Imperial College London
Department of Computing

MEng Final Year Project

Quantitative Models for Retirement Risk in Professional Tennis

Author:
Adam Cutmore

Supervisor:
William Knottenbelt

June 21, 2012

Abstract

Injuries during tennis matches are phenomena that can drastically alter the in-play betting odds of a match even during the course of a single point. In-play tennis betting markets are some of the most heavily traded in the industry and enforce a variety of payout policies. These markets often differ in their odds for a match as only some of them take player retirement into account. We specifically investigate the Betfair Set Betting market, in which all bets are cancelled in the event of a retirement, and the Betfair Match Odds market, which only pays out on retirements if they occur after the first set has been played.

By interpreting the probability a player will retire at some point during the remainder of a given match as a function of any gap in the odds of the two Betfair markets, we create the world's first model of a tennis match that takes into account risk of retirement. We test our model on randomly generated artificial matches to see if we can imitate the expected behaviour of markets that use different retirement payout rules.

We find that we are able to follow the progression of Betfair in-play tennis markets for a number of real-life matches to a good degree of accuracy and can provide a value for the retirement risk of a given player at any point. We also attempt to predict the evolution of a market which pays out on retirements even after just one ball has been played. We find that although this *after one ball* model generally behaves as expected, it is very sensitive to any gap in the Betfair odds which in turn affects our predicted retirement risk. Similarly, we note that larger than expected retirement risk spikes are seen when a player has a low match-winning probability. Such fragilities are due to a heavy dependence on imperfect odds data.

Acknowledgements

First, and most importantly, I'd like to thank Professor William Knottenbelt for coming up with the core idea behind the investigation as well as for the fantastic and unconditional support he has provided me throughout the course of the project. I'd also like to extend my thanks to Andy Murray and his fellow tennis professionals for occasionally getting injured during high-profile tennis matches but, nevertheless, soldiering on and not always immediately retiring (at least for a few more games anyway).

Contents

INTRODUCTION 1

1.1. Player Retirement in Tennis Matches

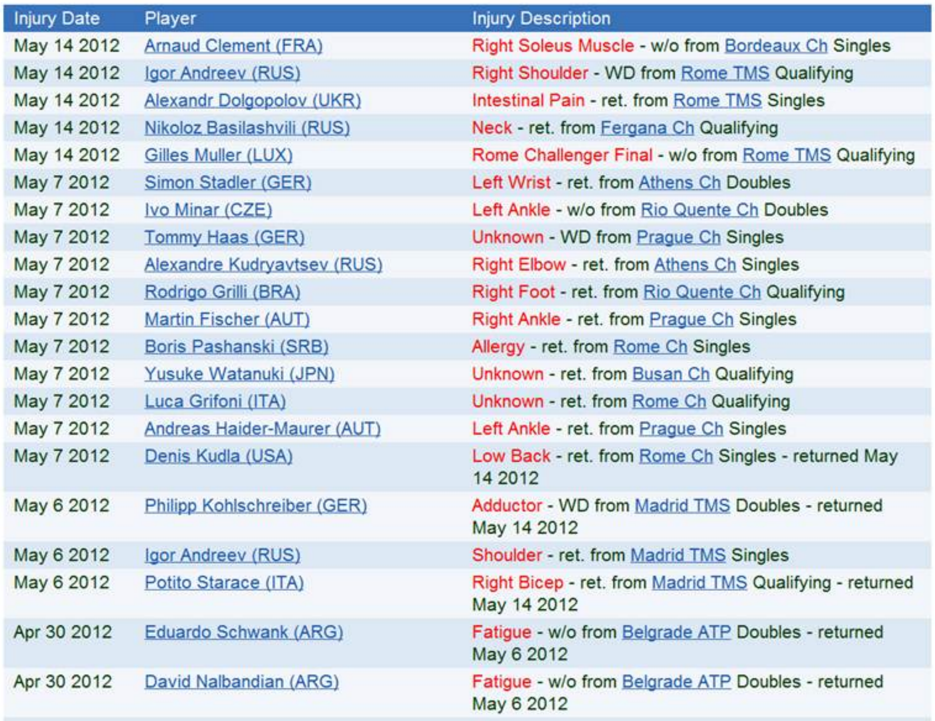
In professional tennis, a player may 'retire hurt' from a match at any time should they feel they are unable to complete the match due to injury or illness, or that it is unwise to continue in case they aggravate their condition. The match is consequently awarded to the opponent regardless of the current match state. A *walkover* occurs when a player withdraws from a match before it has begun.

In-play injuries are common occurrences in tennis as a whole. Between 2000 and 2009, there was a retirement during approximately 3.9% of Grand Slam men's singles matches[1]. At any one time there is likely to be a significant number of professional tennis players that are currently recovering from injuries sustained during a match. The *TennisInsight.com*[1] website records a huge number of facts and interesting statistics about all registered professional tennis players and any matches they play, including tournaments currently being held. Figure 1.1 shows a snapshot of the list of currently sidelined players maintained on *TennisInsight.com*. The majority of the injuries shown were the cause of retirements from matches in tournaments, suggesting that they occurred in-play.

1.2. Rise of the Online Gambling Industry

In recent times, interest in online gambling has seen phenomenal growth due to the widespread availability of the internet, especially with the rapid advances in smartphone technology seen over the last few years. The global online gambling industry grew by 12% during 2010 to reach a breathtaking total of almost $30 billion. Sporting events, in particular, are a favourite of many a gambler and sports betting has seen an increase in popularity to match the industry as a whole, with 41% of revenue attributed to this sector. The online gambling industry is predicted to be worth up to $40 billion by 2014 should current trends continue[2].

[1] http://www.tennisinsight.com



Injury Date	Player	Injury Description
May 14 2012	Arnaud Clement (FRA)	Right Soleus Muscle - w/o from Bordeaux Ch Singles
May 14 2012	Igor Andreev (RUS)	Right Shoulder - WD from Rome TMS Qualifying
May 14 2012	Alexandr Dolgopolov (UKR)	Intestinal Pain - ret. from Rome TMS Singles
May 14 2012	Nikoloz Basilashvili (RUS)	Neck - ret. from Fergana Ch Qualifying
May 14 2012	Gilles Muller (LUX)	Rome Challenger Final - w/o from Rome TMS Qualifying
May 7 2012	Simon Stadler (GER)	Left Wrist - ret. from Athens Ch Doubles
May 7 2012	Ivo Minar (CZE)	Left Ankle - w/o from Rio Quente Ch Doubles
May 7 2012	Tommy Haas (GER)	Unknown - WD from Prague Ch Singles
May 7 2012	Alexandre Kudryavtsev (RUS)	Right Elbow - ret. from Athens Ch Singles
May 7 2012	Rodrigo Grilli (BRA)	Right Foot - ret. from Rio Quente Ch Qualifying
May 7 2012	Martin Fischer (AUT)	Right Ankle - ret. from Prague Ch Singles
May 7 2012	Boris Pashanski (SRB)	Allergy - ret. from Rome Ch Singles
May 7 2012	Yusuke Watanuki (JPN)	Unknown - ret. from Busan Ch Qualifying
May 7 2012	Luca Grifoni (ITA)	Unknown - ret. from Rome Ch Qualifying
May 7 2012	Andreas Haider-Maurer (AUT)	Left Ankle - ret. from Prague Ch Singles
May 7 2012	Denis Kudla (USA)	Low Back - ret. from Rome Ch Singles - returned May 14 2012
May 6 2012	Philipp Kohlschreiber (GER)	Adductor - WD from Madrid TMS Doubles - returned May 14 2012
May 6 2012	Igor Andreev (RUS)	Shoulder - ret. from Madrid TMS Singles
May 6 2012	Potito Starace (ITA)	Right Bicep - ret. from Madrid TMS Qualifying - returned May 14 2012
Apr 30 2012	Eduardo Schwank (ARG)	Fatigue - w/o from Belgrade ATP Doubles - returned May 6 2012
Apr 30 2012	David Nalbandian (ARG)	Fatigue - w/o from Belgrade ATP Doubles - returned May 6 2012

Figure 1.1.: Snapshot of the list of current and recently injured professional tennis players as shown on *TennisInsight.com* (accessed June 2012, midway through the French Open)

Betting Exchanges

In the high street, the traditional bookmaker is the market maker; they quote buy and sell prices, i.e. offer bets for contrasting events. Companies such as William Hill and Paddy Power offer odds on the outcomes of events and accept wagers from customers. However, these bookmakers do not offer what you might call actual odds. Although odds are by their very nature subjective, the odds offered by bookmakers are slightly biased in favour of themselves in order to statistically guarantee that they make a profit. If bookmakers offered actual odds, over a long period of time they would only be able to break even (similar to if one flips a coin many times, one will get approximately heads half the time and tails half the time). This means that these odds are not truly indicative of the probability of a certain event occurring.

Betting exchanges, on the other hand, work differently. They allow customers to trade directly with each other whilst they instead play the role of supervisor or middleman. The customers themselves are allowed to offer as well as place bets; they essentially fulfil the role of bookmaker as well. Whereas successful gamblers may be restricted by a traditional bookmaker, exchanges allow bets of any size and odds - as long as someone is willing to match them! Betting exchanges such as Betfair instead charge a percentage commission on customer net winnings in order to generate revenue. As a result, the odds on an exchange are decided by the user base so we can assume that they more closely reflect the true odds of an event occurring as they represent the combined opinions of a large number of people. This is especially true for popular markets such as tennis, football, and horse-racing, where millions of pounds are traded on individual events.

1.3. In-Play Tennis Betting Markets

The creation of betting exchanges has arrived hand-in-hand with the emergence of in-play betting markets. In-play betting is when customers are able to place bets while an event is still in progress. Some tennis related examples include Set Betting (final score), Most Aces, Match Odds (picking a winner), and Total Games. Since traders can now rely on live data from a match in addition to pre-match estimations, odds fluctuate far more quickly than in the standard pre-match market as traders react to what is happening in real-time, potentially allowing greater profits. Traditional high street bookmakers do offer relatively up-to-date in-play betting odds, for example, next goalscorer odds at half-time in a football match. However, the high volatility of the market increases the risk to the bookmaker, so they may decide to play it safe by offering poor odds. Consequently, in-play betting is much more prevalent on exchanges, where the burden is on the trader and one can therefore find better odds.

Tennis is one of the most heavily traded sports on betting exchanges and continues to grow at a remarkable pace. For example, during the Wimbledon 2006 final between Roger Federer and Rafael Nadal, Betfair processed approximately £25 million worth of matched trades. For comparison, the Wimbledon 2011 men's final between Novak Djokovic and Rafael Nadal matched £40 million worth of bets. Such interest is not limited to Grand Slam finals either; during the women's semi-final of the Sony Ericsson Open 2011 between Maria Sharapova and Andrea Petkovic, £10 million was traded on Betfair[3].

Tennis is well suited to in-play trading on exchanges since points are played at a steady rate, are clearly separated, and are consistently won and lost by both participants, leading to frequent dramatic changes in fortune for players but at relatively predictable intervals. Consequently, the odds can very quickly swing back and forth as traders react to on-court events, generating potential money-making opportunities. This, in combination with the fact that tennis markets have the ability to offer only a few outcomes (e.g. in the Match Odds market, there are only two outcomes, either one player wins or the other does), ensures its popularity. Around 80% of money waged on tennis matches is bet while the match is in progress[4].

Player Retirement Payout Policies

Betting companies take different approaches when dealing with the issue of player retirement. Typically, they fall into one of four categories (with regards to Match Odds markets)[5]:

Category 1: Ball-Served Rule For a bet to stand, at least one ball must have been served, e.g. Ladbrokes.

Category 2: One-Set Rule For a bet to stand, at least one set must have been completed in the match. However, if a player retires from the match before the first set is over, all bets are cancelled and stakes are refunded, e.g. Betfair.

Category 3: Two-Sets Rule For a bet to stand, at least two sets must have been completed in the match, e.g. TheGreek.

Category 4: Match-Completed Rule The entire match must be completed for a bet to stand, e.g. Paddy Power.

Markets offered on the outcomes of individual games or sets are always rendered void unless the result has been unconditionally determined.

Betfair

UK-based betting company Betfair[2] was the world's first betting exchange. It was launched in 2000 and now boasts over 4 million customers, handles in excess of £50 million a week[6], and possesses a dominant 90% share of the betting exchange market[7]. Betfair's popularity should ensure that we have markets with as high *liquidity* (i.e. plenty of willing buyers and sellers) as possible to analyse, where the odds accurately reflect the evolution of a tennis match.

Betfair falls into *Category 2* of the tennis betting retirement payout policies with respect to its in-play Match Odds market[8]. Its Set Betting market is always voided in the event of a retirement since if the match is not finished, we do not have a final score.

[2] http://www.betfair.com

1.4. Premise

The Betfair exchange provides us access to only Set Betting and Match Odds markets. Figure 1.2 displays the evolution of implied match-winning probabilities for Novak Djokovic when he played Rafael Nadal in the US Open 2011 Men's Final. The blue line shows implied probabilities extracted from the Betfair Set Betting market, the red line shows implied probabilities extracted from the Betfair Match Odds market, and the green line shows the positive difference of the Set Betting probability minus the Match Odds probability. As you can see, both markets are closely matched. This is intuitive since the probability of Djokovic winning the match should be the same as the sum of the probabilities of the final score being 3-0, 3-1, or 3-2 in Djokovic's favour. This is especially true in high profile matches such as this one where the markets are very liquid and millions of pounds are being traded in both, leading them to produce very accurate odds.

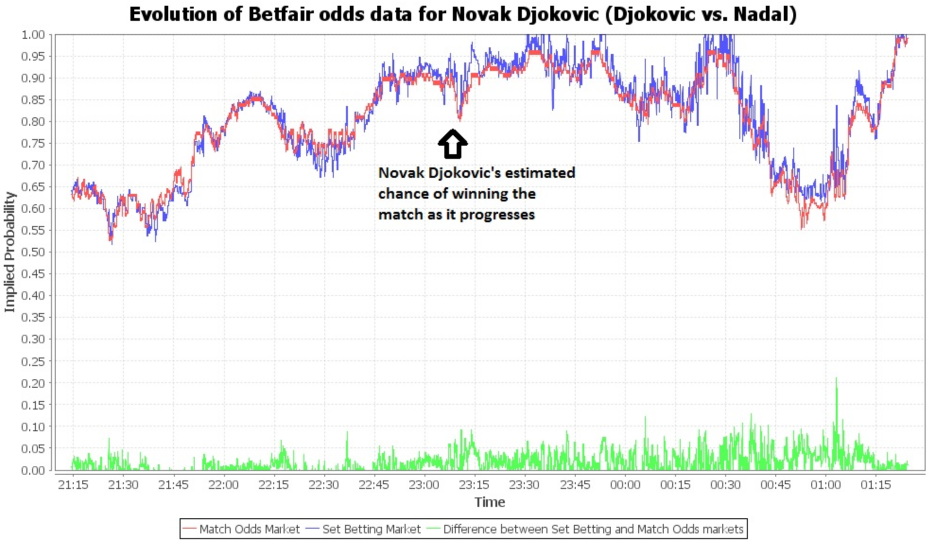


Figure 1.2.: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Novak Djokovic - *Djokovic vs. Nadal (US Open 2011 Men's Final)*

Compare with Figure 1.3 which displays the evolution of implied match-winning probabilities for Andy Murray when he played Michael Berrer in the French Open 2011 Men's Third Round. In particular, observe where the Match Odds probability of Murray winning the match suddenly drops *almost 60%* (pointed out by the arrow). This phenomenon in the odds data occurred during a *single point in the match*. There is no possible event that could have caused this huge swing in the market so quickly other than the *injury* that was suffered by Andy Murray. We quote from the *BBC Sport*[3] live text commentary of the match during this point:

- "Big, big trouble for Andy Murray, who has gone over on his right ankle and looks in real pain. Not sure whether he will be able to continue. Unbelievable."
- "We are going to have a medical time-out while Murray has treatment. He slipped as he ran in to put away a forehand, and the replays are not very pleasant to watch."

In this case, the injury did not cause Murray to retire from the match and he went on to win in straight sets. Nevertheless, it is clear the market reacted to this event and its opinion on Murray's chances of winning the match was severely affected.

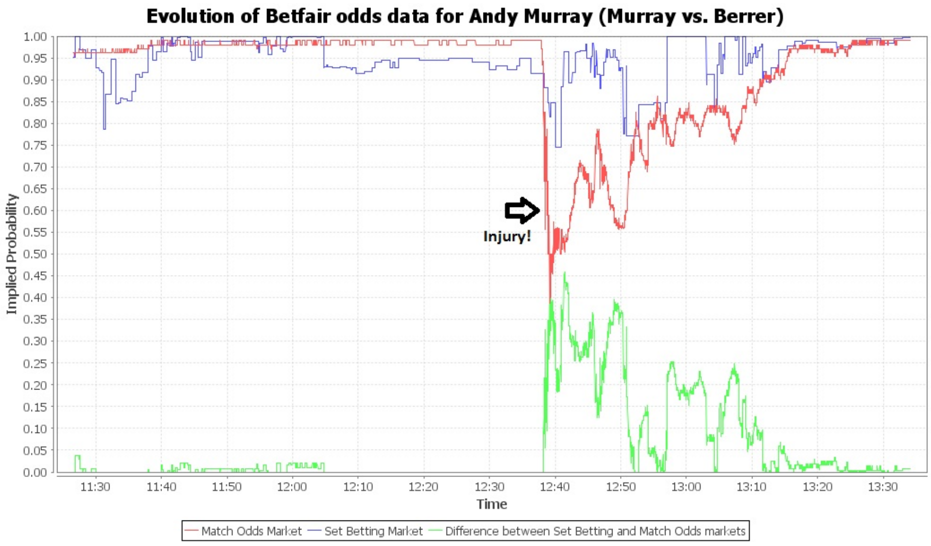


Figure 1.3.: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Andy Murray - *Murray vs. Berrer (French Open 2011 Men's Third Round)*

[3] http://www.bbc.co.uk/sport

1.5. Contribution

The primary goal of this project is to attempt to create a new model for tennis matches that takes into account the risk that a player will retire from the match as it progresses. This is not a concept that has been researched before; this is the world's first model of its kind. This knowledge makes the idea an exciting prospect, but also ensures that the investigation will involve significant trial and error.

We can observe the evolution of the Betfair Set Betting and Match Odds markets in order to help us *quantify* a player's risk of retirement. This will provide us with parameters to input into our model. For example, we can see from the Murray vs Berrer match when the injury occurred that the Match Odds implied probability dropped sharply but the Set Betting implied probability remained relatively stable. If a player becomes injured and concedes the match at any point, traditional bookmakers will usually cancel all bets in the Match Odds and Set Betting markets (they fall into Category 4). Betfair, however, will still pay out on bets to win, but will refund money on final score bets (since the match has a winner but no final score) as long as at least one set has been played. In this way, the Betfair Set Betting market simulates a Match Odds market that ignores risk of retirement whereas the actual Betfair Match Odds market takes into account this important factor, given one set has been played. Consequently, we see discrepancies between the match and final score odds in such scenarios (possibly more so beyond the first set) as we do in the Murray match.

The theory is that the probability of a player retiring *at some point during the remainder of the match* should be somehow encapsulated within the difference between the two markets, i.e. the combined opinions of all the traders betting in those markets (the wisdom of the crowd). All the complex factors which lead to a player deciding to retire such as the state of the match, the seriousness of the injury, the importance of the match, even their past history of retirements, should be *some function* of this gap in the odds. We will examine the possibility of observing when a potential injury might have occurred in a match and extracting a value for the likelihood that a player will retire at any point during the remainder of a match by analysing the difference between these markets in-play.

- **Can we create a more accurate model for predicting the winner of a tennis match than has been achieved previously by taking into account risk of retirement?**
- **Can we use our model and the odds data to predict the evolution of markets that use other player retirement payout policies from any point in a match?**

Applications

The project alone is an attempt to break new ground in the fascinating area of sports modelling and this is our prime motivator. Nevertheless, we hope that the findings will be useful to those with an interest in the growing tennis trading business since injuries are phenomena that have a massive financial impact on in-play betting markets. The ability to predict the evolution of the market (and even markets using different retirement payout policy rules) in the presence of injuries is an important factor that current models are lacking. We could also give valuable insight into the efficiency of the in-play markets we are studying.

Our results could be useful to tennis coaches and personal trainers, who might be interested in the likelihood their players may retire hurt after suffering a potentially serious injury during a match. For example, Jayanthi, O'Boyle, and Durazo-Arvisu (2009)[9] examined over 28,000 matches from US Tennis Association junior national tennis tournaments in 2005 and found a much greater risk of medical withdrawal in matches beyond the fourth round. Such information may also be useful for media coverage of tennis. Imagine if commentators and pundits could estimate the likelihood one of the players might retire and display it to the viewer.

We may also be able to apply our theories to other sports. It could be relatively straightforward to model retirement risk in similar point-based racquet sports such as badminton, table tennis, or squash, given there is the necessary data. It would require much further research if we wanted to look at team sports such as football, rugby, basketball, or cricket, where it is extremely difficult to measure the impact an injury to an individual player has on the team.

There are even parallels to the topic of natural disaster modelling which has a large impact on the insurance industry. One could arguably see an injury as a 'disaster' (the greater the severity, the lower the frequency) or even a player as a precious commodity.

1.6. System Overview

In order to be able to observe the Match Odds and Set Betting in-play markets for a given match, we will require access to Betfair odds data. It would be inefficient to use real-time data; we do not know if an injury will occur in a match yet to be completed. We will make use of a software platform called *Swarm* by *FracSoft*[3], which allows users access to a large database of historical Betfair odds information.

In order to be able to analyse whether there is any relationship between an odds discrepancy in a match and a risk of retirement (and also decide which matches to gather data for), we will also need information concerning whether a player sustained an injury during a given match. *TennisInsight.com* as well as sports news reporting websites such as *BBC Sport* will be of great assistance since they both compile reports on all important events occurring in major tennis tournaments.

The investigation will be divided into four main stages which will hopefully combine to form a coherent story.

1. **Parsing and processing Match and Set Betting historical Betfair odds data** for top-level tennis matches in order to calculate and compare the match-winning probabilities each market produces. We will only study matches for which it was reported that one of the players sustained an injury while playing (Chapter 3).
2. Creating **a new model for a tennis match** that takes into account the probabilities of each player retiring *on each point* by incorporating extra parameters and defining additional outcomes of a match compared to a standard tennis model. We will interpret the probability a player will retire at some point during the remainder of a given match as a function of any gap between the two Betfair in-play markets (Chapter 5).
3. Finding a way of **calculating or approximating suitable values for the parameters required for the model** with the aid of the odds data and real-world averages (Chapter 6).
4. Preparing results data to accompany the corresponding odds data as well as our chosen parameters in order to **evaluate and refine the model** (Chapter 7).

We reason that there is an element of randomness with respect to injury occurrence in a tennis match. Consequently, we investigate the use of a suitable probability distribution when modelling the risk of retirement. We use the well-known *tennis formulae* presented in various academic papers as well as investigating both analytical and numerical solutions in solving the modelling problem and then searching for appropriate parameters.

To test our system on odds data for a real-world tennis match, we also need to know how the given match played out on a point-by-point basis so that we are able to input the current score into our model at any time during the match. We find that the only viable way of entering such data into our system is manually. Furthermore, as with any modern tennis model, we require the point-winning probabilities on serve for each player as input. We estimate these values using a combination of the odds data, real-world averages, and ideas discussed in previous investigations into these variables.

Figure 1.4 below shows a high-level overview of what our system will look like.

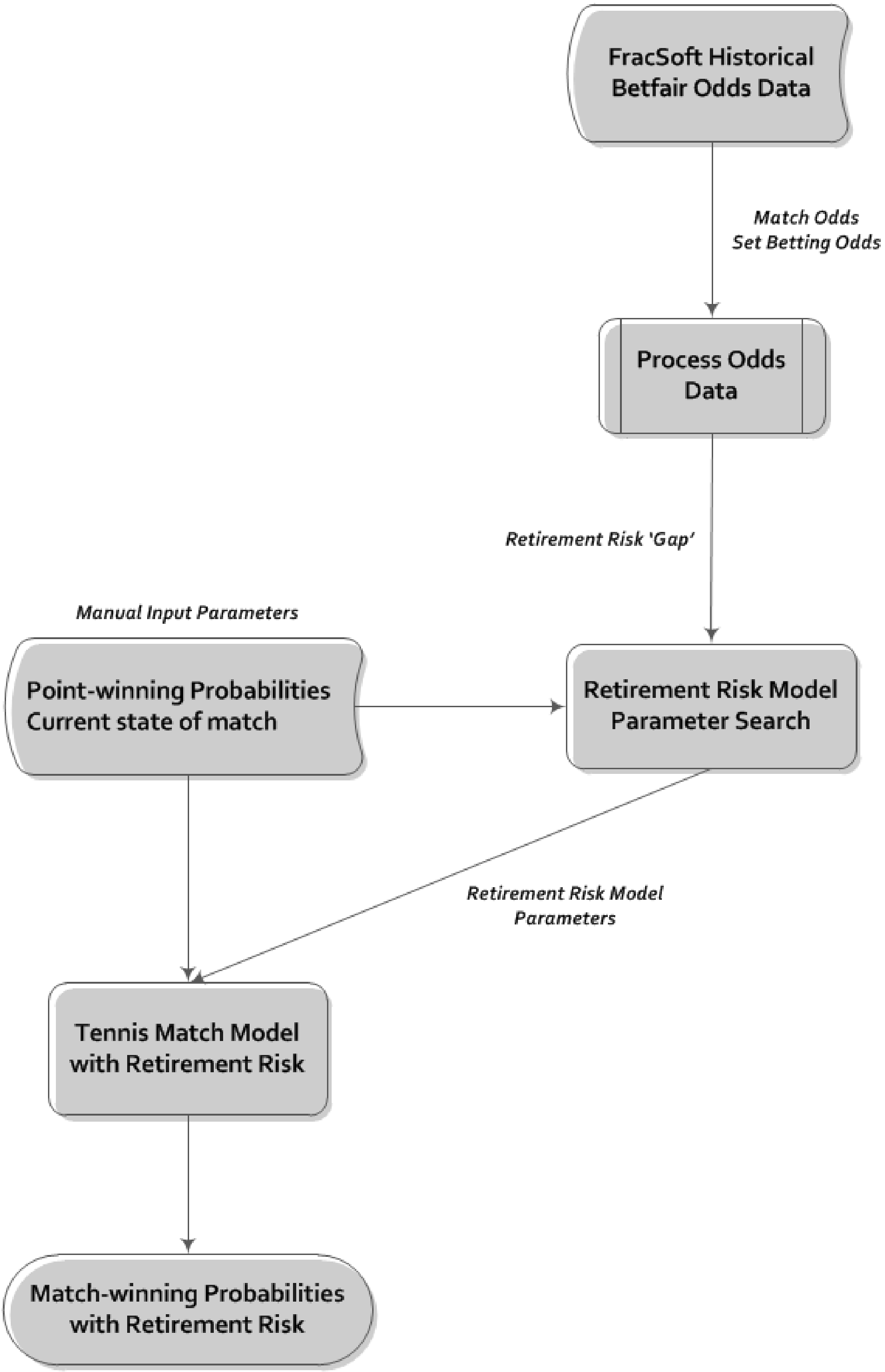

FracSoft Historical Betfair Odds Data
Match Odds
Set Betting Odds
Process Odds Data
Retirement Risk 'Gap'
Manual Input Parameters
Point-winning Probabilities
Current state of match
Retirement Risk Model Parameter Search
Retirement Risk Model Parameters
Tennis Match Model with Retirement Risk
Match-winning Probabilities with Retirement Risk

Figure 1.4.: High-level system overview diagram

Background 2

Tennis is one of the world's most popular individual sports. It would be helpful when reading this report to have a good understanding of how the sport is played so we briefly summarise the scoring system in Appendix A.

Tennis happens to be a relatively simple game to model in comparison with other highly complex sports such as football or cricket, as it is just a series of discrete repeated contests, i.e. points, and calculations essentially boil down to the probability each player has of winning a given point. The scoring system has a fixed number of hierarchical states; points are nested within games, games within sets, and sets make up a match.

2.1. Betting Odds

Equally as important to this investigation are the mechanics of betting odds. Everyday usage of odds in the UK usually comes in the form of the odds against for a particular event or outcome. This is commonly displayed as the ratio of two integers:

$$n/m$$

where n and m represent the relative chance of the event *not occurring* or *occurring* respectively. For example, if you roll a die, the odds of getting a six is 5/1 (5 *to* 1). This is equivalent to the probability:

$$\frac{m}{(m+n)} = \frac{1}{6}$$

of the event occurring. In terms of wagering, a bet of m currency units would return a profit of n units, e.g. an outrageous bet of £10 on rolling a six could return a profit of £50, as well as the original £10 stake, if one is lucky enough.

Betting Exchanges such as Betfair often use a decimal representation of odds, which is more common in Europe. Given a probability, p, of an event occurring:

$$DecimalOdds = \frac{1}{p}$$

Continuing with our previous die example, the decimal odds for rolling a six would be:

$$\frac{1}{\frac{1}{6}} = 6$$

Since:

$$TotalReturn = OriginalStake * DecimalOdds$$

you must subtract the original stake from your return in order to calculate your profit. To convert decimal odds to a percentage chance:

$$PercentageChance = \frac{100}{DecimalOdds}$$

e.g.

$$\frac{100}{6} = 16.67\%$$

2.2. Exchange Trading

Less widely understood are the intricacies of trading on betting exchanges. Users buy and sell (or *back* and *lay*) with respect to the outcome of sporting events. When a user lays, they offer up a bet on the exchange with a stake and odds of their choosing. In essence, they are saying that an outcome will not occur, and another user may back that bet. For example, a lay of £10 at odds of 3.1 gives you a maximum profit of £10 (you just keep the stake if someone takes the bet and loses), and a maximum liability of £21 (you must pay out £10 ×3.1 = £31 should someone take the bet and win). Similarly, you can back a bet and another user can match it with a lay. Just as with shares on a stock exchange, traders create strategies to buy and sell (or trade their position in the market) in such a way that profit is guaranteed regardless of the outcome (achieving a *green book*). For example, say we back Andy Murray to win for £100 at odds of 1.4 against Gael Monfils. Murray wins the first set 6-4 and his price drops to 1.2. We can now hedge our initial bet with a lay for £116.67. As shown in Figure 2.1, this guarantees a net profit of £16.67 before commission[10].

	Odds	**Stake**	**Murray Win Profit**	**Murray Loss Profit**
Back	1.4	£100	£40	-£100
Lay	1.2	£116.67	-£23.33	£116.67

Total Liability	£123.33
Total Return	£140
Net Profit	£16.67

Figure 2.1.: Tables displaying an in-play betting opportunity to guarantee a profit for a match involving Andy Murray

One might notice that backing and laying are logical opposites. In a horse-race, laying one horse is the same as backing any other horse to win. In our example above, laying Andy Murray is the same as backing Gael Monfils to win.

2.3. Interpreting Odds Information

The Betfair tennis odds data supplied by FracSoft is made up of a number of different values. Each specific outcome that you can wager on is known as a *Selection*, e.g. for one of the players to win, or for the final score to be 3-1 to one of the players. The values provided for each selection include:

Last Price Matched This represents the odds of the last back bet that was matched by a corresponding lay bet or vice versa; it is the value of the last odds that were traded.

Best Back and Lay Odds The Betfair Exchange displays the three best available back and lay odds and their respective volumes (stakes). The lowest odds represent the best lay price that someone is willing to offer whereas the highest odds represent the best back price that someone is willing to offer.

Market Percent The proportion of the total volume matched on a given selection.

Total Volume Matched The total amount of money wagered in all matched bets.

We will want to compare the Set Betting and Match Odds markets directly at any timepoint during the match in order to compare them accurately. In order to do this, we could compare the Last Price Matched (LPM) values. This could be seen as an accurate indicator of the true odds since it represents a bet taken rather than just offered. On the other hand, the last price matched is often (naturally) a step behind the best back and lay prices (bets that have been offered but not yet taken).

Another option would be to make use of the best available back and lay prices. The difference between them is known as the *market spread*. This can be thought of as the amount you will lose if you both back and lay the same player at the same time. Consider the following scenario: you back a player to win for £10 at odds of 2.8 and also lay them for £10 at 3.0 simultaneously. As Figure 2.2 shows, at best you could break even and at worse you will make a loss of £2.

	Odds	Stake	Player Win Profit	Player Loss Profit
Back	2.8	£10	£18	-£10
Lay	3.0	£10	-£20	£10

Figure 2.2.: Table illustrating the concept of market spread

This is analogous to *bid-offer spread* for stock traded on financial exchanges. A small spread indicates a less risky market for buyers since prices only have to rise by a small amount before you can sell to make a profit. A large spread indicates uncertainty in the market. Traders offer and accept 'safer' prices because they perceive that it might be more difficult to make a profit later on. For example, say that BP is quoted at 676p-677p. This means you can buy shares in BP at 677p (the *bid* price) each and sell them for 676p (the *offer* price) each. Consequently, if you buy shares at that bid price, the offer price only has to rise marginally for you to be able to then sell them at a profit[11]. This measure could be seen as more indicative to the current state of the match as traders immediately respond to events that are occurring by offering up-to-date odds. This could be helpful to us as we want to monitor instinctive reactions to a possible injury.

2.4. On Injury Risk in Professional Tennis Matches

Many papers have been written on tennis-related injuries. In particular, Johnson and McHugh (2006)[12] attempted to quantify the demands in professional male tennis by analysing the number and type of strokes played per game for 22 players from three Grand Slams. They found that the serve was the predominant stroke played in service games (up to 60%) whereas topspin forehands and backhands were more frequent when receiving. The 2003 US Open winner hit over 1000 serves during his seven matches at the tournament. More strokes are played at the French Open than Wimbledon due to the relative speeds of the clay and grass court surfaces leading to longer rallies at Roland Garros. Importantly, Johnson and McHugh discuss the strain that playing a point inflicts on the body. They report that over 50% of world-class tennis players experience shoulder discomfort during their career and 80% of these cases stem from overuse. Stroke production in tennis involves generating repetitive forces and motions that are of high intensity and short duration. For example, the serve is the most strenuous stroke on the upper extremity with internal rotation velocities of the humerus reaching 2420% for elite players during the acceleration phase. This, coupled with the relentless demands the ATP (Association of Tennis Professionals) and WTA (Women's Tennis Association) tours place on players, suggest that it is no surprise injuries are an issue in the world of professional tennis.

2.5. On Analysis of In-Play Tennis Betting Odds

Easton and Uylangco (2010)[13] modelled a tennis match and compared the calculated match-winning probabilities of each player to those implied by the Betfair in-play match odds on a point-by-point basis (using data from 49 matches played at the 2007 Australian Open). They utilised the mid-point of the favourite's best back and lay prices (as the odds for the favourite usually possess a smaller spread) to calculate the implied match-winning probability. They found an extremely strong correlation between the model and implied probabilities, the only noticeable period of variation coming when one contestant played at a level vastly different from their expected ability for a short time (see Figure 2.3). This property is vital to our work as it signifies a high level of market efficiency, i.e. the implied odds for an outcome are close to the actual odds. Similarly, Servan-Schreiber, Wolfers, Pennock, and Galebach (2004)[14] found that prediction markets are superior to human experts at forecasting the outcomes of NFL football games.

In addition, Easton and Uylangco found that the market prices even incorporated information about the differing importance of various points with the anticipation of service breaks. This bodes well for our hypothesis that market prices encapsulate information about a tennis match, in particular, the probability of player retirement. On the other hand, they also discovered that the market underestimates the tendency for players to lose a greater percentage of first points after conceding a break of service then points lost on average while receiving, instead displaying mechanistic responses to the points, i.e. not anticipating a less than normal probability of winning such points. This suggests that the in-play markets are not flawless and we will bear this in mind during our investigation. There was no such evidence of a biased reaction with respect to the information provided by a player holding serve, however.

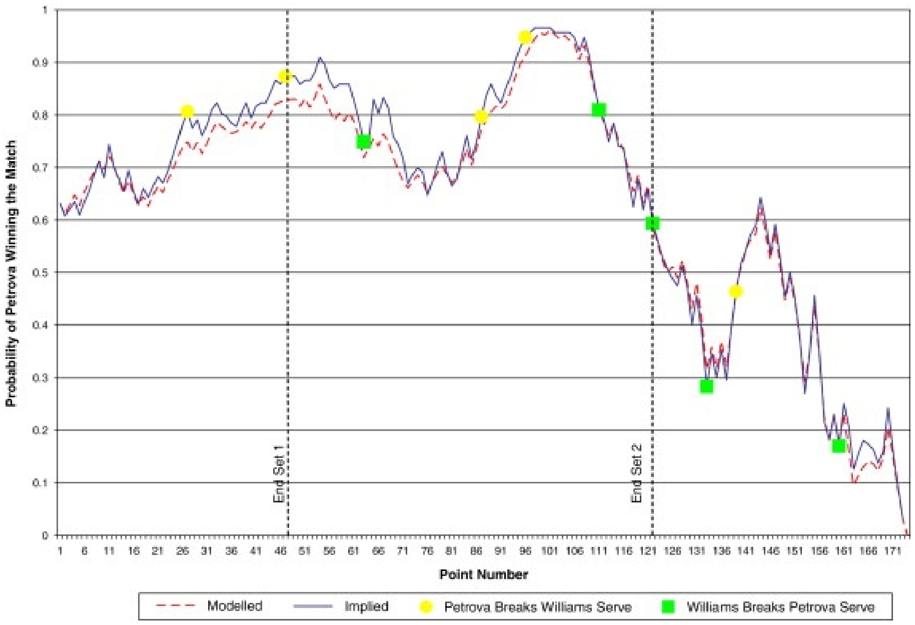


Figure 2.3.: Probability of Nadia Petrova winning the match - *Petrova vs. Williams (Australian Open 2007 Women's Third Round)*

Huang (2011)[15] investigated the possibility of accurately inferring the current score of a tennis match from the in-play match odds. He finds that entire sets can be tracked with very few errors, further indicating that in-play odds successfully reflect the evolution of a tennis match.

2.6. On Modelling Tennis

There have been many past works on the topic of modelling tennis (although there are none that incorporate retirement risk as a factor). For instance, Klaassen and Magnus (2003)[16] create a program dubbed TENNISPROB, which is capable of calculating match-winning probabilities for tennis matches using Wimbledon singles data from 1992-1995, official rankings, and subjective judgement to estimate variables such as the point-winning probabilities.

O'Malley (2008)[17] presents what are considered the *tennis formulae*. The tennis formulae are a hierarchical series of equations that compute the probability a given player will win a tennis match given the probabilities that the given player and his/her opponent will win any of their service points. Combined together are individual formulae for the probabilities of winning games, sets, and tiebreaks. O'Malley's tennis formulae are concise and intuitive and an excellent introduction to the area of research. Although O'Malley focusses on using his models for pre-play scenarios, for example, for a whole match or whole set, he hints at using recursion (conditional probabilities) for generating match-winning probabilities from any given current score in a match. We will be looking to extend or improve upon the performance of such a model.

Newton and Keller (2005)[18] more comprehensively explore the use of recurrence relations to model tennis, utilising them to calculate probabilities of winning tournaments and also proving explicitly that the probability of winning a set or match does not depend on which player serves first. Barnett and Clarke (2002)[19] experiment with the same idea in Microsoft Excel. They investigate using six parameters rather than just the two point-winning probabilities taking into account service faults. Consequently, for each player they input the probability of a successful first serve, the probability of winning a point on first serve, and the probability of winning a point on second serve. This greatly complicates the base model and since we are adding further complexity with the probability of retirement, we shall concentrate on extending the simpler, two parameter version. Barnett, Brown, and Clarke (2003)[20] followed this up with a investigation into player momentum in tennis matches by slightly perturbing point-winning probability depending on how much the given player is leading or trailing the match by (essentially introducing a dependency between points).

Equivalent is the idea of using a hierarchy of Discrete-Time Markov Chains (DTMC) to model a tennis match proposed by Liu (2001)[21]. Each possible score in a match is represented by a state in the system. Certain equivalent states are combined in order to reduce the size of the state space, e.g. 30-30 and deuce. However, Liu assumes that the probability a player wins a point stays the same regardless of whether that player is serving or receiving. It is well known that the server has a significant advantage in tennis (which is why breaking your opponent's serve is such a cause for celebration), so we take this factor into account in our model.

The vast majority of models use the assumption that points played in a tennis match are independent and identically distributed, e.g. the probability of winning the current point is unaffected by any previous point. Klaassen and Magnus (2001)[22] analysed almost 90,000 points played at Wimbledon over 3 years and found that, although points played are not completely independent, the assumption that they are is still a good approximation as the divergence from iid is small. Note that this assumption is necessary in order to avoid violating the *Markov Property* when modelling a tennis match as a DTMC (see Appendix E).

2.7. On Estimation of Point-winning Probabilities

Most models of tennis matches have in common the use of point-winning probabilities on serve for each player as parameters. Many researchers such as Klaassen and Magnus (2003) and Barnett and Clarke (2005)[23] use ranking data, historical statistics, and subjective judgement to estimate point-winning probabilities. Klaassen and Magnus (2000)[24], from the 258 mens and 223 womens Wimbledon singles matches they analysed, found that the average sum of two players' point-winning probabilities on serve was 1.29 for men and 1.12 for women (implying there is greater service dominance in the men's game). Huang suggested a combination of these ideas. If we can use the match odds at any point in the match to estimate the current point-winning probability of Player A, P_A, we can estimate that the equivalent probability for Player B, P_B, is either 1.29 - P_A or 1.12 - P_A. Although point-winning probabilities will vary for a player from match to match for reasons such as surface, fitness, form, strength of opponent, etc, Newton and Aslam (2009)[25] showed that they can be modelled as Gaussian-distributed random variables with relatively low variance. Interestingly, Marek (2011)[26] found that it is the *difference*, δ, between the point-winning probabilities of two players which determines who is more likely to win the match, and not the absolute values. One can express the match-winning probability of Player A, for instance, as a linear function of $\delta = P_A - P_B$, within bounds of $-0.1 \leq \delta \leq 0.1$ with reasonable accuracy. We show this in Figure 2.4.

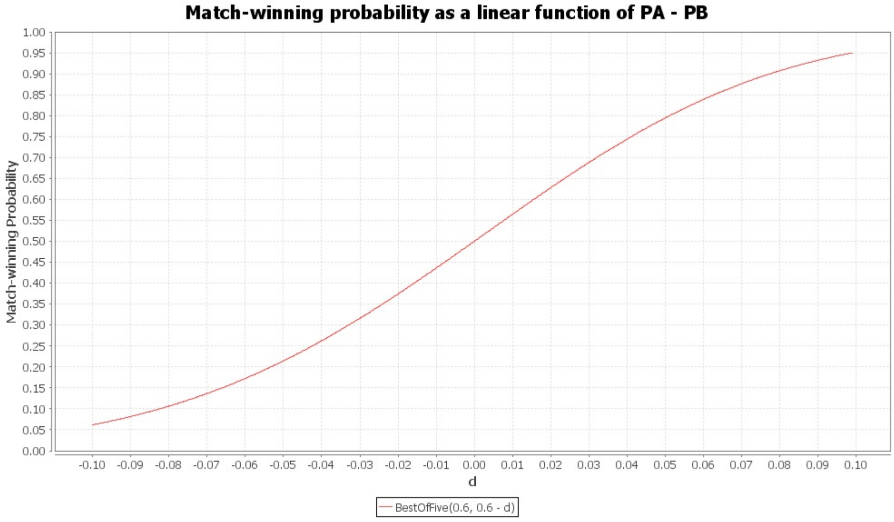


Figure 2.4.: Match-winning probabilities for the situation where $P_A = 0.645$ and $P_B = P_A - \delta$, where δ ranges from -0.1 to 0.1

Acquiring Odds Data 3

There would have been no point proceeding with the investigation if we were not sure that information about a player's risk of retirement from a match was could be found within the betting odds. Attempting to confirm that a significant gap between the Betfair Set Betting and Match Odds markets is created when a player suffers an injury during a match presented a significant challenge.

3.1. Identifying Matches

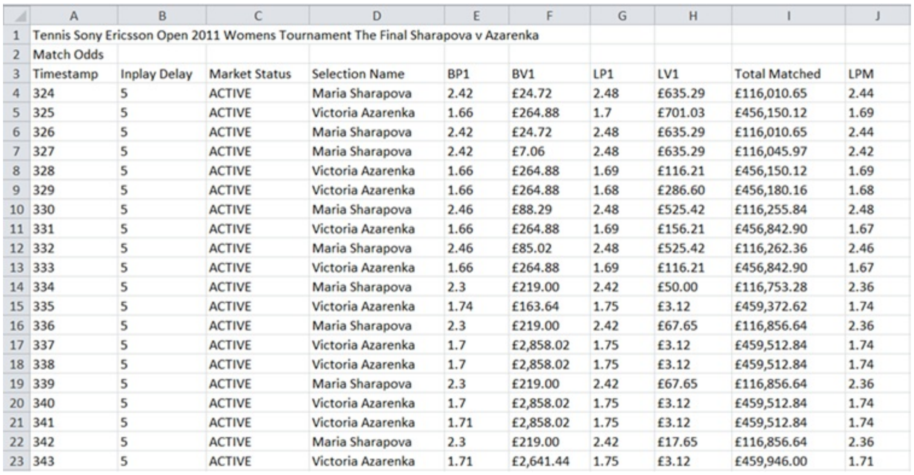
We required odds data from tennis matches where one of the players suffered a clear injury, i.e. they require treatment from a trainer. In order to help find such matches, we scoured the internet for sports articles and news reports describing the events of matches from major tournaments. Particularly useful was the popular *BBC Sport* website, which even provides live text commentary on many sporting events.

As mentioned previously, we used a piece of software called *Swarm* created by *FracSoft* to gather historical Betfair tennis odds data. FracSoft[3] is a UK-based company which seeks to provide trade execution and analysis tools for use with electronic sports trading exchanges such as Betfair. Their *Swarm* software platform provides access to large database of historical Betfair data going as far back as 2006. It was far more straightforward and reliable to analyse the in-play odds data retrieved from Swarm as opposed to recording real-time data. It also allowed us to repeatedly test our software with match data of our choosing as opposed to only matches that are currently being played. In Chapter 7, we test our new tennis match model against a subset of these matches.

Swarm allows the customised exporting of odds data in the *Comma-Separated Values (CSV)* file format such as in Figure 3.1 below. We used Java CSV parser library *OpenCsv*[1] to parse such files. It was helpful to visualise the evolution of the odds data throughout matches when trying to understand and interpret market behaviour. We used open source Java chart library *JFreeChart*[2] to generate graphs of our parsed data against the accompanying timestamps.

[1] http://opencsv.sourceforge.net

[2] http://www.jfree.org/jfreechart/



	A	B	C	D	E	F	G	H	I	J
1	Tennis Sony Ericsson Open 2011 Womens Tournament The Final Sharapova v Azarenka									
2	Match Odds									
3	Timestamp	Inplay Delay	Market Status	Selection Name	BP1	BV1	LP1	LV1	Total Matched	LPM
4	324	5	ACTIVE	Maria Sharapova	2.42	£24.72	2.48	£635.29	£116,010.65	2.44
5	325	5	ACTIVE	Victoria Azarenka	1.66	£264.88	1.7	£701.03	£456,150.12	1.69
6	326	5	ACTIVE	Maria Sharapova	2.42	£24.72	2.48	£635.29	£116,010.65	2.44
7	327	5	ACTIVE	Maria Sharapova	2.42	£7.06	2.48	£635.29	£116,045.97	2.42
8	328	5	ACTIVE	Victoria Azarenka	1.66	£264.88	1.69	£116.21	£456,150.12	1.69
9	329	5	ACTIVE	Victoria Azarenka	1.66	£264.88	1.68	£286.60	£456,180.16	1.68
10	330	5	ACTIVE	Maria Sharapova	2.46	£88.29	2.48	£525.42	£116,255.84	2.48
11	331	5	ACTIVE	Victoria Azarenka	1.66	£264.88	1.69	£156.21	£456,842.90	1.67
12	332	5	ACTIVE	Maria Sharapova	2.46	£85.02	2.48	£525.42	£116,262.36	2.46
13	333	5	ACTIVE	Victoria Azarenka	1.66	£264.88	1.69	£116.21	£456,842.90	1.67
14	334	5	ACTIVE	Maria Sharapova	2.3	£219.00	2.42	£50.00	£116,753.28	2.36
15	335	5	ACTIVE	Victoria Azarenka	1.74	£163.64	1.75	£3.12	£459,372.62	1.74
16	336	5	ACTIVE	Maria Sharapova	2.3	£219.00	2.42	£67.65	£116,856.64	2.36
17	337	5	ACTIVE	Victoria Azarenka	1.7	£2,858.02	1.75	£3.12	£459,512.84	1.74
18	338	5	ACTIVE	Victoria Azarenka	1.7	£2,858.02	1.75	£3.12	£459,512.84	1.74
19	339	5	ACTIVE	Maria Sharapova	2.3	£219.00	2.42	£67.65	£116,856.64	2.36
20	340	5	ACTIVE	Victoria Azarenka	1.7	£2,858.02	1.75	£3.12	£459,512.84	1.74
21	341	5	ACTIVE	Victoria Azarenka	1.71	£2,858.02	1.75	£3.12	£459,512.84	1.74
22	342	5	ACTIVE	Maria Sharapova	2.3	£219.00	2.42	£17.65	£116,856.64	2.36
23	343	5	ACTIVE	Victoria Azarenka	1.71	£2,641.44	1.75	£3.12	£459,946.00	1.71

Figure 3.1.: Sample CSV file exported from Swarm - *Sharapova vs. Azarenka Match Odds (Sony Ericsson Open 2011 Women's Final)*

3.2. Comparing In-Play Tennis Betting Markets

We started off by taking the ideal and most straightforward approach to processing the odds data. We utilise just the Last Price Matched (LPM) values as they give the most accurate indicator of the true odds. To extract a value for the gap between the two markets at a specific given time in a match we:

1. Convert the current LPM value in the Match Odds market for the selection corresponding to our target player into a match-winning probability.
2. Convert the current LPM values in the Set Betting market of all the selections corresponding to a victory for our target player, e.g. 2-0, 2-1, into probabilities and sum to form a match-winning probability.
3. Extract the probability produced by the Match Odds minus the probability produced by the Set Betting odds.

We quickly discovered that much of our exported odds data suffered from the problem of low *market liquidity*, where traders are reluctant to place bets. This can often be the case for early-round or minor tournament matches between two little-known players or the underdog in matches where one player is heavily favoured. Consequently, we find the most accurate data with respect to matches where the heavy favourite suffers an injury or high profile matches where either one of the competitors required treatment. Importantly, we also note that the more complex Set Betting market is often significantly less popular than the straightforward Match Odds market for most matches. For example, in the French Open 2010 first round match between Andy Murray and

Richard Gasquet, almost £20 million was matched in the Match Odds market but only £60,000 was matched in the Set Betting market.

This issue is complicated further by the fact that we are *directly* comparing two different markets throughout their lifetimes, one of which has up to 6 outcomes (i.e. in the Set Betting market for a 5-set match, each player can win 3-0, 3-1, or 3-2). For every LPM value in the Match Odds market during a match that we convert into a match-winning probability for the player in question, we need a corresponding value drawn from the Set Betting market using LPM values that were created from bets matched at *(approximately) the same time*.

This naturally begs the question of how to deal with the situation where we do not have all the corresponding LPM values in the Set Betting market. A possible solution is to make use of the best available back prices, which give the most up-to-date indicator of what the market thinks. In order to acquire the highest quality back prices, we must use the idea of *crossmatching*.

3.2.1. Crossmatching and Virtual Bets

We have previously established that backing and laying are natural opposites, particularly if there are only two possible outcomes. We also know that the majority of money traded is bet on the favourite. One might think that this would make it unreasonably difficult to get a bet on the underdog to be matched. The crossmatching system implemented by Betfair is a set of rules dictating how to match unmatched bets. For instance, if you back Player A at odds of 2.0, Betfair will attempt to match your bet with another trader's lay offer at odds of 2.0 or better or *any possible match that is equivalent to a lay bet on Player A*, e.g. a back on Player B at odds of 2.0 or better. Consequently, bets can either be matched by another bet from an actual trader or a *virtual bet* created by Betfair using the crossmatching rules[27].

The formula for calculating the odds for a virtual back bet on a specific outcome by laying all the other selections as given on the *Betfair Developers Programme*(BDP) website[28] is as follows:

$$BackPrice = \frac{1}{N - \sum_{\iota \in L} \frac{1}{\iota}}$$

N is the number of possible winning outcomes (1 in our case)
L is the set of the best available lay prices of all the other selections

To calculate the amount of money that should be offered at these odds:

$$BackOffer = \frac{min(LayOdds1 * Stake1, LayOdds2 * Stake2, ...)}{BackPrice}$$

In the Match Odds market, there are only two possible selections so crossmatching is simple. Say we lay Victoria Azarenka at odds of 3.5 to win for £10. This is equivalent to backing opponent Laura Robson at odds of:

$$BackPrice = \frac{1}{1 - \frac{1}{3.5}} = 1.4$$

At these odds we can match a stake of:

$$BackOffer = \frac{3.5 * \pounds10}{1.4} = \pounds25$$

This is intuitive since if you back Laura Robson at for £25 at 1.4 you could potentially stand to win a profit of:

$$(\pounds25 * 1.4) - \pounds25 = \pounds10$$

On the other hand, if you back Victoria Azarenka for £10 at 3.5 you could potentially stand to win a profit of:

$$(\pounds10 * 3.5) - \pounds10 = \pounds25$$

Things get a little more complicated when we start to consider more than two selections. Remember that higher odds are better when you are backing as you get a greater payout if you win whereas lower odds are better when you are laying so you pay out less if you lose. We run through an example similar to that given on the BDP website[28]. We omit calculations of the appropriate stakes as we are only interested in the odds information.

Selections	**Back**			**Lay**		
England	1.01	1.5	**1.5**	**2.0**	2.5	1000.0
West Indies	1.01	2.4	**2.5**	**3.0**	20.0	1000.0
The Draw	1.01	3.0	**5.0**	**10.0**	50.0	1000.0

Figure 3.2.: Example state of what an in-play Match Odds market for an England vs West Indies Test Match might look like

Looking at Figure 3.2, say we place a *large* back bet on the The Draw at 1.01. This causes the stake to be split between the odds of 1.01, 3.0, and 5.0, with anything remaining being left unmatched. With crossmatching, we can get better prices. Notice the best available lay offers for England (odds of 2.0 to win) and the West Indies (odds of 3.0 to win). These are previously made back bets that have not yet been matched by a trader willing to lay them. These two back offers can be matched against our bet on The Draw at a price of **6.0**, since odds of 2.0, 3.0, and 6.0 form a 100% book:

$$\frac{1}{2}+\frac{1}{3}+\frac{1}{6}=1$$

Say that after matching these bets with appropriate stakes, the market now looks like this:

Selections	**Back**			**Lay**		
England	1.01	1.5	**1.5**	**2.5**	1000.0	
West Indies	1.01	2.4	**2.5**	**3.0**	20.0	1000.0
The Draw	1.01	3.0	**5.0**	**10.0**	50.0	1000.0

Figure 3.3.: England vs West Indies in-play Match Odds market after the first round of matching

We ended up matching the whole stake that was available on the lay offer at 2.0, hence its absence from the new state. Now we still have back bets on England at 2.5 and the West Indies at 3.0. These two bets can be matched against a back bet on The Draw at a price of **3.75**, since odds of 2.5, 3.0, and 3.75 form a 100% book:

$$\frac{1}{2.5}+\frac{1}{3}+\frac{1}{3.75}=1$$

We now have the best three available back prices. The two virtual bets we have calculated are the bets that would have been matched had we received that (sufficiently) large back bet at 1.01 on The Draw. Figure 3.4 shows the state of the market with the two virtual bets on the back side of the market for The Draw selection.

Selections	**Back**			**Lay**		
England	1.01	1.5	**1.5**	**2.0**	3.0	1000.0
West Indies	1.01	2.4	**2.5**	**3.0**	20.0	1000.0
The Draw	3.75	5.0	**6.0**	**10.0**	50.0	1000.0

Figure 3.4.: England vs West Indies in-play Match Odds market with the two virtual bets

We can apply this same concept to the Set Betting market in tennis. For 3-set matches there are four selections: 2-0, 2-1, 0-2, 1-2, whereas for 5-set matches there are six selections: 3-0, 3-1, 3-2, 0-3, 1-3, 2-3. Now our algorithm becomes:

1. Convert the current LPM value in the Match Odds market for the selection corresponding to our target player into a match-winning probability.
2. Attempt to find 'recent' LPM values in the Set Betting market of all the selections corresponding to a victory for our target player, e.g. 2-0, 2-1 and convert into a match-winning probability.
3. *If we find there is no recent LPM value for one or more of the required Set Betting selections, use crossmatching with the best available lay prices of the remaining Set Betting selections to calculate appropriate estimations of the back prices to use instead.*
4. Take the probability produced by the Match Odds minus the probability produced by the Set Betting odds.

Unfortunately, in extreme cases, we may find that we do not have a best available lay price for a selection in the Set Betting market where we need one. There are not enough up-to-date offers in the market to use our current algorithm. We turn directly to the raw best available back price for the selection in question in order to calculate a match-winning probability. **Step 3** now becomes:

If we find there is no recent LPM value for one or more of the required Set Betting selections, use crossmatching with the best available lay prices of the remaining Set Betting selections to calculate appropriate estimations of the back prices to use instead. ***If we find there is no best available lay price for a Set Betting market selection, then just take the raw best available back price for that selection.***

3.2.2. Correcting for Overround

The concept of overround is essentially how traditional bookmakers make their money. Given 5 possible outcomes, say a bookmaker prices each selection at odds of 4/1 or 5.0. This indicates that each selection has a 20% chance of occurring and we have a 100% book. If the bookmaker were to take an equal amount of money on each selection, he would break even.

Outcome	**Odds**	**Percentage**
A	5.0	20%
B	5.0	20%
C	5.0	20%
D	5.0	20%
E	5.0	20%
		100%

Figure 3.5.: Table displaying five outcomes of an event each with odds of 5.0, creating a 100% book

If the bookmaker were to price each selection at 3/1 or 4.0, then the implied probability of each outcome would change to 25% despite this not being mathematically possible. Now if the bookmaker were to take an equal amount of money on each selection, he would take five 'units' and pay out four. We now have a 125% book with the extra 25% being known as the overround and representing the bookmaker's profit ($25/125 = 20\%$ profit) or 'vigorish'.

Outcome	**Odds**	**Percentage**
A	4.0	25%
B	4.0	25%
C	4.0	25%
D	4.0	25%
E	4.0	25%
		125%

Figure 3.6.: Table displaying five outcomes of an event each with odds of 4.0, creating a 25% overround

What this means is that when working with unmatched bets, i.e. best available back prices, we must account for the overround (lay bets have an equivalent symmetric concept called the underround). **Step 3** of our algorithm now becomes:

If we find there is no recent LPM value for one or more of the required Set Betting selections, use crossmatching with the best available lay prices of the remaining Set Betting selections to calculate appropriate estimations of the back prices to use instead, ***making sure to correct for overround****. If we find there is no best available lay price for a Set Betting market selection, then just take the raw best available back price for that selection.*

In more detail:

(a) Subtract the probability contribution of any corresponding / recent LPM values from 1 in order to find the *remaining probability* that we need to make up using crossmatching.

(b) Calculate the overround (or rather, the *ratio of the total remaining value of the book to the remaining probability*) with respect to the sum of the crossmatched (or if necessary, raw) best available back prices.

Note that the crossmatching formula is now:

$$BackPrice = \frac{1}{RP - \sum_{\iota \in L} \frac{1}{\iota}}$$

RP is the remaining probability
L is the set of the best available lay prices of all the other selections

since we have already accepted some LPM values.

(c) To correct for overround, we divide each chosen back price by the ratio of the total value of the book to the remaining probability that we calculated.

(d) Add the probability contribution of the LPM values and the probability contribution of the sum of the overround-corrected best available back prices to generate the desired match-winning probability.

Note that the accuracy of our direct comparisons between the two markets depends on our definition of 'approximately the same time'. We search in a pre-defined time window around a Match Odds implied probability for Set Betting data. The larger the window, the easier it is to find Set Betting offers, but the less accurate the comparison and consequently, the less reliable the resulting graph. We balanced these concerns by choosing a window of 5 minutes either side of the time of a Match Odds sample.

3.3. Sample Match Data

We now apply our odds processing algorithm to an example real-world match. In the third round of the French Open 2011, Andy Murray faced off against Michael Berrer. Murray was cruising through the match until he dramatically rolled his ankle chasing a drop shot early in the second set and had to receive lengthy treatment. Fortunately, an almost one-legged Murray still managed to see out the match in straight sets and avoid the disappointment of a retirement. Figure 3.7 shows the behaviour of the Match and Set Betting markets for Murray on the Betfair exchange during the match. The blue line is the implied match-winning probability drawn from the Set Betting market, the red line is with respect to the Match Odds market, and the green line is the difference between the two when the Set Betting probability is greater, i.e. information about the risk of retirement of Andy Murray. This match is an excellent example of the 'ideal' type of injury situation; we can see clearly when Murray's injury occurred and the effect it immediately had on the markets (Murray's Match Odds probability instantly drops from almost 1 to 0.4, creating a gap of 0.4). We even see the market's slow realisation that Murray was not going to retire as the Match Odds implied probability climbed back up to meet the Set Betting implied probability at 1 at victory.

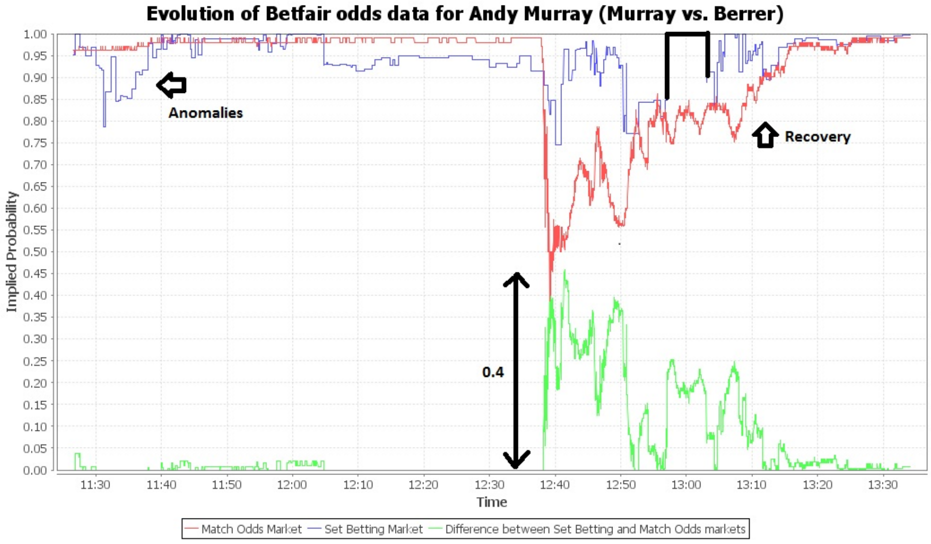


Figure 3.7.: Evolution of extracted Betfair Match and Set Betting markets implied match-winning probabilities as well as the gap between them for Andy Murray - *Murray vs. Berrer (French Open 2011 Men's Third Round)*

We also note that the odds data does not produce 'smooth' results. The Set Betting market represents implied probabilities without risk of retirement whereas (for Betfair), the Match Odds market takes into account risk of retirement after the first set has been played. A totally efficient market (at times where there is no risk of retirement) would cause both the Match Odds and Set Betting markets to generate identical implied probabilities. Prior to the injury, we can see that even though the Match Odds implied probability stays close to 1 as Murray is dominating, the Set Betting implied probability fluctuates. The theory suggests that a lower Set Betting than Match Odds probability represents the opponent's chance of retiring (whereas the opposite is the target player's retirement risk). However, Berrer showed no sign of injury whatsoever in this match.

In a hypothetical match scenario where both players show genuine risk of retirement, each player's Match Odds probability will decrease a certain amount due to their own injury, yet also increase due to their opponent's injury. The difference between the two Betfair markets will then indicate something about each player's retirement risk *relative* to the other player.

Even after the injury, we see wild movements in the Set Betting odds data. There are many possible reasons for such anomalies:

- We have already noted that the Set Betting market is significantly less popular than the Match Odds market. The strictly vertical and horizontal movement at approximately 13:00, for instance, suggests lack of investment in the market. The match took place at a Grand Slam (the highest profile of competitions) but it was only the third round and all signs pointed to a whitewash in favour of Murray. Consequently, there may not have been much interest in the match from traders. The figures from FracSoft say £4 million worth of bets were matched in the Match Odds market and £260,000 in the Set Betting market, which is nothing special.

- We acknowledge that we could have missed flaws in the algorithm we used to process the Set Betting odds due to the complexity of the data and created anomalies in the graphs that otherwise should not be there. However, we are confident in the correctness of our algorithm since our results shows some similarity to graphs generated using *only* LPM values in the Set Betting market (not necessarily in correspondence with the Match Odds LPM values).

- We could just be seeing inefficiencies in the market. Differences between the implied probabilities of the two markets at times when there was no obvious retirement risk (e.g. early in the match) could simply indicate possible arbitrage opportunities.

We also present the same information but for opponent Michael Berrer in Figure 3.8.

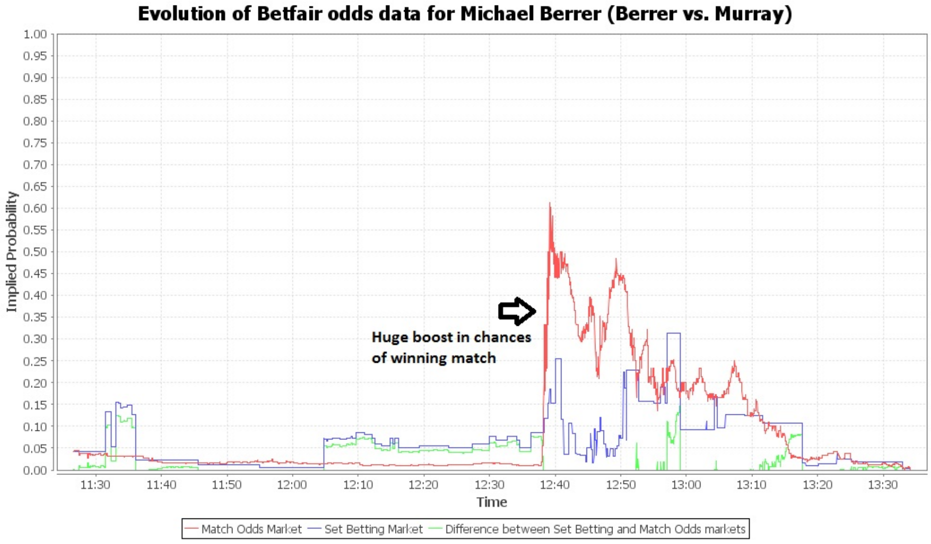


Figure 3.8.: Evolution of extracted Betfair Match and Set Betting markets implied match-winning probabilities as well as the gap between them for Michael Berrer - *Murray vs. Berrer (French Open 2011 Men's Third Round)*

As noted previously, the odds data is never as comprehensive for heavy underdogs. Nevertheless, you can still clearly see how Murray's injury coincides with a large boost in the chances of Berrer winning the match due to a Murray default. Berrer's Set Betting implied probability even rises for a short period, potentially due to an expectation that he might come back to win the match normally against a hobbling Andy Murray.

4 A Base Model for Tennis

We now know that information about the risk of retirement of players in professional tennis matches can be extracted from odds data. We can use the Set Betting market as an imitation of a Match Odds market that provides implied match-winning probabilities without risk of retirement (since we do not have access to a real one). Alternatively, we can use an established tennis match model such as the tennis formulae to represent this market.

4.1. Implementation

The tennis formulae described by O'Malley (2008)[17] calculate the pre-play probabilities of winning games, sets, matches and tiebreaks given the probabilities of each player winning a point on their serve. As an example, we give below O'Malley's formula for the probability of winning a game. The formula combines the probabilities of all the different ways a player can win a game. Note that when the game score reaches deuce, the game does not end until one of the players achieves a two-point advantage. Consequently, an infinite geometric series represents the progression of the match from deuce. The formula for a tiebreak is derived in a similar fashion.

$$\begin{aligned} g(p) &= \mathbb{P}(WinGame) \\ &= p^4+4p^4(1-p)+10p^4(1-p)^2+20p^3(-p)^3.p^2\sum_{i=3}^{\infty}[2p(1-p)]^{i-3} \\ &= p^4\left(15-4p-\frac{10p^2}{1-2p(1-p)}\right) \end{aligned}$$

However, we require the probability of winning a game, set, or match from any given current match state. We code hierarchical, recursive functions that calculate these probabilities for a given player, similar to those previously explored in papers such as Newton and Keller (2005)[18] and Barnett, Brown, and Clarke (2003)[20]. Our input is the probability the target player wins a point on serve, P_A, the probability the opponent player wins a point on serve, P_B, and the current state of the match. Note that for a game, we need only to pass in the probability the target player wins any point in that game (regardless of the server).

You can find pseudocode for our functions in Appendix B.

4.2. Comparison with Previous Works

We compare match-winning probabilities generated by our base model with those published in previous academic papers. We investigate the pre-play probabilities found in such papers as well as, vitally, the probabilities of winning from a given match state. For instance, Figure 4.1 shows the probabilities of winning a game from all possible starting scenarios given a point-winning probability of 0.54 found by Barnett and Clarke (2002)[19]. O'Malley charted 3-set match-winning probabilities as P_A increases from 0 to 1 where $P_B = P_A - 0.02$. Figure 4.2 shows our faithful reproduction of the original graph.

		A Score				
		0	15	30	40	game
	0	0.60	0.74	0.87	0.96	1
	15	0.44	0.59	0.76	0.91	1
B Score	30	0.25	0.39	0.58	0.81	1
	40	0.09	0.17	0.31	0.58	
	game	0	0	0		

Figure 4.1.: Probabilities of Player A winning a game from all possible starting scenarios given a point-winning probability of 0.54

Our base recursive model produces results identical to those published in related works. The model, given a pre-play starting scenario, also agrees with our implementation of the tennis formulae, further adding to our confidence. Furthermore, we were successful in reproducing Figure 4.2 and others like it using our model. However, such results are usually mentioned only in passing and therefore cover just a small subset of the possible range of input parameters. It would be unreasonable to expect to be able to test all possible point-winning probability and starting match state combinations and, regardless, there is no definitive source of correct values to check our output against anyway. Nevertheless, we still felt it would be sensible to search for a way of providing a more thorough verification of our functions.

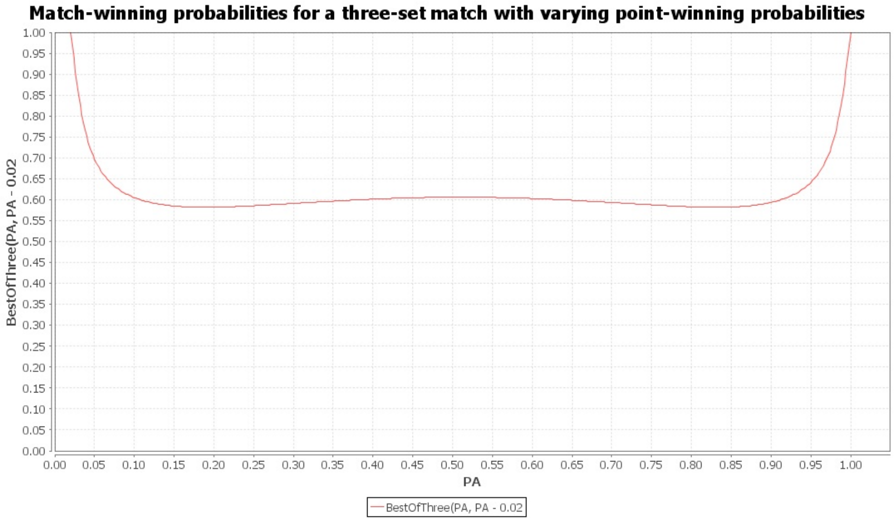


Figure 4.2.: Chart displaying 3-set match-winning probabilities as P_A increases from 0 to 1 where $P_B = P_A - 0.02$

4.3. A Tennis Match Simulator

We decided that another way of verifying the correctness of our model would be to use a different method to calculate match-winning probabilities and see if we get the same results. To this end, we created a tennis match simulator capable of approximating match-winning probabilities (as well as many other statistics, for example, the average number of points in a tiebreak). We input the point-winning probability for each player as well as the current score, and simulate a large number of matches using the same parameters for each match. The simulator is probabilistic but we expect convergence towards exact match-winning probabilities. The greater the number of *runs* (i.e. matches played), the greater the accuracy of the approximation generated by the simulator. The proportion of matches the target player wins out of the total number of runs is an estimation of the chance of winning. Note that we use a single static Mersenne Twister algorithm implementation to decide which player wins each point according to the point-winning probabilities, as it is a fast way of generating very high-quality pseudorandom numbers.

Below we present pseudocode for the simulator.

Algorithm 1 SimulationOutcomes `simulate`(double *pa*, double *pb*, MatchState *initialState*, boolean *isScenario*, double *runs*)

```
SimulationOutcomes outcomes = new SimulationOutcomes(runs)
// When simulating a particular scenario, e.g. a match in progress, we want to
replicate the starting conditions exactly for each run, else we just pick a player
to serve next at random
for each run do
    MatchState result = new MatchState(initialState)
    if !isScenario then
        result.chooseRandomServer()
    end if
    simulateMatch(pa, pb, result)
    outcomes.update(result)
end for
return outcomes
```

Algorithm 2 MatchState `simulateMatch`(double *pa*, double *pb*, MatchState *score*)

```
while !score.matchOver() do
    while !score.setOver() do
        while !score.gameOver() do
            playPoint(pa, pb, score, score.targetPlayerServing())
        end while
        if score.tiebreak() then
            playTiebreak(pa, pb, score)
        end if
    end while
end while
return score
```

Algorithm 3 void `playPoint`(double *pa*, double *pb*, MatchState *score*, boolean *serving*)

```
double point = mersenneTwister.nextDouble()
if (serving && point < pa) || (!serving && point ≥ pb) then
    score.incrementTargetPlayerScore()
else
    score.incrementOpponentScore()
end if
```

Algorithm 4 void `playTiebreak`(double *pa*, double *pb*, MatchState *score*)

```
boolean $serving = score$.targetPlayerServing()
// Service swaps every odd number of points
while !$score$.tiebreakOver() do
    playPoint($pa, pb, score, serving$)
    if $score$.isOddPoint() then
        $serving = !serving$
    end if
end while
```

Fortunately, we find that the tennis match simulator agrees with our base recursive model. This gives us confidence in moving forwards with creating a model for retirement risk. However, we remain aware of the possibility that we could have made exactly the same or equivalent mistakes when coding both the base model and the simulator. In this case, both systems would generate the same probabilities, hiding any such errors.

Modelling Retirement Risk 5

5.1. Point-level Retirement Risk

Our approach to this problem took an intuitive angle. We decided to maintain the granularity of the model to be on a point-by-point basis as this is compatible with the structure of standard tennis models. We could then try to incorporate this model into a main model for calculating the probability of winning a tennis match. As examined by Johnson and McHugh (2006)[12], when a tennis player plays a point (or even on each stroke), he or she puts great strain on their body. This strain naturally leads to a chance of an injury occurring. For the vast majority of points played, the strain is perfectly manageable and does not lead to injury. For example, Grand Slam winners will hit over 1000 serves during the course of the tournament without issue. Occasionally however, the body fails to cope with the strain, or the strain is for some reason much more acute than normal (e.g. remember Andy Murray twisting his ankle), and an injury occurs. When a player does get injured during a match, it does not always end in retirement. Players often soldier on at least for a few points and may even recover from the injury as the match progresses.

Here we present an elegant model for the probability a player will retire *on a given point* in a tennis match:

$$\begin{aligned} r_0 &= 0 \\ r_{n+1} &= min(\rho r_n + X, 1) \end{aligned}$$

where r_n is the given player's risk of retirement on point n of the match, $0 \leq \rho \leq 1$, and X is a random variable. Do not confuse r_n with what we are hoping to eventually calculate which is R_n, the risk of retiring at some point during the remainder of the match from point n. We make the reasonable assumption that players start a match with zero probability of retiring ($r_0 = 0$), else they would choose to not participate and allow their opponent a walkover. This is not strictly true since players sometimes play matches despite possessing niggling injuries. We have a decay parameter, ρ, which models the idea that players recover from injuries as matches progress. For the purposes of avoiding too complex a model, we make the simplifying assumption that ρ and the distribution of X is the same for both players. When setting r_{n+1}, we make sure to take the minimum of the calculated r_{n+1} and 1, since retirement risk is a probability and cannot be greater than 1. The random variable, X, models the rationale that each point played causes strain on a player and therefore risk of injury. Since we are modelling probability, r_n must always be between 0 and 1, so when choosing an appropriate distribution for X we had the following requirements:

- $F(0) = 0$ and $F(1) = 1$, where $F(x)$ is the Cumulative Distribution Function(CDF) of the distribution (i.e. the *support* of the distribution is $0 \leq x \leq 1$, where x is a member of the distribution).
- The distribution is *heavy-tailed* so that the majority of points cause very little injury risk but occasionally a point may cause a more significant risk of retirement.
- The distribution requires as few unknown parameters as possible meaning fewer approximations and a more accurate model.

There are many probability distributions which display such characteristics. We explore the use of a *bounded Pareto distribution* and an *exponential distribution*.

5.1.1. A Bounded Pareto Distribution

The bounded (or truncated) Pareto distribution is a special case of the standard Type I Pareto Distribution where you supply an additional parameter, H, which is the upper bound of the distribution. Pareto distributions also take a lower bound parameter, L, and a parameter, α, which determines the shape of the distribution. Pareto distributions are heavy-tailed and often used to model situations where there is some sort of balance in the distribution of the 'small' to the 'large', for example, phenomena such as the distribution of wealth (most people are not particularly wealthy but a few are super rich) or the size of human settlements (few cities but many villages). In this sense, the distribution appears to be intuitively suitable for our model.

Although we can take $H = 1$, the Bounded Pareto distribution requires that $L > 0$ which means we cannot directly use it in our model as we need the support of the distribution to be between 0 and 1 inclusive. A solution is to shift the Bounded Pareto so that this is the case by subtracting L from any value sampled from the distribution and setting $H = L+1$. By also setting $L = 1$ we can simplify the inverse transform of the CDF used to sample from the distribution. Figure 5.1 shows a possible CDF of such a modified distribution.

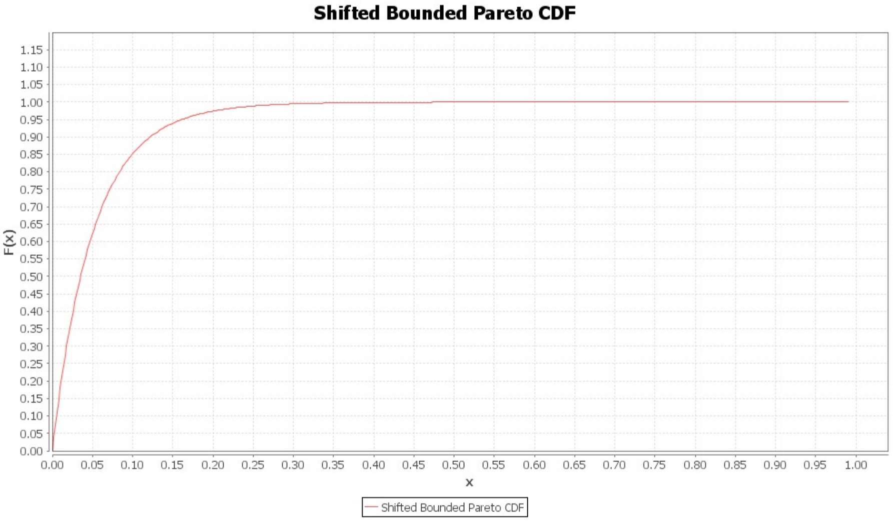


Figure 5.1.: The CDF of a Pareto distribution with upper bound 1, shifted to possess a lower bound of 0, and with shape parameter $\alpha = 20$

Note that the *Lomax distribution* is a Pareto Type I distribution which has been shifted so that $L = 0$ (and also happens to be a special case of the Pareto Type II distribution), but it does not have the necessary upper bound. The Pareto distribution is also closely related to the exponential distribution, which we investigate next.

5.1.2. A Truncated Exponential Distribution

The well-known exponential distribution is also heavy-tailed. It is used to model the time between events which occur continuously and independently at a constant average rate, for example, such phenomena as the time until a radioactive particle decays or the time until your next phonecall. Although this description is less intuitively suitable for our model, all we are really looking for is a CDF that is easy to sample and where all the probability is contained in members in the range [0, 1] (and most of it in the bottom end of this range). The support for the exponential distribution is bounded below with $x \geq 0$ which is an advantage over the Pareto but it is also unbounded above which is a problem.

The exponential distribution is one of the easiest to sample from as it has a CDF very amenable to the inverse transform method. Given a uniformly distributed random variable $U \sim U(0,1)$, the random variable:

$$T = F^{-1}(U)$$

is exponentially distributed, where:

$$F^{-1}(y) = \frac{-\ln(1-y)}{\lambda}$$

is the inverse of the CDF of the exponential distribution. Since $(1-U) \sim U(0,1)$ as well, we have:

$$T = \frac{-\ln(U)}{\lambda}$$

Usefully, if you take a random variable $Y \sim exp(\lambda)$, and extract the fractional part $\{Y\}$ of Y, we find that $\{Y\}$ is drawn from a truncated exponential distribution with upper bound 1! A proof and reference for this result can be found in Appendix D. In order to sample from a truncated exponential distribution, we sample from a normal exponential distribution (once again using our global Mersenne Twister to generate the uniform random variable required) and subtract the integer part of the value drawn. As you can see from Figure 5.2, the CDF of a truncated exponential distribution is very similar to that of our modified Pareto distribution above. However, we find that sampling from the truncated exponential is faster and simpler than using a bounded Pareto distribution that we have had to shift and so we define in our retirement risk model our random variable:

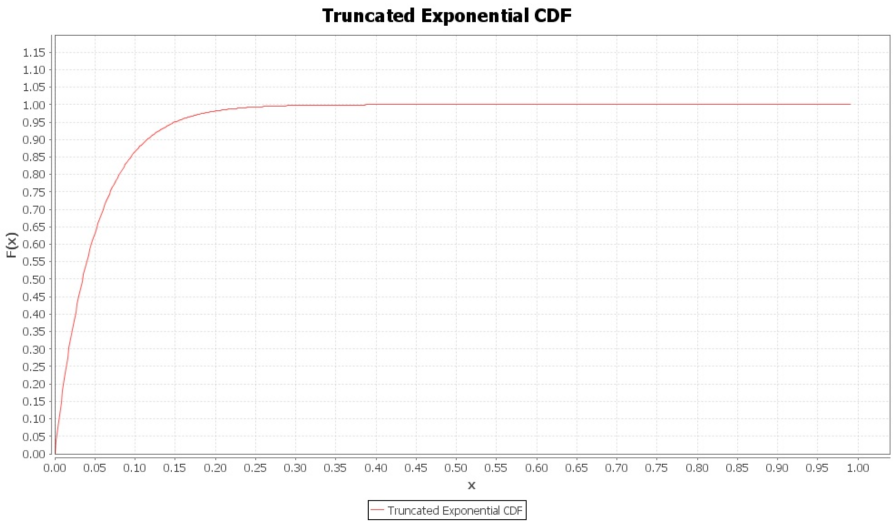


Figure 5.2.: The CDF of an exponential distribution with rate $\lambda = 20$ and upper bound 1

As we saw in the Murray vs Berrer sample match data, serious injuries can produce a gap in the odds of at least 0.4. This implies that our truncated exponential distribution requires a rate λ small enough such that sampling values that are not trivially tiny are possible but very unlikely (such injuries should be rare). However, the distribution should still yield very very small samples the vast majority of the time as on most points there is no injury. We could not find a value for λ that allowed the distribution to successfully meet both these criteria so we now turn to a *hyper-exponential distribution*.

5.1.3. A Hyper-Exponential Distribution

A hyper-exponential distribution has probability density function with respect to a random variable Y:

$$f_Y(y) = \sum_{i=1}^{n} f_{Z_i}(y) p_i$$

where Z_i is an exponential random variable with rate parameter λ_i and p_i is the probability that Y will take on the form of Z_i.

In our case, we can say that X takes on the form of an exponential distribution with rate parameter λ with probability drawn from a Bernoulli distribution with success parameter c. So a sample from a Bernoulli distribution takes value 1 with probability c and value 0 with probability $1 - c$. Since the probability of injury on any point is so small, we can safely model it as 0. Occasionally (with probability c), a player suffers an injury while playing a point and the magnitude of this injury is modelled using an exponential distribution with rate λ. Note that our distribution still needs to be truncated as detailed previously. In addition, although we do not let r_n be greater than 1, we will generally use a high enough value for λ such that it is unlikely we will sample a value for X that causes the point-level retirement risk to rise above 1. We define in our retirement risk model our random variable:

$$X \sim TrHypExp(c, \lambda)$$

In order to find out whether we have created a reasonable model for retirement risk, we need to find a way to incorporate our specific model into a overarching model for the probability of winning a tennis match. The simplest way of doing this to begin with was to modify the tennis match simulator that we conveniently created earlier.

5.2. A Modified Tennis Match Simulator

We use our simulator to run a large number of matches, each with identical starting scenarios, in order to approximate the match-winning probability of each player. Previously, we had only two outcomes to each point; either Player A won the point or Player B won the point. Now we have two additional outcomes; Player A can retire from the match or Player B can retire from the match. This requires two more parameters in addition to the point-winning probabilities of each player. We now also have r_n^A and r_n^B, the probabilities that Player A and Player B retire from the match on point n, respectively. These probabilities are updated each point in accordance with the model we defined above.

Algorithm 5 shows the pseudocode of our new *playPointWR()* method:

Algorithm 5 void `playPointWR`(double *pa*, double *pb*, RetirementRisk *risk*, MatchState *score*, boolean *serving*)

```
risk.ra = risk.ra * $\rho$
risk.ra = risk.ra + $X_A \sim TrHypExp(c, \lambda)$
risk.rb = risk.rb * $\rho$
risk.rb = risk.rb + $X_B \sim TrHypExp(c, \lambda)$
// p is the probability Player A wins the point whereas q is the probability Player B wins the point. These probabilities must be normalised.
double p, q
if serving then
    p = pa / (1 + risk.ra + risk.rb)
    q = (1 - pa) / (1 + risk.ra + risk.rb)
else
    p = (1 - pb) / (1 + risk.ra + risk.rb)
    q = pb / (1 + risk.ra + risk.rb)
end if
double point = mersenneTwister.nextDouble()
if point < p) then
    score.incrementTargetPlayerScore()
else if $point \geq p$ && $point < p + q$ then
    score.incrementOpponentScore()
else if $point \geq p + q$ && $point < p + q + ra$ then
    score.targetPlayerRetires()
else if $point \geq p + q + ra$ then
    score.opponentRetires()
end if
```

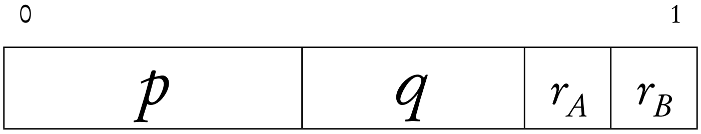


Figure 5.3.: The four possible outcomes of a point in our modified tennis match simulator

We use a Mersenne Twister random number generator to decide which player wins each point. Looking at Figure 5.3, the bin the random number falls into is the outcome of the point. The match ends if it falls in one of the retirement bins (which is why r_A and r_B are usually 0 else you will rarely be able to get through even a single match without retiring), otherwise one of the players wins the point and the match continues (unless it was match point). Now our simulator calculates four results. The probabilities of each player winning the match by achieving the required number of sets and the probabilities of each player retiring from the match.

Although the simulator produces results to a good degree of accuracy and reasonably fast (see Appendix C), ideally we would build a new mathematical model for a tennis match that incorporates retirement risk and generate exact solutions. The problem is that such a model would be analytically very difficult to solve (although we do detail the theory behind its possible creation in Appendix E) so we decide to focus on making our simple and efficient modified simulator as our main model for a tennis match. The use of the simulator also makes it straightforward to calculate the probability each player retires *in a particular set* (splitting up the two retirement outcomes), allowing us to predict the evolution of markets that follow the different tennis betting retirement payout policies.

5.3. Modelling Markets

To imitate the Betfair Set Betting market (or equivalently, a Match Odds market using a Paddy Power-style *match-completed* payout policy), we can use our base recursive model which ignores retirement risk (the *No Risk* column in Figure 5.4). We use our modified simulator to imitate a Match Odds market as it might behave using the different retirement betting payout policies by calculating the remaining probabilities in the right-most three columns.

Player	No Risk	Normal Win With Risk	Retirement in 1st set	Retirement after 1st Set
A	W_A	W'_A	R^1_A	R^2_A
B	W_B	W'_B	R^1_B	R^2_B

Figure 5.4.: Probabilities that can be closely approximated by our modified tennis match simulator

W'_A, for example, is the probability of Player A winning the match normally by achieving 3 sets and not via Player B retiring.

Note that we can (re-)calculate the match-winning probability for Player A ignoring retirement risk (W_A) as a sanity check by computing:

$$\frac{W'_A}{W'_A + W'_B}$$

This is because we assume that the point-winning probabilities of each player are unaffected by injury. Consequently, the ratio of W_A to W_B is the same as the ratio of W'_A to W'_B.

We can imitate a Betfair-style *after one set* payout policy market for Player A by computing:

$$\frac{(W'_A + R^2_B)}{(W'_A + R^2_B) + (W'_B + R^2_A)}$$

which is the probability that Player A wins normally plus the probability that Player B retires after the first set, normalised by the sum of the probabilities that either player wins the match normally and either player retires after the first set.

Similarly, we can imitate a Ladbrokes-style *after one ball* payout policy market for Player A by computing:

$$\frac{(W'_A + R_B)}{(W'_A + R_B) + (W'_B + R_A)}$$

where $R_A = R^1_A + R^2_A$ and $R_B = R^1_B + R^2_B$. This is essentially a re-calculation of the *Normal Win With Risk* column.

Parameterising the Model 6

Figure 6.1 displays a table describing the majority of the variables we have introduced thus far. An appropriate parameterisation of the model would be values for the unknown input variables, c, λ, and ρ, that generate $R_A \approx R_B \approx 1.95\%$ when input into our modified simulator for a whole match with $P_A = P_B = 0.6$. We believe this is justified as it corresponds to the 3.9% average retirement rate in Grand Slam men's singles matches as well as a common point-winning probability on serve of a top-level professional tennis player that we discovered through research. The here goal is to *fit* the model to accurately reflect real-world events.

The problem now is that we essentially have *three* unknowns but only *one* equation (the simulator). We can choose a reasonable value to fix ρ at but this still leaves us with two unknowns so we gain little by guessing the decay constant. One possible solution is to only consider the retirement risk of one player at a time, for example, Player A, and fix $R_n^B = 0$ (i.e. assume Player B has no chance of retiring) since it is rare that both players in a match come to be at serious risk of retiring. However, this would not provide a symmetric model. We want W_A' and W_B' to be logical opposites of each other. Consequently, we have no choice but to estimate suitable c, λ and ρ that will give us the retirement rates that we want.

6.1. The Nelder-Mead Simplex Method

The Nelder-Mead Simplex Method is a multivariate direct search optimisation algorithm primarily designed for statistical parameter estimation problems such as ours. The method uses the idea of a *simplex*, a polytope of $N + 1$ vertices given N dimensions, i.e. a line segment on a line, a triangle on a plane, a tetrahedra in 3D space, etc. The user defines an initial non-degenerate simplex and an *objective function* that the algorithm attempts to minimise. It does this by iteratively trying a sequence of three operations (reflection, expansion, contraction) with the vertices of the simplex to generate a new vertex which is input into the objective function.

Variable	Known	Input / Output	Description
P_A	Yes	Input	The probability Player A wins a point on serve (assumed for the moment)
P_B	Yes	Input	The probability Player B wins a point on serve (assumed for the moment)
W_A	Yes	Input	The probability Player A wins the match using the standard base tennis model
W_B	Yes	Input	The probability Player B wins the match using the standard base tennis model
G_A	Yes	Input	The Betfair Set Betting implied probability minus the Betfair Match Odds implied probability for Player A for $G_A \geq 0$
G_B	Yes	Input	The Betfair Set Betting implied probability minus the Betfair Match Odds implied probability for Player B for $G_B \geq 0$
-	Yes	Input	The current score in the match
W'_A	No	Output	The probability Player A wins the match normally given the possibility of retirement
W'_B	No	Output	The probability Player B wins the match normally given the possibility of retirement
R_A	No	Output	The probability Player A retires at some point during the remainder of the match (can be categorised by set)
R_B	No	Output	The probability Player B retires at some point during the remainder of the match (can be categorised by set)
c	No	Input	The Bernoulli success probability parameter for the truncated hyper-exponential distribution representing the chance a player suffers an injury on any given point
λ	No	Input	The rate parameter for the truncated hyper-exponential distribution dictating the magnitude of the injury suffered by a player should such an event occur.
ρ	No	Input	The decay constant representing recovery from injuries in our retirement risk equation

Figure 6.1.: Table describing the variables used in our system

If the new vertex produces a value smaller than the worst of the current vertices, than it replaces that vertex to create a new simplex. If the new value is not better, than we have stepped across a 'valley' and we create a new simplex by shrinking the old one using pre-determined *step sizes* (which decrease on each iteration). Ideally, the points of the simplex should eventually converge on the minimum of the objective function (or as close to it as the user decides is sufficient). For example, Figure 6.3 shows the evolution of the Nelder-Mead method executed on a suitable initial simplex for the simple bivariate function shown in Figure 6.2[29].

The Nelder-Mead method is quite easy to understand and simple to use; we found readily available Java code for it. Furthermore, the method requires no derivative information about the given objective function, which we do not have. The method is also reasonably quick, requiring only one or two evaluations of the objective function per iteration, rather than N as with other search algorithms. Due to the nature of our modified simulator, which delivers only approximations, we will look for a satisfactory rather than precise solution. Fortunately, the first few iterations tend to give significant improvement on the value of the objective function.

A disadvantage of the Nelder-Mead method is that it can converge to a non-stationary point. In fact, there is little convergence theory at all with regards to the algorithm. We hope to avoid such situations by choosing carefully our initial simplex and starting step sizes as well as defining a maximum number of iterations. There are more modern techniques for multivariate optimisation, but we feel that the Nelder-Mead algorithm will be sufficient.

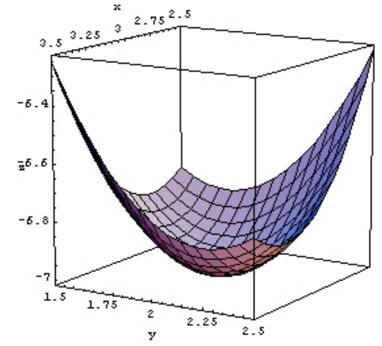


Figure 6.2.: Visualisation of the bivariate function $f(x,y) = x^2 - 4x + y^2 - y - xy$

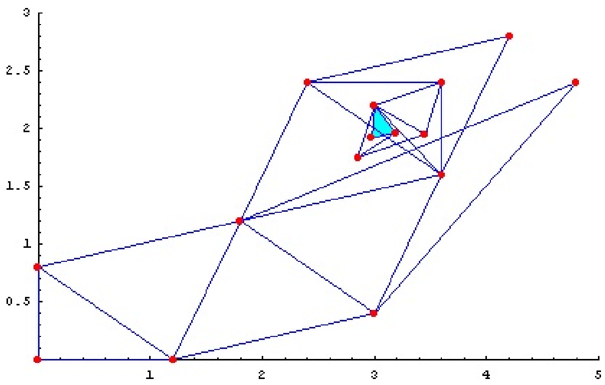


Figure 6.3.: An evolution of a simplex when executing the Nelder-Mead method on bivariate function $f(x,y) = x^2 - 4x + y^2 - y - xy$

Defining an Objective Function

Vitally important to the success of the Nelder-Mead method is the choice of objective function to minimise. In our case our modified simulator is essentially the function, but we must still define what it means to minimise it. This is the main way we mould the algorithm to solve our problem.

$$\left|W_A - (W'_A + R_A)\right| + \left|W_B - (W'_B + R_B)\right|$$

After the algorithm has completed, the first term ensures that we have recreated the correct gap in the odds for Player A with the second term being the same idea for Player B. So in this case, at the minimum, we would have:

$$|0.5 - (0.4805 + 0.0195)| + |0.5 - (0.4805 + 0.0195)| = 0$$

An Initial Simplex

In our case, we require a tetrahedra simplex since we have three variables we are trying to approximate. Choosing an initial simplex is an important part of the process. For instance, choosing too small an initial simplex can lead to a local search, causing the algorithm to get stuck in some sub-optimal hole. We can use our intuition to pick the vertices for the initial simplex, which is shown in Figure 6.4. For c, we know that it will be very small and sensitive as thousands of normal points can be played between injuries. For λ, we have seen that in an extreme case that the gap in the odds for a player can jump to at least 0.4 so we can hypothesise that the mean magnitude of a point-level retirement risk will be something lower than this. For ρ, we judge from inspecting the odds data of numerous matches that a likely value might be $0.5 < \rho < 1.0$. Note that we used a *constrained* version of the Nelder-Mead algorithm in order to ensure logical bounds $0 < c, \rho < 1$ and $\lambda > 0$.

	Parameters		
Vertex	c	λ	ρ
1	0.0001	5.0	0.50
2	0.00025	10.0	0.65
3	0.00035	15.0	0.80
4	0.0005	25.0	0.95

Figure 6.4.: The vertices of the initial simplex we input into the Nelder-Mead method

We also defined suitable initial step sizes according to the magnitudes of the numbers that made up the initial simplex (as well as choosing a *tolerance level* indicating when convergence has happened). We tried to strike a balance between avoiding the risk of stepping too far between simplex iterations and potentially falling into sub-optimal minima, and search speed.

The Results

We ran the Nelder-Mead method a number of times under the same conditions and took the means of the approximations found for our three parameters each time as our chosen values. Remember, c is the per point injury probability, λ is the point-level retirement risk magnitude exponential distribution rate parameter, and ρ is the injury recovery factor.

- $c = 0.000115$ (implying an injury approximately every 8500 points)
- $\lambda = 10.0$ (implying injuries cause a point-level retirement risk of 0.1 on average when they occur)
- $\rho = 0.95$ (point-level retirement risk is scaled by 0.95 each point)

Note that in practice there are likely to be many combinations of values for these parameters that would generate the retirement rates we require. For example, the effect of decreasing ρ (quicker recovery) could be countered by increasing λ (injuries are more severe) or increasing c (injuries are more common).

6.2. An Artificial Tennis Match

In order to find out what our model is capable of, we try it out in a totally artificial environment where we control all the variables. We ran a single simulated match many times with our set of parameters $P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, and $\rho = 0.95$. We assume that P_A and P_B remain unchanged throughout the match. We were looking for 'ideal' scenarios, e.g.

where Player A receives an injury during the match but does not retire and still goes on to achieve victory. When we generate such a match, we recorded into a CSV file the score, the server, and the point-level retirement risks, r_A and r_B, at each point during the match. We then read this information back in, calculating match-winning probabilities at each point in the match.

Figure 6.5 shows the effect on simulated markets with varying retirement payout policies of Player A suffering an injury in the second set but still managing to go on to win the match. We see a sharp drop in the red and orange lines (Match Odds markets with *after one set* and *after one ball* payout policies, respectively) when the injury occurs. This is followed by recovery, similar to the Murray vs. Berrer example. The blue line (Match Odds market with *match-completed* policy) ignores retirement risk and therefore does not react to the injury. Figure 6.6 shows the evolution of both the point-level (magenta line) and match-level (green line) retirement risks for Player A in this match. They are clearly inversely related to the behaviour of the markets; where match-winning probability falls, retirement risk rises, and where match-winning probability recovers, retirement risk decays.

Figure 6.7 shows a similar situation but where the injury occurs in the first set. This time we can see the orange line approach the red as we near the end of the first set. The models the idea that if an injury occurs in the first set of a match, traders in such a market will display increased reluctance to back the player in question as it becomes more likely he or she will finish the set (in case they then decide to default afterwards and payouts happen).

Figure 6.9 shows a match where Player A decided not to continue. Player A does attempt to play one or two more points after the injury and the market anticipates recovery but those hopes are soon dashed. Note that any discrepancies you might see between the three markets are due to the fact that the modified simulator provides only approximations whereas the base recursive model provides exact solutions (particularly noticeable at deuce or 6-6 in a tiebreak).

The model appears to produce markets that behave in the theoretically correct manner. We have sudden, sharp drops in match-winning probability corresponding to an injury on a point, followed by gradual recovery as the traders realised retirement might not happen. We move on to testing it against Betfair odds data from real-world top-level tennis matches.

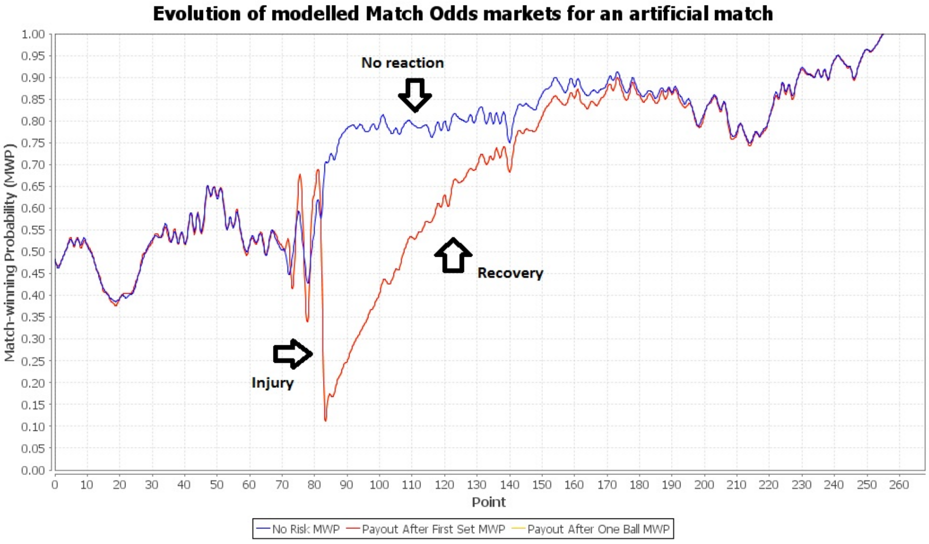


Figure 6.5.: Evolution of Match Odds markets with a variety of retirement payout policies for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury in the second set but rallies and continues on to victory

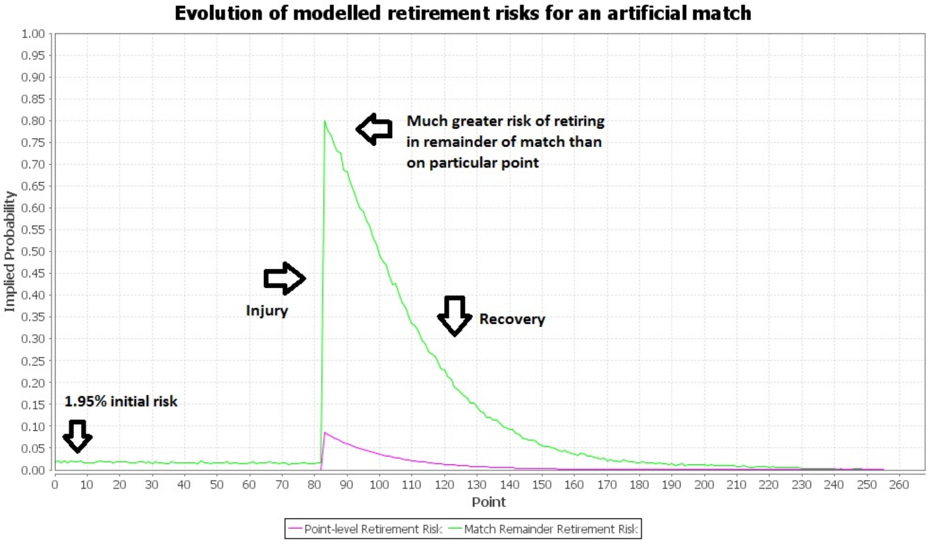


Figure 6.6.: Evolution of point-level and match-level retirement risks for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury in the second set but rallies and continues on to victory

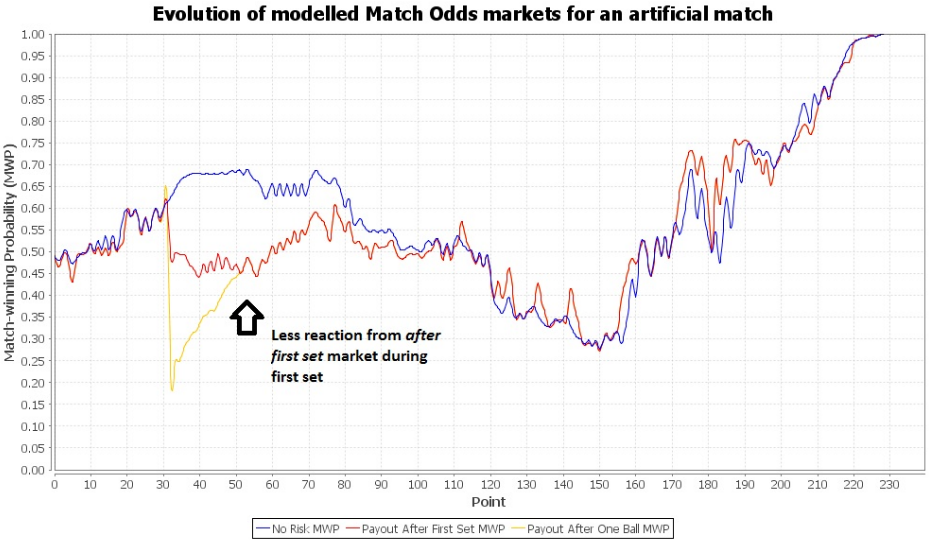


Figure 6.7.: Evolution of Match Odds markets with a variety of retirement payout policies for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury in the first set but rallies and continues on to victory

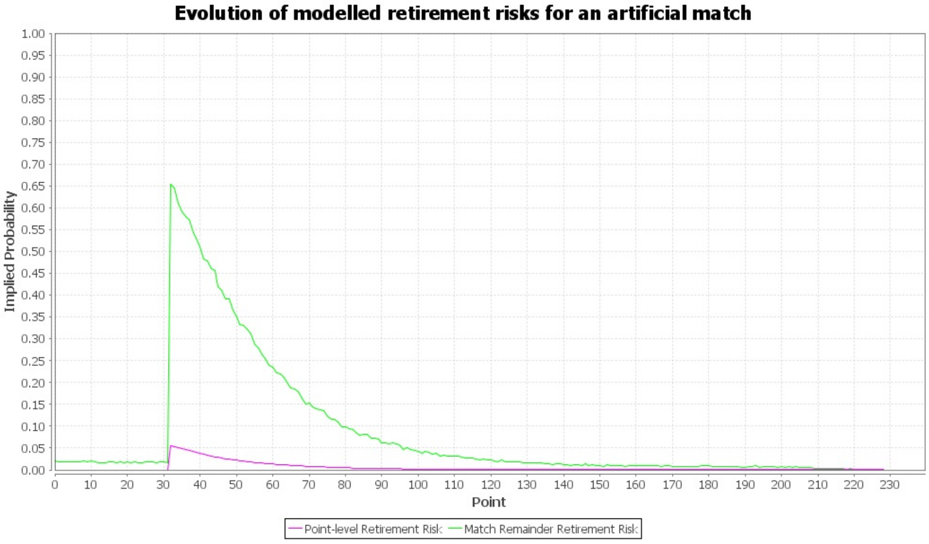


Figure 6.8.: Evolution of point-level and match-level retirement risks for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury in the first set but rallies and continues on to victory

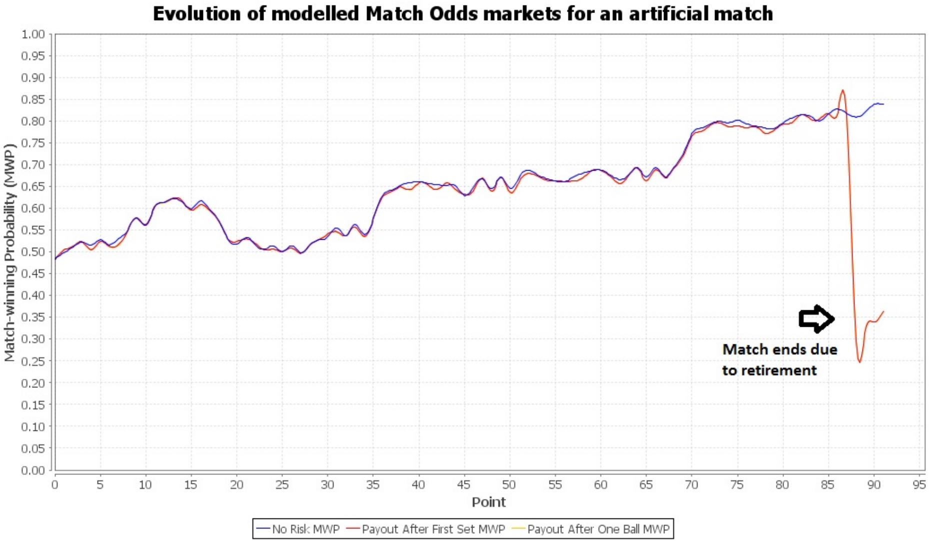


Figure 6.9.: Evolution of Match Odds markets with a variety of retirement payout policies for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury and has to retire

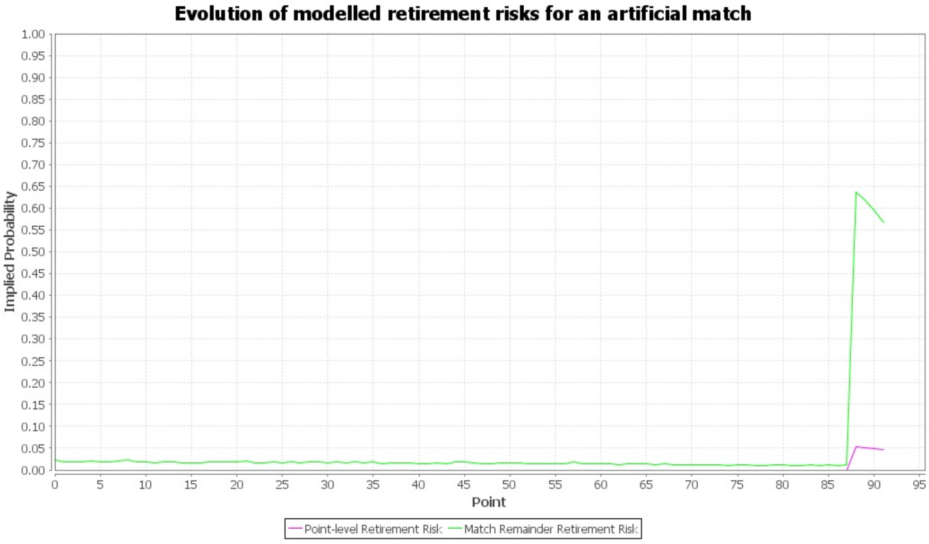


Figure 6.10.: Evolution of point-level and match-level retirement risks for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury and has to retire

6.3. Fitting the Model to the Odds Data

Our model appears to produce reasonable results from the beginning of a match but this is only because of our assumption that players begin matches with no risk of retiring ($r_0^A = r_0^B = 0$). Our model must also be applicable for any given match state, including the situation where an injury has recently occurred but the affected player is continuing to play. At this stage, we know the gap in the odds for each player (G_A and G_B), but how can we go *backwards* from this and calculate the retirement risk for the *current point*, r_n^A and r_n^B (which will naturally be a lot smaller)? Well once again, we can use the Nelder-Mead method to approximate these values for each point in the match. Given this information, our simulator will be able to calculate the match-level retirement risk for each player at each point (R_n^A and R_n^B).

Vertex	r_n^A	r_n^B
1	0.001	0.001
2	0.1	0.01
3	0.01	0.1

Figure 6.11.: The Nelder-Mead initial simplex we use when approximating the point-level retirement risks for a given current match score

Remember, we can imitate a Betfair-style *after one set* payout policy market for Player A by computing:

$$W_A^{''} = \frac{(W_A' + R_B^2)}{(W_A' + R_B^2) + (W_B' + R_A^2)}$$

and similarly for Player B.

Therefore, we redefine our objective function as:

$$\left| W_A - (W_A^{''} + G_A) \right| + \left| W_B - (W_B^{''} + G_B) \right|$$

now that we are dealing with the gaps in the odds rather than the retirement risks to be output. We now also have all the information we need to be able to predict the evolution of a market with an *after one ball* payout policy for a real match using just the two Betfair markets!

Figure 6.12 below displays our variables are where they are used in the system.

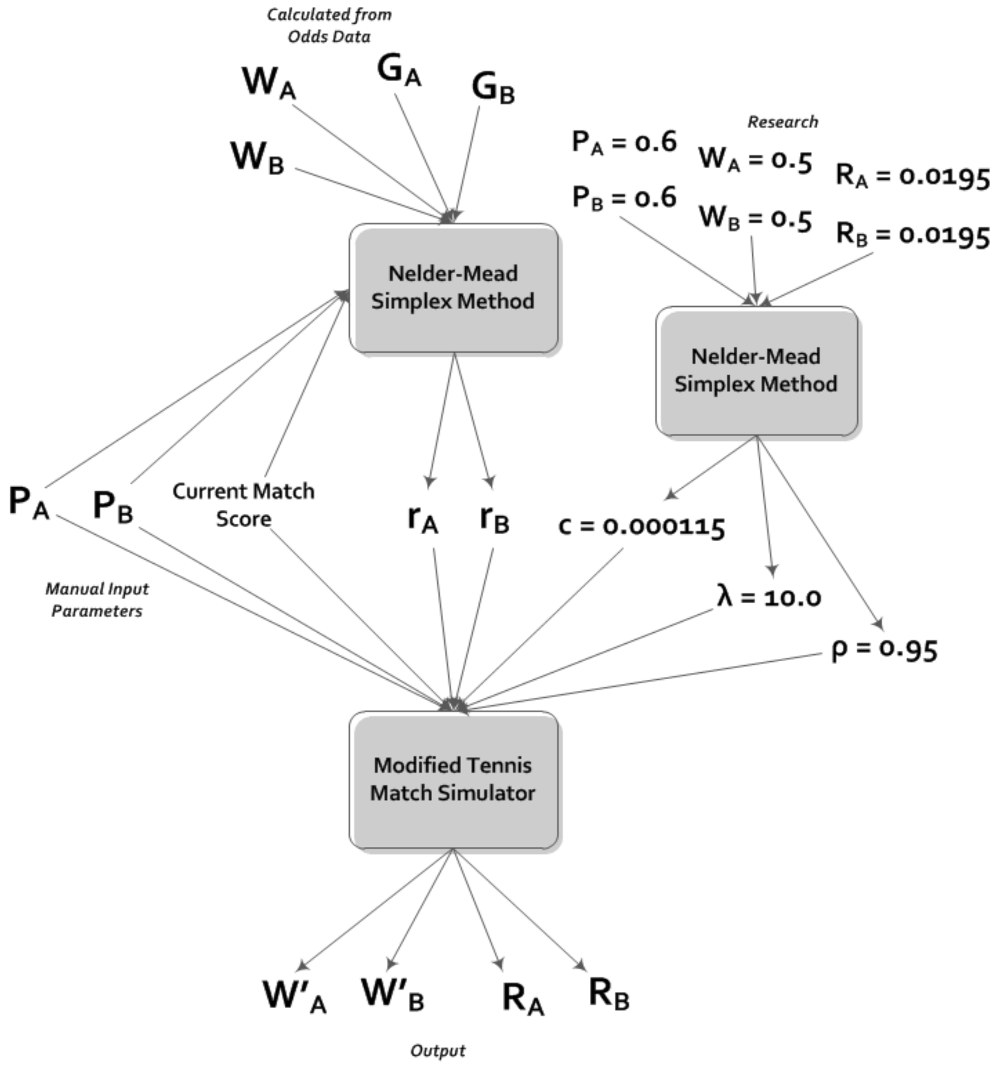


Figure 6.12.: A diagram displaying the variables of the whole system and how they are used. Note that c, λ, and ρ are constant whereas r_A and r_B are per point per match

Case Studies 7

7.1. Preparing the Match Data

Obtaining the Betfair odds data for a match was only the first step. We also need to acquire other match-specific information such as a pointstream and a point-winning probability for each player.

Acquiring Real Point-by-Point Match Data

In order to test our model against odds data from real-world matches, we need to be able to input the current score into the model at any stage in the match. We could not find any resource which archives point-by-point data in a suitable format so we entered it manually into CSV files using the historical point-level live commentary provided by website *TennisEarth.com*[1]. Unfortunately, the *TennisEarth.com* point-level archive is not comprehensive, so we could not study all the matches we wanted to (notably our running example, the Murray vs. Berrer French Open 2011 Third Round match) and we did not have time to try and procure videos of such matches in order to record the score.

Estimating Point-Winning Probabilities

Up until this point we have assumed that we will be able to manually input the point-winning probabilities on serve of both players in the match. We noted previously Klaassen and Magnus' (2000)[24] findings that the average point-winning probability on serve for a top-level professional tennis player, γ, was 0.645 for men and 0.560 for women. We also know, according to Marek (2011)[26], that the important thing in determining the winner of a match is the *difference*, δ, between the point-winning probabilities of each player and not the absolute values. Consequently, if we can find what δ *should* be, we can choose appropriate values for P_A and P_B by using the constraint that their average equals γ. We must find both; we cannot, for example, fix P_A to γ and linearly search for P_B, since the model would cease to be symmetric for both players. We know the current implied match-winning probability of the Set Betting (no retirement risk) market for both players and so we can do a

[1] http://www.tennisearth.com

simple binary search to find a value for δ (and therefore values for P_A and P_B) for which our base recursive model generates the same results as the Set Betting market. Marek also states that for one to be able to express match-winning probability as a function of δ, one must constrain $-0.1 \leq \delta \leq 0.1$. Since the market's opinion of each player's point-winning probability changes throughout the match, we recalculate P_A and P_B on every point to reflect this. Note that this does not invalidate the assumption that points are iid as no dependency between points is introduced.

Aligning the Odds Data with the Correct Score

A challenge we had to overcome was how to match the odds data to the score at each point. For instance, we may have many thousands of odds data lines for a match but only a couple of hundred points. Although we have a timestamp for each line of odds data, we do not know the exact time that each point was played. To overcome this challenge, we make the reasonable assumption that points happen at regular intervals. For example, if we have 9000 lines of odds data for a match and 300 points were played, we would sample every $9000/300 = 30^{th}$ line. Consequently, our modelled markets will appear more sparse than the original odds data as we only have as many points as there are points in the match.

We now test our system on five real-world matches. We consider only the evolution of the in-play markets from the point of view of the player that was injured in the match (it is extremely rare that both players in a match receive significant injuries).

7.2. Rafael Nadal vs. David Ferrer

An all-Spanish Australian Open 2011 Men's Quarter Final saw Rafael Nadal battle veteran and fellow clay court specialist David Ferrer. A mystery injury early on in the first set meant Nadal struggled throughout the match but he refused to retire and allowed his opponent the three-love win.

Figure 7.1 displays processed odds data as in Chapter 3. In Figure 7.2, we use our system to model Match Odds markets under three different retirement payout policies using the current score, estimated point-winning probabilities, and the positive gap created by subtracting the Betfair Match Odds implied probabilities on each point from the Betfair Set Betting probabilities, as input. As you can see from the graph, our system reproduces the Betfair Match Odds market (red line) quite accurately with the *after one set* modelled market. This shows that the Nelder-Mead algorithm was successful in finding point-level retirement risks that would recreate the gap between the Betfair Set Betting and Match Odds markets. The orange line shows our predicted *after one ball* modelled market. As we would expect (since the injury occurred in the first set), this market anticipates an even greater drop in Nadal's match-winning probability around the time of his injury than the Betfair Match Odds market. In the event of a Nadal retirement, such a market would pay out for a David Ferrer win as long as one ball has been played so at this point, traders would be very reluctant to back Nadal if he showed signs he might retire. As the match moves past the first set, the *after one ball* and *after first set* lines merge since potential injuries now contribute to both markets equally.

Bear in mind that there will always be a certain amount of variation due to the inexact nature of the simulator and our method of aligning our point-by-point data with the odds data.

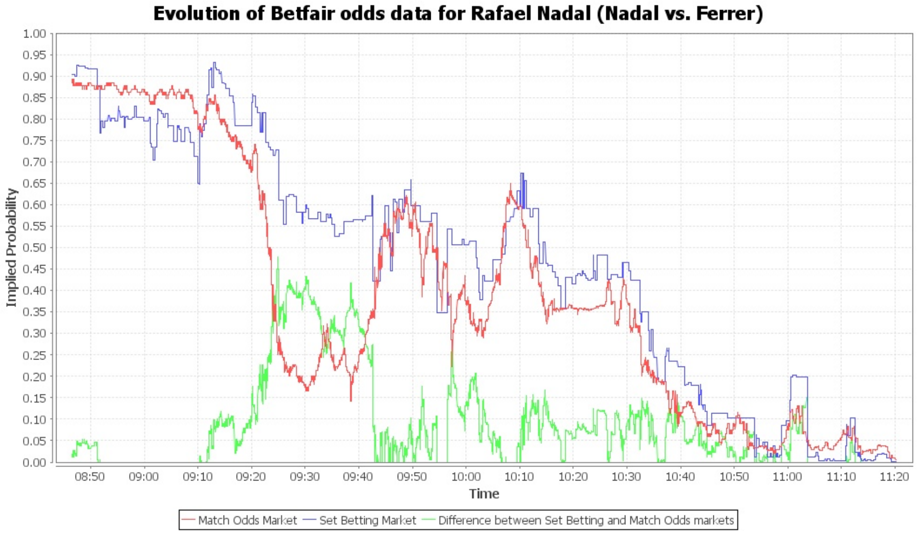


Figure 7.1.: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Rafael Nadal - *Nadal vs. Ferrer (Australian Open 2011 Men's Quarter Final)*

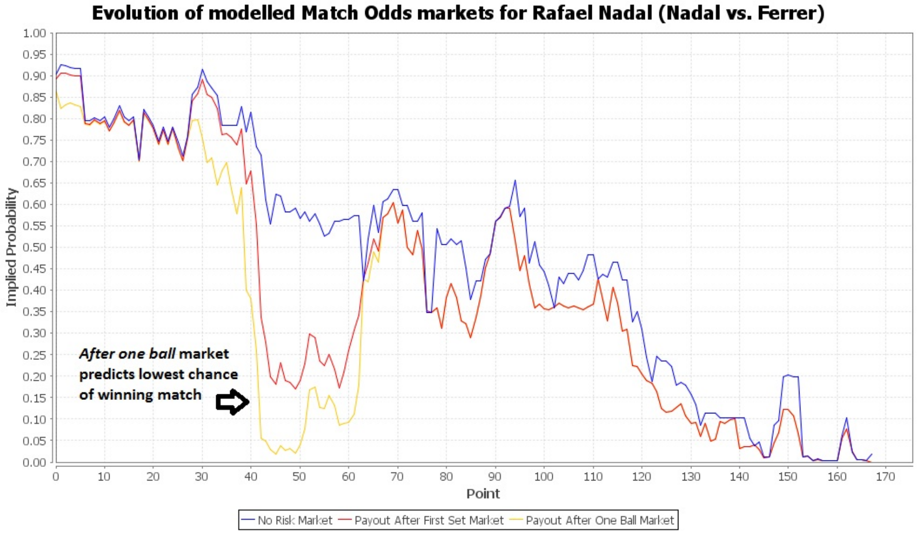


Figure 7.2.: Evolution of modelled Match Odds markets under three different retirement payout policies for Rafael Nadal - *Nadal vs. Ferrer (Australian Open 2011 Men's Quarter Final)*

Figure 7.3 shows the evolution of Rafael Nadal's risk of retirement throughout the match. The green line follows the probability of retirement during the remainder of the match from the given point (R_n), whereas the magenta line gives the probability of retiring on a particular point itself (r_n). Our model predicts a high peak match-level retirement risk of 96% during the injury period. This coincides with a peak point-level retirement risk of 18%.

Furthermore, as the match progresses and Nadal falls further behind, we see spikes in his retirement risk despite there being no large gaps between the Betfair markets. This happens when the Match Odds market gives a player little chance of winning the match. We illustrate how this occurs with an example. Say that on a particular point, the Set Betting market tells us that the implied probability of Nadal winning the match is 0.5. This means that we can assign $P_A = P_B = 0.645$. We also happen to know that the difference between the Betfair Set Betting and Match Odds for Nadal is 0.4. When we run the Nelder-Mead method, we will be looking for a value of r_A (retirement risk of Nadal on this point) such that $W_A^{''}$ (probability of winning the match normally with retirement risk after the first set only) is only 0.1. Since we have $P_A = P_B$, we have $W_B^{''} = 0.1$ and so the total probability of either player winning the match normally is only 0.2. Making the assumption that r_n^B is close to zero, we must have that $R_A \approx 0.8$ (retirement risk of Nadal in the match). This can happen in any situation where $W^{''}$ is small and there is a risk of retirement, like towards the end of this quarter final.

Such scenarios, as well as our choice of matches where only one player was injured, help to explain why our predicted retirement risk is generally much larger than the gap in the Betfair markets.

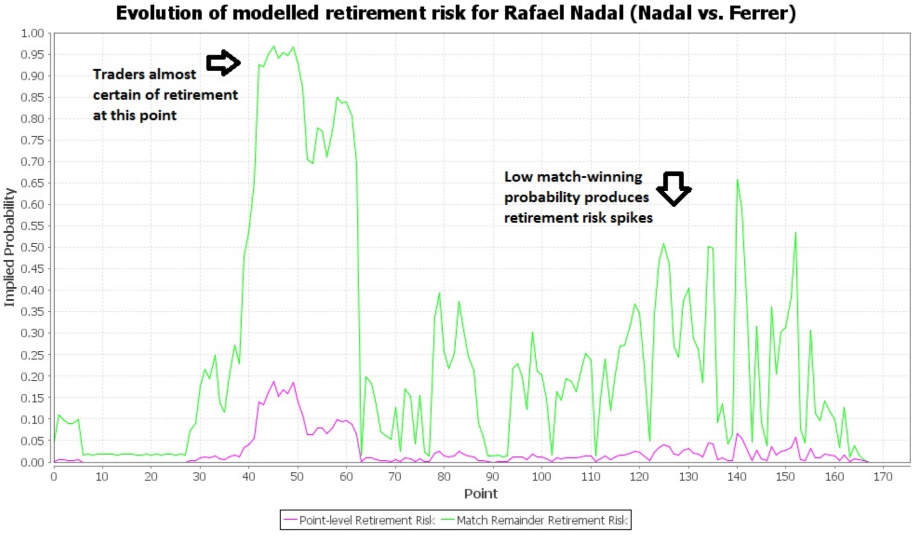


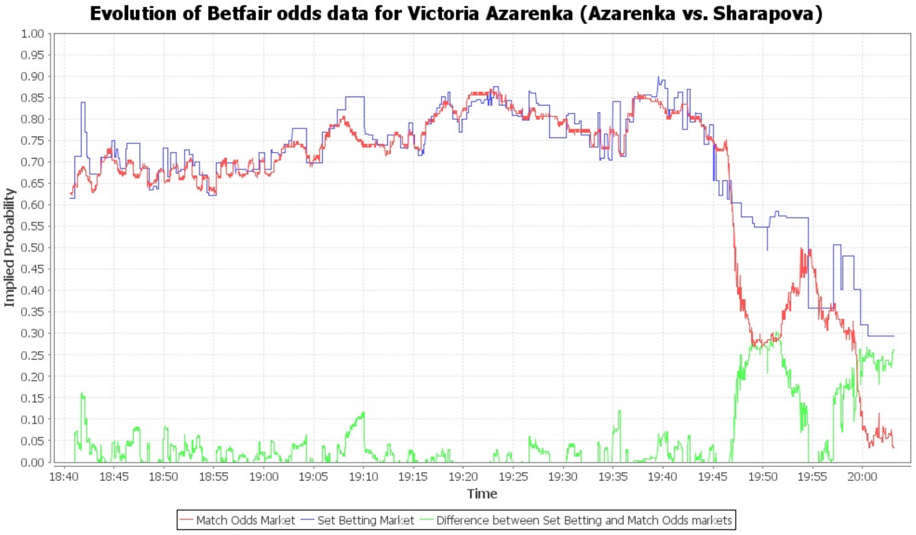
Figure 7.3.: Evolution of modelled point-level and remainder of match retirement risks for Rafael Nadal - *Nadal vs. Ferrer (Australian Open 2011 Men's Quarter Final)*

7.3. Victoria Azarenka vs. Maria Sharapova

Victoria Azarenka faced off against Maria Sharapova in the Rome Masters 2011 Quarter Final. Unfortunately, Azarenka suffered a hand injury early in the second set while leading one set to love and was unable to continue the match.

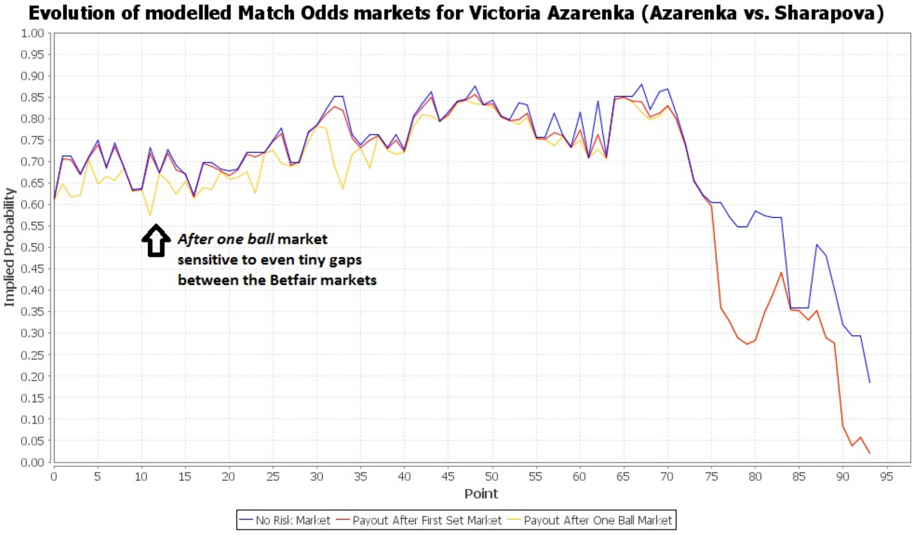
As you can see in Figure 7.6, we have spikes of retirement risk during the first set which correspond with sporadic drops in the implied probability of our predicted *after one ball* market. These anomalies appear to coincide correctly with small gaps created where the *after one set* market model is beneath the Set Betting market. However, the tiniest of these gaps can correspond to a significant drop in the implied probability of *after one ball* market. Whenever we have such a gap, the Nelder-Mead algorithm attempts to find a point-level probability (r_n) that will recreate this gap. Since the Betfair Match Odds market only takes into account retirement after the first set, only retirements after the first set can affect our model of this market. However, the system has no direct control over retirements *after* the first set if we are still in the first set. The more you try to increase the immediate point-level retirement risk, the more likely the given player will retire straight away (still in the first set). Lower it so the player will survive past the first set and he or she probably will not retire. The consequence of this is that our *after one ball* market model (and therefore our predicted retirement risk) is very sensitive to the Betfair odds source data, particularly towards the beginning of the first set.

Nonetheless, this match is still a good example of how unpredictable injury occurrence is. Azarenka's match-level retirement risk still stays relatively low until the injury occurs on point 75. The risk shoots up to 52%, peaking at 90% at retirement on point 93 after a brief dalliance with the possibility of recovery.



Evolution of Betfair odds data for Victoria Azarenka (Azarenka vs. Sharapova)

Figure 7.4.: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Victoria Azarenka - *Azarenka vs. Sharapova (Rome Masters 2011 Women's Quarter Final)* [ended in retirement]



Evolution of modelled Match Odds markets for Victoria Azarenka (Azarenka vs. Sharapova)

After one ball **market sensitive to even tiny gaps between the Betfair markets**

Figure 7.5.: Evolution of modelled Match Odds markets under three different retirement payout policies for Victoria Azarenka - *Azarenka vs. Sharapova (Rome Masters 2011 Women's Quarter Final)* [ended in retirement]

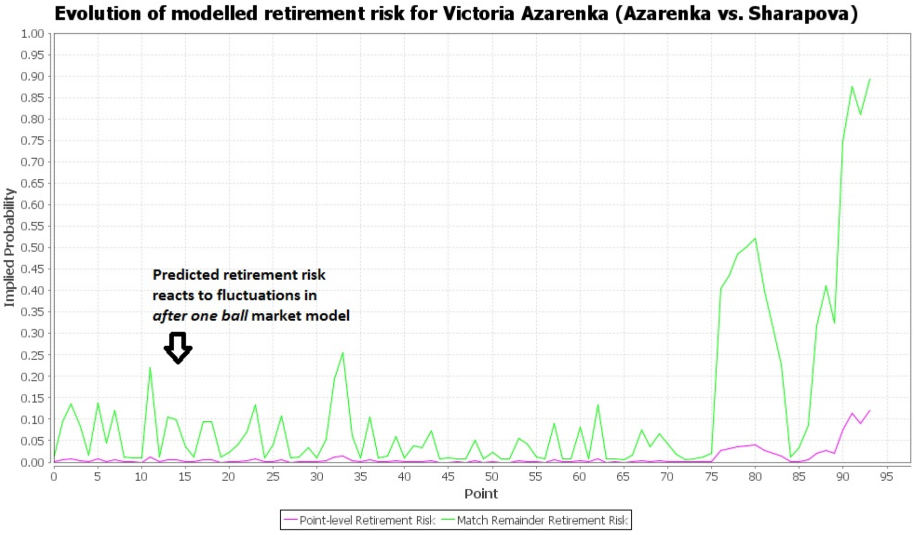


Figure 7.6.: Evolution of modelled point-level and remainder of match retirement risks for Victoria Azarenka - *Azarenka vs. Sharapova (Rome Masters 2011 Women's Quarter Final)* [ended in retirement]

7.4. Andy Murray vs. Jarkko Nieminen

During Andy Murray's recent French Open 2012 second round match against Jarkko Nieminen, he appeared to suffer a quite serious lower back injury in the first set of this match and looked very unlikely to be able to continue. Murray decided to struggle on and conceded the first set.

This can be seen in Figure 7.8, where Murray's Set Betting implied probability falls as traders suspect he might lose the match normally. As Murray recovers from his injury, the market realises he is unlikely to default and his risk of retirement dies away (Figure 7.9).

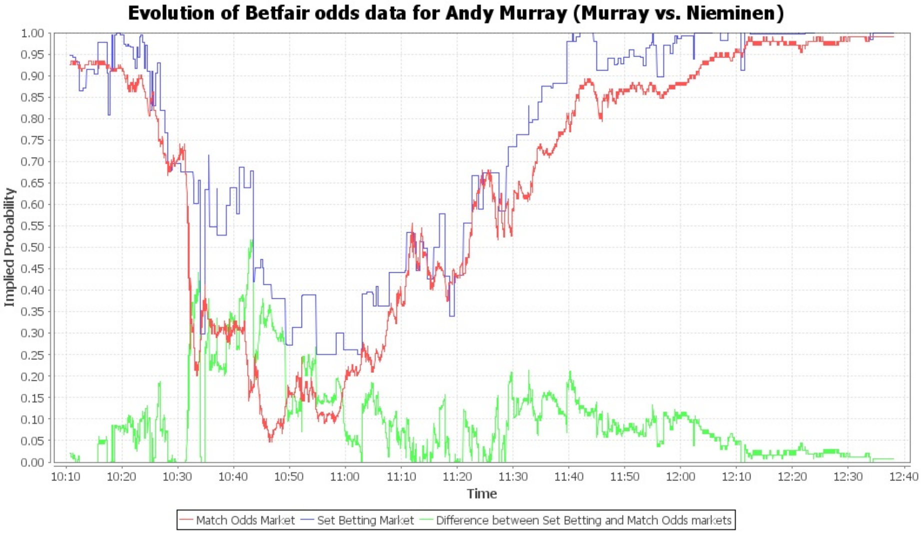


Figure 7.7.: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Andy Murray - *Murray vs. Nieminen (French Open 2012 Men's Second Round)*

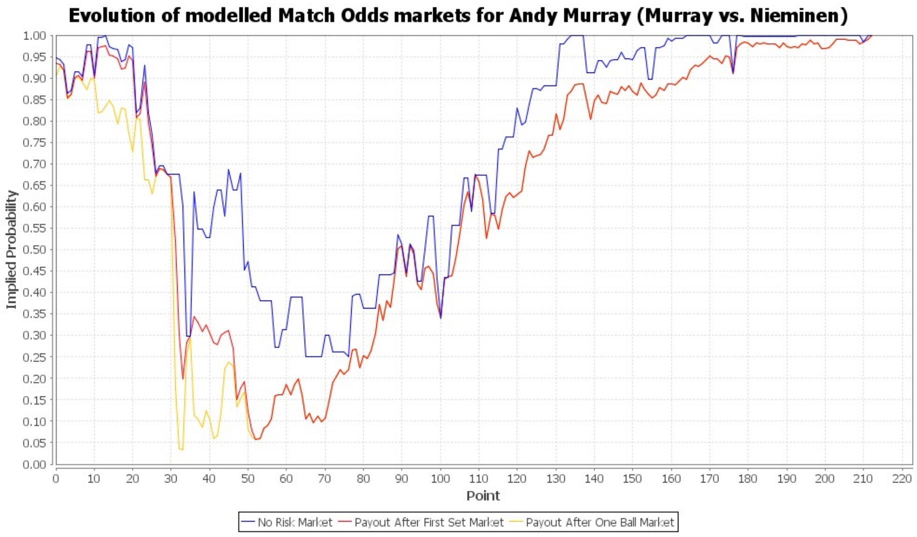


Figure 7.8.: Evolution of modelled Match Odds markets under three different retirement payout policies for Andy Murray - *Murray vs. Nieminen (French Open 2012 Men's Second Round)*

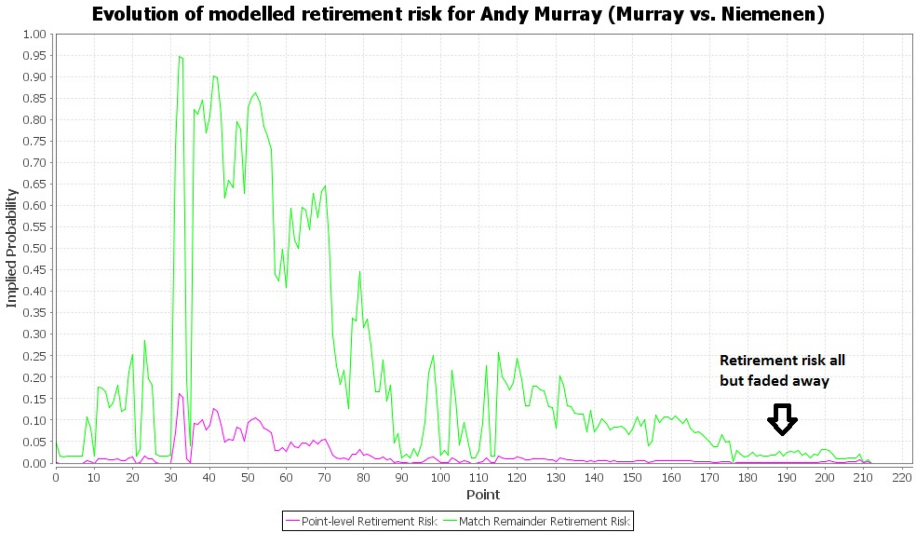


Figure 7.9.: Evolution of modelled point-level and remainder of match retirement risks for Andy Murray - *Murray vs. Nieminen (French Open 2012 Men's Second Round)*

7.5. Andy Roddick vs. Lleyton Hewitt

In the Second Round of the Australian Open 2012, Andy Roddick played Lleyton Hewitt in front of a partisan crowd. Roddick started strongly and went one-set up, but unfortunately suffered a hamstring injury after an awkward lunge early in the second set. He valiantly played on through to completion of the third set until retiring to avoid doing himself further damage.

Looking at Figure 7.12, we once again see that the gap between the Betfair markets has a notable effect on our *after one ball* prediction. You can see how Roddick's retirement risk spikes towards 100% as his Match Odds implied probability drops towards zero at the end of the match.

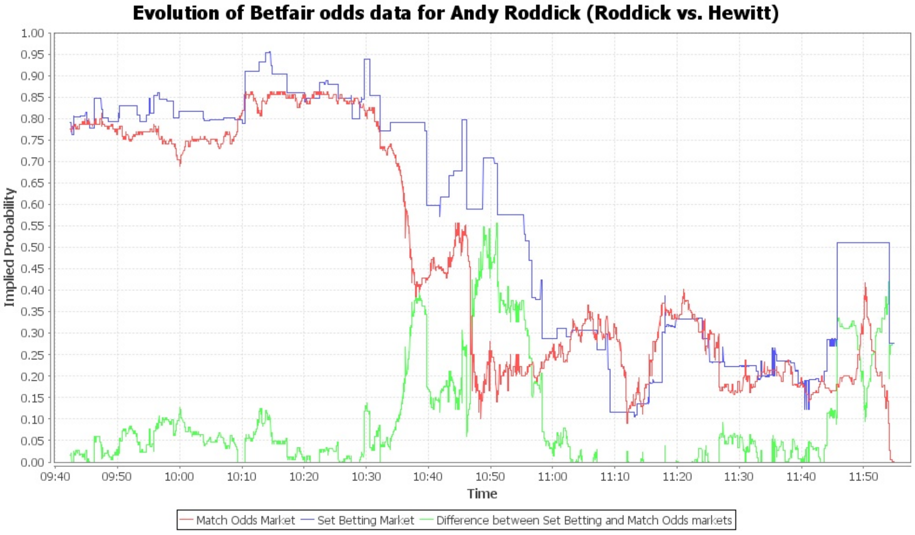


Figure 7.10.: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Andy Roddick - *Roddick vs. Hewitt (Australian Open 2012 Men's Second Round)* [ended in retirement]

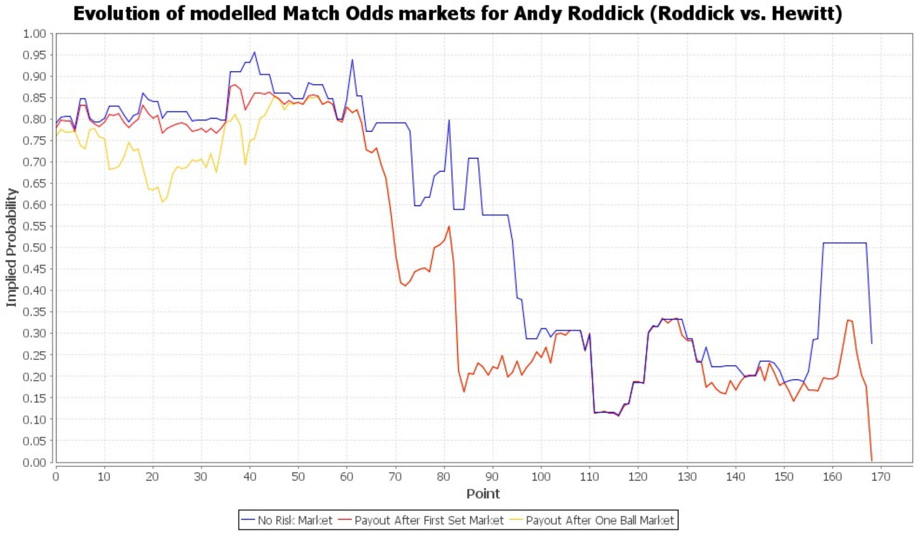


Figure 7.11.: Evolution of modelled Match Odds markets under three different retirement payout policies for Andy Roddick - *Roddick vs. Hewitt (Australian Open 2012 Men's Second Round)* [ended in retirement]

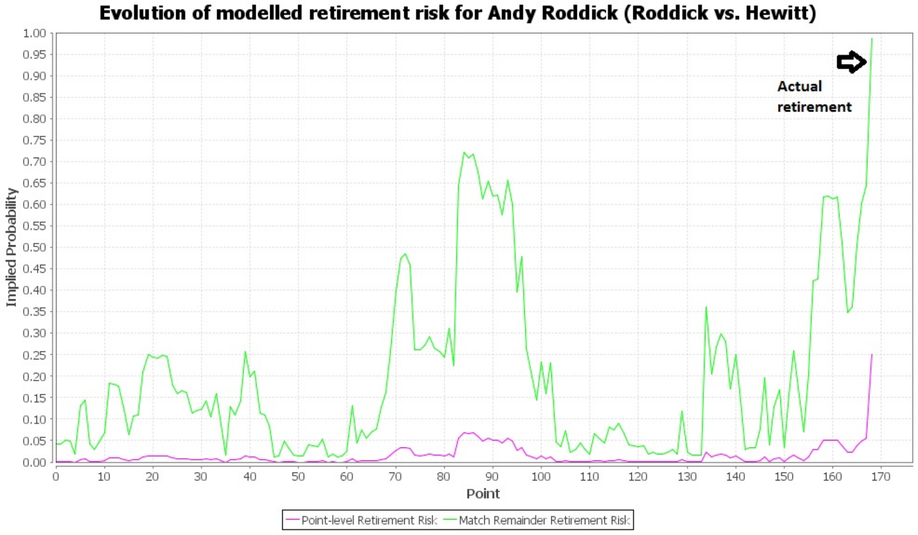


Figure 7.12.: Evolution of modelled point-level and remainder of match retirement risks for Andy Roddick - *Roddick vs. Hewitt (Australian Open 2012 Men's Second Round)* [ended in retirement]

7.6. Rafael Nadal vs. Juan Martin del Potro

In the Fourth Round at Wimbledon 2011, Rafael Nadal was matched up against Argentine Juan Martin del Potro. Nadal required treatment on his left foot just before the first set tiebreaker and despite initially limping between points, managed to edge the first set. He became stronger as he ran off the injury and went on to notched a straight sets victory.

Figure 7.15 shows that this match-level retirement risk indicates a slightly less severe injury than those we have seen previously, peaking at 65%. This match was the longest and most heavily traded of all our case studies. Consequently, our modelled markets and retirement risks are more dense and potentially more accurate than in the other matches.

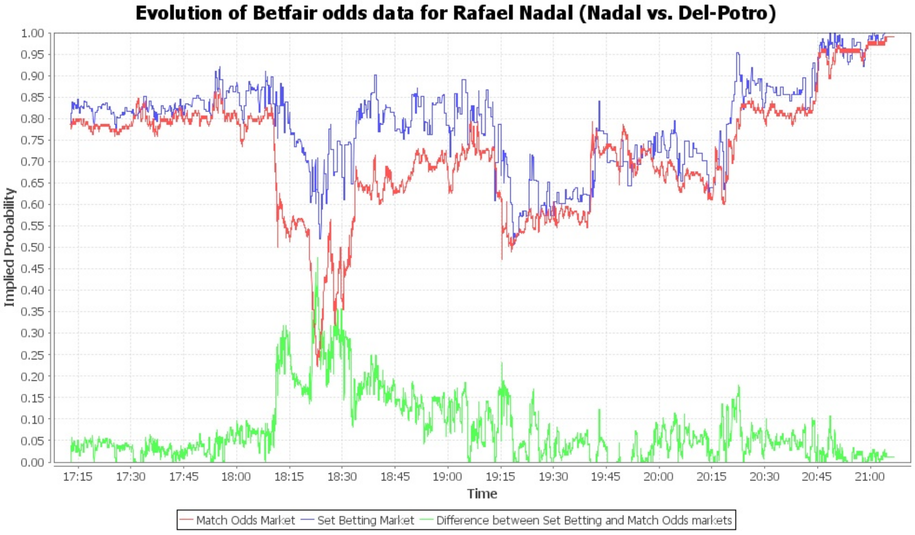


Figure 7.13.: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Rafael Nadal - *Del Potro vs. Nadal (Wimbledon 2011 Men's Fourth Round)*

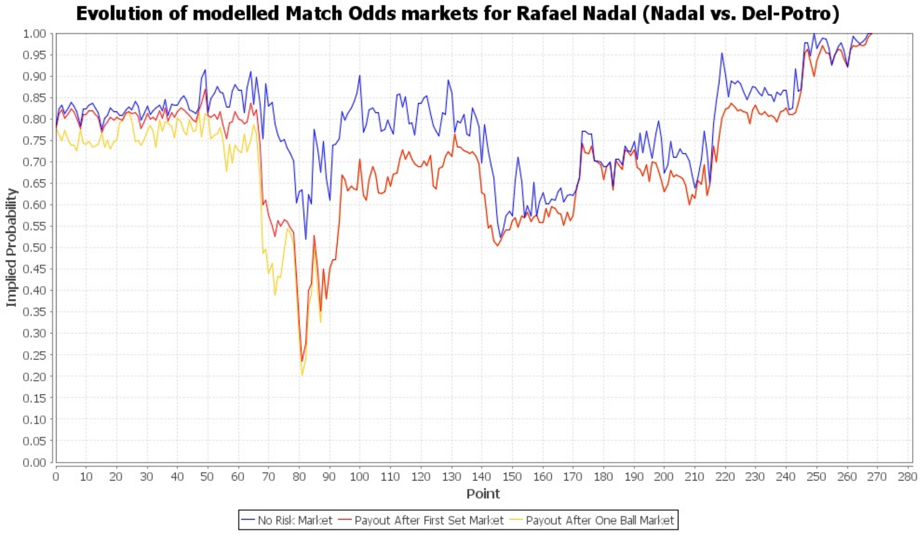


Figure 7.14.: Evolution of extracted Betfair Match and Set Betting markets implied match-winning probabilities as well as the gap between them for Rafael Nadal - *Del Potro vs. Nadal (Wimbledon 2011 Men's Fourth Round)*

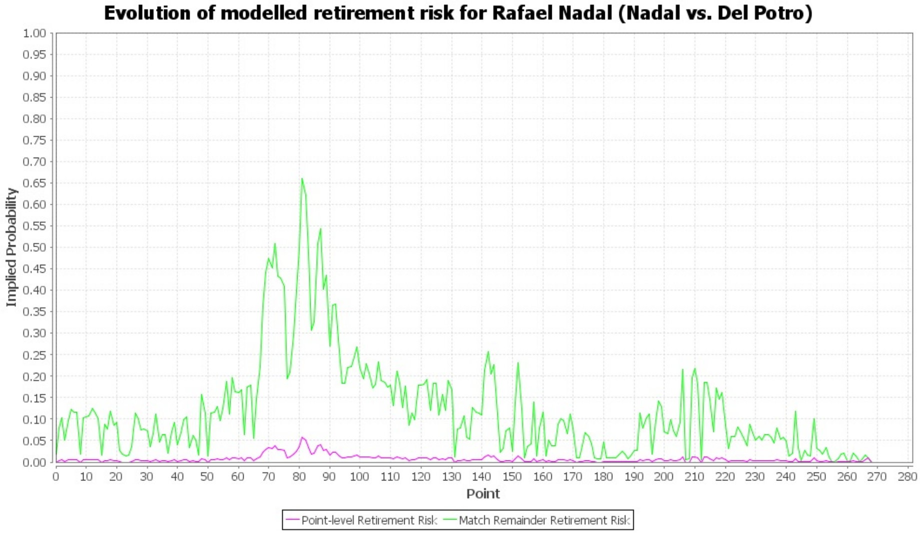


Figure 7.15.: Evolution of modelled point-level and remainder of match retirement risks for Rafael Nadal - *Del Potro vs. Nadal (Wimbledon 2011 Men's Fourth Round)*

Conclusions 8

Observing the Effect of Injuries on In-Play Tennis Betting Markets

We confirm that it is possible to observe the occurrence of an injury in a given tennis match by observing the evolution of the in-play betting odds. We examine historical Betfair odds for a number of real-life matches and find that a gap between the markets is, in fact, created in correspondence with the occurrence of an injury in the match. We note that injuries have a drastic effect on the odds data, in many occasions turning the overwhelming favourite into the underdog in an instant.

An Enhanced Model of Tennis incorporating Retirement Risk

We have created a new model for tennis in the form of a tennis match simulator which takes the retirement risk of players into account by incorporating extra parameters approximated using real-world averages and betting odds data as well as additional match outcomes compared to standard tennis models. This is the world's first attempt to create such a model.

We tested the model in a totally artificial environment and found it was able to mimic the expected patterns of in-play betting markets with different retirement payout policies for matches with varying injury scenarios.

Quantifying Retirement Risk in Professional Tennis Matches

We conclude that a given player's risk of retirement at some point during the remainder of a match is a function of the difference between the odds of a Match Odds market for that player that ignores injuries and the odds of a Match Odds market for that player that takes into account risk of retirement, at any given point in the match. Our system can provide a value for the retirement risk of a given player at any point in a match.

We applied the model to a number of real-world matches (from the point of view of an injured player) using Betfair odds data and produced imitations of the in-play betting markets following different player retirement payout rules. We find that we can mimic to a good degree of accuracy the progress of the Betfair Set Betting and Match Odds in-play markets throughout matches, i.e. a *match-completed* market and an *after first set* market, respectively. We attempt to use the Betfair odds data to predict the evolution of an *after one ball* market. We find that such a market generally correctly produces slightly higher retirement risks than the Betfair Match Odds market when the given match is still in the first set, although it can be somewhat erratic. The retirement risk values are very sensitive to any gap between the Betfair markets due to the lack of control our system has over retirement risk *after* the first set when the match is still in the first set. Furthermore, we notice that the system generates retirement risk spikes when the player in question has a low match-winning probability. The problem underlying these fragilities in our predictions is that the Betfair in-play tennis betting markets are not perfect. Since this odds data is vital input for our system, it is no surprise that fluctuating, anomalous, and sparse data heavily influences the output of our system.

9 Future Work

The Effect of Injury on Point-winning Probability

In theory, the Betfair Set Betting market ignores the risk of retirement and therefore should be a true indicator of a player's match-winning probability regardless of any injury. However, though the Set Betting market does not react as quickly as the Betfair Match Odds market to injury events, it *does* react. This suggests that injuries affect a player's point-winning probability, a factor which our model does not take under consideration. This is intuitive since a player limping around the court will be less able to win points than they are in normal circumstances. A further extension to this investigation would be to try and incorporate the effect of injury on point-winning probabilities into the model. Success in this endeavour would affect our updating of point-winning probability estimates on each point since they will also be modified by the model itself.

Accounting for Player Individuality

It is clear that some players are more injury prone than others, just as in any sport. For example, as of August 2011, 31-year-old Michael Llodra had retired from 25 matches and withdrawn from 2. His career total of 27 defaults is the highest of any current player. In contrast, Roger Federer has only withdrawn once and has never retired in almost 1000 career matches. Mischa Zverev, on the other hand, has already racked up 22 defaults in under 150 career matches at the tender age of 24[30].

We accept that we have made the generalising assumption here that injury rates are the same for all players, which is a weakness in the model. It would, however, be quite possible to tailor these parameters for individual players using their own past histories of retirement if one wanted to add further complexity to the model. For example, if we find that Samantha Stosur has retired from 2.8% of her matches on grass, than we can use the Nelder-Mead method to find personalised Bernoulli success probability c, point-level retirement risk magnitude rate λ, and recovery parameter ρ that produce this initial retirement rate for Samantha Stosur playing on grass (as well as another set of values for her opponent). The challenge with this approach is finding enough data about a player to allow for accurate parameterisation. It is somewhat rare that a specific player will suffer an injury in a match (Andy Murray and Rafael Nadal are *not* always getting injured despite the impression one might get from this report!), and even more unlikely that they will retire. If we also categorise matches by court surface, we will find it very difficult to piece together comprehensive data for any player.

Dependence on the Current State of the Match

We have assumed that retirement risk is homogeneous over all possible states of a tennis match. In reality, this assumption may not hold. There has been recent debate over whether players are far more ready to give up when carrying an injury and losing than they used to be. With the hectic ATP and WTA tour schedules and ever-intensifying demands of the sport itself, players may value avoiding further injury and maximising precious rest periods over playing 'lost cause' matches to completion just for the benefit of the viewers and sponsors. This implies that players may be more likely to retire if they are in a losing position compared to when they sense victory. There could also be other match situations which encourage retirement such as the end of a set or a switchover between games.

We note that our simulator is already much more sensitive to retirement when the match-winning probability of the given player is low, potentially modelling the expectation of traders that a player is more likely to retire as it becomes more difficult for them to win the match (although this was not an intended effect and thus is not under our control).

An 'After Two Sets' Match Odds Market

Just as we have created an *after one ball* Match Odds market, we could model an *after two sets* Match Odds market in a similar fashion. This would require our simulator to produce two more values, R_A^3 and R_B^3, which are the probabilities that Player A and Player B retire after the second set has been played, respectively. We would then compute:

$$\frac{(W_A' + R_B^3)}{(W_A' + R_B^3) + (W_B' + R_A^3)}$$

which is the probability that Player A wins normally plus the probability that Player B retires after the second set, normalised by the sum of the probabilities that either player wins the match normally and either player retires after the second set. We expect this to be a less popular market, however, since some matches can finish after two sets such as a straights victory in the women's game.

Parallelisation

The main disadvantage of our system is that it is very slow. This is mainly due to the execution of the Nelder-Mead method on every point. We know that a single 10,000 run simulation is no great drain on computing power (see Appendix C) but when we have over 200 points in a match, all these approximations take their toll. This could limit the real-time application of our model although we expect that the system would be able to keep up with the rate at which points are played in top-level tennis.

We can speed up the use of the Nelder-Mead method by increasing the convergence tolerance level or limiting the maximum number of iterations of the algorithm, but only at the cost of accuracy and therefore unwanted fluctuations in our modelled markets. A solution could be the use of parallelisation. There is no reason why we could not process each point in parallel as points do not depend upon each other. This would speed up the system significantly although it would require access to multiple machines or CPU cores. Further optimisations could be made by improving the code and making use of helpful features of programming languages other than Java.

Betfair provides a free API which allows users to programmatically connect to its exchanges in order to observe markets and place bets using their own software. It is likely that much of the trading that takes place on Betfair can be attributed to automated trading software rather than individual gamblers. This API would be available to us for gaining access to the real-time odds from the Set Betting and Match Odds markets should we require it.

An Exact Model

We have used a simulator that can only produces approximations as our model. As we touched upon earlier, we could try to create a mathematical state transition system model for tennis that incorporates retirement risk in order to calculate exact solutions. Such a model may also be faster than a simulator. Furthermore, we may be able to use such a model to investigate in more depth the relationship linking the gap between the Betfair Set Betting and Match Odds markets and our estimated match-level retirement risk values. In Appendix E, we attempt to explain the theory that would form the basis behind this future research.

A The Scoring System of Tennis

A singles match in tennis concerns only two players and consists of a best-of-three *set* or best-of-five *set* contest. Sets consist of a number of *games*. A player who has won one, two, or three *points* in a game, has the score 15, 30, or 40, respectively. A game is won only when one player has at least four points and has at least a two-point lead. 40-all is known as *deuce*. Whichever player wins the next point when the score is at deuce, gains *advantage*. However, if the opposing player then wins the following point, the score returns to deuce. Consequently, a game could, in theory, last forever. The players take turns serving each game, with the initial server determined by a coin toss. If the server is one point away from winning a game, the point is known as *game point*. If the returner is one point away from winning a game, the point is known as *break point*, i.e. he/she is on the verge of 'breaking the opponent's serve'. The score of a player who has no points in a game is known as *love*.

A set is won when one player has at least 6 games and at least a two-set lead. Alternatively, if the score is 6-all, there is a change of serve as per normal and then a *tiebreak* is (usually[1]) played. In a tiebreak, the player who serves first is the designated server for the 'game' but service changes on every odd point, e.g. after point 1, 3, 5, etc. A tiebreak is won when one player has at least 7 points and at least a two-point lead. The server of the first game of the set following a tiebreak is the player who served second in the tiebreak itself, i.e. no player is allowed to serve two games in a row[31].

[1] Today, only the final set in singles matches at three of the four Grand Slam tournaments (the Australian Open, the French Open, and Wimbledon), as well as in Davis Cup ties, do not use the tiebreak system.

B Base Tennis Model Pseudocode

Algorithm 6 double `game`(double p, CurrentGameScore *gameScore*)

```
if target player has won the game then return 1
end if
if opponent has won the game then return 0
end if
if score is deuce then return $p^2/[1-2p(1-p)]$
end if
return $p$ * $\mathbb{P}$(Win game after winning next point)
        + (1 - $p$) * $\mathbb{P}$(Win game after losing next point)
```

Algorithm 7 double `set`(double *pa*, double *pb*, CurrentSetScore setScore, CurrentGameScore *gameScore*, boolean *servingNext*)

```
if target player has won the set then return 1
end if
if opponent has won the set then return 0
end if
if tiebreaker then
    return tiebreak(pa, pb, new CurrentGameScore(), servingNext)
end if
double g
if servingNext then
    g = game(pa, gameScore)
else
    g = game(1 - pb, gameScore)
end if
return g * $\mathbb{P}$(Win set after winning next game)
        + (1 - g) * $\mathbb{P}$(Win set after losing next game)
```

Algorithm 8 double `tiebreak`(double *pa*, double *pb*, CurrentGameScore *gameScore*, boolean *servingNext*)

```
if target player has won the tiebreak then return 1
end if
if opponent has won the tiebreak then return 0
end if
if score is 6-6 then
    return ($pa$ * (1 - $pb$)) / (1 - ($pa$ * $pb$ + (1 - $pa$) * (1 - $pb$)))
end if
double $p$, $q$
if servingNext then
    $p = pa$
    $q = 1 - pa$
else
    $p = 1 - pb$
    $q = pb$
end if
// Note that service changes player on every odd point
return $p$ * $\mathbb{P}$(Win tiebreak after winning next point)
        + $q$ * $\mathbb{P}$(Win tiebreak after losing next point)
```

Algorithm 9 double `match`(double *pa*, double *pb*, CurrentMatchScore matchScore, CurrentSetScore setScore, CurrentGameScore *gameScore*, boolean *servingNext*, int *numSetsToWin*)

```
if target player has won the match then return 1
end if
if opponent has won the match then return 0
end if
double $s$ = set($pa$, $pb$, setScore, gameScore, servingNext)
return $s$ * $\mathbb{P}$(Win match after winning next set)
        + (1 - $s$) * $\mathbb{P}$(Win match after losing next set)
```

C Tennis Match Simulator Analysis

We experimented with the number of runs required to produce accurate results using our tennis match simulator. We repeatedly simulated a 5-set match with point-winning probabilities $P_A = P_B = 0.6$. As one might expect, the match-winning probability given by the mathematical model for such a scenario is 0.5. On each iteration we increment the number of runs by 100 up to 25,000. As Figure C.1 shows, anything at or above a mere 10,000 runs appears to consistently give results within 0.5% of the exact solution - sufficient for our needs.

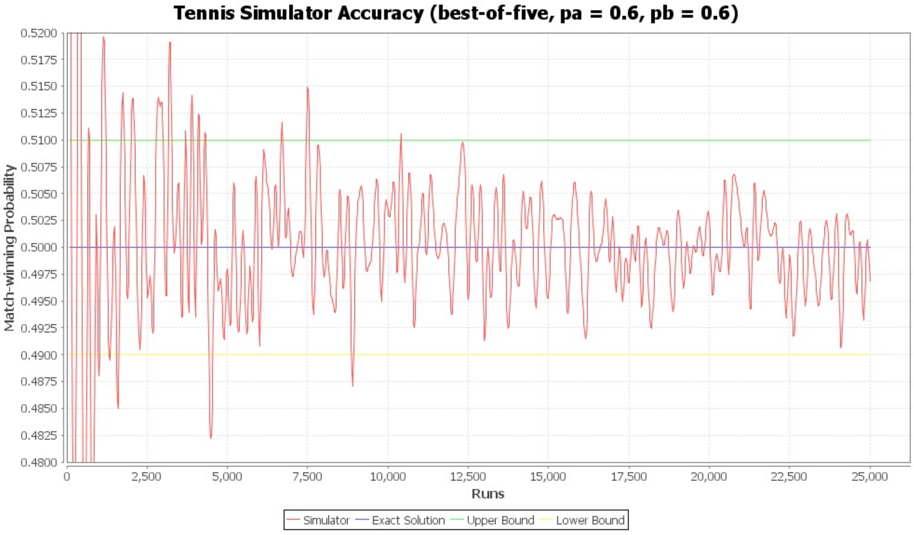


Figure C.1.: An experiment concerning the accuracy of our tennis match simulator

We also note that running 10,000 full matches on our modified tennis match simulator with $P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$ and $\rho = 0.95$ takes approximately 0.328 seconds on a state-of-the-art laptop.

D Truncated Exponential Distribution Proof

Say we have random variable $Y \sim exp(\lambda)$. We take $\lfloor Y \rfloor$ as the integer part of Y and $\{Y\}$ as the fractional part. We find the joint distribution:

$$\begin{aligned}\mathbb{P}(\lfloor Y \rfloor = k, \{Y\} \leq y) &= \mathbb{P}(k \leq X < k + y) \\ &= e^{-\lambda k} - e^{-\lambda(k+y)} \\ &= e^{-\lambda k}(1 - e^{-\lambda y})\end{aligned}$$

If we fix y and sum over all possible values of k we have:

$$\begin{aligned}\mathbb{P}(\{Y\} \leq y) &= (1 - e^{-\lambda y})\sum_{k=0}^{\infty} e^{-\lambda k} \\ &= \frac{1 - e^{-\lambda y}}{1 - e^{-\lambda}}\end{aligned}$$

$F(y) = 1 - e^{-\lambda y}$ is the CDF of an exponential distribution and $1 - e^{-\lambda}$ is the probability that an exponential random variable with rate λ is no greater than 1[32].

E A Possible State Transition System

Tennis matches are amenable to being modelled using Discrete Time Markov Chains as demonstrated by Liu (2001)[21] and Huang (2011)[15]. Any system that is modelled as a Markov Chain must fulfil the Markov Property:

$$\mathbb{P}(X_{n+1} = j | X_n = x_n, ..., X_0 = x_0) = \mathbb{P}(X_{n+1} = j | X_n = x_n)$$

for n, $j = 0, 1, ...$ and where X_n represents the state of the system at time n. In other words, the next state the Markov Chain transitions to is independent of any state the system was in prior to the current state. Unfortunately, our model cannot fulfil the Markov Property. On point n, each player has an updated retirement risk parameter r_n that is dependent on its value at point $n - 1$. The retirement parameters help to determine the subsequent state as they contribute to the normalised probabilities of each player winning the point or one of the players retiring. Consequently, the next state in the system is dependent on states prior the current state.

So we may not have a Markov Chain that we can study, but we do still have a labelled state transition system that we can reason about. An *absorbing state* in a state transition system is a state that once entered, is never left. A *transient state* is a state that is only visited by the system a finite number of times. In addition to the two absorbing states found in standard Markov Chain models tennis, i.e. either one of the players wins the match normally by achieving the required number of sets, we have two more: either one of the players can retire from the match on any point (just as with our simulator). If we also wanted to know the probability of a player retiring in a particular set, we would have to introduce further absorbing states (and complexity!). Every other state in our system is transient as each of these states represents a different possible score in the match and you never visit the same score twice in any given tennis match. We explain theory as described by Bause and Kritzinger (2002)[33] that could be applied to our model.

Say we have a set S_t of n_t transient states and a set S_a of n_a absorbing states. We number the states such that the n_a absorbing states occur first and write the one-step transition probability matrix:

$$P = \begin{pmatrix} I & 0 \\ B & Q \end{pmatrix}$$

I is the identity matrix with all element $p_{ii} = 1$ since once you enter an absorbing state you stay there. B is an $n_t \times n_a$ matrix describing movement from the transient to the absorbing states. Q is an $n_t \times n_t$ matrix describing the movement amongst transient states. $\mathsf{0}$ is the $n_a \times n_t$ zero matrix since you cannot move from absorbing states to transient states. We define the n-step transition probability matrix:

$$P^n = \begin{pmatrix} I & \mathsf{0} \\ N_n B & Q^n \end{pmatrix}$$

where $N_n = \sum_{i=1}^{n} Q^{i-1}$. Since as $n \to \infty$, $Q^n \to \mathsf{0}$ (intuitive since transient states eventually will not be revisited), $N_n \to (I - Q)^{-1}$ and the matrix $(I - Q)$ is invertible, we can say $N = (I - Q)^{-1}$ is the fundamental matrix of the system. So now we have:

$$\lim_{n \to \infty} P^n = \begin{pmatrix} I & \mathsf{0} \\ NB & \mathsf{0} \end{pmatrix}$$

We should be able to use this result to calculate the probability of reaching any of our absorbing states (any way the match can end) from any of our transient states (any given current match score). Unfortunately, the matrices B and Q are very large (there are a lot of possible match scores) and would take a long time to produce correctly.

Bibliography

[1] The Grand Slam Tennis Archive (accessed May 2012). http://www.tennis.ukf.net/stats15.htm.

[2] Global Betting and Gambling Consultants (GBGC). http://www.gbgc.com/2011/02/e-gaming-to-pass-us30bn-ggy-in-2011, February 2011.

[3] FracSoft. http://www.fracsoft.com.

[4] Peter Webb. Why Tennis is Big Business for Bookmakers. http://www.sportspromedia.com/guest_blog/peter_webb_why_tennis_is_big_business_for_bookmakers, January 2011.

[5] What are the bookmakers' different tennis rules? http://rebelbetting.com/faq/tennis-rules, July 2011.

[6] Wikipedia - Betfair (accessed January 2012). http://en.wikipedia.org/wiki/Betfair.

[7] Wikipedia - Betting Exchanges (accessed January 2012). http://en.wikipedia.org/wiki/Betting_exchange.

[8] Betfair Rules and Regulations (accessed January 2012). http://www.betfair.com/aboutUs/Rules.and.Regulations/.

[9] N. Jayanthi, J. O'Boyle, R.A. Durazo-Arvisu. "Risk Factors for Medical Withdrawals in United States Tennis Association National Tennis Tournaments: A Descriptive Epidemiologic Study". *Sports Health: A Multidisciplinary Approach*, 1(3):231–235, 2009.

[10] Exchange Betting Basics Part 1 - How Betfair Works (accessed January 2012). http://www.tennisbetting365.com/articles/7-exchange-betting-basics-part-1-how-betfair-works.

[11] Stock Market Spreads. http://www.thefti.com/stockmarketspreads.html, June 2008.

[12] C.D.Johnson, M.P.McHugh. "Performance demands of professional male tennis players". *British Journal of Sports Medicine*, 40(8):696–699, August 2006.

[13] S. Easton, K. Uylangco. "Forecasting outcomes in tennis matches using within-match betting markets". *International Journal of Forecasting*, 26(3):564–575, 2010.

[14] E. Servan-Schreiber, J. Wolfers, D.M. Pennock, B. Galebach. "Prediction Markets: Does Money Matter?". *Electronic Markets*, 14(3):243–251, 2004.

[15] X. Huang. "Inferring Tennis Match Progress from In-Play Betting Odds". (final year project), Imperial College London, South Kensington Campus, London, SW7 2AZ, June 2011.

[16] F.J.G.M. Klaassen, J.R. Magnus. "Forecasting the Winner of a Tennis Match". *Discussion Paper - Tilburg University, Centre for Economic Research*, 2001. 2001-38.

[17] J. O'Malley. "Probability Formulas and Statistical Analysis in Tennis". *Journal of Quantitative Analysis in Sports*, 4(2):Article 15, 2008.

[18] P.K. Newton, J.B.Keller. "Probability of Winning at Tennis I: Theory and Data". *Studies in Applied Mathematics*, 114(3):241–269, April 2005.

[19] T.J. Barnett, S.R.Clarke. "Using Microsoft Excel to Model a Tennis Match". `http://www.strategicgames.com.au/excel.pdf`, 2002.

[20] T.J. Barnett, A.Brown, S.R.Clarke. "Developing a Model that reflects Outcomes of Tennis Matches". *Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport*, pages 178–188, July 2006.

[21] Y. Liu. "Random Walks in Tennis". *Missouri Journal of Mathematical Sciences*, 13(3), 2001. www.math-cs.ucmo.edu/ mjms/2001.3/Yliuten.pdf.

[22] F.J.G.M. Klaassen, J.R. Magnus. "Are Points in Tennis Independent and Identically Distributed?". *Journal of the American Statistical Association*, 96(454):500–509, 2001.

[23] T.J. Barnett, S.R.Clarke. "Combining player statistics to predict outcomes of tennis matches". *IMA Journal of Mathematical Statistics*, 16(2):113–120, 2005.

[24] F.J.G.M. Klaassen, J.R. Magnus. "How to reduce the service dominance in tennis? Empirical results from four years at Wimbledon". Open Access publications from Tilburg University, 2000.

[25] P.K.Newton, K.Aslam. "Monte Carlo Tennis: A Stochastic Markov Chain Model". *Journal of Quantitative Analysis in Sports*, 5(3), 2009.

[26] A. Marek. "Applying a Hierarchical Markov Model to Tennis Matches". (final year project), Imperial College London, South Kensington Campus, London, SW7 2AZ, June 2011.

[27] Betfair Developers Programme - Why are the prices displayed on the website different from what I see in my API application? (ac-

cessed May 2012). http://bdp.betfair.com/index.php?option=com_content&task=view&id=242&Itemid=68.

[28] Betfair Developers Programme - What is the algorithm for displaying prices inline with the website? (accessed February 2012). http://bdp.betfair.com/index.php?option=com_content&task=view&id=237&Itemid=62.

[29] Nelder-Mead Example (accessed May 2012). http://math.fullerton.edu/mathews/n2003/neldermead/NelderMeadMod/Links/NelderMeadMod_lnk_5.html.

[30] Tennis' Biggest Quitters - ATP Retirements / Withdrawals. http://www.tennisbetting365.com/articles/7-exchange-betting-basics-part-1-how-betfair-works, September 2011.

[31] International Tennis Federation - Rules (accessed January 2012). http://beta.itftennis.com/about/organisation/rules.aspx.

[32] B. Eisenberg. "On the expectation of the maximum of IID geometric random variables". *Statistics and Probability Letters*, 78(2):135–143, February.

[33] F.Bause and P. Kritzinger. *"Stochastic Petri Nets - An Introduction to the Theory"*. Bause and Kritzinger, 2002.