

Imperial College London
Department of Computing

Professional Tennis: Quantitative Models and Ranking Algorithms

Demetris Spanias

30 September 2014

Supervised by William Knottenbelt

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College London
and the Diploma of Imperial College London

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Declaration of Originality

I declare that the content of this thesis was composed by myself and unless otherwise stated, the work presented is my own.

Abstract

Professional singles tennis is a popular global sport that attracts spectators and speculators alike. In recent years, financial trading related to sport outcomes has become a reality, thanks to the rise of online betting exchanges and the ever-increasing development and deployment of quantitative models for sports.

This thesis investigates the extent to which the outcome of a match between two professional tennis players can be forecast using quantitative models parameterised by historical data. Three different approaches are explored, each having its own advantages and disadvantages. Firstly, the problem is approached using a Markov chain to model a tennis point, estimating the probability of a player winning a point while serving. Such a probability can be used as a parameter to existing hierarchical models to estimate the probability of a player winning the match. We demonstrate how this probability can be estimated using varying subsets of historical player data and investigate their effect on results.

Averaged historical data over varying opponents with different skill sets, does not necessarily provide a fair basis of comparison when evaluating the performance of players. The second approach presented is a technique that uses data, which includes only matches played against common opponents, to find the difference between the modelled players' probability of winning a point on their serve against each common opponent. This difference in probability for each common opponent is a "transitive contribution" towards victory for the match being modelled. By combining these "contributions" the "Common-Opponent" model overcomes the problems of using average historical statistics at the cost of a shrinking data set.

Finally, the thesis ventures into the field of player rankings. Rankings provide a fast and simple method for predicting match winners and comparing players. We present a variety of methods to generate such player rankings, either by making use of network analysis or hierarchical models. The generated rankings are then evaluated using their ability to correctly represent the subset of matches that were used to generate them as well as their ability to forecast future matches.

I would like to dedicate this thesis first and foremost to my parents, who supported me both financially and mentally throughout my PhD. Secondly, my supervisor, William, who has been my guide and mentor to the world of academia and research. Last but not least, my partner, Sofia, who motivated me through the tough days and kept me going. Had it not been for their understanding and support I would not have reached as far as I did.

Contents

1. Introduction	17
1.1. Motivation	17
1.2. Objectives	18
1.3. Contributions	20
1.3.1. Point Model	20
1.3.2. Common-Opponent Model	20
1.3.3. Ranking Systems	21
1.4. Thesis Structure	22
1.5. Publications	24
2. Background	25
2.1. The Game of Tennis	25
2.1.1. Rules	25
2.1.2. Scoring System and Order of Serve	28
2.1.3. Tournaments	31
2.1.4. The Official Ranking Systems	33
2.2. Theoretical Methods	36
2.2.1. Probability Theory	36
2.2.2. Common Distributions	37
2.2.3. Significance Testing	42
2.2.4. Stochastic Processes	44
2.3. Literature Overview	46
2.3.1. Hierarchical Match Models	46
2.3.2. Independence of Points	51
2.3.3. Ranking Models	51
2.3.4. Using Rankings as Predictive Tools	52
2.3.5. Other Tennis Model Uses	53

3. Expanding the Hierarchical Tennis Model	60
3.1. Match Markov Chain	60
3.2. Set Markov Chain	62
3.3. Game Markov Chain	64
3.4. Tiebreaker Markov Chain	65
3.5. Point Markov Chain	67
3.6. Service and Rally Markov Chains	68
3.7. Forecasting the Outcome of a Match	71
3.7.1. Collecting the Data	71
3.7.2. A Closer Look at the Data	73
3.7.3. Estimating the Probability of Winning Service Points	78
3.7.4. Combining Historical Data for Doubles	82
3.8. Selecting Historical Data	83
3.8.1. Age of Match Played	84
3.8.2. Surface of Match Played	84
3.9. Evaluating the Performance of Tennis Models	85
3.9.1. A Tennis Model Performance Rating, ρ	85
3.9.2. The Random Model	86
3.9.3. Back-testing Using Real Data	86
3.9.4. Comparing Models Against the Random Model	91
3.9.5. Uncombined Model vs. Combined Model	92
3.9.6. Barnett Model vs. Combined Model	93
3.9.7. Combined Model vs. Bookmaker Models	93
3.10. Conclusions	96
4. A Common-Opponent Based Model	98
4.1. Introduction	98
4.2. The Concept of Transitivity	99
4.3. Relationship between the probabilistic difference of winning service points and winning the match	99
4.4. Match Probabilities Using Common-Opponent Model	100
4.5. Evaluating Model Performance	105
4.5.1. Back-testing Results	105
4.5.2. Common-Opponent Model vs. Random Model	107
4.5.3. Common-Opponent Model vs. Uncombined Model	108
4.5.4. Common-Opponent Model vs. Combined Model	108

4.5.5. Common-Opponent Model vs. Bookmaker Models	109
4.6. Conclusion	110
5. Ranking Systems for Tennis Players	112
5.1. Introduction	112
5.2. The PageRank Approach	112
5.3. Set, Game and Point PageRank Approach	115
5.4. Comparing PageRank Approaches	116
5.5. SortRank	125
5.6. The LadderRank Algorithm	125
5.7. SortRank and LadderRank Performance	127
5.8. Forecasting Match Outcomes Using Ranking Systems	133
5.9. Match Probabilities from Rankings	136
5.10. PageRank Set rankings vs. Bookmakers	137
5.11. Conclusion	138
6. Conclusions and Further Research	139
6.1. Achievements	139
6.1.1. Tennis Point Markov Model	139
6.1.2. Common-Opponent Model	139
6.1.3. Ranking Systems and Forecasting	140
6.2. Applications	141
6.3. Further Research	142
6.3.1. Analysis of Data	142
6.3.2. Application on Doubles Matches	142
6.3.3. Analysis of Rallies and Serving	142
6.3.4. Back-testing with WTA Matches	143
6.3.5. 2-tier Common-Opponent Model	143
6.3.6. Expand Data-set to Include Challenger Data	144
Bibliography	144
Appendices	155
A. Selected Ranking Figures and Tables	156

List of Tables

2.1. ATP ranking points structure for larger tournaments (excludes Challenger and Futures tournaments, the Olympics and Tour Finals) . . .	34
2.2. WTA ranking points structure for larger tournaments (excludes ITF Circuit tournaments, the Olympics and Tour Finals)	36
2.3. A simple website example for a two-sample Z-test.	44
3.1. Results from a 3-month all surface back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 6551 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	88
3.2. Results from a 6-month all surface back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 7184 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	88
3.3. Results from a 12-month all surface back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 7211 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	89
3.4. Results from a 12-month surface filtered back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 6501 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	90
3.5. Results from a 24-month surface filtered back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 6916 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	90

3.6. Results from a 36-month surface filtered back-test using the ‘un-combined’, ‘combined’ and Barnett’s models to predict 7051 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	91
3.7. A two-sample Z-test using results from a 12-month all surface back-test of a random model and the uncombined model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.	91
3.8. A two-sample Z-test using results from 12-month all surface back-test of a random model and the combined model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.	92
3.9. A two-sample Z-test using results from 12-month all surface back-test of a random model and Barnett’s model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.	92
3.10. A two-sample Z-test using results from 12-month all surface back-test of the uncombined model and the combined model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.	92
3.11. A two-sample Z-test using results from 12-month all surface back-test of Barnett’s model and the ‘combined’ model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.	93
4.1. Statistical data on matches played with common opponents for Andy Murray and Leonardo Mayer in the second round of 2013 US Open. The data includes all common opponent ATP matches played within 12 months of the modelled match.	103
4.2. Probability of Andy Murray winning against Leonardo Mayer, given data on each of the common opponent match combinations separately.	104
4.3. Results from 3, 6 and 12 month all surface back-tests using the Common-Opponent model to predict the outcome of 7938 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	105

4.4.	Results from a 24-month all surface back-test using the uncombined, combined, Barnett's and Common-Opponent models to predict 7938 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	106
4.5.	Results from a 36-month all surface back-test using the uncombined, combined, Barnett's and Common-Opponent models to predict 7938 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	106
4.6.	Results from 12, 24 and 36 month surface filtered back-tests using the Common-Opponent model to predict the outcome of 7938 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.	107
4.7.	A two-sample Z-test using results from a 24-month all surface back-test of a random model and the Common-Opponent model for 7000 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.	108
4.8.	A two-sample Z-test using results from a 24-month all surface back-test of the 'uncombined' model and the Common-Opponent model for 7000 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.	108
4.9.	A two-sample Z-test using results from a 24-month all surface back-test of the 'combined' model and the Common-Opponent model for 7000 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.	109
5.1.	Comparing different PageRank tennis ranking approaches to the official ATP Tour rankings as they were on 01/01/2013. The rankings contain a total 303 players who participated in ATP Tour matches over the year of 2012.	117
5.2.	Comparing different PageRank tennis ranking approaches to the official ATP Tour rankings as they were on 01/01/2014. The rankings contain a total 305 players who participated in ATP Tour matches over the year of 2013.	117

5.3. Comparing different SortRank and LadderRank tennis ranking approaches to the official ATP Tour rankings as they were on 31/12/2012. The rankings contain players who participated in ATP Tour matches over the year of 2012.	132
5.4. Prediction results of all presented ranking systems for 2669 matches played in year 2013.	134
5.5. P-Values using two-sample Z-Tests between the prediction success rates of all ranking systems. All significant results using the 95% confidence level are marked in bold.	135
A.1. Top 100 players as ranked by Match PageRank using matches from 2013.	156
A.2. Top 100 players as ranked by Set PageRank using matches from 2013.	159
A.3. Top 100 players as ranked by Game PageRank using matches from 2013.	162
A.4. Top 100 players as ranked by Point PageRank using matches from 2013.	165
A.5. Top 100 players as ranked by LadderRank Combined system with parameter $X=1$ using matches from 2013.	168
A.6. Top 100 players as ranked by LadderRank Combined system with parameter $X=3$ using matches from 2013.	171
A.7. Top 100 players as ranked by LadderRank Combined system with parameter $X=3$ and a minimum of 5 matches played using historical data from 2013.	174

List of Figures

1.1. The Common-Opponent model has inspired a new approach which is featured in tennisinsight.com headlines.	21
1.2. Screenshot of the tennisinsight.com statistics feature using the Common-Opponent inspired approach.	22
2.1. The tennis court schematic.	26
2.2. Commemorative plaque in Wimbledon for the longest match in the history of the tournament.	30
2.3. Uniform distribution pdf and cdf plots for $a = 0$ and $b = 1$	38
2.4. Bernoulli distribution pmf and cdf plots with parameter $p=0.4$	39
2.5. Binomial distribution pmf and cdf plots with parameters $n=100$ and $p=0.4$	40
2.6. Normal distribution pdf and cdf plots with parameters $\mu=0$ and $\sigma^2=1$	41
2.7. A Discrete-Time Markov Chain	46
2.8. Random walk from Deuce using a single parameter p which is the probability of Player A winning a point.	47
2.9. The distribution of Rafael Nadal's probability of winning points while serving and returning over 124 matches played within January 2012 to December 2013 ATP tournaments.	50
3.1. Markov Chain of a Tennis Match	61
3.2. Markov Chain of a Tennis Tiebreaker Set (the tiebreaker game is defined in Figure 3.4)	63
3.3. Markov Chain of a Tennis Game	64
3.4. Markov Chain of a Tiebreaker Game	66
3.5. Markov Chain of a Tennis Point	67
3.6. Markov Chain of a Serve	69
3.7. Markov Chain of a Rally	70

3.8. A screenshot of the statistics of a single match between Jo-Wilfried Tsonga and Andy Murray as presented by the ATP World Tour website.	71
3.9. A screenshot of the collective statistics of Jo-Wilfried Tsonga as presented by the TennisInsight.com website.	72
3.10. Distributions of Roger Federer’s statistics of 127 matches played in a two year period.	74
3.11. Distributions of Novak Djokovic’s statistics of 147 matches played in a two year period.	75
3.12. Distributions of Gael Monfils’ statistics of 83 matches played in a two year period.	76
3.13. Distributions of Marcos Baghdatis’ statistics of 91 matches played in a two year period.	77
3.14. Distributions of John Isner’s statistics of 115 matches played in a two year period.	78
3.15. The cumulative profit of 7132 bets with exposure 1 unit against the best match opening odds from 5 bookmakers and 4 other bookmaker’s over-round-corrected opening odds.	94
3.16. The cumulative profit of 6973 bets with exposure 1 unit against the best match opening odds from 5 bookmakers.	95
4.1. Probability of the dominant player winning a best-of-three-sets tennis match with fixed differences of 0.01, 0.02, 0.05 and 0.10 in the two players’ probability of winning a point on serve.	100
4.2. Parameters of the Common-Opponent Model.	101
4.3. The cumulative profit of 6923 bets with exposure 1 unit each, against the best match opening odds from 5 bookmakers and against the over-round-corrected odds of 4 individual bookmakers over all ATP Tour matches in the years 2011–2013.	110
5.1. The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the Match PageRank system.	121
5.2. The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the Set PageRank system. .	122
5.3. The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the Game PageRank system.	123

5.4.	The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the Point PageRank system.	124
5.5.	LadderRank pseudocode	126
5.6.	The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the QuickSort Uncombined system.	128
5.7.	The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the LadderSort Combined system with $X=1$	129
5.8.	The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the LadderSort Combined system with $X=3$	130
5.9.	The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the LadderSort Combined system with $X=3$ and minimum of 5 matches played.	131
5.10.	The cumulative profit of 2364 bets with exposure 1 unit each, against the best match opening odds from 5 bookmakers and against the over-round-corrected odds of 4 individual bookmakers over 2562 ATP Tour matches in the year 2013.	137
6.1.	Two tier common opponent example.	143
A.1.	The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the Match PageRank system.	178
A.2.	The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the Set PageRank system. .	179
A.3.	The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the Game PageRank system.	180
A.4.	The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the Point PageRank system.	181
A.5.	The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the LadderRank Combined system with $X=1$	182
A.6.	The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the LadderRank Combined system with $X=3$	183

A.7. The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the LadderRank Combined system with $X=3$ and a minimum of 5 matches. 184

1. Introduction

1.1. Motivation

Tennis has developed, over the years, to one of the most popular spectator sports. Professional tennis players tour all over the world, during the tennis season, competing in very prestigious and at the same time profitable events. The sport's player-rivalry sagas, as well as its apparent unpredictability and potential for dramatic match turnarounds, attract world-wide interest and engages spectators like no other sport. Wherever there is such crowd engagement there is almost always a market for bets and it is needless to say that the market is thriving for tennis.

With the introduction of live online betting, financial markets related to tennis have proliferated allowing traders to speculate on numerous outcomes – e.g. the likely winner of a match or the expected number of aces. In the above context, quantitative models of tennis have gained importance as they provide an understanding of the sport and allow traders to make educated guesses for any outcome.

A good quantitative model can benefit the online sports trader by providing the knowledge required to build a successful betting strategy. Indeed a number of hedge funds, such as the Priomha Capital Sports hedge fund, have turned their attention to sports markets in order to exploit the short-term market inefficiencies that arise. In fact, some of the hedge funds claim significant growth in the past years, e.g. Sports Trading Club boasts figures such as 61% trading profits in the first quarter of 2014 [1]. It appears to be highly possible to make profit using algorithmic trading in online betting exchanges. The sports trader who utilises statistical analysis to manage risk, can take advantage of the fact that many participants are driven by emotion rather than logic which creates opportunities to make profit. This is not without risk though; the Centaur Galileo fund was a sport based hedge fund that was forced to liquidate because, since it opened in 2010, lost \$2.5 million dollars in investments [2].

Of course, the uses of quantitative tennis models are not limited to sports trading. They provide invaluable tools for bookmakers who can use them to estimate odds,

detect fraudulent activities, such as fixed matches, and even provide tools for their customers to make betting suggestions, expanding their services.

Statistical analysis may also be used by the players themselves to reveal a player's weaknesses and strengths and to consequently devise game strategies against particular opponents or focus their own training to specific areas. Taken to extremes quantitative models can be used to simulate match strategies and see what their effect would be on particular opponents.

Additionally, statistical models for tennis are also valuable to broadcasting stations as they can be used to estimate match duration and hence arrange their broadcasting schedule accordingly. They can also be used by sports commentators to increase the interest of the broadcast information.

The International Tennis Federation has used quantitative models in the past to assess rule change proposals by simulating what would happen if a rule would change. There are in fact a number of articles in literature that use quantitative analysis to propose new scoring systems and tournament structures. Quantitative models have also been used in other tennis areas, such as the creation of new equipment and even the prevention of player injuries or the prediction of retirement age.

It is fact then, that good quantitative tennis models are desirable for a range of applications. This dissertation develops a number of quantitative tennis models in an attempt to provide some insight towards the result of professional tennis matches and therefore is directly applicable to sports traders and bookmakers. With a currently booming market of online sports-betting and future predictions (by Bank of America and Merrill Lynch) that the on-line sports market will be worth over \$500 billion by 2015, models like the ones proposed in this thesis are in high demand [3].

1.2. Objectives

The present research aims to develop quantitative models which yield insights into the outcome of professional tennis matches using existing and freely available historical data. A fair amount of research has already been directed towards this goal.

A number of authors have attempted to quantitatively model tennis in the past. The most widely used approach in literature is to model tennis as sequence of independent contests. By modelling a tennis game as a sequence of independent and identically distributed points, one can model a game as a sequence of points, a tennis set as a sequence of games and a match as a sequence of sets. This hierarchi-

cal Markov chain approach appears throughout the literature and is the underlying idea whether authors end up with closed form equations or conditional probability equations. Some authors argue that, in fact, tennis points are neither independent nor identically distributed and present solutions taking that into account. A few papers outline more unique approaches to modelling tennis, for example, by constructing quantitative models using player rankings or even using machine learning techniques. Even less publications, though, discuss how to parameterise the models using existing historical data. A complete analysis of the literature can be found in Chapter 2.

Exploring the literature, it was evident that there was no attempt to model a tennis point, in detail, for the purpose of estimating the probability of winning a point while serving – a probability which is used as a parameter in most hierarchical quantitative models in the literature. Also, few papers discuss how to use historical data to represent specific contests between two players. In an attempt to fill this gap, we develop a Markov chain of a tennis point and discuss how to combine one player’s serving performance with the opponent’s returning performance to improve on the parameters’ representation of a specific contest. Additionally, we introduce a novel approach which we named the “Common-Opponent” model, which limits the use of the available data to a few related matches which are more representative of the match being modelled.

Objectively we will attempt to achieve the development of novel quantitative tennis models, which provides insight towards the outcome of professional tennis matches by:

- Modelling a single tennis point as a Markov chain and using it as a parameter to existing hierarchical Markov tennis models.
- Experimenting with different subsets of historical data and arriving at conclusions on the impact they have on the prediction accuracy of the models.
- Exploiting the transitive component of tennis matches, using statistical data from a small set matches which only include opponents faced by both players being modelled.
- Introducing new algorithms for generating professional tennis player rankings and using the generated rankings for prediction.

1.3. Contributions

This dissertation approaches the problem of forecasting professional tennis results from three different angles. Firstly, we introduce a method to estimate the probability of players winning points while serving against particular opponents, expanding existing models which use this parameter extensively. Secondly, we propose a completely novel approach to modelling a tennis match, using statistics from matches which are linked to the players via common opponents. Finally, we explore tennis rankings, propose new techniques of ranking players and use these generated rankings to quickly predict match results.

1.3.1. Point Model

The probability of players winning points on their serve is the basic parameter in the majority of tennis match models. While there are multiple publications about modelling tennis matches using the probability of winning points, there are few that discuss how to calculate this probability.

We solve this problem by analysing the tennis point, creating a Markov chain in the process and generating closed form equations for calculating the probability of a player winning a service point. These equations use probabilities of specific events occurring during the point, such as the probability of serving an ace, the probability of entering a rally and winning it and various others. We then show how to calculate the probabilities of those events occurring, using publicly available statistics, and adjusting them for specific opponents.

Further analysing historical data, we discuss and analyse the impact (on prediction efficiency) of using different subsets of match statistics and identify possible pitfalls. Expanding our point model to predict match results, using existing techniques, we are able to compare its performance with industry-standard models to find that it is of similar quality.

1.3.2. Common-Opponent Model

In an attempt to eliminate the pitfalls of using averaged statistics over a wide variety of opponents, we created a completely novel approach to forecasting the result of a professional tennis match. The Common-Opponent model avoids using such averaged statistics by comparing differences in the performance of players against their common opponents and combining them to predict match outcomes. We



Figure 1.1.: The Common-Opponent model has inspired a new approach which is featured in tennisinsight.com headlines.

have, therefore, contributed a model which exploits the transitive component of tennis to predict match results. When compared to other predictive models, including industry standard models, the Common-Opponent model is of similar performance, while still having an unexploited potential for improvement and further research.

This method has, in fact, been enthusiastically adopted by the community. In particular, it has inspired a new range of statistical analysis options on the tennis website, tennisinsight.com, whereby common opponent statistics can be viewed across matches (see Figures 1.1 and 1.2).

1.3.3. Ranking Systems

Diversifying our approach to tennis outcome predictions, we focused on tennis rankings. A number of authors have created models which make use of tennis rankings to estimate the probability of professional players winning matches against particular opponents but few have focused on improving existing rankings.

In this dissertation we explore the existing PageRank ranking system for tennis and improve on it, as well as introduce a new concept for a versatile ranking algorithm which can make use of any predictive model to generate rankings. Namely, we introduce new PageRank ranking systems which use quantities of Sets, Games and Points lost as weights in their network. Also we introduce the SortRank and LadderRank algorithms used in conjunction with the low-level point model we

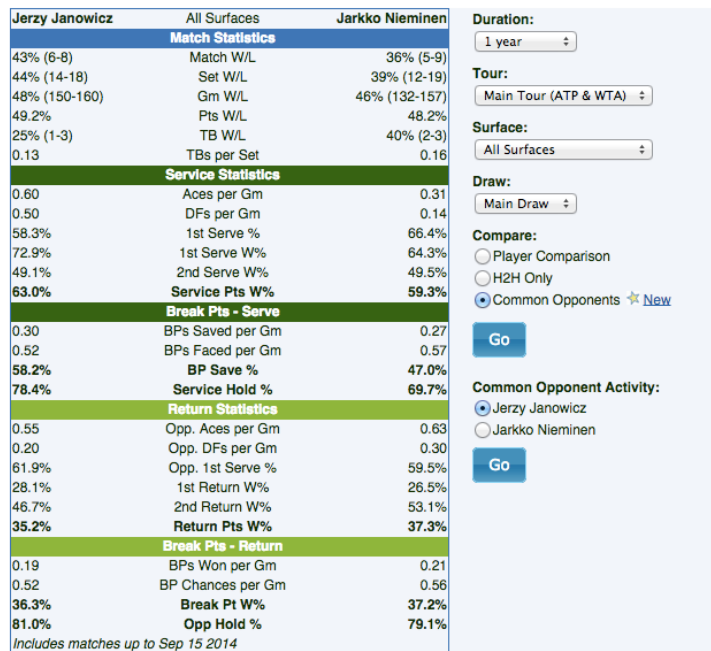


Figure 1.2.: Screenshot of the tennisinsight.com statistics feature using the Common-Opponent inspired approach.

generated earlier.

We compare all these ranking systems to the official ATP Rankings and identify their strengths as well as their weaknesses. In various cases, the new introduced ranking systems significantly outperform the official ATP Rankings in representing the subset of data that was used to generate them. Furthermore, we test the ability of these ranking systems to predict future matches to find that two of the ranking systems, the PageRank Match and Set, perform significantly better than the official ATP Rankings.

1.4. Thesis Structure

The thesis is structured in six chapters.

- **Chapter 2** is divided into three main sections. The first section describes the sport of tennis, introducing the rules of the game and the structure of professional competitions. The second section provides an overview of the theoretical background required in understanding the remainder of this dis-

sertation. Finally the third section explores the literature and the already existing work related to mathematical models involving the sport of tennis.

- **Chapter 3** begins by introducing the idea of hierarchical Markov chains and how they can be used to model the sport of tennis in great detail making the important assumption that individual tennis points are identically distributed and independent from each other. It then continues by analysing individual points as Markov chains by assuming that individual strokes are independent and identically distributed. Once the theory has been introduced, we explain how to put it in practice by showing how to collect the data which will be used to parameterise our models. We discuss how one can use this data and how different subsets of data can affect the predictive power of the model. Finally, we introduce a method to evaluate tennis models and by implementing our model we evaluate predicted results by comparing them against real match results.
- **Chapter 4** approaches professional tennis outcome prediction from a new perspective attempting to improve on some disadvantages of the models discussed in the previous chapters. By taking advantage of the transitive element in tennis, we introduce a model that combines a subset of data, which includes only common opponent matches, to estimate the probability of winning a match. The chapter begins by introducing the shortcomings of past models and explaining the notion of transitivity in sport. It then continues to introduce the work of O'Malley, which is vital in the understanding of the Common-Opponent model. It then continues to explain the Common-Opponent model and in-turn evaluate it comparing its performance against other models. Finally it concludes by explaining the results and problems of this approach.
- **Chapter 5** turns focus to the area of tennis rankings in attempt to find the better performing players and in turn use these rankings to simplify match winner prediction. Three main algorithms are discussed in this chapter, PageRank, SortRank and LadderRank. For the PageRank algorithm we present four different techniques of generating rankings and discuss their effectiveness by evaluating how well the final set of rankings represents the set of matches used to generate them. We take similar approaches for SortRank and LadderRank algorithms. Finally, we show how these rankings can be

used to predict match results and present results based on real world data.

- **Chapter 6** concludes the dissertation by discussing the work done over the years and the results produced. It also introduces work which can still be done to further improve the models presented in this thesis.

1.5. Publications

The following articles were published as part of the research related to this thesis and will be closely referenced in the duration of this dissertation.

The first publication [4] presented at the **3rd IMA International Conference on Mathematics in Sport** introduces a Markov chain which models a tennis point and presents some preliminary results which show promise. This publication was later evolved and published as a journal article [5] in a special edition of the **IMA Journal in Management Mathematics**, presenting extended results. Chapter 3 explores this journal paper in greater detail and presents more recent results.

Continuing the work on tennis prediction, we published another article in the journal of **Computers and Mathematics with Applications** [6], in which we discuss a new approach which uses common opponent matches to exploit the transitive element of tennis. Chapter 4 is based on this journal publication presenting new evidence on the effectiveness of this model for match outcome prediction.

Turning our attention towards methods of ranking players in tennis, we explored the application of the PageRank algorithm to ranking players. Our publication [7] presented in the **9th UK/European Performance Evaluation Workshop (UKPEW/EPEW 2012)**, expands the work of Radicchi [8] by presenting a more efficient algorithm and evaluating new data.

Further research in the topic of tennis player rankings resulted in a another publication [9] at the **4th International Conference on Mathematics in Sport**. This paper presents a new tennis player ranking algorithm that combines quantitative tennis models with a sports ladder. Chapter 5 of this dissertation explores tennis rankings and discusses all these techniques in greater depth. During the **4th International Conference on Mathematics in Sport** we also presented a second tennis related paper [10] which introduces the possibility of inferring live tennis match score-lines using only a live feed of betting exchange odds. Results presented in this paper demonstrate the feasibility of score inference from betting odds in tennis and simultaneously indicates the efficiency of current exchange odds.

2. Background

This chapter introduces the game of tennis by describing the rules, the scoring system and its variations, the way professional tournaments work and how professional players are ranked. It then briefly introduces basic probability theory, statistical testing and stochastic processes required in the understanding of the topic of tennis modelling and of the work contributed by this thesis. Finally the chapter includes a comprehensive literature review which covers the most important work done related to mathematical modelling of tennis.

2.1. The Game of Tennis

This section will introduce the rules and scoring system of the game of tennis while defining tennis specific terminology in the process. The aim of this section is to ensure the reader has complete knowledge of the specifics of the sport of tennis. This section will also discuss tournament structure and identify rules and scoring system changes which apply to specific tournaments. To find further details on the rules of tennis, it is advisable to read Exhibit I of the Official ATP Rulebook [11].

2.1.1. Rules

This section will attempt to briefly describe the rules of tennis aiming to give the reader an understanding of the flow of the sport and acquire some of the terminology used in tennis. The rules described in this section are a summary of the tennis rules as described in the 2014 Official ATP Rulebook.

The sport of tennis is played in a rectangular court of dimensions 23.77m in length and 8.23m in width. In the case of a doubles tennis match, the court widens further to 10.97m. The court is split in the middle lengthwise by a net which stands at 0.914m height in the center and 1.07m at the poles on which it is supported. The baselines of the court are defined as the two lines which define the ends of the court which run in parallel to the net. The sidelines of the court are the four lines

that run perpendicular to the net along the length of the court and define the sides of the singles and doubles courts.

As it can be observed in Figure 2.1, each side of the court also has two service courts or service boxes. These service boxes define the area in which a service must land before it is struck by the returner. In a singles tennis game, each player stands on opposite sides of the net and one player serves while the other receives the ball. The server starts by serving from the right half side of the court behind the baseline and is to serve the ball to the service box diagonally from him where the receiver expects to return the ball. The service location alternates between right and left halves after every point. Who will serve first is decided by a coin toss before the start of the match. The winner of the toss can choose whether he serves or receives first. The server changes with every new game except in the case of a tie-breaker. More will be discussed about the serving order when the scoring system is introduced.

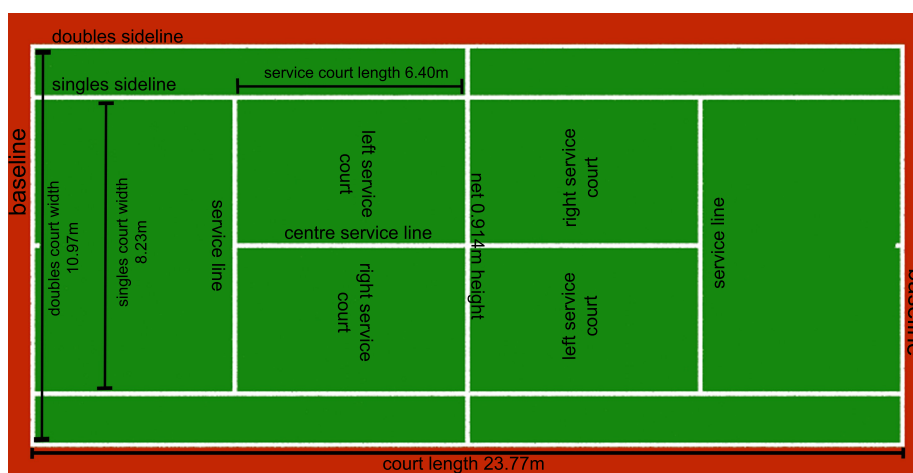


Figure 2.1.: The tennis court schematic.

During the serve a service fault may occur for any one of three reasons:

- The player is not standing in the correct half of the court and at rest before starting the service motion or does not execute the service motion correctly. The service motion begins when the server releases the ball into the air with his hand and ends when the server strikes the ball with his racket.
- The ball does not land within the confines or on the lines of the desired service box.

- A foot fault – this may be caused when during the service motion the server walks or runs out of position, touches the baseline or inside the court with either foot or touches an area on the left or right hand side of the half of the court the server is standing in.

In the case of a service fault, a first serve is followed by a second serve from the same half of the court unless that half was the wrong one to begin with. If a service fault occurs during the second serve, the server forfeits the point.

A service may be repeated in the case a let occurs during the service. A let may occur if the ball, which has been served, bounces off the top of the net or otherwise touches the net, and then lands in the correct service box, or if the ball touches the receiver before hitting the ground or, finally, if the ball is served before the receiver is ready to receive.

The aim of of the two competing teams or players is to win a point at the end of every exchange. A player will forfeit a point if:

- as a server he succumbs to two consecutive service faults; or
- as a receiver he returns the service before it bounces off the ground; or
- the player does not return a ball in-play before it bounces twice; or
- the player fails to return the ball and achieve a good return (defined later on);
or
- the player catches the ball or touches it more than once before returning it in-play; or
- the player touches the net or opponent's side of the court while the ball is in-play; or
- the player hits the ball before the ball passes over the net; or
- the ball while in-play, touches anything other than the player's racket or the player's racket when not held by the player; or
- the player changes the shape of the racket in any way when the ball is in-play.

A player can achieve a good return when the returned ball lands inside the confines of the opponent's court or on the baseline and sidelines in the opponent's

half of the court. The return is still considered good even if, before it lands on the ground, it touches the net (provided that it passes over it). The player's racket may even pass over the net after hitting the ball on the player's side of the court and be considered a good return given it lands correctly.

This section covers the basic and most common occurrences of events and rules within a tennis match. The ATP Rulebook [11] covers a much more detailed account of the rules and addresses more obscure events which are beyond the scope of this thesis. For the purpose of understanding the nature and flow of the sport of tennis the rules included here will suffice.

Having achieved an understanding of how a player can win or lose a point while serving, the only issue remaining in order to understand how a player can win a tennis match is to describe the scoring system that is used and the order of serve.

2.1.2. Scoring System and Order of Serve

The game of tennis has a very granular and hierarchical scoring system which is also the reason why it is such a great candidate for creating mathematical models to represent it. Granularity provides an abundance of data which makes it easier to historically assess the performance of individual players and the hierarchical nature of the scoring system allows mathematicians to simplify models by making them hierarchical.

A tennis match is won by winning a number of sets with the number being dependant on the tournament. The sets are won by winning a number of games and in some cases tiebreakers depending on the tournament and set. Games and tiebreakers are in turn won by winning a number of points.

Game Scoring

A tennis game is played as a sequence of tennis points. For the duration of a singles tennis game, one player is the server and the other one is the receiver. The score-line starts with 0-0 or as often referred in tennis terms, "*love all*". Now let's assume that the server wins all the points the score-line will develop as follows: 15-0, 30-0, 40-0 and finally victory for the server – notice that the server's score is always the first number on the score-line. The same score-line progress applies to the receiver. The victor of the game is the player that wins a point while having a score of 40 given his opponent has a score of 30 or less. In the case the score-line is either 40-30 and the receiver wins the point or 30-40 and the server wins the point,

the score-line becomes *Deuce*. From *Deuce*, a player needs to win two consecutive points in order to win the game. When the first point is won from *Deuce* it is said that the player has an *Advantage*. When a player has an *Advantage* and loses the point the score-line reverts to *Deuce*.

The origin of this non-conformative scoring system is unknown but one theory is that the score-line used to follow quarters of the minute clock-hand with the 45 later simplified to 40.

Tiebreaker Scoring

A tiebreaker is a special type of game that is played as the decider of the result of a tiebreaker set. A player can win a tiebreaker by being the first to win 7 points given that the opponent has won 5 points or less. In the case the score-line reads 6-6 then the tiebreaker proceeds until one player has won two consecutive points. Given the tiebreaker is the decider of the tiebreaker set, both players need to serve in order for it to be fair. In fact serving begins with the receiver of the last game, then proceeds with the other player and alternates every two points. Given that players A and B participate in the tiebreaker and player A serves first the serving sequence is as follows: *A B B A A B B A A . . .* The first number on the tiebreaker score-line is always the score of the player who served first.

Set Scoring

There are two types of set scoring systems, the advantage set and the tiebreaker set. In order to win an advantage set a player must be the first one to reach a score of 6 games or more with a gap of two games or more from the opponent. Advantage sets can sometimes take a very long time to complete and as a result they are used in exceptional cases depending on the tournament rules. An example of a very long match, due of an advantage set, is the famous John Isner vs. Nicolas Mahut match during the 2010 Wimbledon. In Wimbledon, if the match score reaches 2 sets all, the final set played is an advantage set. The final score of this match was 6-4, 3-6, 6-7(7-9), 7-6(7-3), 70-68 and it lasted a total of 11 hours and 5 minutes spanning over 4 days of play. A commemorative plaque of this match is on display on the grounds at Wimbledon (Figure 2.2).

To avoid very long matches, tiebreaker sets were introduced. A tiebreaker set's conclusion differs since when a score of 6 all is reached, the winner of the set is decided using a special type of game called a tiebreaker, described in Section 2.1.2.



Figure 2.2.: Commemorative plaque in Wimbledon for the longest match in the history of the tournament¹.

Match Scoring

The match score is a simple count of the number of sets won by either player. A player can win a match by being the first to win 2 sets. Exceptions exist according to tournament rules, e.g. the men's Grand Slams for which it is a requirement to win 3 sets in order to win a match.

Serving Order

Serving is important in tennis as it provides the player who serves an advantage in winning the point. This is because a strong serve can set up the rally to follow to the advantage of the server.

During each game of a match, one player is always the server unless that game is a tiebreaker. Players alternate serve after each game. That is, the receiver of the previous game becomes the server of the current game. In the case of a tiebreaker, the receiver of the previous game is the server of the first point of the tiebreaker.

¹Image source: wikipedia.org - uploaded by user Jonotennis on the 2nd of July 2011.

This alternation of serve continues through sets – i.e. the receiver of the last game of a set becomes the server of the first game of the following set. The server of the tiebreaker is considered to be the player who served first in the tiebreaker and as such the receiver of the first point in the tiebreaker will serve first in the following set.

2.1.3. Tournaments

Tennis tournaments are split into categories based on the points they award towards player rankings. In this section we will focus on the ATP World Tour tournaments and will discuss how these tournaments are set up and specific rules that govern the most important tournaments.

The most prestigious tournament is the World Tour Finals where only 8 players are allowed to participate. These eight players are selected from a priority list composed of the top players in the ATP Official Rankings at the end of the tennis season. These eight players are split into two groups of four who face each other in three best-of-three round-robin matches. The top two seeds are placed in different groups. From these two groups, the best two players emerge from each group to face each other in a knock-out phase which again comprises of best-of-three tiebreaker-set matches. Each round-robin victory awards players 200 points towards the rankings, a semi-final victory awards another 400 points and a final victory awards 500 points. An undefeated champion therefore has the opportunity to amass 1500 points from this tournament².

The four Grand Slams award the highest amount of points (2000 to the champion) and money to the winner. They are highly prestigious and at the same time competitive tournaments. Comprised of 128 very skilled participants, Grand Slams are always crowd and player favourites. Each Grand Slam is played on a different surface, requiring different skill sets from the participants. Wimbledon is played on a natural grass surface which is a fast, low bounce surface. Roland Garros is played on red clay surface which is a slow, high bounce surface. The U.S. Open is played on a blue hard-court surface called DecoTurf which is a fast, medium bounce surface. Finally, the Australian Open is played on Plexicushion which is a medium speed, medium bounce, hard-court surface [12]. Men's Grand Slams are the only tournaments where a best-of-5 tiebreaker-set victory is required to

²Information retrieved from <http://www.barclaysatpworldtourfinals.com/en/event/rules-and-format> on 18/09/2014

progress to the next round. In all Grand Slams with the exception of the U.S. Open, the 5th set is an advantage set [13].

The third category of ATP tournaments are the ATP Masters series which award 1000 points to the winner. There is a total of 8 Masters tournaments in the ATP World Tour, each comprising of draws varying from 96 to 48 participants.

ATP 500 series tournaments award, as the name suggests, 500 points to the winner. In 2014 there was a total of 11 ATP 500 Series tournaments scheduled. The number of participants allowed in an ATP 500 tournament is officially 32 but in reality it varies from 48 to 32 as tournaments can petition for increased draw sizes.

Finally, ATP 250 tournaments feature less prestigious events which award 250 points to the winner. The draw size is officially 28 participants but can vary, depending on the tournament, from 56 participants to 28. In 2014 a total of 40 ATP 250 tournaments were scheduled all over the world.

Seeded Players

Most tournaments, if not all, have seeded players. A number of top ATP Ranked players (number depends on the draw size) who are participating in the tournament are awarded seeded positions in the draw. These players are strategically placed in the knock-out draw so that if they achieve victories they will face each other in the latter rounds of the tournament. If the tournament is concluded in 7 rounds (which is the case in Grand Slams which have 128 participants and the Masters which have 96 participants) then there are 32 seeded players. Tournaments which have 6 rounds, (i.e. tournaments with 56 or 48 participants) usually have 16 seeded positions and finally tournaments with 5 rounds (32 or 28 participants) usually have 8 seeded positions in the draw.

When the draw is constructed, the seeded players are placed in the predefined positions of the draw, and then the remaining players (direct acceptances, qualifying round winners, wild cards and lucky losers) are usually drawn randomly and placed in order in the remaining lines of the draw.

Direct Acceptances

Direct acceptances are players who are ranked high enough in the ATP Rankings to be accepted directly into the tournament with no need to pass through the qualifying tournament. There is a fixed number of Direct Acceptances which depends

on the tournament category and on the draw size.

Qualifiers

Qualifiers are players who have won the final round of the qualifying tournament of the main event. The qualifying tournament is a full tournament with seeded players which is played before the main tournament. Players who manage to qualify are then placed in the main draw of the main event. There is a fixed number of Qualifiers which varies depending on the tournament category and on the draw size but has a minimum of 4.

Lucky Losers

Lucky losers are players who have played in the qualifying tournament of the event but lost in the final round. If for some reason a player already participating in the main draw is unable to attend the tournament, a lucky loser takes his/her place in the draw.

Wild Cards

Wild Cards are players who, at the discretion of the tournament organisers, are allowed to participate in the tournament with no need to pass through the qualifying rounds or have a high enough ranking to be accepted in the draw directly. Wild Cards are usually given to players who have in the past performed well but have dropped in ranking. They may also be given to local talent or local favourite players. The amount of Wild Cards allowed varies, from 3 to 6, depending on tournament type and draw size.

2.1.4. The Official Ranking Systems

Rankings have always been the focal point of both the fans and the players. They are a representation of the players' ability to win matches and persevere over time. Also, they are used to determine seeding positions in tournaments as well as the participation of players in the World Tour finals, which indicate the season champion. This section discusses how these rankings are calculated by both the ATP and the WTA.

ATP

The Emirates ATP Rankings is the official ranking system ATP used in 2013. As the official ATP World Tour website states: it is "... a historical objective merit-based method used for determining entry and seeding in all tournaments ...". The ranking is generated using a summation of points that players acquire while proceeding within tournaments. Tournaments themselves are split into categories with some tournaments awarding more points than others.

The ranking points of a player is the summation of the points awarded over a maximum of 18 tournaments played within the previous 52 weeks (19 tournaments if the player qualifies to the World Tour Finals). From these 18 tournaments, four are the Grand Slam tournaments, eight are the compulsory ATP World Tour Masters 1000, and the rest are the best six results from the ATP 500, 250 and other tournaments (a minimum of 4 ATP 500 tournament attendances are required). Additionally, players who have finished within the top eight positions, in the official ATP rankings, at the end of the ATP tennis season, automatically qualify to play at the Barclays World Tour Finals to earn points that count towards crowning the final champion of the year. In those years where the Olympics occur, the players also win extra points for the position they get in the Olympics.

Table 2.1.: ATP ranking points structure for larger tournaments (excludes Challenger and Futures tournaments, the Olympics and Tour Finals)

	W	F	SF	QF	R16	R32	R64	R128	Qual. ³
Grand Slams	2000	1200	720	360	180	90	45	10	25
Masters 1000	1000	600	360	180	90	45	10(25)	(10)	25
ATP Tour 500	500	300	180	90	45	20	-	-	20
ATP Tour 250	250	150	90	45	20	(10)	-	-	12

Points awarded by playing in the ATP Challenger Tour vary from tournament to tournament depending on the tournament's prize money and hospitality. Points for the overall winner range from 75 to 125. For playing in the Futures Series, players are awarded even less points, with points earned by the overall winners ranging from 18 to 35.

The ATP Ranking score is therefore the summation of the points awarded from:

- The four so-called Grand Slam tournaments (Australian Open, French Open,

³Points awarded for qualifying subject to adjustment depending on tournament type

Wimbledon US Open)

- The eight mandatory ATP World Tour Masters 1000 tournaments,
- The previous Barclays ATP World Tour Finals count until the Monday following the final regular-season ATP event of the following year.
- The best six results from all ATP World Tour 500, 250, ATP Challenger Tour, and Futures Series tournaments played in the calendar year (a minimum of 4 ATP 500 tournaments must be included).

In those years when the Olympics are held, results from the Olympics also count towards a player's world ranking.

Table 2.1 shows the points awarded according to the tournament type and round (beginning with Qualifying, and ending with the Final) in which a player is eliminated – or if they win the tournament.

WTA

Similarly to ATP rankings, a player's WTA ranking is computed over the immediate past 52 weeks, and is based on the total points a player accrues at a maximum of 16 tournaments. As shown in Table 2.2, points are awarded according to the round in which a player is eliminated in or for winning the tournament. The tournaments that count towards the ranking are those that yield the highest ranking points. These must include:

- The four Grand Slam tournaments (Australian Open, French Open, Wimbledon US Open)
- Premier Mandatory tournaments (Indian Wells, Miami, Madrid, Beijing)
- The WTA Championships (Istanbul)

The qualifying points awarded for the tournaments in Table 2.2 may vary depending on the tournament's draw size.

For top 20 players, their best two results at Premier 5 tournaments (Doha, Rome, Cincinnati, Montreal, Toronto and Tokyo) also count. Like in the ATP tour, in those years when the Olympics are held, results from the Olympics also count towards a player's world ranking.

Table 2.2.: WTA ranking points structure for larger tournaments (excludes ITF Circuit tournaments, the Olympics and Tour Finals)

	W	F	SF	QF	R16	R32	R64	R128	Qual.
Grand Slams	2000	1400	900	500	280	160	100	5	60
Premier Mandatory	1000	700	450	250	140	80	50(5)	(5)	30
Premier 5	800	550	350	200	110	60(1)	(1)	-	30
Premier	470	320	200	120	60	40(1)	(1)	-	20
International	280	200	130	70	30	15(1)	(1)	-	16

2.2. Theoretical Methods

This section is meant to provide a brief introduction to the theoretical methods and fundamental knowledge used through-out this thesis. First, we introduce basic probability theory and some commonly used probabilistic distributions. Following that, we describe the fundamentals of significance testing and finally stochastic processes and in particular Discrete-Time Markov chains.

2.2.1. Probability Theory

Probability theory provides the fundamental building blocks for both significance testing and stochastic processes. This section introduces succinctly, the basic formulae and terminology used in probability theory.

A *sample space*, S , is the full set of outcomes of an experiment. E.g. The sample space of flipping a coin consists of the outcomes (events), Heads and Tails. Two events are *mutually exclusive* when their intersection is the empty set.

$$A \cap B = \emptyset \quad (2.1)$$

When the events are *mutually exclusive exhaustive*, then each event in the set is mutually exclusive to all other events in the set and the union of all of the events of the set is equal to the full set itself.

$$A_i \cap B_j = \emptyset \text{ for all } i \neq j \quad (2.2)$$

$$A_1 \cup A_2 \cup \dots \cup A_n = S \quad (2.3)$$

A fundamental concept of probability theory is *conditional probability* i.e. the

probability of an event occurring given another event is known to have occurred. The probability of event A occurring given event B is known to have taken place is denoted by $P(A | B)$.

$$P(A | B) := \frac{P(AB)}{P(B)} \quad (2.4)$$

where $P(B) \neq 0$

Two events can be called *independent* when the probability of both events happening is the probability of one event multiplied by the probability of the other.

$$P(A \cap B) = P(A)P(B) \quad (2.5)$$

The conditional probability of two independent events A and B is then the probability of each event happening independently of the other event.

$$P(A | B) = P(A) \quad (2.6)$$

$$P(B | A) = P(B) \quad (2.7)$$

Two other important theorems in probability theory is the theorem of Total Probability and Bayes' theorem [14]. *Total Probability* states: *For an event B and a set of mutually exclusive exhaustive events A_1, A_2, \dots, A_n , if event B occurs it must occur with exactly one of the mutually exhaustive events A_i then*

$$P[B] = \sum_{i=1}^n P[A_i B] \quad (2.8)$$

Bayes' theorem states: *For a set of mutually exclusive and exhaustive events A_i then*

$$P[A_i | B] = \frac{P[B | A_i]P[A_i]}{\sum_{j=1}^n P[B | A_j]P[A_j]} \quad (2.9)$$

2.2.2. Common Distributions

Uniform Distribution

The uniform distribution in probability and statistics is used to describe uniform random variables. These are variables that can take any value within a range (a, b) with identical probability. The uniform distribution is usually denoted as $unif(a, b)$. The probability density function (pdf) and the cumulative density func-

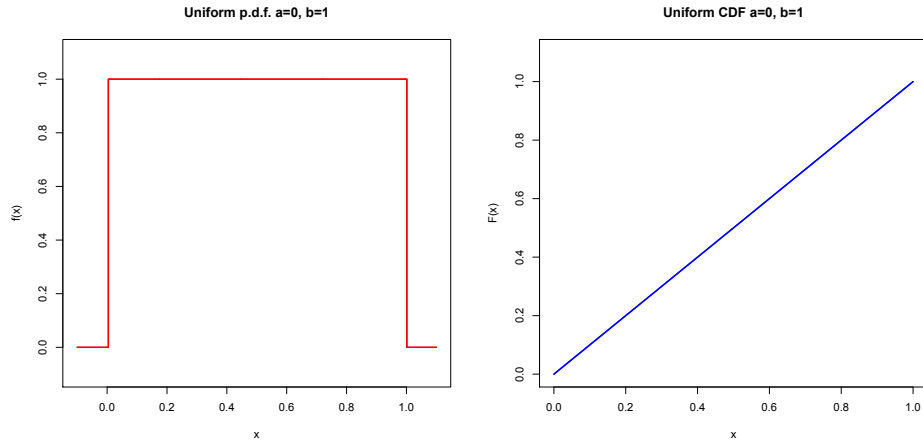


Figure 2.3.: Uniform distribution pdf and cdf plots for $a = 0$ and $b = 1$.

tion (cdf) of the uniform distribution are described by Equations 2.10 and 2.11 respectively and demonstrated by Figure 2.3.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \quad (2.10)$$

$$F(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } a \leq x < b, \\ 1 & \text{for } x \geq b \end{cases} \quad (2.11)$$

The mean, $E(X)$ and variance, $\text{Var}(X)$ of the uniform distribution are described by Equations 2.12 and 2.13.

$$E(X) = \frac{1}{2}(a+b) \quad (2.12)$$

$$\text{Var}(X) = \frac{1}{12}(b-a)^2 \quad (2.13)$$

Bernoulli Distribution

A random variable which takes the value of 1 with a probability p and otherwise the value of 0, can be described by the Bernoulli probability distribution. This random variable is most commonly used as a binary success/failure variable. It takes the value 1 for a success with a probability of success, p . It is implied that

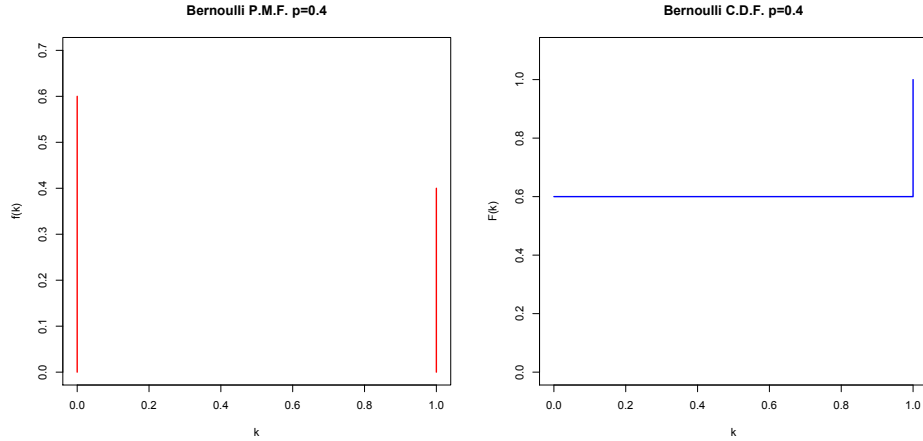


Figure 2.4.: Bernoulli distribution pmf and cdf plots with parameter $p=0.4$.

the probability of failure is therefore $q = 1 - p$. The probability mass function (pmf) and cumulative distribution function (cdf) of the Bernoulli distribution with success probability p are described by Equations 2.14 and 2.15 [15].

$$f(k) = \begin{cases} p & \text{for } k = 1, \\ 1 - p & \text{for } k = 0 \end{cases} \quad (2.14)$$

$$F(k) = \begin{cases} 0 & \text{for } k < 0, \\ 1 - p & \text{for } 0 \leq k < 1, \\ 1 & \text{for } k = 1 \end{cases} \quad (2.15)$$

The mean, $E(X)$ and variance, $\text{Var}(X)$ of the Bernoulli distribution are described by Equations 2.16 and 2.17.

$$E(X) = p \quad (2.16)$$

$$\text{Var}(X) = p(1 - p) \quad (2.17)$$

Binomial Distribution

The binomial distribution is related to the Bernoulli random variable. It represents the discrete probability distribution of achieving k successes in n successive independent and identically distributed binary experiments with probability of success p . That means that when $n=1$ then the binomial distribution is identical to the

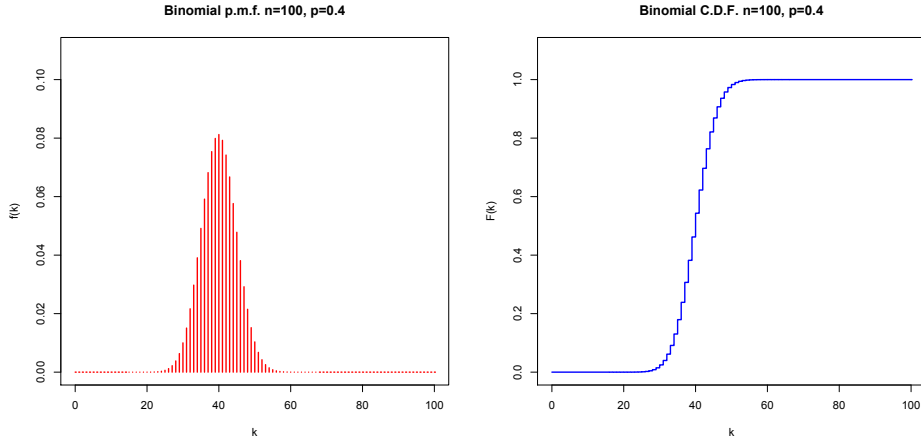


Figure 2.5.: Binomial distribution pmf and cdf plots with parameters $n=100$ and $p=0.4$.

Bernoulli distribution. An example of a binomial distribution with 100 trials and a probability of success 0.4 is demonstrated in Figure 2.5.

The binomial distribution, with parameters: the number of trials, n , and probability of success, p , for k successes, is thus represented by the probability mass function shown in Equation 2.18.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (2.18)$$

where:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (2.19)$$

The cumulative density function of the binomial distribution is a summation of all the discrete values of the Binomial pmf up to k as shown in Equation 2.20.

$$F(k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i} \quad (2.20)$$

The mean, $E(X)$ and variance, $\text{Var}(X)$ of a binomial distribution are shown in Equations 2.21 and 2.22 respectively [15].

$$E(X) = np \quad (2.21)$$

$$\text{Var}(X) = np(1 - p) \quad (2.22)$$

Normal Distribution

The Normal distribution is a very important distribution in statistics. It is most commonly used to describe real-world variables with unknown distributions whose average and standard deviation can be estimated. Additionally, the central limit theorem, which is described in detail in Section 2.2.3, states that the means of independent samples drawn from the same distribution are normally distributed.

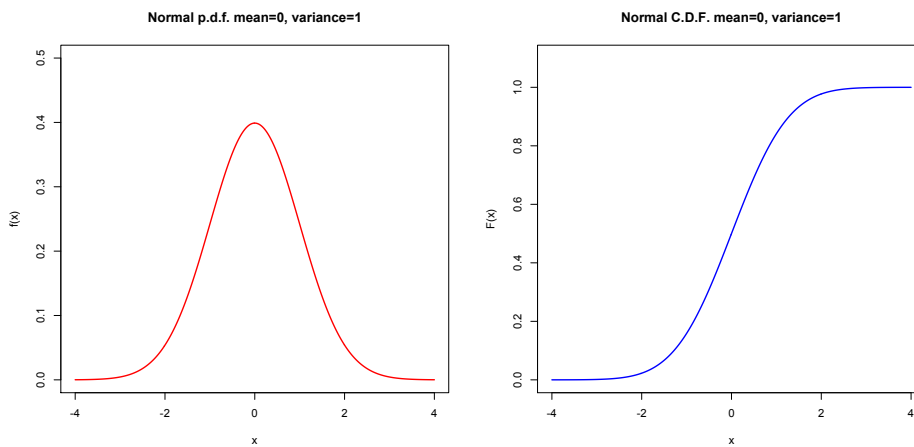


Figure 2.6.: Normal distribution pdf and cdf plots with parameters $\mu=0$ and $\sigma^2=1$.

The Normal distribution takes two parameters, the mean, μ , and the standard deviation σ (or sometimes the variance, σ^2 , instead). The probability density function of the normal distribution with parameters μ and σ is shown in Equation 2.23. The standard normal distribution which is used very frequently is a normal distribution with parameters, $\mu = 0$ and $\sigma = 1$. The standard normal distribution is demonstrated in Figure 2.6.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.23)$$

The cumulative distribution function of the normal distribution is the integral of the normal pdf from $-\infty$ to x which is equivalent to Equation 2.24.

$$F(x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x - \mu}{\sigma\sqrt{2}} \right) \right] \quad (2.24)$$

where

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt \quad (2.25)$$

The mean, $E(X)$, and variance, $\operatorname{Var}(X)$, are in fact the parameters of the normal distribution and are shown below in Equations 2.26 and 2.27 respectively.

$$E(X) = \mu \quad (2.26)$$

$$\operatorname{Var}(X) = \sigma^2 \quad (2.27)$$

2.2.3. Significance Testing

Testing for statistical significance is important in any experiment which uses samples of a population. A significance test is used to make sure that a hypothesis is true with a probability greater than a significance level that has been set. The significance level used to ensure reproducible results is most often the 95% threshold.

In this thesis we will use simple A/B testing or Split tests to compare whether the results of different models are actually different. Since all of our experiments use a high number of samples which have known distribution and hence the variance can be calculated with fair accuracy, we only use the Z-Score for testing.

Central Limit Theorem

The central limit theorem is what allows us to use the normal distribution to calculate the Z-score and subsequently the p-value of our simple tests.

Consider N samples which contain n values each, $\{X_1, \dots, X_n\}$, all generated independently from the same distribution which has a mean μ and a variance σ^2 . For each of the N samples we calculate the sample average, $S_i = \frac{1}{n} \sum_{k=1}^n X_k$ for $i = \{1, \dots, N\}$. The central limit theorem states that the distribution of the sample averages S_i , approximates a normal distribution with a mean equal to the source distribution's mean, μ and variance equal to the the source distribution's variance over the number of values in each sample (i.e. $\frac{\sigma^2}{n}$). This is true regardless of the source distribution [16].

The central limit theorem therefore provides us with a tool to understand how

sample averages can deviate from the population mean and also calculate the probability of a particular deviation occurring.

Z-Score

The Z-Score or standard score is simply a signed measure of standard deviations a data point, x , differs from the mean, μ , of a normal distribution with standard deviation σ . This is calculated using Equation 2.28.

$$z = \frac{x - \mu}{\sigma} \quad (2.28)$$

The Z-Score can be used in conjunction with the standard normal distribution to find the p-value of the Z-Test. The p-value is the probability of getting a sample average with a specific deviation or more from the mean of the population distribution.

For example, let's say we have retrieved a sample with $n=100$ values from a distribution which we know has a mean $\mu = 3$ and standard deviation of $\sigma = 1$. Our sample has an average of $x = 2.98$ and we want to know the probability of getting that average or lower, given we know the mean and standard deviation of the distribution.

According to the central limit theorem we know that the sample average has a normal distribution with mean 3 and variance $\frac{\sigma^2}{n} = \frac{1}{n}$. The z-score of our sample average is therefore $\frac{2.98-3}{1/10} = -0.2$. As a result, we know that the sample average is -0.2 standard deviations away from its mean. Depending on whether we want to do a one-tailed or two-tailed test we can then use the standard normal distribution to calculate the probability of getting a sample average of 0.2 or more standard deviations away from the population mean.

Split Test

The split test (or A/B test) is designed to test whether the results of two experiments differ from one another. Experiment A and experiment B may have different averages but may not be in fact different. This test allows us to distinguish whether the difference occurs due to the natural variation of the sample averages or because in fact the two experiments have different means.

A very common usage for a split test is website feature testing. The experiment is very simple, randomly present website visitors either the original website or

the modified website and observe whether they convert by reaching a target (e.g clicking the check-out button). One can then use a split test on the number of conversions for each of the two versions of the website to determine whether the change observed in conversions is actually statistically significant.

Throughout this thesis we will use two-sample Z-tests for split testing because sample sizes are sufficiently large in all our experiments and therefore the variances of distributions can be estimated with a fair amount of accuracy.

Example: The original website received 1000 visitors out of which 50 purchased something. The modified version of the website received 1000 visitors out of which 59 purchased something. Is this because of chance or is it because of the changes made to the website?

The conversions on a version of the website can be modelled as n Bernoulli trials with expected value equal to the probability of conversion (in this case a purchase) and n equal to the number of visitors.

For such a large n , it is safe to assume that p is a good estimate for the probability of conversion. Therefore for the original website the probability of conversion is 0.05 and for the modified it is 0.059. Knowing the mean, and standard deviation of the Bernoulli distribution makes testing easy since we can calculate the z-score. Since we are comparing two samples for differences in the mean, we are asking the question: Is the mean of the first sample equal to the mean of the second sample? Table 2.3 provides the answer to this question.

Table 2.3.: A simple website example for a two-sample Z-test.

Website	n	p	Standard Error	Z-Score
Original	1000	0.050	$\sqrt{\frac{0.05 \times 0.95}{1000}}$	-
Variation	1000	0.059	$\sqrt{\frac{0.059 \times 0.941}{1000}}$	$\frac{0.059 - 0.05}{\sqrt{\frac{0.05 \times 0.95}{1000} + \frac{0.059 \times 0.941}{1000}}} = 0.8867$

For a Z-Score of 0.8867 using a two-tailed test and a significance threshold of 95%, the result is not significant which means that the means of the two samples can be considered equal.

2.2.4. Stochastic Processes

To define a stochastic process we must first define a random variable. A random variable is a variable which can take an uncertain value. A *discrete random variable* is a variable which has an uncertain value but at the same time its value is one

of a countable set of possible values. The sum of the probabilities of the variable having each value is equal to 1. For example a discrete random variable x can take either one of 3 values x_1, x_2, x_3 with probabilities p_1, p_2 and p_3 . By definition

$$\sum_{i=1}^3 p_i = 1 \quad (2.29)$$

A stochastic process is a random variable which includes the dimension of time. i.e. a stochastic process has an uncertain value at one point in time and another uncertain value at another point in time. An example of a stochastic process is readings of the temperature for every hour of the day. Stochastic processes are classified in terms of three things: “their *state space*; the *nature of the time parameter* and the *statistical dependencies* among the random variables” at different times [14]. If the number of states of a stochastic process is countable then the process is a discrete-state process or chain. If the stochastic process is sampled at a countable intervals then the stochastic process is a discrete-time process. The random variables themselves can either be independent or dependent on each other over time.

Discrete-Time Markov Chains

Discrete-Time Markov Chains are stochastic processes which have a countable number of states in their state space, are sampled at discrete points in time and the random variables are only dependent on their immediately previous state and independent from any other previous states. Therefore a stochastic process is a DTMC given that

$$P[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0] = P[X_{n+1} = x_{n+1} | X_n = x_n] \quad (2.30)$$

for $n \in \mathbb{N}$

A Markov chain can be visualised using a state diagram such as the one in Figure 2.7 . The circles indicate the different states of the Markov chain and the arrows connecting them show the possible paths the chain can follow with the probability that they are followed. For example from state 1, the sequence can move to state 2 with a probability of $\frac{2}{3}$ or to state 3 with a probability of $\frac{1}{3}$.

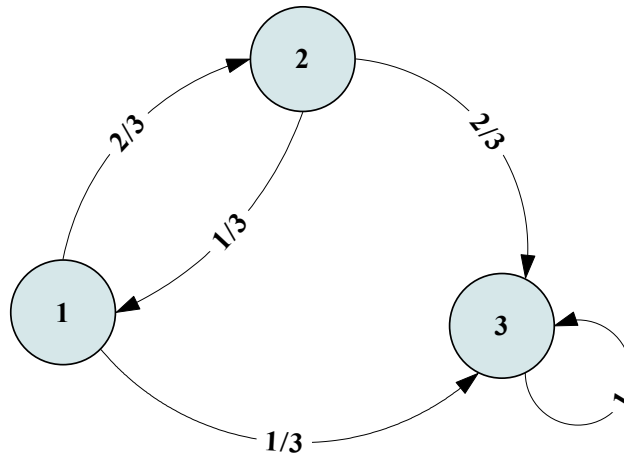


Figure 2.7.: A Discrete-Time Markov Chain

2.3. Literature Overview

Tennis is considered a sport which lends itself to mathematical modelling. In fact, the nature of the granular scoring system which tennis uses, the abundance of match statistics and the popularity of the sport has led many authors to use tennis as an example for teaching applied statistics [17, 18, 19]. Even so, this does not make the problem of modelling a tennis match, a simple matter. Over the years, dozens of authors have published work which attempts to provide an answer to this problem or related aspects. The most important work in the field is discussed in this section.

2.3.1. Hierarchical Match Models

The hierarchical approach to modelling tennis is the most popular among the literature. Kemeny and Snell [17] were some of the earliest authors to model the game of tennis using a hierarchical approach utilising Markov Chains. The hierarchical idea is relatively simple: in order to win the match, a player must win a number of sets, in order to win a set, a player must win a number of games and in order to win the game, a player must win a number of points. Using the probability of a player winning a point, one can hierarchically model the match by making the assumption that points are independent of each other and identically distributed (i.i.d.). In their book, Kemeny and Snell [17] model a tennis game using a single parameter which is a constant probability of a player winning a point throughout

the match, regardless of who is serving. This simple game model was made up of a preliminary Markov chain followed by a random walk (in the case of a deuce) of 5 states with two absorbing states as demonstrated in Figure 2.8.

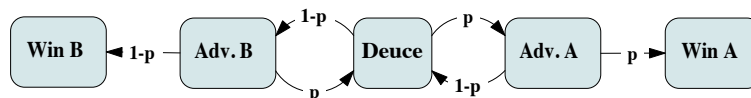


Figure 2.8.: Random walk from Deuce using a single parameter p which is the probability of Player A winning a point.

The simplistic model developed by Kemeny and Snell [17], to demonstrate a Markov chain, disregards the advantage of serve when extended to a match model, something which plays a vital role in tennis. Hsi and Burych [20] stress the importance of the advantage of serve and calculate the algebraic expressions for the probability of winning tennis games, sets and matches by taking this into account. They use a two parameter model and assign two different probabilities of winning a point, one for Player A serving and one for Player B serving.

Carter and Crews [21] later develop a single parameter Markov chain model which uses a constant probability throughout the match. This single parameter is calculated from the average of each player's probability of winning a point on their serve, thus combining the two parameters. Fischer [22] develops a model of winning a match for both tiebreaker sets and advantage game sets but ignores the serving advantage because, as he points out, it averages out through the match.

In the interest of analysing the efficiency of the scoring systems of various sports, including tennis, Miles [23] adopts Bernoulli-type models to represent sequences of contests. Miles uses both a one parameter model (unipoint model) and a two parameter model (bipoint model) for tennis and compares the two. Miles agrees with the "averaging assumption" used by Carter and Crews [21]. This discussion about the disaggregation of statistics continues to this day and it appears that disaggregating statistics makes little difference on the higher level probability of winning the match as it does in fact "average" out. On the other hand when modelling lower level probabilities such as the probability of winning the point or the probability of winning the game, disaggregation of model parameters is important.

Croucher discusses the impact tie-breakers have on a tennis match [24] and introduces the idea that the probability of winning a tennis game changes throughout the game [25]. This is done by presenting conditional probabilities to winning a

single game from any score-line while making some deductions about the most “important points” in a tennis game. Morris [26] also discusses the most “important points” in tennis and proves that they are equally important for both competing players. The importance of points is a concept which appears in a broad variety of papers and has originated from these two papers.

Riddle [27] models a tiebreaker set as a sequence of contests (games) between the two players, using three parameters (the probability that Player A/B wins a game given they are serving and the probability Player A wins a tiebreaker). He then models a tennis game and a tiebreaker as a sequence of contests (points) and formulates equations for the probability of a player winning a game and a tiebreaker when serving first. Riddle [27] also presents empirical evidence that the assumption that the probability of winning a point on serve remains constant throughout the match is valid since the factors that may affect that probability are cancelled out over many matches. Though this is a valid argument for match outcome prediction (since over the many points in the match the factors affecting the probability of winning a point may cancel out) when trying to predict the winner of individual points or even games the factors affecting the probability of winning points become more important.

Liu [28] also used finite Markov chains to derive equivalent closed form equations for calculating the probabilities of winning games, sets and finally the match. He models deuce as a 5 state random walk similarly to Kemeny and Snell [17] and proceeds to find the steady state probabilities of the game Markov chain. Liu [28] also demonstrates a Gambler’s ruin style approach which yields the same solution.

Klaassen and Magnus [29] among other important research, proposed a method that uses a closed form equation of the probability of a player winning the game, for forecasting winners of a tennis match. Using their program and data from Wimbledon they were able to find the probability of a player winning the match not only prior to but during the match. Their method was later used by Easton and Uylangco1 [30] to generate point-by-point probabilities and compare them to betting exchange implied probabilities during live matches. They concluded that odds presented by the exchange are in fact closely related to the model as they found extremely high correlation between the model and exchange probabilities. In fact, betting exchanges respond extremely well to the changing realities of matches given sufficient liquidity and do provide a good guideline towards validation of tennis models.

Barnett and Clarke [31] use the work of Riddle and show how to use a spread-

sheet to predict the outcome of a tennis match. This spreadsheet clearly shows the probability of winning the game/set/match from any point in the match taking as input the probability of each player winning a point on their serve. In a later publication, Barnett and Clarke [32] formulate the conditional probabilities for game, set and match given the probability of winning a point on serve and in the same paper demonstrating a gambling strategy using the model. Barnett and Clarke [33] also present a means to calculate the probability of each player winning a point on their serve by using publicly available statistics. They do that by first calculating the percentage of points won on serve, f_i , for each player i , as well the percentage of points won while returning serve g_i . Calculating f_i from the statistics is straightforward, g_i though is not that simple as the number of first serves that were in play is not known. To overcome that they use the average 1st serve percentage of the top 200 players of the ATP. They then combine f_i with g_i to come up with a probability of player i winning a point on serve when playing against player j . This is done by adding the probability $(f_i - f_{av})$ and subtracting the probability $(g_j - g_{av})$ to the tournament average probability of winning a point on serve. This approach combines the capability of a player winning a point on his serve with the capability of his opponent winning a point while returning serve. This is the first approach that actually accounts for the abilities of the server and the returner and as such, the predictions are much more accurate.

Newton and Keller [34] unify some of the previous literature and derive the probabilities of winning a point, set and match hierarchically. They present evidence that winning a set depends on who will serve first and also calculate the probability of a player to win in a 128-player tournament such as a Grand Slam. They also discuss possible solutions to the non-i.i.d effects of points in tennis. O'Malley [35] explores the properties of the probability function of winning a game on serve by plotting the derivative and integral of the probability function. He concludes that a 0.01 increase in probability of winning a point on serve when it is at 0.5 affects the probability of winning a game much more than when the probability is around 0.7. He also calculates the probability functions of winning the set and match and plots their distributions giving a complete account on their properties. O'Malley [35] also demonstrates the interesting fact that the probability of winning a match is highly dependant on the difference between the two players' probabilities of winning a point on serve and not so much on the values of the individual probabilities.

Newton and Aslam [36] demonstrate the importance of looking at the distri-

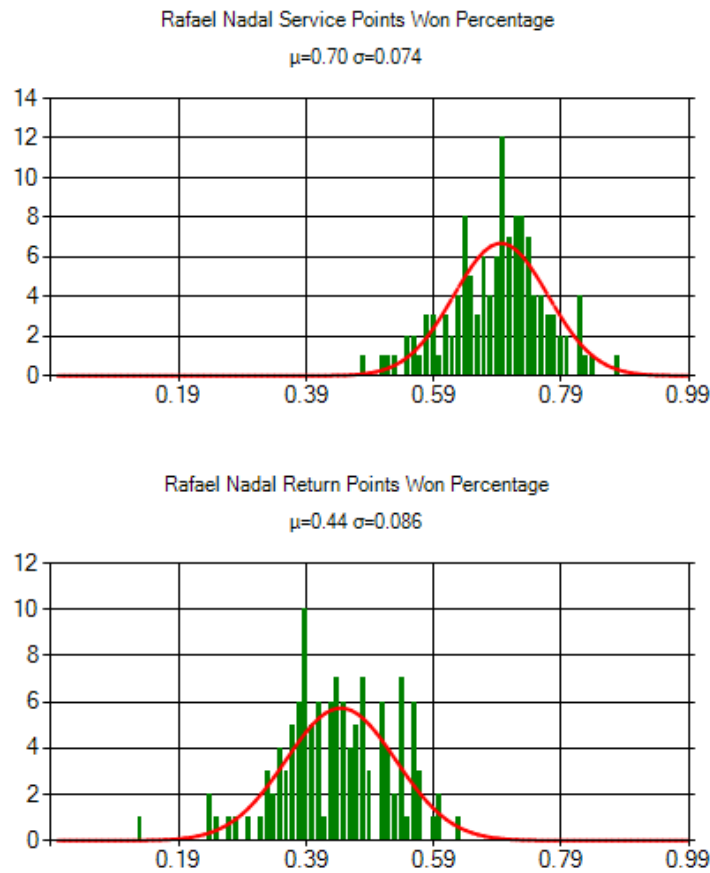


Figure 2.9.: The distribution of Rafael Nadal’s probability of winning points while serving and returning over 124 matches played within January 2012 to December 2013 ATP tournaments.

butions of a player’s probability of winning points and not just their means as in previous literature. They examine how the player’s probability of winning a point on serve and while receiving changes over multiple matches and show that they can be modelled as normally distributed random variables as seen in the case of Rafael Nadal in Figure 2.9. They then build a four parameter model which uses the means of those Gaussian distributions and their standard deviation to model the probability of winning a tennis match. They do that by firstly combining the probabilities of winning a point on serve with the opponents’ probabilities of returning a point in a similar way as Barnett and Clarke [33] propose. They then use those adjusted probabilities as the mean and the player’s standard deviation in

order to generate a truncated gaussian distribution of the match. Next, they sample a random probability from the distribution and use it as the input to a game Markov chain model running Monte Carlo simulations of the entire chain to calculate the probability of a player winning the match. Newton and Aslam [36] also point out that the standard deviation can also be used as a measure of the player's consistency from match to match and over the different surfaces which is a concept which can be explored further, especially when applied to selecting possible wagers.

2.3.2. Independence of Points

The hierarchical models discussed in Section 2.3.1 make the assumption that winning a point on serve during a tennis match is an independent and identically distributed process. This is an important assumption and it simplifies the hierarchical tennis models significantly. It is widely accepted though that factors such as fatigue, injuries and psychological factors based on the importance of the point and winning streaks [37, 38] can greatly impact the probability of winning a point in tennis during a match. Therefore in reality, the probability of winning points on service is both dependant to other points being won and not identically distributed.

Klaassen and Magnus in 1998 [39] investigated, using data from Wimbledon, whether points are i.i.d. and have concluded that in fact they are not. They further propose an extended logit model which attempts to correct for the interdependency of points. In their paper in [40] they also conclude that even though the probability of winning a point during a match is not i.i.d., for professional players the deviation from i.i.d. is small and for the stronger players even smaller. For applications such as forecasting professional tournaments therefore. it is not unacceptable to assume that they are in fact i.i.d.

Newton and Keller [34] also discuss evidence of the non-i.i.d. nature of tennis points and present how one can adjust the probabilities of points won depending on "point importance" [26] and player characteristics.

2.3.3. Ranking Models

As mentioned in Section 2.1.4 official ranking tables are developed to assess the yearly performance of professional players. We have in the past argued that these ranking tables suffer from bias [9] and are unfair to poorly ranked players.

Radicchi [8] demonstrates a technique which enables the ranking of 3500 tennis

players from 1960 to 2011 in a quest to find the best player of all time. Radicchi adapted the PageRank [41] algorithm, which was originally developed to rank websites, to rank professional tennis players. He creates a network out of the meetings of the players and assigns a weight to each connection depending on how many defeats each player had from its paired player. Radicchi then calculates a prestige score for each player from those connections. The prestige score of a player is calculated as follows:

$$P_i = (1 - q) \sum_j P_j \frac{w_{ji}}{s_j^{out}} + \frac{q}{N} + \frac{1 - q}{N} \sum_j P_j \delta(s_j^{out}) \quad (2.31)$$

where:

P_i is the prestige score of player i

P_j is the prestige score of player j

w_{ji} is the weight of the directional connection from player j to player i . (i.e. the number of times player j has been defeated by player i)

s_j^{out} is the out-strength of player j , that is $s_j^{out} = \sum_i w_{ji}$

q is a damping factor where $q \in [0, 1]$

and finally $\delta()$ is a function which takes the value of 1 for zero input and 0 for all others.

Radicchi then goes on to generate the prestige ranking of a player from a single tournament as an example. Radicchi concludes by ranking all tennis players from 1960 according to their prestige score and finds Jimmy Connors to be the best player of all time.

Baker and McHale [42] present a new, more generalised, closed form of Stern's gamma comparison model [43]. They then use this model with barycentrically interpolated player strengths, which adapt over time, in a quest to discover the best tennis player since 1968. The answer they provide to the question is Roger Federer, closely followed by Bjorn Borg and Jimmy Connors. Using their generated ranking they found that it closely follows a ranking of players which is based on the number of Grand Slams won.

2.3.4. Using Rankings as Predictive Tools

Various other authors approach the problem of predicting match outcomes by making use of player rankings and figuring out methods of extracting probabilities from the rankings.

Boulier and Stekler [44] construct a probit regression model which uses the differences in player rankings to make predictions on the tennis match outcome, finding rankings as useful predictors. Clarke [45] predicts the outcomes of matches of top players by using exponential smoothing on official ATP player rankings and later on uses that approach [46] to simulate major tournaments such as Wimbledon.

Klaassen and Magnus [47] use a logit model based on the differences between the competing players' rankings to estimate an initial probability of winning a point on serve which they then adapt during the match to provide in-game analysis.

Corral and Rodriguez [48] use a different approach to investigate whether the difference in player rankings is a good predictor. They suggest three different probit models which include differences in player rankings and use them to evaluate their forecasting accuracy by comparing the predicted probabilities with actual match results. They also study the effect of ranking differences on prediction for varying player gender.

An issue with the above approaches is not the models they develop but the underlying ATP rankings which provide the data for the models. It can be argued that the official ATP rankings do not represent true player ability but rather the ability of players to compete within the rules of the ranking system itself. This affects the performance of these models in a negative way.

McHale and Morton [49] overcome this problem by using the number of games won against opponents, exponentially decayed over time (accounting for players' recent form), to calculate the players' ability to win games. They then generate rankings based on that ability and use it as a parameter to a Bradley-Terry type model [50] to calculate the probability of a player winning a game against another player. This game winning probability can be expanded using hierarchical models to the probability of winning the match.

2.3.5. Other Tennis Model Uses

Result Prediction and Sport Analysis

Hierarchical models and ranking systems are not the only way to predict tennis results. This section includes all literature which focuses on different methods for predicting tennis results as well as general analysis of various factors related to the sport of tennis.

Richardson et al. [51] analyse the effect of psychological momentum in tennis across genders and ability. They concluded that psychological momentum depends

on individual players and that gender and ability does not make a significant difference.

Bosscherr et al. [52] present an analysis of the correlation between the country of origin of the player and their success, achieving a greater understanding of the socio-economic impact on the sport.

A different type of solution to the problem of predicting tennis results, is using neural modelling, as introduced by Somboonphokkaphan et al. [53]. They use a multi-layer perceptron to predict the match result with three different techniques, each progressively better than the last one. Their results are impressive as their TimeSeries model can predict from 70% to 81% of the matches of each tournament tested. The downside of this “black-box” approach is that it contributes little to the mathematical understanding of the game of tennis and in the case it under-performs it offers no conceptual explanation as to why it would be the case.

Scarf and Shi [54] measure the importance of a match quantitatively for any sport. They do this in a similar way to the methodology Morris [26] uses measure the importance of points in tennis (i.e. by measuring the impact of winning the match has on the probability of achieving a goal). This is done using Monte Carlo simulations because of the complexity.

Djurovic et al. [55] utilise statistical tennis match data of 128 matches played on hard courts to perform factor analysis based on a component model. They identify five significant factors which account for 83.38% of the variability of matches. Those factors are the total number of break points, total number of first serve points won, the average and fastest serve speeds, the number of net approaches and win percentage of net approaches and finally unforced errors and double faults. Identifying the contributing factors to the variability of match results is useful in prediction as it can simplify models. These results are reasonable but the sample of data used to come to these conclusions is insufficient and biased (as it is only for hard courts). Ma et al. [56] also published a paper presenting a logistic regression model with 16 variables, constructed using match statistics, player characteristics and match characteristics. Their data included statistics from 9144 matches and their model explains 79.4% of the variance. Their results also confirm the importance of serving, receiving and break-points to the final outcome of matches.

Giltsdorf and Sukhatmeb [57] use Rosen’s tournament model [58], which includes tournament incentives among other parameters, in an attempt to measure the impact the prize money of a tournament has on the probability that the favourite player will win. They found that for WTA tournaments, the prize money has a pos-

itive impact on the favourite's probability of victory. The reason for this could be that more skilled players participate in the tournaments with higher incentives. As a result there is a wider gap between the seeded players and the unseeded ones and as a result there is an overall greater chance of the favourites to win matches in the tournament as a whole.

Leitner et al. [59] investigate the effect of Rafael Nadal's absence from tournaments on Roger Federer. In particular, they use bookmaker's expectations on the winner of Wimbledon 2009 and analyse how these change given Rafael Nadal's sudden withdrawal. They find that the probability of Roger Federer, Andy Murray and Tommy Haas winning the tournament increase disproportionately compared to everyone else. Although an apparently interesting discovery, there appears to be little data to support a pattern.

Malueg and Yates [60] construct an economic model and use data from equally matched players to understand and measure the effort exerted by tennis players in a best-of-three match. They find that the winner of the first set exerts greater effort in the second set because the reward is greater and thus a best-of-three set is more likely to end in 2 straight sets. This discovery goes against the idea of momentum which is a widely accepted concept and further investigation into this type of analysis would be interesting.

Vis et al. [61] apply pattern mining to tennis. They succeed in identifying sequences of strokes that occur frequently during rallies for both individual players and in general. This paper sets up a framework for further analysis into pattern detection for tennis.

Scheibehenne and Broder [62] created a study to measure the recognition of tennis players' names and used their recognition ranking results to predict match outcomes. Herzog and Hertwig present an interesting paper [63] in which they use the "wisdom of ignorant crowds" to predict outcomes of various sports, including tennis. They use crowd recognition of players to forecast the outcomes of matches and they found that ranking players by recognition can have the same predictive power as official rankings. Crowd recognition could be a powerful predictor but could also offer pitfalls for specific players, as recognition does not distinguish the reasons for players' fame or infamy.

Nevill et al. [64] and later on Holder and Nevill [65] compare world rankings with Grand Slam rankings using logistic regression to investigate whether playing at home affects the performance of players, only to find little evidence that it does. Koning [66] approaches the same problem with a probit model to quantify the

home advantage in tennis matches. He discovered that playing at home affects men's performance but does not significantly affect women's performance.

Knight and O'Donoghue [67] analyse the probability of winning break-points in Grand Slams and compare it to the probability of winning other non-break points. They find that there is a significant increase in the probability of the receiver winning a break-point when compared to other points and as a result they conclude that the probability of winning a point depends on the match score at the time.

Competitive Balance

Klaassen and Magnus [68] investigate how to reduce the dominance of serve in tennis to make matches more interesting and more competitive. Their analysis indicates that abolishing the second serve altogether will eliminate the serving advantage and will make the server and receiver more equal.

Du Bois and Heyndels [69] investigate competitive balance in women's and men's professional tennis. According to their findings, there is higher inter-seasonal as well as long term uncertainty in men's tennis suggesting that the ATP Tour is more competitive.

Corral [70] also investigates competitiveness in tennis. He composed a paper investigating the effect seeded tournaments have on the competitiveness of tennis in both men and women. He proposes a method of measuring competitiveness in tournaments based on the seed position of players and how far along they proceed in tournaments. The conclusion presented is that seeded tournaments reduce competitive balance in men but do not make a significant difference in women's tennis.

Sunde [71] explores whether heterogeneity in tennis tournaments affects the effort exerted by competing players. The heterogeneity of players is measured in terms of their difference in ATP Ranking before the match. The findings in this paper suggest that players exert greater effort when facing opponents who are closer to them in ranking. This is a reasonable conclusion and it would be interesting to investigate the underlying cause. For example, it could be because players will face other players who are closer to them in ranking more often in tournament finals (because of seeded tournaments), so perhaps the underlying cause of greater effort is that they are closer to their goal.

Halkos and Tzeremes [72] calculate an efficiency indicator which includes 9 performance indicators to evaluate the efficiency of 229 professional players over

their entire career. They find that tennis is highly competitive with 39 players appearing to be efficient.

Optimising Player Strategies

Gale [73] was one of the earliest authors to touch on using models to improve strategy. He published a simple model for the probability of winning a point by splitting the point into first and second serves. He then used that model to comment on the optimum serving strategy. George [74] also comments on the serving strategy by developing a similar conditional model on first and second serves, discovering, using real data from tournaments, that the usual strategy of strong then weak serve may not be optimal. Norman [75] uses a dynamic programming approach to address the serving strategy problem and provide the conditions of when to serve fast and when to serve slow in both serves. Pollard [76] comes to the conclusion that the risk taken during serve has a quadratic relationship with the chance of winning the point and hence players should manage service risk accordingly. Pollard et al. [77] also composed an article outlining how match statistics can be used during play by players to manage that risk.

Klaassen and Magnus [78] also touch on the issue of serving and service strategy using more recent data from Wimbledon. They found that in general the serving strategy of top players is not optimal but inefficiencies are small. They expand these results to estimate the effect service inefficiency has on the probability of players winning matches and (for Wimbledon) the monetary loss that they cause players.

O'Donoghue and Ingram [79] analyse singles events to determine the impact player sex and the surface of the court have on the top players' strategy in terms of rally length and the quantity of baseline rallies. Their results show that both parameters have a significant influence on the players' strategy. O'Donoghue [80] also mentions the significance of including those same parameters in the measurement of the importance of points.

Chiu and Chiao [81] mathematically analyse positioning of players and the defence space with the purpose of creating mathematical models to optimise player positioning before the stroke.

Since the introduction of the "Hawkeye" system in tennis, players have been able to challenge umpire calls. Pollard et al. [82] discuss how this problem opens up potential analysis for efficient use of player challenges. Using the importance

of points, the expected number points remaining in the set, and the player's probability of getting challenges right, one could develop a model to assist players in the decision of when to challenge. Nadimplali and Hasenbein [83] suggest a strategy about when players should challenge a call in tennis using a simple Markov decision process. The parameters which define the decision to challenge are: the number of challenges remaining, the confidence of the player that the call was wrong, the current score, the outcome of a successful challenge and the outcome of the point.

Equipment Improvement and Injury Prevention

Some literature deals with measuring impacts on tennis equipment and players with the aim of preventing injuries and creating new and better equipment such as rackets and shoes. Brody [84] published one of the earliest papers measuring impacts on rackets using a piezoelectric foil on the racket's handle converting force into an electric signal.

Cross [85] mathematically models the swing of a racket and the forearm as a double pendulum and proves that the speed of the racket swing is dependent primarily on the racket's moment of inertia.

Cutmore [86] develops a tennis match model which takes into account the risk of retirement of the players at any point in the game. This risk is modelled as a function of the gap between bookmakers set markets and the match market.

Automated Annotation by Video and Audio Analysis

A great deal of work has been done in automating the retrieval of statistics using video and audio analysis of tennis broadcasts. As the results of this line of research become more and more reliable, it will enable more detailed statistics to be gathered and publicised which will in turn make tennis models even more powerful.

Using court dimensions and camera geometry, Sudhir et al. [87], were able to create an algorithm that is able to detect court lines and track tennis players from video feeds. This information is then analysed and linked to high-level tennis events such as the detection of baseline shots, passing-shots, serve-and-volley and net-games.

Petkovic et al. [88] introduce another method of analysing TV broadcast videos using Hidden Markov Models and an image segmentation algorithm to recognise tennis strokes in an attempt automate retrieval of tennis statistics. Bloom and

Bradley [89] attempted to solve the same problem by tracking tennis players and recognising different strokes. This kind of research can automate tennis metadata capture and provide detailed data for statistical analysis.

Kolonias, Christmas et al. [90, 91, 92] also describe a system for automated tennis match annotation from video. Their system uses a hidden Markov model, the evolution of a point and other higher level models to detect the outcome of individual shots and annotate video with fair accuracy. It achieves this using court reconstruction, player and ball-tracking and following the grammar rules set by the restricted state machine in the background.

Hunter et al. [93, 94] analyse the audio from tennis feeds and using a Markov chain to simulate points, they are able to detect events and predict in-point sequences based on stroke sounds alone.

Jiang et al. [95] developed a system which will automatically detect and reconstruct a tennis court based on the court lines, it will then detect and track players knowing the colours of their uniform. The system also extracts the player figure completely and includes a shadow removal algorithm which opens research possibilities for stroke type detection.

Dang et al. [96] develop a robust framework of real-time video analysis for tennis player detection and tracking. This framework boasts court line detection and a player tracking system which uses an underlying tennis model. Another system designed by Connaghan et al. [97] can automatically detect the beginning of a tennis game, a change of ends by the players and a tennis serve. It does this by using player position and visual characteristics of the players to recognise them.

3. Expanding the Hierarchical Tennis Model

Our research builds on Barnett's research [98], by modelling the point itself in more detail. Modelling the point as a Markov chain allows us to combine statistics individually at a lower level than the point itself; e.g. combine the server's ace ability with the receiver's vulnerability to aces or the server's second serve win % with the receiver's second serve return win %. We then demonstrate how one can use player statistics to calculate the probability of a player winning a point on serve and how to combine player statistics to account for the returning skills of the opposing player.

The first part of this chapter describes how the match, set, game and point are modelled hierarchically as Markov Chains and how conditional probability formulae can be generated. The match chain model uses the probabilities derived from the set chain model, the set chain model uses the probabilities derived from the game and tiebreaker chain models and finally the game and tiebreaker chain models use the probabilities derived from the point chain model. The match, set and game models presented here are identical to the ones presented by Barnett et al. [98] with the exception of slight modifications on the order of serve within the tiebreaker and the order of serve within the set model.

3.1. Match Markov Chain

The match chain model is a very simple Markov Chain as shown in Figure 3.1. To improve readability of the diagram the probabilities are not shown on the lines but the layout has been designed such that any movement to a state which is higher in the diagram happens when Player A wins a set and movement to a lower state happens when Player A loses a set.

The players alternate serve with each passing game so we must also keep in mind who serves first in each set. To illustrate how this affects the probabilities

Tennis Match (Best of 5 Sets)

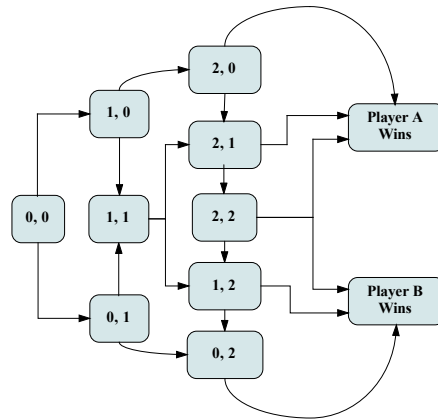


Figure 3.1.: Markov Chain of a Tennis Match

used we will explain using an example. If we assume that Player A starts to serve first then from state (0,0) we will move to state (1,0) with a probability p_A^{**} (the probability that Player A wins a set in which he serves the first game) and we will move to state (0,1) with a probability of $1 - p_A^{**}$. In the second set that is played, the receiver of the last game serves the first game. Assuming Player A has won the first set and player B received last in the first set and we are now in state (1,0), then we can move to state (1,1) with a probability of p_B^{**} (the probability that Player B wins a set in which he serves the first game) and to state (2,0) with a probability of $1 - p_B^{**}$. In the case the last set ended with a tiebreaker, then the player who received the first point of that tiebreaker will be the server of the first game in following set.

A best-of-3 sets match model can be designed in a similar way by removing the third column of states and moving directly from the second column to the winning states. The probabilities p_A^{**} and p_B^{**} can be calculated from the Set chain model which is described in the next section.

Assuming that Player A is the player who serves the first game of the match, the

following conditional probabilities hold.

$$P^m(x,y) = p_A^{**}P^m(x+1,y) + (1 - p_A^{**})P^m(x,y+1) \text{ when A serves 1}^{\text{st}} \quad (3.1)$$

$$P^m(x,y) = p_B^{**}P^m(x,y+1) + (1 - p_B^{**})P^m(x+1,y) \text{ when B serves 1}^{\text{st}} \quad (3.2)$$

Where $P^m(x,y)$ is the probability that Player A wins the match from match score x, y .

The boundary values of $P^m(x,y)$ for a best-of-5 set match are:

$$P^m(x,y) = 1 \text{ for } x = 3, y < 3$$

$$P^m(x,y) = 0 \text{ for } y = 3, x < 3$$

$$P^m(x,y) = p_A^{**} \text{ for } x = 2, y = 2$$

Similarly for a best-of-3 set match:

$$P^m(x,y) = 1 \text{ for } x = 2, y < 2$$

$$P^m(x,y) = 0 \text{ for } y = 2, x < 2$$

$$P^m(x,y) = p_A^{**} \text{ for } x = 1, y = 1$$

3.2. Set Markov Chain

There are two types of sets in tennis, the advantage set and the tiebreaker set. An advantage set ends only when some player has a score of 6 or higher and a difference of at least 2 games from his opponent. The tiebreaker set is similar to an advantage set with the difference that the game at 6-6 is a tiebreaker game which settles the set. A tiebreaker set is shown in Figure 3.2.

In the same way as in the match model we will move to a higher state if Player A wins the game and to a lower state if Player A loses the game. Assuming that Player A is the player who serves in the first game, then we will use p_A^* when even number of games have been played and p_B^* when odd number of games have been played. In the case of a tiebreaker we use p_A^{T*} which is the probability that Player A wins a tiebreaker game in which he starts serving first. p_A^* and p_B^* are the probabilities that Player A wins a game as a server and Player B wins a game as a server respectively. p_A^* and p_B^* are calculated using the game chain model whereas p_A^{T*} can be calculated using the tiebreaker chain model both described in

Tennis Set - Tiebreaker

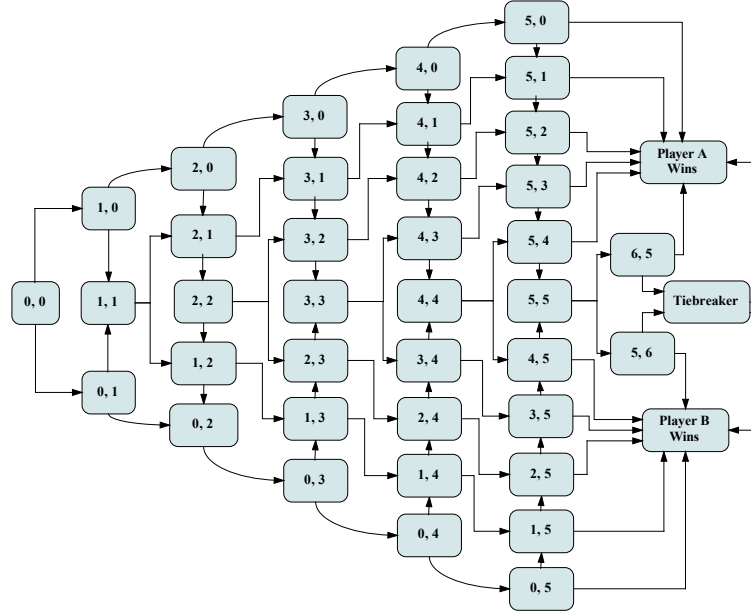


Figure 3.2.: Markov Chain of a Tennis Tiebreaker Set (the tiebreaker game is defined in Figure 3.4)

subsequent sections.

Assuming that Player A is the first player to serve in the set then the equations for the probability of winning a set are:

$$P^S(x, y) = p_A^* P^S(x + 1, y) + (1 - p_A^*) P^S(x, y + 1) \text{ for even } (x + y) \quad (3.3)$$

$$P^S(x, y) = p_B^* P^S(x, y + 1) + (1 - p_B^*) P^S(x + 1, y) \text{ for odd } (x + y) \quad (3.4)$$

The boundary values for $P^S(x, y)$ in the case of a tiebreaker set are:

$$P^S(x, y) = 1 \text{ if } x \geq 6, x - y \geq 2$$

$$P^S(x, y) = 0 \text{ if } y \geq 6, y - x \geq 2$$

$$P^S(x, y) = p_A^{T*} \text{ if } x = 6, y = 6$$

The boundary values in the case of an advantage set are:

$$P^S(x, y) = 1 \text{ if } x \geq 6, x - y \geq 2$$

$$P^S(x, y) = 0 \text{ if } y \geq 6, y - x \geq 2$$

$$P^S(x, y) = \frac{p_A^*(1 - p_B^*)}{p_A^*(1 - p_B^*) + (1 - p_A^*)p_B^*} \text{ if } x = 5, y = 5$$

The same equations can apply when Player B is serving first in the set by simply substituting p_A^* with p_B^* and p_B^* with p_A^* .

An interesting state to further investigate, in the case of the advantage set, is the 5-5 state. The probability of winning the set from this state is the same as the probability of winning the set from any tied score greater than 5. In order to calculate this probability, one only needs to consider the problem as a random walk with two states. Player A can win the set by winning two games in a row with a probability of $(P_A^*)(1 - P_B^*)$. Keeping in mind that the probability of going back to a tie is $P_A^*P_B^* + (1 - P_A^*)(1 - P_B^*)$, one can find the probability of Player A winning the set from this state is $P^S(5, 5) = (P_A^*)(1 - P_B^*) + (P_A^*P_B^* + (1 - P_A^*)(1 - P_B^*))P^S(5, 5)$. Solving for $P^S(5, 5)$ gives the boundary value described above [99].

3.3. Game Markov Chain

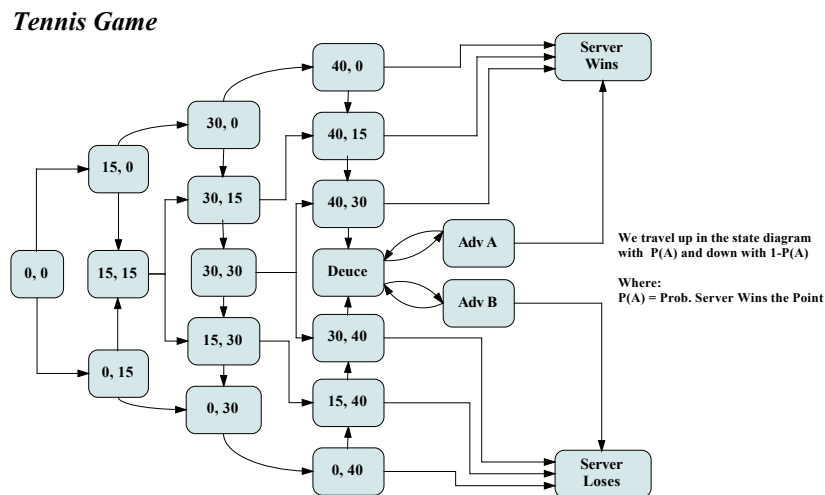


Figure 3.3.: Markov Chain of a Tennis Game

Figure 3.3 shows the Markov chain of a tennis game. Similarly to the previous chains we move to a higher state in the diagram when the server wins the point and to a lower one when the server loses a point. In a game the server is always the

person serving first, therefore we only need the probability of the server winning a point to evaluate the game probability. We define p_A and p_B to be the probabilities of Player A winning a point while serving and Player B winning a point while serving respectively. p_A and p_B can be calculated in a variety of ways depending on how the point itself is modelled.

$P^G(x,y)$ denotes the probability of the server winning the game from score x, y . For simplification, we will use the value 1 for x and y when the score is 15, the value 2 for a score of 30 and the value 3 for a score of 40. Assuming Player A is serving for the duration of a game $P^G(x,y)$ can be calculated as follows:

$$P^G(x,y) = p_A P^G(x+1,y) + (1-p_A) P^G(x,y+1) \quad (3.5)$$

The boundary values are:

$$P^G(x,y) = 1 \text{ when } x = 4, x - y \geq 2$$

$$P^G(x,y) = 0 \text{ when } y = 4, y - x \geq 2$$

$$P^G(x,y) = \frac{p_A^2}{p_A^2 + (1-p_A)^2} \text{ when } x = 3, y = 3$$

These equations are also valid for Player B serving first by substituting p_A with p_B .

The boundary value for deuce is approached similarly to the method used to calculate the $P^S(5,5)$ boundary value in the previous section. Using the method proposed by Haigh [99] one can *jump* directly from Deuce to either victory or loss, skipping the advantage states. Using this approach then the probability to win directly from deuce is p_A^2 , to lose directly from deuce is $(1-p_A)^2$ and to go from deuce back to deuce is $2 \times p_A(1-p_A)$. This means that the probability $P^G(3,3) = p_A^2 + 2p_A(1-p_A)P^G(3,3)$. Solving for $P^G(3,3)$ gives us the result we have above.

3.4. Tiebreaker Markov Chain

Like the previous chain models this one has also been designed such that whenever Player A wins a point we follow the model upwards and whenever Player B wins a point we follow the model downwards. Tiebreakers have a complication however, as players serve alternatively every two points. In fact assuming Player A serves first, Player B will serve the following two points, then Player A will start serving and they will alternate serving every two points. (i.e. the sequence of serving will be ABBAABBAA...BBAA). This complicates the equations for $P^T(x,y)$ which

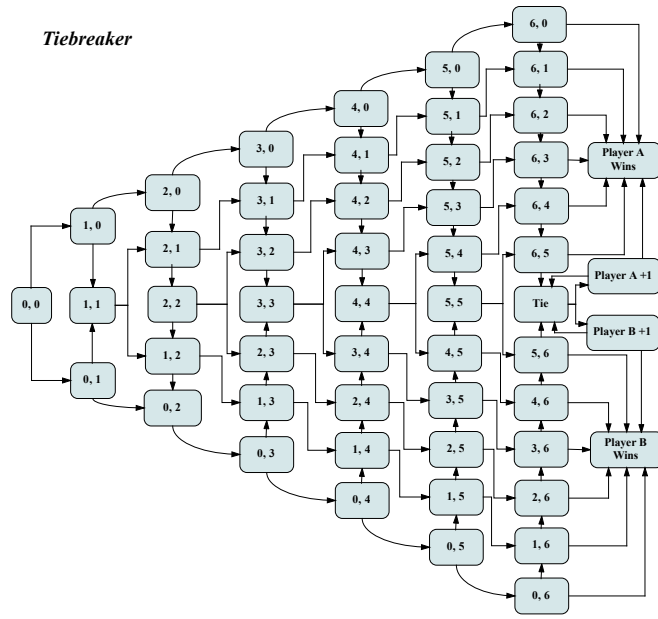


Figure 3.4.: Markov Chain of a Tiebreaker Game

denotes the conditional probability of the player who serves first in a tiebreaker to win the tiebreaker from score x, y .

Assuming Player A serves first, the formula for $P^T(x, y)$ is as follows:

$$P^T(x, y) = p_A P^T(x+1, y) + (1-p_A) P^T(x, y+1) \text{ for } 2 \leq (x+y+3) \bmod 4 \leq 3 \quad (3.6)$$

$$P^T(x, y) = p_B P^T(x, y+1) + (1-p_B) P^T(x+1, y) \text{ for } 0 \leq (x+y+3) \bmod 4 \leq 1 \quad (3.7)$$

The boundaries are:

$$P^T(x, y) = 1 \text{ when } x = 7, x - y \geq 2$$

$$P^T(x, y) = 0 \text{ when } y = 7, y - x \geq 2$$

$$P^T(x, y) = \frac{p_A(1-p_B)}{p_A(1-p_B) + (1-p_A)p_B} \text{ when } x = 6, y = 6$$

These equations are also valid for Player B serving first by substituting p_A with p_B and p_B with p_A . The approach used to calculate the boundary $P^T(6, 6)$ is the same as the one used to calculate the boundary $P^S(5, 5)$ of the advantage set. Also worth noting is that $P^T(5, 5) = P^T(6, 6)$ therefore one could move the boundary value to the score point 5-5 and simplify the model further. We chose to leave it at the score point 6-6 for greater clarity.

3.5. Point Markov Chain

Point on Serve

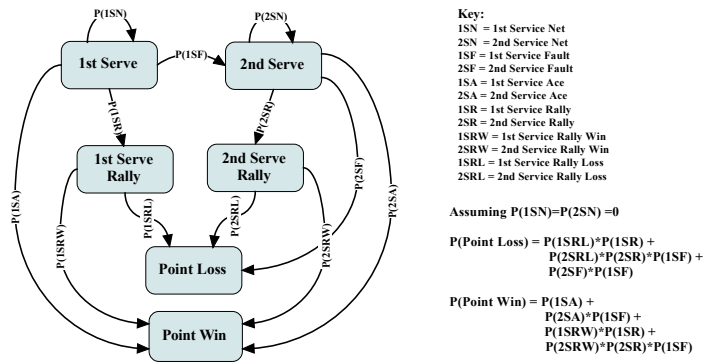


Figure 3.5.: Markov Chain of a Tennis Point

The single player point model is our approach to calculating the probabilities p_A and p_B as used in the previous models and this is where our research expands existing literature. From the perspective of the server, a point in tennis can be broken down into a set of states. A tennis point starts with the first serve. From the first serve, the server can either repeat first serve by getting a net (let) call, proceed into a second serve by a fault, proceed into a rally by successfully placing the serve and be returned or directly win the point with an ace. Similarly, while in the second serve the server can either repeat second serve by getting a net (let) call, lose the point by another fault, go into a rally by successfully placing the serve or win the point with an ace. Following a rally, the server can either win or lose the point. This translates into the Markov chain shown in Figure 3.5. For the duration of this chapter we will use the following abbreviations: SN for Service Net, SF for Service Fault, SA for Service Ace, SR for Service Rally, SRW and SRL for Service Rally Win and Service Rally Loss respectively. A number before an abbreviation denotes whether the event has occurred during a first or second service. e.g. 1SF denotes 1st Service Fault.

From the Markov chain we can easily compose equations for the probability of the server winning the point. Due to lack of statistical data on net calls and because they occur rarely, we assume that $P(1SN) = 0$ and $P(2SN) = 0$. This results in the simplified Equations 3.8 and 3.9 which are the probabilities of the server winning and losing the point respectively. It must also hold that $P(\text{PointWin}) = 1 - P(\text{PointLoss})$ as they are mutually exclusive events.

$$P(\text{PointWin}) = P(1SA) + P(1SR)P(1SRW) + P(1SF)P(2SA) + P(1SF)P(2SR)P(2SRW) \quad (3.8)$$

$$P(\text{PointLoss}) = P(1SR)P(1SRL) + P(1SF)P(2SF) + P(1SF)P(2SR)P(2SRL) \quad (3.9)$$

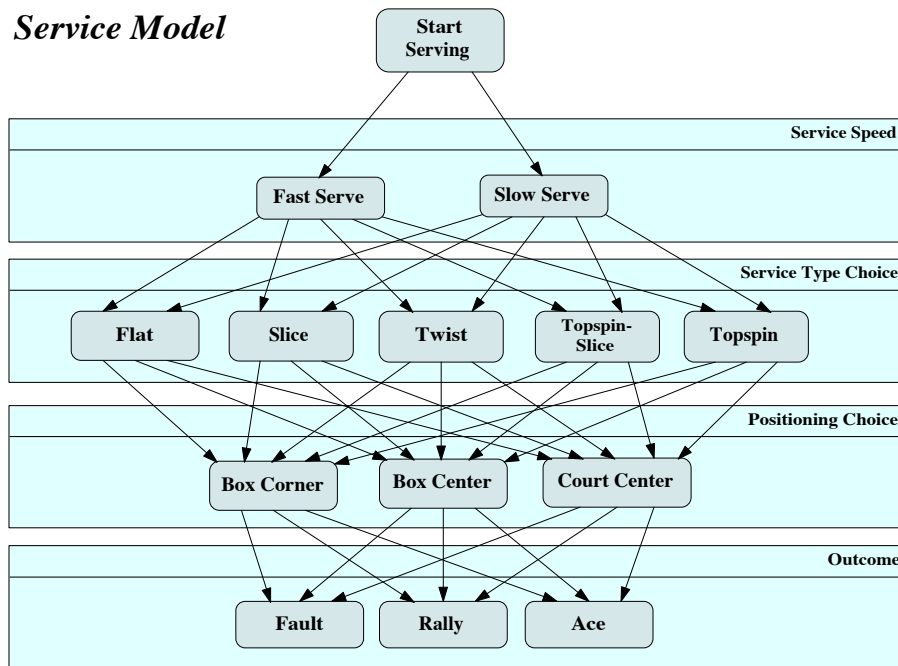
The probability $P(\text{PointWin})$ can be embedded in higher level Markov chains for games, sets and matches as p_A or p_B depending on which player is the server of the point. Equation 3.8 however is of little use without knowing how to estimate the probabilities that it takes as input. We can estimate these input probabilities in a number of ways, whether it is a method that only uses the average long-term statistics of the server against the average player he has faced, one which combines the long-term average statistics of both server and receiver or even an approach which combines the statistics of four players for a doubles match. Whichever approach is adopted, the quality of the results will depend on the availability of accurate statistical values.

3.6. Service and Rally Markov Chains

Although at this time, the statistical data required for even deeper modelling of a point is not available to the public, it is very possible that advances in computer vision like the work of Kolonias et al. [100] and the wider spread of the Hawkeye system in tennis courts, will provide a means to collect them in the future. In such a case models which analyses stroke placement and type for both services and rallies will become very useful as they will model the strengths and weaknesses of players in particular strokes.

A model similar to the one in Figure 3.6 can be used to estimate the probabilities $P(1SR)$, $P(1SA)$, $P(1SF)$, $P(2SR)$, $P(2SA)$ and $P(2SF)$ (given adequate data on the

Service Model



Probabilities of Choices can be calculated using statistical data from previous matches for each player. Probabilities of outcomes may also be calculated using statistical data of both server and receiver.

Figure 3.6.: Markov Chain of a Serve

speed of each serve, the type of serve, the service positional choices of players and the Win/Loss percentages of each combination of the above). The model firstly splits into two different power levels (fast and slow serves), then into possible types of service shot (a flat shot, a slice, a twist, a topspin slice and finally a topspin). Statistics for the choice of service can possibly be collected by tracking the motion of the player’s racket while hitting the ball or alternatively by following the ball and depending on the curve it follows assign a shot type. The shot is then further categorised in terms of where it lands on the court. By splitting the service box to three equal vertical sections we can categorise the shots’ target as box corner, box center and court center. Finally the service shot has three outcomes. It can either be an ace, a fault or it can be returned resulting in a rally. This allows for detailed analysis of the player’s serving habits and strengths in the first serve and in the second serve. This can also be further combined with the opponent’s returning strengths for each particular type of serving shots to estimate the probabilities of the outcomes of service against particular opponents.

Expanding this type of analysis to a rally, a model similar to the one in Fig-

Figure 3.7 can be used to estimate the probabilities $P(1SRW)$, $P(1SRL)$, $P(2SRW)$ and $P(2SRL)$ given adequate data on stroke choices of players and positions of stroke as well as return win percentages of each type of stroke within a rally. The shot has been divided into six different strokes: forehand, backhand, volley, half-volley, drop-shot and lob. Each of these strokes can have one of three court targets – it can either be on the left court, the center or the right court. Gathering statistics for a player’s winning and returning abilities in every one of these combinations of shot choice and shot placement, can allow one calculate the probability of a player winning a rally against a particular opponent.

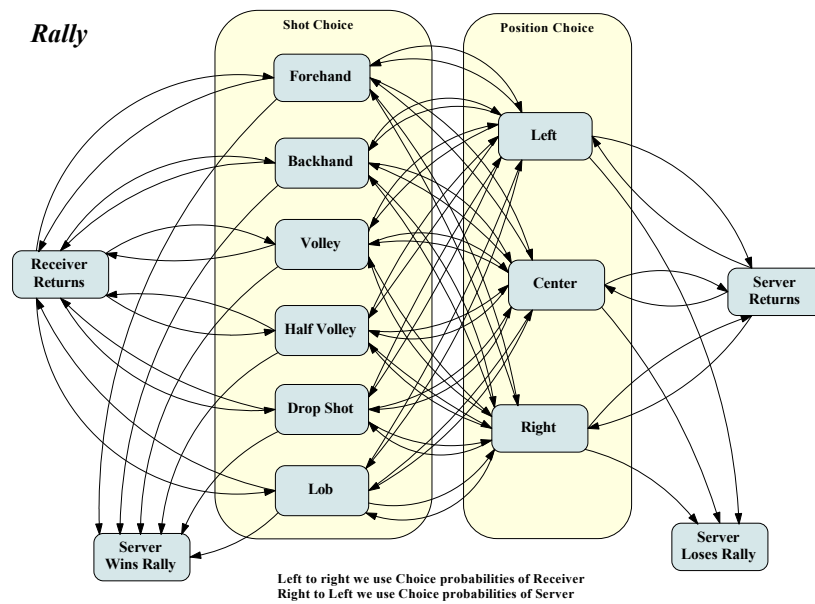


Figure 3.7.: Markov Chain of a Rally

3.7. Forecasting the Outcome of a Match

3.7.1. Collecting the Data

At present, detailed statistical data about individual strokes are not available to the public. Basic statistical information regarding player performance in individual matches, however, is available to the public through a number of online services.



Match	
Tournament	
US Open	
Round	
R16	
Time	
155 minutes	
Winner	
Andy Murray	
Players	
Jo-Wilfried Tsonga	Andy Murray
Nationality	
	
France	Great Britain
Statistics on Service	
Aces	
5	11
Double Faults	
6	3
1st Serve	
59% (67/113)	49% (49/99)
1st Serve Points Won	
64% (43/67)	83% (41/49)
2nd Serve Points Won	
52% (24/46)	58% (29/50)
Break Points Saved	
61% (8/13)	60% (3/5)
Service Games Played	
17	17
Statistics on Return	
1st Serve Return Points Won	
16% (8/49)	35% (24/67)
2nd Serve Return Points Won	
42% (21/50)	47% (22/46)
Break Points Converted	
40% (2/5)	38% (5/13)
Return Games Played	
17	17

Figure 3.8.: A screenshot of the statistics of a single match between Jo-Wilfried Tsonga and Andy Murray as presented by the ATP World Tour website.

The ATP World Tour website is a useful source of statistics as they provide individual match statistics for almost all the matches played in ATP 250, ATP 500, Masters and Grand Slam tournaments since 1999. The statistics they provide are number of aces and double faults, first service % when serving, first service point win % when serving, second service point win % when serving, first service point win % when returning and second service point win % when returning. Also their Live Scores applet will display the score, aces, double faults, 1st serve %, 1st

and 2nd serve point win % for each player in a live match or in any match of the active tournament. Figure 3.8 shows a screenshot of a single match as provided by the ATP World Tour website. Unfortunately these statistics are not presented in a useful collective way.

TennisInsight.com on the other hand readily provide collective statistics on the Top 200 ATP players which can be filtered by date and surface type. Additionally, their statistics include a measure for Aces per game and Double Faults (DF) per game as well as Opponent Aces per game and Opponent DF per game which are not provided by the ATP World Tour website. Figure 3.9 demonstrates the interface used to retrieve these statistics from the TennisInsight.com website.

Jo-Wilfried Tsonga	All Surfaces	ATP Tour
Match Statistics		
68% (41-19)	Match W/L	50% (2626-2626)
62% (96-58)	Set W/L	50% (6756-6756)
53% (837-728)	Gm W/L	50% (66202-66202)
51.6%	Pts W/L	50.0%
68% (21-10)	TB W/L	50% (1220-1220)
0.20	TBs per Set	0.18
Service Statistics		
0.78	Aces per Gm	0.51
0.18	DFs per Gm	0.23
61.8%	1st Serve %	61.3%
76.8%	1st Serve W%	71.9%
54.3%	2nd Serve W%	50.8%
68.2%	Service Pts W%	63.7%
Break Pts - Serve		
0.28	BPs Saved per Gm	0.32
0.40	BPs Faced per Gm	0.53
69.4%	BP Save %	60.9%
87.7%	Service Hold %	79.4%
Return Statistics		
0.43	Opp. Aces per Gm	0.51
0.18	Opp. DFs per Gm	0.23
61.9%	Opp. 1st Serve %	61.3%
28.7%	1st Return W%	28.1%
45.6%	2nd Return W%	49.2%
35.1%	Return Pts W%	36.3%
Break Pts - Return		
0.18	BPs Won per Gm	0.21
0.48	BP Chances per Gm	0.53
37.5%	Break Pt W%	39.1%
81.9%	Opp Hold %	79.4%

Includes matches up to Sep 15 2014

Duration:

Tour:

Surface:

Draw:

Figure 3.9.: A screenshot of the collective statistics of Jo-Wilfried Tsonga as presented by the TennisInsight.com website.

We have built tools which retrieve the statistics from TennisInsight.com for any requested player. We have also created a tool which constructs and maintains

a database of the statistics of individual matches which can be used locally to generate player statistics on demand. The database currently holds over 35000 individual ATP matches. This allows us to perform extensive analysis and tests on our models.

The following variables are calculated from the publicly available statistical data and are used in the experiments to follow.

$$a = \frac{\text{Total Aces}}{\text{Total Serves}} = \text{Ace probability} \quad (3.10)$$

$$b = \text{1st Serve \%} \quad (3.11)$$

$$c = \text{1st Serve Win \%} \quad (3.12)$$

$$d = \frac{\text{Total Double Faults}}{\text{Total Serves}} = \text{Double fault probability} \quad (3.13)$$

$$e = \text{2nd Serve Win \%} \quad (3.14)$$

$$f = \frac{\text{Total Opp. Aces}}{\text{Total Returns}} = \text{Opp. Ace Probability} \quad (3.15)$$

$$g = \text{Opp. 1st Serve \%} \quad (3.16)$$

$$h = \text{1st Serve Return Win \%} \quad (3.17)$$

$$i = \text{2nd Serve Return Win \%} \quad (3.18)$$

$$j = \frac{\text{Total Opp. Double Faults}}{\text{Total Returns}} = \text{Opp. Double Fault probability} \quad (3.19)$$

3.7.2. A Closer Look at the Data

Our detailed database of matches allows us to study the distributions of the details of match statistics. In this section, we will use an approach similar to the one presented by Newton and Aslam [36] and we will model individual statistical attributes as normally distributed variables commenting on the variance of the distributions as a measure of the stability of players.

Throughout this section we will work with matches played during the two year period starting on the 1st of January 2012 to the 31st of December 2013. Firstly, we will discuss the distributions of statistics of Novak Djokovic and Roger Federer, two players who are considered extremely stable in their performance. We will then analyse the distributions of statistics of players like Marcos Baghdatis and Gael Monfils whose performance is known to fluctuate. Finally we will consider John Isner, known for his strong first serve.

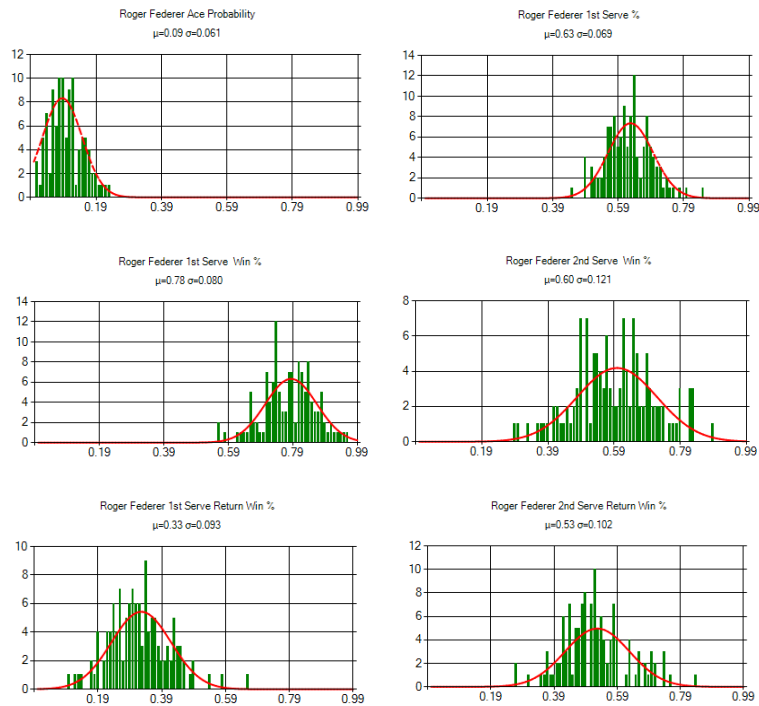


Figure 3.10.: Distributions of Roger Federer’s statistics of 127 matches played in a two year period.

Figures 3.10 and 3.11 show the frequency of occurrence of different statistics of Federer and Djokovic, two players who at the end of 2013 were ranked in positions 6 and 2 of the official ATP Rankings respectively. The Normal distribution appears to approximate these distributions well. For every statistic, the mean and variance is calculated and used to plot a Normal distribution over the bar charts to demonstrate this. The smaller the variance the more consistent the player is at the aspect of the game represented by the statistic. For example we can see in Figure 3.10 that Federer can hit an Ace with a mean probability of 0.09 and standard deviation of 0.061. Djokovic is slightly more stable at hitting aces but not necessarily better. As it is evident from Figure 3.11 Djokovic has a mean probability of 0.08 of hitting aces with a standard deviation 0.056. Federer also appears to be slightly more consistent and better at winning first and second serve points although a true comparison of these statistics should be done with significance testing. As the purpose of this section is to demonstrate the amount of information we can get when we examine the distribution of statistics we will not proceed along that path.

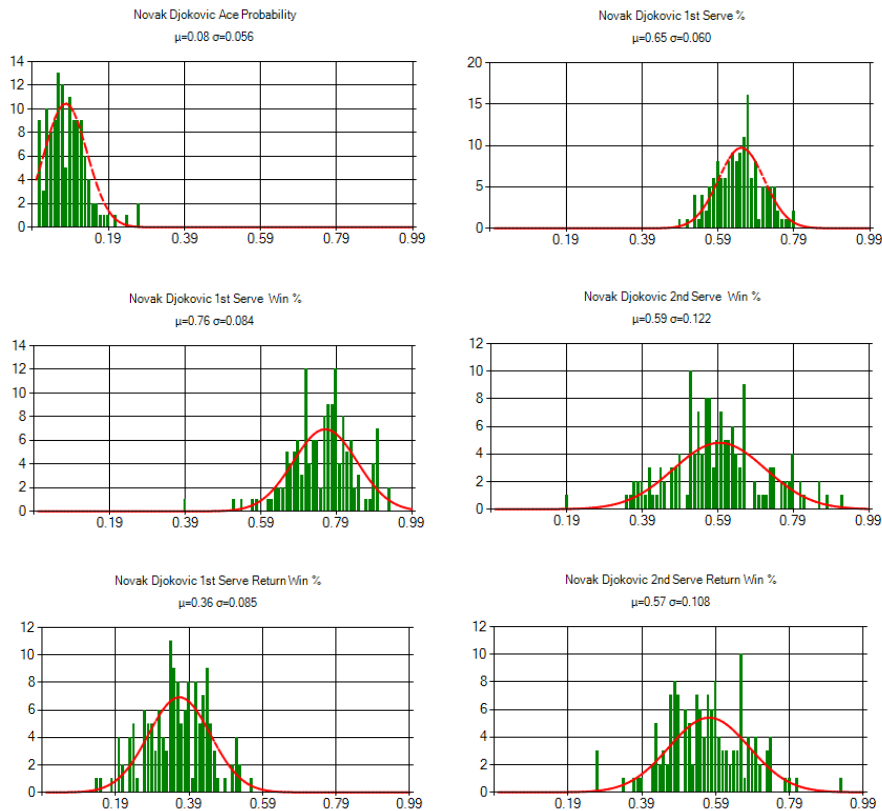


Figure 3.11.: Distributions of Novak Djokovic's statistics of 147 matches played in a two year period.

To demonstrate the notion of how the variance of the distribution represents the instability of the players we chose two players who are notorious for their instability. Firstly, Gael Monfils who has a number of victories against the likes of Roger Federer and Rafael Nadal, both No.1 players at some point in their career, and at the same time has suffered losses from Lukasz Kubot (ranked 42 at the time). Figure 3.12 demonstrates this instability with the wide Normal curves in all statistics. Comparing the Monfils' first serve percentage with Djokovic's we can see that in the latter case the standard deviation is 0.06 whereas in the first case it is 0.088 making the Normal curve wider. This is the case with all the distributions of Monfils' statistics. We can also use the variance of these statistics as a measure of how predictable the performance of a player will be during a match.

Another example of a player with erratic performance is Marcos Baghdatis, who

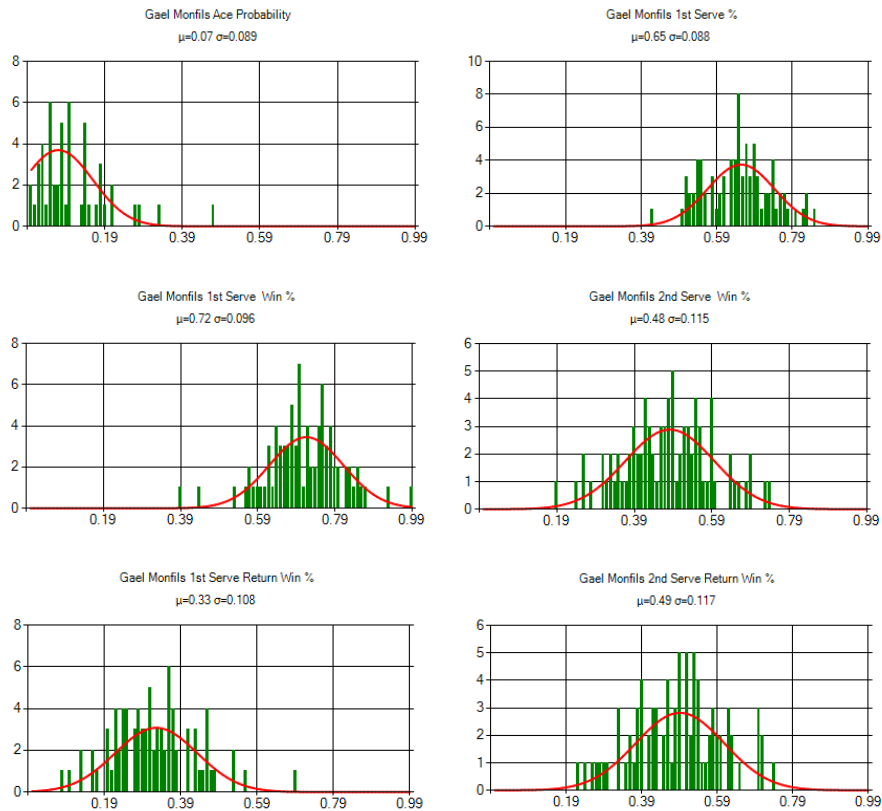


Figure 3.12.: Distributions of Gael Monfils' statistics of 83 matches played in a two year period.

at the very beginning of his professional career in the ATP reached the final of the Australian Open and also has victories against Rafael Nadal and Roger Federer and loses against low ranked players. Figure 3.13 seems to agree with this as in all statistics except the first serve percentage we see high variance.

To demonstrate the advantage of disaggregation (using all statistics) as opposed to just using a generalised statistic of Service Point Win Percentage and Return Point Win Percentage – as most models in literature suggest – we will include a special player. John Isner has a very powerful weapon in his arsenal – his first serve. Being a very tall player he is able to hit angles and speeds which are very uncommon during serve and as a result he serves a lot of Aces and has a high probability of winning first serve points. This can clearly be seen from the distributions of his statistics in Figure 3.14. The ace probability has a mean of 0.14 (compare

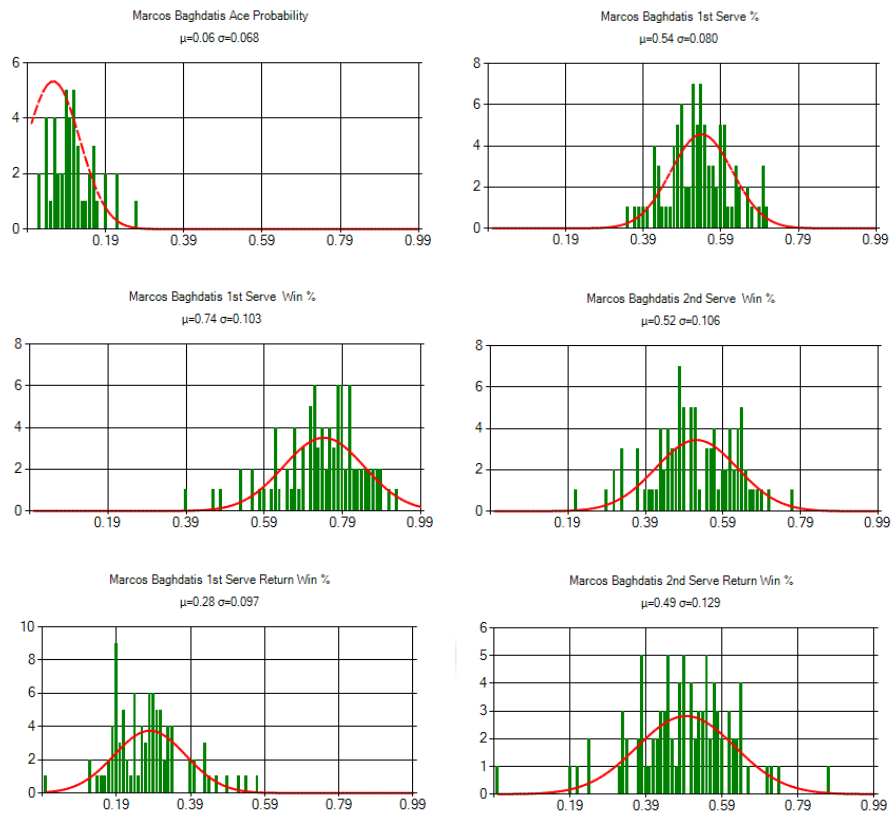


Figure 3.13.: Distributions of Marcos Baghdatis' statistics of 91 matches played in a two year period.

that to Federer which is 0.09) and a relatively high variance. Also, Isner is very consistent with achieving and winning first serves but his second serve winning capability and returns of serve seem very unstable and lower than the top players'. Being able to compare statistics at this level, has the added advantage that individual attributes of players can be compared with returning attributes of their opponents. For example, if John Isner is pitted against another very strong server but poor receiver, one can expect very few breaks of serve. Referring back to the match of John Isner vs. Nicolas Mahut match during the 2010 Wimbledon which ended with a final set score of 70-68 as just such an example. Simply using the probabilities of winning points on serve would not include such information in the model. Another scenario to consider is the case of a good returner being the opponent of Isner. It can be expected, in such a case, that the good returner will have

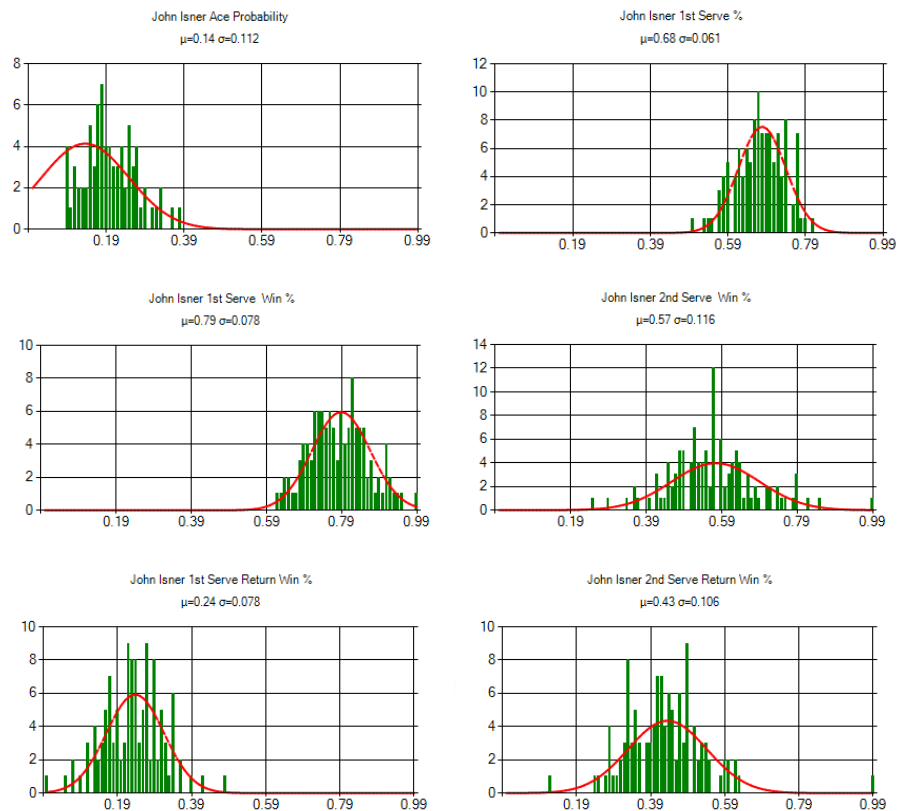


Figure 3.14.: Distributions of John Isner's statistics of 115 matches played in a two year period.

Isner's main weapon effectively neutralised. This should affect the probabilities of him winning such a match.

3.7.3. Estimating the Probability of Winning Service Points

The missing piece of the puzzle in estimating the probability of a tennis player winning a point while serving is the connection between the statistics available and the point model provided in Section 3.5. This section suggests various approaches that can be used to estimate the probabilities required by the point model, given data available to us.

Using ‘Uncombined’ Average Historical Data

A simple way of calculating the probability of a player winning a service point is to use the player’s average statistics, as described in Section 3.7.1, over a set of matches. From Equation 3.8 an estimation of, $P(1SA)$, $P(1SR)$, $P(1SRW)$, $P(1SF)$, $P(2SA)$, $P(2SR)$ and $P(2SRW)$, is required in order to find the probability of winning a point while serving. As demonstrated in Equations 3.20 through to 3.27, these probabilities can be derived from the available statistics. Since our data does not clarify whether an ace has occurred in the first or second serve, we are forced to make the assumption that all aces occur during the first serve. This is not an unreasonable assumption as a second serve is usually more cautiously struck.

$$P(1SA) = a \quad (3.20)$$

$$P(1SR) = b - a \quad (3.21)$$

$$P(1SRW) = \frac{bc - a}{b - a} \quad (3.22)$$

$$P(1SF) = 1 - b \quad (3.23)$$

$$P(2SA) = 0 \quad (3.24)$$

$$P(2SF) = d \quad (3.25)$$

$$P(2SR) = 1 - d \quad (3.26)$$

$$P(2SRW) = e \quad (3.27)$$

Variables a , b , c , d and e are as shown in Equations 3.10 through to 3.19. An interesting equation to focus on is 3.22 which may require further clarification. As the available statistics only provide us with the first service win percentage (which includes aces), it does not reflect the probability of winning a first service rally (given we are already in a rally). In order to calculate this probability we need to mathematically arrive to $P(1SRW) = P(\text{Win} \mid 1SR)$. Therefore, since $P(\text{Win} \mid 1SR) = \frac{P(\text{Win} \cap 1SR)}{P(1SR)}$ and since we know $P(1SR) = b - a$ the only unknown is $P(\text{Win} \cap 1SR)$. We also know that all points won in the first serve are either aces or first service rally wins, which means that $P(\text{Win} \cap \text{First Serve}) = P(\text{Win} \cap 1SA) + P(\text{Win} \cap 1SR) = P(\text{Win} \mid \text{First Serve})P(\text{First Serve}) = cb$. Also knowing that $P(\text{Win} \mid 1SA) = 1$, then $P(\text{Win} \cap 1SA) = P(1SA) = a$. There-

fore, $P(\text{Win} \cap 1SR) = bc - a$ and hence $P(\text{Win} | 1SR) = \frac{bc-a}{b-a}$.

The data used to generate these statistical probabilities greatly affect the resulting probability of winning a point on serve. The data can be filtered to include subsets of the available matches to controllably affect the outcome. Filtering available matches according to the period of time they were played in, or according to surface type they were played on can influence how well certain aspects of the players are represented in the average statistics retrieved and it warrants extensive research. Taking it even further, one can take weighted averages of statistics according to predefined parameters such as how recently the match was played or how many sets were played or even the ranking of the opponent. For the production of the results presented in subsequent sections, we used averaged data over various period lengths and also tested filtering the data according to their surface.

Combining Historical Data – ‘Combined’ Model

The ‘*uncombined*’ method just discussed has a significant drawback since it models how the server would play against the average player he has faced, rather than the specific opponent being modelled. In order to correct this and improve our model further, the opponent’s returning capabilities need to be included somehow in our probabilities.

Barnett [33] estimates the overall probability of a player winning a service point over the average player and then compares this with the average ATP player and adjusts for the opponent’s probability of winning a point while returning in comparison to the average player. We use a similar approach to estimate $P(1SA)$, $P(1SR)$, $P(1SRW)$, $P(1SF)$, $P(2SA)$, $P(2SR)$ and $P(2SRW)$ as used in Equation 3.8. This allows us to compare individual strengths of the two opponents and thus model the point more accurately. We will demonstrate the concept we used to estimate these probabilities by showing how to estimate the combined probability of hitting an ace, $P(1SA)$.

Let $P(1SA_A)$ be the probability of player A serving an ace against the average opponent and $P(1OSA_B)$ be the probability that player B will receive an ace from the average opponent. Also let $P(1SA_{av})$ be the probability an average player serves and receives an ace (by definition must be equal). By these terms $P(1SA_A) - P(1SA_{av})$ defines how much more probable Player A is to serve an ace than the average player and similarly $P(1OSA_B) - P(1SA_{av})$ defines how much more probable player B is to receive an ace from the average player. Adding these two differences

to the average ace probability approximates the probability of Player A serving an ace against Player B as shown below:

$$P(ISA) = P(ISA_{av}) + ((P(ISA_A) - P(ISA_{av})) + (P(IOSA_B) - P(ISA_{av}))) \quad (3.28)$$

which simplifies to:

$$P(ISA) = P(ISA_A) + P(IOSA_B) - P(ISA_{av}) \quad (3.29)$$

where $P(ISA_A)$, $P(IOSA_B)$ and $P(ISA_{av})$ can be calculated using the available statistics. Using a similar method, the rest of the combined probabilities used in our model can be retrieved from the combined statistics of the two opponents. Equations 3.30 through to 3.34 define our new combined statistic variables marked with a subscripted c .

$$a_c = a_A + f_B - a_{av} \quad (3.30)$$

$$b_c = b_A + g_B - b_{av} \quad (3.31)$$

$$c_c = 1 + c_A - h_B - c_{av} \quad (3.32)$$

$$d_c = d_A + j_B - d_{av} \quad (3.33)$$

$$e_c = 1 + e_A - i_B - e_{av} \quad (3.34)$$

Equations 3.35 through to 3.42 show how these combined statistics can be used in our tennis point model.

$$P(ISA) = a_c \quad (3.35)$$

$$P(ISR) = b_c - a_c \quad (3.36)$$

$$P(ISRW) = \frac{b_c c_c - a_c}{b_c - a_c} \quad (3.37)$$

$$P(ISF) = 1 - b_c \quad (3.38)$$

$$P(2SA) = 0 \quad (3.39)$$

$$P(2SF) = d_c \quad (3.40)$$

$$P(2SR) = 1 - d_c \quad (3.41)$$

$$P(2SRW) = e_c \quad (3.42)$$

3.7.4. Combining Historical Data for Doubles

Modelling the outcome of a doubles match is considerably more complicated than for a singles match for several reasons. Firstly, there are different variations of rules for doubles matches which depend on the tournament. Additionally, the serving and receiving order of the players has to be considered in the game, set and match Markov chains. The serving team chooses who will be the player who serves in the first game of the set and they then alternate turns every new game the team is serving for the entire set. Similarly the receiving team chooses who will receive in the first game and must alternate with receiving games within the set. The player who receives in the current game will be the player who will serve in the next game. This section only touches upon the point model and for this purpose we will suppose that Player AA is the server of the point, Player AB is the server's partner, Player BA is the receiver of the service and Player BB is the partner of the receiver.

Traditionally doubles tournaments and singles tournaments had distinct sets of players. As a more recent trend however, top singles tennis players are joining forces and are forming teams for doubles tournaments. The media has raised the question whether traditional teams like the Bryan brothers are better than newly formed teams made up of the top singles players. These newly formed teams are causing problems when modelling because of the lack of data of playing together as a team. Doubles modelling so far has treated a doubles team similarly to a singles player and not as two distinct players. For traditional teams on the other hand there are no statistics available for individual players. This section demonstrates how to adapt the probabilities in Equation 3.8 developed earlier to estimate the probability of winning a doubles point for both traditional doubles teams and newly formed ones by combining player statistics.

Of course a big part of the effectiveness of doubles teams is the teamwork of the players and how well they cooperate. This section does not touch upon this and only uses individual performance of players in the model.

Modelling Newly Formed Teams

A point in a doubles match begins with Player AA serving, Player BA receiving and then proceeding into a rally where both players of both teams can strike the ball. From this we can conclude that in order to estimate the probability of winning the point on serve we need to combine the abilities of our players. We need to

combine Player AA's serving ability with Player BA's receiving ability and then also combine Player AA's and AB's ability to win points within a serving rally with Player BA's and BB's ability to win points within a receiving rally.

With reference to Equation 3.8 we can approximate our probabilities in a way such that it accounts for this. In order to do this we need to make the assumption that when the ball is in-play, each player within a team has a probability of 0.5 to be the one that returns the ball to the opponent. By this assumption the following Equations 3.43 to 3.47 describe the new combined probabilities which can be used with Equations 3.35 to 3.42.

$$a_c = a_{AA} + f_{BA} - a_{av} \quad (3.43)$$

$$b_c = b_{AA} + g_{BA} - b_{av} \quad (3.44)$$

$$c_c = 1 + 0.5(c_{AA} + c_{AB}) - 0.5(h_{BA} + h_{BB}) - c_{av} \quad (3.45)$$

$$d_c = d_{AA} + i_{BA} - d_{av} \quad (3.46)$$

$$e_c = 1 + 0.5(e_{AA} + e_{AB}) - 0.5(i_{BA} + i_{BB}) - e_{av} \quad (3.47)$$

Modelling Traditional Doubles Teams

For traditional doubles teams there is a wealth of statistics of the players playing together as a team. Unfortunately, individual player performances are not available but it is usually assumed that the team statistics reflect the average of the abilities of the two players. Taking this into account, we can substitute the individual player statistics of both team players with the statistics of the team and thus estimate our probabilities as mentioned in Equations 3.43 to 3.47 above. For example, if team A is a traditional team and team B is a newly formed team, players AA and AB will use identical statistical data which match the statistical data of team A and players BA and BB will use their individual statistical data from the singles tournaments.

3.8. Selecting Historical Data

The quality of historical data used to estimate the parameters of a tennis model can greatly impact the resulting predictions. For tennis in particular we can identify two important factors which affect the quality of the data – age and match surface. The parameters of the models discussed in this chapter are estimated using

historical statistics averaged over a set of matches. In this section we discuss how choosing different subsets of the available data can impact the accuracy of model predictions.

3.8.1. Age of Match Played

Player performance varies over time. This is a fact as players can improve their stamina and skills as they practise more or see their abilities decay as they get older. To take this into account, more recently played matches should be used to parametrise models as it can be assumed that recent matches better represent current form than older matches. Using recent matches though, introduces a trade-off in the sense that as more weight is given to recent matches, the risk of the model parameters being affected by recent outlier matches increases. On top of that, the more restricted the data is on age, the smaller the sample size will be and therefore the average statistics will be further away from the true average performance of the player. Therefore the trade-off, data age versus data sample size, must be considered with care and be taken into account when parametrising our models.

3.8.2. Surface of Match Played

Professional tennis is played on a variety of court surfaces. The most commonly used surfaces are clay, grass, and hard courts where hard courts can either be indoor or outdoor. The court surface can affect ball speed and bounce which in turn favours some players and makes things more difficult for others. This has a direct impact on the performance of individual players on particular court surfaces, which is something which is reflected in the statistics of the matches played on those courts. When selecting the dataset which will be used to parametrise our model, one can restrict the sample to include matches played on the same surface as the match being modelled. Similarly to restricting the match age, a trade-off is introduced between representing the surface bias and decreasing the accuracy by shrinking the sample size.

3.9. Evaluating the Performance of Tennis Models

3.9.1. A Tennis Model Performance Rating, ρ

In a quest to evaluate how accurately various models perform in predicting tennis outcomes, we have defined a performance rating value, ρ . ρ can take values between -1 and 1 and is a measure of the average information a model reveals about the outcomes of a set of matches. A perfect model would have a value of $\rho = 1$, whereas a model which reveals no information yields a $\rho = 0$. For example a completely random model is expected to have a ρ close to 0. A model with a $\rho = -1$ is a model which always predicts the winner to be the actual loser of a match with absolute certainty therefore is as good a model as the model with $\rho = 1$.

In order to evaluate ρ , a set of N matches, M , for which there is a known outcome is used. For each of these matches, M_i , where $1 \leq i \leq N$, a model is used to predict the outcome of M_i using only the historical data which would be known immediately prior to the match. This process is known as back-testing in financial terms. Assuming the model outputs the predicted winner of the match along with a probability of winning, p_i , then one can penalise the model for getting a wrong prediction by subtracting the probability provided from a total, T , and reward it for getting a correct prediction by adding the probability of winning to the same total. In cases where the model does not make a prediction (i.e. $p_i = 0.5$ for either player) then a value of 0 is added to the total.

$$T = \sum_{i=1}^{i=N} \begin{cases} 0, & \text{if } p_i = 0.5 \\ p_i, & \text{if } p_i \neq 0.5 \text{ and prediction correct} \\ -p_i, & \text{if } p_i \neq 0.5 \text{ and prediction incorrect} \end{cases} \quad (3.48)$$

Using this approach, T would be equal to N in the case where the model is perfect (i.e. predicts all outcomes with a probability of 1) and $T = -N$ for a model which predicts all results wrong with probability 1. In the case where the model offers no information for any match T would be equal to 0. Therefore T , offers all that is needed to know about the predicting power of tennis models except in cases where models have been tested on different sets of matches which have different sizes. This is where ρ is useful which is simply T scaled down by the factor N .

$$\rho = \frac{T}{N} \quad (3.49)$$

To demonstrate how ρ is evaluated let's assume we have 4 players, A, B, C and D. These players have played 4 matches between them and for which we used 3 different models, X, Y and Z, to predict the outcomes of their matches. The results are presented in the following table.

Match	Winner	Winner X/Prob	Winner Y/Prob	Winner Z/Prob
A vs B	A	A / 1.0	A / 0.8	A / 0.65
A vs C	C	C / 1.0	C / 0.55	C / 0.6
B vs C	B	B / 1.0	B / 0.6	C / 0.55
C vs D	C	C / 1.0	D / 0.55	C / 0.55

Model X is the perfect model therefore ρ is expected to be equal to 1 which is true as $\rho_X = \frac{1+1+1+1}{4} = 1$. Models Y and Z both get one prediction wrong with Model Y giving slightly better probabilities in the correct predictions. One can say that Model Y performs slightly better than Model Z even though they predict the same number of correct results. Evaluating the performance rating for each of these models, $\rho_Y = \frac{0.8+0.55+0.6-0.55}{4} = 0.7$ and $\rho_Z = \frac{0.65+0.6-0.55+0.55}{4} = 0.625$, we can confirm that model Y has in fact a larger value of ρ and is deemed a better performing model than Z.

3.9.2. The Random Model

In order to assess whether our models provide us with any useful information about an upcoming match, we devised a model which is random and provides absolutely no information about the match. This Random Model will be used as a comparison to other models in conjunction with split testing as described in Section 2.2.3.

The Random model provides the probability of Player A winning the match using samples from a uniform distribution with parameters $a = 0$ and $b = 1$. This uniform distribution is identical to the one described in Figure 2.3. This model uses no information in modelling the match and simply provides a random probability. We will later on generate samples from this distribution to compare with the probabilities generated using the models described in this chapter.

3.9.3. Back-testing Using Real Data

In order to test the performance of our models we will perform a process known, in the financial world, as back-testing. For a number of already known real matches we will attempt to predict their result using our models as if we only have data up

to and not including the match modelled. This way we avoid over-estimating our models and knowing the result of the match we can assess whether the prediction was successful.

All the back-tests presented in the following sections were done on a total of 7938 ATP Tour matches played in the period beginning from 01/01/2011 to 31/12/2013. We present several different back-testing runs on this data, each time varying the subset of data which was used to generate the parameters of the models. More specifically we vary the period over which statistics are averaged for players as well as filter the statistics according to surface in an attempt to observe the effects on the results as discussed in Section 3.8.

Tables 3.1, 3.2 and 3.3 show the back-testing results using 3, 6 and 12 months of available data respectively. It is noticeable that even though 7938 matches were played in the period tested, our models only attempted to predict a fraction of those matches. For example, when using 3 months of statistical data going back from the match tested, only 6551 matches were tested. While testing, matches may be skipped for two main reasons. Firstly, if a match being tested is a match that was terminated abnormally, i.e. by a retirement or a walkover, that match is skipped. Secondly, if any of the two players have not played and finished matches in the period assigned before the match (i.e. 3 months) then no statistics are available for those players and thus testing is not possible. In fact it can be observed that as the period of background data is extended, more and more matches are attempted (for 3 months 6551 matches, for 6 months 7184 matches and for 12 months 7211 matches). Therefore the first advantage of using more data to model matches is immediately apparent – the more data we have the more matches we can model.

For every back-testing run, we observe four results as shown in the columns of Tables 3.1, 3.2 and 3.3:

- The success percentage, which is the percentage of correct predictions each model makes (a prediction is considered correct when the winner of the match is given a probability of winning greater than 0.5).
- The average probability of a model which is the average probability the model gives for predicted winners over all the predictions. This is a useful measure as it can help understand why the prediction percentage is low. For example if the model outputs a 0.6 probability of someone winning, then one can expect that it will fail in its prediction 40% of the time.
- The total, T , which is described in Section 3.9.1 and is used to evaluate the

performance of the model. The bigger T is, the better the model.

- The value ρ , which is T divided by the number of matches tested, to account for differences in sample sizes. The closer ρ is to 1 the better the model, as explained in Section 3.9.1

Table 3.1.: Results from a 3-month all surface back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 6551 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Model	Success %	Avg. Prob.	Total, T	ρ
Uncombined	61.8226	0.5782	1324.51	0.2022
Combined	64.8451	0.6074	1709.85	0.2610
Barnett	64.7535	0.6070	1701.55	0.2597

To generate results for Table 3.1 6551 ATP Tour matches were predicted using three different models, using as input statistics from matches played in the period between three months prior the match being predicted and up to but not including the match itself. This way we ensure that there is no over-estimating of the data and that in fact we only use data that would be known at the time the match was played. The three models used is Barnett’s model [33], and the ‘uncombined’ and ‘combined’ point models presented in this chapter. Using three months of background statistics affects results greatly. As most average statistics (used as input to the models) are based on few matches they do not approximate the population means well. This is evident in the results where we see low success percentages in all models. We can also observe that the performance in terms of success percentage and ρ is lower for the ‘uncombined’ model than the other two models. Barnett’s model and the ‘combined’ point model seem to be performing on a similar level.

Table 3.2.: Results from a 6-month all surface back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 7184 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Model	Success %	Avg. Prob.	Total, T	ρ
Uncombined	62.7506	0.5766	1520.98	0.2117
Combined	65.4649	0.6054	1907.83	0.2656
Barnett	65.4092	0.6051	1910.19	0.2659

Increasing the period of data the back-testing algorithm is allowed to use, from three months to six months prior the match simulated, makes a big difference on

the amount of matches that can be tested (increasing from 6551 to 7184). Table 3.2 summarises the results of the 6 month all surface back-test. It can be observed that performance increases slightly for all three models but the performance of the ‘uncombined’ model compared to the other two models remains significantly lower.

Table 3.3.: Results from a 12-month all surface back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 7211 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Model	Success %	Avg. Prob.	Total, T	ρ
Uncombined	63.2506	0.5757	1558.29	0.2161
Combined	66.5511	0.6049	2002.12	0.2776
Barnett	67.2722	0.6046	2047.43	0.2839

Further increasing the period of statistics used to twelve months prior to the match being simulated, further increases the performance of all models. The ‘uncombined’ model still performs worse than the other two models and Barnett’s model seems to be outperforming the ‘combined’ model in both success percentage and ρ measure. In reality though the increase in success percentage is not statistically significant using a 95% confidence level two-sample z-test which yields a p-value of 82.13%.

From the results, it is evident that the quantity of data available plays a vital role in the efficiency of the models. We may even go as far as to say that having more data is more important than filtering data by surface to capture the surface specific performance of players or limiting the back-period of data to capture the more recent fluctuations in the performance of players. It seems that accurate representation of the mean statistics of the population is more important.

Further expanding the period of data to 24 and 36 months decreases ρ more and more therefore it appears a good time period to use for data is in fact the full year of statistics before the match being simulated. The results of unfiltered 24 month and 36 month back-testing for these models are presented in the results section of Chapter 4 where they are compared with the Common-Opponent model. Instead, here we will present 12, 24 and 36 months surface filtered back-testing runs.

Tables 3.4, 3.5 and 3.6 show the results retrieved when running back-tests with 12, 24 and 36 months of match statistical background filtered according to the surface of the match being modelled. For example, to model a match which was played on grass on the 01/06/2012 using a 12 month surface filtered back-test,

the statistics of all matches in which the two players being modelled have played on grass during the period 01/06/2011 to 31/05/2012 will be used to generate the averaged statistics feeding the models.

Table 3.4.: Results from a 12-month surface filtered back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 6501 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Model	Success %	Avg. Prob.	Total, T	ρ
Uncombined	62.4658	0.5793	1388.49	0.2111
Combined	65.7191	0.6119	1809.35	0.2751
Barnett	65.5366	0.6108	1787.68	0.2718

In the case of 12 month surface filtered tests, the amount of matches tested was fairly low at 6501 compared to the 12 month unfiltered test which was 7211. The success percentages of the models also suffer from filtering the data according to surface as it can be observed in Table 3.4. Both results suffer because of the reduced number of available matches to average over, resulting in a poor representation of the population means of the players’ statistics, once again confirming that quantity of background matches is important.

Table 3.5.: Results from a 24-month surface filtered back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 6916 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Model	Success %	Avg. Prob.	Total, T	ρ
Uncombined	63.1001	0.5778	1493.51	0.2159
Combined	66.3245	0.6119	1950.32	0.2820
Barnett	66.4546	0.6112	1953.13	0.2824

In an attempt to maintain the surface filtering to capture surface specific performance of players, the period which was used to average matches over was increased to 24 months. Table 3.5 show the results. The amount of matches tested was increased significantly as more players now have background data and the success percentages also seem to increase for all three models.

Further increasing the time period of surface filtered back-tests to 36 months made little difference in both the amount of matches tested and the success percentages of the models themselves. Regardless, the results of all three models are, even in the slightest, improved and are now comparable to the 12 month unfiltered

Table 3.6.: Results from a 36-month surface filtered back-test using the ‘uncombined’, ‘combined’ and Barnett’s models to predict 7051 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Model	Success %	Avg. Prob.	Total, T	ρ
Uncombined	63.7356	0.5765	1562.63	0.2216
Combined	66.8274	0.6100	2001.05	0.2838
Barnett	66.4870	0.6088	1968.77	0.2792

backtests. In all three cases, the behaviour of the models relative to each other remain the same, the ‘uncombined’ model seems to be trailing behind in terms of predictive performances when compared to the ‘combined’ and Barnett’s model which seem to be closely matched.

3.9.4. Comparing Models Against the Random Model

In this section we will compare the ‘*uncombined*’ model, the ‘*combined*’ model (both described in Section 3.7.3) and Barnett’s model [33] against the random model using back-testing methodology. We will then demonstrate with significance testing that all models offer some information towards predicting the outcome of matches.

Using a two-sample Z-test as described in Section 2.2.3 we test whether the success rate of the Random model, which as expected is near 0.5, is different from the success rates of the 12 month all surface back-tests of our three models.

Table 3.7.: A two-sample Z-test using results from a 12-month all surface back-test of a random model and the uncombined model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.

Model	n	p	Standard Error	p-value
Random	7938	0.49358	0.005611	-
Uncombined	7211	0.63251	0.005678	0.000

Table 3.7 shows the results of a split test using as a control the success rate of the Random model and as the treatment the success rate of the ‘uncombined’ model. It is obvious that the difference is significant with a p-value very close to 0. The test therefore suggests that the average success rates of the two samples are in fact different with a probability near 1. This means that without a doubt the ‘uncombined’ model is a better predictor of match results than the random model.

Table 3.8.: A two-sample Z-test using results from 12-month all surface back-test of a random model and the combined model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.

Model	n	p	Standard Error	p-value
Random	7938	0.49358	0.005611	-
Combined	7211	0.66551	0.005556	0.000

Table 3.9.: A two-sample Z-test using results from 12-month all surface back-test of a random model and Barnett’s model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.

Model	n	p	Standard Error	p-value
Random	7938	0.49358	0.005611	-
Barnett	7211	0.67272	0.005526	0.000

Tables 3.8 and 3.9 show a similar analysis for the ‘combined’ model and Barnett’s model respectively. Both tests result in a confirmation that both models display increased success rate in predicting match results in comparison to the Random model.

This serves as a confirmation that models work regardless of how they compare to one another, but the question of which model performs better still remains.

3.9.5. Uncombined Model vs. Combined Model

In an attempt to test whether the apparent performance advantage of the ‘combined’ model in relation to the ‘uncombined’ model is statistically significant we perform yet another two-sample Z-test.

Table 3.10.: A two-sample Z-test using results from 12-month all surface back-test of the uncombined model and the combined model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.

Model	n	p	Standard Error	p-value
Uncombined	7211	0.63251	0.0056775	-
Combined	7211	0.66551	0.0055561	0.000016

Table 3.10 shows the results of this test on the success rates of the two models’ 12 month all surface back-tests. The ‘combined’ model success rate is better than the success rate of the ‘uncombined’ model with 0.99984 confidence which

satisfies the 95% confidence level. Therefore we can conclude that the combined statistics point model in fact performs better than the uncombined one.

3.9.6. Barnett Model vs. Combined Model

In Table 3.3 we can observe that while back-testing Barnett’s model we achieved a success percentage in predicting the results of matches of the order of approximately 0.7% higher than our ‘Combined’ model. In this section we test this higher percentage for significance.

Table 3.11 shows analytically the two-sample Z-test that was used to test whether success percentages of the two models are in fact different with a confidence level of 95%. In fact the test shows that there is a probability of 0.17872 that the two sample averages are the same. Therefore the test does not satisfy the confidence level of 95% and we must in fact accept that the two models have the same level performance.

Table 3.11.: A two-sample Z-test using results from 12-month all surface back-test of Barnett’s model and the ‘combined’ model for 7211 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.

Model	n	p	Standard Error	p-value
Barnett	7211	0.67272	0.005526	-
Combined	7211	0.66551	0.005556	0.17872

Despite the fact that the two models appear to perform on the same level, the ‘combined’ model still has the advantage of further detail. It improves upon its predecessor the Barnett model in the sense that it can resolve probabilities within point play. For example the model can output a probability of an ace being struck at any particular serve or it can adjust the output probability based on the fact that the first serve was lost and the point is now in play with a rally. The ‘combined’ model also sets up the ground for further in-point analysis which will be made possible with increased availability of more detailed statistics.

3.9.7. Combined Model vs. Bookmaker Models

Bookmakers make it their business to generate odds to provide to punters for the purpose of betting. The models that generate those odds can be considered as tried and tested from the wide audience that they receive. For professional tennis, these

odds, which are published prior to the match (opening odds), can be converted into probabilities of players winning tennis matches but unfortunately would not provide a useful comparison for the performance of models in the same way as other models. The odds published include a profit margin for the bookmaker which can be as high as 10%. For example in the 2011 Brisbane first round match between Michael Berrer and Dudi Sela the bookmaker Bet365 had opening decimal odds of 1.83 for both players when in fact for equal decimal odds one would expect odds of 2 for each player. Odds of 1.83 translate to a probability of 0.546 for both players winning due to the fact that Bet365 added approximately 9.2% profit margin to their odds depending on volumes bet on each player outcome. This percentage over the “true” odds is known as the bookmaker’s over-round.

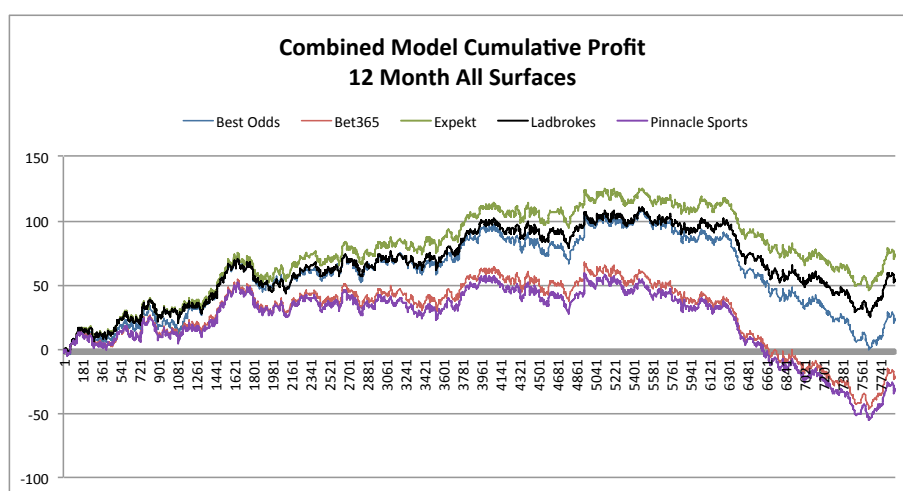


Figure 3.15.: The cumulative profit of 7132 bets with exposure 1 unit against the best match opening odds from 5 bookmakers and 4 other bookmaker’s over-round-corrected opening odds.

In order to fairly measure how our ‘combined’ model would perform against bookmaker models, we used two different methods. Firstly, we combined the opening odds of 5 leading bookmakers (Bet365, Ladbrokes, Expekt, Pinnacle Sports and Stan James) to retrieve the best (maximum) decimal odds available and simulated betting with those odds on the predicted winner of the match. Secondly, we generated new over-round-corrected odds for each of four popular bookmakers and proceeded to simulate the same bets for each individual bookmaker. To correct for this over-round which is included in the bookmakers odds, one can adjust these odds proportionally so that the implied probabilities of both winner and loser of a

match add up to 1. The initial odds were provided by tennis-data.co.uk who hold a record in .csv files of a great wealth of ATP Tour tournaments and their opening odds.

Our betting simulation wagers bets of value 1 unit on all available predictions from a 12 month, all surface back-test run. The cumulative profit after 7132 bets on best odds was 23.98 units, a value which represents 0.28% return on investment. The outcome of simulating bets against the over-round-corrected odds of individual bookmakers, ranges from 0.432% loss (Pinnacle Sports) to 1.03% profit (Expekt). The differences in these results is can be attributed to the varying profit margins each bookmaker uses. It is evident from the results that Expekt use higher profit margins and when correcting the odds for over-round this affects the cumulative profit in a positive manner, as the odds are increased more to correct for this profit margin. Figure 3.15 shows the cumulative profit plotted against the number of bets in chronological order for the best odds, as well as for the four over-round-corrected bookmaker's odds.

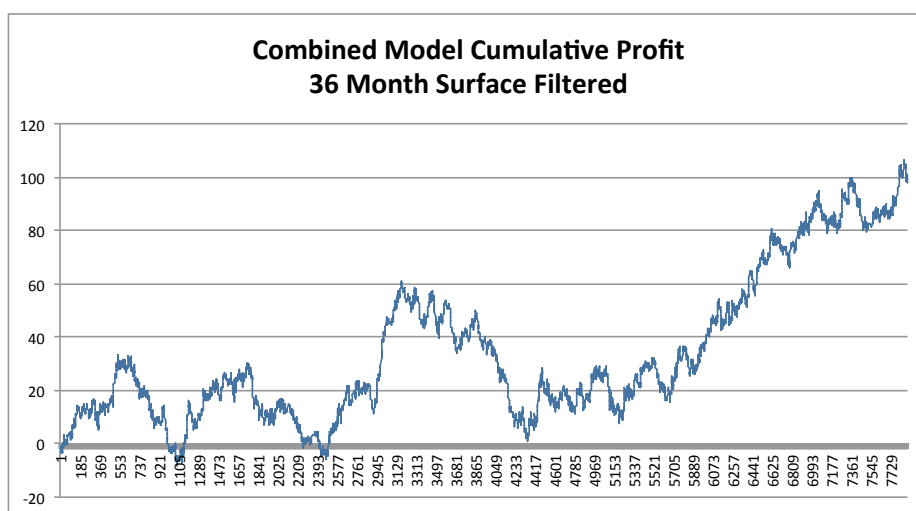


Figure 3.16.: The cumulative profit of 6973 bets with exposure 1 unit against the best match opening odds from 5 bookmakers.

The best odds seem to provide a good comparison that is not affected by the bookmakers' profit margins. In the case of the best odds, the 'combined' model was able to achieve a small profit with no betting strategy. It is therefore reasonable to assume that the model is of at least comparable performance as the models that are active in the professional world. In fact, by using the predicted winners

of the 36 month surface filtered back-test with the same simulated betting strategy against best opening odds, we were able to achieve a return on investment of 1.45%. Figure 3.16 shows how the cumulative profit develops over 6973 bets, ordered chronologically.

3.10. Conclusions

In this chapter, we introduced a Markov chain model for the probability of a tennis server winning a point. This model is parameterised using two different approaches. In the first approach, average statistics, from the past performance of the player serving the point, are used to calculate the parameters of the point model. In the second approach, average statistics from both the server and the receiver are combined and used to calculate the parameters of the point model. The resulting tennis match models from these two approaches were named the ‘*uncombined*’ model and the ‘*combined*’ model respectively.

We then discussed the statistics themselves and how using different subsets of matches over which averaged statistics are estimated can affect the outcome of the point model. We focused on two factors which affect the selected subsets of historical matches – the age and the surface of the historical matches with respect to the modelled match. We concluded that the subset of matches used to generate average statistics must firstly have a sufficient quantity of matches to ensure that sample averages approach the population means of statistics. If that is achieved then and only then should match age and surface filters be applied.

Retrieving different subsets of historical matches, we tested the performance of our two models against a random model, a recent literature model and industry standard models. It was found that all the models significantly outperform the random model. This means that all the models contribute some information towards the outcomes of matches. The ‘*combined*’ model significantly outperforms the ‘*uncombined*’ model but has no significant improvement in performance from Barnett’s model or from the industry models.

We have therefore introduced a model which analyses the tennis point and provides the user with the capability to retrieve detailed in-point probabilities which other models in literature do not touch upon. Additionally, our ‘*combined*’ model performs as well as other models in literature and in the industry.

Having said that, there is a lot of room for improvement and future work. For example, instead of using averaged statistics one can use weighted statistics, weigh-

ing them in terms of recency (to account for player form), surface (to account for surface relevance) or even according to opponent's ranking (to account for poorer performance when facing harder opponents or better performance when facing easy opponents).

In the future, the model can be further developed to estimate rally victory and serving probabilities. With all the research time dedicated to automated annotation and statistics generation from video feeds and tools like Hawkeye which are installed in tennis courts all over the world, more detailed statistics are around the corner. These can be utilised in the future to break down the point model even further. More work should also be involved with analysing the model's performance in women's professional tennis and in doubles matches.

4. A Common-Opponent Based Model

4.1. Introduction

In Chapter 3, an approach for estimating the probability of a player winning a service point was introduced. This approach makes use of a Markov model to analyse the point as it evolves. The state transition probabilities of this Markov model are estimated using average historical player statistics adapted according to the opponent faced.

Although this approach is intuitively appealing, it is not perfect. Average historical statistics are likely to contain a bias because of the way tournaments are structured. A good player is more likely to advance in a given tournament and face other strong players. Additionally, a strong player is more likely to be ranked in the top of the rankings and thus have a seeded position in tournament draws. This means that in the early rounds of tournaments, a strong player is likely to face a much weaker player. At the same time a weaker player will tend to drop out early in tournaments as the early opponents are more likely to be top performing players. This distorts average historical statistics over a set of matches in the sense that weaker players tend to face very strong players in the majority of their matches which on average makes them perform more poorly. Strong players face players from the entire spectrum of player ability. From this it can be concluded that for weak players the notion of the “average opponent” can be quite different than for strong players because the “average opponent” a weak player will face will be more skilled than the “average opponent” of a strong player.

The Common-Opponent model introduced in this chapter attempts to overcome exactly this problem. The model is designed to take advantage of the transitive element of tennis and work with data which is common to both players modelled. It achieves this by finding opponents faced by both players being modelled and then use statistical data from matches that were played against those common opponents. This automatically ensures that the level of the “average opponent” faced by both players is approximately equal.

4.2. The Concept of Transitivity

The Common-Opponent model utilises the transitive component of tennis. Tennis though is not a completely transitive sport. Complete transitivity in any competitive activity would assume that if a player can beat another player in a competition and that other player can beat a third player in the same competition, then it must hold that the first player can beat the third player. In other words, had there been a perfect ranking of players with the first player being the best and the last player being the worst at the activity, in a completely transitive activity it must always hold that the better ranked player will always win a lower ranked player.

It is obvious that tennis is not an absolutely transitive sport as there are a lot of examples where Player A has won a match against Player B, Player B has won a match against Player C and Player C has won a match against Player A, but it is safe to assume that there is a transitive element to the game. It is this transitive element that conceptually allows us to compare the performance of two players against a common third one and come to a conclusion as to which of the two players has a better chance of winning when facing each other. This intuition is the basis of the Common-Opponent model.

In the case of professional singles tennis matches, this reasoning has the potential to be especially fruitful, because of the limited number of players involved. There are roughly only 150 active professional tennis players in each of the ATP and WTA tours. Within each tour, players frequently play against each other in a variety of tournaments. Although there are a limited number of head-to-head encounters for any two given players, many pairs of players share a rich set of common opponents.

4.3. Relationship between the probabilistic difference of winning service points and winning the match

A brief discussion of a concept introduced by Klaassen and Magnus [29] and later further investigated by O'Malley [35], is essential in order to comprehend the reasoning behind the Common-Opponent algorithm. In this section, we briefly discuss the important insight that the probability of a player winning a match is closely related to the difference of the two players' probability of winning a point while serving.

Using O'Malley's equations which estimate the probability of winning a match

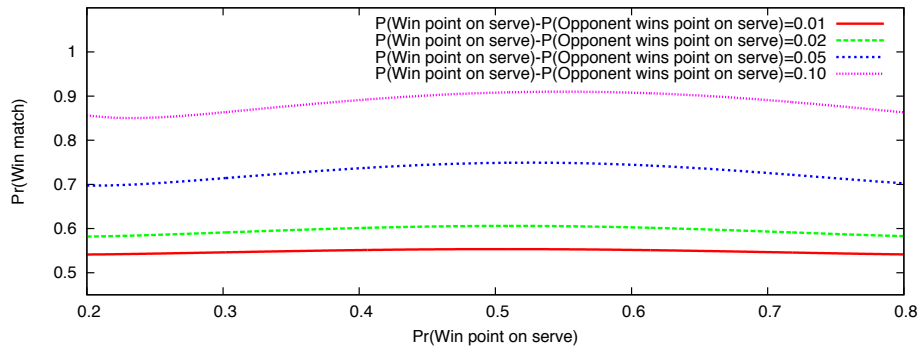


Figure 4.1.: Probability of the dominant player winning a best-of-three-sets tennis match with fixed differences of 0.01, 0.02, 0.05 and 0.10 in the two players' probability of winning a point on serve.

using the player's probabilities of winning points on serve, we replicated a graph presented in his paper (Figure 4.1). This graph demonstrates the relationship between the difference in the probability of winning points on serve of the two players and the probability of the dominant player winning the match. By plotting the dominant player's probability of winning the match while varying the player's probability of winning a point on serve, we show different plots for various fixed differences in the two players' probability of winning a point on serve. Note the x-axis runs from 0.2 to 0.8 corresponding to the domain of values likely to be encountered in professional tennis.

It is evident from this plot that there is only a small deviation from the initial probability of winning the match, even if the probabilities of winning a point on serve are varied, as long as the difference between the dominant player's probability of winning a point on serve and his opponent's probability of winning a point on serve remains constant. It is this relationship that allows to generate Equation 4.2 later on.

4.4. Match Probabilities Using Common-Opponent Model

Let players A and B be the two players playing in the match we wish to model. Also, let C_i for $1 \leq i \leq N$ be the N common opponents they have faced in the past. For each C_i we denote $spw(A, C_i)$ as the proportion of service points won by A against C_i and $spw(B, C_i)$ as the proportion of service points won by B against

C_i . Similarly, $rpw(A, C_i)$ is the proportion of returning points won by A against C_i and $rpw(B, C_i)$ is the proportion of returning points won by B against C_i . This is illustrated in Figure 4.2. In cases where either A or B has faced the common opponent C_i in multiple matches during the period of the data set, then $spw(A, C_i)$, $rpw(A, C_i)$, $spw(B, C_i)$ and $rpw(B, C_i)$ can either represent the averages over those matches or they can be added as different common-opponent contributions (more in the example which follows in Table 4.1).

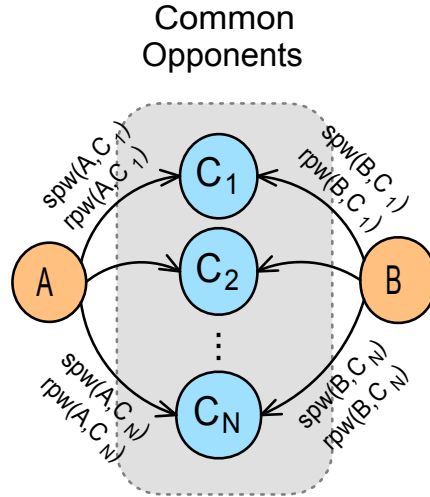


Figure 4.2.: Parameters of the Common-Opponent Model.

As discussed earlier, following O'Malley's findings [35], the difference in service points won can be used as an indicative measure of the probability of a player winning the match against an opponent. In order to model how A and B would play against each other through their common opponents, C_i , we first need to calculate the differences in service points won by A and B against those opponents. We can then additively combine those differences to come up with an indication of how well A would perform against B.

For each common opponent, C_i , we compute Δ_i^{AB} which represents a measure of the advantage (or if negative, disadvantage) Player A has over Player B in terms of the proportion of service points won against opponent C_i , as follows:

$$\Delta_i^{AB} = (spw(A, C_i) - (1 - rpw(A, C_i))) - (spw(B, C_i) - (1 - rpw(B, C_i))) \quad (4.1)$$

This value can be used to additively influence an arbitrary probability of winning

a point on serve for player A or player B in any hierarchical model. Let $M_3(p, q)$ represent a function which estimates the probability of a player winning a best-of-three tennis match, using parameters p and q where p is the probability of the player winning points while serving, and q is the probability of the player winning points while returning. Using this function we show how one can approximate the probability of Player A winning a best-of-three-sets match against Player B based on their past performances against common opponent C_i :

$$\Pr(A \text{ beats } B \text{ via } C_i) \approx \frac{M_3(0.6 + \Delta_i^{AB}, (1 - 0.6)) + M_3(0.6, (1 - (0.6 - \Delta_i^{AB})))}{2} \quad (4.2)$$

In Equation 4.2, we calculate the match probabilities twice: once by positively influencing Player A's probability of winning a service point and once by negatively influencing Player B's probability of winning a service point. Subsequently, we average the two values. We arbitrarily use the value 0.6 because it is the approximate average probability of a professional player winning a point on serve when playing against another professional player (recall that, as shown in Figure 4.1, the exact value is not critical). Also, note that in cases where Δ_i^{AB} is less than -0.6 or either greater than 0.4 the input values to Equation 4.2 become invalid. To make sure this does not happen, we need to cap the values within the boundaries $-0.6 \leq \Delta_i^{AB} \leq 0.4$ for all i .

To combine all the available data from all common opponents, we calculate the average of $\Pr(A \text{ beats } B \text{ via } C_i)$ over all common opponents, C_i , to estimate the probability of player A winning the match as follows:

$$P_{avg}^{AB} = \frac{\sum_{i=1}^N \Pr(A \text{ beats } B \text{ via } C_i)}{N} \quad (4.3)$$

To illustrate the Common-Opponent approach we will model an example from the second round of the 2013 US Open played by Andy Murray vs. Leonardo Mayer.

Table 4.1 presents the percentage spw and rpw against the opponents for each player. You will notice that in this particular example the names of Mikhail Youzhny and Marcel Granollers appear twice as common opponents. The reason for this is that Andy Murray has played two matches with those two particular opponents in the past 12 months each match having a different spw and rpw for Andy Murray. There are two approaches to deal with cases like this. This first approach involves averaging the statistics of Andy Murray over all the matches where he faced the

Common Opponent	Murray <i>spw</i>	Murray <i>rpw</i>	Mayer <i>spw</i>	Mayer <i>rpw</i>
Kei Nishikori	73%	50%	56%	34%
Robin Haase	66%	54%	65%	35%
Carlos Berlocq	62%	46%	44%	7%
Juan Martin Del Potro	59%	33%	59%	27%
Bernard Tomic	74%	51%	57%	19%
Andreas Seppi	76%	44%	59%	36%
Florian Mayer	71%	36%	48%	29%
Tommy Robredo	75%	41%	70%	38%
Mikhail Youzhny	66%	44%	63%	53%
Mikhail Youzhny	67%	52%	63%	53%
Marcel Granollers	64%	42%	49%	28%
Marcel Granollers	51%	44%	49%	28%

Table 4.1.: Statistical data on matches played with common opponents for Andy Murray and Leonardo Mayer in the second round of 2013 US Open. The data includes all common opponent ATP matches played within 12 months of the modelled match.

same common opponent and using those as the *spw* and *rpw* readings in our model. The second approach would be to combine each match with all possible combinations of the second player's matches with that particular common opponent which is what is shown in Table 4.1. Mikhail Youzhny occupies two rows, each with different *spw* and *rpw* values for Andy Murray but the same values for Leonardo Mayer since there was only one match of Youzhny vs Mayer. Each of these two rows will contribute its own component towards the probability of Andy Murray winning Leonardo Mayer.

We combine the above to estimate the advantage or disadvantage, Δ , Murray has over Mayer using Equation 4.1. Subsequently we calculate the probability of Murray winning a match with Mayer, given the information inferred from each of the rows in Table 4.1. These results are presented in Table 4.2.

Averaging out the results in Table 4.2 using Equation 4.3 gives us an estimated probability of 0.8825 of Andy Murray winning the match. In the event Murray won by 3 sets to 1, (7-5, 6-1, 3-6, 6-1) which was a comfortable victory.

One might note that there can be quite some variation between the probabilities

Opponent	Δ	Probability of Murray beating Mayer
Kei Nishikori	0.34	1
Robin Haase	0.2	1
Carlos Berlocq	0.57	1
Juan Martin Del Potro	0.06	0.83
Bernard Tomic	0.49	1
Andreas Seppi	0.25	1
Florian Mayer	0.29	1
Tommy Robredo	0.09	0.92
Mikhail Youzhny	-0.06	0.15
Mikhail Youzhny	0.03	0.69
Marcel Granollers	0.29	1
Marcel Granollers	0.18	1
Overall average		0.8825

Table 4.2.: Probability of Andy Murray winning against Leonardo Mayer, given data on each of the common opponent match combinations separately.

of winning in Table 4.2, as estimated from different common opponents. This suggests that predictions made with only a small number of common opponents should be treated with caution. However, our experience is that matches between active professional tennis players usually feature a sufficiently rich set of common opponents to yield a stable estimate. Also, notice that the Δ values for Carlos Berlocq and Bernard Tomic exceed the boundary value of 0.4. In the equation which calculates match probability (Equation 4.2), the value is capped to 0.4 to avoid invalid calculations.

Note that this approach can use any underlying function which estimates the probability of winning a match using the two probabilities of players winning service points (whether that is O'Malley's equations, or any other model in literature). In fact, Table 4.2, as well as all results later on, were generated using Barnett's conditional equations (described in Chapter 3) in combination with our estimation of the players' probabilities of winning a point on serve.

4.5. Evaluating Model Performance

In Chapter 3 we introduced three models, Barnett’s tennis model, the ‘uncombined’ statistics point model and finally the ‘combined’ statistics point model. We compared the models and concluded that even though there is no statistically significant improvement from Barnett’s model, the higher analytical power of the ‘combined’ model is an improvement by itself. We presented back-testing results for various subsets of historical data and showed that the models themselves provide some insight towards predicting the results of matches by comparing them to the random model. In this section we will do the same with the Common-Opponent model as well as present new back-tests for previous models.

4.5.1. Back-testing Results

Firstly, we run back-tests using the Common-Opponent approach searching for common opponent matches from 3, 6 and 12 month, all-surface, background data. That means that from the date of the match, we compile a list of all matches the two players have participated in the past 3, 6 or 12 months and search for matches where they faced common opponents. We then execute the Common-Opponent algorithm using statistics from those matches.

Table 4.3.: Results from 3, 6 and 12 month all surface back-tests using the Common-Opponent model to predict the outcome of 7938 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Back-test data period	Prediction Attempts	Success Percentage	Average Probability	Total, T	ρ
3 months	3484	60.2181	0.5839	674.14	0.1935
6 months	6540	61.3456	0.5786	1312.75	0.2007
12 months	6578	64.3509	0.5774	1507.30	0.2291

Table 4.3 shows the results of these three back-test runs. What is notable is that when using 3 months of background data, only for 3484 matches (out of a total of 7938) a prediction was even attempted. For the majority of matches there were no common opponents in the past three months therefore the algorithm could not even be executed. The number of attempted predictions increases significantly when we increase the background period to 6 and 12 months but the prediction success percentage remains quite low when compared to the performances in Ta-

bles 3.2 and 3.3. The performance of the Common-Opponent model is actually comparable to the ‘uncombined’ model. The reason for this is that even though we use data which go back to 12 months in time, the amount of matches which we actually use is mostly quite limited and as we showed in the last chapter, a critical amount of matches need to be collected for the models to reach their highest level of performance.

Table 4.4.: Results from a 24-month all surface back-test using the uncombined, combined, Barnett’s and Common-Opponent models to predict 7938 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Back-test data period	Prediction Attempts	Success Percentage	Average Probability	Total, T	ρ
Uncombined	7316	63.4090	0.5743	1589.00	0.2172
Combined	7316	66.8125	0.6043	2043.68	0.2793
Barnett	7316	67.1679	0.6037	2064.36	0.2822
Com. Opp.	7000	65.8000	0.5754	1700.23	0.2429

Following our own advice, we decided to increase the period, over which we allow data to be collected, to 24 months and all surfaces. Table 4.4 shows the results of running back-tests for all four models using 24 months of background data including all surfaces. It is evident from this table that increasing the background data increases the performance of the Common-Opponent model significantly while the other models have similar performances as in the 12 month all-surface back-tests. We also see a significant increase in attempted predictions in the case of the Common-Opponent model.

Table 4.5.: Results from a 36-month all surface back-test using the uncombined, combined, Barnett’s and Common-Opponent models to predict 7938 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Back-test data period	Prediction Attempts	Success Percentage	Average Probability	Total, T	ρ
Uncombined	7348	63.0920	0.5730	1559.71	0.2123
Combined	7348	66.7665	0.6028	2044.92	0.2783
Barnett	7348	67.2156	0.6019	2064.56	0.2810
Com. Opp.	7095	65.5814	0.5722	1698.91	0.2395

Increasing the historical data even more to 36 months with all surfaces, we no-

tice a fairly stable behaviour from all models with insignificant variations in success percentages. These results are shown in Table 4.5.

Table 4.6.: Results from 12, 24 and 36 month surface filtered back-tests using the Common-Opponent model to predict the outcome of 7938 ATP Tour matches played between the 1st of January 2011 and the 31st of December 2013.

Back-test data period	Prediction Attempts	Success Percentage	Average Probability	Total, T	ρ
12 months	4102	62.5792	0.5886	907.38	0.2212
24 months	5285	63.4437	0.5832	1193.11	0.2258
36 months	5765	63.3131	0.5794	1276.04	0.2213

Finally, we test what happens to the Common-Opponent model when we limit the historical data subset according to the surface of the match being modelled. It is understandable that by filtering the data according to the surface, limits the number of historical matches that can be used. When using the Common-Opponent model, this limits them even more as we search for common opponent matches played on the same surface as the match being modelled. It is because of this, that the number of attempts as shown in Table 4.6, are quite few. Even more, the model performs poorly and when compared to the performances of the back-tests shown in Table 3.6 it is again comparable to the ‘uncombined’ model performance.

4.5.2. Common-Opponent Model vs. Random Model

The first question that needs to be answered is whether the Common-Opponent model offers some insight towards the prediction of match results. This, as in the past, can be answered by comparing the Common-Opponent model performance with the performance of samples generated by the random model (defined in Section 3.9.2). To do this we chose the best Common-Opponent back-test (i.e the one with 24 months background from all surfaces) and we compared the success percentage it achieved with the success percentage of the random model.

Table 4.7 shows the calculation of a two-sample Z-test to compare the mean performance of the 24 month Common-Opponent back-test with the random model. The two samples have a probability close to zero of having the same means, something that can be interpreted as the fact that the Common-Opponent model does offer significant insight towards the outcome of matches.

Table 4.7.: A two-sample Z-test using results from a 24-month all surface back-test of a random model and the Common-Opponent model for 7000 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.

Model	n	p	Standard Error	p-value
Random	7938	0.4936	0.005611	-
Com. Opp.	7000	0.6580	0.005670	0.000

4.5.3. Common-Opponent Model vs. Uncombined Model

The uncombined statistics model introduced in Section 3.7.3 was the lowest performing of the models we tested. We will check whether the Common-Opponent model has achieved some improvement in performance from the ‘uncombined’ model by performing a Split test on back-testing runs of the two models which used background data of 24 months and all surfaces.

Table 4.8.: A two-sample Z-test using results from a 24-month all surface back-test of the ‘uncombined’ model and the Common-Opponent model for 7000 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.

Model	n	p	Standard Error	p-value
Uncombined	7316	0.6309	0.005632	-
Com. Opp.	7000	0.6580	0.005670	0.00139

Table 4.8 shows the results of this split test. We can observe that the p-value is 0.00139 which satisfies the 95% threshold set for this test. Therefore we can with 99.9% certainty say that the Common-Opponent model is an improvement in performance from the ‘uncombined’ model. This demonstrates that the novel approach behind the Common-Opponent model has some merit to it.

4.5.4. Common-Opponent Model vs. Combined Model

When we compared the 24-month, all-surface Common-Opponent back-test with the same back-test of the combined statistics point model (introduced in Section 3.7.3), we discovered that the models to have no significant difference in performance.

Table 4.9 shows the results of a two-sample Z-test comparing the success percentages retrieved from a 24-month all surface back-test run using the Common-Opponent approach and the combined statistics approach. The ‘combined’ model

Table 4.9.: A two-sample Z-test using results from a 24-month all surface back-test of the ‘combined’ model and the Common-Opponent model for 7000 ATP Tour matches played from 1st of January 2011 to the 31st of December 2013.

Model	n	p	Standard Error	p-value
Combined	7316	0.6681	0.005505	-
Com. Opp.	7000	0.6580	0.005670	0.1001

had 7316 attempts at prediction with success percentage at 66.8% whereas the Common-Opponent model had 7000 attempts with success percentage 65.8%. For those results the p-value is calculated to be 0.1001 which does not satisfy the 95% threshold and therefore we must consider the two models of equal performance.

4.5.5. Common-Opponent Model vs. Bookmaker Models

Finally, we investigate how the Common-Opponent approach to predicting professional tennis match outcomes compares to industry standard models. Similarly to the approach adopted in Section 3.9.7, we chose to place a 1 unit virtual bet, for every prediction the model outputs, against the best pre-match opening odds provided by 5 large bookmakers and finally aggregate the total winnings over three years of bets. Additionally, we also placed virtual bets of 1 unit, against over-round-corrected odds of 4 popular bookmakers.

The results obtained here were similar to the results presented in Section 3.9.7 in the sense that the model does the worse against the odds retrieved from Pinnacle Sports and best against the odds offered by Expekt. Once again this variation in the apparent cumulative profit can be attributed to the varying profit margins of the different bookmakers.

Figure 4.3 shows the cumulative profits from 6923 bets placed over 3 years using pre-match predictions of the Common-Opponent model having 24 months of historical data spanning over all surfaces. In the case of best odds, the Common-Opponent model ends up with a profit of 18.51 units and a return on investment of 0.27%. The final profit against the individual, over-round-corrected, bookmaker odds ranges from -36.34 units (Pinnacle Sports) to 61.04 units (Expekt).

The conclusion that can be drawn from this test is that the Common-Opponent model is at least comparable to the current industry standard models being able to compete and even achieve a small profit over 6923 bets.

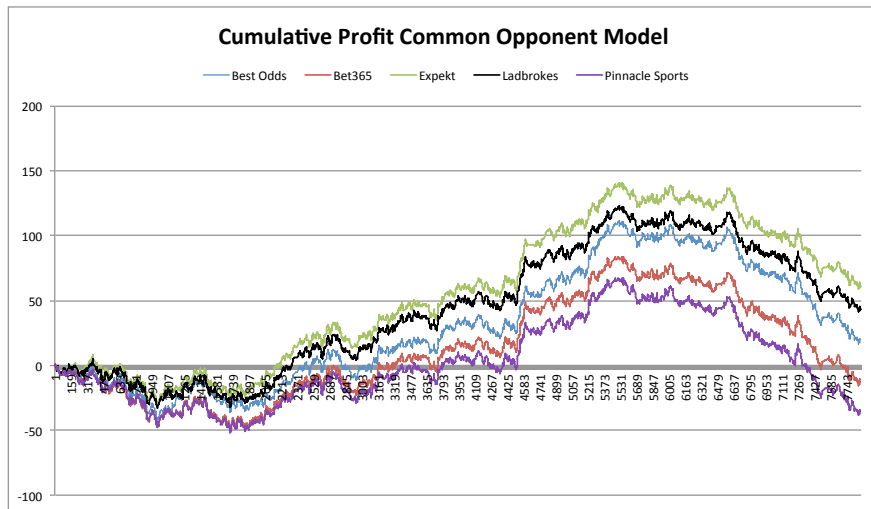


Figure 4.3.: The cumulative profit of 6923 bets with exposure 1 unit each, against the best match opening odds from 5 bookmakers and against the over-round-corrected odds of 4 individual bookmakers over all ATP Tour matches in the years 2011–2013.

4.6. Conclusion

In an attempt to improve on models which use averaged statistics to predict tennis results, we invented the Common-Opponent model. The Common-Opponent model exploits the transitive component inherent in tennis by comparing how players perform against opponents, which they both faced in the past, to come to a conclusion how the players will fare against each other in the present.

The Common-Opponent model, therefore, increases the relevance of historical data with the downside of decreasing the quantity of data. This is apparent from the results retrieved from back-tests, where using 3 month and 6 month background periods for historical data has a detrimental effect on prediction attempts and success rate. A period of 24 months had to be used to have comparable performance with other models.

Even though the results of this model were not as impressive as the ‘combined’ model’s when increasing the background period the difference between the two models was insignificant. Therefore the idea behind the model appears to be solid but needs to be improved by solving the problem diminishing background data.

This is what our suggestions of future work will focus on. As the results seem to be dependent on the number of common opponents in the sample investigated,

it might be interesting to investigate a recursive approach to the problem to expand that number of common opponents. In this case, we would extend the algorithm by considering common opponents between both modelled players, as well as the common opponents of those common opponents. There is scope for research with respect to determining the optimal depth of recursion as well as an appropriate limit on the number of common opponents considered at each stage. There is also room for investigation into using indirect common opponents, or second tier common opponents, to expand the data-set used for analysis. That is, opponents of the two main players who are linked via common opponents only and not directly.

Adding exponential time weights to the statistics of common opponents might also be an option to help increase the background data set. One could expand the background period to include more matches but weigh favourably on the more recent matches.

To conclude, evidence was provided to show that the Common-Opponent idea has merit as well as room for improvement. The novelty of the approach has attracted some attention and has even been implemented on a popular community website (namely <http://www.tennisinsight.com>).

5. Ranking Systems for Tennis Players

5.1. Introduction

Having discussed two different approaches to model the outcome of professional tennis matches in the last two chapters, this chapter shifts the topic of the thesis towards a new direction – that of player rankings. Official player rankings play a vital role in the perception the wider public has on player performance. Climbing to the top of the professional rankings is a very sought-after achievement for the players themselves. Reaching the top positions of professional rankings automatically gives an advantage to the players in terms of gaining seeded positions in tournaments. This is in addition to the fact that, by the end of the tennis season, the the top eight male tennis players qualify to play in the very prestigious events of the ATP World Tour Finals respectively, for hefty cash prizes and ranking points.

This chapter introduces various algorithmic approaches to generating professional player rankings and compares them with the official ATP rankings in an attempt to understand how well the rankings represent the set of matches used to generate them. We introduce three distinct approaches of generating player rankings (PageRank, SortRank and LadderRank) each with a few variations.

5.2. The PageRank Approach

The PageRank tennis ranking system was first introduced by Radicchi in 2011 [8] and further investigated by Dingle, et al. [7]. It is an effective ranking system for tennis players, which, in this dissertation, we vary in an attempt to improve it and use it as a good comparison for other ranking systems introduced. The system is based on the Google PageRank algorithm summarised by Brin & Page [41] which is used for ranking websites. This section explains the original formulation of the PageRank equation for the case of web-graphs and then explains how Radicchi applied it to the calculation of tennis rankings.

The original formulation of PageRank uses a random surfer model to measure

the relative importance of web-pages. The central idea of the algorithm is that pages which have a large number of incoming links from other pages are regarded as being more important than those with fewer incoming links; a surfer clicking through links on web-pages at random is therefore more likely to land on the more important web-pages. What is more, a link coming from an already important website “carries” more weight to the receiving site in the sense that if a popular site links to another site, then that means that the other site must be important as well.

For a web-graph with N pages, PageRank constructs an $N \times N$ matrix R that encodes a surfer’s behaviour in terms of the matrices W , D and E , which are now described.

The first behaviour modelled is a surfer who randomly clicks on links on a given page to move to another page. The corresponding matrix W has elements w_{ij} given by:

$$w_{ij} = \begin{cases} \frac{1}{deg(i)} & \text{if there is a link from page } i \text{ to page } j \\ 0 & \text{otherwise} \end{cases}$$

where $deg(i)$ denotes the total number of links out of page i . Therefore, w_{ij} is the probability of randomly picking a link which links page i to page j .

The second behaviour modelled is when a surfer encounters a page that has no outgoing links. In this case the surfer will randomly jump to any other page in the web-graph. This is described by the matrix $D = du^T$, where d and u are column vectors:

$$d_i = \begin{cases} 1 & \text{if } deg(i) = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$u_i = 1/N \quad \forall i, 1 \leq i \leq N$$

We note that other probability distributions for u are possible; here we consider only a uniformly distributed choice. The two behaviours are then combined into a single-step transition matrix $W' = W + D$.

The third and final behaviour modelled, is that of a surfer deciding to ignore the links on the current page and to surf instead to some other random page. This is captured in a dense matrix E with elements $e_{ij} = u_j \quad \forall i, j$.

The surfer’s overall behaviour is determined by whether or not they choose to follow the link structure of the web-graph or to jump about at random. The balance

between the two is controlled by the scalar parameter α ($0 \leq \alpha \leq 1$). The overall one-step PageRank Discrete-Time Markov Chain (DTMC) transition matrix R is therefore defined as:

$$R = (1 - \alpha)W' + \alpha E \quad (5.1)$$

which is a dense matrix due to the presence of E . The PageRank of the web-graph can be calculated by solving the DTMC steady-state problem:

$$x = xR \quad (5.2)$$

Radicchi's formulation of the problem is equivalent to the matrix-based description of PageRank given above. When using PageRank to model tennis, the pages in the web-graph now become records of the players and the outgoing links are now represented by defeats in head-to-head encounters of the players. Therefore instead of N pages as nodes, the graph in the Radicchi implementation has N players. The major difference in the tennis PageRank model from the original is that the original disregards multiple outgoing links from a single source page to a given target page, while in the tennis model the number of times a single player loses to each of their opponents is counted and used as a weight on the edges of the graph.

Each player (node) in the network is assigned a "prestige score" which is passed on to other players through weighted edges. The prestige scores, P_i in a network of N nodes, can be found by solving the system of equations given by Equation 5.3.

$$P_i = (1 - \alpha) \sum_j P_j \frac{w_{ji}}{s_j^{out}} + \frac{\alpha}{N} + \frac{(1 - \alpha)}{N} \sum_j P_j \delta(s_j^{out}) \quad (5.3)$$

for $i = 1, \dots, N$ with the constraint $\sum_i (P_i) = 1$.

In Equation 5.3:

- w_{ji} is the outgoing weight from player j to player i (i.e. the number of defeats player j has suffered against player i),
- s_j^{out} is the total out-strength of player j (i.e. $s_j^{out} = \sum_i w_{ji}$),
- α is a damping parameter where $0 \leq \alpha \leq 1$ and
- N is the total number of players in the network.

The δ function takes a value of 1 for an input of 0 and a value of 0 otherwise.

Creating an equivalent matrix representation of Radicchi's equation we can re-define Equation 5.1 by defining the (i, j) th entry of W , denoted w_{ij} , as the number of

matches player i has lost to player j normalised over the total number of matches player i has lost. Just as web-pages linked to by a large number of other pages will achieve a high PageRank score, so too will players who defeat a large number of other players.

The definitions of D , E and α are unchanged but we interpret them differently. We need D in the cases where a player has no defeats recorded against them – in reality this is unlikely to occur, but it may be the case in our data-sets given that we only have access to results from a limited time period. In this case, we assume the player is equally likely to lose to all other players given the absence of any information to the contrary.

Just as a surfer may disregard the links on a current page and surf to a random page, we believe that it is possible for any player to lose to any other (due to a variety of unpredictable external factors) and this is how we interpret E . The scalar parameter α lets us decide how likely we think it is that this will happen. In the experiments that follow in this chapter we set α to 0.01 unless stated otherwise.

5.3. Set, Game and Point PageRank Approach

As described in the last section, Radicchi's PageRank for tennis algorithm utilises a count of match defeats as the weights for the outgoing edges of the player graph. In this section we propose to alter this to use counts of sets, games or even points lost to another player as the outgoing weight. In other words instead of having w_{ji} represent the amount of matches player j lost to player i , it will now represent the amount of sets, games and points player j lost to player i .

One of the problems of PageRank for ranking players occurs when there is a limited number of matches for particular players. In the case where those players have no victories over other players, they have no incoming links. This automatically causes the prestige score of those players to be equally distributed between them and results in having a number of players at the bottom of the rankings with equal rank (for example in rankings generated for 2013 the last 104 players are all equally ranked). In an attempt to correct this, a measure of the number of sets, games or even points lost to other players helps expand the data-set and as a result there are far fewer players with no incoming edges at the bottom of the rankings. When using Sets lost as the weights in PageRank 2013 rankings, the number of equally ranked players is reduced to 60 from 104. When using Games lost as a weight in the same period the number is further reduced to 0 and the problem is

eradicated.

This is not without disadvantages though. As we move to more granular weights (i.e. to using games and points lost) to generate the prestige rankings a side-effect emerges. Players who, in the case of matches and sets, did not receive any prestige from the high ranked players because simply they did not achieve match and set victories against them, suddenly are able to receive some prestige because they have, most probably, won games and points against them. This results in low ranked players being boosted to better rankings simply because they played against top ranked players. On the other hand, players who have achieved a single victory against top ranked players but nothing else, are now losing positions in the rankings. As the data-set increases, the importance of that one victory decreases. A balance needs to be achieved between the two, keeping in mind that winning points does not necessarily translate to a better quality of player.

As it becomes evident from the results of experimentation with the various types of PageRank approaches, using sets lost as a measure of weight seems to be the best approach. This makes sense as it is a player's ability to win sets which defines his ability to win matches. The ability to win individual games and points does not reflect attributes such as the player's response to crucial points and breaking serve ability.

5.4. Comparing PageRank Approaches

In order to evaluate how well the rankings represent the set of matches which were used to generate them we propose a simple test. For each of the matches we count the number of times a winner is better ranked than the loser of the match. Such a match is successfully represented by the ranking system and the success percentage over all matches will offer a measure of the performance of the ranking system. We will also use the official ATP Tour rankings as a control ranking system.

Tables 5.1 and 5.2 show the success percentages achieved by the different PageRank approaches described, for the years of 2012 and 2013.

In Table 5.1 the success percentages were generated using the rankings as they were generated in the end of the year 2012. Since the ATP Official Rankings are generated using the last 52 weeks of play, the data-set which was used to generate the PageRank tennis rankings were also of the same time period. The success percentage is the percentage of matches within that period, whose winners were

Table 5.1.: Comparing different PageRank tennis ranking approaches to the official ATP Tour rankings as they were on 01/01/2013. The rankings contain a total 303 players who participated in ATP Tour matches over the year of 2012.

Ranking System	Match Attempts	Success Percentage	DataSet	Standard Error	p-value from ATP
ATP	2538	69.6612	-	0.009125	-
Match PR	2673	71.9416	2673	0.008690	0.0352
Set PR	2673	71.9791	6647	0.008687	0.0329
Game PR	2673	70.4826	65017	0.008822	0.2588
Point PR	2673	69.5847	409906	0.008898	0.5239

Table 5.2.: Comparing different PageRank tennis ranking approaches to the official ATP Tour rankings as they were on 01/01/2014. The rankings contain a total 305 players who participated in ATP Tour matches over the year of 2013.

Ranking System	Match Attempts	Success Percentage	DataSet	Standard Error	p-value from ATP
ATP	2488	70.2572	-	0.009165	-
Match PR	2615	70.4398	2615	0.008923	0.4433
Set PR	2615	71.1281	6326	0.008862	0.2473
Game PR	2615	71.0134	61916	0.008872	0.2767
Point PR	2615	70.3251	389927	0.008933	0.4789

ranked higher than the loser by each ranking system. The DataSet column displays the total number of outgoing edges each PageRank system uses demonstrating how the use of more granular scoring increases the available data. Finally the Table shows a p-value from a two-sample Z-test from the control ranking system (the Official ATP Ranking). We can see in Table 5.1 that the Match PageRank system and the Set PageRank system both significantly improve the representation of matches when compared to the ATP Rankings of 2012. The Game PageRank and the Point PageRank generated rankings seem to perform similarly to the ATP Official rankings with no significant difference observed.

In 2013 results are very different as demonstrated in Table 5.2. It appears that in 2013 there is no significant improvement in model representation from any ranking system when compared to the ATP Official Rankings even though the highest percentage was still achieved by the Set PageRank system.

To improve our understanding of how the different ranking systems affect individual player rankings we constructed Figures 5.1, 5.2, 5.3 and 5.4. These figures show the Match, Set, Game and Point PageRank generated rankings of the Top 100 ATP Players as ranked by the end of 2013 plotted against their respective ATP Rankings at the time. This enables us to spot differences in rankings easily. Players who appear below the $Y=X$ line are players who PageRank considers should be ranked better than their ATP Ranking. Players who appear above the line are players for which PageRank suggests a worse ranking than the one assigned by the ATP. The greater the distance from the line the bigger the difference of the PageRank suggested ranking and the ATP Ranking.

As a generic observation, in all PageRank systems it appears that the distances from the line increase with increasing ATP Ranking. There is more disagreement in the correct rank of players who are poorly ranked by the ATP. This is to be expected and there are two main reasons for this.

The most prominent reason is that, in the official ATP Rankings, points from Challenger tournaments also count which make a big difference to low performance players. The data-set we use, unfortunately, does not include Challenger matches as a result the PageRank system uses only ATP Tour tournaments to generate rankings.

Another possible reason is the fact that tournaments are seeded. Seeded tournaments create a bias which works in the advantage of top players. Since worse ranked players are matched up with the top players in the first rounds of the tournaments, it makes it very difficult for those players to proceed in the latter rounds

to gain the points needed to climb the rankings. This also means that worse ranked players, play with top players in the majority of their matches and have fewer matches with players of their own calibre. Because very few matches exist in the ATP Tour with players ranked greater than 64 facing each other, it is hard to compare the performance between them and as a result the generated rankings vary greatly. Since in most tournaments either the top 16 or top 32 players are seeded, these players face-off relatively frequently. This is the reason why the Match and Set PageRank systems are in rough agreement with the ATP for the rankings of the Top 32 players. From an ATP Rank of 32 onwards the disagreement increases.

An example of how using Matches, Sets and Games to generate PageRankings affects the result is Horacio Zeballos. In Match PageRank he was ranked in position 27, in Set PageRank he was ranked in position 55 and in Game PageRank in position 67. It so happens, that Horacio Zeballos managed to achieve a victory against Rafael Nadal during the final of Vina del Mar, Chile in early 2013. This is the reason he was boosted up to position 27 in the Match PageRank system. ATP Ranked him in position 56 at the end of 2013, a ranking very similar to the Set PageRank system (position 55). The reason Set PageRank gives a worse ranking than Match PageRank is that, because of the increased number of incoming and outgoing links in the network, the 2 sets that Horacio won against Nadal have shrunk in significance. Also, Horacio conceded a set to Rafael during that same match. Game PageRank ranks Horacio even lower for the same reasons as the Set PageRank, i.e. in a greater data-set, a single victory against a top player loses significance. The actual score of the match was 6-7(2), 7-6(6), 6-4. A tight victory where the number of games almost balance out (with only 2 games in favour of Zeballos). This further decreases the importance of this victory. This example demonstrates the trade-offs of using more and more granular data to generate PageRankings. It is our opinion that the best ranking system out of the 4 PageRank systems introduced here is the Set PageRank. It expands the data-set without greatly impacting the importance of victories and without penalising players for not playing a lot of matches.

Another example is Lleyton Hewitt who appears to be ranked lower than he deserves by the ATP Rankings. Investigating further into the history of the matches he played in 2013, Hewitt has achieved a number of victories over top 20 ATP Ranked players. In fact out of the 24 victories he had over the year, 7 of them were against players ranked in the top 20. The ATP ranks Hewitt lower than he deserves merely because he did not proceed to the latter stages of tournaments and this does

not reflect the player's true skill but rather the player's ability to play well within the rules of the ranking system.

Finally, to demonstrate how winning points does not necessarily translate to a higher quality player we use the example of David Ferrer. David Ferrer was ranked as number 3 by the ATP and number 1 by the Point PageRank system. The Set and Game PageRank systems also rank David Ferrer as number 3 and agree with the ATP. In the case of points, David Ferrer has won a total of 6822 points over 2013 but only 1146 games and 143 sets. When compared to Novak Djokovic, who has won 6483 points, 1133 games and 156 sets in the same period, it becomes clear that winning points does not mean winning sets and matches. Novak Djokovic is ranked as number 1 by Set and Game PageRank systems and number 2 by the Point PageRank system. David Ferrer was able to overtake Novak Djokovic in the point PageRank rankings because he has won more points against highly ranked players, thus getting a better prestige score.

It can be observed that in Figures 5.3 and 5.4, which compare Game and Point PageRank systems with the ATP Rankings, there is a greater amount of players which appear above the line. As discussed earlier, Game and Point PageRank systems are vulnerable to rank players higher just because these low ranked players receive PageRank contributions from top ranked players since they can win games and points against them but not necessarily sets and matches. The reason why more players are ranked worse than they should is because a lot of players, having a ranking greater than 100 by the ATP, are ranked in the Top 100 positions by Game and Point PageRank pushing the rest lower in the rankings.

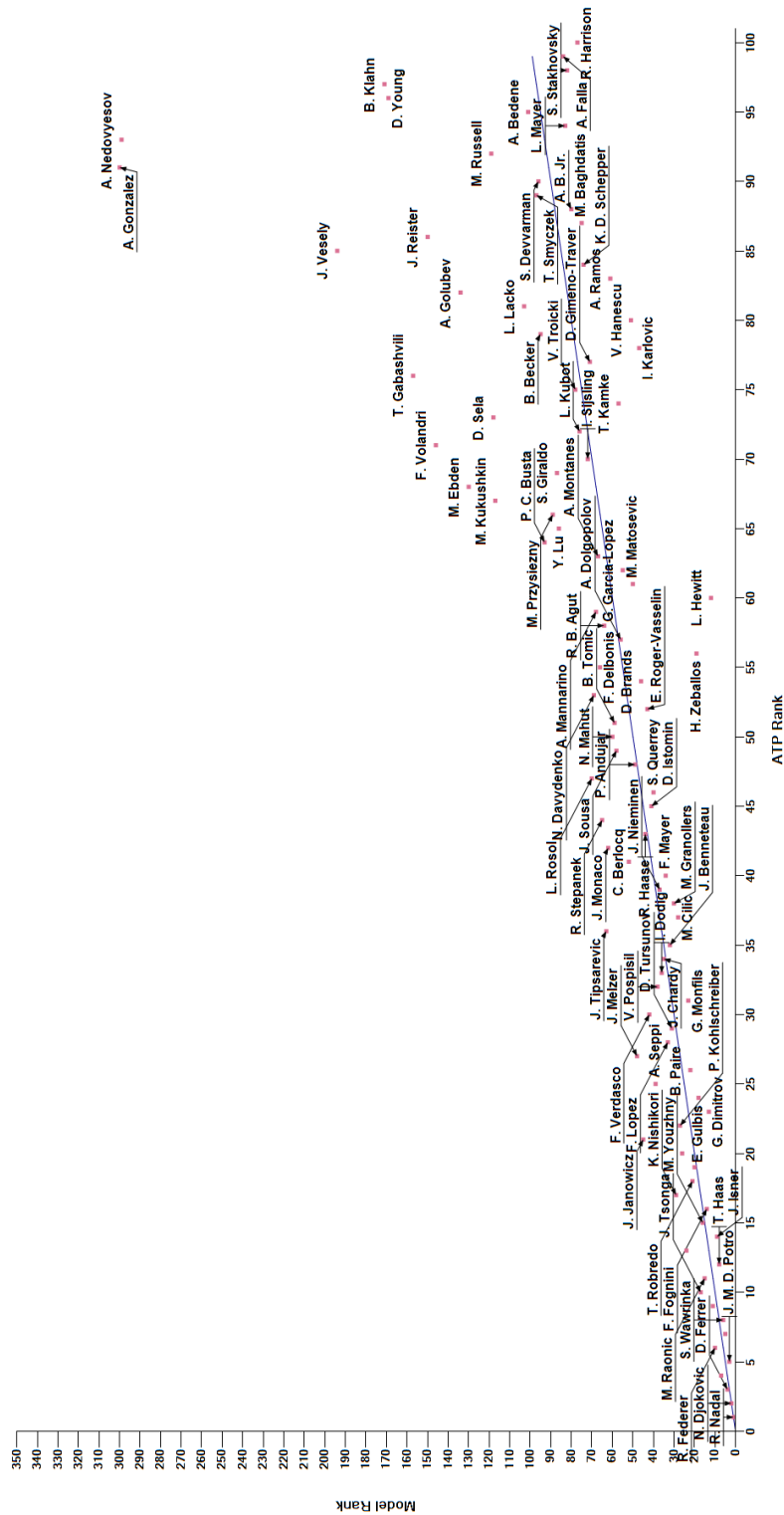


Figure 5.1.: The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the Match PageRank system.

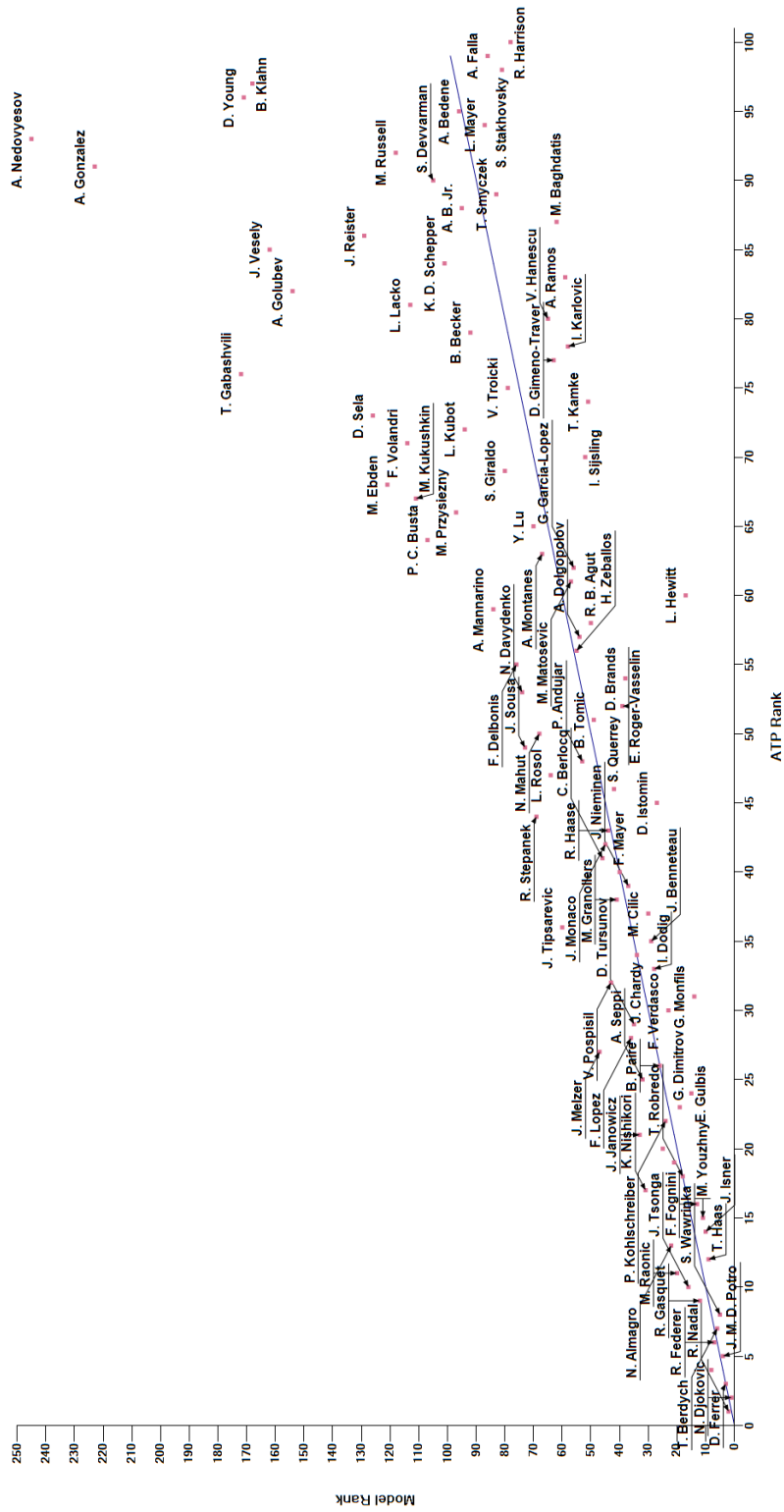


Figure 5.2.: The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the Set PageRank system.

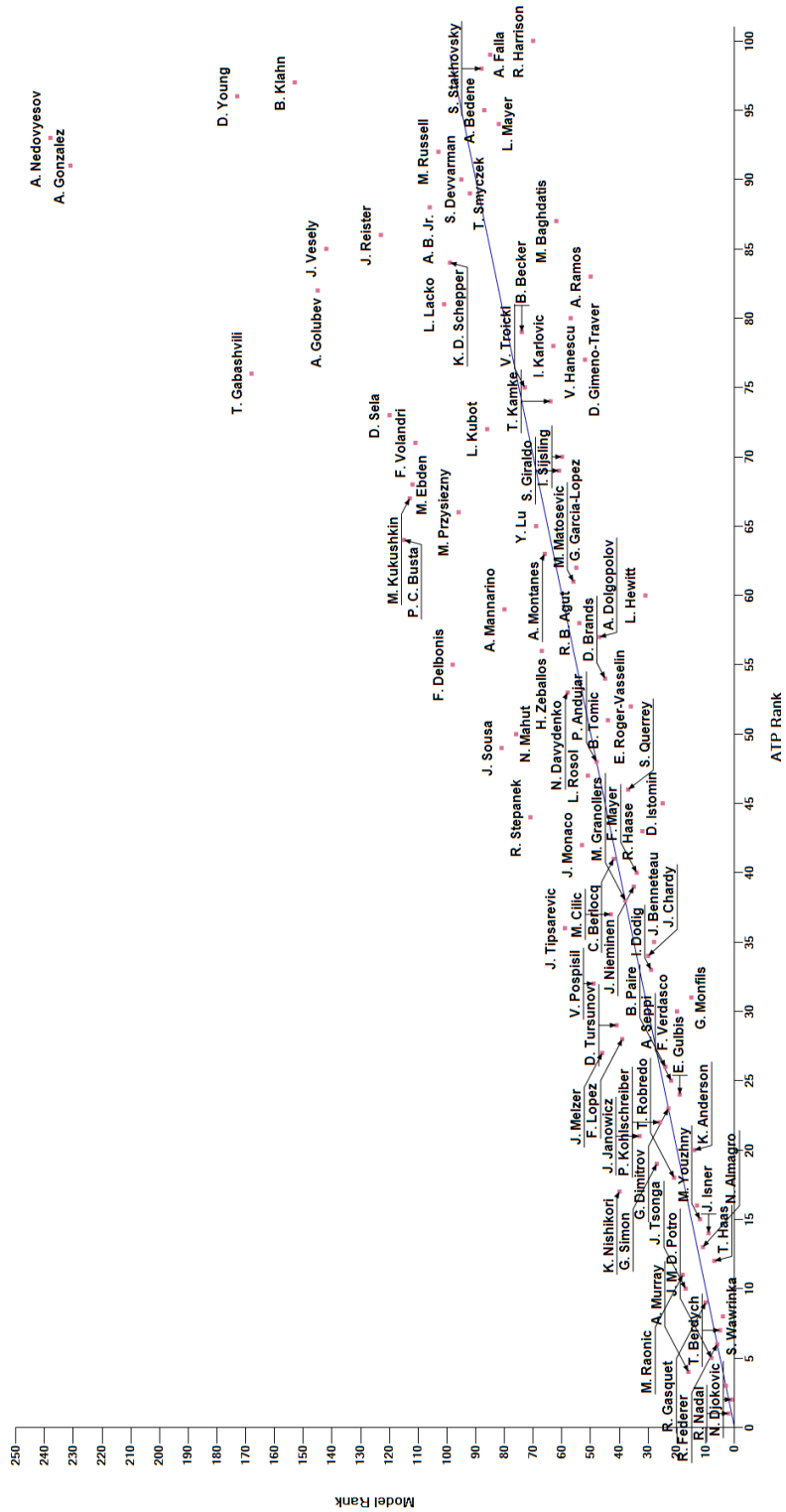


Figure 5.3.: The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the Game PageRank system.

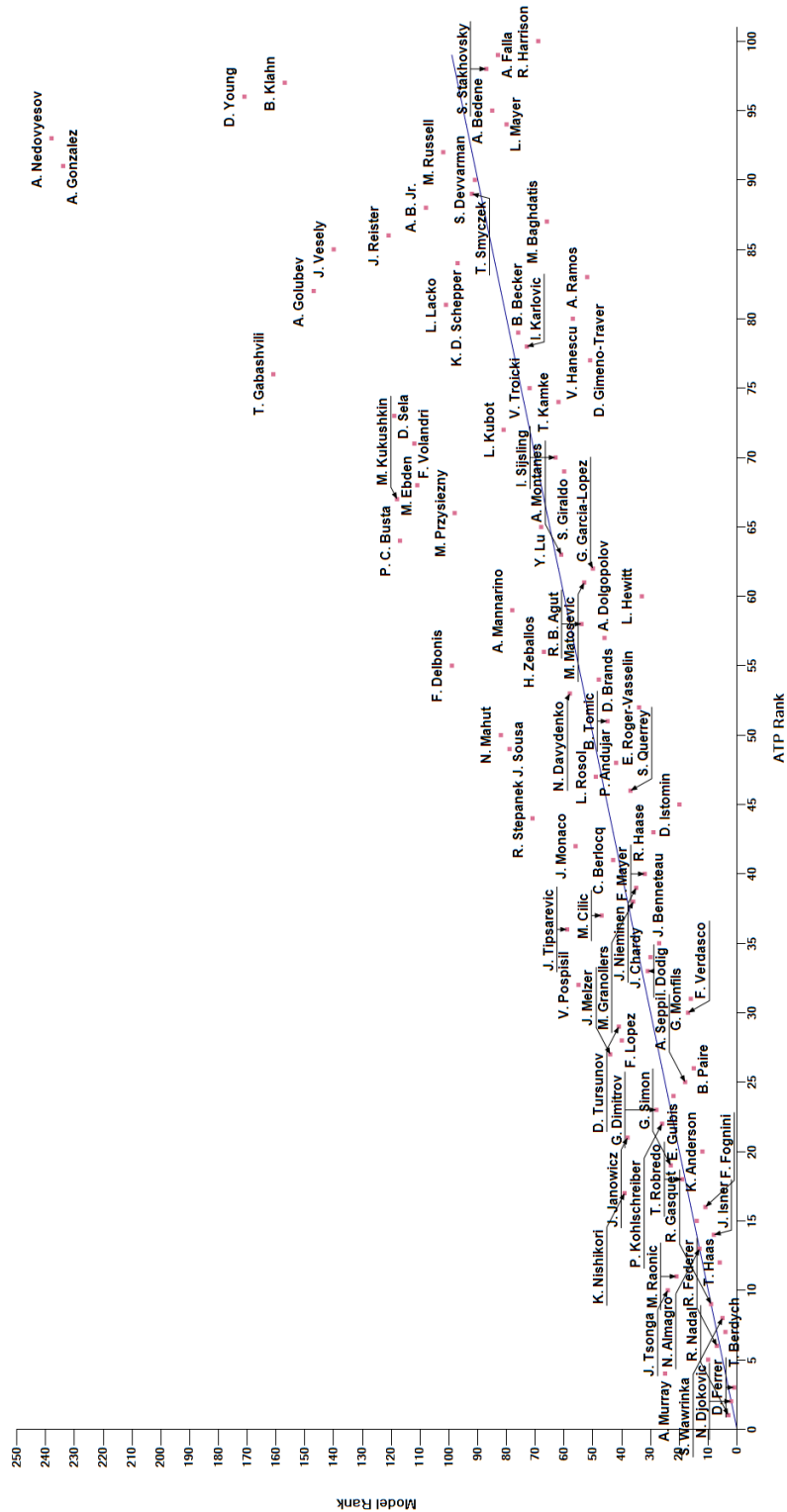


Figure 5.4.: The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the Point PageRank system.

5.5. SortRank

SortRank and LadderRank are two novel ideas which were born in an attempt to use knowledge gained from tennis models to generate player rankings. The concept behind SortRank is very simple: take any tennis model, convert it into a binary model and then use it as the comparison criterion of a sorting algorithm. For example: lets assume that we have a list of players to rank. A sorting algorithm such as QuickSort, as described by Hoare [101], can be used to sort this list of players by using a binary model which outputs a comparison criterion between players.

A limitation of any sorting algorithm is that it assumes absolute transitivity. This means that if Player A can beat Player B and Player B can beat Player C then it must always hold that Player A can beat Player C. As a consequence, any model that is used as the comparison criterion should also be absolutely transitive to ensure replicable results.

An example of a fully transitive model is the “uncombined” model mentioned in Chapter 3. This model is transitive by definition as the opponent is not taken into consideration when estimating the parameter of a player winning a point while serving. Therefore, the output of any probability from the model is always compared against the constant “average” player. This “uncombined” model can be converted into a binary model by using the resulting probability of Player A winning a match against Player B. If this probability is greater than 0.5 then the binary model returns “true”, if the probability is less than or equal to 0.5 then it returns “false”.

This binary model can be joined with any sorting algorithm to generate a ranking. For this to happen, the sorting algorithm, when comparing two players, A and B, should use the binary model as the comparison criterion. That is, if the binary model returns “true” for Player A winning a match against Player B, the sorting algorithm places Player A above Player B in the rankings. By completing the algorithm for the entire list of players, the end result is a sorted list of players based on their performance, with the best player at the top of the list – thus a ranking.

5.6. The LadderRank Algorithm

To overcome the limitation of absolute transitivity, SortRank was evolved to a new algorithm that does not assume the comparison criterion is absolutely transitive.

This algorithm is inspired by normal “sports-ladders”. In a “sports-ladder” there is an initial ranked list of players, and each of those players is allowed to challenge another player that is ranked up to X positions higher. If the challenger is victorious in the challenge, then he/she overtakes the player challenged and pushes everyone in-between one position down. The resulting algorithm is described by the pseudocode in Figure 5.5.

For this algorithm to function correctly it must be provided with these crucial variables: the ‘iterations’, the ‘challenge_offset’ and the ‘ranking_list’. To ensure complete ranking of the players the ‘iterations’ must always be larger than the number of players being ranked. The ‘challenge_offset’ defines the number of positions in the ranking list that any player is allowed to jump after any challenge. Finally the ‘ranking_list’ is the list of players ranked in an initial order.

```

for (int i = 0; i < iterations; i++) {
  foreach (current_player in ranking_list) {
    if (current_player.ranking > 0) {
      x = challenge_offset
      if (x < current_player.ranking) { x = current_player.ranking }
      for(int position = x; position > 0; position--) {
        PlayerA = PlayerWithRanking(current_player.ranking - position)
        PlayerB = current_player
        if (Compare(PlayerA, PlayerB) == false) {
          //if player A loses the match-up move player B
          //above A and push all players in-between 1 spot down
          MovePlayerToRanking(PlayerB, PlayerA.ranking)
          position = 0 //stop challenging
        }
      }
    }
  }
}

```

Figure 5.5.: LadderRank pseudocode

The function *PlayerWithRank*(integer) which appears in the algorithm retrieves the player which has the ranking provided by the integer parameter. The function *MovePlayerToRanking*(player, integer) changes the ranking of the player to the integer value provided and shifts all rankings of players which were between the

player and the new ranking by 1 position towards the direction of the players current ranking. For example in a list of three players, A, B and C ranked as 1, 2 and 3 respectively, the function `MovePlayerToRanking(C, 1)` will change the rankings of A, B and C to 2, 3, 1 respectively.

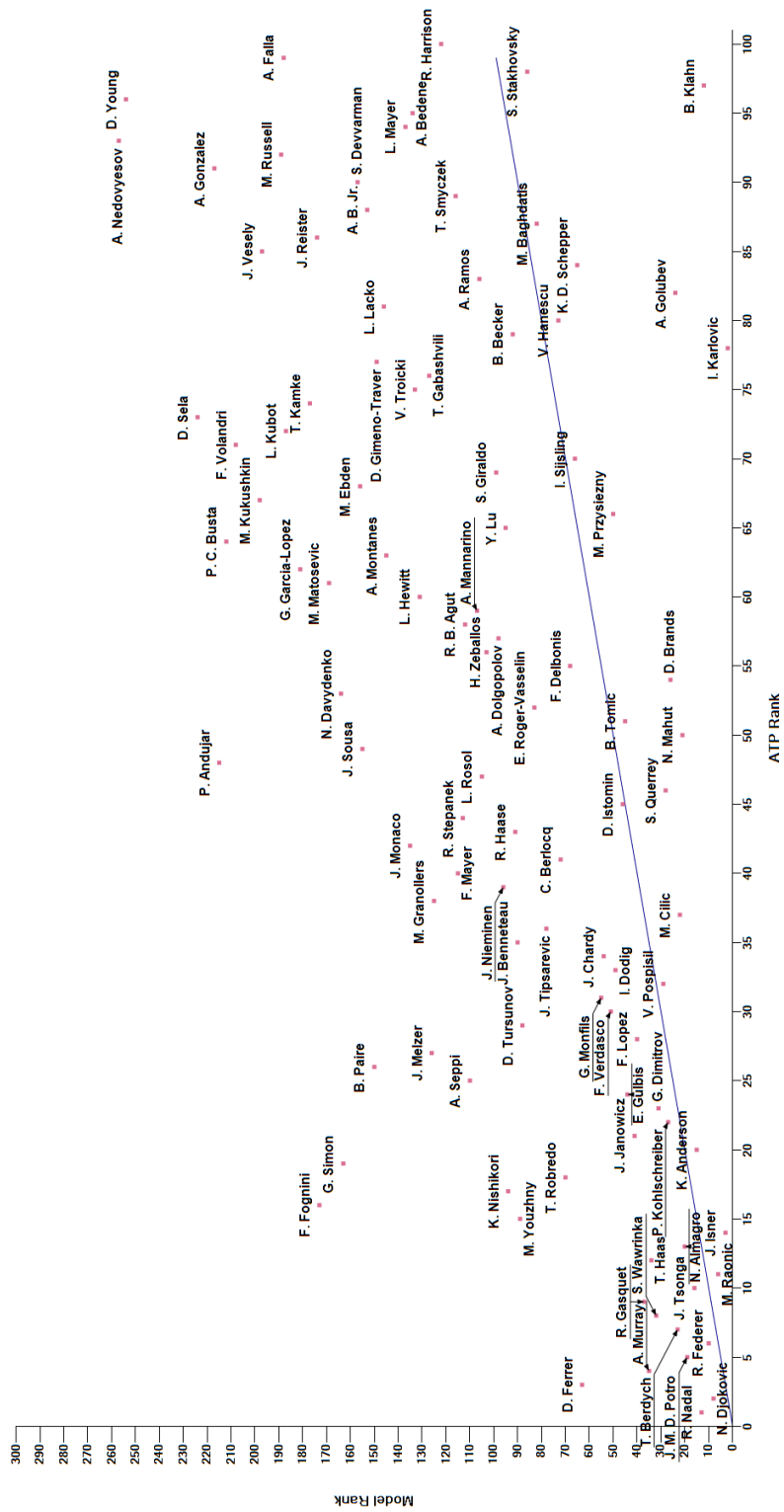
5.7. SortRank and LadderRank Performance

Using using the SortRank and LadderRank algorithms we generated a few variations of rankings for the years 2012 and 2013. We used the SortRank algorithm together with the only fully transitive model we described – the uncombined statistics model to generate rankings for both years. Figure 5.6 shows the rankings assigned by the SortRank ‘uncombined’ to the Top 100 ATP Players of 2013 using statistical averages over all the matches of the same year.

The LadderRank algorithm with parameter $X=1$ (i.e. the number of positions players are allowed to “move” above their rank is 1) was used in combination with the combined statistics model to generate rankings for years 2012 and 2013. Figure 5.7 shows the resulting rankings generated using this algorithm for the Top 100 players, as ranked by the ATP at the end of 2013. It is noticeable that, like the Point PageRank graph, there are a lot more players above the $Y=X$ line than below. Also, they tend to be further away from the $Y=X$ line, the further along we are on the ATP rankings axis. This, as discussed in an earlier section, is a sign that poorly ranked players, by the ATP, tend to climb the rankings generated by the algorithm. Further investigation revealed that players who had limited matches during the year and performed well in those matches resulted in very good statistical averages and were interpreted by the ‘combined’ model as high performance players. Once again, the problem of small samples resulting in averaged statistical data which do not represent the population means adequately, was the source of this problem.

Increasing the number of positions players were able to challenge above their ranking did not solve this problem. Figure 5.8 demonstrates how the rankings, generated for those same players, change when the parameter X is increased to 3. We can observe that some players have moved closer to the $Y=X$ line but others in fact moved further away. Again, the overall tendency is for more players to appear above the $Y=X$ line.

Finally, Figure 5.9 shows the rankings assigned to the Top 100 ATP Players using LadderRank with parameter $X=3$ and the ‘combined’ model, while removing players with 4 or fewer matches played in 2013 (i.e. having parameter $D=5$). By



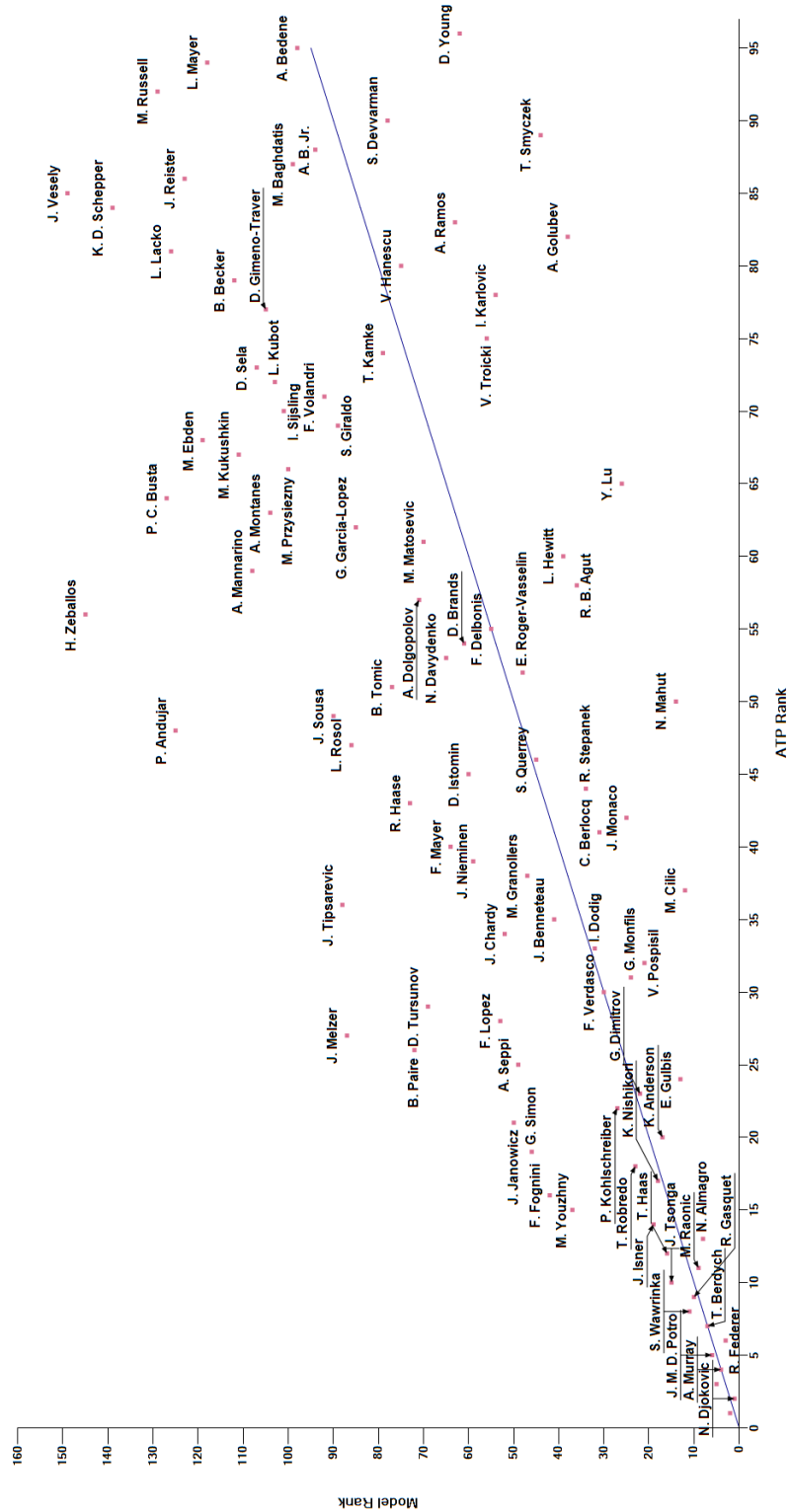


Figure 5.9.: The rankings of the Top 100 ATP players at the end of 2013 compared to their ranking generated using the LadderSort Combined system with $X=3$ and minimum of 5 matches played.

removing these players, we eliminate ‘*outliers*’ who had high performances in a few matches and were highly ranked by the ‘combined’ model. As a result, we observe a much more condensed graph with players being much closer to the $Y=X$ line. Also, players are now more evenly spread above and below the $Y=X$ line.

To understand how the different ranking systems really perform, we tested them by counting the percentage of matches in the year 2012 whose winners were actually better ranked by the system than the losers of the match. We generated rankings for the year 2012 using matches from the same year, and then compared how well those rankings represented the same set of matches.

Table 5.3.: Comparing different SortRank and LadderRank tennis ranking approaches to the official ATP Tour rankings as they were on 31/12/2012. The rankings contain players who participated in ATP Tour matches over the year of 2012.

Ranking System	Match Attempts	Success Percentage	Players	Standard Error	p-value from ATP
ATP	2538	69.661	300	0.009125	-
SR-Uncom.	2673	68.575	303	0.008979	0.80198
LR-Com. X=1	2673	72.204	303	0.008665	0.02168
LR-Com. X=3	2673	72.241	303	0.008662	0.02016
LR-Com. X=3 D=5	2470	71.457	177	0.009087	0.08153

Table 5.3 shows the results of these tests. In the case of the ATP Official rankings at the end 2012, out of the 2538 matches for which we had rankings for both players, a percentage of 69.7% had winners whose rankings were better than the losers. The SortRank ‘uncombined’ had a smaller percentage at 68.6% and the three variations of LadderRank ‘combined’ systems had higher success percentages.

Performing a split test comparison between the success percentage of the official ATP Rankings and the rest of the models, we discovered that the improvement of success percentage the LadderRank, with the ‘combined’ statistics model, achieves is, in fact, significant when there is no limitation on the number of matches played by the players. What is more, LadderRank ‘combined’ with $X=3$ is the best performing algorithm even when compared to the Set PageRank algorithm which was presented in Table 5.1. Even though LadderRank ‘combined’ appears to be performing better than the rest, a closer look at the rankings indicates a problem where low performing players appear in the top rankings simply because of outlier statistics. The requirement of having a minimum of 5 matches to be included in the

rankings seems to solve this problem and the rankings generated by it appears to be much more reasonable. This is evident in the tables presented in Appendix A. The Appendix also includes a closer look at the Top 32 players of 2013.

5.8. Forecasting Match Outcomes Using Ranking Systems

Comparing the ranking systems for their ability to represent the data sets, which were used to create them, does not necessarily reflect on their ability to predict new matches. One should expect worse results when predicting matches as the ranking systems will no longer be overfitted.

To check the performance of ranking systems in predicting matches we propose a simple approach. We will generate two rankings using the algorithms. One at the end of 2012 and one exactly 6 months after that. Using the first ranking we will attempt to predict the matches played in the first 6 months of 2013 and using the second ranking we will predict the final 6 months of 2013 and finally aggregate the results. This way reduce the risk of over-fitting and we gain an understanding of how successful the ranking systems are at predicting match results. Another approach, which we reserve for future research, would be to generate weekly rankings and predict the following week's matches over the entire year.

The predictions we generate do not involve a probability but only a binary value (i.e. the player with the better ranking is the winner). Such a probability can be retrieved from rankings using a simple logit model with a method similar to the one described by Clarke and Dyte [46]. This method is discussed in the subsequent section.

Table 5.4 presents the results of all rankings algorithms presented in this chapter. In this table, PR abbreviates PageRank, QSU abbreviates QuickSort Uncombined, and LRC abbreviates LadderRank Combined. X and D are the LadderRank variables as described in Section 5.6. The first row of the table shows the number of matches available for 2013, the second row shows the number of matches that were tested for prediction, the third row shows the number of successful predictions which were made by the algorithm, the fourth row shows the Standard Error calculated using the Bernoulli approach and finally the last row shows the percentage of success of the ranking algorithms.

Some matches were skipped because of missing player rankings. Our database holds only the Top 300 players of the ATP Rankings and we can only generate rankings for players who have had activity 12 months prior to the date of the rank-

Table 5.4.: Prediction results of all presented ranking systems for 2669 matches played in year 2013.

	ATP	PR Match	PR Set	PR Game	PR Point	QSU	LRC X1	LRC X3	LRC X3 D5
Matches	2669	2669	2669	2669	2669	2669	2669	2669	2669
Tested	2458	2508	2508	2508	2508	2508	2508	2508	2366
Successes	1596	1696	1696	1633	1618	1594	1664	1653	1555
Stand. Error	9.6249E-3	9.3433E-3	9.3433E-3	9.5171E-3	9.5542E-3	9.6101E-3	9.4353E-3	9.4652E-3	9.7578E-3
Percentage %	64.931	67.624	67.624	65.112	64.514	63.557	66.348	65.909	65.723

Table 5.5.: P-Values using two-sample Z-Tests between the prediction success rates of all ranking systems. All significant results using the 95% confidence level are marked in bold.

p-values	ATP	PRM	PRS	PRG	PRP	QSU	LRCX1	LRCX3	LRCX3D5
ATP	0.5	2.235E-2	2.235E-2	0.4469	0.6208	0.8438	0.1466	0.2343	0.2817
PRM	0.9776	0.5	0.5	0.9702	0.9900	0.9988	0.8317	0.9013	0.9203
PRS	0.9776	0.5	0.5	0.9702	0.9900	0.9988	0.8317	0.9013	0.9203
PRG	0.5531	2.982E-2	2.982E-2	0.5	0.6713	0.8749	0.1782	0.2762	0.3270
PRP	0.3792	9.975E-3	9.975E-3	0.3287	0.5	0.7600	8.598E-2	0.1497	0.1880
QSU	0.1562	1.205E-3	1.205E-3	0.1251	0.2400	0.5	1.911E-2	4.0576E-2	5.687E-2
LRCX1	0.8534	0.1683	0.1683	0.8218	0.9140	0.9809	0.5	0.6286	0.6774
LRCX3	0.7657	9.868E-2	9.868E-2	0.7238	0.8503	0.9594	0.3714	0.5	0.5545
LRCX3D5	0.7183	7.971E-2	7.971E-2	0.6730	0.8120	0.9431	0.3226	0.4455	0.5

ings, therefore not all players have an available ranking.

The results in Table 5.4 show that the best performing algorithms in terms of successfully predicting match results are the PageRank Match and Set algorithms with 67.624% over 2508 tested matches. The ATP scores the second lowest percentage in successful predictions with 64.931%. To analyse whether these differences in success percentages are actually significant statistically, we analyse the results using two-sample Z-tests. The p-values of these tests are presented in Table 5.5 where all algorithms are compared to each other. The significant results are highlighted with bold lettering.

It is evident from Table 5.5, that PageRank Match and Set are significantly better at predicting match results than the ATP Rankings, PageRank Game and Point and QuickSort Uncombined. Also, LadderRank Combined with X=1 and X=3 performs significantly better from the QuickSort Uncombined approach. These results provide further evidence that the PageRank Set algorithm represents player performance better than the ATP Official Rankings.

5.9. Match Probabilities from Rankings

In their paper, Clarke and Dyte [46] discuss how a logit model can be used to estimate the probability of a player winning a tennis match against another player using their official ATP Ranking difference. In this section, we apply the same technique to the PageRank Set rankings.

Firstly, we generated a PageRank Set ranking at the end of 2012. Then we aggregated, over all matches in 2012, the number of wins the best ranked player has for each difference in rankings to find the probability of the best ranked player to win given a particular difference in ranking with his opponent. Having all the required data all that is left is to fit the logit model (shown in Equation 5.4). To model the probability of the best ranked player winning a match, p , given a difference in ranking, x , the log of the odds ratio (logit) of p can be modelled linearly as follows.

$$\ln\left(\frac{p}{1-p}\right) = a + bx \quad (5.4)$$

For an equal ranking (therefore a difference, $x = 0$), p should be equal to 0.5. This directly gives an answer for the parameter $a = 0$, since $\ln(0.5/0.5) = a$.

Therefore,

$$p = \frac{1}{1 + e^{-bx}} \quad (5.5)$$

A value for b was fitted using Microsoft Solver and Microsoft Excel by maximising the log-likelihood between the real 2012 data and the estimated values of the logit model. This resulted in the value of $b = 0.010808$ which completes the model. For any difference in PageRank Set Rankings, x , the probability that the best ranked player will emerge victorious, p , is $\frac{1}{1 + e^{-0.010808x}}$.

5.10. PageRank Set rankings vs. Bookmakers

Using the same technique to develop the forecasts of match outcomes in Section 5.8, we will now simulate 1 unit match bets against the best pre-match odds of 5 different bookmakers (Bet 365, Expekt, Pinnacle Sports, Ladbrokes and Stan James). Additionally, we will simulate the same bets against the individual over-round-corrected odds of Bet 365, Expekt, Pinnacle Sports and Ladbrokes.

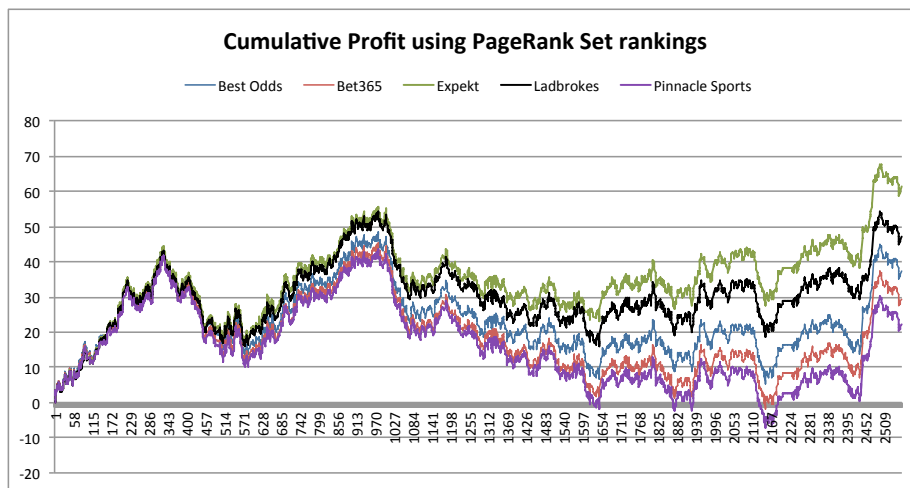


Figure 5.10.: The cumulative profit of 2364 bets with exposure 1 unit each, against the best match opening odds from 5 bookmakers and against the over-round-corrected odds of 4 individual bookmakers over 2562 ATP Tour matches in the year 2013.

Figure 5.10 shows the cumulative profit for each of these simulations. Approximately 2360 bets were placed in each simulation depending on the availability of the odds of each bookmaker. It can be observed that all simulations return a

positive return on investment (ROI). The results range from 0.9% ROI (Pinnacle Sports) to 2.6% ROI (Expekt), something which, as discussed in the previous chapters, can be attributed to the variation of the profit margins of the bookmakers.

These results serve as strong evidence towards the idea that PageRank Set rankings are as effective in predicting tennis results as professional models which are used by bookmakers.

5.11. Conclusion

In this chapter, we presented four different variations of Google's PageRank algorithm for generating professional tennis player rankings. By substituting web-pages by players and outgoing links by matches lost (as originally demonstrated by Radicchi), we were able to convert the original PageRank algorithm to a ranking system for professional tennis players. Radicchi's [8] PageRank algorithm for tennis was examined and modified to create a more efficient algorithm and then tested using Sets, Games and Points lost as outgoing links between players. We found that when using Sets lost as the weights of the PageRank edges, the final rankings system is a better alternative to Matches since it does not alter how well player ability is represented but it increases the data-set upon which the rankings are based. As a result of the increased data-set, there is a significant reduction in the number of ties at the bottom of the rankings.

We also introduced a novel and flexible concept for ranking professional tennis players which can convert any tennis match model to a ranking. This algorithm, which we called 'LadderRank', simulates a common 'sports-ladder' using any tennis model to determine the winner of a contest. We found that, using the 'combined' model introduced in Chapter 3, with 'LadderRank' we get comparable performance when it comes to player ability representation as in the PageRank Set approach but not as good a performance on match prediction.

We further tested the prediction capabilities of all presented ranking algorithms and found that the PageRank Match and Set algorithms have significantly better performance at predicting matches than the ATP Official Rankings whereas all other algorithms failed to show any significant improvement over the ATP Rankings.

6. Conclusions and Further Research

6.1. Achievements

This thesis has addressed the problem of modelling and predicting professional tennis matches. The problem has been approached using three distinct methods: a hierarchical Markov Chain approach, a novel transitive approach and via the generation of player rankings.

6.1.1. Tennis Point Markov Model

The first model that was introduced achieved the expansion of current hierarchical models. Hierarchical models in literature have been using the probability of winning a point while serving as a parameter to a game model which in turn was the parameter to a set model and finally, that was the parameter to a match model.

The problem of calculating this probability of winning a point on serve was scarcely addressed in the literature. The first part of our research was focused on exactly this problem. We have therefore contributed a Markov Chain model of a tennis point and we have shown how to adapt publicly available statistical data to calculate this probability.

We have tested this model using thousands of real world matches and found that the final model produced, which we named the ‘combined’ model, performs as well as models which have been deployed to the industry. We also investigated the effect of using different subsets of statistical data on the results and identified the problems of using averaged statistical data.

6.1.2. Common-Opponent Model

Most models in the literature make use of averaged statistics over various classes of opponents to estimate the probability of winning points while serving. Some exceptions address the problem of varying skill levels of opponents by adding weights to the statistics.

We wanted to contribute a new approach to modelling tennis matches which also solved the problems of using averaged statistical data across different skill levels of opponents. Our novel model exploits the transitive component inherent in tennis, by using the probability of winning service and return points against common opponents alone. Finding the differences that the two modelled players have in those probabilities, we combine them to get the probabilities of the players winning a match between them.

This rather unorthodox approach was tested against thousands of real data matches and was found to perform as well as other approaches when there is a sufficient amount of background information. One of the problems which was apparent with this Common-Opponent model was a diminishing set of background matches.

6.1.3. Ranking Systems and Forecasting

A selection of publications address the problem of forecasting tennis results using traditional or generated rankings together with logit and probit models. For the final chapter of our research we turned to the investigation of tennis rankings and their ability to predict tennis match results.

During our research we identified a problem with the existing PageRank tennis ranking system which was published by Radicchi [8]. The system would tie at the bottom of the rankings all players who had no victories over the period which was modelled. We therefore proposed the alternative of using Set losses as the weights in the PageRank algorithm, something which improved on the problem without affecting the quality of the rankings. We also experimented with using more granular scoring as weights, such as games and points lost, but determined these to deteriorate the quality of the rankings.

We also introduced an algorithm, based on the traditional sports-ladder, which enables one to convert any predictive tennis model to a ranking. We named this the ‘LadderRank’ algorithm. Using the ‘LadderRank’ algorithm in combination with the ‘combined’ tennis model we developed, we were able to generate rankings which are of better quality compared to the official ATP Ranking system.

Using the ranking systems described, we tested their ability to predict match results, using real ATP matches from 2013 and found that the PageRank Match and Set systems significantly outperform the official ATP Rankings. In fact they were able to achieve a successful prediction percentage similar to the ‘combined’ model.

6.2. Applications

This thesis presents research dedicated to the prediction of professional tennis match results. For this purpose we developed various quantitative models as well as ranking systems.

The prediction of tennis results is directly applicable to investors seeking financial gain from sports betting exchanges. By making informed decisions based on model predictions, investors can exploit market inefficiencies for a profit. Both the quantitative models and rankings are useful for this purpose.

The ‘combined’ and Common-Opponent models both provide probabilities of tennis players winning matches. These can be of use to bookmakers in the process of generating the odds they offer to punters. These same probabilities can be used to detect foul play, by detecting extreme upsets. Probabilities of winning matches may also be calculated from the ranking algorithms we presented using a simple logit model as presented in Clarke and Dyte [46].

The ‘combined’ model uses a point model to estimate the outcome of points. Players’ coaches may use this model to assess how a change in a player’s statistic will affect the probability of winning matches against particular opponents. Therefore it is applicable to deciding training and playing strategies for players.

The analysis provided by the ‘combined’ model may also be used to find the most likely score-lines, giving broadcasting stations an idea on the duration of matches. Tennis commentators may use the model during an in-play match to make informed comments on the match outcome as well as how a player can change strategy to affect the game.

The ‘combined’ model, employing a point model, may also be used in the decision process of systems analysing video and audio feeds for automated annotation in tennis, improving them by using probabilities retrieved from historical statistics.

The ranking systems themselves are applicable to quick comparison of player performance as their quality is superior to official ATP Rankings. This does not suggest, though, that they may be used to replace the ATP Rankings. ATP Rankings are simple and understandable to the general public and to the players. Complicated systems such as PageRank or LadderRank may cause disputes and are also eligible to exploitation by players. Also the ranking systems presented in this dissertation are designed to measure player ability alone. Other factors, such as the ability to win prestigious tournaments, are not included. These factors are part of what makes the sport interesting to the public.

6.3. Further Research

Future research based on the algorithms and models presented in this thesis may further improve their performance and increase their effectiveness in the aforementioned applications. This section is a discussion of possible directions the research may take in the future.

6.3.1. Analysis of Data

Quantitative models for tennis have been the focus of research in the field for many years. Little research has been devoted, on the other hand, to analysing historical data in tennis. The results of quantitative models are very much dependent on the quality of the input parameters.

It is therefore prudent to analyse the historical data which is used to generate those parameters to ensure more accurate estimation of their true values. An example of this would be an investigation of the effect of using weighted averages of match statistics on the performance of models. The weights of matches could be generated in various methods: exponentially decaying the importance of the statistics in time, or according to the relevance of the surface on which the match was played on, or even the difference in ranking of the two players in the match.

Giving greater weight to statistics of matches which are more relevant to the match being modelled may increase the performance of the model using them, whether that is the ‘uncombined’, ‘combined’ or even the Common-Opponent model.

6.3.2. Application on Doubles Matches

Section 3.7.4 has introduced the possible application of the ‘combined’ model in modelling doubles tennis matches. This was merely an introduction and further research can be undertaken towards that direction. Additionally, this research will need to be backed up by evidence using a similar approach of back-testing with real doubles matches.

6.3.3. Analysis of Rallies and Serving

As more and more detailed historical data become available to the public, research may be undertaken regarding the effectiveness of further analysing a tennis point. This can be done by deeper analysis of rallies and serving. Section 3.6 hints towards possible Markov Chains that can be used to estimate the probabilities of

various outcomes of serve as well as the two outcomes of a rally. In the future, it may become possible to test these theories and further develop the point model.

6.3.4. Back-testing with WTA Matches

This thesis has not tested the applicability of our models in women’s professional tennis. In the future, retrieval of a WTA match database with statistics will allow an investigation on the performance that can be achieved in women’s tennis by our introduced models and rankings.

6.3.5. 2-tier Common-Opponent Model

One of the problems that was mentioned with the Common-Opponent model was the reduction in available historical data due to the strict use of relevant matches. A possible solution to this problem would be the expansion of the Common-Opponent model to a recursive Common-Opponent model.

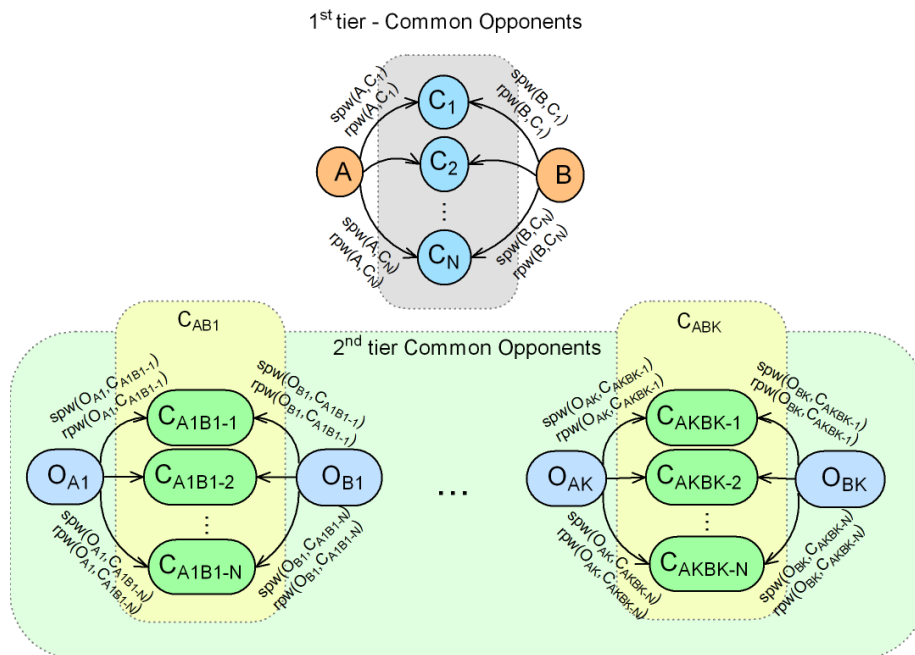


Figure 6.1.: Two tier common opponent example.

There is a variety of ways to do this and they all justify research time. In this section, we will propose one possible method as an example. Figure 6.1 demonstrates how one may expand the Common-Opponent model to include a second

level of common opponents. In this example, player O_{A1} is player A's opponent which has not been faced by player B but has common opponents with player O_{B1} which has been faced by player B. This is repeated for all players player A has faced but have not been faced by player B. This creates a second tier of common opponents which can be related via two level transitive results. What we mean by two level transitive results is best demonstrated by an example: if player A can beat player O_{A1} who can beat player O_{B1} and player O_{B1} can beat player B then player A can beat player B. Of course this is not always true but some component of it exists in tennis. It is this component which the 2-tier Common-Opponent model proposed will attempt to exploit.

Further research is needed to understand how to relate second tier opponents to the players being modelled and whether adding even more tiers makes sense. This method has potential to increase the data-set, which is available, greatly but testing needs to be done to investigate whether this will diminish the model's performance.

6.3.6. Expand Data-set to Include Challenger Data

In Chapter 5, we mentioned that our database only includes matches played in the ATP tour. Official ATP Rankings use points retrieved from matches played in the Challenger and Futures tournaments which are not included in our database and also not used to generate rankings using the PageRank systems and LadderRank algorithms. It would be interesting to include these matches in our database and generate new rankings for comparison to the ATP Rankings. This will most likely increase the fidelity in the rankings of poorly ranked players.

Bibliography

- [1] The Jakarta Post, “Sports Trading Club announces record 61% trading profit for the first quarter of 2014.” <http://prnw.cbe.thejakartapost.com/news/2014/sports-trading-club-announces-record-61-trading-profit-for-the-first-quarter-of-2014-2.html>, April 2014. Accessed on: 18/09/2014.
- [2] WagerMinds, “Centaur Galileo sports betting hedge fund collapses.” <http://www.wagerminds.com/blog/uncategorized/centaur-galileo-sports-betting-hedge-fund-collapses-3944/>, January 2012. Accessed on: 18/09/2014.
- [3] Priomha Capital PTY LTD, “Information Document - Fund I.” http://www.priomha.com/pdf/Priomha_Capital_Information_JANUARY_2013.pdf, January 2013. Accessed on: 18/09/2014.
- [4] D. Spanias and W. Knottenbelt, “Quantitative modelling of singles and doubles tennis matches,” in *Proceedings of the 3rd IMA International Conference on Mathematics in Sport*, 2011.
- [5] D. Spanias and W. J. Knottenbelt, “Predicting the outcomes of tennis matches using a low-level point model,” *IMA Journal of Management Mathematics*, 2012.
- [6] W. J. Knottenbelt, D. Spanias, and A. M. Madurska, “A common-opponent stochastic model for predicting the outcome of professional tennis matches,” *Computers & Mathematics with Applications*, vol. 64, no. 12, pp. 3820–3827, 2012. Theory and Practice of Stochastic Modeling.
- [7] N. Dingle, W. Knottenbelt, and D. Spanias, “On the (page) ranking of professional tennis players,” in *Computer Performance Engineering* (M. Tribastone and S. Gilmore, eds.), vol. 7587 of *Lecture Notes in Computer Science*, pp. 237–247, Springer Berlin Heidelberg, 2013.

- [8] F. Radicchi and M. Perc, “Who is the best player ever? A complex network analysis of the history of professional tennis,” *PLoS ONE*, vol. 6, no. 2, p. e17249, 2011.
- [9] D. Spanias and W. Knottenbelt, “Tennis player ranking using quantitative models,” in *Proceedings of the 4th International Conference on Mathematics in Sport*, 2013.
- [10] G. I. Dumitrescu, X. Huang, W. Knottenbelt, D. Spanias, and J. Wozniak, “Inferring the score of a tennis match from in-play betting exchange markets,” in *Proceedings of the 4th International Conference on Mathematics in Sport*, 2013.
- [11] ATP Tour, Inc, “The 2014 ATP Official Rulebook,” 2014.
- [12] International Tennis Federation, “ITF rules of tennis 2014,” 2014.
- [13] Grand Slam Board, “2014 Official Grand Slam Rule Book,” 2014.
- [14] P. S. K. Falko Bause, *Stochastic Petri Nets*. Friedrich Vieweg & Sohn Verlag, 2002.
- [15] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill series in electrical engineering: Communications and signal processing, Tata McGraw-Hill, 2002.
- [16] J. Rice, *Mathematical Statistics and Data Analysis*. Advanced series, Cengage Learning, 2006.
- [17] J. G. Kemeny and J. L. Snell, *Finite Markov chains*. Van Nostrand Publishing Company, 1960.
- [18] E. Y. Rodin, “Modular applied mathematics for beginning students,” *American Mathematical Monthly*, pp. 555–560, 1977.
- [19] R. Noubary, “Teaching mathematics and statistics using tennis,” *Mathematics and Sports*, no. 43, p. 239, 2010.
- [20] B. P. Hsi and D. M. Burych, “Games of two players,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 20, no. 1, pp. 86–92, 1971.

- [21] J. Carter, Walter H. and S. L. Crews, "An analysis of the game of tennis," *The American Statistician*, vol. 28, no. 4, pp. 130–134, 1974.
- [22] G. Fischer, "Exercise in probability and statistics, or the probability of winning at tennis," *American Journal of Physics*, vol. 48, no. 1, pp. 14–19, 1980.
- [23] R. E. Miles, "Symmetric sequential analysis: The efficiencies of sports scoring systems (with particular reference to those of tennis)," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 46, no. 1, pp. 93–108, 1984.
- [24] J. S. Croucher, "The effect of the tennis tie-breaker," *Research Quarterly for Exercise and Sport*, vol. 53, no. 4, pp. 336–339, 1982.
- [25] J. S. Croucher, "The conditional probability of winning games of tennis," *Research Quarterly for Exercise and Sport*, vol. 57, no. 1, pp. 23–26, 1986.
- [26] C. Morris, "The most important points in tennis," *Optimal Strategies in Sport*, pp. 131–140, 1977.
- [27] L. H. Riddle, "Probability models for tennis scoring systems," *Applied Statistics*, pp. 63–75, 1988.
- [28] Y. Liu, "Random walks in tennis," *Missouri Journal of Mathematical Sciences*, vol. 13, no. 3, 2001.
- [29] F. J. Klaassen and J. R. Magnus, *Forecasting in tennis*. Tennis Science and Technology 2., 2003.
- [30] S. Easton and K. Uylangco, "Forecasting outcomes in tennis matches using within-match betting markets," *International Journal of Forecasting*, vol. 26, no. 3, pp. 564–575, 2010.
- [31] T. J. Barnett and S. R. Clarke, "Using Microsoft Excel to model a tennis match," *6th Australian Conference on Mathematics and Computers in Sport (G. Cohen ed.)*, pp. 63–68, 2002.
- [32] B. A. Barnett, Tristan and S. Clarke, "Developing a model that reflects outcomes of tennis matches," *Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport*, pp. 178–188, 2006.

- [33] T. Barnett and S. R. Clarke, “Combining player statistics to predict outcomes of tennis matches,” *IMA Journal of Management Mathematics*, vol. 16, no. 2, pp. 113–120, 2005.
- [34] P. K. Newton and J. B. Keller, “Probability of winning at tennis i. theory and data,” *Studies in Applied Mathematics*, vol. 114, no. 3, pp. 241–269, 2005.
- [35] A. J. O’Malley, “Probability formulas and statistical analysis in tennis,” *Journal of Quantitative Analysis in Sports*, vol. 4, no. 2, p. 15, 2008.
- [36] K. Newton Paul and A. Kamran, “Monte Carlo tennis: a stochastic Markov chain model,” *Journal of Quantitative Analysis in Sports*, vol. 5, no. 3, pp. 1–44, 2009.
- [37] D. Jackson and K. Mosurski, “Heavy defeats in tennis: Psychological momentum or random effect?,” *Chance*, vol. 10, no. 2, pp. 27–34, 1997.
- [38] F. Klaassen and J. R. Magnus, *Analyzing Wimbledon: The power of statistics*. Oxford University Press, 2014.
- [39] F. J. Klaassen and J. R. Magnus, “On the independence and identical distribution of points in tennis,” 1998.
- [40] F. J. G. M. Klaassen and J. R. Magnus, “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, pp. 500–509, 2001.
- [41] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998.
- [42] R. D. Baker and I. G. McHale, “A dynamic paired comparisons model: who is the greatest tennis player?,” *European Journal of Operational Research*, 2014.
- [43] H. Starn, “A continuum of paired comparisons models,” *Biometrika*, vol. 77, no. 2, pp. 265–273, 1990.
- [44] B. L. Boulier and H. Stekler, “Are sports seedings good predictors?: an evaluation,” *International Journal of Forecasting*, vol. 15, no. 1, pp. 83–91, 1999.

- [45] S. R. Clarke, "An adjustive rating system for tennis and squash players," in *2nd Conference on Mathematics and Computers in Sport*, pp. 43–50, 1994.
- [46] S. R. Clarke and D. Dyte, "Using official ratings to simulate major tennis tournaments," *International Transactions in Operational Research*, vol. 7, no. 6, pp. 585–594, 2000.
- [47] F. Klaassen and J. Magnus, "Forecasting the winner of a tennis match," *European Journal of Operational Research*, vol. 148, no. 2, pp. 257–267, 2003.
- [48] J. del Corral and J. Prieto-Rodriguez, "Are differences in ranks good predictors for grand slam tennis matches?," *International Journal of Forecasting*, vol. 26, no. 3, pp. 551–563, 2010.
- [49] I. McHale and A. Morton, "A Bradley-Terry type model for forecasting tennis match results," *International Journal of Forecasting*, vol. 27, no. 2, pp. 619–630, 2011.
- [50] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.
- [51] P. A. Richardson, W. Adler, and D. Hanks, "Game, set, match: Psychological momentum in tennis.," *Sport Psychologist*, vol. 2, no. 1, 1988.
- [52] V. De Bosscher, P. De Knop, and B. Heyndels, "Comparing tennis success among countries," *International Sports Studies*, vol. 25, no. 1, pp. 49–68, 2003.
- [53] A. Somboonphokkaphan, S. Phimoltares, and C. Lursinsap, "Tennis winner prediction based on time-series history with neural modeling," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2009.
- [54] P. A. Scarf and X. Shi, "The importance of a match in a tournament," *Computers & Operations Research*, vol. 35, no. 7, pp. 2406–2418, 2008.
- [55] N. Djurovic, V. Lozovina, and L. Pavicic, "Evaluation of tennis match data-new acquisition model," *Journal of Human Kinetics*, vol. 21, no. 1, pp. 15–21, 2009.

- [56] S.-M. Ma, C.-C. Liu, Y. Tan, and S.-C. Ma, “Winning matches in Grand Slam men’s singles: An analysis of player performance-related variables from 1991 to 2008,” *Journal of Sports Sciences*, vol. 31, no. 11, pp. 1147–1155, 2013.
- [57] K. F. Gilsdorf and V. A. Sukhatme, “Tournament incentives and match outcomes in women’s professional tennis,” *Applied Economics*, vol. 40, no. 18, pp. 2405–2412, 2008.
- [58] S. Rosen, “Prizes and incentives in elimination tournaments,” Working Paper 1668, National Bureau of Economic Research, July 1985.
- [59] C. Leitner, A. Zeileis, and K. Hornik, “Is Federer stronger in a tournament without Nadal? an evaluation of odds and seedings for Wimbledon 2009,” *Austrian Journal of Statistics*, vol. 38, no. 4, pp. 277–286, 2009.
- [60] D. A. Malueg and A. J. Yates, “Testing contest theory: evidence from best-of-three tennis matches,” *The Review of Economics and Statistics*, vol. 92, no. 3, pp. 689–692, 2010.
- [61] J. K. Vis, W. A. Kusters, and A. Terroba, “Tennis patterns: Player, match and beyond,” in *22nd Benelux Conference on Artificial Intelligence (BNAIC 2010)*, Luxembourg, pp. 25–26, 2010.
- [62] B. Scheibehenne and A. Bröder, “Predicting Wimbledon 2005 tennis results by mere player name recognition,” *International Journal of Forecasting*, vol. 23, no. 3, pp. 415–426, 2007.
- [63] S. M. Herzog and R. Hertwig, “The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition,” *Judgment and Decision Making*, vol. 6, no. 1, pp. 58–72, 2011.
- [64] A. M. Nevill, R. L. Holder, A. Bardsley, H. Calvert, and S. Jones, “Identifying home advantage in international tennis and golf tournaments,” *Journal of Sports Sciences*, vol. 15, no. 4, pp. 437–443, 1997. PMID: 9293420.
- [65] R. L. Holder and A. M. Nevill, “Modelling performance at international tennis and golf tournaments: Is there a home advantage?,” *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 46, no. 4, pp. pp. 551–559, 1997.

- [66] R. H. Koning, "Home advantage in professional tennis," *Journal of sports sciences*, vol. 29, no. 1, pp. 19–27, 2011.
- [67] G. Knight and P. O'Donoghue, "The probability of winning break points in grand slam men's singles tennis," *European Journal of Sport Science*, vol. 12, no. 6, pp. 462–468, 2012.
- [68] F. Klaassen and J. Magnus, "How to reduce the service dominance in tennis? empirical results from four years at Wimbledon," *Open Access publications from Tilburg University*, 2000.
- [69] C. Du Bois and B. Heyndels, "It's a different game you go to watch: competitive balance in men's and women's tennis," *European Sport Management Quarterly*, vol. 7, no. 2, pp. 167–185, 2007.
- [70] J. del Corral, "Competitive balance and match uncertainty in grand-slam tennis effects of seeding system, gender, and court surface," *Journal of Sports Economics*, vol. 10, no. 6, pp. 563–581, 2009.
- [71] U. Sunde, "Heterogeneity and performance in tournaments: a test for incentive effects using professional tennis data," *Applied Economics*, vol. 41, no. 25, pp. 3199–3208, 2009.
- [72] G. Halkos and N. Tzeremes, "Evaluating professional tennis players' career performance: A data envelopment analysis approach," 2012.
- [73] D. Gale, "Optimal strategy for serving in tennis," *Mathematics Magazine*, pp. 197–199, 1971.
- [74] S. L. George, "Optimal strategy in tennis: A simple probabilistic model," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 22, no. 1, pp. 97–104, 1973.
- [75] J. Norman, "Dynamic programming in tennis-when to use a fast serve," *Journal of the Operational Research Society*, pp. 75–77, 1985.
- [76] G. Pollard, "What is the best serving strategy," *J Med Sci Tennis*, vol. 13, no. 2, pp. 34–38, 2008.
- [77] G. Pollard, G. Pollard, T. Barnett, and J. Zeleznikow, "Applying tennis match statistics to increase serving performance during a match in

- progress,” *Journal of Medicine and Science in Tennis*, vol. 14, no. 3, pp. 16–19, 2009.
- [78] F. J. Klaassen and J. R. Magnus, “The efficiency of top agents: An analysis through service strategy in tennis,” *Journal of Econometrics*, vol. 148, no. 1, pp. 72–85, 2009.
- [79] P. O’Donoghue and B. Ingram, “A notational analysis of elite tennis strategy,” *Journal of sports sciences*, vol. 19, no. 2, pp. 107–115, 2001.
- [80] P. G. O’Donoghue, “The most important points in grand slam singles tennis,” *Research quarterly for exercise and sport*, vol. 72, no. 2, pp. 125–131, 2001.
- [81] C.-H. Chiu and S.-Y. Tsao, “Mathematical model for the optimal tennis placement and defense space,” *International Journal of Sport and Exercise Science*, vol. 4, no. 2, pp. 25–36, 2012.
- [82] G. Pollard, G. Pollard, T. Barnett, and J. Zeleznikow, “Applying strategies to the tennis challenge system,” *Medicine and Science in Tennis*, vol. 15, no. 1, pp. 12–15, 2010.
- [83] V. K. Nadimpalli and J. J. Hasenbein, “When to challenge a call in tennis: A markov decision process approach,” *Journal of Quantitative Analysis in Sports*, vol. 9, no. 3, pp. 229–238, 2013.
- [84] H. Brody, “Models of tennis racket impacts.,” *International Journal of sport biomechanics*, vol. 3, no. 3, 1987.
- [85] R. Cross, “A double pendulum model of tennis strokes,” *American Journal of Physics*, vol. 79, no. 5, pp. 470–476, 2011.
- [86] A. C. Cutmore and W. J. Knottenbelt, “Quantitative models for retirement risk in professional tennis,” Master’s thesis, Imperial College of Science, 2012.
- [87] G. Sudhir, J. C.-M. Lee, and A. K. Jain, “Automatic classification of tennis video for high-level content-based retrieval,” in *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pp. 81–90, IEEE, 1998.

- [88] M. Petkovic, W. Jonker, and Z. Zivkovic, "Recognizing strokes in tennis videos using hidden markov models.," in *VIIP*, pp. 512–516, 2001.
- [89] T. Bloom and A. P. Bradley, "Player tracking and stroke recognition in tennis video," in *APRS Workshop on Digital Image Computing (WDIC'03)*, vol. 1, pp. 93–97, The University of Queensland, 2003.
- [90] I. Kolonias, W. Christmas, and J. Kittler, "Automatic evolution tracking for tennis matches using an hmm-based architecture," in *Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*, pp. 615–624, IEEE, 2004.
- [91] W. Christmas, A. Kostin, F. Yan, I. Kolonias, and J. Kittler, "A system for the automatic annotation of tennis matches," *Report on the current state of the tasks and recommendations for future directions*, p. 135, 2005.
- [92] I. Kolonias, J. Kittler, W. Christmas, and F. Yan, "Improving the accuracy of automatic tennis video annotation by high level grammar," in *Image Analysis and Processing Workshops, 2007. ICIAPW 2007. 14th International Conference on*, pp. 154–159, IEEE, 2007.
- [93] G. Hunter, A. Shihab, and K. Zienowicz, "Modelling tennis rallies using information from both audio and video signals," in *Proceedings of the IMA International Conference on Mathematics in Sport*, pp. 103–108, Manchester, UK: The Institute of Mathematics and Its Applications, 2007.
- [94] G. J. Hunter, K. Zienowicz, and A. I. Shihab, "The use of mel cepstral coefficients and markov models for the automatic identification, classification and sequence modelling of salient sound events occurring during tennis matches," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3431–3431, 2008.
- [95] Y.-C. Jiang, K.-T. Lai, C.-H. Hsieh, and M.-F. Lai, "Player detection and tracking in broadcast tennis video," *Advances in Image and Video Technology*, pp. 759–770, 2009.
- [96] B. Dang, A. Tran, T. Dinh, and T. Dinh, "A real time player tracking system for broadcast tennis video," *Intelligent information and database systems*, pp. 105–113, 2010.

- [97] D. Connaghan, P. Kelly, and N. E. O'Connor, "Game, shot and match: Event-based indexing of tennis," in *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pp. 97–102, IEEE, 2011.
- [98] T. J. Barnett, *Mathematical Modelling in Hierarchical Games with Specific Reference to Tennis*. PhD thesis, Swinburne University of Technology, Melbourne, Australia, 2006.
- [99] J. Haigh, *Taking Chances: Winning with Probability*. Oxford University Press, 2003.
- [100] I. Kolonias, W. Christmas, and J. Kittler, "Tracking the evolution of a tennis match using hidden markov models," in *Structural, Syntactic, and Statistical Pattern Recognition* (A. Fred, T. Caelli, R. Duin, A. Campilho, and D. de Ridder, eds.), vol. 3138 of *Lecture Notes in Computer Science*, pp. 1078–1086, Springer Berlin / Heidelberg, 2004.
- [101] C. A. R. Hoare, "Algorithm 64: Quicksort," *Commun. ACM*, vol. 4, p. 321, July 1961.

Appendices

A. Selected Ranking Figures and Tables

Table A.1.: Top 100 players as ranked by Match PageRank using matches from 2013.

Rank	ATP	Name	PageRank	S _{out}	S _{in}
1	1	Rafael Nadal	4.7885E-2	6	76
2	2	Novak Djokovic	4.4906E-2	9	62
3	5	Juan Martin Del Potro	2.9202E-2	14	51
4	3	David Ferrer	2.7513E-2	21	61
5	7	Tomas Berdych	2.1155E-2	21	46
6	8	Stanislas Wawrinka	2.0300E-2	20	50
7	12	Tommy Haas	1.8569E-2	23	46
8	4	Andy Murray	1.8409E-2	9	38
9	6	Roger Federer	1.7147E-2	16	46
10	14	John Isner	1.6698E-2	22	38
11	9	Richard Gasquet	1.5470E-2	20	42
12	16	Fabio Fognini	1.4405E-2	26	41
13	60	Lleyton Hewitt	1.3161E-2	17	26
14	11	Milos Raonic	1.3035E-2	20	37
15	15	Mikhail Youzhny	1.2641E-2	23	36
16	23	Grigor Dimitrov	1.2614E-2	23	32
17	24	Ernesto Gulbis	1.1905E-2	19	34
18	10	Jo-Wilfried Tsonga	1.1810E-2	16	36
19	13	Nicolas Almagro	1.1736E-2	23	42
20	18	Tommy Robredo	1.1675E-2	19	35
21	31	Gael Monfils	1.1502E-2	22	35
22	19	Gilles Simon	1.1255E-2	21	33
23	26	Benoit Paire	1.0830E-2	30	30
24	20	Kevin Anderson	1.0765E-2	23	37

25	22	Philipp Kohlschreiber	1.0240E-2	24	33
26	17	Kei Nishikori	1.0091E-2	19	33
27	56	Horacio Zeballos	1.0007E-2	23	14
28	38	Marcel Granollers	9.6768E-3	24	30
29	40	Florian Mayer	8.9761E-3	25	29
30	28	Feliciano Lopez	8.9270E-3	19	29
31	29	Dmitry Tursunov	8.7817E-3	22	30
32	37	Marin Cilic	8.7627E-3	13	25
33	34	Jeremy Chardy	8.6249E-3	25	25
34	33	Ivan Dodig	8.4331E-3	23	31
35	35	Julien Benneteau	8.4000E-3	25	27
36	39	Jarkko Nieminen	8.1114E-3	25	27
37	45	Denis Istomin	8.0557E-3	28	27
38	30	Fernando Verdasco	7.8099E-3	25	28
39	32	Vasek Pospisil	7.6993E-3	16	20
40	25	Andreas Seppi	7.6819E-3	27	28
41	164	Steve Darcis	7.5069E-3	7	2
42	46	Sam Querrey	7.4686E-3	22	26
43	52	Edouard Roger-Vasselin	7.3976E-3	26	26
44	43	Robin Haase	7.0370E-3	25	26
45	27	Jurgen Melzer	7.0081E-3	22	23
46	21	Jerzy Janowicz	6.8848E-3	20	23
47	54	Daniel Brands	6.4087E-3	22	21
48	41	Carlos Berlocq	6.3537E-3	24	26
49	48	Pablo Andujar	6.3332E-3	31	18
50	78	Ivo Karlovic	6.3136E-3	13	16
51	80	Victor Hanescu	6.1048E-3	25	20
52	58	Roberto Bautista Agut	6.0361E-3	20	20
53	61	Marinko Matosevic	5.9837E-3	23	19
54	62	Guillermo Garcia-Lopez	5.8457E-3	21	21
55	42	Juan Monaco	5.3583E-3	21	18
56	47	Lukas Rosol	5.2576E-3	23	19
57	74	Tobias Kamke	5.1458E-3	20	15
58	57	Alexandr Dolgoplov	5.1043E-3	26	20

59	51	Bernard Tomic	5.0600E-3	21	20
60	49	Joao Sousa	4.9022E-3	13	14
61	83	Albert Ramos	4.7104E-3	22	19
62	113	Evgeny Donskoy	4.6916E-3	16	12
63	63	Albert Montanes	4.6646E-3	19	18
64	36	Janko Tipsarevic	4.5889E-3	22	15
65	50	Nicolas Mahut	4.5795E-3	8	18
66	44	Radek Stepanek	4.5436E-3	16	14
67	105	Michael Llodra	4.5351E-3	15	12
68	53	Nikolay Davydenko	4.4179E-3	22	18
69	55	Federico Delbonis	4.3952E-3	8	12
70	59	Adrian Mannarino	4.3546E-3	16	10
71	70	Igor Sijsling	4.0587E-3	23	13
72	65	Yen-Hsun Lu	3.9727E-3	17	16
73	77	Daniel Gimeno-Traver	3.8833E-3	26	14
74	69	Santiago Giraldo	3.8317E-3	23	14
75	72	Lukasz Kubot	3.7265E-3	15	10
76	135	Xavier Malisse	3.6604E-3	19	13
77	84	Kenny De Schepper	3.5962E-3	15	9
78	75	Viktor Troicki	3.4812E-3	16	15
79	87	Marcos Baghdatis	3.3773E-3	21	11
80	131	Ricardas Berankis	3.3222E-3	16	11
81	102	Jack Sock	3.2477E-3	13	10
82	98	Sergiy Stakhovsky	3.2388E-3	14	9
83	109	Paolo Lorenzi	3.2331E-3	22	11
84	94	Leonardo Mayer	3.2249E-3	18	12
85	66	Michal Przysiezny	3.1381E-3	12	10
86	100	Ryan Harrison	3.1218E-3	21	11
87	110	David Goffin	2.9945E-3	20	9
88	99	Alejandro Falla	2.9325E-3	18	12
89	95	Aljaz Bedene	2.8932E-3	17	9
90	121	Grega Zemlja	2.8358E-3	17	11
91	146	Thiemo de Bakker	2.7335E-3	10	7
92	88	Alex Bogomolov Jr.	2.6974E-3	9	5

93	90	Somdev Devvarman	2.6669E-3	12	10
94	133	Martin Alund	2.5628E-3	8	6
95	153	James Blake	2.4977E-3	14	9
96	79	Benjamin Becker	2.4650E-3	19	10
97	104	Blaz Kavcic	2.3907E-3	12	10
98	125	Thomaz Bellucci	2.3373E-3	16	7
99	108	Martin Klizan	2.3058E-3	23	12
100	89	Tim Smyczek	2.3004E-3	12	8

Table A.2.: Top 100 players as ranked by Set PageRank using matches from 2013.

Rank	ATP	Name	PageRank	S_{out}	S_{in}
1	2	Novak Djokovic	4.5689E-2	40	156
2	1	Rafael Nadal	4.3766E-2	34	163
3	3	David Ferrer	2.9025E-2	75	143
4	5	Juan Martin Del Potro	2.7559E-2	46	110
5	8	Stan Wawrinka	2.5392E-2	58	114
6	7	Tomas Berdych	2.3029E-2	54	112
7	6	Roger Federer	2.2938E-2	51	115
8	4	Andy Murray	2.1029E-2	30	93
9	12	Tommy Haas	2.0257E-2	58	107
10	14	John Isner	1.8270E-2	64	94
11	15	Mikhail Youzhny	1.7153E-2	61	90
12	9	Richard Gasquet	1.6978E-2	60	102
13	16	Fabio Fognini	1.6222E-2	66	94
14	31	Gael Monfils	1.5980E-2	54	88
15	24	Ernesto Gulbis	1.5156E-2	54	80
16	10	Jo-Wilfried Tsonga	1.4847E-2	42	86
17	60	Lleyton Hewitt	1.3889E-2	54	67
18	18	Tommy Robredo	1.3858E-2	59	85
19	23	Grigor Dimitrov	1.3337E-2	60	71
20	11	Milos Raonic	1.2920E-2	51	79
21	19	Gilles Simon	1.2910E-2	59	76
22	13	Nicolas Almagro	1.2732E-2	58	93

23	30	Fernando Verdasco	1.2493E-2	65	78
24	22	Philipp Kohlschreiber	1.2320E-2	56	75
25	20	Kevin Anderson	1.1759E-2	64	90
26	26	Benoit Paire	1.1178E-2	70	77
27	45	Denis Istomin	1.0751E-2	73	70
28	33	Ivan Dodig	1.0097E-2	59	67
29	35	Julien Benneteau	1.0050E-2	62	68
30	37	Marin Cilic	9.8859E-3	34	61
31	17	Kei Nishikori	9.7791E-3	47	73
32	25	Andreas Seppi	9.6439E-3	77	71
33	21	Jerzy Janowicz	9.3959E-3	52	62
34	34	Jeremy Chardy	9.1694E-3	63	66
35	29	Dmitry Tursunov	9.1338E-3	47	66
36	28	Feliciano Lopez	8.9297E-3	52	62
37	39	Jarkko Nieminen	8.9278E-3	62	63
38	54	Daniel Brands	8.3980E-3	54	50
39	52	Edouard Roger-Vasselin	8.0617E-3	63	58
40	40	Florian Mayer	7.9505E-3	65	64
41	38	Marcel Granollers	7.8247E-3	66	64
42	46	Sam Querrey	7.7114E-3	59	60
43	32	Vasek Pospisil	7.6446E-3	40	50
44	43	Robin Haase	7.5688E-3	72	61
45	42	Juan Monaco	7.2768E-3	46	49
46	41	Carlos Berlocq	7.2436E-3	56	56
47	27	Jurgen Melzer	7.1380E-3	59	53
48	113	Evgeny Donskoy	6.8826E-3	43	38
49	51	Bernard Tomic	6.6193E-3	50	53
50	58	Roberto Bautista Agut	6.5629E-3	45	51
51	74	Tobias Kamke	6.4930E-3	45	42
52	70	Igor Sijsling	6.2098E-3	44	40
53	48	Pablo Andujar	6.0514E-3	73	42
54	57	Alexandr Dolgoplov	5.9617E-3	58	46
55	56	Horacio Zeballos	5.9139E-3	52	34
56	62	Guillermo Garcia-Lopez	5.7936E-3	57	48

57	61	Marinko Matosevic	5.6637E-3	51	44
58	78	Ivo Karlovic	5.6621E-3	33	36
59	83	Albert Ramos	5.5182E-3	53	50
60	36	Janko Tipsarevic	5.4516E-3	50	38
61	105	Michael Llodra	5.4294E-3	28	31
62	87	Marcos Baghdatis	5.3624E-3	49	33
63	77	Daniel Gimeno-Traver	5.2005E-3	61	43
64	47	Lukas Rosol	5.1765E-3	55	52
65	80	Victor Hanescu	5.1196E-3	55	38
66	108	Martin Klizan	5.1024E-3	52	35
67	63	Albert Montanes	4.8919E-3	50	41
68	50	Nicolas Mahut	4.7517E-3	21	40
69	44	Radek Stepanek	4.6174E-3	36	34
70	65	Yen-Hsun Lu	4.5999E-3	38	41
71	135	Xavier Malisse	4.4204E-3	45	34
72	164	Steve Darcis	4.3691E-3	14	7
73	49	Joao Sousa	4.3673E-3	37	34
74	53	Nikolay Davydenko	4.1740E-3	51	41
75	131	Ricardas Berankis	4.0687E-3	39	31
76	55	Federico Delbonis	4.0573E-3	21	27
77	102	Jack Sock	3.8856E-3	28	25
78	100	Ryan Harrison	3.8707E-3	49	30
79	75	Viktor Troicki	3.8524E-3	37	39
80	69	Santiago Giraldo	3.6837E-3	56	36
81	98	Sergiy Stakhovsky	3.6260E-3	34	28
82	162	Peter Gojowczyk	3.3142E-3	7	12
83	89	Tim Smyczek	3.2985E-3	31	26
84	59	Adrian Mannarino	3.2957E-3	40	28
85	104	Blaz Kavcic	3.2013E-3	33	29
86	99	Alejandro Falla	3.1643E-3	37	28
87	94	Leonardo Mayer	3.1335E-3	37	31
88	133	Martin Alund	3.0862E-3	21	17
89	153	James Blake	2.9512E-3	33	24
90	109	Paolo Lorenzi	2.8814E-3	50	27

91	110	David Goffin	2.8705E-3	49	24
92	79	Benjamin Becker	2.8145E-3	45	26
93	121	Grega Zemlja	2.6949E-3	36	28
94	72	Lukasz Kubot	2.6577E-3	39	28
95	88	Alex Bogomolov Jr.	2.5635E-3	24	16
96	95	Aljaz Bedene	2.5185E-3	41	26
97	66	Michal Przysiezny	2.4470E-3	30	22
98	125	Thomaz Bellucci	2.4352E-3	37	21
99	146	Thiemo de Bakker	2.4316E-3	23	14
100	137	Guillaume Rufin	2.4053E-3	31	23

Table A.3.: Top 100 players as ranked by Game PageRank using matches from 2013.

Rank	ATP	Name	PageRank	S_{out}	S_{in}
1	2	Novak Djokovic	2.1346E-2	717	1133
2	1	Rafael Nadal	2.1047E-2	720	1148
3	3	David Ferrer	1.9674E-2	880	1146
4	8	Stan Wawrinka	1.6637E-2	758	933
5	7	Tomas Berdych	1.6444E-2	733	928
6	6	Roger Federer	1.5594E-2	678	906
7	12	Tommy Haas	1.5593E-2	747	889
8	5	Juan Martin Del Potro	1.5283E-2	687	855
9	14	John Isner	1.5005E-2	829	886
10	9	Richard Gasquet	1.4469E-2	716	860
11	13	Nicolas Almagro	1.3120E-2	675	815
12	15	Mikhail Youzhny	1.3038E-2	691	772
13	16	Fabio Fognini	1.2929E-2	719	780
14	20	Kevin Anderson	1.2861E-2	737	803
15	31	Gael Monfils	1.2482E-2	647	748
16	4	Andy Murray	1.2229E-2	498	689
17	10	Jo-Wilfried Tsonga	1.2072E-2	575	705
18	11	Milos Raonic	1.2038E-2	625	712
19	24	Ernesto Gulbis	1.1925E-2	585	703

20	30	Fernando Verdasco	1.1806E-2	695	726
21	18	Tommy Robredo	1.1777E-2	643	727
22	25	Andreas Seppi	1.1445E-2	733	717
23	23	Grigor Dimitrov	1.1426E-2	611	671
24	26	Benoit Paire	1.1348E-2	724	733
25	45	Denis Istomin	1.1284E-2	708	700
26	22	Philipp Kohlschreiber	1.1245E-2	632	683
27	19	Gilles Simon	1.0649E-2	641	658
28	35	Julien Benneteau	1.0596E-2	649	658
29	33	Ivan Dodig	1.0383E-2	598	633
30	34	Jeremy Chardy	1.0280E-2	641	651
31	60	Lleyton Hewitt	1.0178E-2	569	612
32	43	Robin Haase	1.0042E-2	663	656
33	21	Jerzy Janowicz	9.9527E-3	580	593
34	40	Florian Mayer	9.9059E-3	628	621
35	39	Jarkko Nieminen	9.8632E-3	596	615
36	52	Edouard Roger-Vasselin	9.5655E-3	600	617
37	46	Sam Querrey	9.5321E-3	621	616
38	38	Marcel Granollers	9.4855E-3	628	602
39	28	Feliciano Lopez	9.4190E-3	569	582
40	17	Kei Nishikori	9.2931E-3	519	595
41	29	Dmitry Tursunov	9.0732E-3	542	590
42	41	Carlos Berlocq	8.6027E-3	526	546
43	37	Marin Cilic	8.5344E-3	440	520
44	51	Bernard Tomic	8.5012E-3	534	529
45	54	Daniel Brands	8.2956E-3	524	512
46	27	Jurgen Melzer	8.1849E-3	547	522
47	57	Alexandr Dolgoplov	7.9264E-3	529	491
48	48	Pablo Andujar	7.8521E-3	600	497
49	32	Vasek Pospisil	7.6217E-3	442	486
50	83	Albert Ramos	7.4139E-3	503	486
51	47	Lukas Rosol	7.3808E-3	519	504
52	77	Daniel Gimeno-Traver	7.3772E-3	538	476
53	42	Juan Monaco	7.3721E-3	431	471

54	58	Roberto Bautista Agut	7.3508E-3	456	467
55	62	Guillermo Garcia-Lopez	7.1979E-3	508	482
56	61	Marinko Matosevic	7.1006E-3	490	451
57	80	Victor Hanescu	6.8469E-3	486	443
58	53	Nikolay Davydenko	6.7576E-3	443	439
59	36	Janko Tipsarevic	6.7257E-3	448	422
60	70	Igor Sijsling	6.5175E-3	432	409
61	69	Santiago Giraldo	6.4514E-3	473	430
62	87	Marcos Baghdatis	6.4137E-3	416	388
63	78	Ivo Karlovic	6.3844E-3	381	387
64	74	Tobias Kamke	6.3727E-3	415	411
65	113	Evgeny Donskoy	6.2918E-3	407	388
66	63	Albert Montanes	6.2461E-3	451	413
67	56	Horacio Zeballos	6.1768E-3	447	401
68	108	Martin Klizan	6.1344E-3	445	388
69	65	Yen-Hsun Lu	6.0250E-3	371	389
70	100	Ryan Harrison	5.8409E-3	429	363
71	44	Radek Stepanek	5.7901E-3	344	356
72	135	Xavier Malisse	5.5707E-3	403	365
73	75	Viktor Troicki	5.5625E-3	371	364
74	79	Benjamin Becker	5.2909E-3	389	326
75	131	Ricardas Berankis	5.1890E-3	350	333
76	50	Nicolas Mahut	5.1307E-3	277	335
77	110	David Goffin	5.0283E-3	381	311
78	109	Paolo Lorenzi	4.8767E-3	397	328
79	105	Michael Llodra	4.7889E-3	287	290
80	59	Adrian Mannarino	4.7681E-3	339	311
81	49	Joao Sousa	4.7663E-3	336	312
82	94	Leonardo Mayer	4.5996E-3	347	316
83	153	James Blake	4.5097E-3	284	286
84	121	Grega Zemlja	4.4896E-3	342	294
85	99	Alejandro Falla	4.4515E-3	328	303
86	72	Lukasz Kubot	4.4419E-3	351	310
87	95	Aljaz Bedene	4.3396E-3	346	298

88	98	Sergiy Stakhovsky	4.3395E-3	316	296
89	104	Blaz Kavcic	4.3297E-3	296	289
90	129	Paul-Henri Mathieu	4.3275E-3	326	268
91	102	Jack Sock	4.2836E-3	270	257
92	89	Tim Smyczek	4.2178E-3	281	273
93	134	Andrey Kuznetsov	4.0823E-3	333	275
94	137	Guillaume Rufin	4.0235E-3	272	258
95	90	Somdev Devvarman	4.0028E-3	289	272
96	66	Michal Przysieszny	3.9362E-3	274	254
97	125	Thomaz Bellucci	3.8917E-3	320	264
98	55	Federico Delbonis	3.7525E-3	238	239
99	84	Kenny De Schepper	3.7153E-3	296	263
100	149	Jesse Levine	3.2164E-3	252	206

Table A.4.: Top 100 players as ranked by Point PageRank using matches from 2013.

Rank	ATP	Name	PageRank	S_{out}	S_{in}
1	3	David Ferrer	1.7901E-2	6005	6822
2	2	Novak Djokovic	1.7689E-2	5156	6483
3	1	Rafael Nadal	1.7363E-2	5250	6409
4	7	Tomas Berdych	1.4789E-2	4975	5560
5	8	Stan Wawrinka	1.4764E-2	5037	5527
6	12	Tommy Haas	1.4110E-2	4934	5312
7	6	Roger Federer	1.4071E-2	4561	5348
8	14	John Isner	1.3780E-2	5063	5288
9	9	Richard Gasquet	1.3417E-2	4725	5167
10	5	Juan Martin Del Potro	1.3402E-2	4537	5025
11	16	Fabio Fognini	1.2776E-2	4780	4910
12	20	Kevin Anderson	1.2567E-2	4639	4923
13	13	Nicolas Almagro	1.2296E-2	4380	4815
14	15	Mikhail Youzhny	1.2189E-2	4514	4682
15	26	Benoit Paire	1.1760E-2	4689	4689
16	31	Gael Monfils	1.1742E-2	4268	4529

17	30	Fernando Verdasco	1.1598E-2	4380	4528
18	25	Andreas Seppi	1.1531E-2	4591	4533
19	18	Tommy Robredo	1.1405E-2	4208	4446
20	45	Denis Istomin	1.1158E-2	4399	4367
21	11	Milos Raonic	1.1094E-2	3936	4244
22	24	Ernesto Gulbis	1.1065E-2	3876	4245
23	19	Gilles Simon	1.0869E-2	4217	4257
24	10	Jo-Wilfried Tsonga	1.0786E-2	3765	4110
25	4	Andy Murray	1.0774E-2	3449	4025
26	22	Philipp Kohlschreiber	1.0758E-2	3956	4159
27	35	Julien Benneteau	1.0585E-2	4141	4163
28	23	Grigor Dimitrov	1.0584E-2	3827	4040
29	43	Robin Haase	1.0481E-2	4244	4208
30	34	Jeremy Chardy	1.0287E-2	4000	4065
31	33	Ivan Dodig	1.0007E-2	3776	3886
32	40	Florian Mayer	9.9367E-3	3974	3894
33	60	Lleyton Hewitt	9.9029E-3	3692	3813
34	52	Edouard Roger-Vasselin	9.7149E-3	3843	3864
35	39	Jarkko Nieminen	9.6874E-3	3811	3799
36	38	Marcel Granollers	9.6501E-3	3878	3807
37	46	Sam Querrey	9.6471E-3	3823	3844
38	21	Jerzy Janowicz	9.5212E-3	3614	3668
39	17	Kei Nishikori	9.4164E-3	3483	3706
40	28	Feliciano Lopez	9.1996E-3	3560	3595
41	29	Dmitry Tursunov	9.1135E-3	3564	3655
42	48	Pablo Andujar	8.8209E-3	3776	3484
43	41	Carlos Berlocq	8.6605E-3	3307	3419
44	27	Jurgen Melzer	8.3699E-3	3456	3312
45	51	Bernard Tomic	8.3451E-3	3356	3263
46	57	Alexandr Dolgoplov	8.0596E-3	3231	3143
47	37	Marin Cilic	8.0531E-3	2866	3126
48	54	Daniel Brands	7.9900E-3	3174	3144
49	47	Lukas Rosol	7.9702E-3	3297	3223
50	62	Guillermo Garcia-Lopez	7.8461E-3	3245	3167

51	77	Daniel Gimeno-Traver	7.7905E-3	3282	3103
52	83	Albert Ramos	7.7028E-3	3132	3097
53	61	Marinko Matosevic	7.6591E-3	3103	3019
54	58	Roberto Bautista Agut	7.5991E-3	2930	2964
55	32	Vasek Pospisil	7.5240E-3	2867	3007
56	42	Juan Monaco	7.3775E-3	2787	2929
57	80	Victor Hanescu	7.2316E-3	3021	2866
58	53	Nikolay Davydenko	7.1191E-3	2836	2825
59	36	Janko Tipsarevic	6.9318E-3	2799	2722
60	69	Santiago Giraldo	6.7924E-3	2888	2729
61	63	Albert Montanes	6.6073E-3	2786	2669
62	74	Tobias Kamke	6.5670E-3	2680	2632
63	70	Igor Sijsling	6.5005E-3	2671	2568
64	108	Martin Klizan	6.4549E-3	2748	2561
65	113	Evgeny Donskoy	6.4207E-3	2593	2508
66	87	Marcos Baghdatis	6.3961E-3	2596	2473
67	56	Horacio Zeballos	6.2631E-3	2708	2510
68	65	Yen-Hsun Lu	6.2500E-3	2412	2471
69	100	Ryan Harrison	6.1114E-3	2604	2395
70	135	Xavier Malisse	5.9556E-3	2453	2389
71	44	Radek Stepanek	5.8632E-3	2240	2283
72	75	Viktor Troicki	5.8236E-3	2335	2322
73	78	Ivo Karlovic	5.7923E-3	2248	2254
74	109	Paolo Lorenzi	5.6136E-3	2425	2267
75	110	David Goffin	5.5171E-3	2314	2153
76	79	Benjamin Becker	5.4277E-3	2278	2115
77	131	Ricardas Berankis	5.4034E-3	2217	2160
78	59	Adrian Mannarino	5.1149E-3	2142	2027
79	49	Joao Sousa	5.054E-3	2099	2017
80	94	Leonardo Mayer	4.9758E-3	2142	2027
81	72	Lukasz Kubot	4.9727E-3	2156	2044
82	50	Nicolas Mahut	4.9368E-3	1818	1971
83	99	Alejandro Falla	4.8872E-3	2068	1970
84	121	Grega Zemlja	4.8715E-3	2083	1937

85	95	Aljaz Bedene	4.7678E-3	2038	1939
86	134	Andrey Kuznetsov	4.6436E-3	2032	1875
87	98	Sergiy Stakhovsky	4.6000E-3	1931	1874
88	104	Blaz Kavcic	4.5861E-3	1820	1841
89	153	James Blake	4.5737E-3	1818	1813
90	105	Michael Llodra	4.5662E-3	1771	1763
91	90	Somdev Devvarman	4.5526E-3	1883	1837
92	89	Tim Smyczek	4.5323E-3	1807	1801
93	129	Paul-Henri Mathieu	4.4492E-3	1964	1763
94	125	Thomaz Bellucci	4.3814E-3	1913	1783
95	137	Guillaume Rufin	4.2892E-3	1741	1690
96	102	Jack Sock	4.0899E-3	1618	1573
97	84	Kenny De Schepper	4.0707E-3	1790	1676
98	66	Michal Przysiezny	3.9782E-3	1654	1579
99	55	Federico Delbonis	3.9598E-3	1530	1552
100	149	Jesse Levine	3.5948E-3	1555	1413

Table A.5.: Top 100 players as ranked by LadderRank Combined system with parameter $X=1$ using matches from 2013.

Rank	ATP	Name	Matches Lost	Matches Won
1	2	Novak Djokovic	9	62
2	1	Rafael Nadal	6	76
3	6	Roger Federer	16	46
4	4	Andy Murray	9	38
5	3	David Ferrer	21	61
6	5	Juan Martin Del Potro	14	51
7	NA	Chris Guccione	1	1
8	223	Miloslav Mecir	1	1
9	7	Tomas Berdych	21	46
10	11	Milos Raonic	20	37
11	NA	Christian Harrison	1	1
12	13	Nicolas Almagro	23	42
13	9	Richard Gasquet	20	42

14	8	Stan Wawrinka	20	50
15	37	Marin Cilic	13	25
16	91	Alejandro Gonzalez	1	0
17	24	Ernesto Gulbis	19	34
18	157	Maximo Gonzalez	1	1
19	10	Jo-Wilfried Tsonga	16	36
20	227	Greg Jones	1	1
21	50	Nicolas Mahut	8	18
22	20	Kevin Anderson	23	37
23	17	Kei Nishikori	19	33
24	97	Bradley Klahn	3	1
25	14	John Isner	22	38
26	154	Paul Capdeville	1	1
27	31	Gael Monfils	22	35
28	270	Alexander Kudryavtsev	1	1
29	23	Grigor Dimitrov	23	32
30	143	Michael Berrer	3	2
31	42	Juan Monaco	21	18
32	115	Stephane Robert	2	2
33	41	Carlos Berlocq	24	26
34	NA	Emilio Gomez	1	1
35	NA	Prakash Amritraj	1	1
36	12	Tommy Haas	23	46
37	162	Peter Gojowczyk	2	4
38	222	Gerard Granollers	1	0
39	32	Vasek Pospisil	16	20
40	18	Tommy Robredo	19	35
41	65	Yen-Hsun Lu	17	16
42	22	Philipp Kohlschreiber	24	33
43	148	Ruben Bemelmans	4	3
44	30	Fernando Verdasco	25	28
45	33	Ivan Dodig	23	31
46	NA	Christian Garin	1	1
47	111	Dustin Brown	3	4

48	44	Radek Stepanek	16	14
49	229	David Nalbandian	5	6
50	82	Andrey Golubev	4	3
51	139	Dominic Thiem	3	4
52	58	Roberto Bautista Agut	20	20
53	181	Marton Fucsovics	1	1
54	15	Mikhail Youzhny	23	36
55	35	Julien Benneteau	25	27
56	16	Fabio Fognini	26	41
57	104	Blaz Kavcic	12	10
58	89	Tim Smyczek	12	8
59	19	Gilles Simon	21	33
60	52	Edouard Roger-Vasselin	26	26
61	25	Andreas Seppi	27	28
62	28	Feliciano Lopez	19	29
63	21	Jerzy Janowicz	20	23
64	78	Ivo Karlovic	13	16
65	55	Federico Delbonis	8	12
66	NA	Brian Baker	4	3
67	45	Denis Istomin	28	27
68	54	Daniel Brands	22	21
69	39	Jarkko Nieminen	25	27
70	60	Lleyton Hewitt	17	26
71	46	Sam Querrey	22	26
72	195	Yuki Bhambri	1	2
73	105	Michael Llodra	15	12
74	38	Marcel Granollers	24	30
75	34	Jeremy Chardy	25	25
76	172	Samuel Groth	4	3
77	NA	Riccardo Ghedin	1	0
78	96	Donald Young	3	2
79	83	Albert Ramos	22	19
80	150	Daniel Evans	3	4
81	29	Dmitry Tursunov	22	30

82	57	Alexandr Dolgoplov	26	20
83	26	Benoit Paire	30	30
84	40	Florian Mayer	25	29
85	53	Nikolay Davydenko	22	18
86	107	Jan-Lennard Struff	9	4
87	122	Frank Dancevic	3	2
88	153	James Blake	14	9
89	75	Viktor Troicki	16	15
90	51	Bernard Tomic	21	20
91	NA	Joachim Johansson	1	1
92	61	Marinko Matosevic	23	19
93	43	Robin Haase	25	26
94	113	Evgeny Donskoy	16	12
95	62	Guillermo Garcia-Lopez	21	21
96	27	Jurgen Melzer	22	23
97	36	Janko Tipsarevic	22	15
98	190	John Millman	3	1
99	49	Joao Sousa	13	14
100	NA	Karen Khachanov	2	3

Table A.6.: Top 100 players as ranked by LadderRank Combined system with parameter $X=3$ using matches from 2013.

Rank	ATP	Name	Matches Lost	Matches Won
1	2	Novak Djokovic	9	62
2	1	Rafael Nadal	6	76
3	6	Roger Federer	16	46
4	4	Andy Murray	9	38
5	3	David Ferrer	21	61
6	5	Juan Martin Del Potro	14	51
7	NA	Chris Guccione	1	1
8	223	Miloslav Mecir	1	1
9	7	Tomas Berdych	21	46
10	11	Milos Raonic	20	37

11	NA	Christian Harrison	1	1
12	13	Nicolas Almagro	23	42
13	9	Richard Gasquet	20	42
14	8	Stan Wawrinka	20	50
15	37	Marin Cilic	13	25
16	91	Alejandro Gonzalez	1	0
17	24	Ernesto Gulbis	19	34
18	222	Gerard Granollers	1	0
19	157	Maximo Gonzalez	1	1
20	10	Jo-Wilfried Tsonga	16	36
21	227	Greg Jones	1	1
22	50	Nicolas Mahut	8	18
23	12	Tommy Haas	23	46
24	20	Kevin Anderson	23	37
25	154	Paul Capdeville	1	1
26	17	Kei Nishikori	19	33
27	97	Bradley Klahn	3	1
28	14	John Isner	22	38
29	115	Stephane Robert	2	2
30	NA	Prakash Amritraj	1	1
31	270	Alexander Kudryavtsev	1	1
32	162	Peter Gojowczyk	2	4
33	32	Vasek Pospisil	16	20
34	23	Grigor Dimitrov	23	32
35	18	Tommy Robredo	19	35
36	31	Gael Monfils	22	35
37	42	Juan Monaco	21	18
38	65	Yen-Hsun Lu	17	16
39	22	Philipp Kohlschreiber	24	33
40	143	Michael Berrer	3	2
41	148	Ruben Bemelmans	4	3
42	30	Fernando Verdasco	25	28
43	41	Carlos Berlocq	24	26
44	NA	Emilio Gomez	1	1

45	33	Ivan Dodig	23	31
46	NA	Christian Garin	1	1
47	111	Dustin Brown	3	4
48	44	Radek Stepanek	16	14
49	229	David Nalbandian	5	6
50	82	Andrey Golubev	4	3
51	139	Dominic Thiem	3	4
52	58	Roberto Bautista Agut	20	20
53	181	Marton Fucsovics	1	1
54	15	Mikhail Youzhny	23	36
55	35	Julien Benneteau	25	27
56	16	Fabio Fognini	26	41
57	NA	Riccardo Ghedin	1	0
58	195	Yuki Bhambri	1	2
59	104	Blaz Kavcic	12	10
60	89	Tim Smyczek	12	8
61	60	Lleyton Hewitt	17	26
62	46	Sam Querrey	22	26
63	19	Gilles Simon	21	33
64	38	Marcel Granollers	24	30
65	52	Edouard Roger-Vasselin	26	26
66	25	Andreas Seppi	27	28
67	34	Jeremy Chardy	25	25
68	28	Feliciano Lopez	19	29
69	21	Jerzy Janowicz	20	23
70	105	Michael Llodra	15	12
71	78	Ivo Karlovic	13	16
72	172	Samuel Groth	4	3
73	263	Rui Machado	1	0
74	55	Federico Delbonis	8	12
75	NA	Brian Baker	4	3
76	45	Denis Istomin	28	27
77	75	Viktor Troicki	16	15
78	54	Daniel Brands	22	21

79	39	Jarkko Nieminen	25	27
80	96	Donald Young	3	2
81	83	Albert Ramos	22	19
82	40	Florian Mayer	25	29
83	53	Nikolay Davydenko	22	18
84	107	Jan-Lennard Struff	9	4
85	150	Daniel Evans	3	4
86	NA	Joachim Johansson	1	1
87	153	James Blake	14	9
88	29	Dmitry Tursunov	22	30
89	61	Marinko Matosevic	23	19
90	57	Alexandr Dolgoplov	26	20
91	26	Benoit Paire	30	30
92	43	Robin Haase	25	26
93	NA	Stephane Bohli	1	0
94	98	Sergiy Stakhovsky	14	9
95	122	Frank Dancevic	3	2
96	51	Bernard Tomic	21	20
97	90	Somdev Devvarman	12	10
98	74	Tobias Kamke	20	15
99	131	Ricardas Berankis	16	11
100	113	Evgeny Donskoy	16	12

Table A.7.: Top 100 players as ranked by LadderRank Combined system with parameter $X=3$ and a minimum of 5 matches played using historical data from 2013.

Rank	ATP	Name	Matches Lost	Matches Won
1	2	Novak Djokovic	9	62
2	1	Rafael Nadal	6	76
3	6	Roger Federer	16	46
4	4	Andy Murray	9	38
5	3	David Ferrer	21	61
6	5	Juan Martin Del Potro	14	51

7	7	Tomas Berdych	21	46
8	13	Nicolas Almagro	23	42
9	11	Milos Raonic	20	37
10	9	Richard Gasquet	20	42
11	8	Stan Wawrinka	20	50
12	37	Marin Cilic	13	25
13	24	Ernesto Gulbis	19	34
14	50	Nicolas Mahut	8	18
15	10	Jo-Wilfried Tsonga	16	36
16	12	Tommy Haas	23	46
17	20	Kevin Anderson	23	37
18	17	Kei Nishikori	19	33
19	14	John Isner	22	38
20	162	Peter Gojowczyk	2	4
21	32	Vasek Pospisil	16	20
22	23	Grigor Dimitrov	23	32
23	18	Tommy Robredo	19	35
24	31	Gael Monfils	22	35
25	42	Juan Monaco	21	18
26	65	Yen-Hsun Lu	17	16
27	22	Philipp Kohlschreiber	24	33
28	143	Michael Berrer	3	2
29	148	Ruben Bemelmans	4	3
30	30	Fernando Verdasco	25	28
31	41	Carlos Berlocq	24	26
32	33	Ivan Dodig	23	31
33	111	Dustin Brown	3	4
34	44	Radek Stepanek	16	14
35	229	David Nalbandian	5	6
36	58	Roberto Bautista Agut	20	20
37	15	Mikhail Youzhny	23	36
38	82	Andrey Golubev	4	3
39	60	Lleyton Hewitt	17	26
40	139	Dominic Thiem	3	4

41	35	Julien Benneteau	25	27
42	16	Fabio Fognini	26	41
43	104	Blaz Kavcic	12	10
44	89	Tim Smyczek	12	8
45	46	Sam Querrey	22	26
46	19	Gilles Simon	21	33
47	38	Marcel Granollers	24	30
48	52	Edouard Roger-Vasselin	26	26
49	25	Andreas Seppi	27	28
50	21	Jerzy Janowicz	20	23
51	105	Michael Llodra	15	12
52	34	Jeremy Chardy	25	25
53	28	Feliciano Lopez	19	29
54	78	Ivo Karlovic	13	16
55	55	Federico Delbonis	8	12
56	75	Viktor Troicki	16	15
57	172	Samuel Groth	4	3
58	NA	Brian Baker	4	3
59	39	Jarkko Nieminen	25	27
60	45	Denis Istomin	28	27
61	54	Daniel Brands	22	21
62	96	Donald Young	3	2
63	83	Albert Ramos	22	19
64	40	Florian Mayer	25	29
65	53	Nikolay Davydenko	22	18
66	107	Jan-Lennard Struff	9	4
67	150	Daniel Evans	3	4
68	153	James Blake	14	9
69	29	Dmitry Tursunov	22	30
70	61	Marinko Matosevic	23	19
71	57	Alexandr Dolgoplov	26	20
72	26	Benoit Paire	30	30
73	43	Robin Haase	25	26
74	98	Sergiy Stakhovsky	14	9

75	80	Victor Hanescu	25	20
76	122	Frank Dancevic	3	2
77	51	Bernard Tomic	21	20
78	90	Somdev Devvarman	12	10
79	74	Tobias Kamke	20	15
80	NA	Karen Khachanov	2	3
81	NA	Mardy Fish	5	4
82	131	Ricardas Berankis	16	11
83	113	Evgeny Donskoy	16	12
84	135	Xavier Malisse	19	13
85	62	Guillermo Garcia-Lopez	21	21
86	47	Lukas Rosol	23	19
87	27	Jurgen Melzer	22	23
88	36	Janko Tipsarevic	22	15
89	69	Santiago Giraldo	23	14
90	49	Joao Sousa	13	14
91	137	Guillaume Rufin	13	8
92	71	Filippo Volandri	12	5
93	142	Alex Kuznetsov	5	2
94	88	Alex Bogomolov Jr.	9	5
95	219	Jan Hernych	4	5
96	102	Jack Sock	13	10
97	99	Alejandro Falla	18	12
98	95	Aljaz Bedene	17	9
99	87	Marcos Baghdatis	21	11
100	66	Michal Przysiezny	12	10

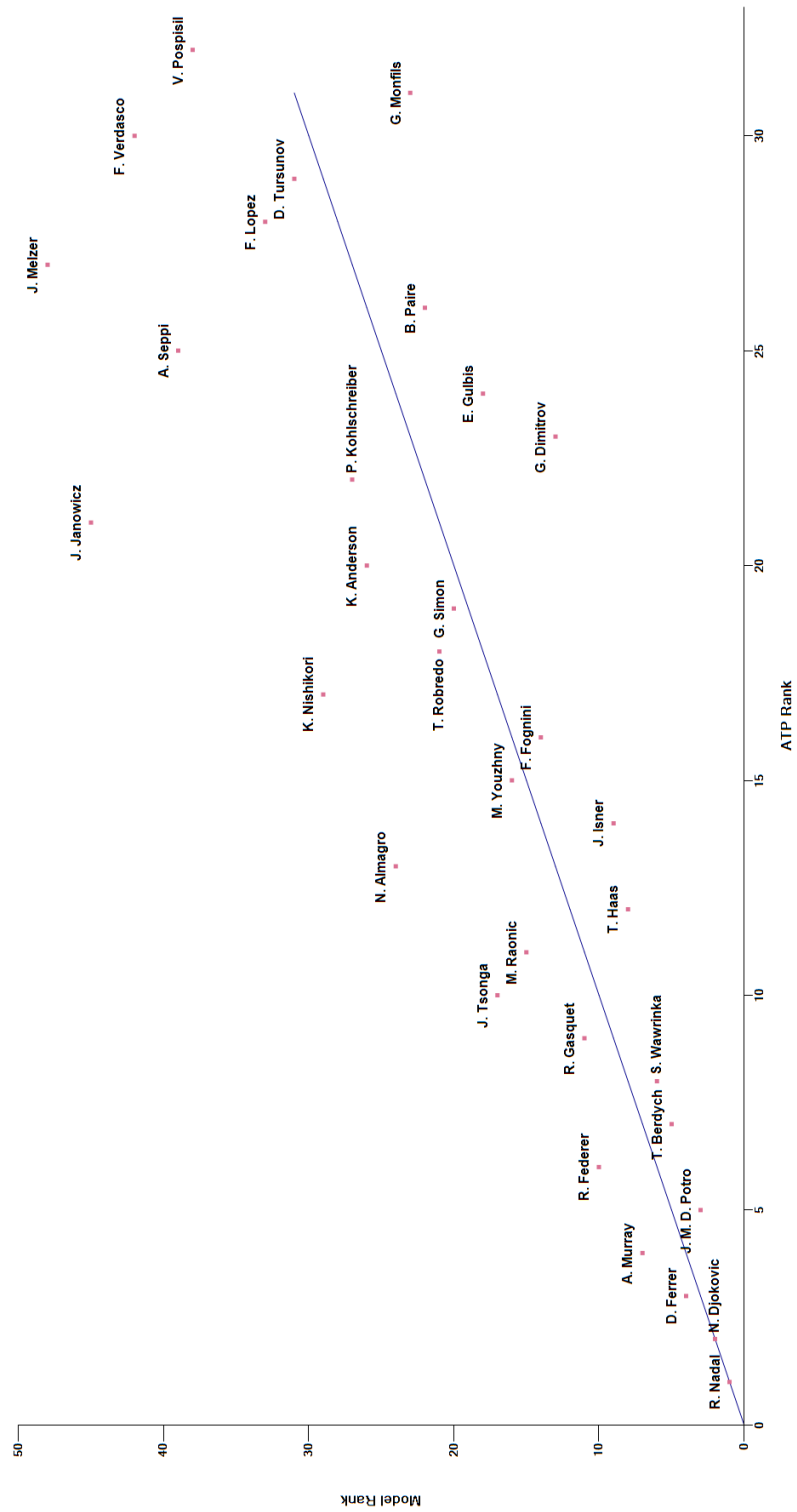


Figure A.1.: The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the Match PageRank system.

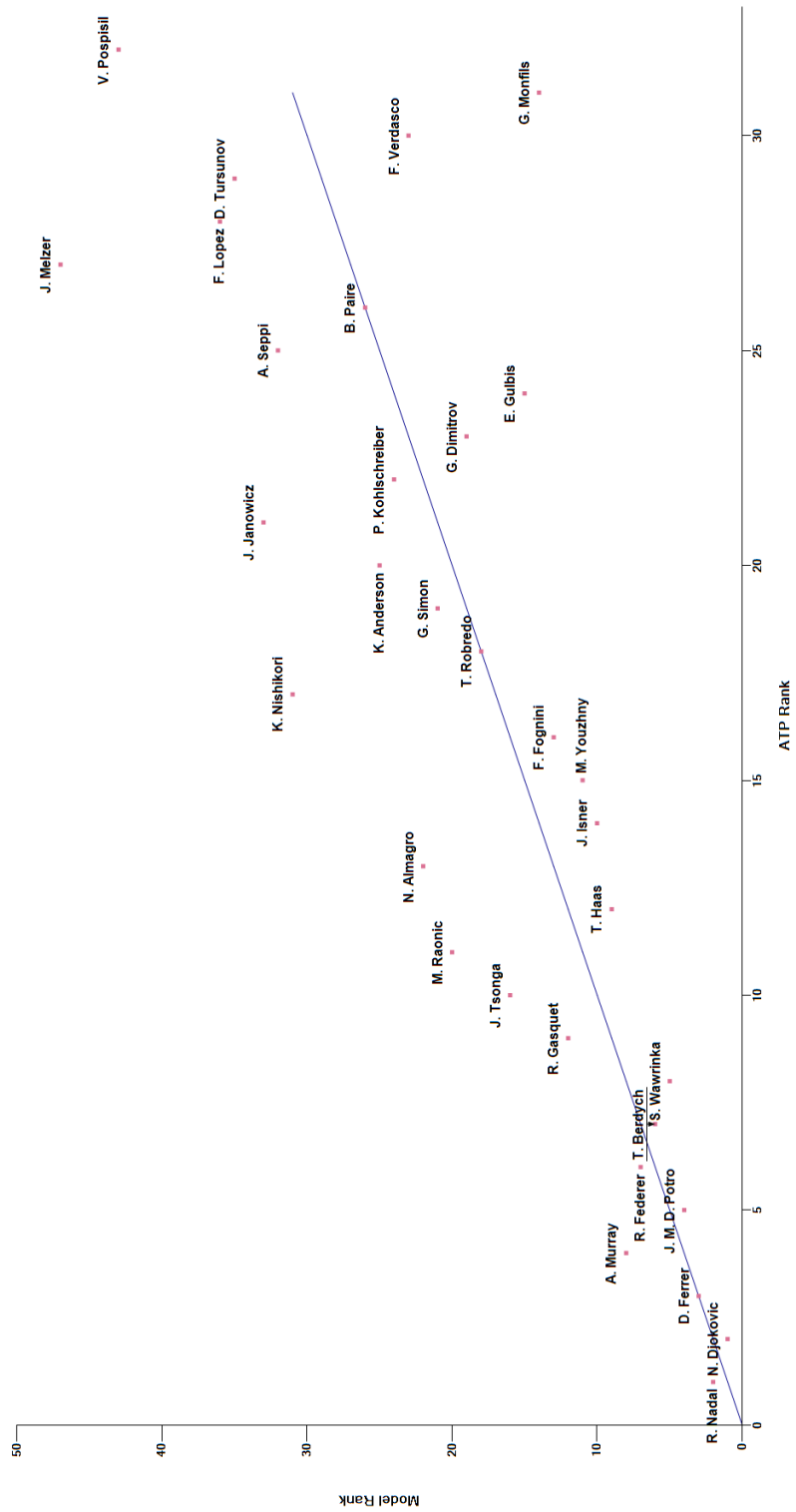


Figure A.2.: The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the Set PageRank system.

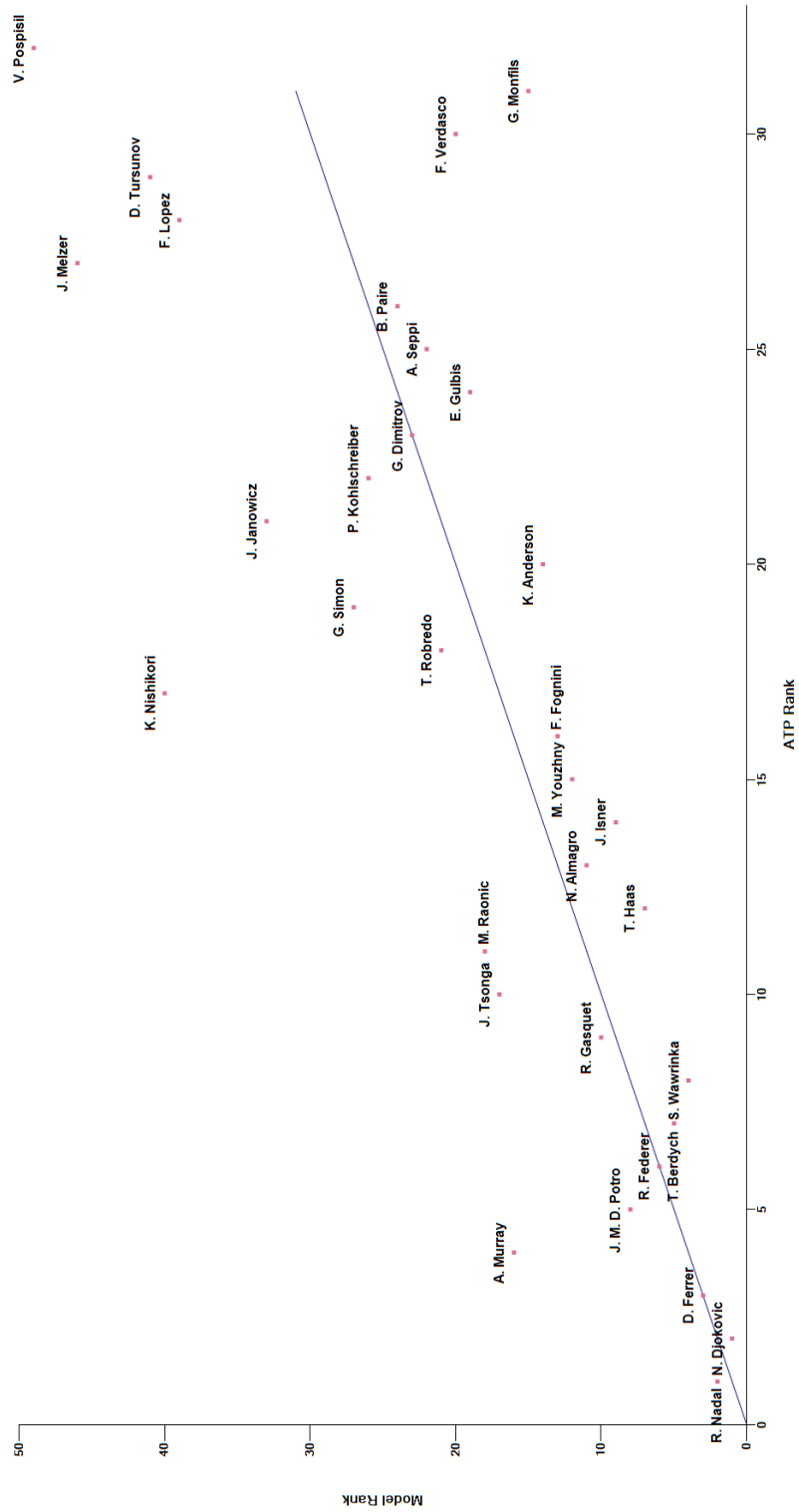


Figure A.3.: The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the Game PageRank system.

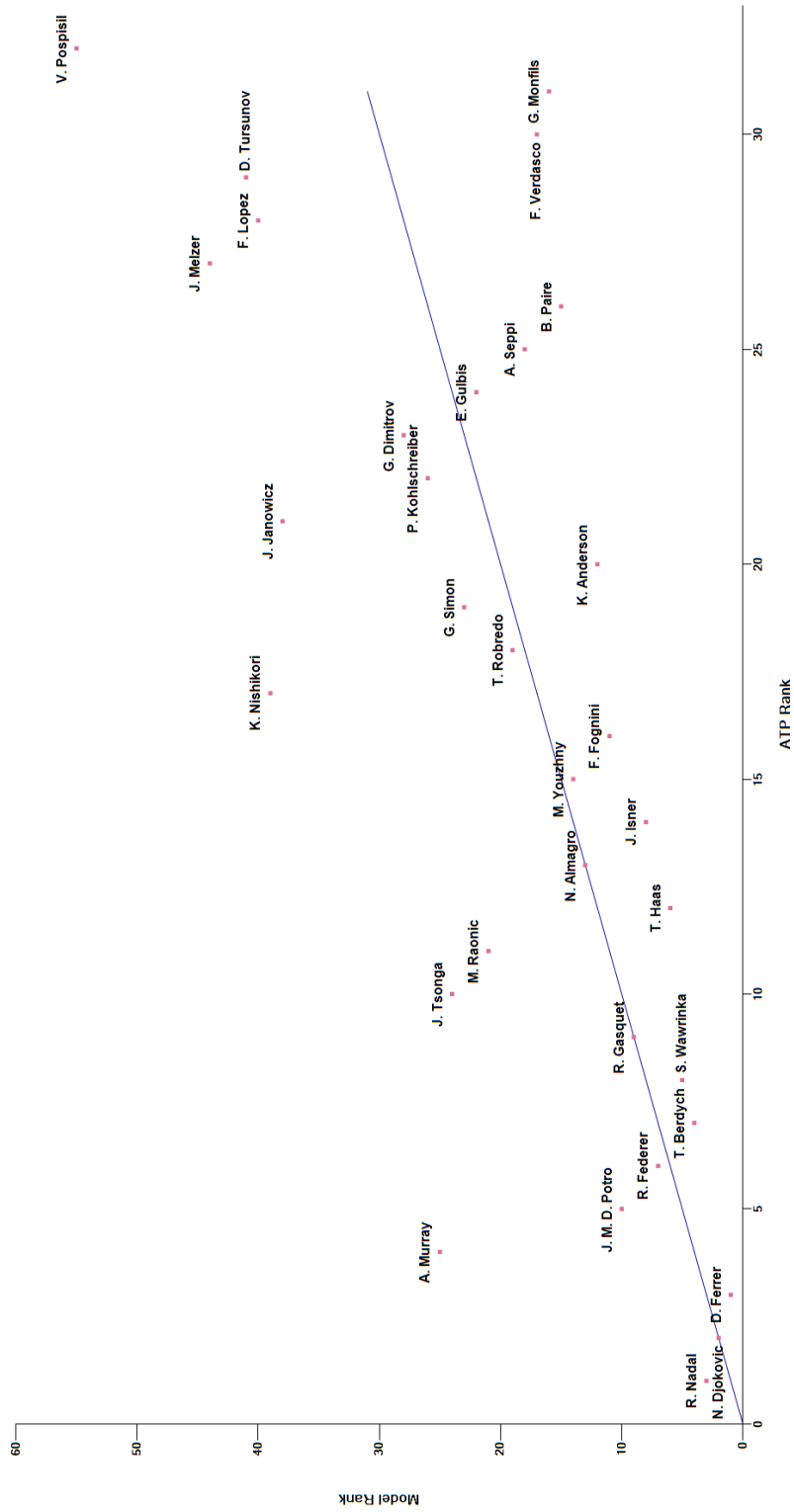


Figure A.4.: The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the Point PageRank system.

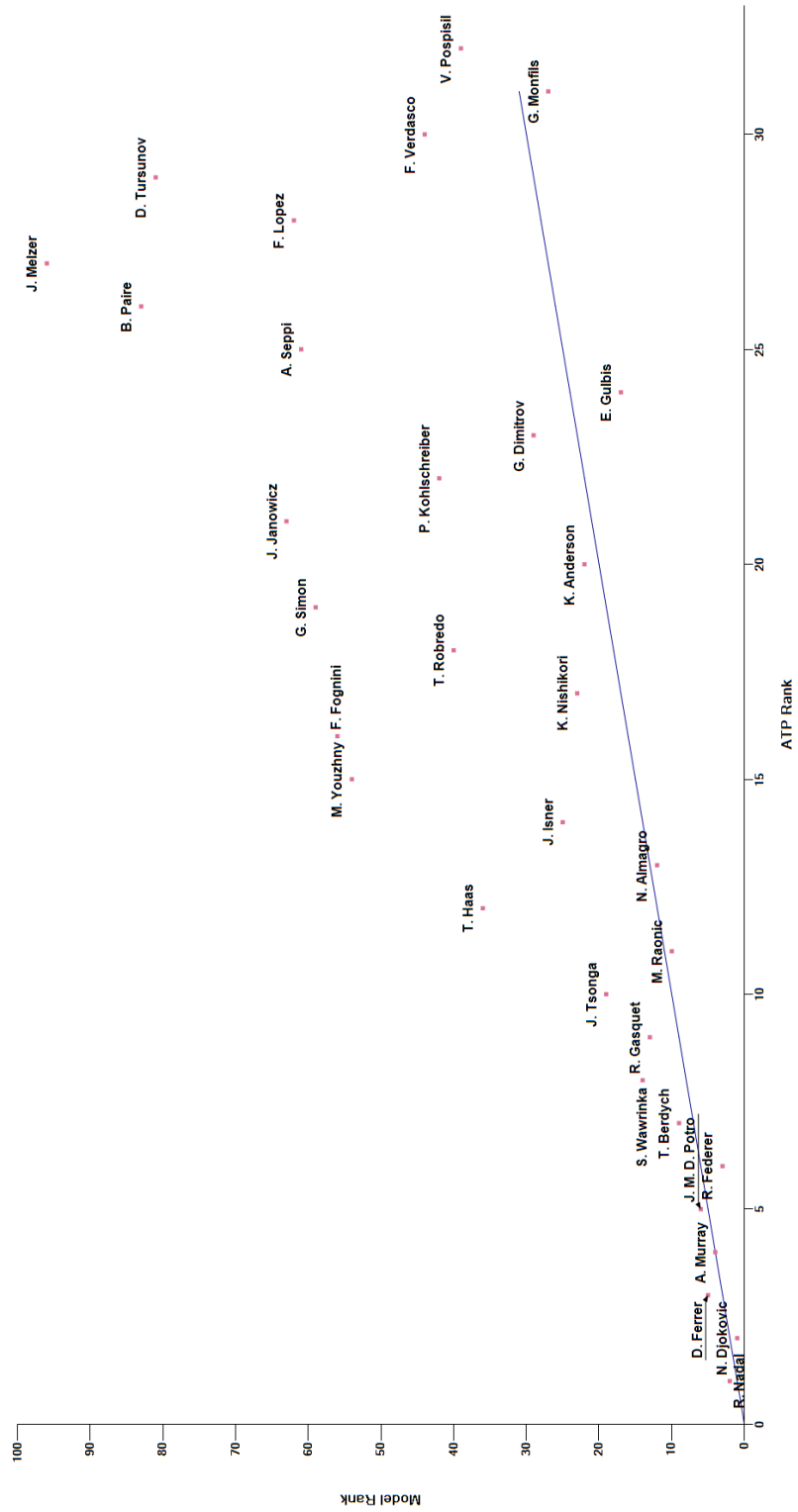


Figure A.5.: The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the LadderRank Combined system with $X=1$.

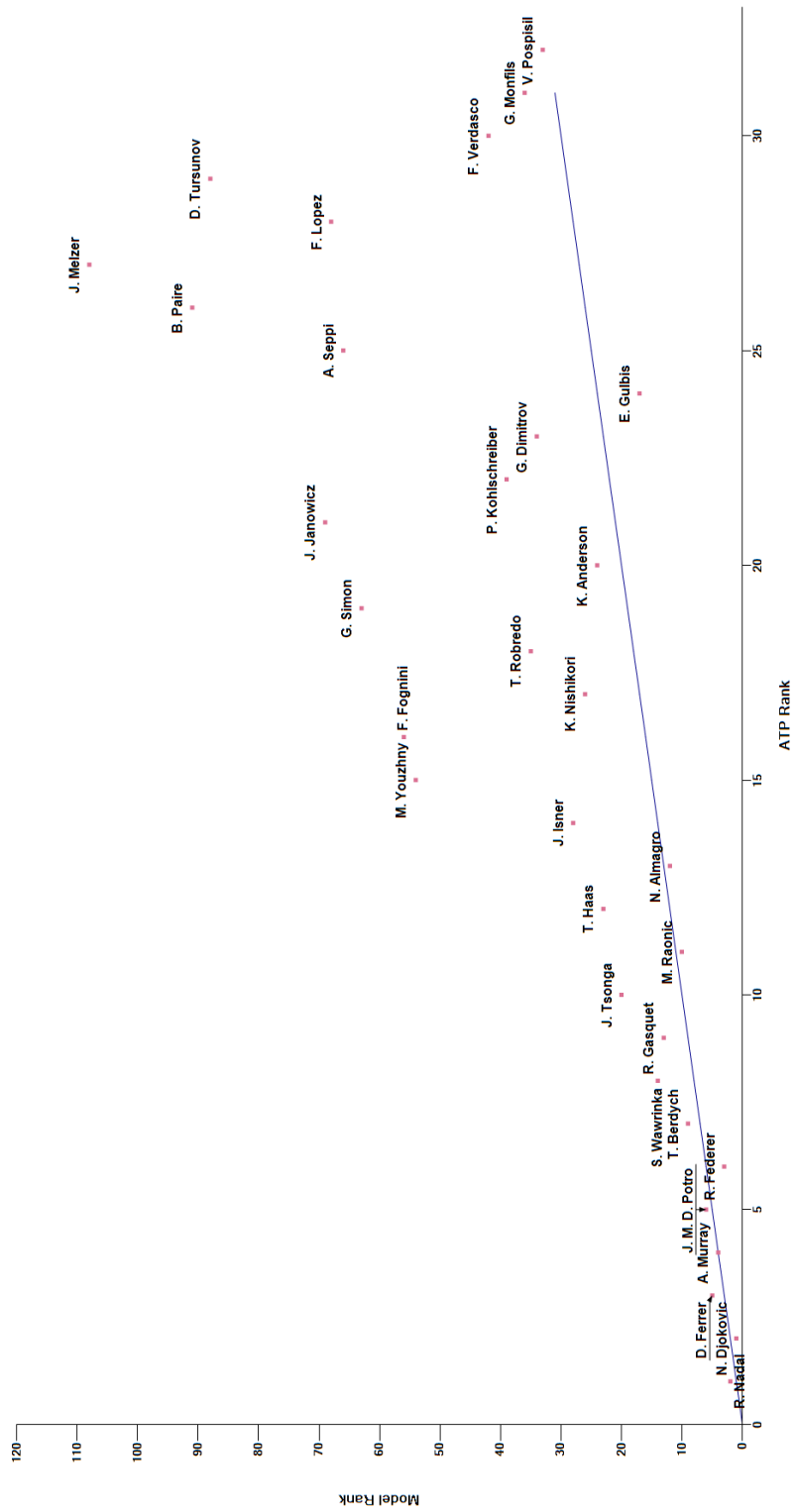


Figure A.6.: The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the LadderRank Combined system with $X=3$.

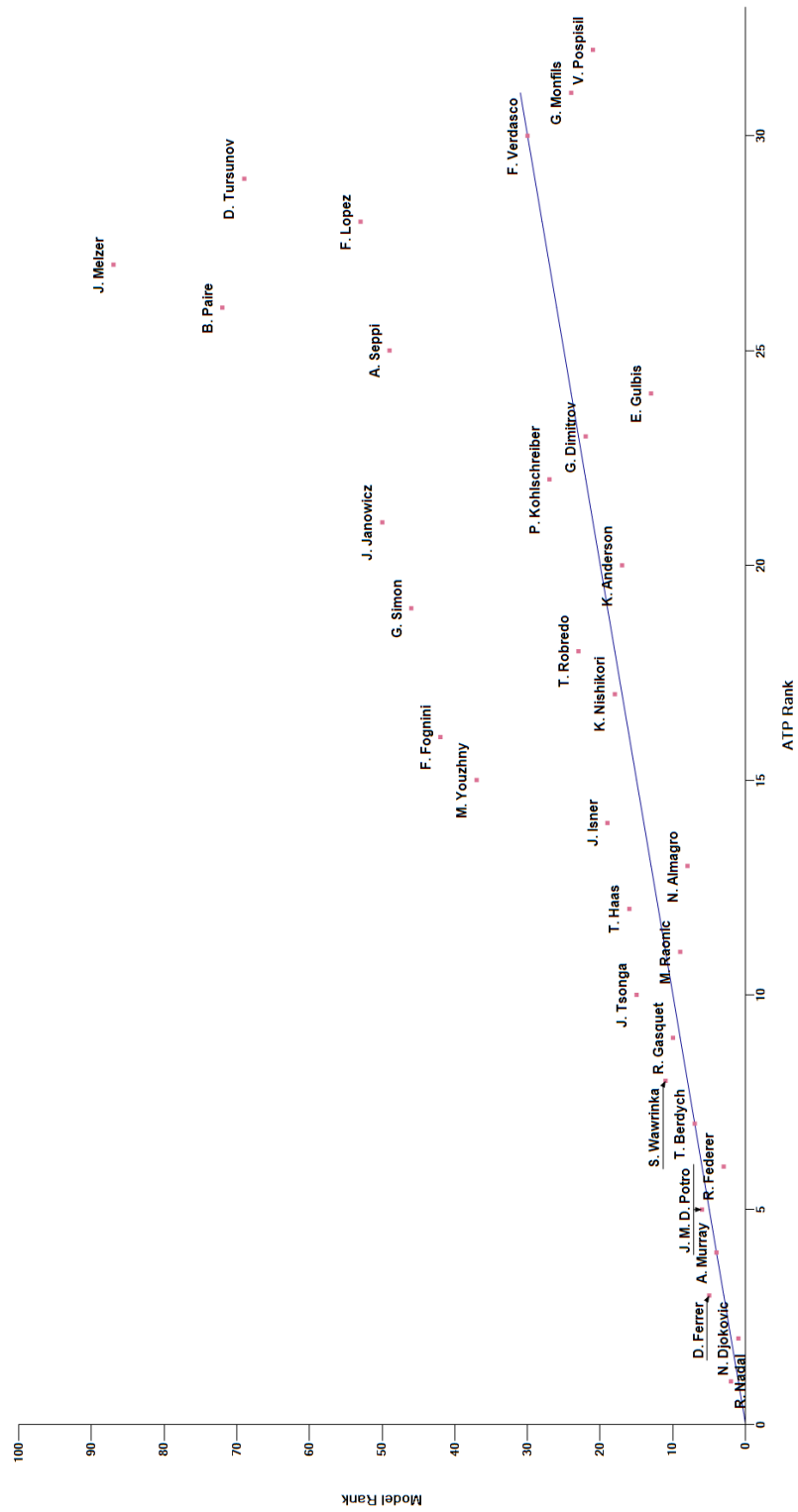


Figure A.7.: The rankings of the Top 32 ATP players at the end of 2013 compared to their ranking generated using the LadderRank Combined system with $X=3$ and a minimum of 5 matches.