IMA Journal of Management Mathematics (2005) **16**, 113–120
doi:10.1093/imaman/dpi001

Combining player statistics to predict outcomes of tennis matches

TRISTAN BARNETT† AND STEPHEN R. CLARKE‡
School of Mathematical Sciences, Swinburne University, PO Box 218, Hawthorn, Victoria 3122, Australia

With the growth in sports betting, it is possible to bet, both before and during a match, on a range of outcomes in tennis. This paper shows how the standard statistics published by the ATP can be combined to predict the serving statistics to be obtained when two given players meet. These statistics are then used in a spreadsheet model to predict further match outcomes, such as the length of the match and chance of either player winning. These calculations can be updated as the match progresses. The method is demonstrated by focusing on a very long men's singles match between Roddick and El Aynaoui played at the 2003 Australian Open.

Keywords: tennis; Australian Open; scoring systems; Excel; gambling.

1. Introduction

There are four major grand slam tennis events played each year, all exhibiting their own unique set of characteristics. Wimbledon is played on grass, the French Open on clay and the US and Australian Open on hard court. Even the scoring systems are different. Most tour events are played as best of 3 sets, but in men's singles, the grand slam events are played over 5 sets. However, the US Open plays a tie-breaker game at 6 games all in the fifth set while the other grand slams play an advantage fifth set. A tie-breaker game is won by a player that first reaches at least 7 points and is ahead by at least 2 points. An advantage set is won by a player that first reaches at least 6 games and is ahead by at least 2 games.

With the growth in sports betting, it is now possible to bet on a range of outcomes in tennis. Bookmakers such as Ladbrokes offer odds not just on the eventual winner, but on match score lines (3 sets to love, 2 sets to 1, etc.). Index betting, where the total number of games in the match is bought or sold at a given value, is also possible. In many cases the odds are updated as the match progresses, e.g. in one product offered by Ladbrokes, bets are made on the point score for each game, with new odds posted when the player's previous service game is concluded.

In planning daily draws, tournament organizers also have an interest in predicting the chances of each player advancing in the draw, and the probable length of matches. In recent years there have been a number of grand slam matches decided in long fifth sets. In the third round of the 2000 Wimbledon men's singles, Philippoussis defeated Schalken 20–18 in the fifth set. Ivanisevic defeated Krajicek 15–13 in the semi-finals of Wimbledon in 1998. In the quarter-finals of the 2003 Australian Open men's singles, Andy Roddick defeated Younes El Aynaoui 21–19 in the fifth set, a match taking 83 games to complete and lasting a total duration of 5 h. The night session containing this long match required the following match to start at 1 am. Long matches require rescheduling of following matches, and also create scheduling problems for media broadcasters. They arise because of the advantage set, which gives more chance of winning to the better player (Pollard & Noble, 2002), but has no upper bound on

†Email: tbarnett@swin.edu.au
‡Email: sclarke@swin.edu.au

© The authors 2005. Published by Oxford University Press on behalf of the Institute of Mathematics and its Applications. All rights reserved.

the number of games played. It is clearly in the interests of broadcasters and tournament organizers to be able to predict when they are likely to occur.

Various authors have produced tennis models that require as input the chances of each player winning their serve. Once these are known, and various assumptions about independence of points are made, the chances of each player winning, winning by a given score and expected lengths of a game, set or match can be calculated. Klaassen & Magnus (1998) use point-by-point data from matches completed during the Wimbledon championships 1992–1995 to calculate input probabilities. Pollard & Noble (2002) assume a range of values near overall averages. However, for the applications discussed above, estimates which take into account not only the various scoring systems in use but also the playing characteristics of the two players and the surface on which the match is played are necessary. Estimates are required prior to a match, not after. In our case we rely on player statistics as published on the web to make predictions before the match.

This paper shows how player statistics can be combined to predict the outcome of tennis matches. The method is illustrated by analyzing a long match played at the 2003 Australian Open and investigates whether long matches can be predicted. A Markov chain model set up in Excel (Barnett & Clarke, 2002) is used to compute the predicted outcomes. The predicted and actual match statistics for the Roddick–El Aynaoui match are compared to the other men's singles matches and some interesting findings confirm why this particular match had the foundations for a very long match.

2. Method

2.1 *Collecting the data*

There is a plethora of statistics now collected and published on tennis. However, a lack of any scientific basis for many of the tables reduce their usefulness for any serious analysis, or even as a measure of a player's ability, e.g. in a table from the official Wimbledon site, http://championships.wimbledon.org/en_GB/scores/extrastats/brk_pts_con_ms.html, that ranks players on break points converted, the top six players, with 100% conversion rate, all lost their first round match. They have top ranking simply because they converted the only break point they obtained. The rankings bear almost no relationship to a player's ability to convert a break point. Not all published statistics are as useless as this table, but because they are often averages taken over many matches and surfaces, it takes some manipulation to gain insights into a particular upcoming match.

Each week from the beginning of the year, the ATP tour web site, www.atptour.com/en/media/rankings/matchfacts.pdf, provides data on the top 200 players in the champions race. Of interest to us are the statistics on winning percentages for players on both serving and receiving. Let a_i = percentage of first serves in play for player i, b_i = percentage of points won on first serve given that first serve is in for player i, c_i = percentage of points won on second serve for player i, d_i = percentage of points won on return of first serve for player i and e_i = percentage of points won on return of second serve for player i. There are three problems associated with using these statistics as inputs to a prediction model for a particular match.

Firstly, unless the match is in the first round, the statistics will be slightly out of date. While the same statistics could be used throughout the tournament, it is also possible to update for the matches played in the tournament since the statistics were published. We use a method of updating the statistics as the tournament progresses which gives more weight to more recent matches, and so attempts to make allowance for current form. The player statistics obtained in this manner for the Roddick–El Aynaoui match are given in columns 2–6 of Table 1, along with the average statistics for the top 200 players.

TABLE 1 *ATP tour statistics for Roddick and El Aynaoui*

Player (i)	a_i (%)	b_i (%)	c_i (%)	d_i (%)	e_i (%)	f_i (%)	g_i (%)
Roddick (1)	62.2	80.7	55.7	29.5	48.1	71.3	37.2
El Aynaoui (2)	65.2	75.2	50.9	29.5	48.9	66.7	37.5
Average (av)	58.7	69.2	49.2	28.7	49.0	61.6	38.4

Unfortunately, it is not possible to put exact standard errors on these estimates. The shortcomings of the statistics provided by the ATP have already been mentioned, and a further problem is that they do not give the total number of points on which these statistics are based. However, the estimates for both Roddick and El Aynaoui are based on over 70 matches. Since a 3 set match averages about 165 points, we can estimate that their statistics are based on about 12000 points. This gives a standard error of less than half a percentage point. The average tour statistics are based on 5794 matches, which result in an estimated standard error of less than 0.05 of a percentage point. Thus, we can say that the individual player statistics are correct to within 1 percentage point, and the overall tour averages to within 0.1 percentage point. The statistics clearly show the serving superiority of Roddick and El Aynaoui. Both players, but particularly El Aynaoui, get a higher percentage than average of first serves into play. Both players, but particularly Roddick, win a higher percentage of points on their first serve when it goes in, and both players win a higher percentage of points on their second serve. On the other hand, both players have only average returning statistics.

A second problem is that these statistics are too detailed for our purposes. We only require the percentage of points won on serve and return of serve for each player, and this requires some manipulation.

Calculating the percentage of points won on serve is quite straightforward. A player wins a point on serve by getting his first serve in and winning the point, or by missing his first serve and winning on his second serve. This results in

$$f_i = a_i b_i + (1 - a_i)c_i,$$

where f_i = percentage of points won on serve for player i.

The chance of winning a point on return of serve is calculated in a similar manner, except that the percentage of first serves in play is not taken from an individual player's statistics, but rather an average player. Thus, we use the averages for the top 200 players (as shown in Table 1) for the chances that the player's opponent gets his first serve in to play. Unfortunately, the ATP does not publish averages for all players. However, the top 200 is probably more suitable in this case as this is more indicative of the standard of opponent likely in a grand slam and we get the following result:

$$g_i = a_{\text{av}} d_i + (1 - a_{\text{av}})e_i,$$

where g_i = percentage of points won on return for player i. The subscript av denotes the ATP tour averages, so a_{av} = first serve percentage for ATP tour averages = 58.7%.

If we let $i = 1$ represent Roddick and $i = 2$ represent El Aynaoui, then the above formulas result in $f_1 = 71.3\%$, $g_1 = 37.2\%$, $f_2 = 66.7\%$, $g_2 = 37.5\%$. These are shown in columns 7 and 8 of Table 1, again along with the tour averages. The tour averages have been normalized, as clearly on average the percentage won on serve and return of serve must sum to 100%. These statistics show that while both players win slightly less than an average percentage of their opponent's serves, they win a much higher percentage of their own serves than the average player. However, Roddick is clearly the better player.

These statistics can be used as input to our model to predict the outcomes of matches between Roddick or El Aynaoui against the average player, e.g. from a spreadsheet model described in Barnett & Clarke (2002), they imply Roddick would win 93.3% of the best of 5 set matches, and El Aynaoui 79.0%. But they are not yet in a form to predict matches between these two players on a particular surface. The third problem is to combine the individual player's statistics to produce expected statistics when two players meet on a given surface.

2.2 *Combining player statistics*

While we expect a good server to win a higher proportion of serves than average, this proportion would be reduced somewhat if his opponent is a good receiver. This is a common problem in modelling sport, e.g. in cricket, what is the expected outcome when a bowler who gains a wicket every 20 runs bowls against a batsman who loses his wicket every 50 runs? For application in a cricket simulator, Dyte (1998) used a multiplicative method that compared a player's average to the overall average for estimating dismissal rates when a particular batsman faced a particular bowler. Here we have the added complication caused by the symmetry that one player's serving statistics are the complement of his opponent's receiving statistics, so the two percentages must add to 100%. For this reason an additive approach was necessary. We also have the complication that we expect all players to win a higher percentage of serves on (say) grass than other surfaces, e.g. at the 2002 Australian Open, 61.7% of points were won on service, whereas at the 2002 Wimbledon championships this rose to 63.8%. Such statistics are usually available on the official web site corresponding to the grand slam tournament.

In simple terms, we take the percentage of points a player wins on serve as the overall percentage of points won on serve for that tournament (this takes account of court surface), plus the excess by which a player's serving percentage exceeds the average (this accounts for player's serving ability), minus the excess by which the opponent's receiving percentage exceeds the average (this accounts for opponent's returning ability). A similar argument is used for percentage of points won on return of serve.

More formally, letting the subscript t denote the particular tournament averages, f_{ij} = the combined percentage of points won on serve for player i against player j, g_{ji} = the combined percentage of points won on return for player j against player i:

$$f_{ij} = f_t + (f_i - f_{\text{av}}) - (g_j - g_{\text{av}}), \tag{1}$$

$$g_{ji} = g_t + (g_j - g_{\text{av}}) - (f_i - f_{\text{av}}). \tag{2}$$

Note that formulas (1) and (2) are symmetrical. Since $f_t + g_t = 1$, it is easily shown that $f_{ij} + g_{ji} = 1$ for all i, j as required. It is also clear that averaging statistics over all possible players and opponents produces the tournament average.

Now f_{av} and g_{av} were obtained from the ATP web site. f_t and g_t were obtained from the 2002 Australian Open match statistics. Applying these formulas to the Roddick–El Aynaoui quarter-final match played at the 2003 Australian Open gives Roddick to win 72.3% of his serves and 32.0% of El Aynaoui's serves, with El Aynaoui winning 68.0% of his serves and 27.7% of Roddick's serves.

2.3 *Predictions*

Next we use the combined statistics to predict the outcomes of tennis matches using a Markov chain model set up in Excel as detailed in Barnett & Clarke (2002). We start by looking at a single game where we have two players, A and B, and player A is serving with constant probability p of winning a point.

TABLE 2 *The conditional probabilities of Roddick winning a service game against El Aynaoui from various score lines with* $p = 0.723$

		Roddick score				
		0	15	30	40	Game
El Aynaoui score	0	0.926	0.963	0.986	0.997	1
	15	0.829	0.902	0.957	0.990	1
	30	0.638	0.757	0.872	0.965	1
	40	0.330	0.456	0.630	0.872	
	Game	0	0	0		

The assumption of independence was investigated in Klaassen & Magnus (2001). They concluded that although points are not independent and identically distributed (i.i.d.), the deviations from i.i.d. are small and hence the i.i.d. assumption is justified in many applications, such as forecasting. We set up a Markov chain model of a game where the state of the game is the current game score in points (thus 40–30 is 3–2). With probability p the state changes from a, b to $a + 1, b$ and with probability $1 - p$ it changes from a, b to $a, b + 1$. Thus, if $P(a, b)$ is the probability that player A wins when the score is (a, b), we have

$$P(a, b) = pP(a + 1, b) + (1 - p)P(a, b + 1).$$

The boundary values are $P(a, b) = 1$ if $a = 4$, $b \leqslant 2$, $P(a, b) = 0$ if $b = 4$, $a \leqslant 2$. The boundary values and formula can be entered on a spreadsheet. It is easily shown that the chance of winning from deuce is

$$\frac{p^2}{p^2 + (1 - p)^2}$$

Advantage server is logically equivalent to 40–30, as in both cases, if the server wins the next point he wins the game, and if he loses the next point the score is 40–40 (deuce). A similar argument shows that 30–40 is equivalent to advantage receiver and 30–30 is equivalent to deuce.

Table 2 shows the results obtained when Roddick is serving with $p = 0.723$. The table illustrates the difficulty of breaking serve. Roddick with a 72.3% chance of winning a point on serve has a 92.6% chance of winning the game. At 0–30 and even 30–40, Roddick has at least a 60% chance of winning the game, and will even win one-third of the games from the worst position of 0–40. A similar spreadsheet can be set up for a game when El Aynaoui is serving and for a tie-breaker game. A slightly more complicated sheet can be set up for a set, where the chance of winning a game depends on who is serving and comes from the previous game sheet. Finally, a similar sheet for a match uses the chance of winning a set as calculated by the set sheet. The same procedures can also be applied for calculating the chances of reaching score lines in games, sets and matches, and mean lengths with associated standard deviation of games, sets and matches. Table 3 represents some resultant predicted statistics for the match between Roddick and El Aynaoui. The mean number of games in a set and the associated standard deviation are calculated for each player serving first in the set.

3. Results

Both players are above the ATP tour averages for percentage of points won on serve and just below the ATP tour averages for percentage of points won returning serve. When the player's statistics are combined together we find that both players are still above the tournament averages for percentage of

TABLE 3 *Predicted parameters for the Roddick–El Aynaoui match played at the* 2003 *Australian Open*

Parameter	Scoring unit	Roddick	El Aynaoui
Probability of winning	Point on serve	72.3%	68.0%
	Game on serve	92.6%	87.5%
	Tie-breaker game	57.5%	42.5%
	Tie-breaker set	63.1%	36.9%
	Advantage set	65.5%	34.5%
	Tie-breaker match	73.4%	26.6%
	Advantage match	74.2%	25.8%
Mean number of games	Tie-breaker set	10.8	10.9
	Advantage set	14.6	14.7
	Tie-breaker match	43.8	43.8
	Advantage match	45.0	45.0
Standard deviation of number of games	Tie-breaker set	1.9	1.8
	Advantage set	9.0	8.9

TABLE 4 *Chances of reaching a score line from* 6 *games all in an advantage set for the Roddick–El Aynaoui match*

Score line	Chances (%)
6–6	100.0
7–7	81.9
8–8	67.1
9–9	55.0
10–10	45.1
11–11	36.9
12–12	30.3
13–13	24.8
14–14	20.3
15–15	16.7
16–16	13.7
17–17	11.2
18–18	9.1
19–19	7.5

points won on serve and below the tournament averages for percentage of points won returning serve. From Table 3, Roddick is expected to win 72.3% of points on serve and El Aynaoui is expected to win 68.0% of points on serve. Roddick is expected to win 92.6% of games on serve and El Aynaoui 87.5%. This means that it will be difficult for either player to break serve and if the match does reach 6 games all in the advantage fifth set, there is a possibility it will go on for a long time. Table 4 gives the chances of an advantage set reaching various score lines from 6 games all. There is a 37.2% chance the set will reach 6 games all. Conditional on the set reaching 6 games all, there is a $0.926 \times 0.875 + 0.074 \times 0.125 = 81.9\%$ chance it will reach 7–7, $(0.926 \times 0.875 + 0.074 \times 0.125)^2 = 67.1\%$ chance of reaching

8–8 and so on (where 0.926 and 0.875 are the probabilities of Roddick and El Aynaoui winning games on serve, respectively).

Klaassen & Magnus (1998) show that while the chance of a player winning is dependent on $f_{ij} - f_{ji}$, the expected length of the match is highly dependent on $f_{ij} + f_{ji}$. The Roddick–El Aynaoui match stood out amongst the other men's singles matches played at the 2003 Australian Open, as this match had the highest predicted total for the combined percentages of points won on serve, given as $72.3\% + 68.0\% = 140.3\%$. The match also had the highest expected number of games for an advantage set (14.6–14.7) along with the highest standard deviation on the number of games played in an advantage set (8.9–9.0). For this reason we can conclude that if there was going to be a long fifth set played at the 2003 Australian Open men's singles, it would most likely come from the Roddick–El Aynaoui match. In the actual match both players actually served slightly better than predicted, with Roddick winning 75.8% and El Aynaoui 70.6% of serves. This total of (146.4%) was the highest total for the probabilities of points won on serve from all the men's singles matches played at the 2003 Australian Open, and easily exceeded the average of 123.2%.

4. Comparison of scoring systems

Punters or bookmakers betting on tennis need to have a clear idea of the effect of different scoring systems. The US Open plays a tie-breaker game at 6 games all in the fifth set, whereas other majors play an advantage fifth set. From Table 3, depending on who starts serving, the expected number of games (standard deviation) for the Roddick–El Aynaoui match is 10.8 (1.9) or 10.9 (1.8) for a tie-breaker set and 14.6 (9.0) or 14.7 (8.9) for an advantage. Clearly, the type of set is of paramount importance if betting on the length of a set. The large standard deviation for advantage sets shows that index betting, where the payoff depends on the difference between the expected and actual length, would be more risky for both punter and bookmaker. On the other hand, the expected length alters only marginally depending on who serves the first game, which would allow a bookmaker to set odds well before the set began. Interestingly, the effect of a tie-breaker fifth set on the length of a match is much less than on a set, since it is not certain a fifth set will be played. Playing a tie-breaker set also reduces slightly the favourite's chances of winning. In this case Roddick has a 74.2% chance of winning the 5 set advantage match, compared to 73.4% if the tie-breaker is applied at 6 games all in the fifth set. However, this small difference magnifies as the match progresses. From 2 sets all going in to the final set, Roddick had a 65.5% chance of winning the match, compared to 63.1% if a tie-breaker set is played. From 6 games all in the final set, Roddick has a 64.0% chance of winning the match compared to only 57.5% if a tie-breaker game is played. The very small virtually negligible advantage to the better player at the start of the match gradually increases the nearer the state of the match approaches 6 all in the final set. At the start of the match there is a trade-off between an extra 0.8% chance of winning versus an expected 1.2 games. By the start of the fifth set it is 2.4% versus 3.8 games. At 6–6 in the fifth set the trade-off is between 6.5% versus 10.1 games. A punter betting as the game progresses would need to understand such subtleties.

5. Conclusions

It is possible to manipulate published player statistics so they can be used to predict head to head matches. We have demonstrated this by using a long match at the 2003 Australian Open. Whenever two players with dominant serves but relatively poor returns of serve meet, there is always a chance that if the match reaches a fifth set, it can go on for a long period of time. This was precisely the scenario

for the Roddick–El Aynaoui match. Furthermore, we have shown from pre-match predictions that this match was likely to go longer than any other men's singles match played at the 2003 Australian Open.

It is well known that some players favour and perform better on particular surfaces. Serve and volleyers do relatively better on grass, while baseliners usually prefer clay. While the method outlined here allows for players generally winning more serves at (say) Wimbledon than the French Open, it does not allow for particular player preference. Unfortunately, the ATP player statistics do not differentiate between the surfaces, which would reflect how different players perform on different surfaces. However, the International Tennis Federation web site keeps a database on players win/loss records partitioned into the four different playing surfaces (grass, hard court, clay, carpet). Further work could involve altering the ATP player statistics to reflect how a player is likely to perform on a particular surface, e.g. a player recording their best results on grass would gain an increase in their service percentage statistics when playing at Wimbledon.

While we have used an interesting long match as an illustration, the methods outlined here can be applied to a match between any two players. Moreover, the Excel spreadsheet used is easily adapted to more complicated models. While we have assumed a constant probability for the server throughout the remainder of the match, this could easily be altered to depend, say, on the point score in a game, or game score in a set. A punter might do this to reflect known player behaviour, or a bookmaker to reflect the opinion of the betting public. As the number and type of bets on sport continue to grow, the use of sophisticated mathematical models to assist punters and bookmakers will become more common. Hopefully the collection and publication of player statistics will also become more sophisticated and better support the use of such models.

REFERENCES

BARNETT, T. & CLARKE, S. R. (2002) Using Microsoft Excel to model a tennis match. *6th Conference on Mathematics and Computers in Sport* (G. Cohen ed.). Queensland, Australia: Bond University, pp. 63–68.

DYTE, D. (1998) Constructing plausible test cricket simulation using available real world data. *4th Conference on Mathematics and Computers in Sport* (N. de Mestre ed.). Queensland, Australia: Bond University, pp. 153–159.

KLAASSEN, F. J. G. M. & MAGNUS, J. R. (1998) Forecasting the winner of a tennis match. *Eur. J. Oper. Res.*, **148**, 257–267.

KLAASSEN, F. J. G. M. & MAGNUS, J. R. (2001) Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *J. Am. Stat. Assoc.*, **96**, 500–509.

POLLARD, G. & NOBLE, K. (2002) The characteristics of some new scoring systems in tennis. *6th Conference on Mathematics and Computers in Sport* (G. Cohen ed.). Queensland, Australia: Bond University, pp. 221–226.