Predicting a tennis match in progress for sports multimedia

Article *in* OR Insight · June 2011

DOI: 10.1057/ori.2011.7 · Source: DBLP

Anthony Bedford
RMIT University

Original Article

Predicting a tennis match in progress for sports multimedia

Tristan Barnett[a,*], Darren O'Shaughnessy[b] and Anthony Bedford[c]

[a]School of Mathematics and Statistics, University of South Australia, 1/11 Findon Street, Hawthorn, VIC 3122, Australia.
E-mail: strategicgames@hotmail.com
[b]Ranking Software, 3 Wootoona Court, Mulgrave 3170, Australia.
[c]School of Mathematical and Geospatial Sciences, RMIT University, PO BOX 71, Plenty Rd, Bundoora VIC 3083, Australia.

*Corresponding author.

Abstract This article demonstrates how spreadsheets can generate the probability of winning a tennis match conditional on the state of the match. Previous models treat games, sets and matches independently. We show how a series of interconnected sheets can be used to repeat these results. The sheets are used in multimedia to predict outcomes for a match in progress, where it is shown how these predictions could benefit the spectator, punter, player and commentator. The development of the predictions could also form an interesting and useful teaching example, and allow students to investigate the properties of tennis scoring systems.
OR Insight (2011) **24,** 190–204. doi:10.1057/ori.2011.7; published online 1 June 2011

Keywords: Markov processes; sports; probability; forecasting

Received 7 December 2010; accepted 20 April 2011

Introduction

Profiting from sports betting is an obvious application of predicting outcomes in sport. Although sports betting was originally restricted to betting before the start of the match, it is now possible and common to be betting throughout a match in progress. However, the appeal of predictions throughout a match may not just be restricted to punters. In-play sports predictions could be used in sports multimedia, and hence could be appealing to the spectator, coach or technology buff without involving actual betting.

According to Wikipedia: 'Multimedia is media and content that includes a combination of text, audio, still images, animation, video, and interactivity content forms. Multimedia is usually recorded and played, displayed or accessed by information content processing devices, such as computerized and electronic devices'. Cadability Pty. Ltd. is a sports multimedia organization that specializes in delivering online information during live matches. As well as displaying the scoreboard, live predictions are in operation for tennis, cricket, Australian Rules football and soccer. This information is available on a standard PC and iPhone (as an iPhone app), with current developments for delivering the information on an iPad and TV (in the form of a widget).

There are many ways the predictions through multimedia could be used. Spectators could engage with live predictions through multimedia for entertainment when watching a live match. If a spectator was to place a bet, then live predictions through multimedia could be used as a decision support tool as to when and how much to bet on a particular event, and hence the combination of betting and multimedia becomes a powerful form of entertainment.

Another interesting application of sports predictions in multimedia is in teaching mathematical concepts, which students often relate to sporting events and hence may be stimulated in learning mathematics through an activity of personal interest. In tennis, for example, the chance of winning from deuce can be calculated by the sum of an infinite series. Another application is in using the predictions as a coaching tool. For example, it is common for players and coaches to watch a replayed match to discuss strategies for upcoming matches. The graphical and visual aspects of the predictions could enhance the TV replay. A further application is in using the predictions for TV commentary. Klaassen and Magnus (2003) demonstrate how a time series plot of the probability of winning a tennis match in progress can be useful for TV commentary by supporting his/her discussion on the likely winner of the match. Commentators could also use the graph to evaluate the match after completion to identify turning points and key shifts in momentum.

In this article, a tennis prediction model during a match in progress is constructed. Data analysis is then carried out to determine the parameters for the model. Following on, a Markov chain model is developed and a rule presented for updating prior estimates of what has occurred during the match. Finally, the methodology is applied to the men's 2010 US Open tennis final as an illustration of the value of live sports predictions through multimedia.

Data Analysis

By assigning two parameters, the constant probabilities of player A and player B winning a point on serve, the probability of winning the match can be determined using a Markov Chain model as described later. Beforehand we derive the probabilities of winning on serve when two players meet on a particular surface. This is achieved by collecting, combining and updating player serving and receiving statistics.

Collecting player statistics

OnCourt (www.oncourt.info) is a software package for all tennis fans, containing match results for men's and woman's tennis, along with statistical information about players, tournaments or histories of the head-to-head matches between two players. Match statistics can be obtained for the majority of Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA) matches going back to 2003. The ATP is the governing body of men's professional tennis for the allocation of player rating points in matches to determine overall rankings and seedings for tournaments. The WTA is used similarly in women's professional tennis. Table 1 gives the match statistics broadcast from the US Open 2010 men's final where Rafael Nadal defeated Novak Djokovic in four sets. Notice that the Serving Points Won is not given directly in the table. This statistic can be derived from the Receiving Points Won such that Serving Points Won for Nadal and Djokovic are $1-36/112=67.9$ per cent and $1-60/143=58.0$ per cent, respectively. Alternatively, the Serving Points Won can be obtained from a combination of the first Serve per cent, Winning per cent on first Serve and Winning per cent on second Serve such that Serving Points Won for Nadal and Djokovic are $75/112\times 55/75+(1-75/112)\times 21/37=67.9$ per cent and $95/143\times 61/95+(1-95/143)\times 22/$

Table 1: Match statistics for the men's 2010 US Open final between Rafael Nadal and Novak Djokovic

	Rafael Nadal	*Novak Djokovic*
First serve (%)	75 of 112=67	95 of 143=66
Aces	8	5
Double faults	2	4
Unforced errors	31	47
Winning per cent on first serve	55 of 75=73	61 of 95=64
Winning per cent on second serve	21 of 37=56	22 of 48=46
Winners (including service)	49	45
Break point conversions (%)	6 of 26=23	3 of 4=75
Receiving points won (%)	60 of 143=42	36 of 112=32
Net approaches (%)	16 of 20=80	28 of 45=62
Total points won	136	119
Fastest serve	212 KPH	201 KPH
Average first serve speed	186 KPH	188 KPH
Average second serve speed	141 KPH	151 KPH

Source: http://2010.usopen.org/en_US/scores/stats/day21/1701ms.html

$48 = 58.0$ per cent, respectively. Note that the Winning per cent on first Serve is conditional on the first Serve going in whereas the Winning per cent on the second Serve is unconditional on the second Serve going in. These calculations could be used as a teaching exercise in interpreting and analyzing data, and in conditional probabilities. Many more calculations can be obtained from broadcast match statistics as outlined in Bedford *et al* (2010).

Combining player statistics

Combining player statistics is a common challenge in sport. Although we would 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. Using the method developed by Barnett and Clarke (2005) we can calculate the percentage of points won on serve when player i meets player j on surface s (f_{ijs}) as follows:

$$f_{ijs} = f_{is} - g_{js} + g_{avs} \tag{1}$$

where f_{is} is the percentage of points won on serve for player i on surface s; g_{is} is the percentage of points won on return of serve for player i on surface s; g_{avs} represents the average (across all ATP/WTA players) percentage of points won on return of serve on surface s.

The surfaces are defined as: $s = 1$ for grass, $s = 2$ for carpet, $s = 3$ for hard and $s = 4$ for clay.

The average percentage of points won on return of serve across all players on each of six different surfaces (grass, hard, indoor hard, clay, carpet and acrylic) was calculated from OnCourt and represented in Table 2. Note that the serving averages for carpet and indoor hard are approximately the same and are therefore combined as the one surface. Similarly, hard and acrylic are combined as the one surface.

For example, suppose male player i with $f_{i1} = 0.7$ and $g_{i1} = 0.4$ meets male player j with $f_{j1} = 0.68$ and $g_{j1} = 0.35$ on a grass court surface. Then the estimated percentage of points won on serve for player i and player j are given

Table 2: The average probabilities of points won on return of serve for men's and women's tennis

Surface	*Men*	*Women*
Grass	0.347	0.420
Carpet–I.hard	0.358	0.430
Hard–Acrylic	0.375	0.448
Clay	0.400	0.464

by $f_{ij1} = 0.7 - 0.35 + 0.347 = 69.7$ per cent and $f_{ji1} = 0.68 - 0.4 + 0.347 = 62.7$ per cent, respectively.

Updating player statistics

The general form for updating the rating of a player as given by Clarke (1994) is

$$\begin{aligned}&\text{New rating} = \text{Old rating}\\&\quad + \alpha[\text{actual margin} - \text{predicted margin}]\\&\text{for some } \alpha.\end{aligned}$$

Using serving and receiving player statistics as ratings, we get

$$f_{is}^{\mathrm{n}} = f_{is}^{\mathrm{o}} + \alpha_s[f_{is}^{\mathrm{a}} - f_{ijs}] \tag{2}$$

$$g_{is}^{\mathrm{n}} = g_{is}^{\mathrm{o}} + \alpha_s[g_{is}^{\mathrm{a}} - g_{ijs}] \tag{3}$$

where f_{is}^{n}, f_{is}^{o} and f_{is}^{a} represent the new, old and actual percentage of points won on serve for player i on surface s; g_{is}^{n}, g_{is}^{o} and g_{is}^{a} represent the new, old and actual percentage of points won on return of serve for player i on surface s; α_s is the weighting parameter for surface s.

Using the method developed by Dowe *et al* (1996) to properly reward predictions reveals that $\alpha_{\mathrm{s}} = 0.05$ is a suitable weighting parameter for all surfaces. Further, every player is initialized with surface averages as given in Table 2.

Equations (2) and (3) treat each surface independently. A more advanced approach is to update the serving and receiving statistics for each surface when playing on a particular surface. For example, if a match is played on grass, then how are the other surfaces of clay, carpet and hard court updated on the basis of the player's performances on the grass?

This more complicated approach is given as follows:

$$f_{ist}^{\mathrm{n}} = f_{is}^{\mathrm{o}} + \alpha_{st}[f_{is}^{\mathrm{a}} - f_{ijs}] \tag{4}$$

$$g_{ist}^{\mathrm{n}} = g_{is}^{\mathrm{o}} + \alpha_{st}[g_{is}^{\mathrm{a}} - g_{ijs}] \tag{5}$$

where f_{ist}^{n} represents the new expected percentage of points won on serve for player i on surface t when the actual match is played on surface s; g_{ist}^{n} represents the new expected percentage of points won on return of serve for player i on surface t when the actual match is played on surface s; α_{st} is the weighting parameter for surface t when the actual match is played on surface s.

We do not propose to estimate these α_{st} parameters in this article.

Markov Chain Model

The basic principles involved in modelling a tennis match are well-known, and a Markov chain model with a constant probability of winning a point was set up by Schutz (1970). Although such a model is acceptable within a game, a model that allows a player a different probability of winning depending on whether they are serving or receiving is essential for tennis. Statistics of interest are usually the chance of each player winning, and the expected length of the match. Croucher (1986) looks at the conditional probabilities for each player winning a single game from any position. Pollard (1983) uses a more analytic approach to calculate the probability for each player winning a game or set along with the expected number of points or games to be played with their corresponding variance.

Most of the previous work uses analytical methods, and treats each scoring unit independently. This results in limited tables of statistics. Thus, the chance of winning a game and the expected number of points remaining in the game is calculated at the various scores within a game. The chance of winning a set and the expected number of games remaining in the set is calculated only after a completed game and would not show, for example, the probability of a player's chance of winning from three games to two, 15–30.

This article discusses the use of spreadsheets to repeat these applications using a set of interrelated spreadsheets. This allows any probabilities to be entered and the resultant statistics automatically calculated or tabulated. In addition, more complicated workbooks can be set up, which allow the calculation of the chance of winning a match at any stage of the match given by the point, game and set score. These allow the dynamic updating of player's chances as a match progresses. Alternatively, these algorithms could be converted into a programming language for automatic integration into multimedia as live scores are received.

Game

We explain the method by first looking at a single game where we have two players, A and B, and player A has a constant probability p_A of winning a point on serve. 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_A, the state changes from a, b to $a+1$, b and with probability $q_A = 1 - p_A$, it changes from a, b to a, $b+1$. Thus, if $P_A^{pg}(a, b)$ is the probability that player A wins the game when the score is (a, b), we have:

$$P_A^{pg}(a, b) = p_A P_A(a+1, b) + q_A P_A(a, b+1)$$

The boundary values are

$$P_A^{pg}(a, b) = 1 \text{ if } a = 4,\ b \leqslant 2,\ P_A^{pg}(a, b) = 0 \text{ if } b = 4,\ a \leqslant 2.$$

The boundary values and formula can be entered on a simple spreadsheet. The problem of deuce can be handled in two ways. As deuce is logically equivalent to 30–30, a formula for this can be entered in the deuce cell. This creates a circular cell reference, but the iterative function of Excel can be turned on, and Excel will iterate to a solution. In preference, an explicit formula is obtained by recognizing that the chance of winning from deuce is in the form of a geometric series

$$P_A^{pg}(3,3) = p_A^2 + p_A^2 2p_A q_A + p_A^2(2p_A q_A)^2 + p_A^2(2p_A q_A)^3 + \cdots\cdots\cdots$$

where the first term is p_A^2 and the common ratio is $2p_A q_A$.

The sum is given by $p_A^2/(1-2p_A q_A)$ provided that $-1 < 2p_A q_A < 1$. We know that $0 < 2p_A q_A < 1$, as $p_A > 0$, $q_A > 0$ and $1-2p_A q_A = p_A^2 + q_A^2 > 0$.

Therefore, the probability of winning from deuce is $p_A^2/(1-2p_A q_A)$. As $p_A + q_A = 1$, this can be expressed as follows:

$$P_A^{pg}(3,3) = p_A^2/(p_A^2 + q_A^2)$$

Excel spreadsheet code to obtain the conditional probabilities of player A winning a game on serve is as follows:

Enter $\mathbf{p_A}$ in cell D1
Enter $\mathbf{q_A}$ in cell D2
Enter **0.60** in cell E1
Enter=**1−E1** in cell E2
Enter **1** in cells C11, D11 and E11
Enter **0** in cells G7, G8 and G9
Enter=**E1^2/(E1^2+E2^2)** in cell F10
Enter=**E1*C8+E2*D7** in cell C7
Copy and Paste cell **C7** in cells D7, E7, F7, C8, D8, E8, F8, C9, D9, E9, F9, C10, D10 and E10

Table 3: The conditional probabilities of A winning the game from various score lines

		B score				
		0	15	30	40	Game
A score	0	0.74	0.58	0.37	0.15	0
	15	0.84	0.71	0.52	0.25	0
	30	0.93	0.85	0.69	0.42	0
	40	0.98	0.95	0.88	0.69	
	Game	1	1	1		

Notice the absolute and relative referencing used in the formula = \$E\$1*C8 + \$E\$2*D7. By setting up an equation in this recursive format, the remaining conditional probabilities can easily and quickly be obtained by copying and pasting.

Table 3 represents the conditional probabilities of player A winning the game from various score lines for $p_A = 0.60$. It indicates that a player with a 60 per cent chance of winning a point has a 74 per cent chance of winning the game. Note that as advantage server is logically equivalent to 40–30, and advantage receiver is logically equivalent to 30–40, the required statistics can be found from these cells. Also worth noting is that the chances of winning from deuce and 30–30 are the same.

Similar equations can be developed for when player B is serving such that p_A and p_B represent constant probabilities of player A and player B winning a point on their respective serves. Also $P_A^{pg}(a, b)$ and $P_B^{pg}(a, b)$ represent the conditional probabilities of player A winning a game from point score (a, b) for player A and B serving in the game, respectively.

Tiebreak game

To extend the preceding results to so-called tie-break games and beyond, we introduce some additional notation – after first noting that:

A tennis match consists of four levels – points, games, sets, match. Games can be standard games (as above) or tiebreak games, sets can be advantage or tiebreak, and matches can be the best-of-5 sets or the best-of-3 sets. To win a set a player needs six games with at least a two game lead. If the score reaches 6 games-all, then a tiebreak game is played in a tiebreak set to determine the winner of the set, otherwise standard games continue indefinitely until a player is two games ahead and wins the set. This latter scoring structure is known as an advantage set and is used as the deciding set in the Australian Open, French Open and Wimbledon. In some circumstances we may be referring to points in a standard or tiebreak game and other circumstances points in a tiebreak or advantage set. It becomes necessary to represent:

points in a game as pg,
points in a tiebreak game as pg^T,
points in an advantage set as ps,
points in a tiebreak set as ps^T,
points in a best-of-5 set match (advantage fifth set) as pm,
points in a best-of-5 set match (all tiebreak sets) as pm^T,
games in an advantage set as gs,
games in a tiebreak set as gs^T,
sets in a best-of-5 set match (advantage fifth set) as sm, and
sets in a best-of-5 set match (all tiebreak sets) as sm^T.

As the chance of a player winning a tiebreak game depends on who is serving, two interconnected sheets are required, one for when player A is serving and one for when player B is serving. The equations that follow for modelling a tiebreak game, set and match are those for player A serving in the game. Similar formulae can be produced for player B serving in the game.

Let $P_A^{pgT}(a, b)$ and $P_B^{pgT}(a, b)$ represent the conditional probabilities of player A winning a tiebreak game from point score (a, b) for player A and player B serving in the game, respectively.

Recurrence formulae:

$$P_A^{pgT}(a, b) = p_A P_B^{pgT}(a+1, b) + q_A P_B^{pgT}(a, b+1),\quad \text{if } (a+b) \text{ is even}$$

$$p_A^{pgT}(a, b) = p_A P_A^{pgT}(a+1, b) + q_A P_A^{pgT}(a, b+1),\quad \text{if } (a+b) \text{ is odd}$$

Boundary values:

$$P_A^{pgT}(a, b) = 1 \text{ if } a = 7,\ 0 \leqslant b \leqslant 5$$

$$P_A^{pgT}(a, b) = 0 \text{ if } b = 7,\ 0 \leqslant a \leqslant 5$$

$$P_A^{pgT}(6, 6) = p_A q_B / (p_A q_B + q_A p_B)$$

where $q_B = 1 - p_B$.

Table 4 represents the conditional probabilities of player A winning the tiebreak game from various score lines for $p_A = 0.62$ and $p_B = 0.60$, and player A serving. Table 5 is represented similarly with player B serving.

Table 4: The conditional probabilities of player A winning the tiebreak game from various score lines for p_A=0.62 and p_B=0.60, and player A serving

		B score							
		0	1	2	3	4	5	6	7
A score	0	0.53	0.44	0.29	0.20	0.10	0.04	0.01	0
	1	0.67	0.53	0.43	0.27	0.17	0.07	0.02	0
	2	0.76	0.68	0.53	0.42	0.24	0.13	0.03	0
	3	0.87	0.77	0.69	0.53	0.40	0.20	0.08	0
	4	0.93	0.89	0.80	0.72	0.52	0.37	0.13	0
	5	0.98	0.95	0.92	0.83	0.75	0.52	0.32	0
	6	0.99	0.99	0.98	0.96	0.89	0.82	0.52	
	7	1	1	1	1	1	1		

Table 5: The conditional probabilities of player A winning the tiebreak game from various score lines for p_A=0.62 and p_B=0.60, and player B serving

		B score							
		0	1	2	3	4	5	6	7
A score	0	0.53	0.39	0.29	0.17	0.10	0.03	0.01	0
	1	0.62	0.53	0.37	0.27	0.14	0.07	0.01	0
	2	0.76	0.63	0.53	0.35	0.24	0.10	0.03	0
	3	0.83	0.77	0.63	0.53	0.33	0.20	0.05	0
	4	0.93	0.86	0.80	0.65	0.52	0.29	0.13	0
	5	0.97	0.95	0.89	0.83	0.67	0.52	0.21	0
	6	0.99	0.99	0.98	0.93	0.89	0.71	0.52	
	7	1	1	1	1	1	1		

Note how the calculations are obtained by the interconnection of both sheets. For example,

$$\begin{aligned} P_A^{pgT}(0,\,0) &= p_A P_B^{pgT}(1,0) + q_A P_B^{pgT}(0,\,1) \\ &= 0.62 \times 0.62 + 0.38 \times 0.39 = 0.53 \end{aligned}$$

Tiebreak set

Formulae are now given for a tiebreak set. Similar formulae can be obtained for an advantage set.

Let $P_A^{gsT}(c,\,d)$ and $P_B^{gsT}(c,\,d)$ represent the conditional probabilities of player A winning a tiebreak set from game score (c, d) for player A and player B serving in the game, respectively.

Recurrence formula:

$$P_A^{gsT}(c,\,d) = P_A^{pg}(0,\,0)P_B^{gsT}(c+1,\,d) + [1 - P_A^{pg}(0,\,0)]P_B^{gsT}(c,\,d+1)$$

Boundary Values:

$$P_A^{gsT}(c, d) = 1 \quad \text{if } c = 6,\ 0 \leqslant d \leqslant 4 \text{ or } c = 7,\ d = 5$$

$$P_A^{gsT}(c, d) = 0 \quad \text{if } d = 6,\ 0 \leqslant c \leqslant 4 \text{ or } c = 5,\ d = 7$$

$$P_A^{gsT}(6, 6) = P_A^{pgT}(0, 0)$$

Notice how the cell $P_A^{pg}(0, 0)$, which represents the probability of winning a game, is used in the recurrence formula for a tiebreak set. Using the formulae given for a game and a tiebreak game conditional on the point score, and a tiebreak set conditional on the game score, calculations are now obtained for a tiebreak set conditional on both the point and game score as follows.

Let $P_A^{psT}(a, b{:}c, d)$ represent the probability of player A winning a tiebreak set from (c, d) in games, (a, b) in points and player A serving in the set. This can be calculated by:

$$\begin{aligned} P_A^{psT}(a, b : c, d) = {} & P_A^{pg}(a, b)P_B^{gsT}(c + 1, d) \\ & + [1 - P_A^{pg}(a, b)]P_B^{gsT}(c, d + 1), \quad \text{if}(c, d) \neq (6, 6) \end{aligned}$$

$$P_A^{psT}(a, b : c, d) = P_A^{pgT}(a, b), \quad \text{if } (c, d) = (6, 6)$$

Match

Formulae are now given for a best-of-5 set match, where all sets are tiebreak sets. Similar formulation can be obtained for a best-5 set match, where the deciding fifth set is advantage. Formulation can also be obtained for a best-of-3 set match.

Let $P^{smT}(e, f)$ represent the conditional probabilities of player A winning a best-of-5 set tiebreak match from set score (e, f).

Recurrence Formula:

$$\begin{aligned} P^{smT}(e,f) = {} & P_A^{gsT}(0, 0)P^{smT}(e + 1, f) \\ & + [1 - P_A^{gsT}(0, 0)]P^{smT}(e, f + 1) \end{aligned}$$

Boundary Values:

$$P^{smT}(e, f) = 1 \quad \text{if } e = 3, f \leqslant 2$$

$$P^{smT}(e, f) = 0 \quad \text{if } f = 3,\ e \leqslant 2$$

Notice how the cell $P_A^{gsT}(0, 0)$, which represents the probability of winning a tiebreak set, is used in the recurrence formula for a best-of-5 set match. Using the formulae given for a tiebreak set conditional on the point and game score, and a best-of-5 set tiebreak match conditional on the set score, calculations are obtained for a best-of-5 set tiebreak match conditional on the point, game and set score as follows.

Let $P_A^{pmT}(a, b{:}c, d{:}e, f)$ represent the probability of player A winning a tiebreak match from (e, f) in sets, (c, d) in games, (a, b) in points and player A serving in the match. This can be calculated by:

$$P_A^{pmT}(a, b : c, d : e, f) = P_A^{psT}(a, b : c, d)P^{smT}(e + 1, f) + [1 - P_A^{psT}(a, b : c, d)]P^{smT}(e, f + 1)$$

Excel spreadsheet code was given earlier to obtain the conditional probabilities of player A winning a game on serve. Building on this, spreadsheets can be developed for a game with player B serving, tiebreak game, tiebreak set conditional on the game score and a best-of-5 set tiebreak match conditional on the set score. By assigning a value for p_B to a cell (cell E3, for example), the probability of winning a match from the outset can be obtained for any probability value of p_A and p_B by changing the probability values given in cells E1 and E3. By adding additional formulae to the spreadsheet for a tiebreak set conditional on the point and game score and for a best-of-5 set tiebreak match conditional on the point, game and set score, the chances of player's winning the set and match can be obtained conditional on who is currently serving, point score, game score and set score. An interactive tennis calculator to reflect this methodology is available at www.strategicgames.com.au.

Updating Rule for Serving Statistic Estimation

Although prior estimates of points won on serve may be reliable for the first few games or even the first set, it would be useful to update the prior estimates with what has actually occurred. We will use an updating system of the form where the proportion of initial serving statistics (X) is combined with actual serving statistics (Y) to give updated serving statistics (Z) at any point within the match.

$$Z = [ac/(ac + b)]X + [b/(ac + b)]Y \tag{6}$$

where a represents the expected number of games remaining in the match; b represents the number of games played; c is a constant.

Experimental results reveal that $c = 2.5$ is a suitable constant for best-of-5 and best-of-3 set matches. Note that the updating process occurs after each point. This method is outlined in Carlin and Louis (2000) in relation to Bayesian analysis and applied to tennis in Barnett (2006).

By assigning these values to a, b the following important properties are met:

1. more weighting on initial estimates towards the start of the match;
2. the weighting increases for the actual statistics as the match progresses;
3. the weighting towards the end of the match is asymptotic to the actual match statistics.

On the basis of earlier formulae for the chance of winning, the expected number of points remaining in a game, the expected number of games remaining in the set and the expected number of sets remaining in the match could be developed. From these, the value of parameter 'a' required in equation (6) could then be determined.

Sports Multimedia

The latter results are now used to demonstrate how tennis predictions can be generated and made available during a live match (www.sportsflash.com.au/). Two feature prediction products have been devised – Crystal Ball and Looking Glass. The Crystal Ball provides the chances of winning the match in progress in the form of a pie chart. The Looking Glass (similar format to a stock market chart) plots the chances of winning the match on a game-by-game basis (as in tennis) or every 1 min time interval (as in soccer). The graphical, visual and interactive properties of the Crystal Ball and Looking Glass could encourage spectators to engage with the predictions throughout a match in progress.

Figure 1 represents the predictions through the Looking Glass for the US Open men's final between Nadal and Djokovic. From the outset, Djokovic had a 68 per cent chance of winning the match. After Nadal won the first Set 6–4, the chances of Djokovic to win the match decreased to 36.7 per cent. Djokovic won the second Set 7–5 and the chances to win the match increased to 61 per cent. This value is represented in Figure 1 by using the interactive mouse-over feature, where solid dots are given for breaks of serve.

Conclusions

On the basis of a Markov chain formulation, this article demonstrates how spreadsheets can efficiently generate the probability of winning a tennis

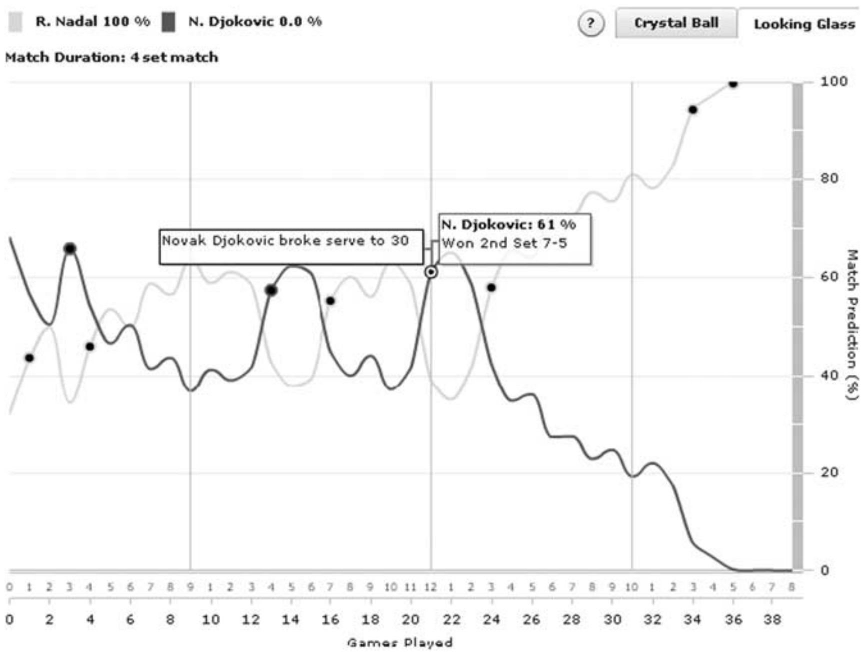

Figure 1: Live predictions for the US Open men's final between Rafael Nadal and Novak Djokovic.

match conditional on the state of the match. In addition, we show how player performance can be routinely predicted for competitors across a range of standard playing surfaces. Following on we present details of a method for iteratively revising player ratings based on a simple linear model where prior estimates are combined with actual match statistics to give updated probabilities of points won on serve. The resultant spreadsheets have been used in multimedia to predict outcomes for a match in progress. There are many applications as to how these predictions could be used. An obvious application is in sports betting and the live predictions could provide a decision support tool to the punter. Another application is in using the predictions as a coaching tool for when players and coaches discuss strategies for upcoming matches on a replayed match. A further application is in using the live predictions for TV commentary by supporting the commentators' discussion on the likely winner of the match. The development of the predictions could also form an interesting and useful teaching example, and allow students to investigate the properties of tennis scoring systems.

About the Authors

Tristan Barnett is an adjunct research fellow in operations research at the University of South Australia. His research interests include sports performance modelling, casino mathematics and dispute resolution. He is also managing director for Strategic Games, a website developed for research in the mathematics of sport, gambling and conflicts (www.strategicgames.com.au).

Darren O'Shaughnessy is an IT consultant to national sporting organizations and clubs in a variety of sports under the brand Ranking Software. His interests include mathematical modelling and simulation of sport.

Anthony Bedford is deputy head, learning and teaching at the School of Mathematical and Geospatial Sciences, RMIT University, Australia. He is also chair of MathSport Australasia and head of the RMIT Sports Statistics Research Group. He has published over 50 papers on simulation in sport, modelling in sport and performance analysis.

References

Barnett, T. (2006) Mathematical modelling in hierarchical games with specific reference to tennis. PhD thesis, Swinburne University of Technology, Melbourne.

Barnett, T. and Clarke, S.R. (2005) Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics* 16(2): 113–120.

Bedford, A., Barnett, T., Pollard, G.H. and Pollard, G.N. (2010) How the interpretation of match statistics affects player performance. *Journal of Medicine and Science in Tennis* 15(2): 23–27.

Carlin, B.P. and Louis, T.A. (2000) *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn., London: Chapman & Hall/CRC.

Clarke, S.R. (1994) An adjustive rating system for tennis and squash players. In: N. de Mestre (ed.) *Proceedings of the Second Australian Conference on Mathematics and Computers in Sport*. Queensland/Gold Coast: Bond University Press, pp. 43–50.

Croucher, J.S. (1986) The conditional probability of winning games of tennis. *Research Quarterly for Exercise and Sport* 57(1): 23–26.

Dowe, D., Farr, G.E., Hurst, A.J. and Lentin, K.L. (1996) Information-theoretic football tipping. In: N. de Mestre (ed.) *Proceedings of the Third Australian Conference on Mathematics and Computers in Sport*. Queensland/Gold Coast: Bond University Press, pp. 233–242.

Klaassen, F.J.G.M. and Magnus, J.R. (2003) On the probability of winning a tennis match. *Journal of Medical Science in Tennis* 8(3): 10.

Pollard, G.H. (1983) An analysis of classical and tie-breaker tennis. *Australian Journal of Statistics* 25: 496–505.

Schutz, R.W. (1970) A mathematical model for evaluating scoring systems with specific reference to tennis. *Research Quarterly for Exercise and Sport* 41: 552–561.

