, kratkoročni zamah se gubi, a sa servisom započinje druga momčad.

   Uzmimo sada za primjer tenis. Ukratko, tenis je igra u kojoj je potrebno osvojiti dovoljan broj poena kako bi se osvojio i gem. Osvajanjem dovoljnog broja gemova, igrač osvaja set, a u konačnici osvajanjem dovoljnog broja setova igrač osvaja meč. Tijekom cijelog

*Pokrivanjem domene odbojke i tenisa, pokriven je problem definiranja parametara kratkoročnog i dugoročnog zamaha za širok spektar sportova s bodovnim ograničenjem. Odbojka, naime, predstavlja ostale sportove s bodovnim ograničenjem u kojima momčad nastavlja servirati sve dok, u nadigravanju, osvaja poen na vlastitom servisu. Tenis, s druge strane, predstavlja sportove u kojima igrač/par zadržava servis unutar određenog perioda, čak i ako izgubi poen na vlastitom servisu.

gema servira isti igrač. Set se sastoji od gemova, a unutar jednog seta igrači se međusobno izmjenjuju na servisu (detaljno objašnjenje strukture bodovanja u tenisu može se naći u [9]). Upravo ovakva pravila i struktura bodovanja dovode do potpuno drugačije dinamike teniskog meča u odnosu na dinamiku odbojkaške utakmice - u cijelom teniskom gemu servira isti igrač pa gubitak poena na vlastitom servisu igrača neće naglo promijeniti dinamiku meča. Iz tog razloga, kratkoročni zamah definiran na način opisan u ovoj doktorskoj disertaciji ne bi trebao utjecati na dinamiku teniskog meča u mjeri u kojoj on utječe na dinamiku odbojkaške utakmice - iako ga je, bez izmjena, moguće ugraditi u model za predviđanje ishoda teniskog meča. U domeni tenisa zanimljivije je analizirati kratkoročni zamah na razini gema ili seta. Formulacija kratkoročnog zamaha i u tom slučaju može ostati nepromijenjena, a vrijednost zamaha (m_i) tada se može izračunati kao razlika u vjerojatnosti osvajanja poena u gemu/setu koji se igra nakon što je igrač osvojio prethodni gem/set i u vjerojatnosti osvajanja poena u gemu/setu koji se igra nakon što je igrač prethodno izgubio gem/set.

Parametar dugoročnog zamaha, u obliku definiranom u ovoj doktorskoj disertaciji, primjenjiv je i na druge sportove s bodovnim ograničenjem. Međutim, prilikom modeliranja parametra dugoročnog zamaha u druge sportove s bodovnim ograničenjem, potrebno je biti oprezan s omjerom težina povijesne i simulacijske statistike. Taj omjer moguće je odrediti na sličan način kako je to objašnjeno i u slučaju dvoranske odbojke (više detalja u poglavlju 8.5).

2. **Profiliranje igrača/momčadi**. Nakon što se definiraju matematički modeli parametara sportske dinamike, problem procjene varijabli potrebnih za izračun tih parametara kao i problem izračuna ostalih parametara potrebnih za izgradnju kvalitetnog profila igrača odnosno momčadi svodi se isključivo na problem veličine i kvalitete podatkovnog skupa (više detalja u poglavlju 8.3).
3. **Inkorporiranje matematičkih modela identificiranih parametara sportske dinamike u Monte Carlo simulaciju**. Svaki sport ima svoje specifičnosti i pravila, te će, zbog toga, implementacija Monte Carlo simulacije varirati od sporta do sporta. Kako bi se u nju ugradili parametri sportske dinamike, potrebno je, tijekom simulacije, identificirati događaje koji uzrokuju promjenu dinamike utakmice te ih kroz odgovarajuće formulacije (korak 1) ugraditi u Monte Carlo simulaciju.

8.5 Optimizacija parametara napredne metode ažuriranja vjerojatnosti

Hiperparametri su parametri čije se vrijednosti koriste za kontrolu procesa učenja u prediktivnim modelima. U predloženoj naprednoj metodi ažuriranja vjerojatnosti, hiperparametri su parametri linearne funkcije, l i k, koji su potrebni za implementaciju dugoročnog zamaha (više detalja u poglavlju 7.2). Ostali parametri predložene napredne metode ažuriranja vjerojatnosti direktno su izračunati iz povijesnih podataka.

U svrhe optimizacije hiperparametara predložene napredne metode ažuriranja vjerojatnosti odabran je popularni pristup pretraživanja mreže predefiniranih parametara (engl. *grid search*). Radi se o iscrpnoj pretrazi parametara koja se provodi u prostoru specifičnih vrijednosti. Doktorska disertacija predlaže optimalnu kombinaciju parametara l i k kao onu kombinaciju parametara koja minimizira ukupnu grešku simulacije, odnosno maksimizira površine poklapanja histograma realnih i simulacijskih podataka za sve grupe mečeva, kako za predviđanje hendikepa tako i za predviđanje ukupnog broja poena koji se očekuje u utakmici [9] (više detalja u poglavlju 9).

S obzirom na prirodu optimizacijskog problema odabrana je dvorazinska L2 norma odnosno mjera Euklidske udaljenosti [9] (formule 8.3, 8.4, 8.5). L2 norma, u osnovi, minimizira zbroj kvadrata razlike između ciljane vrijednosti i predviđene vrijednosti, odnosno zbroj kvadrata simulacijskih pogrešaka. Ona kod problema minimizacije pogreške predviđanja daje bolje performanse od L1 norme poznatije pod pojmom *najmanje apsolutno odstupanje* (engl. *least absolute deviations*, LAD). L2 norma osjetljiva je na iskočnice, a na taj način sprječava prilagođavanje hiperparametara samo određenoj grupi utakmica - velika simulacijska pogreška u određenoj grupi utakmica ne smije biti kompenzirana malim pogreškama u ostalim grupama utakmica, kako kod predviđanja ukupnog broja poena (8.3), tako i kod predviđanja hendikepa (8.4). L2 norma implementirana je na dvije razine budući da se traži jedinstvena kombinacija vrijednosti parametara l i k koja minimizira ukupnu (kombiniranu) simulacijsku pogrešku i hendikepa i ukupnog broja poena u utakmici [9] (8.5).

$$E_{ukBrPoena} = \sqrt[2]{E^2_{ukBrPoena1} + E^2_{ukBrPoena2} + ... + E^2_{ukBrPoena8}}, \tag{8.3}$$

$$E_{hendikep} = \sqrt[2]{E^2_{hendikep1} + E^2_{hendikep2} + ... + E^2_{hendikep8}}, \tag{8.4}$$

$$E = \sqrt[2]{E^2_{ukBrPoena} + E^2_{hendikep}}. \tag{8.5}$$

Parametri formula 8.3, 8.4, 8.5 su:

$E_{ukBrPoena}, i \in [1,8]$ - greška u predviđanju ukupnog broja poena u utakmici za grupu i.
$E_{hendikepi}, i \in [1,8]$ - greška u predviđanju hendikepa u utakmici za grupu i.
$E_{ukBrPoena}$ - ukupna greška u predviđanju ukupnog broja poena (za sve grupe).
$E_{hendikep}$ - ukupna greška u predviđanju hendikepa (za sve grupe).
E - ukupna greška modela.

Za pretraživanje optimalnih vrijednosti parametara l i k korišten je podatkovni skup za treniranje (70% ukupnog podatkovnog skupa). Optimizacijskim postupkom konačno su odabrane vrijednosti $k = 1$ i $l = 300$. Testni skup (30% ukupnog podatkovnog skupa) je korišten za evaluaciju i prezentiranje rezultata.

Poglavlje 9

Rezultati i diskusija

Uključivanje parametara sportske dinamike u modele za predviđanje ishoda sportskih događaja intuitivan je korak u procesu poboljšavanja prediktivnih modela koji je podržan i znanjem domenskih eksperata [48, 63, 120, 121, 122]. Međutim, podržavaju li i rezultati detaljnih analiza uključivanje dinamičkih parametara u modele za predviđanje utakmica sportova s bodovnim ograničenjem? Odgovor na ovo pitanje dat će ostatak ovog poglavlja.

Doktorska disertacija primarno analizira utjecaj parametra kratkoročnog i dugoročnog zamaha na predviđanje hendikepa i ukupnog broja poena koji se očekuje u utakmici. Klađenje na hendikep i ukupan broj poena koji se očekuje u utakmici dva su najpopularnija tipa klađenja u odbojci (više detalja u poglavlju 2.2). Oni opisuju omjer snaga između suprotstavljenih momčadi (hendikep) i duljinu utakmice (ukupan broj poena koji će se odigrati u utakmici).

U nastavku poglavlja dana je usporedba stvarnih distribucija hendikepa i ukupnog broja poena u različitim grupama utakmicama testnog skupa podataka s distribucijama koje se dobiju iz različitih Monte Carlo simulacija. Parametri sportske dinamike, izračunati iz odgovarajuće grupe utakmica trening skupa podataka, u simulacije se uvode postepeno kako bi se mogao analizirati njihov utjecaj na predviđanja. Detaljnije, prvo se uspoređuje realna distribucija s distribucijom koja se dobije iz simulacije temeljene na jednolikoj i neovisnoj distribuciji poena. Zatim se u tu simulaciju uključuje parametar kratkoročnog zamaha te se analizira njegov utjecaj na predviđanja. Potom se parametar kratkoročnog zamaha isključuje, a uključuje se parametar dugoročnog zamaha i analizira se utjecaj upravo tog parametra na točnost predviđanja. Konačno se realna distribucija hendikepa i ukupnog broja poena uspoređuje s distribucijom koja se dobije iz Monte Carlo simulacije koja uključuje i kratkoročni i dugoročni zamah. Rezultati su prikazani grafički, a što je veće poklapanje grafa realne i simulacijske distribucije, to je bolje predviđanje.

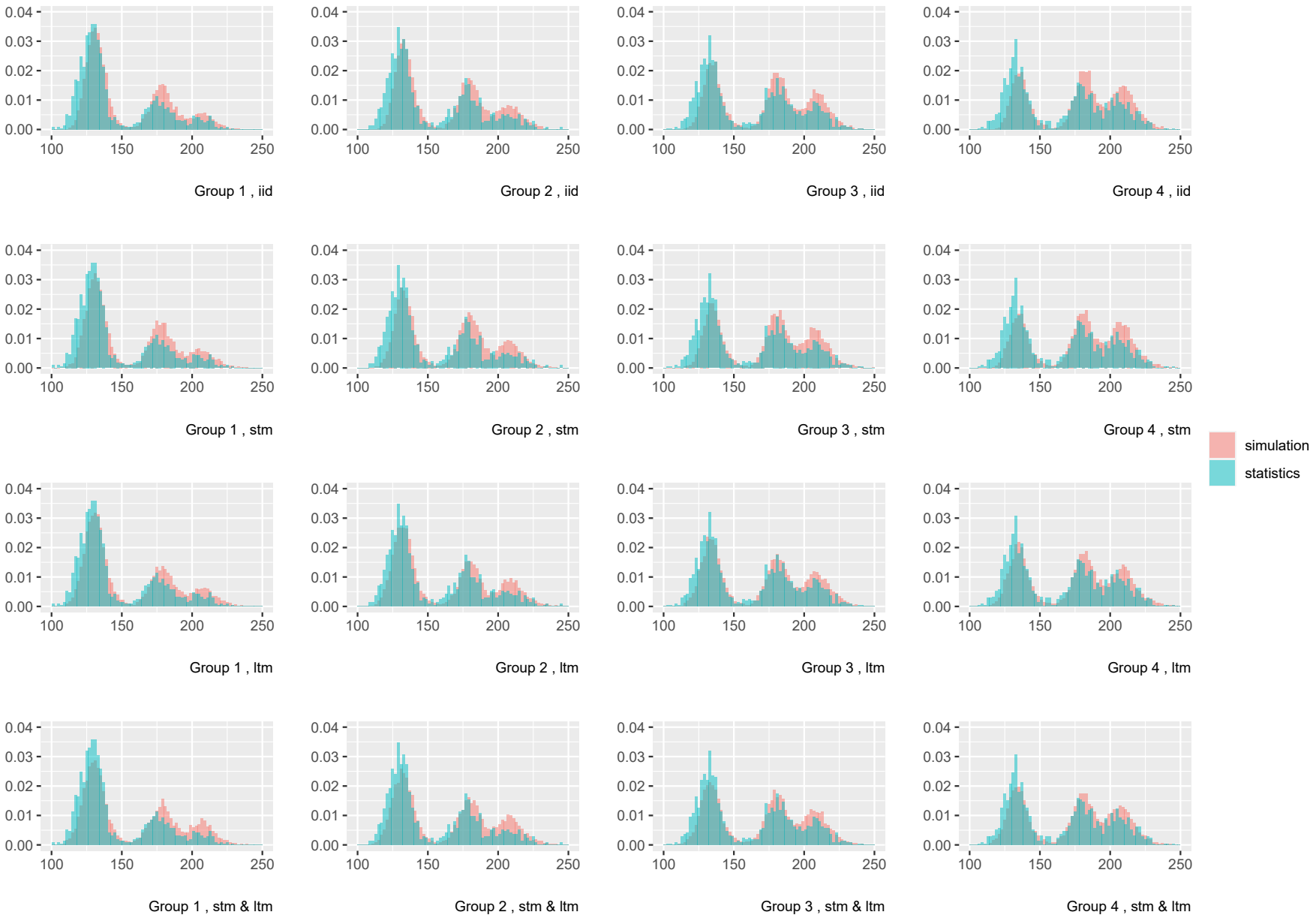
Na temelju svih utakmica u jednoj grupi testnog skupa podataka i na temelju utakmica simuliranih odgovarajućom Monte Carlo simulacijom s ugrađenim odgovarajućim profilom grupe, kreirani su histogrami preklapanja ukupnog broja poena te hendikepa što je ponovljeno za sve

grupe utakmica. Konačni rezultati prikazani su slikama 9.1, 9.2, 9.3 i 9.4 [116].

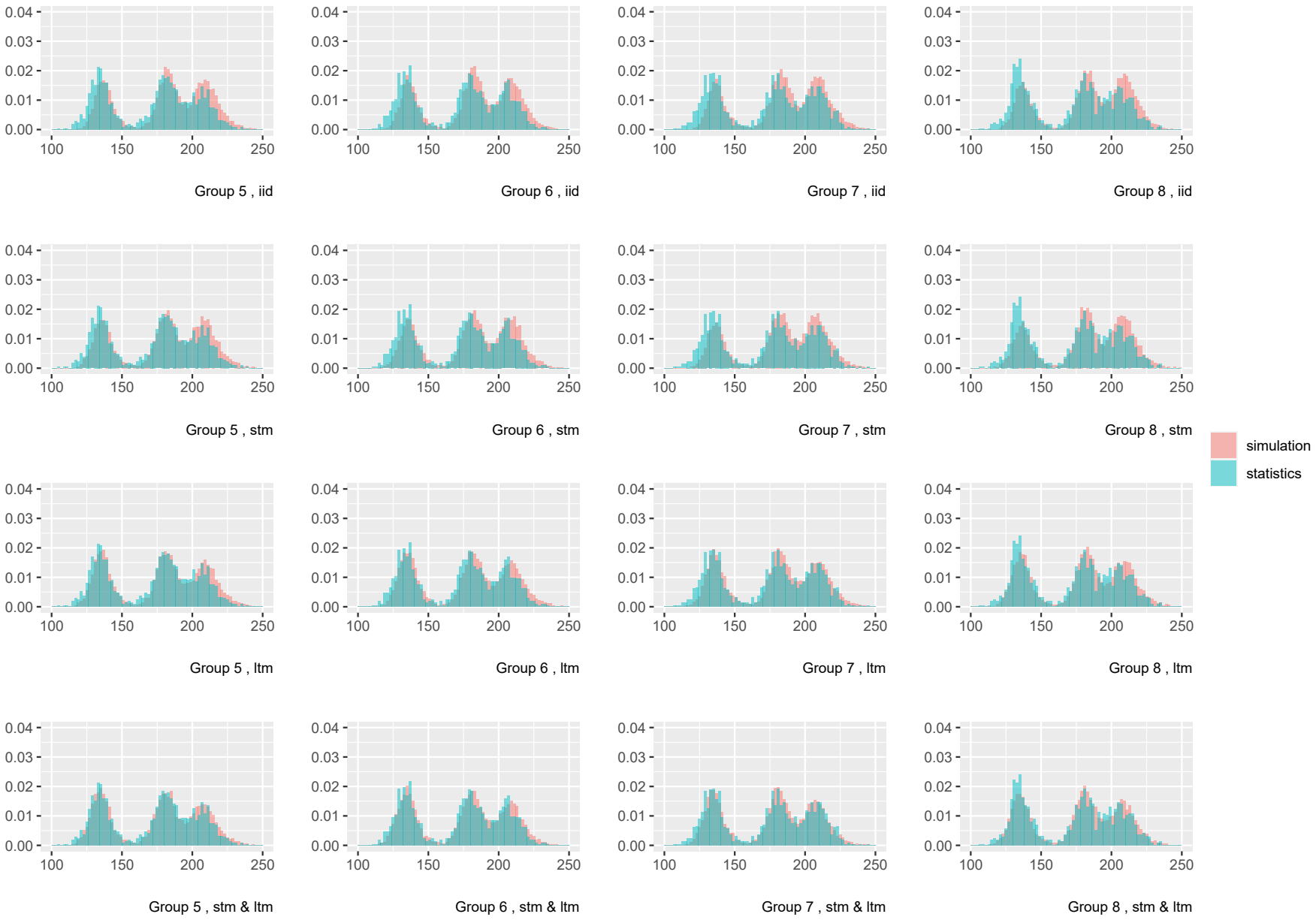
Slika 9.1 [116] prikazuje preklapajuće histograme raspodjele ukupnog broja poena u utakmicama za prve četiri grupe utakmica. Histogrami plave boje prikazuju stvarnu raspodjelu ukupnog broja poena za odgovarajuću grupu utakmica, a histogrami crvene boje prikazuju distribucije ukupnog broja poena dobivene simulacijama. Svaki stupac na slici 9.1 prikazuje jednu od grupa utakmica (prvi stupac - prva grupa, drugi stupac - druga grupa itd.), a svaki redak prikazuje rezultate jedne od četiriju simulacija. Prvi redak tako uspoređuje realnu distribuciju s distribucijom ukupnog broja poena koja se dobila simulacijama temeljenim na jednolikoj i neovisnoj distribuciji poena (*iid*). Drugi redak prezentira realne distribucije u usporedbi s rezultatima simulacija koje inkorporiraju kratkoročni zamah (*stm*), dok treći redak uspoređuje realne distribucije i rezultate simulacija koje uključuju dugoročni zamah (*ltm*). Konačno, zadnji redak prikazuje rezultate simulacija koje uključuju i kratkoročni i dugoročni zamah (napredna metoda ažuriranja vjerojatnosti, *stm i ltm*) u usporedbi s realnim distribucijama.

Slično tome, slika 9.2 [116] prikazuje preklapajuće histograme raspodjele ukupnog broja poena za grupe utakmica u intervalu od 5 do 8 (obje granice su uključene). Slika 9.3 prikazuje preklapajuće histograme distribucija hendikepa za grupe utakmica od 1 do 4 (obje granice su uključene), a slika 9.4 prikazuje preklapajuće histograme distribucija hendikepa za grupe utakmica u intervalu od 5 do 8 (obje granice su uključene).

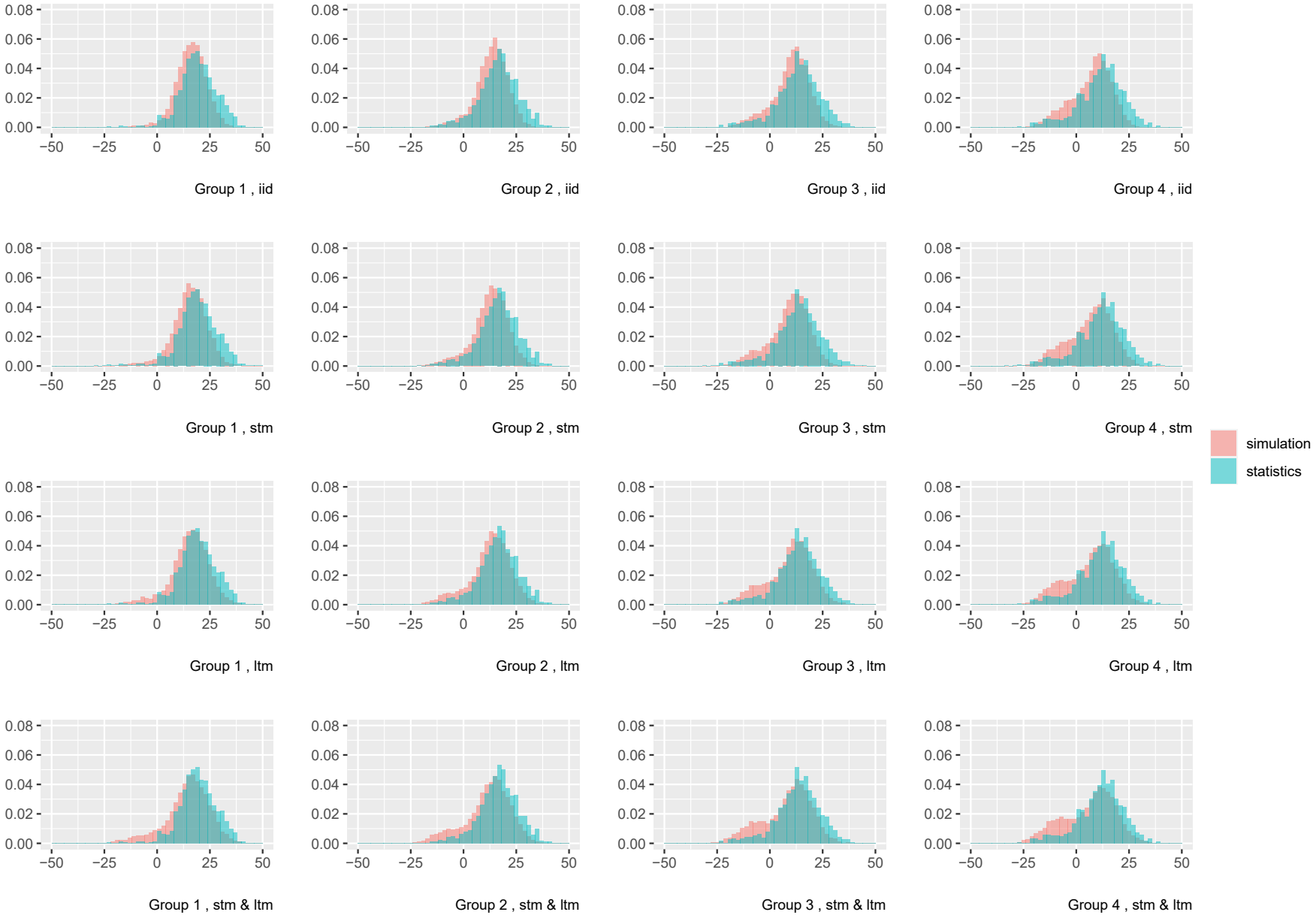
Gleda li se prvi redak histograma na slikama 9.1, 9.2, 9.3 i 9.4 evidentno je da pretpostavka o jednolikoj i neovisnoj distribuciji poena nije u potpunosti dobra u predviđanju ukupnog broja poena u utakmicama niti u predviđanju hendikepa. Naime, površina na kojoj se histogrami realne i simulacijske distribucije ne preklapaju relativno je velika. Kada se uvede dinamika u vidu kratkoročnog i dugoročnog zamaha rezultati se popravljaju, a tome svjedoči veće poklapanje histograma (redak 2, 3 i 4 na slikama 9.1, 9.2, 9.3 i 9.4).



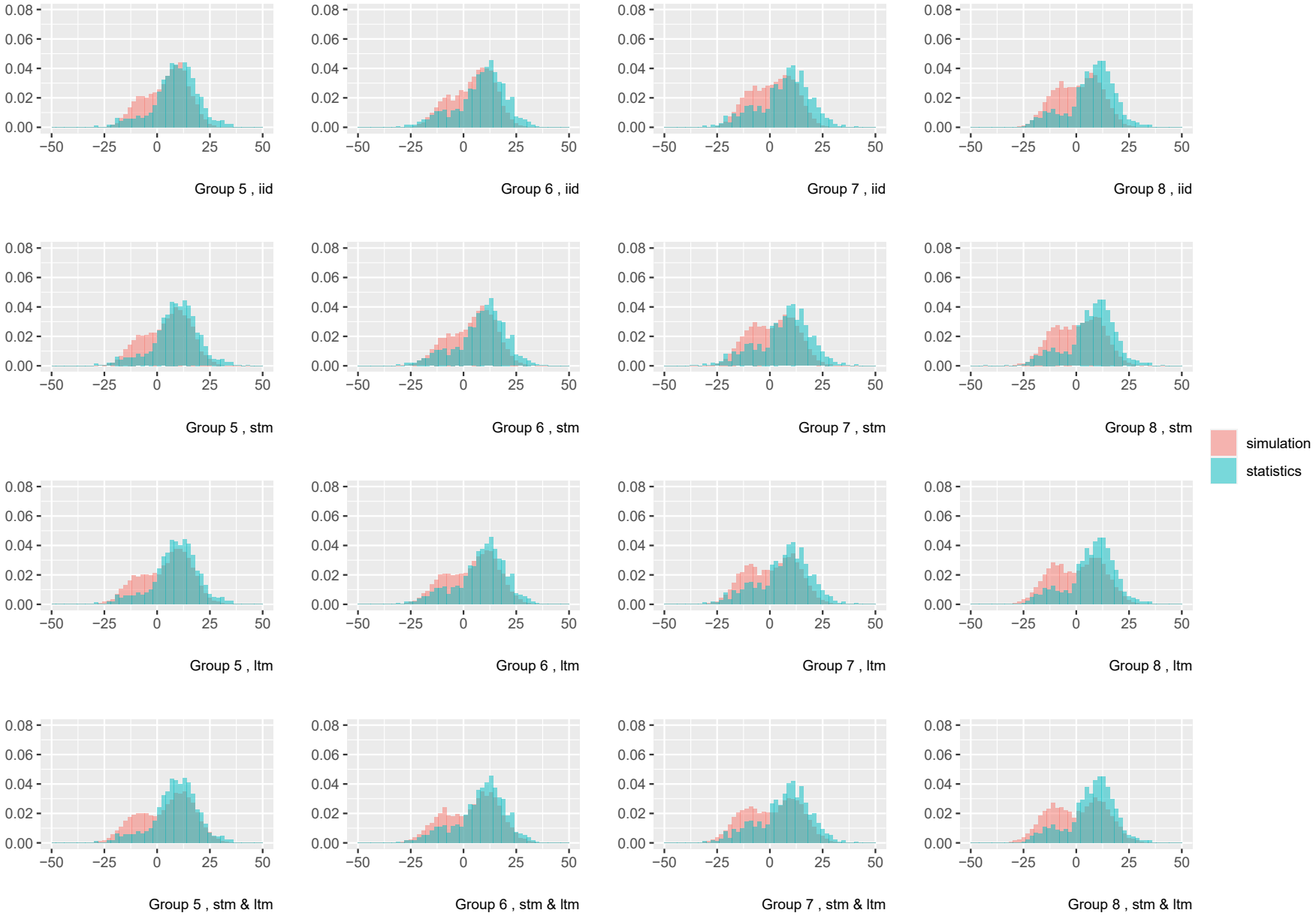
Slika 9.1: Realna i simulirana razdioba ukupnog broja poena u utakmici za grupe 1 do 4



Slika 9.2: Realna i simulirana razdioba ukupnog broja poena u utakmici za grupe 5 do 8



Slika 9.3: Realna i simulirana razdioba hendikepa za grupe 1 do 4



Slika 9.4: Realna i simulirana razdioba hendikepa za grupe 4 do 5

Grafički rezultati dodatno su potkrijepljeni numeričkim vrijednostima prikazanim u tablicama 9.1 i 9.2. Tablice 9.1 i 9.2 prikazuju ukupnu simulacijsku pogrešku (zbroj apsolutnih vrijednosti razlike gustoće vjerojatnosti realnih i simulacijskih podataka) u predviđanju ukupnog broja poena (tablica 9.1) i hendikepa (tablica 9.2) za svaku grupu utakmica. Prema oznakama uvedenima prethodno u disertaciji, stupac *iid* označava pogrešku Monte Carlo simulacije koja se temelji na jednolikoj i neovisnoj distribuciji poena, stupac *stm* označava pogrešku simulacije koja uključuje kratkoročni zamah, a stupac *ltm* označava pogrešku simulacije koja uključuje dugoročni zamah. Konačno stupac *stm i ltm* označava pogrešku napredne metode ažuriranja vjerojatnosti. Podebljano su označene greške simulacija s najboljim performansama.

Tablica 9.1: Simulacijska greška u predviđanju ukupnog broja poena u utakmici

Grupa	iid	stm	ltm	stm i ltm
1	0,1810	0,1942	**0,1669**	0,1849
2	0,2045	0,2146	**0,1771**	0,1912
3	0,1820	0,1986	**0,1456**	0,1589
4	0,2086	0,1964	0,1652	**0,1487**
5	0,1677	0,1316	0,1126	**0,0959**
6	0,1758	0,1525	0,1363	**0,1093**
7	0,1767	0,1601	0,1279	**0,1077**
8	0,1777	0,1591	0,1366	**0,1114**

Tablica 9.2: Simulacijska greška u predviđanju hendikepa u utakmici

Grupa	iid	stm	ltm	stm i ltm
1	0,1870	**0,1798**	0,1917	0,2024
2	0,1794	0,1889	**0,1617**	0,1648
3	**0,1770**	0,1944	0,2154	0,2091
4	0,2057	**0,1872**	0,2646	0,2658
5	0,1864	**0,1776**	0,1917	0,2024
6	0,1809	0,1749	**0,1617**	0,1648
7	0,2315	0,2259	0,2154	**0,2091**
8	0,2645	**0,2572**	0,2646	0,2658

Analizirajmo za početak greške simulacija koje inkorporiraju samo jedan od dinamičkih parametara. Generalno je vidljivo da uvođenje parametra dugoročnog zamaha konzistentno

popravlja rezultate predviđanja u slučaju predviđanja ukupnog broja poena u utakmicama određene grupe. Parametar kratkoročnog zamaha, u istom tom slučaju, popravlja rezultate predviđanja u svim grupama utakmica osim u prve tri grupe utakmica, gdje u usporedbi s modelom temeljenim na jednolikoj i neovisnoj distribuciji poena, daje nešto lošije rezultate predviđanja. Situacija je drugačija promatra li se greška predviđanja hendikepa. U tom slučaju kratkoročni zamah popravlja rezultate predviđanja u svim grupama utakmica osim u drugoj i trećoj grupi. Parametar dugoročnog zamaha, s druge strane, pokazuje bolje performanse samo u predviđanju hendikepa druge, šeste i sedme grupe utakmica usporedi li se s modelom temeljenim na jednolikoj i neovisnoj distribuciji bodova.

Kombinacija parametara kratkoročnog i dugoročnog zamaha popravlja rezultate predviđanja ukupnog broja poena za sve grupe utakmica, usporedi li se s modelom temeljenim na jednolikoj i neovisnoj distribuciji poena. Iznimka je slučaj predviđanja ukupnog broja poena u utakmicama prve grupe. U slučaju predviđanja hendikepa, kombinacija oba parametra poboljšava rezultate predviđanja samo u utakmicama sedme grupe.

Značajno poboljšanje rezultata napredne metode ažuriranja vjerojatnosti u slučaju predviđanja ukupnog broje poena posebno je vidljivo u posljednjih 5 grupa utakmica, odnosno što je razlika između snaga suprotstavljenih momčadi manja. Detaljnije, u posljednjih 5 grupa utakmica, napredna metoda ažuriranja vjerojatnosti smanjuje pogrešku simulacije kada se ona koristi za predviđanje ukupnog broja poena koji se očekuje u utakmici za 37% u prosjeku, usporedi li se s pristupom koji se temelji na jednolikoj i neovisnoj distribuciji bodova. Najveće poboljšanje u slučaju predviđanja ukupnog broja poena vidljivo je u sedmoj grupi utakmica. Relativna promjena greške predviđanja u toj grupi utakmica iznosi približno 43%. U prve tri grupa utakmica, simulacija koja inkorporira samo dugoročni zamah pokazuje najbolje rezultate u predviđanju ukupnog broja poena u utakmici. Prikazani rezultati sugeriraju zaključak da kratkoročni zamah gubi svoj utjecaj kada postoji značajna razlika u omjeru snaga momčadi, tako da bolja momčad može lakše početi dominirati utakmicom. Iz prikazanih rezultata, može se zaključiti da korištenje parametara sportske dinamike daje bolje rezultate u usporedbi s modelom koji se temelji na jednolikoj i neovisnoj distribuciji poena. U slučaju predviđanja ukupnog broja poena u prve tri grupe utakmica najbolje rezultate daje simulacija koja inkorporira samo parametar dugoročnog zamaha, dok za ostale grupe utakmica najbolje performanse pokazuje napredna metoda ažuriranja vjerojatnosti. U slučaju hendikepa, napredna metoda ažuriranja vjerojatnosti pokazuje najbolje performanse samo u predviđanju distribucije hendikepa u sedmoj grupi utakmica. Simulacija koja inkorporira samo kratkročni zamah pokazuje najbolje performanse u predviđanju distribucije hendikepa u prvoj, četvrtoj, petoj i osmoj grupi utakmica, dok simulacija koja inkorporira samo dugoročni zamah pokazuje najbolje performanse u drugoj i šestoj grupi utakmica.

Poglavlje 10

Zaključak

U doktorskoj disertaciji dovodi se u pitanje uobičajen pristup modeliranja sportova s bodovnim ograničenjem koji se temelji na pretpostavci o jednolikoj i neovisnoj distribuciji poena. Uvedena su dva nova parametra sportske dinamike, parametar kratkoročnog i parametar dugoročnog zamaha. Za oba parametra, predložene su matematičke formulacije temeljene na uvjetnoj vjerojatnosti i empirijskom Bayesovom pravilu ažuriranja. Potom je implementiran prediktivni model temeljen na Monte Carlo simulaciji koji uključuje matematičke formulacije uvedenih parametara. Konačno je analiziran utjecaj definiranih parametara na predviđanje određenih karakteristika odbojkaških utakmica. Radi se o ukupnom broju poena koji se očekuje u određenoj utakmici i o hendikepu.

Metoda je testirana na stvarnom skupu podataka i za slučaj predviđanja ukupnog broja poena u utakmici, rezultati su potvrdili početnu pretpostavku - uvođenjem dinamike u simulacije odbojkaških utakmica moguće je točnije simulirati stvarni tijek utakmica i posljedično dobiti bolje rezultate. Rezultati su lošiji previđa li se hendikep. To je i očekivano. Parametri sportske dinamike mogu kratkotrajno utjecati na performanse suprotstavljenih momčadi, međutim, kvaliteta momčadi konačno će imati puno značajniji utjecaj na predviđanja. Navedeno znači da parametri sportske dinamike, na primjer, mogu "pogurati" autsajdera u kraćem vremenskom periodu utakmice, a to može utjecati na veći broj odigranih poena u utakmici. Međutim, kvaliteta momčadi favorita prevladat će dugoročno i odlučit će o pobjedniku utakmice i o razlici u broju odigranih poena (hendikep). Dugoročni zamah pokazao se važnijim u predviđanju ukupnog broja poena u utakmicama, a u predviđanju hendikepa izbor najboljeg dinamičkog parametra ovisi o grupi utakmica čiji se ishod predviđa. Napredna metoda ažuriranja vjerojatnosti, u usporedbi s modelom temeljenim na jednolikoj i neovisnoj distribuciji poena, popravlja rezultate predviđanja ukupnog broja poena za sve grupe utakmica. Iznimka je slučaj predviđanja ukupnog broja poena u utakmicama prve grupe kada je razlika među momčadima najveća. U slučaju predviđanja hendikepa, napredna metoda ažuriranja vjerojatnosti poboljšava rezultate predviđanja samo u utakmicama sedme grupe. Važno je napomenuti da u prve tri grupe utak-

mica, korištenje samo parametra dugoročnog zamaha pokazuje bolje rezultate pri predviđanju ukupnog broja poena u usporedbi s predloženom naprednom metodom ažuriranja vjerojatnosti. Ovakvi rezultati su očekivani i idu u prilog razmišljanju da kratkoročni zamah gubi svoj utjecaj kada postoji značajna razlika u omjerima snaga dviju momčadi.

Potrebno je još jednom naglasiti da je metoda predložena u ovom radu interpretabilna te se metoda vrlo lako može ugraditi u ekspertne sustave kako bi se dobio uvid u moguće rezultatske sekvence utakmice koje se mogu pojaviti pod različitim okolnostima. Stručnjaci u domeni sporta mogu koristiti takve simulacije za prepoznavanje, razumijevanje i optimiziranje performansi igrača i momčadi.

Sekundarni dio ovog rada, ali ipak neophodan za demonstraciju rezultata, bio je dio profiliranja momčadi. Zbog ograničenja skupa podataka korištenog u ovom istraživanju, doktorska disertacija predlaže metodu profiliranja koja je prilagođena za rad s heterogenim skupovima podataka. Pristup se pokazao učinkovitim te se može koristiti i u drugim sličnim slučajevima kada nije dostupna dovoljna količina podataka za profiliranje individualnih momčadi ili igrača.

Model predložen u disertaciji može se generalizirati na sportove koji koriste sličan sustav bodovanja kao i odbojka. Napredna metoda ažuriranja vjerojatnosti je vrlo fleksibilna i lako se može nadograditi definiranjem drugih parametara sportske dinamike poput psihičkog pritiska i umora, a to je ostavljeno za buduća istraživanja. Dodatne nadogradnje prezentiranog rada planirane u bliskoj budućnosti uključuju analizu promjene vrijednosti parametra kratkoročnog zamaha u ovisnosti o fazi utakmice te implementaciju potencijalne promjene u model. Buduće istraživanje usmjerit će se na individualno profiliranje momčadi što će potencijalno biti omogućeno prikupljanjem većeg podatkovnog skupa ili novim tehnologijama temeljenim na umjetnoj inteligenciji.

Literatura

[1]Percy, D. F., "Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes", Journal of the Operational Research Society, Vol. 66, No. 11, 2015, str. 1840–1849.

[2]Šar čević, A., "Prediktivna analiza i modeliranje teniskih mečeva", Master's thesis, University of Zagreb. Faculty of Electrical Engineering and Computing, 2017.

[3]Schutz, R. W., "A mathematical model for evaluating scoring systems with specific reference to tennis", Research Quarterly. American Association for Health, Physical Education and Recreation, Vol. 41, No. 4, 1970, str. 552–561.

[4]Liu, Y., "Random walks in tennis", Missouri Journal of Mathematical Sciences, Vol. 13, No. 3, 2001, str. 154–162.

[5]Barnett, T. J., Clarke, S. R. *et al.*, "Using microsoft excel to model a tennis match", in 6th Conference on Mathematics and Computers in Sport. Queensland, Australia: Bond University, 2002, str. 63–68.

[6]Barnett, T., Brown, A., Clarke, S., "Developing a model that reflects outcomes of tennis matches", in proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport, Coolangatta, Queensland, 2006, str. 3–5.

[7]Huang, X., Knottenbelt, W., Bradley, J., "Inferring tennis match progress from in-play betting odds", Final year project, Imperial College London, South Kensington Campus, London, SW7 2AZ, 2011.

[8]Barnett, T., Brown, A., The Mathematics of Tennis. Strategic Games, 2012.

[9]Šar čević, A., Vranić, M., Pintar, D., "A combinatorial approach in predicting the outcome of tennis matches", International Journal of Applied Mathematics and Computer Science, Vol. 31, No. 3, 2021.

[10]Barnett, T. J., Brown, A., Jackson, K. *et al.*, "Modelling outcomes in volleyball", in Proceedings of the 9th Australasian Conference on Mathematics and Computers in Sport (9M&CS), 2008, str. 130–137.

[11]Ferrante, M., Fonseca, G., "On the winning probabilities and mean durations of volleyball", Journal of Quantitative Analysis in Sports, Vol. 10, No. 2, 2014, str. 91–98.

[12]Gojsali ć, A., "Primjena metoda dubinske analize u domeni igara na sreću", Master's thesis, University of Zagreb. Faculty of Electrical Engineering and Computing, 2016.

[13]Hsi, B., Burych, D., "Games of two players", Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol. 20, No. 1, 1971, str. 86–92.

[14]Pollard, G., "An analysis of classical and tie-breaker tennis", Australian Journal of Statistics, Vol. 25, No. 3, 1983, str. 496–505.

[15]Newton, P. K., Keller, J. B., "Probability of winning at tennis i. theory and data", Studies in applied Mathematics, Vol. 114, No. 3, 2005, str. 241–269.

[16]O'Malley, A. J., "Probability formulas and statistical analysis in tennis", Journal of Quantitative Analysis in Sports, Vol. 4, No. 2, 2008.

[17]Croucher, J. S., "The conditional probability of winning games of tennis", Research Quarterly for Exercise and Sport, Vol. 57, No. 1, 1986, str. 23–26.

[18]Iso-Ahola, S. E., Mobily, K., ""Psychological momentum": A phenomenon and an empirical (unobtrusive) validation of its influence in a competitive sport tournament", Psychological Reports, Vol. 46, No. 2, 1980, str. 391–401.

[19]Bar-Eli, M., Avugos, S., Raab, M., "Twenty years of "hot hand" research: Review and critique", Psychology of Sport and Exercise, Vol. 7, No. 6, 2006, str. 525–553.

[20]Milinovi ć, M., Službena odbojkaška pravila 2017.-2020., 2018.

[21]Štrumbelj, E., "On determining probability forecasts from betting odds", International journal of forecasting, Vol. 30, No. 4, 2014, str. 934–943.

[22]Clauset, A., "A brief primer on probability distributions", 2011.

[23]Aldridge, M., MATH2750 Introduction to Markov Processes. University of Leeds, 2021.

[24]Pishro-Nik, H., "Introduction to probability, statistics, and random processes", 2016.

[25]Topi ć, T., "Bayesova statistika i procjena vrijednosti ulaganja", Master's thesis, University of Zagreb. Faculty of Science. Department of Mathematics, 2016.

[26]Dobrow, R. P., Introduction to stochastic processes with R. Hoboken, New Jersey: John Wiley & Sons, 2016.

[27]Kovalchik, S. A., "Searching for the goat of tennis win prediction", Journal of Quantitative Analysis in Sports, Vol. 12, No. 3, 2016, str. 127–138.

[28]Peters, J., "Predicting the outcomes of professional tennis matches", Master's thesis, School of Informatics University of Edinburgh, 2017.

[29]Šar čević, A., Vranić, M., Pintar, D., Krajna, A., "Predictive modeling of tennis matches: a review", in 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE, 2022, str. 1099–1104.

[30]Knottenbelt, W. J., Spanias, D., Madurska, A. M., "A common-opponent stochastic model for predicting the outcome of professional tennis matches", Computers & Mathematics with Applications, Vol. 64, No. 12, 2012, str. 3820–3827.

[31]Vra čar, P., Štrumbelj, E., Kononenko, I., "Modeling basketball play-by-play data", Expert Systems with Applications, Vol. 44, 2016, str. 58–66.

[32]Albert, J., "Baseball data at season, play-by-play, and pitch-by-pitch levels", Journal of Statistics Education, Vol. 18, No. 3, 2010.

[33]Aalbers, B., Van Haaren, J., "Distinguishing between roles of football players in play-by-play match event data", in International Workshop on Machine Learning and Data Mining for Sports Analytics. Springer, 2019, str. 31–41.

[34]Grassetti, L., Bellio, R., Fonseca, G., Vidoni, P., "Estimation of lineup efficiency effects in basketball using play-by-play data", Book of Short Papers SIS2019, 2019, str. 363–370.

[35]Albert, J., "Using play-by-play baseball data to develop a better measure of batting performance", Bowling Green State University., 2001.

[36]"Dubinska analiza podataka", https://www.fer.unizg.hr/_download/repository/DAP_1_predavanje.pdf, accessed: 2022-12-15. 2021/2022.

[37]Jovi ć, A., "Postupci dubinske analize podataka", Qualifying examination, University of Zagreb. Faculty of Electrical Engineering and Computing.

[38]Han, J., Kamber, M., Pei, J., Data mining: concepts and techniques. Morgan kaufmann, 2012.

[39]The Business Research Company, “Sports global market report 2022 – by type (spectator sports, participatory sports), by revenue source (media rights, merchandising, tickets, sponsorship), by ownership (chained, standalone) – market size, trends, and global forecast 2022-2026”, https://www.thebusinessresearchcompany.com/report/sports-global-market-report#:~:text=The%20global%20sports%20market%20grew,least%20in%20the%20short%20term, accessed: 2022-12-15. 2022.

[40]Goud, P. S. H. V., Roopa, Y. M., Padmaja, B., “Player performance analysis in sports: With fusion of machine learning and wearable technology”, in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019, str. 600–603.

[41]Baclig, M. M., Ergezinger, N., Mei, Q., Gül, M., Adeeb, S., Westover, L., “A deep learning and computer vision based multi-player tracker for squash”, Applied Sciences, Vol. 10, No. 24, 2020, str. 8793.

[42]Bendtsen, M., “Regimes in baseball players’ career data”, Data mining and knowledge discovery, Vol. 31, No. 6, 2017, str. 1580–1621.

[43]Kautz, T., Groh, B. H., Hannink, J., Jensen, U., Strubberg, H., Eskofier, B. M., “Activity recognition in beach volleyball using a deep convolutional neural network”, Data Mining and Knowledge Discovery, Vol. 31, No. 6, 2017, str. 1678–1705.

[44]Andrienko, G., Andrienko, N., Budziak, G., Dykes, J., Fuchs, G., von Landesberger, T., Weber, H., “Visual analysis of pressure in football”, Data Mining and Knowledge Discovery, Vol. 31, No. 6, 2017, str. 1793–1839.

[45]The Business Research Company, “Online gambling global market report 2022 – by game type (betting, casino, lottery, poker, online bingo), by device (desktop, mobile), by component (software, services) – market size, trends, and global forecast 2022-2026”, https://www.thebusinessresearchcompany.com/report/online-gambling-global-market-report#:~:text=The%20global%20online%20gambling%20market,least%20in%20the%20short%20term., accessed: 2022-12-15. 2022.

[46]European Gaming and Betting Association, “European online gambling key figures 2021 edition”, https://www.egba.eu/uploads/2021/12/European-Online-Gambling-Key-Figures-2021-Edition.pdf, accessed: 2022-12-15. 2021.

[47]Wallace, H. M., Baumeister, R. F., Vohs, K. D., “Audience support and choking under pressure: A home disadvantage?”, Journal of sports sciences, Vol. 23, No. 4, 2005, str. 429–438.

[48]Gilovich, T., Vallone, R., Tversky, A., “The hot hand in basketball: On the misperception of random sequences”, Cognitive psychology, Vol. 17, No. 3, 1985, str. 295–314.

[49]Tversky, A., Gilovich, T., “The cold facts about the “hot hand” in basketball”, Chance, Vol. 2, No. 1, 1989, str. 16–21.

[50]McCallum, J., “Hot hand, hot head”, Sports Illustrated, Vol. 78, 1993, str. 22–24.

[51]Attali, Y., “Perceived hotness affects behavior of basketball players and coaches”, Psychological science, Vol. 24, No. 7, 2013, str. 1151–1156.

[52]Albright, S. C., “A statistical analysis of hitting streaks in baseball”, Journal of the american statistical association, Vol. 88, No. 424, 1993, str. 1175–1183.

[53]Jackson, D. A., “Independent trials are a model for disaster”, Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol. 42, No. 1, 1993, str. 211–220.

[54]Richardson, P. A., Adler, W., Hankes, D., “Game, set, match: Psychological momentum in tennis”, The Sport Psychologist, Vol. 2, No. 1, 1988, str. 69–76.

[55]Silva, J. M., Hardy, C. J., Crace, R. K., “Analysis of psychological momentum in intercollegiate tennis.”, Journal of Sport & Exercise Psychology, 1988.

[56]Weinberg, R., Jackson, A., “The effects of psychological momentum on male and female tennis players revisited”, Journal of Sport Behavior, Vol. 12, No. 3, 1989, str. 167.

[57]Jackson, D., Mosurski, K., “Heavy defeats in tennis: Psychological momentum or random effect?”, Chance, Vol. 10, No. 2, 1997, str. 27–34.

[58]Ransom, K., Weinberg, R., “Effect of situation criticality on performance of elite male and female tennis players”, Journal of Sport Behavior, Vol. 8, No. 3, 1985, str. 144.

[59]Klaassen, F. J., Magnus, J. R., “Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model”, Journal of the American Statistical Association, Vol. 96, No. 454, 2001, str. 500–509.

[60]Newton, P. K., Aslam, K., “Monte carlo tennis”, SIAM review, Vol. 48, No. 4, 2006, str. 722–742.

[61]Schilling, M. F., “Does momentum exist in competitive volleyball?”, Chance, Vol. 22, No. 4, 2009, str. 29–35.

[62]Stanimirovic, R., Hanrahan, S. J., “Efficacy, affect, and teams: Is momentum a misnomer?”, International Journal of Sport and Exercise Psychology, Vol. 2, No. 1, 2004, str. 43–62.

[63]Raab, M., Gula, B., Gigerenzer, G., “The hot hand exists in volleyball and is used for allocation decisions.”, Journal of Experimental Psychology: Applied, Vol. 18, No. 1, 2012, str. 81.

[64]Clarke, S., Kovalchik, S., Ingram, M., “Adjusting bookmaker’s odds to allow for overround”, American Journal of Sports Science, Vol. 5, No. 6, 2017, str. 45–49.

[65]Koning, R. H., Zijm, R., “Betting market efficiency and prediction in binary choice models”, Annals of Operations Research, 2022, str. 1–14.

[66]Shin, H. S., “Measuring the incidence of insider trading in a market for state-contingent claims”, The Economic Journal, Vol. 103, No. 420, 1993, str. 1141–1153.

[67]Jullien, B., Salanié, B., “Measuring the incidence of insider trading: A comment on shin”, The Economic Journal, Vol. 104, No. 427, 1994, str. 1418–1419.

[68]“Kontinuirane slu čajne varijable”, https://www.pmf.unizg.hr/_download/repository/PREDAVANJE9.pdf, accessed: 2022-12-15.

[69]Walpole, R., Myers, R., Myers, S., Ye, K., “Probability & statistics for engineers & scientists, mylab statistics update”, 2017.

[70]“Stohasti čki procesi”, https://www.fer.unizg.hr/_download/repository/04_-_Stohasticki_procesi[2].pdf, accessed: 2022-12-15. 2021.

[71]Coleman, R., “What is a stochastic process?”, in Stochastic Processes. Springer, 1974, str. 1–5.

[72]Revuz, D., Markov chains. North-Holland, Amsterdam: Elsevier, 1984.

[73]Doob, J. L., Stochastic Processes. New York: John Wiley & Sons, 1953.

[74]Chung, K. L., Markov Chains: With Stationary Transition Probabilities. Berlin: Springer, 1967.

[75]Casella, G., “An introduction to empirical bayes data analysis”, The American Statistician, Vol. 39, No. 2, 1985, str. 83–87.

[76]Fink, D., “A compendium of conjugate priors”, Environmental Statistics Group, Department of Biology, Montana State Univeristy, Vol. 46, 1997.

[77]Carlin, B. P., Louis, T. A., Bayes and Empirical Bayes Methods for Data Analysis, Second Edition. Chapman and Hall/CRC, 2000.

[78]Harrison, R. L., “Introduction to monte carlo simulation”, in AIP conference proceedings, Vol. 1204, No. 1. American Institute of Physics, 2010, str. 17–21.

[79]Men čík, J., “Monte carlo simulation method”, in Concise Reliability for Engineers. IntechOpen, 2016.

[80]Sopa, I., Szabo, D., “Comparison between statistical parameters of attack and defence in high volleyball performance (csm volei alba blaj in the cev champions league final four 2018)”, Bulletin of the Transilvania University of Braşov. Series IX: Sciences of Human Kinetics, 2020, str. 92–102.

[81]Conejero, M., Claver, F., González-Silva, J., Fernández-Echeverría, C., Moreno, P., “Analysis of performance in game actions in volleyball, according to the classification.”, Revista Portuguesa de Ciências do Desporto, 2017.

[82]Gabrio, A., “Bayesian hierarchical models for the prediction of volleyball results”, Journal of Applied Statistics, Vol. 48, No. 2, 2021, str. 301–321.

[83]Wozniak, J., “Inferring tennis match progress from in-play betting odds: Project report”, Imperial College London, South Kensington Campus: London, UK, 2011.

[84]Lee, K. T., Chin, S. T., “Strategies to serve or receive the service in volleyball”, Mathematical Methods of Operations Research, Vol. 59, No. 1, 2004, str. 53–67.

[85]Bradley, R. A., Terry, M. E., “Rank analysis of incomplete block designs: I. the method of paired comparisons”, Biometrika, Vol. 39, No. 3/4, 1952, str. 324–345.

[86]Karwoski, D., “Analysis of gnac volleyball using the bradley-terry model”, QMaster’s Project, University of Alaska, 2020.

[87]Aoki, R. Y., Assuncao, R. M., Vaz de Melo, P. O., “Luck is hard to beat: The difficulty of sports prediction”, in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, str. 1367–1376.

[88]Hunter, D. R., “Mm algorithms for generalized bradley-terry models”, The annals of statistics, Vol. 32, No. 1, 2004, str. 384–406.

[89]McHale, I., Morton, A., “A bradley-terry type model for forecasting tennis match results”, International Journal of Forecasting, Vol. 27, No. 2, 2011, str. 619–630.

[90]Baker, R. D., McHale, I. G., “A dynamic paired comparisons model: Who is the greatest tennis player?”, European Journal of Operational Research, Vol. 236, No. 2, 2014, str. 677–684.

[91]Baker, R. D., McHale, I. G., “An empirical bayes model for time-varying paired comparisons ratings: Who is the greatest women’s tennis player?”, European Journal of Operational Research, Vol. 258, No. 1, 2017, str. 328–333.

[92]Gorgi, P., Koopman, S. J., Lit, R., “The analysis and forecasting of tennis matches by using a high dimensional dynamic model”, Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 182, No. 4, 2019, str. 1393–1409.

[93]“Bradley–terry model”, https://en.wikipedia.org/wiki/Bradley%E2%80%93Terry_model, accessed: 2022-12-15.

[94]Elo, A. E., The rating of chessplayers, past and present. New York: Arco Pub., 1978.

[95]Cesarec, B., “Teorija igara u poslovanju”, Bachelor’s Thesis, University North. University centre Varaždin. Department of Technical and Economic Logistic, 2020.

[96]Angelini, G., Candila, V., De Angelis, L., “Weighted elo rating for tennis match predictions”, European Journal of Operational Research, Vol. 297, No. 1, 2022, str. 120–132.

[97]Glickman, M. E., Hennessy, J., Bent, A., “A comparison of rating systems for competitive women’s beach volleyball”, Statistica Applicata-Italian Journal of Applied Statistics, No. 2, 2018, str. 233–254.

[98]Madurska, A. M., “A set-by-set analysis method for predicting the outcome of professional singles tennis matches”, 4th year Software Engineering MEng project, Imperial College London, Department of Computing: London, UK, 2012.

[99]Barnett, T., Clarke, S. R., “Combining player statistics to predict outcomes of tennis matches”, IMA Journal of Management Mathematics, Vol. 16, No. 2, 2005, str. 113–120.

[100]Newton, P. K., Aslam, K., “Monte carlo tennis: a stochastic markov chain model”, Journal of Quantitative Analysis in Sports, Vol. 5, No. 3, 2009.

[101]Ingram, M., “A point-based bayesian hierarchical model to predict the outcome of tennis matches”, Journal of Quantitative Analysis in Sports, Vol. 15, No. 4, 2019, str. 313–325.

[102]Spanias, D., Knottenbelt, W. J., “Predicting the outcomes of tennis matches using a low-level point model”, IMA Journal of Management Mathematics, Vol. 24, No. 3, 2013, str. 311–320.

[103]Kovalchik, S., Reid, M., “A calibration method with dynamic updates for within-match forecasting of wins in tennis”, International Journal of Forecasting, Vol. 35, No. 2, 2019, str. 756–766.

[104]Gollub, J., “Forecasting serve performance in professional tennis matches”, Journal of Sports Analytics, No. Preprint, 2021, str. 1–11.

[105]Carrari, A., Ferrante, M., Fonseca, G., “A new markovian model for tennis matches”, Electronic Journal of Applied Statistical Analysis, Vol. 10, No. 3, 2017, str. 693–711.

[106]Iso-Ahola, S. E., Dotson, C. O., “Psychological momentum: Why success breeds success”, Review of general psychology, Vol. 18, No. 1, 2014, str. 19–33.

[107]Gula, B., Köppen, J., “Einfluss von länge und perfektion einer „hot-hand”-sequenz auf zuspielentscheidungen im volleyball”, Zeitschrift für Sportpsychologie, Vol. 16, No. 2, 2009, str. 65–70.

[108]Avugos, S., Köppen, J., Czienskowski, U., Raab, M., Bar-Eli, M., “The “hot hand” reconsidered: A meta-analytic approach”, Psychology of Sport and Exercise, Vol. 14, No. 1, 2013, str. 21–27.

[109]Pelechrinis, K., Winston, W., “The ‘hot hand’ is a real basketball phenomenon – but only some players have the ability to go on these basket-making streaks”, https://theconversation.com/the-hot-hand-is-a-real-basketball-phenomenon-but-only-some-players-have-the-ability-to-go-on-these-basket-making-streaks-179082#:~:text=To%20say%20a%20player%20is,and%20fit%20within%20statistical%20norms, accessed: 2022-12-15.

[110]Hendricks, D., Patel, J., Zeckhauser, R., “Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988”, The Journal of finance, Vol. 48, No. 1, 1993, str. 93–130.

[111]Gilden, D. L., Wilson, S. G., “On the nature of streaks in signal detection”, Cognitive Psychology, Vol. 28, No. 1, 1995, str. 17–64.

[112]Serfozo, R., Basics of applied stochastic processes. Springer Science & Business Media, 2009.

[113]Rorrison, B. D., “The growing importance of data analytics in the sports industry”, https://www.linkedin.com/pulse/growing-importance-data-analytics-sports-industry-rorrison, accessed: 2022-12-15.

[114]Christina Gough, “Market size of the sports analytics industry worldwide in 2021, with a forecast for 2030(in billion u.s. dollars)”, https://www.statista.com/statistics/1185536/sports-analytics-market-size/#:~:text=Global%20sports%20analytics%20market%20revenue%202021%2D2030&text=The%20global%20sports%20analytics%20market,billion%20U.S.%20dollars%20by%202030., accessed: 2022-12-15. 2022.

[115]MarathonBet, https://www.marathonbet.com/en/?cppcids=all, accessed: 2021-1-3.

[116]Šar čević, A., Pintar, D., Vranić, M., Gojsalić, A., "Modeling in-match sports dynamics using the evolving probability method", Applied Sciences, Vol. 11, No. 10, 2021, str. 4429.

[117]Šar čević, A., Gojsalić, A., Pintar, D., Vranić, M., "Volleyball data", https://github.com/ana2202/Volleyball-Data, accessed: 2022-12-15.

[118]Magnus, J. R., Klaassen, F. J., "On the advantage of serving first in a tennis set: four years at wimbledon", Journal of the Royal Statistical Society: Series D (The Statistician), Vol. 48, No. 2, 1999, str. 247–256.

[119]Aslam, K., "A stochastic markov chain approach for tennis: Monte carlo simulation and modeling", Doktorski rad, University of Southern California, 2012.

[120]Sipko, M., Knottenbelt, W., "Machine learning for the prediction of professional tennis matches", MEng computing-final year project, Imperial College London, Vol. 2, 2015.

[121]Gao, Z., Kowalczyk, A., "Random forest model identifies serve strength as a key predictor of tennis match outcome", Journal of Sports Analytics, Vol. 7, No. 4, 2021, str. 255–262.

[122]Candila, V., Palazzo, L., "Neural networks and betting strategies for tennis", Risks, Vol. 8, No. 3, 2020, str. 68.

Popis slika

Popis tablica

Životopis

Ana Šarčević rođena je 22. veljače 1993. godine u Tübingenu, Njemačkoj. U Zadru je završila Osnovnu školu Smiljevac te opću gimnaziju Vladimira Nazora. Nakon gimnazije upisala je preddiplomski studij računarstva na Sveučilištu u Zagrebu Fakultet elektrotehnike i računarstva (FER), a 2015. godine na istom fakultetu stekla je prvostupničku diplomu. Te godine upisala je diplomski studij Računarstvo, profil *Računalno inženjerstvo*, također na FER-u. U srpnju 2017. godine diplomirala je s pohvalom (lat. *cum laude*) kakvu dobije tek 10% najboljih studenata u okviru studijskog programa. Kasnije iste godine upisala je poslijediplomski doktorski studij na FER-u, gdje je počela raditi kao istraživač suradnik za Zavodu za osnove elektrotehnike i električna mjerenja u okviru projekta *Razvoj integriranog sustava za zaštitu od kibernetičkih prijevara – IAFS*. Po završetku tog projekta zaposlila se kao mlađi istraživač u sklopu projekta *DATACROSS – Napredne metode i tehnologije o podatcima i kooperativnim sustavima*, također na Zavodu za osnove elektrotehnike i električna mjerenja, FER. Istekom ugovora 2020. godine zaposlila se u firmi NEOS d.o.o. Konačno je sklopila ugovor za mjesto istraživača u sklopu projekta *Sustav umjetne inteligencije za autonomni nadzor i upravljanje sigurnosti cloud okruženja – AI DEFENDER* ponovno na Zavod za osnove elektrotehnike i električna mjerenja, FER, a na istom mjestu zaposlena je i trenutno. U sklopu svih ovih projekata imala je priliku surađivati i s gospodarskim subjektima gdje je aktivno radila na transferu znanja i novih otkrića u industriji. U periodu od upisa poslijediplomskog doktorskog studija obavlja nastavne aktivnosti na preddiplomskom studiju FER-a. U suradnji s mentoricom, bila je izravan voditelj pri izradi nekoliko završnih i diplomskih radova. U istraživačkom smislu bavi se dubinskom analizom podataka, strojnim učenjem i objašnjivom umjetnom inteligencijom. Fokus istraživanja joj je na sportskoj domeni, a između ostalih domena često se bavi i telekomunikacijskom domenom i domenom kibernetičke sigurnosti gdje je ponajviše fokusirana na prediktivno modeliranje, detekciju anomalija i objašnjavanje modela crnih kutija. Objavila je 3 rada u znanstvenim časopisima kategorije A (CC i SCI Expanded) i 3 rada u zbornicima međunarodnih znanstvenih skupova. Članica je IEEE.

Popis objavljenih djela

Radovi u časopisima

1.Šar čević, A., Pintar, D., Vranić, M., Gojsalić, A., "Modeling in-match sports dynamics using the evolving probability method", Applied Sciences, Vol. 11, No. 10, 4429, svibanj 2021.

2.Šar čević, A., Vranić, M., Pintar, D., "A Combinatorial Approach in Predicting the Outcome of Tennis Matches", International Journal of Applied Mathematics and Computer Science, Vol. 31, No. 3, pp. 525-538, rujan 2021.

3.Šar čević, A., Vranić, M., Pintar, D., Krajna A., "Cybersecurity Knowledge Extraction Using XAI", Applied Sciences, Vol. 12, No. 17, 8669, kolovoz 2022.

Radovi na međunarodnim konferencijama

1.Zovak, T., Šar čević, A., Vranić, M,, Pintar, D., "Game-to-Game Prediction of NBA Players' Points in Relation to Their Season Average", 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Hrvatska, 2019, pp. 1266-1270.

2.Šar čević, A., Vranić, M., Pintar, D., Krajna A., "Predictive modeling of tennis matches: a review", 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Hrvatska, 2022, pp. 1099-1104.

3.Krajna, A., Kova č, M., Brčić, M., Šarčević A., "Explainable Artificial Intelligence: An Updated Perspective", 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Hrvatska, 2022, pp. 859-864.

Biography

Ana Šarčević was born on February 22, 1993 in Tübingen, Germany. She attended Smiljevac Elementary School and the Vladimir Nazor Gymnasium in Zadar. Following gymnasium, she enrolled in the *Computing* undergraduate program at the University of Zagreb Faculty of Electrical Engineering and Computing (FER), where she earned a bachelor's degree in 2015. That same year, she enrolled in the *Computer Engineering* profile at FER Graduate Studies. In July 2017, she graduated with honors (lat. *cum laude*), which is received only by the 10% of the best students in the study program. Later that year, she enrolled in postgraduate doctoral studies at FER, where she began working as a research associate for the *Department of Electrical Engineering Fundamentals and Measurements*, FER, on the project *Development of an integrated system for protection against cyber fraud - IAFS*. Following the completion of that project, she was hired as a junior researcher for the *DATACROSS project - Advanced Methods and Technologies in Data Science and Cooperative Systems*, again, at the Department of Electrical Engineering Fundamentals and Measurements, FER. She was hired by NEOS d.o.o. after the contract's expiration in 2020. Finally, she signed a contract for the position of researcher within the project *Artificial intelligence system for autonomous monitoring and management of cloud environment security - AI DEFENDER*, again at the Department of Electrical Engineering Fundamentals and Measurements, FER, where she still works. As part of all of these projects, she had the opportunity to collaborate with economic entities, where she actively worked on the transfer of knowledge and new discoveries in the industry. In the period from the enrollment of the postgraduate doctoral study, she performed teaching activities at the undergraduate study of FER. She worked closely with her mentor to oversee the direct development of a number of bachelor and master's theses. She focuses her research on data mining, machine learning and explainable artificial intelligence. Her research is centered on the sports domain, and among other domains she often deals with the telecommunications domain and the domain of cyber security, where she is mostly focused on predictive modeling, anomaly detection and explaining black box models. She published 3 papers in scientific journals of category A (CC and SCI Expanded) and 3 papers in proceedings of international scientific meetings. She is a member of IEEE.