Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 27 (2011) 619–630
www.elsevier.com/locate/ijforecast

A Bradley-Terry type model for forecasting tennis match results

Ian McHale[a], Alex Morton[b,*]

[a] The University of Salford – Centre for Operational Research and Applied Statistics, Salford, Greater Manchester M5 4WT, United Kingdom
[b] SANSTAT LTD, 483 River Valley Road, #10-07, Singapore 248368, Singapore

Abstract

The paper introduces a model for forecasting match results for the top tier of men's professional tennis, the ATP tour. Employing a Bradley-Terry type model, and utilising the data available on players' past results and the surface of the contest, we predict match winners for the coming week's matches, having updated the model parameters to take the previous week's results into account. We compare the model to two logit models: one using official rankings and another using the official ranking points of the two competing players. Our model provides superior forecasts according to each of five criteria measuring the predictive performance, two of which relate to betting returns.

Keywords: Bradley-Terry model; Logit; Ranking evaluation; Sport; Betting

1. Introduction

Sports forecasting models are often used to inform debate on some wider aspect of research, rather than being the subject of the research themselves. In their most common application, as a tool for assessing the efficiency of betting markets, the published results of sports forecasting models have rarely been found to be successful to the extent of enabling positive returns to be made. This could be a consequence of the proprietary nature of a successful forecasting model, where neither an odds-setter nor a bettor with the capacity to beat the bookmaker would be eager to release his or her formula.

Attempting to win money is not the only possible use for an objective forecasting model. Such models can be used by the media (see for example Finkelstein, Graham, Morton, & Stott, 2002, or Klaasen & Magnus, 2003) to analyse the psychology of betting markets (see for example Dixon & Pope, 2004 or Graham & Stott, 2008), uncover the dynamics of a sport (see for example Dixon & Robinson, 1998, or Holder & Nevill, 1997), construct ranking systems (Macmillan & Smith, 2007), or aid in the design of tournaments (see for example Szymanski, 2003).

In this paper we concentrate on forecasting tennis match results using a Bradley-Terry type model. Previous published papers on the subject

* Corresponding author.
E-mail addresses: i.mchale@salford.ac.uk (I. McHale), sanstat1@hotmail.com (A. Morton).

0169-2070/$ - see front matter
doi:10.1016/j.ijforecast.2010.04.004

of forecasting tennis match results have employed official ATP (Association of Tennis Professionals) rankings or tournament seedings for producing match forecasts, rather than historical results. Klaasen and Magnus (2003) use a function of the ATP rankings of the two competing players to evaluate the probability of a single point being won, then use these to infer the probabilities of each player winning the match. The authors then demonstrate how they can track this probability over the course of the match. Boulier and Stekler (1999) show that the seedings of the top 16 players (with a dummy variable to represent non-seeds) are informative in predicting match outcomes in Grand Slam tennis tournaments. Clarke and Dyte (2000) also focus on Grand Slam tournaments, but use the difference in ATP ranking points between players, rather than the position, as a predictor. More recently, Corral and Prieto-Rodriguez (2010) compared the predictive powers of three types of variables: match characteristics (such as tournament and surface), player characteristics (for example, age and height) and past performance, as measured by the ATP rankings, in a probit model for the match winners in Grand Slam tournaments. The authors find, unsurprisingly, that past performance is the most important of the three types of variables in their series of fitted models.

However, the usefulness of official rankings or seedings as predictors is called into question in the study by Scheibehenne and Broder (2007), who predicted match winners to be the higher (better) ranked player, then compare these predictions with the predictions of randomly selected volunteers with no knowledge of tennis. The volunteers were able to pick winners with the same hit rate as the official ranking predictions, provided that they were familiar with at least one of the players. Meanwhile, the bookmaker's odds identified the winner almost 10% more often than the official rankings. Although this was based on the results of a single tournament, it is quite a damning indictment of the ATP rankings as useful predictors. Clarke and Dyte (2000) justify using the official rankings for prediction purposes, rather than the actual results, by saying that it is too difficult to update a results database continually. However, online resources such as www.tennis-data.co.uk now make the collection and manipulation of data straightforward.

Tennis is not the only sport to have had its officials' rankings called into question as useful predictors of future performance. McHale and Forrest (2007) find that for men's professional golf, additional forecasting power can be provided by adding recent results to a forecasting model which already uses world rankings as a predictor. In a similar study for soccer, McHale and Davies (2007) again find that there is additional information for forecasting match results in the recent results of international teams. Thus, the evidence from tennis, golf and soccer suggests that although official rankings of players and teams are useful as predictors, they do not contain all of the information which is relevant for forecasting results.

Unlike previous papers, we do not restrict our modelling approach to models which use information on official rankings. Our application of the Bradley-Terry model uses historical match results to obtain forecasts, and we show that these forecasts are more accurate, according to several criteria, than the forecasts obtained from standard models employed in the literature. In addition, our model is also used to assess the influence of the surface on match outcomes, with the results suggesting that tennis played on clay is very different to tennis on other surfaces. Lastly, unlike other previous studies, by updating the model as new results are recorded, our model provides out-of-sample forecasts which are good enough to enable positive returns from betting.

The paper is structured as follows. Section 2 presents the data and our model. Section 3 provides a comparison of our model with ranking-based models, and also disaggregates our model and identifies which information contributes to our model outperforming the official rankings based models. Section 4 considers using our model as the basis of a betting strategy, and some closing remarks are given in Section 5.

2. Data and model

We obtained match results on the top tier of men's professional tennis, the ATP tour, for nine seasons from 2000–2008, from www.tennis-data.co.uk. The details given are participants' names and ATP rankings, the match results in games and sets, the date of the match, tournament name, location, surface (hardcourt, carpet, clay or grass) and series. 'Series' relates to a tournament's importance in terms of

the ranking points and prize money available, and is divided here into Grand Slam, Masters Series, International Gold and International. These categories are further subdivided depending on the ranking points and prize money available, as is shown in the archive section of the ATP website.

The ATP rankings are derived from the results of tournaments from the previous 52 weeks of competition. Points are awarded based on the prestige of the tournament and how far through the tournament a player progresses. Points are not weighted according to when in the preceding year a tournament took place.

The tennis-data website also contains records of the bookmakers' odds for each game, which we utilise in a measure of forecasting performance. The data consist of the closing odds from up to six bookmakers per match, and are quoted as the decimal odds (as opposed to fractional odds) of each player winning the match. If the decimal odds are o, then this is the amount paid out for a winning bet with a unit stake, indicating a profit of $o - 1$. The implied probability associated with odds of o is simply $1/o$. Thus, if the two players were equally likely to win, then the fair odds would be $o_1 = 2.00$ and $o_2 = 2.00$. However, a bookmaker would be more likely to quote $o_1 = 1.85$ and $o_2 = 1.85$, in which case the implied probabilities would sum to approximately 1.08. It is this "overround" (in the above example the overround is 7.5%) which allows the bookmaker to make an overall profit regardless of the match outcome, with their expected return being larger, the larger the overround (Graham & Stott, 2008). Scaling the implied probabilities for each match so that they sum to one provides a benchmark against which we compare the model forecast probabilities.

2.1. The model

The model employed here is based on the Bradley-Terry model (Bradley & Terry, 1952), which is a popular model for handling data on paired comparisons. It has been applied to fields as diverse as citation patterns in statistical journals (Stigler, 1994) and the factors which contribute to aggression and duelling ability in chameleons (Stuart-Fox, Firth, Moussalli, & Whiting, 2006). Applied to tennis, we can assume that the probability of a victory for player i over player j is $\alpha_i / (\alpha_i + \alpha_j)$, where α_i and α_j are positive-valued parameters representing each player's ability. This model has previously been applied to tennis results by Glickman (1999), who shows how the player strengths can be updated for new results, without re-estimating them all, by maximising the likelihood. Glickman uses match results to produce an up-to-date men's tennis ranking for the end of the 1995 season, which is very similar to the official ATP ranking, but he does not test its forecasting ability. He also shows that his updating method is equivalent to the popular ELO rating system for a particular choice of a given constant within the ELO updating algorithm. Since we find that the likelihood can be maximised quickly, we do not require his method for the approximate updating of the parameters.

A further advantage of this model over previous models is that we have information on not only the winner of the match, but also the number of games won by each player. Since a player who loses a match 6–0, 6–7, 6–7 has clearly performed better than a player who loses 6–0, 6–0 to the same opponent, it seems sensible to utilise this additional information. Thus, applying the Bradley-Terry model to the probability of winning games (with a tiebreak just counted as a normal game), the contribution to the likelihood of a contest where player i wins g_i games and player j wins g_j games can be written as

$$L(\alpha_i, \alpha_j) \propto \frac{\alpha_i^{g_i} \alpha_j^{g_j}}{(\alpha_i + \alpha_j)^{g_i + g_j}}. \tag{1}$$

In order to account for recent form, we follow the approach of Dixon and Coles (1997) and weight past results in the likelihood using an exponential decay function. We also make further use of this weighting methodology so that matches played on surfaces other than that of the current tournament are less important. We consider forecasting a match at time t on surface S. Let ε control the half-life of the exponential form decay function and S_k be a surface parameter which is set equal to 1 if match k was played on surface S but which may take a smaller value otherwise. ε and S_k are not estimated by the likelihood maximisation, but rather by optimising the predictive ability, as described in Section 2.3; they may depend on S themselves. In order to estimate the parameters for the prediction of matches at time t for fixed values of ε and S_k, we

maximise (with respect to the $\alpha_i\,(t, S)$s)

$$L\left(\alpha_i\,(t, S)\,; i = 1, \ldots, n\right) = \prod_{k \in A_t} \left(\frac{\alpha_i\,(t, S)^{g_i}\,\alpha_j\,(t, S)^{g_j}}{\left(\alpha_i\,(t, S) + \alpha_j\,(t, S)\right)^{g_i + g_j}} \right)^{\exp(\varepsilon(t - t_k))S_k}, \tag{2}$$

where k is an index over the matches, t_k is the time that match k was played, $A_t = \{k : t_k < t\}$, and n is the number of players in the model. Note that the dependence of α_i on the time and surface is suppressed in the notation in the remainder of the paper, since this is assumed.

This model only allows player strengths to vary through time by re-estimating the model (as we do after each week of competition). Each time Eq. (2) is estimated, the player strengths are assumed to be constant in all previous matches. Thus, this approach could be viewed as an *ad hoc* method of quantifying a player's recent form, rather than a full likelihood approach for modelling the temporal dynamics in player quality. The theoretical implications of this method are discussed in some depth by Dixon and Coles (1997). Crowder, Dixon, Ledford, and Robinson (2002) showed how the time dependence in player strength may be modelled explicitly, but their algorithm was extremely computationally intensive, and ultimately provided forecasting results for football which were no better than those of the approximate method which we adopt here.

This model could be used to test for the existence of a home advantage in cases where one of the players is playing in his own country, as is demonstrated by Firth (2005) for a home advantage in baseball. However, Holder and Nevill (1997) assert that a home advantage is not an important factor in tennis, and we do not consider this any further here.

There are two omissions from this model. First, there is no allowance for any dependence between consecutive games; and second, it does not account for the intrinsic structure of tennis, where players are far more likely to win games when serving than when receiving. We return to these issues in the discussion, but for now, we merely note that the forecasting model results in an uncomplicated, well-performing model, despite these omissions.

It is of interest to use the player abilities, α_i, to produce an alternative rankings system and to compare the new rankings with the official ATP rankings. We estimated the model parameters for the period 2000–2008 to obtain a rankings list for December 2008 using a form decay with a half-life of 240 days and weighting matches on all surfaces equally, as is done for the official ATP World Rankings. We also required all players in the model to have played at least 15 matches in the previous three seasons. The maximisation of the likelihood (Eq. (2)) is straightforward, although we imposed the constraint $\alpha_1 = 1$ to prevent over-parameterisation. Because of his constant activity throughout the time span covered by the data, we took Roger Federer as the player corresponding to $i = 1$. The model's top 15 players are shown in Table 1 for comparison with the ATP ranking at the end of 2008.

The top 5 players in the model ranking are the same as in the ATP ranking, but below this there are some large discrepancies, most notably with Lleyton Hewitt. The former World No. 1 is still ranked 9th by the model, compared to 67th by the official rankings. An analysis of his results in 2008 shows that he played very few tournaments in the latter part of the season, which would explain why he had not picked up many ranking points. This example highlights the fact that the official ranking rewards players for their frequent participation in events, rather than for their absolute performance. The model, of course, draws no such distinction, and merely weights past performances with respect to ease of win, opponent quality and how recent the match was.

From the parameter estimates in Table 1 we can deduce, for example, that, with no knowledge of the surface, at the end of 2008 the probability of Rafael Nadal winning a game against Roger Federer is $1.04/(1.04 + 1) = 0.51$. Assuming independence between games, one can evaluate the probabilities of different scores within a set, and hence the match result, depending on the number of sets in the match. For a best-of-three set match, this yields a win probability for Nadal of 0.538.

2.2. The effect of surface

Forecasting tennis is further complicated by the effect of the surface. It is well known that some players have better results on some surfaces than on others. The most celebrated example in recent years has been

Table 1
The top 15 players at the end of 2008 across all surfaces, as ranked by the model using a form decay with a half-life of 240 days, together with the corresponding ATP rankings.

Rank	Player	Model rating	Rating SE	ATP ranking
1	R. Nadal	1.04	0.07	1
2	R. Federer	1.00	0.00	2
3	N. Djokovic	0.90	0.06	3
4	A. Murray	0.86	0.06	4
5	N. Davydenko	0.82	0.05	5
6	A. Roddick	0.81	0.05	8
7	J.M. Del Potro	0.81	0.06	9
8	R. Soderling	0.80	0.06	17
9	L. Hewitt	0.79	0.07	67
10	D. Nalbandian	0.78	0.06	11
11	R. Gasquet	0.76	0.06	25
12	D. Ferrer	0.76	0.05	12
13	J.W. Tsonga	0.76	0.06	6
14	N. Kiefer	0.74	0.06	38
15	T. Berdych	0.74	0.05	20

one of the most successful players of all time, Pete Sampras, who managed 14 Grand Slam tournament victories yet failed to win the French Open, which is the only Grand Slam played on clay. A similar story is true for Roger Federer, arguably the greatest player of all time, who, as of July 2009, had won 15 Grand Slam titles, with just one being at the French Open.

To illustrate the effect of the surface, we estimate the parameters for the end of the 2008 season as in the last section, but now using past results on hardcourt only and clay only. The rankings are shown in Table 2. The top three players are the same in both lists, but there are some differences further down the rankings, with Andy Murray down at 21st on the clay court ranking and Gremelmayr outside the top 100 in the hardcourt ranking. The dominance of Rafael Nadal on clay, with only four losses in the previous four seasons, is also very much in evidence. Continuing the theme of Federer vs. Nadal, the model estimates that at the end of 2008, if the two were to meet on hardcourt, the probability of Federer winning a game is 0.526, which translates to a 0.636 probability of winning a best-of-three set match; while on clay these figures drop to 0.415 and 0.130, respectively.

It is therefore clear that the accuracy of a model for predicting player abilities on a given surface will depend on the comparative weighting of matches on the other surfaces. Experimentation has led us to conclude that player abilities on hardcourt and carpet are fairly similar, and therefore, due to the small number of games played on carpet, we group these tournaments together. This leaves three surface categories to consider: hard/carpet, clay and grass, with two values of S_k to be estimated for each surface. For example, we might take $S_k = 0.5$ for all matches not played on the same surface as the tournament in question.

It is not possible to estimate the weighting parameters by maximum likelihood, since it is clear from Eq. (2) that simply taking large values for S_k and taking $\varepsilon = 0$ will increase the likelihood (see Dixon & Coles, 1997, for further discussion), without necessarily improving the forecasting accuracy. We therefore estimate only the α_is by maximising the likelihood, whilst employing a grid search in three dimensions to estimate the decay parameter ε and the surface decay parameters S_k, which optimise the forecasting accuracy. To calculate the forecasting accuracy we evaluated all match outcome probabilities each week using player parameters estimated from the model based only on results prior to that week. We then accumulated the predictive performance over all weeks from 2001–2008. This process was repeated for each form decay/surface weighting combination. Thus, the player parameters are estimated out-of-sample prior to each tournament, whilst the form decay and surface weightings are estimated once in-sample. In Section 4, we produce pure out-of-sample forecasts, and thus the form decay and surface

Table 2
The top 10 players at the end of 2008 based on results on hardcourt only and on clay only using form decays with a half-life of 240 days.

Rank	Player	Rating: hardcourt only	Player	Rating: clay only
1	R. Federer	1.00	R. Nadal	1.41
2	N. Djokovic	0.93	R. Federer	1.00
3	R. Nadal	0.90	N. Djokovic	0.91
4	A. Murray	0.89	D. Ferrer	0.80
5	N. Davydenko	0.86	R. Stepanek	0.79
6	R. Soderling	0.85	L. Hewitt	0.79
7	D. Nalbandian	0.84	N. Davydenko	0.78
8	J.M. Del Potro	0.84	J.M. Del Potro	0.77
9	A. Roddick	0.83	D. Gremelmayr	0.75
10	J.W. Tsonga	0.78	D. Nalbandian	0.74

weighting parameters are estimated at the end of the previous season. Deciding on suitable criteria to use for ranking different weighting combinations is not straightforward, and is the subject of the next section.

2.3. Estimation of form decay and surface coefficients

We introduce four measures of predictive performance. The first two are given by

$$m_1 = \frac{1}{N}\sum_{k\in B_k} \log p_k^w \quad \text{and} \quad m_2 = \frac{1}{N}\sum_{k\in B_k} p_k^w,$$

where B_k is the set of all matches predicted for which we also have bookmaker's odds, N is the size of this set, and p_k^w is the probability assigned by the model to the winner. m_1 is analogous to the measure Dixon and Coles (1997) suggested for choosing the optimal form decay in football prediction, as it corresponds to maximising the predictive log-likelihood of the match outcome. m_2 is a simple variation which tends to be more correlated with betting returns than m_1. However, it would be possible to achieve a high score using m_2 by simply assigning unit probabilities to all favourites. An alternative measure of forecasting accuracy would be to calculate the proportion of predicted favourites who actually win, as was done by Scheibehenne and Broder (2007). However, such an approach does not test the accuracy of the predicted probabilities.

Performance against the bookmakers is often of primary interest to those constructing sports forecasting models. For example, Forrest and McHale (2007) investigate the efficiency of the betting market for tennis. Recognising the importance of such studies, we also use two predictive criteria based on betting returns. Comparing the model probabilities to the bookmaker's odds, it is straightforward to evaluate the expected return for betting on either player winning. We define m_3 as the return from betting one unit at the average bookmaker odds available when the expected return is positive. Similarly, m_4 is the return from betting one unit at the best odds available. Despite bookmaker overround, the predicted probabilities are often larger than the corresponding bookmaker implied probability, and thus this betting strategy involves wagering on the vast majority of matches. We are not suggesting that this is a sensible betting strategy in practice, but it provides a fairly comprehensive measure of performance against the bookmakers.

Table 3 presents the four measures of predictive performance for optimal values of the surface and form decay factors for each surface. For comparison we also show the predictive performance of the average bookmaker odds. m_3 and m_4 in the bookmakers' odds columns correspond to the expected return from betting randomly.

Table 3 shows that while the model outperforms the bookmakers according to measure m_2, it lags behind according to m_1. This is what would be expected from a market exhibiting a long-shot bias (Forrest & McHale, 2007): the subjective model predictions tend to give a higher probability to favourites, resulting in a higher score on m_2, but wins by outsiders drag down m_1 far more for the model than for the bookmakers' odds.

The returns using average odds are quite poor, but are considerably superior to the expected return from betting randomly. Positive returns are achievable using the best available odds, and given the huge

Table 3
Forecasting performance measures for model predictions and predictions inferred from the bookmakers' odds.

	Model				Bookmakers' odds			
	m_1	m_2	m_3	m_4	m_1	m_2	m_3	m_4
Hard/carpet	−0.620	0.598	−5.2%	1.6%	−0.590	0.585	−9.8%	−4.1%
Clay	−0.615	0.606	−5.2%	0.2%	−0.587	0.586	−10.5%	−5.1%
Grass	−0.575	0.618	−7.3%	−1.2%	−0.549	0.609	−13.0%	−7.6%

Table 4
Optimal values of the form decay half-life and S_k for different surface categories according to different measures of forecast accuracy.

	Measure	Form half-life (days)	Hardcourt/carpet weighting	Clay weighting	Grass weighting
Hardcourt/carpet	m_1	120	1	0.25	0.5
	m_2	240	1	0.01	0.01
	m_3	120	1	0.25	0.5
Clay	m_1	120	0.25	1	0.01
	m_2	240	0.01	1	0.01
	m_3	120	0.01	1	0.01
Grass	m_1	480	0.5	0.01	1
	m_2	180	0.01	0.01	1
	m_3	240	0.25	0.01	1

number of games, this small return corresponds to a considerable profit. However, we must keep in mind that the model in this section is optimised with respect to form decay and surface weighting parameters that have been estimated in-sample, and therefore this does not necessarily indicate a positive profit for out-of-sample prediction. True out-of-sample predictions are generated and discussed in Section 4.

Turning our attention to the optimal surface weightings, the grid search showed that previous results on different surfaces are less important predictors than previous results on the same surface. Evidence for this is shown in Table 4, where we show the optimal surface weightings for three of our measures of predictive accuracy. Quite often the weights on other surfaces in the optimal models are deflated to the extent that they take their minimum value of 0.01, and thus in practice the results on other surfaces can be ignored.

3. Comparison to rankings-based models

We now present two alternative forecasting models and compare them to our model. For comparison with the results in the previous section, all model parameters are updated at the end of each week in order to incorporate new results, and forecasts for the coming week's games are generated. The first year's data were used to obtain initial estimates of the parameters.

3.1. Binary logistic regression models

Unlike a soccer match, which, in general, has three possible outcomes (win, draw, loss), a tennis match has just two, win and loss, and as such, binary logistic regression models are arguably the most obvious candidate for forecasting tennis match results. We reproduce the two model specifications of Boulier and Stekler (1999) and Clarke and Dyte (2000) to be used as benchmark models for the Bradley-Terry model described above.

In their study on the effectiveness of seedings in Grand Slam tournaments, Boulier and Stekler (1999) use binary logistic regression with a probit link to regress the match outcome on the difference between the two players' seedings. We do not restrict our study to Grand Slam tournaments, and therefore, as seedings are produced for Grand Slam tournaments only, use the rankings of the players. However, we expect there to be very little difference, as there is a perfect correlation between tournament seedings and

Table 5
Forecasting performances of different models.

Model	m_1	m_2	m_3	m_4	m_5
Rankings position	−0.659	0.530	−18.1%	−10.0%	0.231
Rankings points	−0.634	0.549	−18.2%	−10.0%	0.222
Surface specific rankings points	−0.620	0.565	−17.1%	−9.2%	0.215
Bradley-Terry	−0.614	0.606	−5.4%	0.9%	0.210

Table 6
Parameter estimates for coefficients of the log ratio of rankings points for logit models for each of the difference surface combinations.

Forecasts for matches on:	Surface predictor based on matches played on:		
	Hard/carpet	Clay	Grass
Hard/carpet	0.54 (26.0)	0.07 (4.6)	0.09 (6.4)
Clay	0.26 (10.1)	0.43 (19.6)	−0.06 (−2.2)
Grass	0.41 (8.6)	0.05 (1.3)	0.30 (8.7)

Note: t-statistics are shown in parentheses.

ATP rankings for all tournaments but Wimbledon. For our data, we find that the logit link function actually provides marginally improved forecasts, and thus is employed here. Estimating the model parameter for the entire data set gives the probability of player i beating player j as

$$\begin{aligned}\text{logit}\,(p_i) &= \log\left(\frac{p_i}{1-p_i}\right)\\ &= 0.0065 \times \left(r_i - r_j\right) \quad (t\text{-stat} = 33.2)\,,\end{aligned}$$

where r_i is player i's ranking position. Unsurprisingly, this shows that the higher ranked player has a higher probability of winning, in agreement with Boulier and Stekler (1999), and the result is highly statistically significant.

We also estimate a second logit model, following the model of Clarke and Dyte (2000), and use the difference in ATP ranking points as a predictor. We estimated

$$\begin{aligned}\text{logit}\,(p_i) &= \log\left(\frac{p_i}{1-p_i}\right)\\ &= 0.62 \times \log q_{ij} \quad (t\text{-stat} = 44.5),\end{aligned}$$

where q_{ij} is the ratio of player i's ranking points to player j's. The predictive performance of the ATP ranking and the points based models across all surfaces are displayed in Table 5. For comparison with other papers, we also include a fifth measure of forecast accuracy, m_5, the Brier Score (as employed by Boulier & Stekler, 1999, for example). According to all of the criteria, these models underperform significantly relative to the Bradley-Terry model. The fact that the rankings based model fares worst is not surprising, since it assumes that the quality gap between adjacently ranked players does not depend on their ranking positions, which seems very unrealistic. Indeed, adding a quadratic term to the Boulier and Stekler model attracted statistical significance here, although the forecasting power was not significantly improved from the linear specification.

3.2. Surface specific rankings based models

One could criticise the above comparison as being somewhat unfair, since our model uses information on the surface and the ranking based models do not. We therefore evaluated the ranking points obtained by each player on each surface. This enabled us to estimate surface-specific regression models using the log-difference in ranking points gained on each surface as match predictors. The parameter estimates are shown in Table 6, and, as with our results-based model, the points accumulated on a particular surface are far more important in predicting player performance on that surface than the points accumulated on other surfaces, with the exception of predicting matches on grass. This is likely to be because there are very few tournaments played on grass and therefore few rankings points are available. Note that for a clay court match between two players with equal ranking points on both hard/carpet and clay surfaces, the

Table 7
Proportion of matches in which the higher ranked player won, for rankings derived from the official ATP rankings and five different Bradley-Terry type models.

	Rankings based on:	Proportion of matches where higher ranked player won (%)
(1)	Official ATP ranking only	64.10
(2)	Surface specific ATP rankings	65.00
Bradley-Terry model incorporating information on:	Winner only	64.30
	Winner and date played	64.70
	Match score	64.70
	Match score and date played	66.00
	Match score, date played and surface	66.90

model predicts that the player with fewer grass court rankings points would be more likely to win.

The predictive performance of this model is also summarised in Table 5. Using surface-specific predictors has significantly improved the match forecasts relative to the other ranking-based models, as measured by m_1 and m_2, but the betting return of this approach is still very poor, worse than that from betting randomly.

3.3. Disaggregating the Bradley-Terry type model's outperformance of the official rankings

In agreement with previous authors, we conclude that the official rankings do contain information which is relevant for predicting match outcomes. However, the match forecasts we generate using the official rankings are poor in comparison to forecasts generated from our Bradley-Terry type model. To investigate the source of this difference in forecasting performance, we compare the official rankings to the model rankings generated as at the end of Section 2 with no surface specific weightings and a form decay parameter corresponding to a half-life of 240 days. The official and model rankings agreed on the higher ranking player in 82.6% of matches (95% CI: 82.1%–83.2%) between 2001 and 2008. For matches where they disagreed, the player ranked higher by the model won on 55.5% of occasions (95% CI: 53.8%–57.2%), and thus the model rankings are significantly better.

The fact that the official rankings are inferior to those obtained from the model can be attributed to their making use of less information, as they: (i) take the results from the previous year only; (ii) do not use information on the dates previous matches were played, and therefore a match played eleven months ago is treated as being as important as one played the previous week; (iii) incorporate only information on who won a match, not the score; and (iv) are uniform across different surfaces.

To quantify the impact of each of these omissions on the forecasting power of the official rankings, we obtained match forecasts for various simplified versions of our Bradley-Terry model. The results are summarised in Table 7, and show that the success rate of the higher ranked player winning a match using the ATP rankings was 64.1%. Using the ranking points obtained on a particular surface to judge the rankings for a match played on that surface increased this success rate to 65%.

It was easy to modify our likelihood algorithm so that only the match winner is accounted for within the model, rather than the number of games won by each player.

Taking only the previous year's results into account restricted the number of players included in the model, and therefore we continued to use the previous three years' results. The resulting Bradley-Terry model based purely on match results over the previous three years and the date of the match produced a ranking success rate of 64.7%, slightly higher than that of the ATP rankings.

Accounting for the recent form within this model and switching to a model based on the match score each yielded a slight increase in ranking accuracy, but combining these two produced a significant improvement, with the ranking success rate climbing to 66%. This was further increased to 66.9% when the surface was accounted for.

Table 8
Returns to betting out-of-sample at the best available and average odds.

Year	Form decay half-life	Clay weighting	Grass weighting	Margin (%)	Number of bets	Betting return (average odds)	Betting return (best odds)
2004	240	0.01	0.75	41	60	−0.8	6.6
2005	240	0.01	0.75	41	107	−17.8	−4.8
2006	240	0.25	0.25	45	54	20.0	29.9
2007	240	0.01	0.75	41	123	−15.4	12.3
2008	240	0.50	0.75	44	59	0.9	10.4

We conclude from this that each of the pieces of information which is omitted when calculating the ATP rankings, i.e. surface, time and ease of win, is important for assessing player quality. It was the combination of information on recent form and match score that was most crucial in producing more accurate forecasts.

4. Application to betting

In this section we discuss whether our simple model could be applied successfully to betting on match winners on the ATP tour. We do this by adjusting our betting strategy on a season by season basis, as one would be likely to do in practice. The return from a betting strategy depends not only on the optimal predictive model through the surface weightings and form decay used, but also on the criteria used for deciding when to bet. Following Dixon and Coles (1997), we consider betting upon any match where the model's expected return is above a given threshold or margin. Thus, based on model forecasts from the 2001–03 seasons, we search for the optimal strategy, as defined by the profit from the best available odds. We then use this strategy to generate betting selections for the 2004 season, following which we update the optimal strategy for the 2005 season, and so on.

We summarise the results of this betting simulation exercise in Table 8, where we concentrate on predictions on a hard/carpet surface, since most matches are played on this surface. The optimal predictive strategies based on previous seasons are shown for each year, together with the performance of the returns from betting one unit stakes.

The betting profits using the best odds are positive, with returns above 10% in every year except 2005, despite using a very cautious strategy in terms of the number of bets. It should be mentioned that the proportion of bets which win is well below 50%, and the overall profit is heavily influenced by winning a few bets at long odds. It is these wins at long odds which cause the large difference between the profits for the average and best odds. Note that the strategy relating to the optimal surface weighting, margin and form decay parameter is fairly stable from year to year.

Of course, obtaining positive returns to betting has more profound implications than merely indicating a good forecasting model. The literature on market efficiency is well established, and many authors have used the sports betting market to test for market efficiency. One definition of market efficiency is that it should not be possible to accrue superior returns (i.e. higher than the bookmaker takeout). Here we obtain superior returns, to the extent that they are positive, suggesting a serious violation of the efficient markets hypothesis, although it should be noted that the positive returns obtained depend on shopping around as much as on the performance of the model itself.

5. Closing remarks

Previous attempts at forecasting match results in tennis have employed information on the official rankings of the players in order to infer the probability of a player winning a match. However, the use of these official rankings based models has been called into question. We specify a forecasting model based on the Bradley-Terry model for paired comparisons that does not rely on either a knowledge of the official rankings nor their efficacy. Each forecast incorporates information on the games won and lost in past matches, the surface of the current match compared with past matches, and the length of time since past match results. According to five separate measures of predictive performance, our model produces forecasts

which are superior to the ranking-based models used previously, and most notably, provides higher returns to betting using a simple betting strategy. Our findings suggest that, although the official rankings do contain information which is relevant to a forecasting model, new rankings that use information on the games won and lost, date of the match and surface, contain more information and produce better forecasts.

Tennis has the distinctive feature that it is a sport which is played in very different environments, namely on different surfaces. Our findings show that clay can almost be regarded as a separate entity in a forecasting model. For example, the rankings table we produce from our model is very different for clay, and furthermore, excluding results for other surfaces did not have a negative impact on the predictive performance of the model for any of the surfaces. This disparity between surfaces could lead one to argue that the official ranking is just a weighted average of the player's quality across different surfaces, and is not relevant when making a prediction for a match on a particular surface.

The deficiencies of the official rankings as predictors actually run deeper, because, as was shown in Section 3.3, the players' strengths, as estimated by our model, provide significantly better rankings than the official rankings, even when surface is not accounted for. The fact that the official rankings do not consider the quality of the opposition is an obvious weakness, as is the apparently arbitrary nature of the point values awarded. One defence of the official rankings is that they are not intended to be an accurate measure of player quality, but a reward for players who compete well across the season, and one which is easy for both players and fans to understand. For example, at the end of 2008 it would be difficult to argue against Serena and Venus Williams being the top two female tennis players in the world following the retirement of Justine Henin, but they were only ranked 2nd and 6th, due to the small number of competitions they entered. In such cases, the average ranking points per tournament may be of greater relevance when forecasting the player quality. Nevertheless, it seems certain that the rankings points allocated could, if desired, be engineered to better reflect ability.

Our model outperforms the models based on the official rankings discussed here. However, we should note that more data are needed for our model than for the simpler models based on the ATP rankings. Although the model performs well here, there is still scope for further work on forecasting tennis results. We have not taken into account the alternating service structure of the game. Since (male) players are far more likely to win games on their own service, we might expect the total number of games in a match, as predicted by the model, to generally be lower than the actual number. However, the opposite is the case, and the model actually over-estimates the average number of games played during matches, although the error is not large. This can be attributed to another deficiency of the model, namely that it does not account for the dependence between games. Clarke and Dyte (2000) report a similar result, as their ranking-based predictive model significantly under-estimates the number of straight set victories, and the authors suggest various reasons for this. The most appealing are that either one of the players is lacking motivation or having an off-day, or the clash of styles favours one of the players. The overall outcome of our model, not accounting for the alternating service and the lack of dependence between games, is that it estimates the competitive balance of a match quite accurately, as the two effects cancel one another out. Nevertheless, a model which could capture both of these features of tennis would ultimately be more desirable.

References

Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors? An evaluation. *International Journal of Forecasting*, *15*, 83–91.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs I: the method of paired comparisons. *Biometrika*, *39*, 324–345.

Corral, J., & Prieto-Rodriguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting*, *26*, 551–563.

Clarke, S. R., & Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, *7*, 585–594.

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English Football League matches for betting. *Statistician*, *51*, 157–168.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, *46*(2), 265–280.

Dixon, M. J., & Pope, P. F. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, *20*, 697–711.

Dixon, M. J., & Robinson, M. E. (1998). A birth process model for association football matches. *The Statistician*, *47*(2), 523–538.

Finkelstein, D., Graham, I., Morton, A., & Stott, H. P. (2002). November and ongoing. In *The fink tank: analysis from the football laboratory* (Saturday edition). London: The Times.

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software*, *12*(1), 1–12.

Forrest, D., & McHale, I. G. (2007). Anyone for tennis (betting)? *The European Journal of Finance*, *13*(8), 751–768.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, *48*(3), 377–394.

Graham, I., & Stott, H. (2008). Predicting bookmaker odds and efficiency for UK football. *Applied Economics*, *40*, 99–109.

Holder, R. L., & Nevill, A. M. (1997). Modelling performance at international tennis and golf tournaments: is there a home advantage? *The Statistician*, *46*, 551–559.

Klaasen, F., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, *148*, 257–267.

Macmillan, P., & Smith, I. (2007). Explaining international soccer rankings. *Journal of Sports Economics*, *8*(2), 202–213.

McHale, I. G., & Davies, S. M. (2007). Statistical analysis of the effectiveness of the FIFA world rankings. In R. Koning, & J. Albert (Eds.), *Statistical thinking in sport* (pp. 77–90). Chapman and Hall.

McHale, I. G., & Forrest, D. (2007). The importance of recent scores in a forecasting model for professional golf tournaments. *IMA Journal of Management Mathematics*, *16*, 131–140.

Scheibehenne, B., & Broder, A. (2007). Predicting Wimbledon 2005 tennis results by mere player name recognition. *International Journal of Forecasting*, *23*, 415–426.

Stigler, S. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science*, *9*, 94–108.

Stuart-Fox, D., Firth, D., Moussalli, A., & Whiting, M. J. (2006). Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour*, *71*, 1263–1271.

Szymanski, S. (2003). The economic design of sporting contests. *Journal of Economic Literature*, *41*, 1137–1187.

Ian McHale is a Senior Lecturer in Statistics at the University of Salford, UK. Having graduated from the University of Liverpool with a B.Sc. (Hons) in mathematical physics, Ian studied extreme value statistics to gain his Ph.D. at the University of Manchester. His current research interests include the analysis of gambling markets and statistics in sport, and he has published a series of papers in these areas. As part of his sports research, Ian was co-creator of the Actim Index, the official player ratings system of the English Premier League.

Alex Morton is founder of the statistics in sport consultancy SANSTAT. Prior to this he gained his Ph.D. in time series analysis at the University of Lancaster and was a postdoctoral researcher at the University of Warwick. During this time he co-founded "The Finktank", a popular weekly column in *The Times* newspaper presenting statistical analyses of football.