TI 2018-009/III
Tinbergen Institute Discussion Paper

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

P. Gorgi[1]
Siem Jan (S.J.) Koopman[2]
R. Lit[3]

1: VU Amsterdam, The Netherlands
2: VU Amsterdam, The Netherlands; CREATES Aarhus University, Denmark; Tinbergen Institute, The Netherlands
3: VU Amsterdam, The Netherlands

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

P. Gorgi[a,b], S.J. Koopman[a,b,c], and R. Lit[a]

[a]Vrije Universiteit Amsterdam, The Netherlands
[b]Tinbergen Institute, The Netherlands
[c]CREATES, Aarhus University, Denmark

January 23, 2018

Abstract

We propose a basic 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 is capable of handling high-dimensional datasets while the number of parameters remains small. In particular, we analyze 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 analyzing tennis matches. We further consider several other extensions including player-specific explanatory variables and the accountance of specific 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 can also be effective in forecasting. We provide evidence that our model significantly outperforms existing models in the forecasting of tennis match results.

1 Introduction

Modeling and predicting the outcome of tennis matches has attracted much attention over the last few years. Statistical models can be useful to describe the main features of tennis matches and elicit the ability level of tennis players in different situations. This can be used to construct rankings and determine entry and seeding of tennis tournaments. Models can also be employed to obtain predictions of matches and tournaments and test the efficiency of betting markets.

The default 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) have considered ATP rankings points to describe the strength level of tennis players. Glickman (1999) has introduced an algorithm to dynamically update the parameter estimates of the Bradley-Terry model within a Bayesian analysis. McHale and Morton (2011) has used a weighted likelihood approach to account for time variation in the ability level of the players within the Bradley-Terry model. Baker and McHale (2014) and Baker and McHale (2017) have 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 strength level of tennis players is one of the key ingredients to properly describe the outcome of tennis matches. The strength of a player typically increases from a young age and reaches a certain peak when he/she is in his/her twenties, followed by a decline until the he/she ends his/her career. However, in all studies so far, time variation is achieved through a modification of an estimation method; it has not been modeled 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 strength levels 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.

Modeling 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 let these coefficients to vary over time, we clearly have an intrinsically difficult problem at 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 surface type has its own characteristics and impacts 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 won 19 Grand Slam tournaments but only one of them was on the clay surface of the French Open. We notice that the French Open is the only Grand Slam tournament played on clay. On the other hand, Nadal won 16 Grand Slam tournaments of which 10 were wins of the French Open 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 modeling of tennis matches. When different strengths for different court surface types need to be specified in the model, then a multiple of strength coefficients are required. In our study we consider three different surfaces: hard court, clay and grass. Hence each player has 4 strength levels: 1 baseline strength plus 1 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 short-falls of the Bradley-Terry model which are typically encountered in empirical work; see, for instance, Baker and McHale (2017). A particular example is that the estimated strength of a player tends towards plus or minus infinity when this player wins or loses all, or almost all, matches in the data set.

The dynamic strengths in our model are specified as score-driven processes. We refer the reader to Creal et al. (2013) and Harvey (2013) for a review on score-driven models, see also Salvatierra and Patton (2015) and Harvey and Luati (2014) for further applications. The resulting dynamic model forms the basis of our analysis of a large data set of ATP (Association of Tennis Professionals) 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 to earlier and simplified versions of the model specification. Also the out-of-sample forecasting performance is rather convincing. Our modeling framework is able to extract four time-varying strengths per player from the data: one baseline strength and three surface specific strengths. Since our 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 underlying and unknown 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 proposed model can also be used to construct surface-specific rankings that are capable of better reflecting the actual abilities of tennis players compared to the ATP point system.

The paper proceeds as follows. Section 2 introduces the modeling framework and several extensions. Section 3 presents the empirical application. 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 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, that is $i, j = 1, \ldots, K$ and $t = 1, \ldots, n$. The outcome $y_{ij,t}$ equals unity if player i wins the match at time t, that is $y_{ij,t} = 1$. The outcome $y_{ij,t}$ equals zero if player j wins the match at time t, that is $y_{ij,t} = 0$. 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 case, the strength $\lambda_{k,t}$ of player k is fixed over time, that is $\lambda_{k,t} = \lambda_k$, 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 permanent decline due to the ageing process. In many sports, especially those which require much physical strain, a player in its twenties is often at its best. This phenomenon applies to most individual sports. Of course, it also applies to team sports when we consider individual players in a team. But when we consider the team, its ageing process can be alleviated via re-selection. For example, in a football team older and weaker players are replaced by young and more talented players with the aim to keep or improve the overall ability of the team. In sport statistics, we treat the team as the same entity over time although its composition typically varies heavily over time. The strength of the team can still vary over time but at a more constant level, 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 while 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 (1); the probability of a win for player j equals $1 - p_{ij,t}$. After the match is 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 strength levels of both players. We formally specify this adjustment process over time 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 (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 the Appendix 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 case 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 ($0.01\,\tau$ and $-0.01\,\tau$). 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 $0.99\,\tau$ while the strength of player j is upgraded by $0.99\,\tau$. The scaling coefficient τ in (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.

Figure 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, that is $\delta_{ij,t}$, and the match outcome, that is $y_{ij,t}$. We find that the functional form of the score innovation is also intuitive with respect to the strength difference $\delta_{ij,t} = \lambda_{i,t} - \lambda_{j,t}$. First, the innovation for player i is positive if he wins the match, that is $y_{ij,t} = 1$, and negative if he loses the match, that is $y_{ij,t} = 1$. Second, if player i wins but he is stronger than player j, that is $\delta_{ij,t} > 0$, 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 Figure 1 we also find that even if a player wins or loses all of his matches, the corresponding strength does not diverge to ± infinity since the score approaches zero 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 encountered with Bradley-Terry models; see the discussions in Baker and McHale (2017).

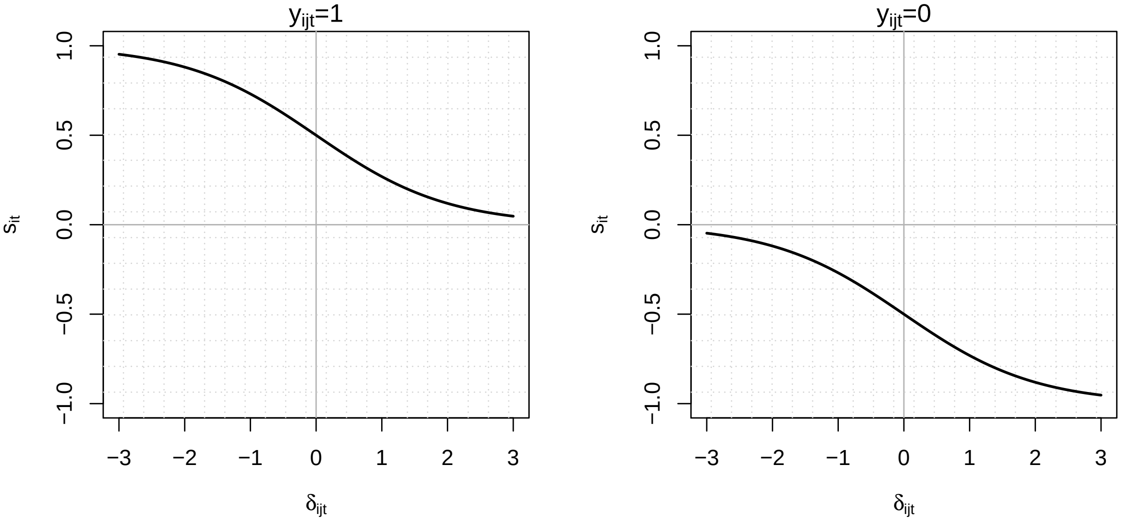

Figure 1: *Impact curve for the score innovation of player i as a function of $\delta_{ij,t}$ and $y_{ij,t}$.*

2.3 Maximum likelihood estimation

The loglikelihood function of the dynamic model is available in closed form via the prediction error decomposition

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

where

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

denotes the pairs of players for which a match takes place at time t, and where ψ is the parameter vector that includes τ. The estimation of ψ relies simply on the numerical maximization of the loglikelihood function with respect to ψ.

Given the specification in terms of predictions, the strengths for each player at time $t = 0$ need to be given initial values. These initial values can be treated as static parameters and estimated by the method of maximum likelihood, jointly with the other parameters. However, this solution requires an additional number of parameters that is equal to the number of players. In our study this number 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 zero. In the empirical application 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 surfaces: 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 are played on hard court, the French Open is played on clay court and Wimbledon is played on 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 player 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 that have a strong defensive game. A notable example is Rafael Nadal who is particularly strong on clay courts. He detains a record of ten French Open titles.

The court surface can be considered one of the crucial ingredients to properly predict tennis matches and assess the strength level of a tennis player. However, in general, the problem is not straightforward from a statistical point of view. Each player should have a different strength level for each surface type. A simple solution would be to include in the model static parameters to account for the surface type. However, this is not a very appealing solution and it may not even be feasible to model a panel dataset with a very large number of players. For instance, in our application this would lead to over two thousand parameters to be estimated. The consequence would be a time-consuming optimization problem as well as very large estimation uncertainty and in-sample overfitting issues. The approach we propose requires only one additional static parameter for each surface type and it allows the model to have a player-specific and dynamic surface effect.

We introduce the surface effect through the following specification

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

where $\lambda_{i,t}^{b}$ represents the baseline strength of player i at time t, $\lambda_{i,t}^{s}$ represents the surface specific strength of player i at time t on surface s and $I_{i,t}^{s}$ is an indicator variable that is equal to 1 if the match of player i at time t is played in surface s. The surface type s belongs to the set $\{h, c, g\}$ 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. We note that $\sum_{s \in \{h,c,g\}} I_{i,t}^{s} = 1$ 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 $\lambda_{i,t} = \lambda_{i,t}^{b} + \lambda_{i,t}^{h}$ if the match is played on hard court, $\lambda_{i,t} = \lambda_{i,t}^{b} + \lambda_{i,t}^{c}$ if the match is played on clay and $\lambda_{i,t} = \lambda_{i,t}^{b} + \lambda_{i,t}^{g}$ if the match is played on grass. We consider a score-driven process for $\lambda_{i,t}^{b}$ and $\lambda_{i,t}^{s}$ as in (7), which leads to the following dynamic equations

$$\begin{aligned} \lambda_{i,t+1}^{b} &= \lambda_{i,t}^{b} + \tau_b s_{i,t}, \\ \lambda_{i,t+1}^{s} &= \lambda_{i,t}^{s} + \tau_s I_{i,t}^{s} s_{i,t}, \quad \text{for} \quad s = h, c, g. \end{aligned} \tag{5}$$

with τ_b, τ_h, τ_c and τ_g part of the parameter vector ψ and where the score innovation $s_{i,t}$ has the same functional form as in (3).

The dynamic specification described in (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 τ_b is equal to zero then the strength of a player on the surface s depends only on matches that are played on that surface. Instead, if τ_s is equal to zero there is no surface effect and the strength of the player depends equally on matches played on different surfaces. We also note that this model nests the basic model when $\tau_s = 0$ for any $s \in \{h, c, g\}$ with surface-specific strengths $\lambda^s_{i,t}$ initialized at zero.

2.5 Explanatory variables

Explanatory variables can be easily included in the model. We denote by $x_{i,t}$ the vector of explanatory variables of player i at time t. The strength $\lambda_{i,t}$ can be specified as

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

where $\lambda_{i,t}$ is the dynamic strength as specified in (4) and $g(\cdot)$ is some parametric function. For instance, in the empirical application we consider the home ground advantage h_i and the age of the players $a_{i,t}$ as explanatory variables. The home ground advantage $h_{i,t}$ is simply a dummy variable that is equal to one if the match at time t is played in the country of player i and zero 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 nonlinear and smooth function of age and therefore we employ a quadratic approximation[1]. The resulting specification is

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

We have no constant terms in the above specification because they are not identified in our modeling framework with player-specific strengths.

Other explanatory variables can be included in the model in a similar fashion. For instance, a dummy variable indicating whether a player is left- or right-handed can be considered. This may be useful to test the hypothesis that left-handed players have an advantage against right-handed players. In our analysis of the ATP tennis matches we do not find empirical evidence that left-handed players have a significant advantage.

2.6 Grand Slam tournaments

It is often observed that good players tend to perform better in Grand Slam tournaments compared to any other standard ATP tournament. This may be due to the fact that to win a Grand Slam match a player needs to win three sets (best of five) compared to 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 to $P(i \text{ wins event}) = 0.648$ in a best of three event.

To take this into account, we specify a model for the probability of winning a set (rather than winning the whole match) and derive the corresponding probability of winning the match under the assumption that the set results are independent. We denote by $\tilde{p}_{ij,t}$ the probability that player

[1] A cubic or higher order approximation can also be used. However, we find that the inclusion of a cubic term is not significant.

i wins a set against player j at time t. It follows that the probability of player i winning a Grand Slam match against player j is given by

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

Similarly, the probability that player i wins a standard ATP match is

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

The probability $\tilde{p}_{ij,t}$ can be specified as in (1) and (2). Furthermore, the surface effect and the explanatory variables can be included in the model in the same way as discussed before.

We note that the resulting score innovations of this model are different from the previous specification. In particular, the score innovation for player i obtained observing the outcome $y_{ij,t}$ of a Grand Slam match is given by

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

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

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

In both cases, the score innovation for player j is $s_{j,t} = -s_{i,t}$.

Figure 2 shows the impact curve for the score innovations based on sets. It is interesting to see that the outcome of a Grand Slam match delivers a larger (in absolute value) score innovation compared to a standard ATP match. The reason for this is that a Grand Slam match has more sets and therefore the outcome of a match is more informative to assess the strength level of the players involved. Concerning the shape of the score innovation, we note that a similar interpretation as discussed for Figure 1 remains valid.

One could opt for an additional strategy of modeling Grand Slam tournaments. A tennis player could be regarded as an economic agent who puts in extra time and effort for tournaments where the price pool is large. The strength of top players could be temporarily higher during Grand Slam tournaments which could be modeled as

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

where γ amplifies the difference in strength between players. This can be formally tested by a t-test or a likelihood ratio test that test for significant difference from 1 of the parameter γ. We incorporated this in our model but found γ to be not significantly different from 1 for the set model described in this section. We conclude that the 'Grand Slam' effect is most likely a statistical effect and not due to extra effort because of a larger price pool.

3 Analysis of ATP tennis match results

3.1 The dataset

The dataset we consider contains tennis match results of Grand Slam tournaments, ATP World Tour Finals, ATP World Tour Masters 1000 and ATP World Tour 500 and 250 series from January

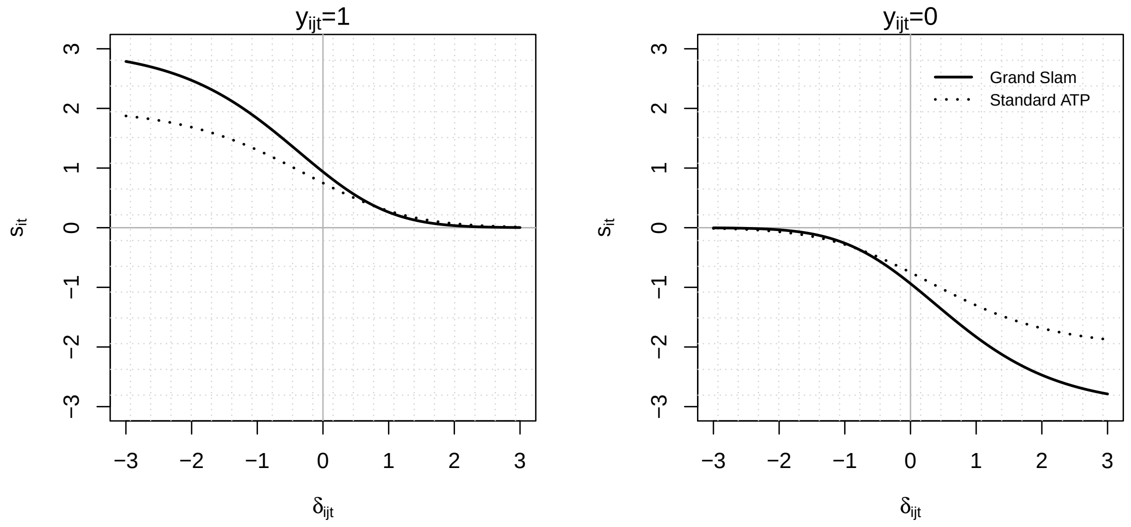

Figure 2: *Impact curve for the score innovations of player i as a function of $\delta_{ij,t}$ and $y_{ij,t}$.*

2000 to February 2017. The dataset contains information on the tennis matches as well as the tennis players. For instance, for each player we have the official ATP ranking points, the ranking position, the age of the player and his country of origin. Instead, for each match we have the date, the location, the type of tournament, the players involved and the outcome of the match. Several matches were excluded from the dataset in a cleaning process. Matches in which a player retired, invalid set and/or match results, and matches with missing ATP points were excluded from the data. We also excluded players that played less than ten games. We noticed that leaving them in leads to roughly the same estimation and forecasting results. We emphasize that we did not remove those matches because of estimation problems, our new model can handle a low number of player matches with ease. After cleaning, the number of players in the dataset is 561 who played a total of 43,175 matches.

3.2 Estimation results

In this section we estimate the models that were introduced in the previous section. Table 1 summarizes the model specifications and shows the number of parameters. In all five model specifications, parameter α is an initialization parameter which is multiplied by the log of the ATP ranking points to initialize the time-varying strengths. The likelihood of the most extensive model, that is Model 5, is optimized in less than one minute on a standard new model laptop with i7 processor. We regard this as very fast given the high dimensionality of the model.

Table 2 reports the parameter estimates of the models. We note that the effect of the surface is highly significant: the additional parameters τ_h, τ_c and τ_g are all significantly different from zero. This can also be seen by comparing the likelihood of Model 1 with the likelihood of Model 2. Furthermore, from the estimation results of Model 2, we can conclude that to predict the strength of a player on a certain surface also the information on matches played in the other surfaces is useful. This finding can be elicited from the significance of the parameter τ_b. In fact, if τ_b is equal

Label	Model description	# parameters
Model 1	Basic model as in (1) - (3)	2
Model 2	Model with surface effect as in (4) and (5)	5
Model 3	Model 2 with home ground advantage	6
Model 4	Model 3 with the variable age	8
Model 5	Set model with surface effect and all variables	8

Table 1: *Model specifications.*

to zero then only matches played on, for example, a clay court are useful to predict the strength of a player on a clay court. The same holds true for the other surface types. Concerning the home ground advantage, we can see that there is a significant and positive effect for players playing in their country of origin. This finding is also coherent with the results in Koning (2011). We excluded so called wild card players from the home ground analysis. Including them made the home ground advantage less pronounced. From the estimation results of Model 4, we can see that the effect of the variable age is significant. Figure 3 shows the plot of the estimated age function. We can see that the performance of players is highest at the age of 25. This means that on average players are their best at the age of 25. This result has an intuitive interpretation: a player strength increases when he is young by gaining experience but then after 25 his strength starts decreasing as his physical skills deteriorate. Finally, we note that modeling the set results instead of directly the match results leads to a better in-sample fit. In particular, Model 5 has the smallest AIC and this indicates a better fit. Note that Model 5 does not nest the other models and therefore a likelihood ratio test cannot be employed. However, the AIC can be a useful means of comparison in this case.

	τ_b	τ_h	τ_c	τ_g	α	β_h	β_{a1}	β_{a2}	llik	AIC
Model 1	0.138 (0.069)	-	-	-	0.120 (0.033)	-	-	-	-26054.3	52113
Model 2	0.117 (0.005)	0.033 (0.006)	0.099 (0.008)	0.134 (0.019)	0.117 (0.019)	-	-	-	-25768.2	51546
Model 3	0.118 (0.005)	0.031 (0.006)	0.098 (0.008)	0.131 (0.019)	0.120 (0.019)	0.228 (0.030)	-	-	-25740.0	51492
Model 4	0.115 (0.005)	0.032 (0.006)	0.098 (0.009)	0.134 (0.019)	0.131 (0.021)	0.226 (0.030)	2.905 (0.484)	-0.591 (0.093)	-25715.1	51446
Model 5	0.045 (0.002)	0.014 (0.002)	0.039 (0.003)	0.045 (0.007)	0.083 (0.013)	0.147 (0.019)	1.847 (0.304)	-0.376 (0.059)	-25658.3	51333

Table 2: *Parameter estimates of the models.*

Figure 4 illustrates the importance of having a player-specific surface. Rafael Nadal is well-known to be very strong on clay courts. He managed to win the French Open ten times. The French Open is the only Grand Slam tournament that is played on a clay court. His rival Roger Federer won more Grand Slam tournaments than Rafael Nadal, but he won the French Open only once in 2009. As we can see in Figure 4, our model suggests that Federer is stronger than Nadal on hard court and grass courts, except for a short period of time around 2014. In contrast, Nadal is stronger than Federer on clay courts. Note that before 2005 Nadal was at the beginning of his

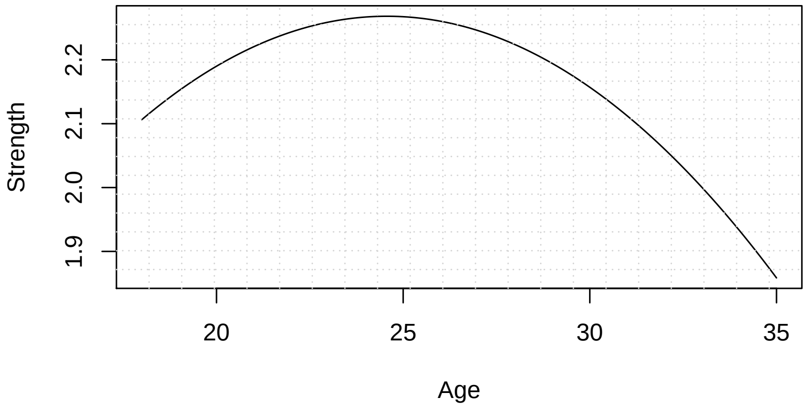

Figure 3: *Estimated age function.*

professional career and for this reason his level of strength is quite lower than the one of Federer. This difference reduces dramatically after 2005 when Nadal won his first Grand Slam tournament.

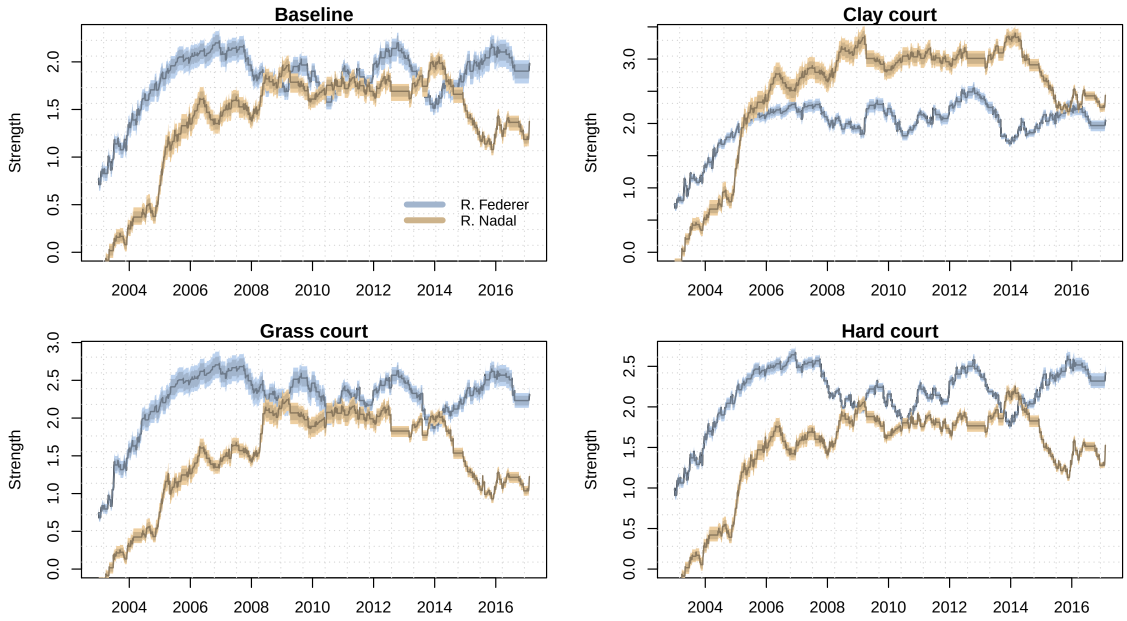

Figure 4: *Dynamic strengths of Roger Federer and Rafael Nadal for each surface type. Confidence bounds for the strengths at* 90% *and* 99% *levels are represented by the shaded area. The confidence bounds are obtained as in Blasques et al. (2016).*

3.3 Out-of-sample comparison

We perform an out-of-sample study to evaluate the forecasting performance of the proposed models. We include two benchmark models in our comparison: one based on the ATP ranking position of the players, Boulier and Stekler (1999), and one based on ranking points, Clarke and Dyte (2000). These models exploit information provided by the rankings to predict a match outcome. The specification of these models are given by

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

where κ is a parameter to be estimated and $r_{i,t}$ is a measure of performance of player i at time t. For the model based on the ATP ranking position, $r_{i,t}$ denotes the position in the ranking, instead for the model based on the ATP ranking points, $r_{i,t}$ is the logarithm of the ranking points. The logarithm is considered in Clarke and Dyte (2000) and it provides a better fit compared to not having any transformation. The estimated coefficient κ of both models is highly significant and equal to -0.0068 and 0.78 for the model based on the ATP ranking position and points, respectively.

	All courts		**Hard court**		**Clay court**		**Grass court**	
	Loss	DM stat.	Loss	DM stat.	Loss	DM stat.	Loss	DM stat.
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
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	-2885.82 (-0.580)	-2.53	-1709.00 (-0.577)	-2.04	-836.22 (-0.585)	-1.52	-340.55 (-0.578)	-0.69
Model 3	-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 4	-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 5	-2870.81 (-0.577)	-	-1700.60 (-0.575)	-	-831.46 (-0.582)	-	-338.79 (-0.575)	-

Table 3: *Log score total loss and average loss (in brackets). The second column of each court type reports the DM statistics of the models against the benchmark model (set model).*

We split the dataset into two sub-samples: a training sample from 2000 to 2014 and a forecasting evaluation sample from 2015 to 2017. We re-estimate all the models at each time point by considering 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 by McHale and Morton (2011) in the context of tennis forecasts. The log score criterion is: $N^{-1} \sum_i^N \log p_i^w$, where p_i^w is the probability of the winner predicted by the model and N is the evaluation sample size. We consider the Diebold-Mariano (DM) test to assess the statistical significance of the predictive ability of the models, Diebold and Mariano (1995). Table 3 reports the out-of-sample results. We can see that Model 5 is the best model for all types of surfaces and it performs significantly better than the ATP points and ranking models. Furthermore, we also note that also our most basic specification, that is Model 1, performs significantly better than the ATP points and ranking models

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 ability level of the players. Furthermore, surface-specific rankings are also useful since the strength of players vary across different surfaces. The effect of the surface is, for instance, considered in the seeding system adopted for Wimbledon.

It is often shown in the literature that statistical methods are often capable of outperforming the ATP scoring system in terms of predictive ability. In the previous section, we have seen how our model produces significantly better predictions than the ATP ranking points and the actual ranking on all surfaces. In this section, we derive rankings based solely on the estimated strength of our model for the different surfaces. In particular, we can sort the players with respect to their strength level on each surface. The baseline strength is used to obtain an overall ranking. We note that the model-based rankings can be considered better than the ATP ranking to sort players in terms of ability but these rankings may lack other desirable features. We refer to Irons et al. (2014) for a discussion on how to construct tennis rankings using statistical models.

Table 4 reports the first ten players in the ATP ranking and the rankings obtained from our model on the 9th of February 2017. We note that there are some similarities but also some differences across the rankings. Seven out of ten players that are in the top ten of the ATP raking are also in the top ten of the baseline ranking. However, there are quite some differences in the order as, for instance, Novak Djokovic is first in the baseline ranking but only third in the ATP ranking. Concerning rankings for different surfaces, as expected, Rafael Nadal is better positioned in the clay ranking compared to the other rankings: 2nd position in clay but outside the top five in all other rankings.

ATP rank	**Baseline**	**Hard court**	**Clay court**	**Grass court**
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

Table 4: *First ten players in each ranking.*

In order 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 one if two rankings are the same and equal to minus one if two rankings are the same but reversed. Figure 5 presents the Kendall correlation between the different rankings. We can see that the ATP ranking is the ranking that is farther apart from all the other rankings with a correlation around 0.5. Focusing on the model-based rankings, we can see that the clay ranking 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 see that the ranking based on the hard court 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 dataset are played on hard court.

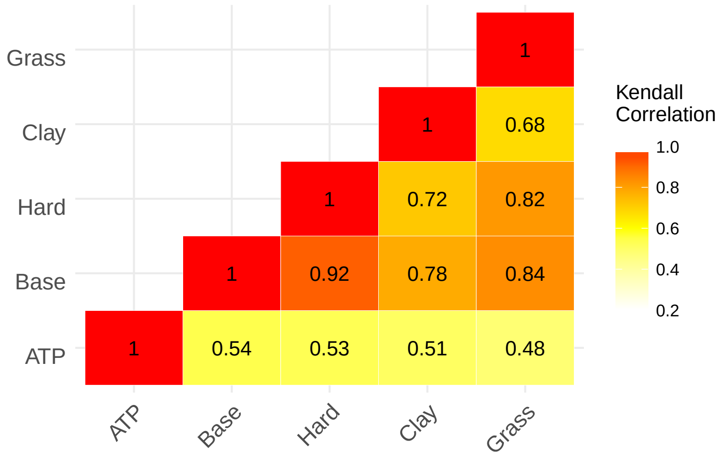

Figure 5: *Kendall correlation between the different rankings.*

4 Conclusion

We have introduced a novel approach to the analysis, modeling and forecasting of tennis matches. Our likelihood-based approach accounts for time-varying strengths and different court surface effects. The proposed modeling framework is able to describe several interesting features of tennis matches and it delivers accurate forecasts. The strength levels for different surface types that are extracted from our model can be used to construct improved rankings of players. These rankings are capable of reflecting the actual abilities of tennis players more accurately when compared to ATP points. These findings are confirmed in out-of-sample experiments. Surface-specific rankings can be useful for entry and seeding of tennis tournaments.

Appendix

A short review of score-driven time series models

Assume 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 has a conditional density function of the form

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

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

$$\lambda_{i,t+1} = \omega_i + \phi_i \lambda_{i,t} + \tau_i s_{i,t}, \tag{6}$$

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 defined by

$$s_{i,t} = S_{i,t} \nabla_{i,t}, \qquad \nabla_{i,t} = \frac{\partial \log p(y_{ij,t} | \lambda_t; \psi)}{\partial \lambda_{i,t}}, \tag{7}$$

with $\nabla_{i,t}$ being the score of the predictive likelihood and $S_{i,t}$ being 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. Alternatively, the scaling can be set equal to one and we can simply consider $s_{i,t} = \nabla_{i,t}$. A more detailed discussion is provided in Creal et al. (2013).

References

Baker, R. D. and McHale, I. (2014). A dynamic paired comparisons model: Who is the greatest tennis player? *European Journal of Operational Research*, 236(2):677–684.

Baker, R. D. and McHale, I. (2017). An empirical bayes model for time-varying paired comparisons ratings: Who is the greatest womens tennis player? *European Journal of Operational Research*, 258(1):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. *International Journal of Forecasting*, 32(3):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(2):325–343.

Boulier, B. L. and Stekler, H. O. (1999). Are sports seedings good predictors?: an evaluation. *International Journal of Forecasting*, 15(1):83–91.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Clarke, S. R. and Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International transactions in operational research*, 7(6):585–594.

Cox, D. R. (1958). The regression analysis of binary sequences (with discussion). *Journal of the Royal Statistical Society: Series B (Methodology)*, 20(3):215–242.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *journal of business and economic statistics*, 13:253–265.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164(1):130–141.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394.

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

Harvey, A. and Luati, A. (2014). Filtering with heavy tails. *Journal of the American Statistical Association*, 109(507):1112–1122.

Irons, D. J., Buckley, S., and Paulden, T. (2014). Developing an improved tennis ranking system. *Journal of Quantitative Analysis in Sports*, 10(2):109–118.

Koning, R. H. (2011). Home advantage in professional tennis. *Journal of Sports Sciences*, 29(1):19–27.

McHale, I. and Morton, A. (2011). A bradley-terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27(2):619–630.

Salvatierra, I. D. L. and Patton, A. J. (2015). Dynamic copula models and high frequency data. *Journal of Empirical Finance*, 30:120–135.