ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2019)
182, *Part* 4, *pp.* 1393–1409

The analysis and forecasting of tennis matches by using a high dimensional dynamic model

P. Gorgi,

Vrije Universiteit Amsterdam and Tinbergen Institute, Amsterdam, The Netherlands

S. J. Koopman

Vrije Universiteit Amsterdam, Tinbergen Institute, Amsterdam, The Netherlands, and Aarhus University, Denmark

and R. Lit

Vrije Universiteit Amsterdam, The Netherlands

[Received February 2018. Revised March 2019]

Summary. We propose a high dimensional dynamic model for tennis match results with time varying player-specific abilities for different court surface types. Our statistical model can be treated in a likelihood-based analysis and can handle high dimensional data sets while the number of parameters remains small. In particular, we analyse 17 years of tennis matches for a panel of over 500 players, which leads to more than 2000 dynamic strength levels. We find that time varying player-specific abilities for different court surfaces are of key importance for analysing tennis matches. We further consider several other extensions including player-specific explanatory variables and the match configurations for Grand Slam tournaments. The estimation results can be used to construct rankings of players for different court surface types. We finally show that our proposed model produces accurate forecasts. We provide evidence that our model significantly outperforms existing models in the forecasting of tennis match results.

Keywords: Association of Tennis Professionals; Bradley–Terry model; Logistic regression; Maximum likelihood; Out-of-sample analysis; Player rankings; Score-driven model; Time varying parameter

1. Introduction

Modelling and predicting the outcomes of tennis matches have attracted much attention over the last few years. Statistical models can be useful to describe the main features of tennis matches and to elicit the level of ability of tennis players in different situations; see Klaassen and Magnus (2001) and Newton and Aslam (2009). Models can be used to construct rankings and to determine entry and seeding of tennis tournaments. Statistical models can also be employed to obtain predictions of matches and tournaments and to test the efficiency of betting markets; see Klaassen and Magnus (2003). The default modelling approach to the statistical analysis of tennis matches is based on the Bradley–Terry model; Bradley and Terry (1952). Boulier and Stekler (1999) and Clarke and Dyte (2000) considered Association of Tennis Professionals (ATP) rank-

Address for correspondence: S. J. Koopman, Department of Econometrics, School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
E-mail: s.j.koopman@vu.nl

ings points to describe the level of strength of tennis players. Glickman (1999) introduced an algorithm to update the parameter estimates of the Bradley–Terry model dynamically within a Bayesian analysis. McHale and Morton (2011) used a weighted likelihood approach to account for time variation in the level of ability of the players within the Bradley–Terry model. Baker and McHale (2014, 2017) adopted a modified version of the Bradley–Terry model to determine the greatest male and female tennis player of all time; in their analyses, the strengths of players were allowed to vary over time by barycentric rational interpolation which compared favourably against spline interpolation methods.

It is widely acknowledged in the literature that time variation in the level of strength of tennis players is one of the key ingredients to describe the outcome of tennis matches properly. The strength of a player typically increases from a young age and reaches a certain peak when he or she is in his or her 20s, followed by a decline until he or she ends his or her career. However, in all studies so far, time variation has been achieved through a modification of an estimation method; it has not been modelled explicitly by means of a fully specified probability measure for the outcome of a tennis match at some time period. Given that the outcome of a match relies mainly on the abilities of the two players, we require to model the strength of each player explicitly. Furthermore, since the strength of a player can vary considerably with the court surface type, the model also needs to identify levels of strength for different surfaces. We propose a fully specified high dimensional dynamic model where the abilities of the players vary over time as stochastic processes. As far as we know, the formulation of a complete dynamic model and the likelihood-based analysis for tennis matches are innovative developments in the literature.

Modelling tennis matches is challenging in many ways. The major challenge is the parameter dimension. To allow for the individual strength of each player, we require as many coefficients as players in the data set. In addition, when we introduce time variation in the coefficients, we clearly have an intrinsically difficult problem on our hands: the vector of strength coefficients is high dimensional and it should be allowed to evolve over time. In our study we consider more than 500 players. Another challenge is to account for the different playing surfaces of tennis courts because each type of surface has its own characteristics and effects on the tennis game and the players. For example, consider two of the strongest tennis players of all times: Roger Federer and Rafael Nadal. Federer has won 20 Grand Slam tournaments but only one of them was on the clay surface of the French Open tournament. We note that the French Open is the only Grand Slam tournament that is played on clay. In contrast, Nadal has won 17 Grand Slam tournaments of which 11 were wins at the French Open tournament on clay. This basic fact strongly suggests that taking into account the ability of a tennis player on different surfaces is important for the effective modelling of tennis matches. When different strengths for different court surface types need to be specified in the model, then a multiple of strength coefficients is required. In our study we consider three different surfaces: hard court, clay and grass. Hence each player has four levels of strength: a baseline strength plus one for each surface type. This yields more than 2000 strength coefficients and all are allowed to vary over time. Finally, our model specification treats some shortfalls of the Bradley–Terry model which are often encountered in empirical work. A particular example is that the estimated strength of a player tends towards $\pm\infty$ when this player wins or loses all, or almost all, the matches in the data set; see also Baker and McHale (2017) for a discussion and for another method that treats this problem.

The dynamic strengths in our model are specified as score-driven processes. We refer to Creal *et al*. (2013) and Harvey (2013) for reviews on score-driven models, and to Harvey and Luati (2014) and Salvatierra and Patton (2015) for interesting illustrations. The resulting dynamic model forms the basis of our analysis of the large ATP data set of world tournament match results, characteristics and player information over a period of 17 years. The in-sample fit

of our model appears promising when compared with earlier and simplified versions of the model specification. Also the out-of-sample forecasting performance is quite convincing. Our modelling framework can extract four time varying strengths per player from the data: one baseline strength and three surface-specific strengths. Since the data set contains information about more than 500 male tennis players, we have more than 2000 unique time varying player strengths. These evolving paths over time are driven by past observations (all realized match results in the past) and a small number of unknown static parameters. Apart from the time varying strengths, our preferred model also includes some features of tennis matches which we capture by the inclusion of explanatory variables (in particular the seniority of a tennis player) and by accounting for the match configuration (five sets) in a Grand Slam tournament. Given the model specification and the likelihood-based analysis, we can properly measure the significance of regression coefficients by means of a standard likelihood ratio test. The model proposed can also be used to construct surface-specific rankings that are capable of better reflecting the actual abilities of tennis players compared with the ATP point system.

The paper proceeds as follows. Section 2 introduces the modelling framework and proposes several extensions with motivations and discussions. Section 3 presents the results of our empirical study of 17 years of ATP tennis match results for more than 500 men tennis players. We discuss the main findings in detail. Section 4 concludes.

2. The model

2.1. The basic Bradley–Terry model

We consider the Bradley–Terry model. Let $y_{ij,t}$ be the outcome of a tennis match that is played between player i and player j at time t. We assume that we have information about K different players over a time period of length n, i.e. $i, j = 1, \ldots, K$ and $t = 1, \ldots, n$. The outcome $y_{ij,t} = 1$ if player i wins the match at time t. The outcome $y_{ij,t} = 0$ if player j wins the match at time t. The conditional probability that $y_{ij,t} = 1$ is given by

$$p_{ij,t} = P(y_{ij,t} = 1 | \delta_{ij,t}) = \frac{\exp(\delta_{ij,t})}{1 + \exp(\delta_{ij,t})}, \qquad \delta_{ij,t} = \lambda_{i,t} - \lambda_{j,t}, \tag{1}$$

where $\lambda_{i,t}$ represents the strength (or ability) of player i at time t and $\lambda_{j,t}$ represents the strength of player j at time t. The conditional probability of $y_{ij,t} = 0$ is instead equal to $1 - p_{ij,t}$. In the case that the strength $\lambda_{k,t}$ of player k is fixed over time, $\lambda_{k,t} = \lambda_k$, and, under the assumption that sufficient observations for player k are available, we can estimate λ_k via logistic regression; see Cox (1958).

2.2. Time varying strength

The strength of a tennis player is, after reaching its peak, inevitably subject to a permanent decline due to the aging process. In many sports, especially those which require much physical strain, a player in his or her 20s is often at their best. This applies to most individual sports. Of course, it applies also to team sports when we consider individual players in a team. But, when we consider the team, its aging process can be alleviated via reselection. For example, in a football team older and weaker players are replaced by young and more talented players with the aim of keeping or improving the overall ability of the team. In sport statistics, we treat the team as the same entity over time although its composition typically varies considerably over time. The strength of the team can still vary over time but typically more slowly, and partly depending on the financial budget of a team. The time variation in the strength of an individual player is

clearly more dramatic. There is no difference for a tennis player. When we consider dynamic processes for time varying strength, we may consider mean reverting processes for team sports whereas non-stationary dynamic processes may be more appropriate for individual sports.

Consider a match between tennis players i and j at time t and assume that the strengths $\lambda_{i,t}$ and $\lambda_{j,t}$ are given such that the probability $p_{ij,t}$ of a win for player i can be computed by expression (1); the probability of a win for player j equals $1 - p_{ij,t}$. After the match has been played, we record the realized outcome $y_{ij,t}$. This observation provides new information about the (relative) strengths of both players i and j. Hence after the match we need to adjust the levels of strength of both players. We formally specify this adjustment process over time by using a dynamic specification for each strength $\lambda_{k,t}$, for $k = 1, \ldots, K$. We consider a simple random-walk process as given by

$$\lambda_{k,t+1} = \lambda_{k,t} + \tau s_{k,t}, \qquad k = i, j, \tag{2}$$

with scaling coefficient $\tau > 0$ and innovation of the dynamic process $s_{k,t}$. After observing the match outcome $y_{ij,t}$, the innovations $s_{i,t}$ and $s_{j,t}$ are given by

$$s_{i,t} = y_{ij,t}(1 - p_{ij,t}) - (1 - y_{ij,t})p_{ij,t}, \qquad s_{j,t} = -s_{i,t}, \tag{3}$$

with $p_{ij,t}$ as defined in expression (1). The innovations $s_{i,t}$ and $s_{j,t}$ equal the score function of the predictive or conditional density function for $y_{ij,t}$, with respect to the strengths $\lambda_{i,t}$ and $\lambda_{j,t}$ respectively. This specification originates from the score-driven time varying parameter models of Creal *et al.* (2013) and Harvey (2013). In Appendix A we provide further details. The use of a score-driven innovation is appealing given its optimality properties in terms of Kullback–Leibler divergence; see Blasques *et al.* (2015). Furthermore, the innovation specification $s_{i,t}$ is realistic. In the case that the strengths of both players are far apart and $p_{ij,t} = 0.99$, then the observation $y_{ij,t} = 1$ is very likely; the resulting score value is 0.01 such that the strengths of both players do not need to be adjusted very much (0.01τ and -0.01τ). However, in the opposite case of $y_{ij,t} = 0$, the score value is -0.99 and the strength of player i is downgraded by 0.99τ whereas the strength of player j is upgraded by 0.99τ. The scaling coefficient τ in expression (2) is the same for each player. Although τ is common to all players, all time varying strengths are unique because the score innovations are player specific. This strict 'pooling' restriction for τ can be relaxed and different τ-coefficients can be considered for different groups or categories of players.

Fig. 1 presents the impact curve for the score innovation $s_{i,t}$ as a function of the difference in strength between player i and j, i.e. $\delta_{ij,t}$, and the match outcome, i.e. $y_{ij,t}$. We find that the functional form of the score innovation is also intuitive with respect to the difference of strength $\delta_{ij,t} = \lambda_{i,t} - \lambda_{j,t}$. First, the innovation for player i is positive if he wins the match, i.e. $y_{ij,t} = 1$, and negative if he loses the match, i.e. $y_{ij,t} = 0$. Therefore, the strength of a player always increases when he wins a match and decreases when he loses. Second, if player i wins but he is stronger than player j, i.e. $\delta_{ij,t} > 0$, then the innovation is attenuated because a win from player i is expected. Similar arguments apply when player i loses the game and when he is weaker than player j. From Fig. 1 we also find that, even if a player wins or loses all of his matches, the corresponding strength does not diverge to $\pm\infty$ since the score approaches 0 for large values of $\delta_{ij,t}$ for $y_{ij,t} = 1$ and vice versa for $y_{ij,t} = 0$. Hence our specification solves one of the practical problems that are encountered with Bradley–Terry models; see the discussions in Baker and McHale (2017).

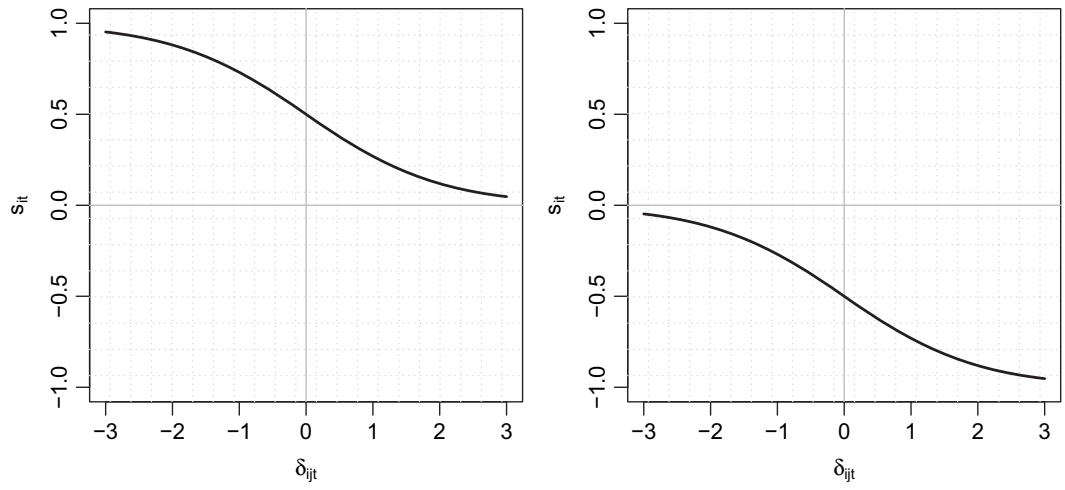

Fig. 1. Impact curve for the score innovation of player i as a function of $\delta_{ij,t}$ and $y_{ij,t}$ (the innovation for player i is positive if he wins the match, i.e. $y_{ij,t} = 1$, and negative if he loses the match, i.e. $y_{ij,t} = 0$; the score innovation approaches 0 for large values of $\delta_{ij,t}$ for $y_{ij,t} = 1$ and vice versa for $y_{ij,t} = 0$): (a) $y_{ij,t} = 1$; (b) $y_{ij,t} = 0$

2.3. Maximum likelihood estimation

The log-likelihood function of the dynamic model is available in closed form via the prediction error decomposition and is given by

$$\mathcal{L}(\psi) = \sum_{t=1}^{T} \sum_{(i,j)\in\mathcal{I}_t} \log\{y_{ij,t} p_{ij,t} + (1 - y_{ij,t})(1 - p_{ij,t})\},$$

where

$$\mathcal{I}_t = \{(i, j) : \text{ a match between player } i \text{ and } j \text{ is played at time } t\}$$

denotes the pairs of players for whom a match takes place at time t, and where ψ is the parameter vector that includes the scaling coefficient τ in expression (2). The estimation of ψ relies simply on the numerical maximization of the log-likelihood function with respect to ψ.

Given the model specification (1)–(2)–(3) that is given in terms of predictions, the strengths for all players at time $t = 0$ need to be given as initial values. The initial values can be treated as static parameters and estimated by the method of maximum likelihood, together with the other parameters in ψ. However, this solution requires an additional number of parameters that is equal to the number of players. In our empirical study below, the number of players exceeds 500. An alternative and more parsimonious solution is to base the initialization on the player ranking points or simply to set all strengths equal to 0. In the empirical study we use the ranking points to initialize the strengths. However, we have found that the other methods lead to very similar results.

2.4. Court surface effects

Tennis matches are played on four types of court surface: hard court, carpet, clay and grass. For instance, the four Grand Slam tournaments, which are the most important tennis tournaments, are played on three different surfaces: the Australian and US Open tournaments are played

on hard courts, the French Open is played on a clay court and Wimbledon is played on a grass court. It is well known that tennis players have different performances when playing on different surface courts. The type of surface affects how the ball bounces as well as the players' movements. This has strong consequences on the characteristics of the match. For instance, the ball tends to bounce slower and higher on a clay surface. This leads to a slower game that favours the so-called baseliners who have a strong defensive game. A notable example is Rafael Nadal who is particularly strong on clay courts. He retains a record of 11 French Open titles.

The court surface can be considered as one of the crucial ingredients to predict tennis matches properly and to assess the level of strength of a tennis player. However, in general, the problem is not straightforward from a statistical point of view. Each player should have a different level of strength for each type of surface. A simple solution would be to include in the model static parameters to account for the type of surface. However, this is not a very appealing solution and it may not even be feasible to model a panel data set with a very large number of players. For instance, in our empirical study this will lead to an increase of more than 2000 parameters that need to be estimated. Consequently, we shall have a time-consuming optimization problem together with high estimation uncertainty and in-sample overfitting issues. The approach that we propose requires only one additional static parameter for each type of surface although the model still contains player-specific and dynamic surface effects.

We introduce the surface effect through the following specification:

$$\lambda_{i,t} = \lambda^{\mathrm{b}}_{i,t} + \sum_{s\in\{\mathrm{h,c,g}\}} I^{\mathrm{s}}_{i,t}\lambda^{\mathrm{s}}_{i,t}, \tag{4}$$

where $\lambda^{\mathrm{b}}_{i,t}$ represents the baseline strength of player i at time t, $\lambda^{\mathrm{s}}_{i,t}$ represents the surface-specific strength of player i at time t on surface s, and $I^{\mathrm{s}}_{i,t}$ is an indicator variable that is equal to 1 if the match of player i at time t is played on surface s and is equal to 0 otherwise. The surface type s belongs to the set $\{\mathrm{h,c,g}\}$ where h denotes hard court, c denotes clay and g denotes grass. Here we merge hard court and carpet court because the characteristics of these surfaces are similar and carpet court is not very common. It follows that $\Sigma_{s\in\{\mathrm{h,c,g}\}} I^{\mathrm{s}}_{i,t} = 1$ because any match is played on one of these three surfaces. The above specification implies that the strength of player i at time t is $\lambda_{i,t} = \lambda^{\mathrm{b}}_{i,t} + \lambda^{\mathrm{h}}_{i,t}$ if the match is played on a hard court, $\lambda_{i,t} = \lambda^{\mathrm{b}}_{i,t} + \lambda^{\mathrm{c}}_{i,t}$ if the match is played on clay and $\lambda_{i,t} = \lambda^{\mathrm{b}}_{i,t} + \lambda^{\mathrm{g}}_{i,t}$ if the match is played on grass. We consider a score-driven process for $\lambda^{\mathrm{b}}_{i,t}$ and $\lambda^{\mathrm{s}}_{i,t}$ as in expression (10) in Appendix A, which leads to the following dynamic equations:

$$\lambda^{\mathrm{b}}_{i,t+1} = \lambda^{\mathrm{b}}_{i,t} + \tau_{\mathrm{b}} s_{i,t}, \qquad \lambda^{\mathrm{s}}_{i,t+1} = \lambda^{\mathrm{s}}_{i,t} + \tau_{\mathrm{s}} I^{\mathrm{s}}_{i,t} s_{i,t}, \qquad \text{for } s \equiv \mathrm{h,c,g}, \tag{5}$$

where τ_{b}, τ_{h}, τ_{c} and τ_{g} are part of the parameter vector ψ and where the score innovation $s_{i,t}$ has the same functional form as given by expression (3).

The dynamic specification that is described in expression (5) is quite intuitive. The strength on a certain surface s depends on two components: one driven only by past matches on that surface and one driven by all past matches. The parameters τ_{b} and τ_{s} determine the relative importance of these two components. If $\tau_{\mathrm{b}} = 0$ then the strength of a player on the surface s depends only on matches that are played on that surface. Instead, if $\tau_s = 0$ there is no surface effect and the strength of the player depends equally on matches played on different surfaces. Therefore, the coefficients τ_{b}, and τ_s, for $s \equiv \mathrm{h,c,g}$, determine the weighing of score innovations, depending on the surface type s. This extended court surface model nests the basic model which is obtained when $\tau_s = 0$ for all $s \in \{\mathrm{h,c,g}\}$ with surface-specific strengths $\lambda^{\mathrm{s}}_{i,t}$ initialized at 0.

2.5. Explanatory variables

Characteristics of player i and/or the match at time t can be treated as explanatory variables and can be included in the model straightforwardly. We denote the vector of explanatory variables of player i at time t by $x_{i,t}$; the corresponding total strength $\tilde{\lambda}_{i,t}$ is then specified by

$$\tilde{\lambda}_{i,t} = \lambda_{i,t} + g(x_{i,t}),$$

where $\lambda_{i,t}$ is the dynamic strength as specified in expression (4) and $g(\cdot)$ is some parametric function. For instance, in our empirical study we consider the home ground advantage $h_{i,t}$ and the age of a player $a_{i,t}$ as explanatory variables. The home ground advantage $h_{i,t}$ is simply a dummy variable that is equal to 1 if the match at time t is played in the home country of player i and is equal to 0 otherwise. Instead, the age $a_{i,t}$ represents the age of player i at time t. We consider the ability of a player to be a non-linear and smooth function of age and therefore we employ a quadratic approximation. A cubic or higher order approximation can also be used. The resulting specification is

$$g(x_{i,t}) = \beta_h h_{i,t} + \beta_{a1} a_{i,t} + \beta_{a2} a_{i,t}^2,$$

where regression coefficient vector $\beta = (\beta_h, \beta_{a1}, \beta_{a2})'$ is part of the overall parameter vector ψ. We have no constant terms in the above specification because they are not identified in our modelling framework with player-specific strengths. Other explanatory variables can be included in the model in a similar fashion. For example, the inclusion of a dummy variable indicating whether a player is left or right handed can be considered. It allows the testing of the hypothesis that left-handed players have an advantage against right-handed players.

2.6. Grand Slam tournaments

It is often observed that good players tend to perform better in Grand Slam tournaments compared with any other standard ATP tournament. This may be because to win a Grand Slam match a player needs to win three sets (best of five) compared with two sets (best of three) for standard tournaments. With more sets played, less randomness is involved. This is easily shown by the following statistical experiment. Assume independence of events and let the probability that player i wins an event against player j be given by $P(i \text{ wins}) = 0.60$. A best-of-five event would result in a winning probability for the whole event of $P(i \text{ wins event}) = 0.683$, compared with $P(i \text{ wins event}) = 0.648$ in a best-of-three event.

To take this effect into account, we specify our model in terms of the probability of winning the set, rather than winning the whole match. Subsequently, we derive the corresponding probability of winning the match under the assumption that the outcomes of the sets within one match are independent. We denote by $\tilde{p}_{ij,t}$ the probability that player i wins a set against player j at time t. It follows that the probability that player i wins a Grand Slam match against player j is given by

$$p_{ij,t} = \tilde{p}_{ij,t}^3 + 3(1 - \tilde{p}_{ij,t})\tilde{p}_{ij,t}^3 + 6(1 - \tilde{p}_{ij,t})^2 \tilde{p}_{ij,t}^3. \tag{6}$$

This result is obtained by summing the probabilities that player i wins all the first three sets, player i loses one of the first three sets and wins the fourth, and player i loses two of the first four sets and wins the fifth. Similarly, the probability that player i wins a standard ATP match is the sum of the probability that he wins both the first two sets and the probability that he loses one of the first two sets and wins the third:

$$p_{ij,t} = \tilde{p}_{ij,t}^2 + 2(1 - \tilde{p}_{ij,t})\tilde{p}_{ij,t}^2, \tag{7}$$

where the probability $\tilde{p}_{ij,t}$ can be specified as in expressions (1) and (2). Furthermore, the surface effect and the explanatory variables can be included in the model in the same way as discussed above. The resulting score innovations of this model are, however, different from the previous specification. In particular, the score innovation for player i that is obtained by observing the outcome $y_{ij,t}$ of a Grand Slam match is given by

$$s_{i,t} = y_{ij,t} \frac{30\tilde{p}_{ij,t}^3(1-\tilde{p}_{ij,t})^3}{p_{ij,t}} - (1-y_{ij,t}) \frac{30\tilde{p}_{ij,t}^3(1-\tilde{p}_{ij,t})^3}{1-p_{ij,t}},$$

whereas the score innovation from a standard ATP match is given by

$$s_{i,t} = y_{ij,t} \frac{6\tilde{p}_{ij,t}^2(1-\tilde{p}_{ij,t})^2}{p_{ij,t}} - (1-y_{ij,t}) \frac{6\tilde{p}_{ij,t}^2(1-\tilde{p}_{ij,t})^2}{1-p_{ij,t}}.$$

In both cases, the score innovation for player j is $s_{j,t} = -s_{i,t}$. These score innovations are specified by applying the generalized auto-regressive score framework and considering the expression of the probabilities given in equations (6) and (7).

Fig. 2 presents the impact curves for the score innovations of matches based on sets. We clearly observe that the outcome of a Grand Slam match delivers a larger (in absolute value) score innovation compared with a standard ATP match. The reason for this difference is that a Grand Slam match has more sets and therefore the outcome of such a match is more informative to assess the levels of strength of both tennis players. The functional forms of these score innovation impact curves are very similar as displayed in Fig. 1; also the intuition behind the functional forms (see the discussion of Fig. 1) remains applicable for the impact curves in Fig. 2.

The perceived higher impact of a Grand Slam tournament can also be accounted for in our model specification. This higher impact can be due to the prestige and the higher prize money. To accommodate this impact, a tennis player can be regarded as an economic agent who puts

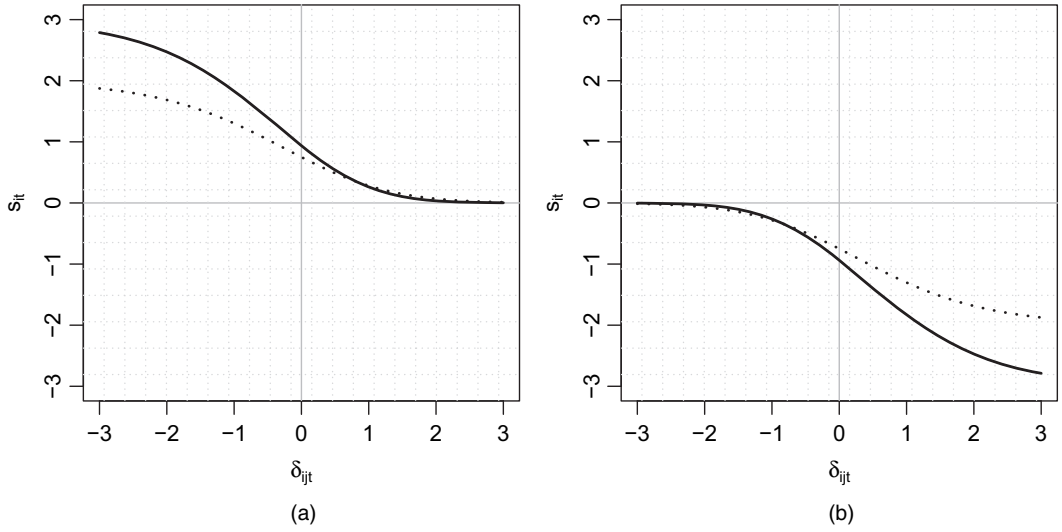

Fig. 2. Impact curves for the score innovations of player i as a function of $\delta_{ij,t}$ and $y_{ij,t}$, for a Grand Slam match (———) and for a standard ATP match (·····) (the figures clearly show that the outcome of a Grand Slam match delivers a larger (in absolute value) score innovation compared with a standard ATP match; the functional forms of the score innovation impact curves are very similar to the score innovation as displayed in Fig. 1; also the intuition behind the functional forms remains applicable for the impact curves in the figure): (a) $y_{ij,t} = 1$; (b) $y_{ij,t} = 0$

in extra time and effort for tournaments where the prestige is high and the prize pool is large with high remunerations. Hence the strength of a top player could be temporarily higher during Grand Slam tournaments, which we can make explicit via the modification

$$\delta_{ij,t} = \gamma(\lambda_{i,t} - \lambda_{j,t}), \tag{8}$$

where scaling coefficient $\gamma > 0$ amplifies the difference in strength between players in Grand Slam matches. This can be formally tested via a t-test or a likelihood ratio test for the null hypothesis of $\gamma = 1$; a significant test statistic for $\gamma > 1$ can provide empirical evidence of a Grand Slam effect.

3. Empirical analysis of tennis match results

3.1. The data set

The data set in our empirical study contains tennis match results by male players in Grand Slam tournaments, the ATP World Tour Finals, ATP World Tour Masters 1000 and ATP World Tour 500 and 250 series from January 2000 to February 2017; see `http://www.tennis-data.co.uk/`. The data set contains information on the tennis matches as well as the tennis players. For each player we have his official ATP ranking points, his ranking position, his age and his country of origin. In contrast, for each match we have the day, the location, the type of tournament (the number of sets in a match), the two players involved and the outcome of the match. Several matches were removed from the final data set after a cleaning process. For example, we removed all matches in which a player retired, matches with invalid sets and/or irregular results, and matches between players where at least one player has no (or a missing) record of his number of ATP points. We also excluded all matches with players who have played only nine games or fewer. When we keep all players in the data set, we obtain roughly the same estimation and forecasting results in our analysis. Hence we have not removed these players and matches because of estimation problems; we believe that it is more appropriate to exclude these 'noisy' observations from the data set. Our statistical methods can handle high numbers of players and matches without numerical problems. After the data cleaning process as described above, the number of players in the data set is 561 and they have played collectively a total of 43175 matches.

3.2. Estimation results

In this section, we discuss the parameter estimation results for the models that we introduced in the previous section. Table 1 summarizes the model specifications and displays the number of parameters. In all six model specifications, an additional parameter α is included for the initial value of the time varying strength of each player which is set equal to α times the logarithm of the ATP ranking points of the player. Models 1.s, 2.s and 3.s are the same as models 1, 2 and 3 respectively, but with a dynamic specification for the set result within a match rather than for the overall match result. The log-likelihood function of the most extensive model, which is model 3.s, is optimized in less than 1 min on a standard laptop computer with an Intel i7 processor and using the software OxMetrics of Doornik (2018). We regard this as very fast given the high dimensionality of the model and the size of the data set.

Table 2 presents the parameter estimates for the six models from which we can draw the following conclusions. First, the surface effect is highly significant: the additional parameters τ_{h}, τ_{c} and τ_{g} in models 2 and 2.s are all estimated to be significantly different from 0. The relative increases of the maximized log-likelihood values for models 2 and 2.s compared with models 1 and 1.s respectively are significant. Hence, for the prediction of the strength of a

Table 1. Model specifications and number of parameters†

Model	*Model description*	*Number of parameters*
1	Basic model as in expressions (1)–(3)	2
2	Model with surface effect as in expressions (4) and (5)	5
3	Model 2 with home ground advantage and age	8
1.s	Basic set model equivalent to model 1	2
2.s	Set model equivalent to model 2	5
3.s	Set model equivalent to model 3	8

†In all model specifications, the logarithms of the ATP ranking points are used to initialize the time varying strengths. Models 1.s, 2.s and 3.s are the same as models 1, 2 and 3 respectively, but with a dynamic specification for the set result within a match rather than for the overall match result.

Table 2. Parameter estimates for six model specifications with standard errors in parentheses, maximized log-likelihood values llik and their corresponding AIC-values

Model	τ_b	τ_h	τ_c	τ_g	α	β_h	β_{a1}	β_{a2}	*llik*	*AIC*
1	0.138	—	—	—	0.120	—	—	—	−26054.3	52113
	(0.069)				(0.033)					
2	0.117	0.033	0.099	0.134	0.117	—	—	—	−25768.2	51546
	(0.005)	(0.006)	(0.008)	(0.019)	(0.019)					
3	0.115	0.032	0.098	0.134	0.131	0.226	2.905	−0.591	−25715.1	51446
	(0.005)	(0.006)	(0.009)	(0.019)	(0.021)	(0.030)	(0.484)	(0.093)		
1.s	0.054	—	—	—	0.076	—	—	—	−26012.0	52028
	(0.020)				(0.027)					
2.s	0.046	0.014	0.040	0.043	0.074	—	—	—	−25712.3	51435
	(0.002)	(0.002)	(0.003)	(0.007)	(0.012)					
3.s	0.045	0.014	0.039	0.045	0.083	0.147	1.847	−0.376	−25658.3	51333
	(0.002)	(0.002)	(0.003)	(0.007)	(0.013)	(0.019)	(0.304)	(0.059)		

player on a certain surface, the information on matches that are played on that surface is most relevant whereas the information from other surfaces is less relevant. This finding can be elicited from the significance of the estimated parameters τ_b. In particular, if $\tau_b = 0$ then only matches played on, for example, a clay court are useful to predict the strength of a player on a clay court. This same implication holds for other types of surface. Second, the home ground advantage in models 3 and 3.s has a significant positive effect on tennis players when they play matches in their country of origin. This finding is coherent with the empirical findings in Koning (2011). We have excluded the so-called wild card players for measuring the home ground effect; applying the home ground effect on those matches as well makes the home ground advantage less pronounced. Third, the estimation results for models 3 and 3.s also reveal that the effect of the age variable is estimated to be significant. Fig. 3 presents the estimated age function. We learn from this analysis that the performance of players is highest at the age of 25 years. In other words, on average and accounting for the other effects, players are at their best at the age of 25 years. This result has an intuitive interpretation: a player's strength increases when

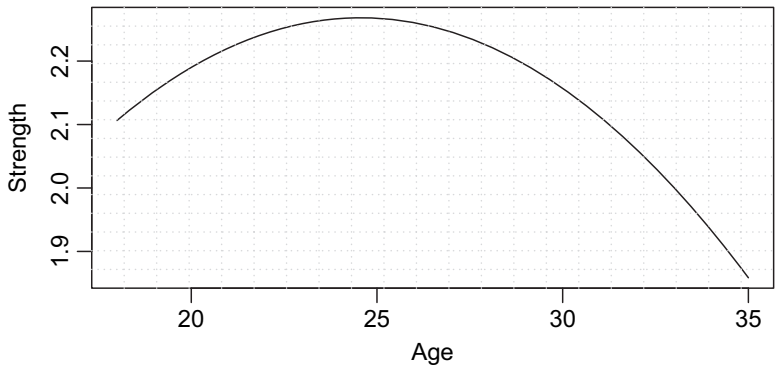

Fig. 3. Estimated strength function by age as given by $\beta_{a1}a + \beta_{a2}a^2$ (y-axis) for a range of age values $a = 18, \ldots, 35$ years (x-axis), for model 3: this effect is treated as common to all tennis players

he is young by gaining experience but then after a certain age his strength starts to decrease as his physical skills deteriorate. Finally, we obtain a better in-sample fit for models 1.s, 2.s and 3.s compared with models 1, 2 and 3 respectively. In particular, the models based on set results have smaller Akaike information criterion AIC compared with their counterparts based on match results. However, the models that are based on sets do not nest the models that are based on matches and therefore a likelihood ratio test cannot be employed to compare these specifications. Nonetheless, AIC can still be useful here since the likelihood of the set models is based on match results through equations (6) and (7).

We further have considered other model extensions that may infer specific empirical features in the data set. First, we have investigated the age effect as displayed in Fig. 3 in more detail by considering other functional forms including linear and cubic polynomials. However, these alternative functional forms do not provide a better fit. Second, we have included other explanatory variables but their regression estimates have not been significant. For example, in our data set of ATP tennis matches we have not found any empirical evidence that left-handed players have a significant advantage against right-handed players. Third, the perceived higher impact of Grand Slam tournaments can be measured by γ in expression (8) in combination with model 3.s, which accounts for the different number of sets in Grand Slam compared with standard ATP matches. However, we have not found an estimate of γ that is significantly larger than 1. Therefore we may conclude that the Grand Slam effect is clearly a statistical effect as evidenced by model 3.s but it appears not to have an additional effect given its prestige or its higher prize pool.

Fig. 4 illustrates the importance of having a player-specific surface. Rafael Nadal is well known to be very strong on clay courts. He has managed to win the French Open tournament 11 times. The French Open is the only Grand Slam tournament that is played on a clay court. His rival Roger Federer has won more Grand Slam tournaments than Rafael Nadal, but he won the French Open tournament only once, in 2009. As we can see in Fig. 4, our model suggests that Federer is stronger than Nadal on a hard court and grass courts, except for a short period of time around 2014. In contrast, Nadal is stronger than Federer on clay courts. For instance, for February 2017 the model predicts that Nadal wins against Federer with a probability of about 29% if the match is played on a hard court; instead, he wins with a probability of about 60% if the match is played on a clay court. This illustrates that the type of court surface can have a massive impact on the winning probability of a match. We also find that, before 2005, Nadal was at the beginning of his professional career and for this reason his level of strength is

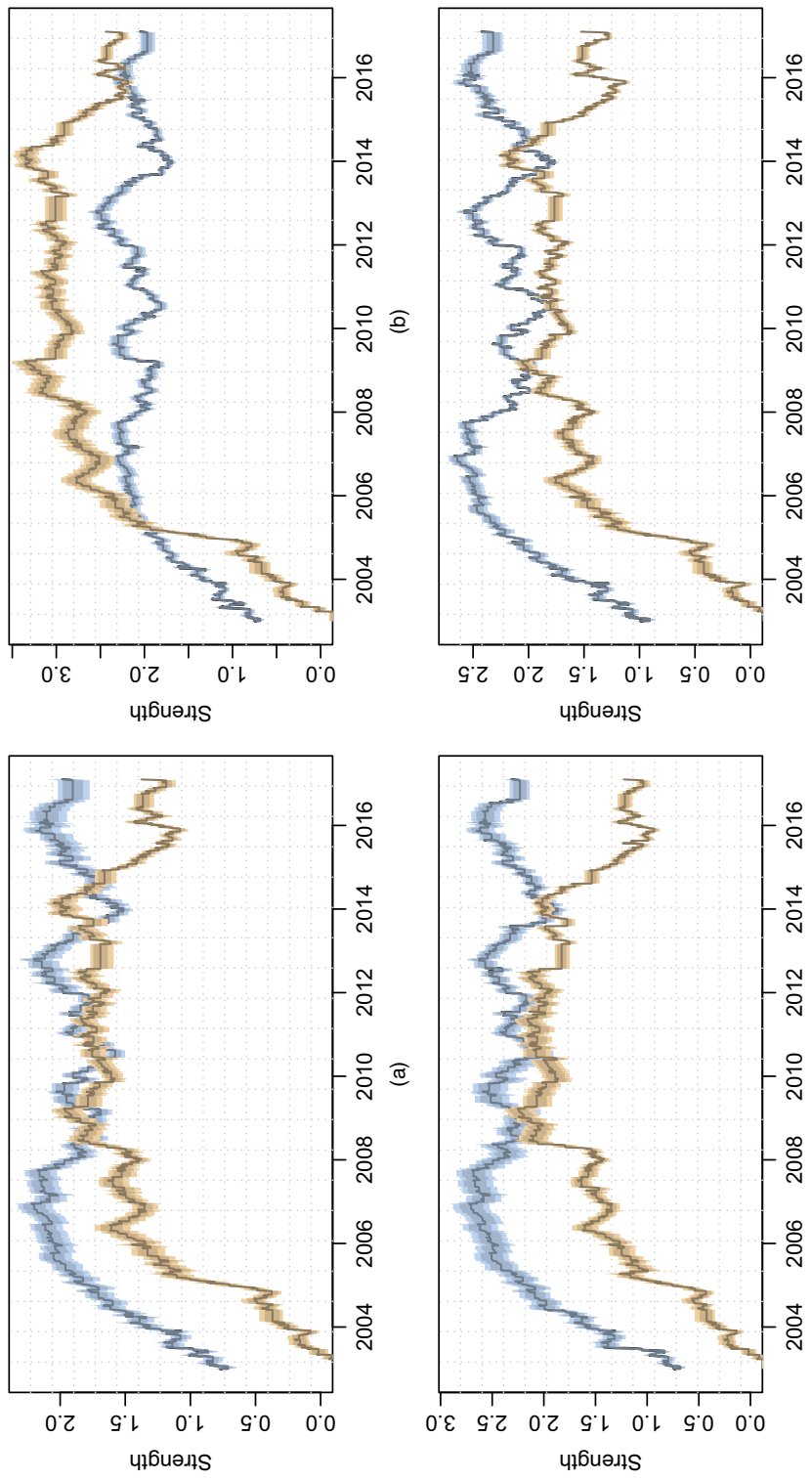

Fig. 4. Dynamic strength levels of Roger Federer (——) and Rafael Nadal (——) for each type of surface (the 90% and 99% confidence bounds for the strength levels (——, ——) are obtained as described in Blasques *et al.* (2016); the bounds take into account parameter estimation uncertainty): (a) baseline; (b) clay court; (c) grass court; (d) hard court

considerably lower than that of Federer. This difference reduces dramatically after 2005 when Nadal won his first Grand Slam tournament.

3.3 Out-of-sample comparison

We have performed an out-of-sample study to evaluate the forecasting performance of the six model specifications together with four benchmark models and the predictions that are implied by the betting market. The first benchmark model is based on the ATP ranking *positions* of the players as proposed by Boulier and Stekler (1999); instead, the second benchmark model is based on ranking *points* as advocated by Clarke and Dyte (2000). These two benchmark models exploit information that is provided by the most recent rankings to predict a match outcome; the specification for both models is given by

$$p_{ij,t} = \frac{\exp(\delta_{ij,t})}{1+\exp(\delta_{ij,t})}, \qquad \delta_{ij,t} = \kappa(r_{i,t} - r_{j,t}),$$

where κ is an unknown parameter and $r_{i,t}$ is a measure of performance of player i at time t. The first benchmark model has the ATP ranking position as $r_{i,t}$ whereas the second model has the logarithm of ATP ranking points as $r_{i,t}$, for player i at time t. The logarithm of ranking points provides a better fit compared with ranking points without transformation; see Clarke and Dyte (2000). The estimated value for κ for both benchmark models (using ATP ranking positions and points) is highly significant and equals -0.0068 and 0.78 respectively for the first and second models. The third benchmark model is an interpolation method where the strength of the players over time is modelled via cubic splines. The number of knots is selected to maximize the log-score criterion and the best performance is achieved with two knots per player. The fourth benchmark is the weighted likelihood approach of McHale and Morton (2011). This model accounts for the surface effect by introducing different weights in the likelihood function depending on the surface where a match is played. The weights cannot be estimated together with the other parameters and they are selected to maximize the log-score in a validation sample over a grid search; see McHale and Morton (2011) for more details on this approach. Finally, the predictions that are implied by the betting market are obtained by using the average market odds that are available from `http://www.tennis-data.co.uk/`. The odds-implied probabilities are obtained from the inverse of the odds and are standardized to sum to 1.

The forecasting study is for the data set that is split into two subsamples: a training sample from 2000 to 2014 and a forecast evaluation sample from 2015 to 2017. We re-estimate the parameters for all models at each time point and consider an expanding window approach. The performance evaluation of the models is based on the log-score criterion as considered by Geweke and Amisano (2011) and, in particular, by McHale and Morton (2011) in the context of tennis forecasts. The log-score criterion is given by $N^{-1}\Sigma_i^N \log(p_i^{\mathrm{w}})$, where p_i^{w} is the probability of the winner predicted by the model and N is the number of matches in the forecast evaluation sample. We consider the Diebold–Mariano test to assess the statistical significance of the predictive ability of the models; see Diebold and Mariano (1995) for further details.

Table 3 reports the out-of-sample results. Among the statistical models, we find that model 3.s is the best model to provide accurate forecasts for all types of surface and it performs significantly better than splines, weighted likelihood, ATP positions and ATP ratings. Furthermore, even our most basic model specification, which is model 1, produces significantly more accurate forecasts than the forecasts that were obtained from the other benchmark models. We also find that models 1.s, 2.s and 3.s perform better than models 1, 2 and 3 respectively. Therefore, we may conclude that the specification that is based on sets delivers more accurate predictions than

Table 3. Log-score total loss and average loss (in parentheses)†

Model	*Results for all courts*		*Results for hard court*		*Results for clay court*		*Results for grass court*	
	Loss	*DM*	*Loss*	*DM*	*Loss*	*DM*	*Loss*	*DM*
Bookmakers' model	−2774.13 (−0.557)	5.91	−1649.02 (−0.557)	4.56	−809.21 (−0.556)	2.94	−315.90 (−0.536)	2.57
ATP position	−3261.19 (−0.655)	−10.88	−1936.65 (−0.654)	−8.83	−924.79 (−0.647)	−4.95	−399.75 (−0.679)	−4.23
ATP points	−3003.82 (−0.603)	−5.95	−1773.78 (−0.599)	−4.78	−863.77 (−0.604)	−2.58	−366.27 (−0.622)	−2.88
Cubic splines	−3076.82 (−0.618)	−6.49	−1811.38 (−0.612)	−4.66	−892.04 (−0.624)	−3.89	−373.39 (−0.634)	−2.97
Weighted likelihood method	−3035.23 (−0.610)	−4.83	−1785.00 (−0.603)	−3.74	−890.14 (−0.623)	−3.23	−360.08 (−0.611)	−1.22
Model 1	−2907.24 (−0.584)	−3.38	−1714.80 (−0.579)	−2.17	−849.32 (−0.594)	−2.66	−343.08 (−0.582)	−0.80
Model 2	−2883.59 (−0.578)	−2.54	−1708.50 (−0.577)	−2.24	−835.91 (−0.585)	−1.63	−339.16 (−0.576)	−0.17
Model 3	−2878.19 (−0.578)	−1.89	−1703.10 (−0.575)	−0.90	−834.64 (−0.584)	−1.58	−340.48 (−0.578)	−0.93
Model 1.s	−2900.31 (−0.583)	−2.94	−1710.46 (−0.578)	−1.82	−847.32 (−0.593)	−2.41	−342.53 (−0.582)	−0.70
Model 2.s	−2878.84 (−0.578)	−1.93	−1706.65 (−0.577)	−2.10	−833.01 (−0.583)	−0.66	−339.19 (−0.576)	−0.26
Model 3.s	−2870.81 (−0.577)	—	−1700.60 (−0.575)	—	−831.46 (−0.582)	—	−338.79 (−0.575)	—

†The second column of each court type reports the Diebold and Mariano (1995) DM-statistics for all considered models against the benchmark model (model 3.s).

the specification that is based on matches. As concerns betting market predictions, the results show that bookmakers' odds have a better performance than all statistical models. This can be explained by the fact that market odds rely on a much wider set of information compared with statistical models. For instance, the models just rely on a limited set of explanatory variables that are available in the data set. Instead, market odds are adjusted according to many other sources of information, including the latest condition of the players. Given that our method significantly outperforms all the other statistical models, we can conclude that our model has a satisfactory forecasting performance.

We have also measured the performance of the models in terms of model accuracy, i.e. the percentage of correct predictions that are produced by a model (or bookmakers). We have obtained that the percentage of correct out-of-sample predictions is very similar across models: between 67% and 69% for all models. This similarity may be because differences in wrong or correct predictions between different models tend to occur only when matches are balanced, in particular the predictions that are close to 50%–50%.

3.4. Ranking tennis players

The ATP ranking is used to determine the entry and the seeding of tennis tournaments. This is of great importance, for instance, to avoid that the two strongest players play against each other in the first stage of a tournament. The ranking should therefore reflect the current level of ability of the players accurately. Furthermore, surface-specific rankings are also useful since

Table 4. First 10 players in the ATP ranking and the rankings obtained from our model on January 9th, 2017†

ATP rank	*Baseline top 10*	*Hard court top 10*	*Clay court top 10*	*Grass court top 10*
Andy Murray	Novak Djokovic	Novak Djokovic	Novak Djokovic	Andy Murray
Novak Djokovic	Roger Federer	Roger Federer	Rafael Nadal	Roger Federer
Milos Raonic	Andy Murray	Andy Murray	Andy Murray	Novak Djokovic
Stanislas Wawrinka	Rafael Nadal	Kei Nishikori	Roger Federer	Ivo Karlovic
Kei Nishikori	Kei Nishikori	Stanislas Wawrinka	Stanislas Wawrinka	Jo Wilfried Tsonga
Gael Monfils	Stanislas Wawrinka	Rafael Nadal	Kei Nishikori	Kei Nishikori
Marin Cilic	Jo Wilfried Tsonga	Milos Raonic	Juan Martin Del Potro	Milos Raonic
Dominic Thiem	Milos Raonic	Jo Wilfried Tsonga	Jo Wilfried Tsonga	Tomas Berdych
Rafael Nadal	Tomas Berdych	Tomas Berdych	Dominic Thiem	Nick Kyrgios
Tomas Berdych	Juan Martin Del Potro	Juan Martin Del Potro	Milos Raonic	Rafael Nadal

†Seven out of 10 players who are in the top 10 of the ATP ranking are also in the top 10 of our model-based baseline ranking. However, there are quite some differences in the order.

the strengths of players vary across different surfaces. The effect of the surface is, for instance, considered in the seeding system that is adopted for the Wimbledon Grand Slam tournament.

Earlier empirical evidence has shown that statistical methods are capable of outperforming the ATP scoring system in terms of predictive ability. In our forecasting study we have given evidence that our model 3.s produces significantly better predictions than those based on ATP ranking points and on actual rankings for all surfaces. Next we derive rankings that are based solely on the estimated strengths from model 3.s for the different surfaces. In particular, we can sort the players with respect to their level of strength on each surface. The baseline strength is used to obtain an overall ranking. The model-based rankings can improve ATP rankings to sort players in terms of overall ability but such rankings may lack other desirable features. We refer to Irons *et al.* (2014) for a discussion on how to construct tennis rankings by using statistical models.

Table 4 reports the first 10 players in the ATP ranking and the rankings that were obtained from our model on January 9th, 2017. There are some similarities but also some differences across the rankings. Seven out of 10 players who are in the top 10 of the ATP ranking are also in the top 10 of our model-based baseline ranking. However, there are quite some differences in the order. For instance, Novak Djokovic is first in the baseline ranking but second in the ATP ranking. Roger Federer is ranked 17th in the ATP ranking whereas he is ranked second in our baseline ranking. For the rankings of different surfaces, we have expected that Rafael Nadal is better positioned in the clay ranking compared with the other rankings: second position on clay but outside the top five in all other rankings.

To evaluate how the rankings are related to each other, we measure their closeness by using the Kendall correlation measure. The Kendall correlation is a measure of correlation between rankings: it is equal to 1 if two rankings are the same and equal to -1 if two rankings are the same but reversed. Fig. 5 presents the Kendall correlation between the different rankings. We find that the ATP ranking is farther apart from all other rankings with a correlation value of around 0.5. When we focus on the model-based rankings, we observe that the ranking for clay is the least correlated with the grass, hard and baseline ranking. This indicates that the clay surface differs most from the other surfaces in terms of players' abilities. Finally, we learn that the ranking based on hard courts is the closest to the baseline ranking. This finding is not surprising since the baseline strength accounts for all surfaces in the same way and the majority of tennis matches in the data set were played on hard courts.

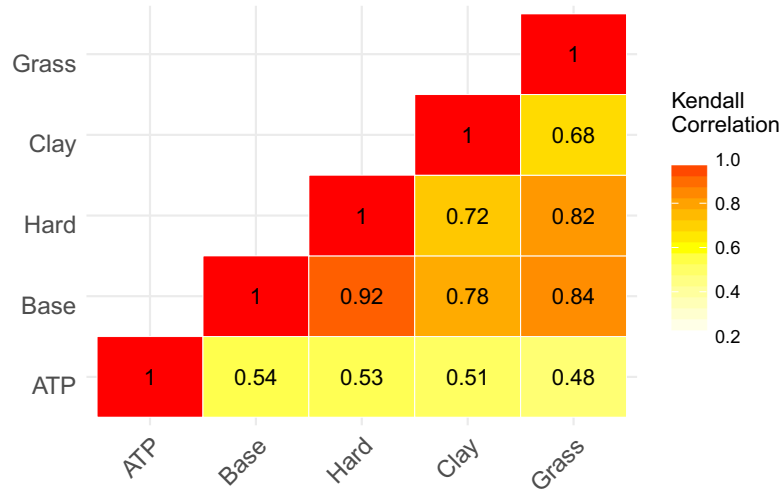

Fig. 5. Kendall correlation between the different rankings: the ATP ranking is farther apart from all other rankings with a correlation value of around 0.5; the clay court ranking is the least correlated with the grass, hard court and baseline ranking; the ranking based on hard courts is the closest to the baseline ranking

4. Conclusion

We have introduced a novel dynamic model for the analysis and forecasting of tennis matches. The model accounts for time varying strengths of the players and different court surface effects for the matches. The likelihood-based analysis can reveal interesting features of tennis matches and can provide accurate forecasts when compared with other modelling approaches. The levels of strength for different types of surface that are extracted from our model can be used to construct improved rankings of players. The rankings can reflect the actual abilities of tennis players more accurately when compared with rankings based on ATP points. Surface-specific rankings can be useful for entry and seeding of tennis tournaments.

Appendix A: A short review of score-driven time series models

Assume that we have a large panel of time series variables denoted $y_{ij,t}$, for $i \neq j = 1, \ldots, M$ and $t = 1, \ldots, T$. All observations at time t are contained in the vector y_t, which in our case consists of binary variables for all match results at time t. We assume that the data have a conditional probability mass function of the form

$$y_t \sim p(y_t|\lambda_t; \psi),$$

where $\lambda_t = (\lambda_{1,t}, \ldots, \lambda_{M,t})'$ is an M-dimensional vector of time varying parameters and ψ is a static parameter vector. The generalized auto-regressive score framework of Creal *et al.* (2013) and Harvey (2013) specifies the dynamics of λ_t as

$$\lambda_{i,t+1} = \omega_i + \phi_i \lambda_{i,t} + \tau_i s_{i,t}, \tag{9}$$

where ω_i, τ_i and ϕ_i are unknown parameters to be estimated, and $s_{i,t}$ is the score innovation of the process that is defined by

$$s_{i,t} = S_{i,t} \nabla_{i,t}, \qquad \nabla_{i,t} = \frac{\partial \log\{p(y_{ij,t}|\lambda_t; \psi)\}}{\partial \lambda_{i,t}}, \tag{10}$$

with $\nabla_{i,t}$ the score of the predictive likelihood and $S_{i,t}$ a scaling factor. A possible choice for the scaling factor is the inverse of the Fisher information to account for the curvature of the likelihood function. An

alternative option is to set the scale equal to 1; we simply have $s_{i,t} = \nabla_{i,t}$. When M is large, the parameters ω_i, ϕ_i and τ_i can be pooled towards a smaller set of parameters. For example, we can have $\omega_i = \omega$, $\phi_i = \phi$ and $\tau_i = \tau$ such that the number of parameters $3M$ is reduced to 3. In our study we have replaced the autoregressive dynamic specification (9) by the random-walk process $\lambda_{i,t+1} = \lambda_{i,t} + \tau s_{i,t}$ with single parameter τ. For instance, considering a basic specification of the Bradley–Terry model, the conditional probability mass function for $y_{ij,t}$ is

$$p(y_{ij,t}|\lambda_t;\psi) = \left\{\frac{\exp(\lambda_{i,t}-\lambda_{j,t})}{1+\exp(\lambda_{i,t}-\lambda_{j,t})}\right\}^{y_{ij,t}} \left\{\frac{1}{1+\exp(\lambda_{i,t}-\lambda_{j,t})}\right\}^{1-y_{ij,t}}.$$

The functional form of the score innovation in expression (3) is obtained immediately by taking the derivative of the logarithm of the above expression and setting the scaling factor to 1, $S_{i,t} = 1$, i.e.

$$s_{i,t} = \frac{\partial \log\{p(y_{ij,t}|\lambda_t;\psi)\}}{\partial \lambda_{i,t}}.$$

A more detailed discussion on score-driven models is provided by Creal *et al.* (2013).

References

Baker, R. D. and McHale, I. (2014) A dynamic paired comparisons model: who is the greatest tennis player? *Eur. J. Oper. Res.*, **236**, 677–684.

Baker, R. D. and McHale, I. (2017) An empirical Bayes model for time-varying paired comparisons ratings: who is the greatest women's tennis player? *Eur. J. Oper. Res.*, **258**, 328–333.

Blasques, F., Koopman, S. J., Lasak, K. and Lucas, A. (2016) In-sample confidence bands and out-of-sample forecast bands for time-varying parameters in observation-driven models. *Int. J. Forecast.*, **32**, 875–887.

Blasques, F., Koopman, S. J. and Lucas, A. (2015) Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, **102**, 325–343.

Boulier, B. L. and Stekler, H. O. (1999) Are sports seedings good predictors?: An evaluation. *Int. J. Forecast.*, **15**, 83–91.

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

Clarke, S. R. and Dyte, D. (2000) Using official ratings to simulate major tennis tournaments. *Int. Trans. Oper. Res.*, **7**, 585–594.

Cox, D. R. (1958) The regression analysis of binary sequences (with discussion) *J. R. Statist. Soc.* B, **20**, 215–242.

Creal, D., Koopman, S. J. and Lucas, A. (2013) Generalized autoregressive score models with applications. *J. Appl. Econmetr.*, **28**, 777–795.

Diebold, F. X. and Mariano, R. S. (1995) Comparing predictive accuracy. *J. Bus. Econ. Statist.*, **13**, 253–265.

Doornik, J. A. (2018) *An Object-oriented Matrix Programming Language Ox 8.0*. London: Timberlake Consultants.

Geweke, J. and Amisano, G. (2011) Optimal prediction pools. *J. Econmetr.*, **164**, 130–141.

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

Harvey, A. C. (2013) *Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series*. New York: Cambridge University Press.

Harvey, A. C. and Luati, A. (2014) Filtering with heavy tails. *J. Am. Statist. Ass.*, **109**, 1112–1122.

Irons, D. J., Buckley, S. and Paulden, T. (2014) Developing an improved tennis ranking system. *J. Quant. Anal. Sprts*, **10**, 109–118.

Klaassen, F. J. and Magnus, J. R. (2001) Are points in tennis independent and identically distributed?: Evidence from a dynamic binary panel data model. *J. Am. Statist. Ass.*, **96**, 500–509.

Klaassen, F. J. and Magnus, J. R. (2003) Forecasting the winner of a tennis match. *Eur. J. Oper. Res.*, **148**, 257–267.

Koning, R. H. (2011) Home advantage in professional tennis. *J. Sprts Sci.*, **29**, 19–27.

McHale, I. and Morton, A. (2011) A Bradley-Terry type model for forecasting tennis match results. *Int. J. Forecast.*, **27**, 619–630.

Newton, P. K. and Aslam, K. (2009) Monte Carlo tennis: a stochastic Markov chain model. *J. Quant. Anal. Sprts*, **5**, no. 3.

Salvatierra, I. D. L. and Patton, A. J. (2015) Dynamic copula models and high frequency data. *J. Empir. Finan.*, **30**, 120–135.