University of Glasgow | School of Computing Science

Predicting the Outcome of Tennis Matches From Point-by-Point Data

Martin Bevc (1006404b)

April 24, 2015

ABSTRACT

Mathematical tennis modelling is increasing in popularity and mostly being driven by recently sparked worldwide interest in data analytics, which is spawning a whole new segment of the sport industry. Open source software packages for computational statistics that are making the application of more advanced algorithms easier even for non specialists which are making their way into sport betting, and the traction of online betting exchanges are also factors for a strong interest in sports predictions.

Bettors use predictive modelling to estimate the probability of a player winning a match and place bets based on their predictions. Millions of pounds are invested in different tennis betting markets around the world at any time of the tennis season.

The majority of published papers use a Hierarchical Markov chain model to describe a tennis match, to which estimates of a player's probability of winning a point on serve are passed as parameters. Estimates are generally based on overall match statistics working under the assumption that the parameters do not change during the match and are not updated with live point by point data.

This paper presents possible improvements to the common tennis model by deviating from the common opponent model and basing predictions on point by point data. The question of what is the optimum combination of historical and current data for making match outcome predictions is explored, and experimental solutions are presented. By knowing how much weight should be put on historical data better match outcome predictions can be made. The methods described are shown to perform better than methods used in previous published research which rely on historical data only.

1. INTRODUCTION

Tennis is one of the most popular sports in the world. The format of the game has made tennis one of the most heavily traded sports in betting markets, and with an opportunity for big profits, interest in tennis predictions is high among professional traders and recreational gamblers.

The game is played between two players with only two possible outcomes to a match as well as to every point played. In doubles tournaments tennis is played between four players, but this paper focuses on singles competition. The scoring system is relatively fine grained and reliably reflects events on court and the progression of the match. This is in contrast to sports like football where modelling in-play dynamics mathematically might be a substantially harder problem to solve and is probably also one of the reasons for the growth of many in-play betting markets for tennis.

There are some irregularities to tennis rules however which complicate modelling slightly. Lengths and formats of matches can vary because slightly different rules can be used at different tournaments.

At a late stage of a given match, it is usually easy to predict the winner. Any person familiar with the rules of tennis can correctly predict who is more likely to win at match point for example. At that stage there is sufficient current data available - score, number of serves served and points won on serve for both players in the match, as well as other information such as break points won etc. to make a highly informed prediction based just on this current data - data acquired during the current match.

Making a good prediction about the outcome of the match becomes increasingly more difficult the further from the end of the match one tries to make it, since there is less current data available to make good predictions from. The rules of tennis require players to perform well over time, and different factors affect player's performance. It is thus very hard to make a good prediction about the outcome of the match after observing, say the first point of the match.

To bypass this restriction historic data is often used to make up for the lack of current data. Data from many previous matches is aggregated to construct a profile of a player's performance over a longer time frame than one match, in the hope that when this data is fed to the model the predictions will be more accurate and the effects of partial current data minimized.

A side effect of this solution can be a failure of the model to recognize and include strong signals from the current data and appropriately base predictions on them. For example if a top player has a "bad day" and is about to lose in straight sets, the model can still favor him late in the match based on the aggregated historic data of good performances. It displays low variance and high bias.

The predictive power of the model could therefore improve if the right balance between historic and current data could be used to better asses the parameters passed to the model.

This paper proposes a solution for this problem and evaluates it against methods used in previous research.

2. BACKGROUND

A tennis match consists of sets, which consist of games, which in turn consist of points. To win the match a player therefore has to win the sequence of points which yields the required number of games and sets. This structure makes it possible to model a match as a hierarchical Markov model, which consists of all possible Markov chains for a particular event. A Markov chain is a construction of a sequence of random variables which represent possible states of the modelled event. The transitions in the chain are the probability a of player winning a point on their serve, or the probability of the opponent winning a point on return. These two probabilities must sum up to 1 as each point has two possible outcomes. O'Malley [1] demonstrated that a Markov chain can be derived for any tennis match.

Since the points are represented as random variables in the model, they are considered to be independent and identically distributed (IID) thus satisfying the Markov property - future events are independent of the past. Klaassen and Magnus [2] have shown that the IID assumption does not hold exactly in tennis, but the deviation from the ideal case is small. The majority of previous research relies on this hypothesis for its simplicity.

Figure 1 shows a graph representation of a tennis game modeled as a Markov Chain. It can be seen that in a tennis game a player has an equal probability of winning the game at the score 30-30 and Deuce, as well as at 40-30 (30-40) and Advantage. p is the probability of a player winning a point on their serve and *(1 - p)* represents the probability of the opponent winning the point on their return. To represent the whole match this model is scaled up to represent a set of games and match of sets in equal fashion with the exception of transitions then representing the probability of a player winning a game and set respectively. The Markov chain representation of a set is presented in figure 2, figure 3 shows the tiebreak model and figure 4 shows the chain for a best of 3 match.

Hence in this model the probability of a player winning the match is directly dependent on his probability of winning a point on serve. A useful observation is also that different values of p can be fed to the model at each transition without violating the general ideas behind it. This research takes advantage of this fact.

By solving the chain model, a probability of winning a match can be obtained for a chosen player. Theoretically a game can be played indefinitely from the score of deuce onward if each player wins one point alternately (The same can occur in a tiebreak or in the 5th set of some grand slam tournaments). This makes it hard to solve the model analytically, although approximations in form of equations do exist [4]. Therefore it can sometimes be easier to implement the model so that the results are obtained by simulation.

Considering we have a model which relies solely on p to make predictions, it becomes important to estimate it accurately, or at least estimate p - q correctly. q is defined as

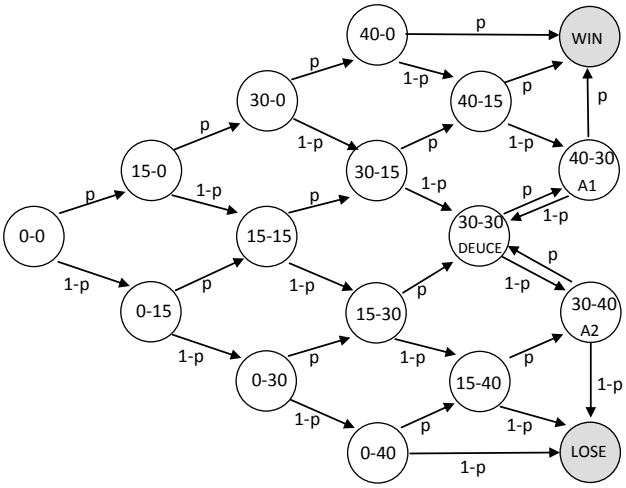


Figure 1: *Adapted from [3]. A Markov model of a tennis game.*
The nodes represent all possible scores in a game. p is the probability of the player serving to win the point. p-1 is the probability of the player returning to win a point. The WIN and LOSE nodes are terminal points in a game. After one or the other state is reached a new game begins at 0-0.

the probability of of player 2 to win a point on his serve, the same as p is for player 1. This is done by gathering historical data on overall match statistics and computing estimates relative to a players performance against the average past opponent in majority of published research.

Some use a slight variation of this approach. For example Barnett [5] uses an updating function to adjust the values for different surfaces and Newton and Aslam [6] use the variance in a player's points won on serve and return to adjust the opponent's serve winning probability. Knottenbelt et al [7] develop a common-opponent approach by using only the subset of historical data containing past opponents that both players have previously encountered. Klaassen and Magnus [2] use all the data available to compute the average probability of a player winning a point on serve - what they call the field value.

2.1 Klaassen & Magnus

Klaassen and Magnus build a deterministic Markov chain model described in section 3.1. They develop a method for estimating p and q before the start of a match from a dataset of played matches. They then use the obtained p and q as inputs to the model, to calculate the players' probability of winning the match from any given score. It is important to note that p and q remain fixed during the match - they are based solely on historical data and do not include new information from the currently playing match in the probability estimates at any point in the match.

This method is used as one of the baselines in this paper. However, there is a downside to the described method. By keeping p and q fixed during the match, there is a chance of overestimating or underestimating a player's probability to win a match. This method shows poor performance in

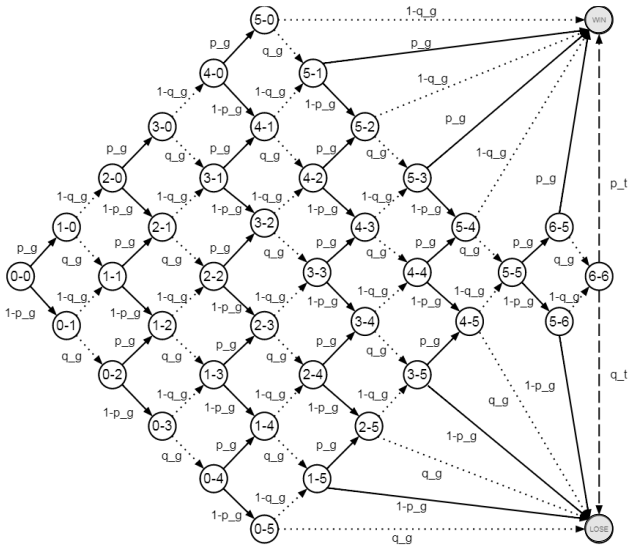


Figure 2: *Adapted from [3]. A Markov model of a tennis set. Nodes represent games. p_g represents the probability of a player to win a game on his serve, and q_g represents the probability of the opponent to win a game his serve. p_t is the probability of player winning a tiebreak, and q_t the probability of the opponent to win a tiebreak.*

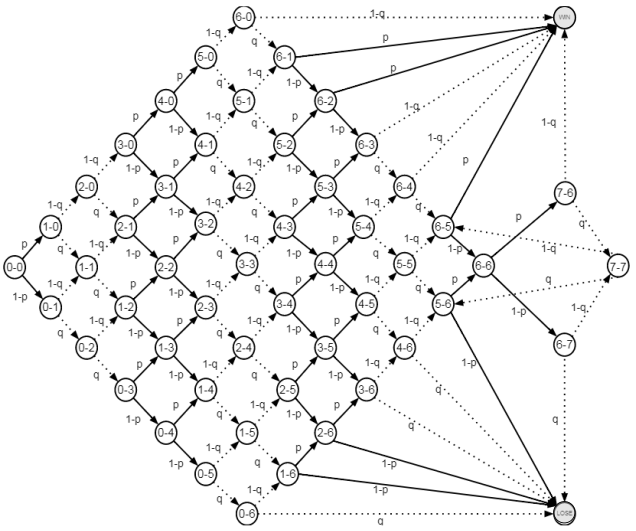


Figure 3: *Adapted from [3]. A Markov model of a tiebreak game.*
Nodes are possible scores. p and q are defined the same as in model 1.

cases where a player's performance in the current match differs greatly from historical performance, which can be made worse further by using field data to estimate p and q in the first place, since a single player's performance on serve in the current match could diverge greatly from the historical field estimates. In this case the model will still favor the losing player late in a match, or the opposite, will not favor the winning player, since none of the current match data are taken into the account in the process of estimating the winning probabilities.

This paper suggest a solution to this issue in Section 3, and compares the performance of the proposed method against the one described above in Section 4.

2.2 Newton & Aslam

Netwon and Aslam recognize that a player's probability of winning a point on serve is not constant throughout a tournament but varies from match to match. They conclude that it is therefore better modelled as a random variable whose probability density function closely resembles a Gaussian [6]. They then use a combination of the analytical formulas to solve the chain model and Monte Carlo simulations to obtain a probability density function (pdf) for a player to win a game on serve, which they use in further head to head simulations to obtain the pdfs for a player to win a match. In their further analysis they focus on ranking related analysis. Similar to Klaassen and Magnus, they also use their data to define a field of players when performing their analysis.

3. APPROACH

The model and all the experiments were developed in the Python[1] programming language and made use of Pandas[2], Numpy[3] and libraries among others. The majority of exploratory analysis was done in the IPython Notebook[4] computational environment.
Many great libraries for data analysis and scientific computing have evolved in the Python ecosystem and were particularly helpful in this research. Pandas allowed for easy manipulation of the dataset and together with the rapid prototyping capabilities of the language and author's experience with it the choice of tools was obvious.

3.1 Deterministic Markov Model

Based on the details in section 2 a Markov chain model was built. The model takes three input parameters, p and q - the probabilities of players winning a point on their serve, and a score. The model outputs the probability of the player with serve winning probability p winning the match from a given score. This model replicates the model built by Klaassen and Magnus [8].

Barnett [4] defines the backward recursion formula for the Markov chain model with the following notation. Players A and B have a constant probability p_A and p_B of winning a point on their serve. With probability p_A the state changes from the score a,b to a+1,b and with probability $q_A = 1 - p_A$ it changes from a,b to a,b+1. a,b represent the current score for player A and B respectively (30-30 = 2-2). Hence P_A is the probability that player A wins the game on serve when the score is (a,b) and we have the following recursion formula:

$$P_A = p_a P_A(a+1, b) + q_A P_A(a, b+1) \quad (1)$$

The boundary values are:
$P_A(a, b) = 1$, if $a = 4$ and $b \leq 2$ and

[1] www.python.org
[2] http://pandas.pydata.org/
[3] http://www.numpy.org/
[4] www.ipython.org/notebook

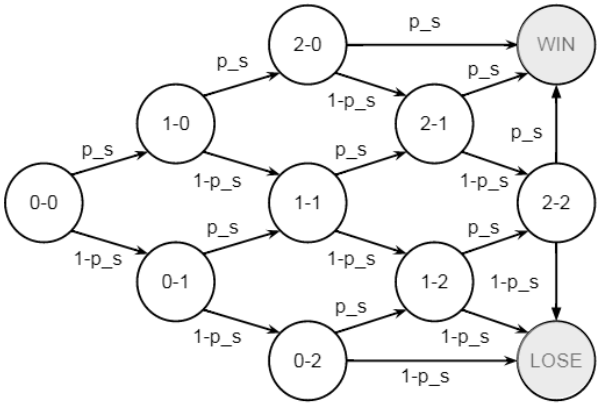


Figure 4: *Adapted from [3]. A Markov model of a best of 3 sets tennis match.*
Nodes represent sets p_s is a probability of a player to win a set. It can be seen how the game model is extended to sets and match, and how the probability of winning a match is dependent on probabilities of winning a set, a game and a point.

$P_A(a,b) = 0$, if $a \leq 2$ and $b = 4$.
Barnett then defines an explicit formula for handling the deuce score, realizing that the chance of winning from deuce equals the form of a geometric series and the equation can be expressed as:

$$P_A(3,3) = \frac{p_A^2}{p_A^2 + q_A^2} \tag{2}$$

For player A serving first the conditional probabilities to win the tiebreaker $P^T(x,y)$ from score (x,y) is:

$$P_{tiebreaker}(x,y) = p_A P_{tiebreaker}(x+1,y) + (1-p_A)P_{tiebreaker}(x,y+1) \quad \text{for } 2 \leq (x+y+3) \bmod 4 \leq 3 \tag{3}$$

$$P_{tiebreaker}(x,y) = p_B P_{tiebreaker}(x+1,y) + (1-p_B)P_{tiebreaker}(x,y+1) \quad \text{for } 0 \leq (x+y+3) \bmod 4 \leq 1 \tag{4}$$

The boundary values are:
$P_{tiebreaker}(7,y) = 1$ when $x - y \geq 2$,
$P_{tiebreaker}(x,7) = 0$ when $y - x \geq 2$ and
$P_{tiebreaker}(6,6) = \frac{p_A(1-p_B)}{p_A(1-p_B)+(1-p_A)p_B}$.

Assuming that player A serves first in the set the probabilities of winning a set from set score (x,y) are:

$$P_{set}(x,y) = p_A^{game} P_{set}(x+1,y) + (1-p_A^{game})P_{set}(x,y+1) \text{ for even } (x+y) \tag{5}$$

$$P_{set}(x,y) = p_B^{game} P_{set}(x+1,y) + (1-p_B^{game})P_{set}(x,y+1) \text{ for odd } (x+y) \tag{6}$$

Where p_A^{game} is the probability of player A winning a game from score (0,0) while serving and p_B^{game} is the same for player B.

The boundary values in this case are:
$P_{set}(x,y) = 1$ if $x \geq 6$, $x - y \geq 2$,
$P_{set}(x,y) = 0$ if $y \geq 6$, $y - x \geq 2$ and
$P_{set}(6,6) = p_{tiebreaker}$ where $p_{tiebreaker}$ is the probability of player A winning a tiebreaker from score (0,0) while serving first.

Finally Barnett derives the following equation:

$$P_{match}(x,y) = p_A^{set} P_{match}(x+1,y) + (1-p_A^{set})P_{match}(x,y+1) \text{ for even } (x+y) \tag{7}$$

$$P_{match}(x,y) = p_B^{set} P_{match}(x+1,y) + (1-p_B^{set})P_{match}(x,y+1) \text{ for odd } (x+y) \tag{8}$$

Where $P_{match}(x,y)$ is the probability of player A winning the match from match score (x,y), p_A^{set} is the probability of player A winning a set from score (0,0) while serving first and p_B^{set} is the same for player B.

The boundary values in this case are:
$P_{match}(3,y) = 1$ for $y < 3$,
$P_{match}(x,3) = 0$ for $x < 3$ and
$P_{match}(2,2) = p_A^{set}$.
This set of boundary values holds for a 5 set match. They are similar for a 3 set match.

3.2 Stochastic Model

A module which implements the rules of tennis and allows for simulation of a match on a point by point basis was developed. The module can correctly keep score and calculate and update statistics on percentage of serving points won for each player while the match is being simulated. A score and sequence of point outcomes relative to the player serving with an additional parameter to choose the serving player can be accepted as inputs, as well as p and q. The first is useful for back-testing matches and for the purpose of verification of the module's operation. The latter is used for simulations.

Hence the module also allows for replaying of a match based on historical point by point data for an arbitrary number of points, then simulating the match forward from a given point in the match. This feature is used in the experiments this research is based on.

3.3 Evaluation Metrics

In predictive modeling an increase in accuracy often comes at the expense of simplicity. Therefore the performance gains have to warrant the increase in the complexity of the model in order for the building of the model to make sense. A simple baseline was established for the purpose of measuring these gains and comparing the performance of methods used in published research to the ones developed as part of this research. The baseline used is "the higher ranked player always wins the match".
Rank being the ATP singles ranking of a player at the time of the match [5]. All experiments are compared against this baseline in Table 3.

3.4 Replication of Klaassen and Magnus

Using the Wimbledon dataset detailed in Table 3.7 and the Markov chain model described in section 3.1 we attempted to replicate the results of Klaassen and Magnus

[5] http://www.atpworldtour.com/Rankings/

as closely as the descriptions of their procedure would allow with the intention of benchmarking our newly proposed improvements to the models which are the state of the art in published research.

Match winning probabilities were computed for every score of every match contained in the dataset. p and q used as inputs are set to *beta_i* and *beta_j* dataset fields, which were precomputed by Klaassen and Magnus for use in their research. They are the probabilities of players winning a point on their serve and are kept constant in their experiments.

3.5 Experiment 1

Our first experiment uses the same setup as described in section 3.4, but the use of historical data is omitted in prediction making. For every match, the current statistics on players' serving get calculated by observing the first 30 points of the match. p then gets calculated using equation 9. q is obtained by the identical equation with the difference of serving statistics for player 2 being used.

$$p = \frac{points\_won\_when\_player\_1\_serves}{total\_points\_served\_by\_player\_1} \tag{9}$$

p and q are passed as inputs to the model and match winning probabilities are obtained from the model. These results can be obtained by either the deterministic model our by our stochastic model. The results from the stochastic model approach those of the deterministic model when increasing the number of simulations we make. Since this consumes unnecessary CPU cycles every match was simulated once and for every score *P(win)* was computed using the deterministic model.
If the stochastic model is used the match winning probability at a given point for the player with point winning probability p is determined by Equation 10.

$$P(win) = \frac{matches\_won}{matches\_simulated} \tag{10}$$

3.6 Experiment 2

The second experiment is based on the hypothesis that assumes sampling points from a probability distribution could be an equal or better alternative to calculating point estimates of serve winning probabilities. This builds on previous research by Newton and Aslam [6].

A beta distribution is assumed to be the best option as we are predicting binary outcomes.
To see why a beta distribution is the correct solution to our problem consider the following scenario. We want a good estimate of p for a player. As defined in equation 9 we might call it conversion rate on serve. If we obtain a player's conversion rate over a long period of time, say a season, or several seasons, we can reason quite well about the player's ability to convert points on serve. But for reasons described in section 2, we might want to put more weight on current data. In case when we don't use any historical data, equation 9 will be a bad measure at the start of a match. If the player converts the 1st point, his conversion rate will briefly be 1, and if he does not it will be 0. But this is unlikely to be realistic, as historic data shows such extremes are unlikely to occur over longer periods of time. Hence we know that conversion rate is a bad predictor at the start of the match.

```
pl1_total_serves = 0
pl1_won_serves   = 0
pl2_total_serves = 0
pl2_won_serves   = 0

for point in match:
    if empirical start reached:
        p = pl1_won_serves/pl1_total_serves
        q = pl2_won_serves/pl2_total_serves
    if pl1 serving:
        P(win) = match_probability(p,q,
    score)
        if pl1 wins point:
            pl1_won_serves += 1
        pl1_total_serves += 1
    if pl2 serving:
        P(win) = 1 - match_probability(q,p,
    score)
        if pl2 wins point:
            pl2_won_serves += 1
        pl2_total_serves += 1
```

Figure 5: *A pseudocode snippet implementing part of the experiment 1. p,q are set to Klaassen's p,q at the start. When up to "start" points are observed, p,q get reevaluated and use only current data from point "start" forward.*

Historic data is useful in giving us prior expectations.
Conversion rate can be represented with a binomial distribution - a series of won or lost points on serve and the best way to represent the prior is with a Beta distribution. The domain of a beta distribution is [0, 1], the same as p.
A Beta distribution can be expressed as:

$$\beta(\alpha, \beta) = Kp^{(\alpha-1)}(1-p)^{(\beta-1)} \tag{11}$$

The parameter α can in this case be interpreted as points a player has served and won, where $\alpha + \beta$ gives total points served by the player.

The expected value of the Beta distribution is defined as:

$$E = \frac{\alpha}{\alpha + \beta} \tag{12}$$

In the experiments, each player starts of with his own beta distribution which is initialized such that

$$p = \frac{\alpha_p}{\alpha_p + \beta_p} \text{ and } q = \frac{\alpha_q}{\alpha_q + \beta_q} \tag{13}$$

In all experiments using the beta distribution we set E to equal p at the start. At the initialization stage p and q are sourced from the data set and set to equal *beta_i* and *beta_j*.

3.6.1 Beta Experiment 1

After p and p are initialized as described previously, alpha values are set to 20 for both players (This number was chosen randomly for the first experiment, see table 2 for a suggestion of an optimal value for this parameter). Beta values can then be worked out from this setup. The match is allowed to be played out for 50 points during which alpha and beta values for the player who is serving are updated as follows:

$$\alpha' = \alpha + n \begin{cases} n = 0 & \text{if player loses point} \\ n = 1 & \text{if player wins point} \end{cases}$$
$$\beta' = \beta + (1 - n)$$

p and q parameters are then set to equal E of the updated beta distribution, and fed to the model. The match is then simulated a 100 times from every next point in the match. At any given point in the match *P(win)* of the serving player is determined by Equation 10.

3.6.2 Beta Experiment 2

The 2nd experiment is set up the same way as the 1st case, with the exception of p and q getting sampled once from the updated beta distribution at the start of every simulated match.

3.6.3 Beta Experiment 3

The 3rd experiment going one step further, samples p and q at every point of every match we simulate. Comparison of the results from all three methods are presented in the Evaluation section.

3.7 Dataset

All experiments use the Wimbledon dataset[6]. The dataset contains match data at point level over four years, 1992-1995, for the Wimbledon Grand Slam tournament. 481 matches are recorded for men's and women's singles matches. It contains a total of 88883 points. Match data was recorded only on the 5 most important courts of the tournament and the amount collected accounts for approximately half of the matches played during those 4 years. For every match, the players and their rankings are known, as well as the exact sequence of points played. 1st and second serve details were also recorded, as was data on whether the point was decided through an ace or double fault. A summary of the data is provided below.

	Men	Women
Matches	258	223
Sets	950	503
Final Sets	51	57
Games (excl tiebreaks)	9367	4486
Tiebreaks	177	37
Points	59466	29417

Some overall statistics on the dataset are presented in Table 1.

Statistic	Men	Women
Average points per match	230	131
Longest match (in points)	453	240
Shortest match (in points)	115	61

Table 1: Overall dataset statistics.

Wimbledon is a tournament played on grass which is a very fast surface. This generally serves well powerful servers who are able to win many points with aces. It tends to be the tournament where most aces are scored by players. Hence it is interesting to look at some stats about aces and compare them between men and women, who often play with a different style and tactics.

[6]http://www.janmagnus.nl/misc/file508545.xlsx

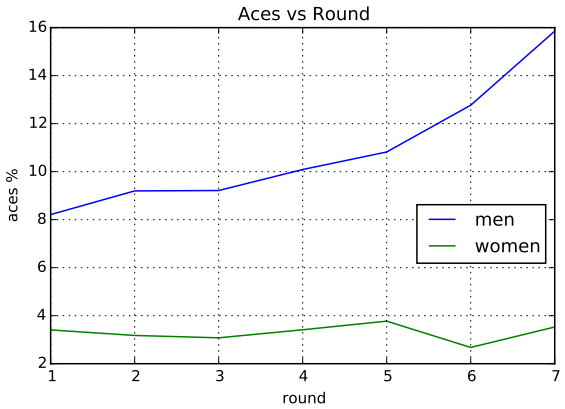


Figure 6: Points won with an ace as a percentage of points served, aggregated by round.

Figure 6 shows that the serve is a bigger weapon in men's game. 16% of all served points were won with an ace in the final rounds of the tournament and less than 4% in women's finals. Another interesting observation is the rate of aces rising with every round. That can be explained by the fact that top players tend to have powerful serves, and players with powerful serves are likely to win matches at Wimbledon an hence fight their way through to the later stages of the tournament.

Another dataset was manually collected from a video broadcast of the 2014 ATP Tour Finals in London. It is not used in this research however. Tour Finals matches are played to a best of 3 sets. The tournament also has a different format than Wimbledon, being a Round Robin event. Furthermore, the final was canceled because of an injury of one of the players and an exhibition event was played instead. As it would be hard to draw meaningful comparisons with the Wimbledon dataset, analysis of the smaller dataset is omitted, it might however prove to be useful in future research.

4. EVALUATION

Table 3 contains the performance benchmark results for all methods used in this research. The baseline, although very simple performs quite well. This would suggest that ATP player rankings are actually a relatively good predictor of match winners, and are representative of players' skills and performance on tour.
However, there could be more reasons for good baseline performance. Our dataset only includes data from one Grand Slam tournament - Wimbledon. Being one of the most prestigious tournaments in tennis the majority of top players in the world are sure to compete. At this level the rank difference between two players can represent a much bigger difference in skill than between two lower ranked players. There is usually a bigger performance difference between players ranked 1 and 4 than between 250 and 259. This effect could be further emphasized by the fact that detailed point by point data in our dataset was only recorded on the few main courts, where matches with top players are usually prioritized when scheduling, because they draw a bigger audience. On the other hand the baseline is a fairly crude method of making predictions - it is a binary prediction at any point before or during the match.

The method used by Klaassen and Magnus outperforms the baseline while it also allows for expression of confidence in one's predictions at any given point during the match. Instead of binary the predictions are on an interval [0,1]. Nevertheless, as explained in section 3 this approach has the downside of greatly favoring historic data over current data when making predictions. An example of this occurring is presented in Figure 10(c). If only relying on current match data the model is able to make better predictions (in this case 150 points into the match).

4.1 Empirical

To eliminate the bias on historic data we next look into how the same model performs if only fed current data. As presented in Plot 10, at a later stage current data can be a better predictor of match outcome than historic data. Table shows that at the end of 2nd set (men) or 1st set (women) the model performs better with current data only.
Plot 7 shows the median as well as 25th and 75th percentiles for value of *P(win)* across all matches played by men for Magnus (historic data) versus Empirical (current data) for every point $>= 150$. It can be observed that after point 210 the model performs significantly better with current data only, while that claim cannot be made for points played earlier in the match.

Even Earlier in the match however when there is not yet enough current data observed the model makes poor predictions. Figure 10 shows an example.

4.1.1 When Current Data Fails

To demonstrate how using only current data can lead to bad predictions lets examine the next case. Figure 10 shows the 1992 quarter final round played between Andre Agassi and Boris Becker. Becker was ranked 5th in the world at the time while Agassi was ranked 14th. By our simple baseline Becker is therefore the favorite. The *P(win)* probabilities are plotted from the perspective of Becker. A total of 305 points were played in the match. Becker, with a better serve winning probability is also the favorite at the start of the match if using the Klaassen and Magnus method.

Using the empirical method - only relying on current data, we observe the match for 38 points (subplot 1). During these points Becker manages to win most of points on his serve. At point 38 Becker's p is 0.8, while Agassi's equals 0.63. This is a relatively big difference and the result is obvious. Solely based on the observation of 38 points from the current match the model predicts Becker will win with probability almost equal to 1. The confidence of the model in this prediction is flawed however as it is unlikely that Becker can continue winning points on his serve with $p = 0.8$ for the rest of the match.

The middle subplot of Figure 10 shows how the results change if we observe a 100 points of the match first. Enough current data is observed by this point to better reflect the reality on the court. The model even makes a slightly better prediction based just on current data.

The rightmost subplot reflects the true power of observing enough current data however. When observing 180 points before making any predictions based on the empirical method, the model is now certain that Becker is losing. In contrast it can be seen that the model still favors Becker to win at point 250 with probability close to 0.7 when using the Klaassen and Magnus method.

Since we want to take advantage of the good performance of the empirical method once enough current data is observed while at the same time avoid the incredibly poor performance when data is unbalanced and too sparse to reflect the real events on court, we turn to the Beta experiments.

4.2 Beta Experiments

Table 3 shows the performance of all Beta Experiments with their α values initialized to 20. Alpha values control the balance between the current and historic data. By setting α to different values we control how much our prior expectations influence the shape of the beta distribution we later sample from. The higher we set alpha to be initially the less influence will current data have on the shape of the distribution. For low values of α we favor current data more.

4.3 Setting alpha parameter

By setting the α parameter when initializing the beta distributions we control the balance between historical and current data. As mentioned in section 3.6 $\alpha + \beta$ can be interpreted as the number of points served by the player. This effectively means that when we initialize the beta distribution we set the number of points observed on a player's serve before the current match starts with $\alpha+\beta$. For example setting alpha to 2 means two points where the player served and won were observed. Hence with alpha and beta parameters we set our prior. From this it follows that if we initialize the distribution with a low alpha, the effect of current data with which we update the alpha and beta parameters during the match on the distribution will be large. In contrast, if we initialize the distribution with $\alpha = 10000$ the effect of current data will be small, since the average match in the dataset is 230 points long.

Table 2 show performance of the model at different values of α.

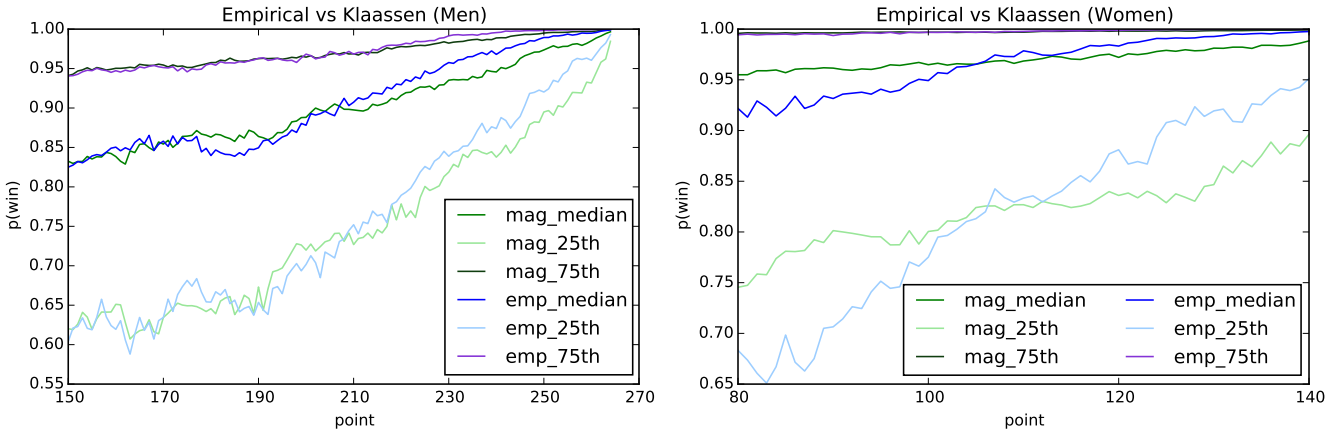
alpha	run 1	run 2	Average model performance
2.0	88.37%	87.21%	87.790%
10.0	87.21%	87.60%	87.405%
200.0	88.76%	89.15%	88.955%
400.0	88.76%	89.15%	88.955%
500.0	89.15%	88.37%	88.760%
700.0	88.76%	87.98%	88.370%
1,000	89.15%	87.6 %	88.375%
10,000	87.98%	89.15%	88.565%

Table 2: Alpha vs Average prediction accuracy.

4.3.1 Example

The effect of different alpha values described in Section 4.3 is best demonstrated on an example. Figure 9 shows the different distributions for the 1992 1st round match between J. Courier and M. Zoecke. The solid lines represent the initial distributions. The dotted lines are the updated distributions. They are updated as described in Section 3.6.1.

When $\alpha = 10$ the shape of the distribution represents our uncertainty in out prior beliefs. After we observe 50 points



(a) Men

(b) Women

Figure 7: *Empirical vs Magnus percentiles. Figure 7a shows the performance of the Empirical method vs that of Klaassen and Magnus over all matches in the dataset for men. Figure 7b shows the same for all women's matches. The 25th, 50th and 75th percentiles are plotted at every points in matches. To account for different lengths of matches, the starting point for our calculations is the length of the shortest match from the end of every match (115 points for men, 61 for women). Every match contains at least 115 or 61 data points for men and women respectively, counting backward from the end of every match. That is how we ensure data aggregated at each point contains every match in the set.*

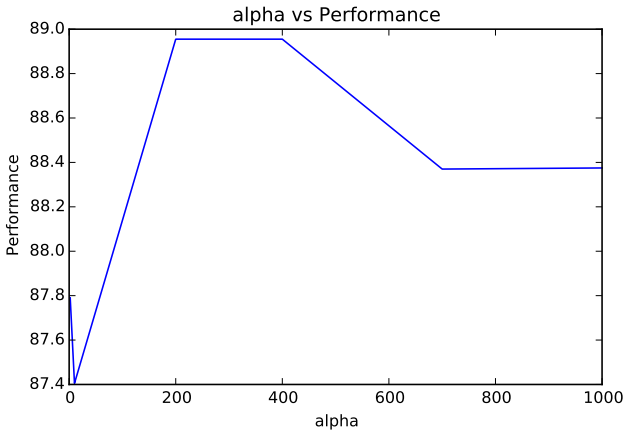


Figure 8: *Values of α subject to test that give the best performance are between 200.0 and 400. This results is interesting as suggests a balance of historical and current data gives the best prediction accuracy. If a much higher value of α would give the best results it would mean historical data is more important to good prediction making, and if the value was lower it would mean that current data is more important.*

of the match the distributions our we can make a more confident estimate of the true values of players' p and q as we have observed more data. Therefore the width of the distributions is reduced.

The second and third plot of Figure 9 show the same process with the exception of the alpha parameter being initialized to 200 and 10000 respectively. Because we initialize the distribution with more "observed points prior to the start of the match", our confidence in our prior beliefs are higher than in the previous case. This is reflected in the shape of the distributions.

4.4 Overall Results

Table 3 shows the percentage of correct match outcome predictions after 2 sets played for men and 1 set played for women, for different methods. The baseline is "the highest ranked player always wins". The rest of the methods are counted "correct" when P(win) > 0.5 for the player that ends up winning the match. Correct predictions are aggregated over the whole dataset and divided by total number of matches in the set. It is interesting to observe that the baseline performs better for the women's dataset, which would suggest that higher ranked female players were consistently more dominant. This is a curious result as from observing the modern game at the time of writing one would conclude the opposite. Top ranked male players are consistently winning matches with few upsets while the women's game is much more volatile.

Method	Men	Women
Baseline	78.3%	82.5%
Klaassen & Magnus	88.4%	84.3%
Empirical	86.8%	85.2%
Beta Experiment 1	89.5%	87.0%
Beta Experiment 2	89.5%	87.9%
Beta Experiment 3	89.1%	84.3%

Table 3: Prediction accuracy for different methods (P(win) > 0.5).

Furthermore table 3 shows our Beta method outperforms the Baseline andEmpirical method which is expected, however it also performs better than the Klaassen method, although by a small amount. To be able to conclude which Beta method performs best with more certainty more tests should be performed with different values of α.

5. FUTURE WORK

This paper improved upon the model commonly used in previous research in terms performance in predictive ability

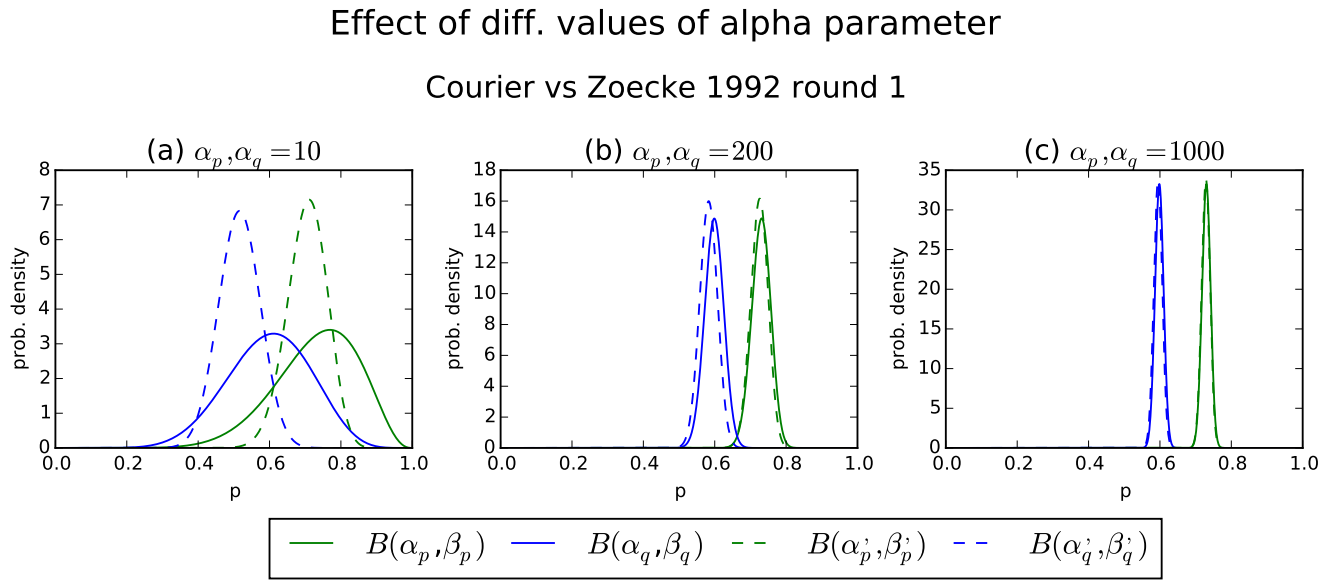


Figure 9: *Shows effect of different settings of alpha on beta distributions on the example of the 1992 1st round match between Courier and Zoecke. The figure shows how the balance between historic and current data is controlled by setting α. The solid lines represent the distributions when they are initialized with their respective value of alpha (10 in plot (a), 200 in plot (b), etc.). β values are set as described by equation 13. The parameters are then updated as defined in equation 3.6.1. It can be observed from the plots that if α is set to a small number like in (a) the effect of the current data with which we update the parameters is bigger than if the value of α is increased. This means that in case (a) we rely more on the current data observed so far when we start simulating the match, while in case (b) and even more in case (c) we rely more on the historical data.*

of match outcomes.
The next logical step is to back test the improved model on actual historic point by point betting market data. It has proved impossible to obtain such data for our data set as markets such as Betfair did not exist when the Wimbledon data set was recorded. It is easier to record current betting market data, although some information like live score streams are still quite hard to obtain since they are valuable and companies often restrict free access.

A further step is testing the model with different betting strategies. It would be interesting to consider the performance of our enhanced model together with strategies like hedging, as well as accounting for commission on net profit and other real world scenarios.

It would also be interesting to run a bigger number of simulations in our Beta experiments, as well as comprehensively test different Beta experiments across a range of α values and also on women's dataset. Because of time constraints the number of simulations per point had also been set relatively low. Code improvements for better performance would also be beneficial.

Moreover, systems like HawkEye[7] collect numerous parameters about matches with high precision and incredible detail. Sadly we were unable to obtain any datasets with in play data on movement of players, speed and spin of the ball etc. However, the rich detail of these datasets would likely allow for a better and even more interesting insight into the game of tennis. It could also be possible to enhance the predictive model to a sub point level, where we could make predictions about point outcomes from relative positions of players on the court, the power of their strokes etc. It could be possible to infer a player's physical condition during the match by comparing their reaction speed, distance covered on court and the speed of the ball leaving their rackets.

6. CONCLUSIONS

Making match outcome predictions from a combination of in play and historical data is shown to be more effective than relying on a single method of either lumping together all available historical data or using just the current in-play data. Historical data contains useful information and can prove beneficial especially in early stages of matches when current in play point by point data is sparse. Using point sampling from a beta distribution can offer the right balance between historic and current data. Optimizing the α parameter for a given data set can further improve performance.

During research for this paper we encountered the common pitfalls of data analysis - data being hard to access, incomplete or in formats hard to scrape.

However the proposed method of sampling from a beta distribution offers a way of circumventing some of the issue of lacking data as it is shown that it is possible to achieve a good performance without comprehensive historical data, hence potentially making our method good for real world application in betting. Further research with betting data from a betting exchange like Betfair would be compelling.

[7] http://www.hawkeyeinnovations.co.uk/

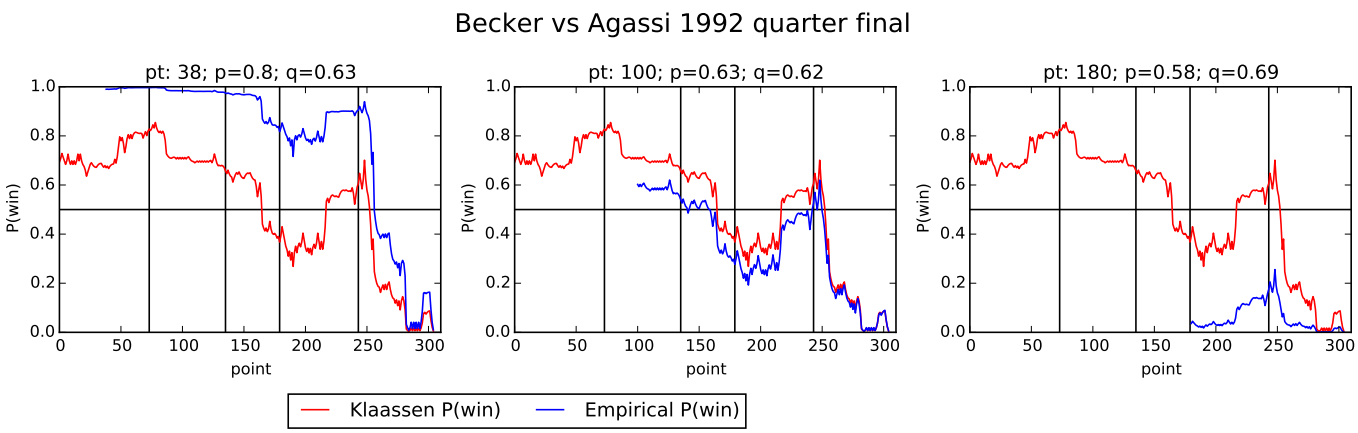


Figure 10: *"pt" is the number of points observed before p,q are calculated with the empirical method. Calculated with the Klaassen method p = 0.68 (Becker) and q = 0.65 (Agassi). They stay fixed through the match. The plots show P(win) for Becker. Hence, Becker was the favorite but lost the match. The figure compares the performance of empirical and Klaassen methods on a single match between Becker and Agassi. (a) Shows the downside of using only current data early in the match (blue line). Not enough points have been observed yet to make accurate predictions and the empirical method vastly overestimates Becker's probability to win the match. (b) demonstrates how the empirical method improves as we collect more data from the match. If p and q are estimated at point 100 instead of 38 as in (a) the model outputs a slightly better prediction than the Klaassen method which relies solely on historical data. (c) shows the outcome prediction for the empirical method gets even better if p and q are estimated even later in the match, at point 180. At the same time it demonstrates the downside of using only historical data to make the prediction as Klaassen's method still favors Becker at point 250 with probability close to 0.7. To try to take advantage of the superior performance of the empirical method later in the match while avoiding its bad performance at the start of the match when current data is sparse we design the Beta experiments.*

7. REFERENCES

[1] A. J. O'Malley, "Probability formulas and statistical analysis in tennis," *Journal of Quantitative Analysis in Sports*, vol. 4, no. 2, 2008.

[2] F. J. Klaassen and J. R. Magnus, "Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 500–509, 2001.

[3] A. M. Madurska, "A set-by-set analysis method for predicting the outcome of professional singles tennis matches," 2012.

[4] T. J. Barnett, *Mathematical modelling in hierarchical games with specific reference to tennis*. PhD thesis, Ph. D. Thesis, Swinburne University of Technology, Melbourne, Australia, 2006.

[5] T. Barnett and S. R. Clarke, "Combining player statistics to predict outcomes of tennis matches," *IMA Journal of Management Mathematics*, vol. 16, no. 2, pp. 113–120, 2005.

[6] N. P. K and A. Kamran, "Monte Carlo Tennis: A Stochastic Markov Chain Model," *Journal of Quantitative Analysis in Sports*, vol. 5, pp. 1–44, July 2009.

[7] W. J. Knottenbelt, D. Spanias, and A. M. Madurska, "A common-opponent stochastic model for predicting the outcome of professional tennis matches," *Computers & Mathematics with Applications*, vol. 64, no. 12, pp. 3820–3827, 2012.

[8] F. Klaassen and J. R. Magnus, *Analyzing Wimbledon: The power of statistics*. Oxford University Press, 2014.