Informative prior distributions for a binomial model to predict professional tennis results

Pierre Colin, Aurélien Bechler

Journal de la Société Française de Statistique
Vol. 156 No. 2 (2015)

Informative prior distributions for a binomial model to predict professional tennis results

Titre: Prédiction des résultats de matchs de tennis professionnel par un modèle binomial avec des lois a priori informatives

Pierre Colin[1,2] and Aurélien Bechler[3,4]

Abstract: Tennis is a sport, as many others, that appears to be quite simple in the type of results (victory of one of the two players) but rather quite complex in factors that leads to this binary outcome. The perpetual evolution and increase of the way to collect data leads to more and more accurate available information about professional tennis matches. We studied the predictive properties of the binomial model representing the victory of one player against the other. Bayesian framework enables the updating of an informative prior distribution on the probability of winning (Beta distribution) by the collected information. After model calibration on the years 2011-2012, we test on the result 2013 of the ATP tour three methodologies for the choice of prior. The two firsts are based on latent variable models (Elo and Bradley-Terry). The third one is a point-by-point game simulation method based on the MatchFact statistics of the ATP. Each method is separated in two steps: specify the mean of the a priori distribution based on gathered data, and then its variance according to predictive characteristics. The second part of this article deals with possible uses of these methods for match result predictions, for whole tournament simulations or to propose a new ranking system for professional tennis players.

Résumé : Le tennis, comme de nombreux sports, a pour caractéristiques d'être à la fois simple dans le type de résultat obtenu (victoire de l'un des deux joueurs) et complexe dans les facteurs explicatifs de ce résultat. La collecte des données liées aux matchs de tennis professionnel ne cessant d'augmenter, l'information disponible est de plus en plus précise. Nous avons étudié les propriétés prédictives d'un modèle binomial représentant la victoire d'un joueur sur un autre. Le cadre d'inférence bayésien permet d'utiliser un prior informatif sur la probabilité de victoire (une loi Bêta) afin d'inclure cette information collectée. Nous avons comparé sur l'année 2013 du circuit ATP (et ajusté sur les années 2011-2012) trois méthodes de choix de prior. Les deux premières sont basées sur des modèles à variables latentes (Elo et Bradley-Terry). La troisième est une méthode de simulation de chaque point joué pendant un match reposant sur les statistiques MatchFacts de l'ATP. Chaque méthode est séparée en deux étapes : déterminer la moyenne de la loi a priori sur la base d'information collectée, puis sa variance sur la base des propriétés prédictives du modèle. La deuxième partie de cet article propose plusieurs utilisations possibles de ces méthodes, que cela soit pour la prédiction de matchs, de tournoi ou pour proposer un nouveau système de classement des joueurs.

Keywords: tennis, Bayesian, prior, binomial model, effective sample size, prediction, ranking

Mots-clés : tennis, bayésien, prior, modèle binomial, nombre équivalent d'observations, prédiction, classement

AMS 2000 subject classifications: 62F15, 91C15

[1] UMR518 Mathématiques et Informatiques Appliquées, AgroParisTech, Paris
[2] Statistical Sciences and Modeling, Sanofi R&D, Chilly-Mazarin
E-mail: pierre.colin@sanofi.com
[3] UMR518 Mathématiques et Informatiques Appliquées, INRA, Paris
[4] LSCE-IPSL, Centre d'Etudes de Saclay, Gif-sur-Yvette
E-mail: abechler@hotmail.com

 ISSN: 2102-6238

1. Motivating example: Tennis

Tennis is considered as an individual sport even if it can be played in double. This statement is more and more valid these past years, because of the increasing physical aspect of the modern sport which does not allow top players to compete in both single and double tournaments over the season. Thus in this article, single career is predominantly followed even if top players are occasionally engaged in both single and double draws of a same tournament. We aim to predict the victory of a player against another using a Bayesian model of the binary outcome of a tennis match. We especially focus on the ATP tournaments from 2011 to 2013.

One particularity of this sport (shared with others as tennis table or badminton) is that the scores are sequentially incremented, which is convenient from a modeling point of view. The winner is the first player who wins 2 sets (or 3 for the major tournaments) and a set is won by the player who reaches 6 games (with a gap of 2 games). If it does not happen, a tie-break is played at $6-6$. A game is won by the first player who reach 4 points (with a difference of 2 points as well). Nevertheless, tennis points are counted in a singular way with four possible scores: 15, 30, 40 and Advantage (origins of this way of scoring are unknown and are supposed to stem from a monetary system as well as from the palm game rules). Figure 1 details the score counting process.

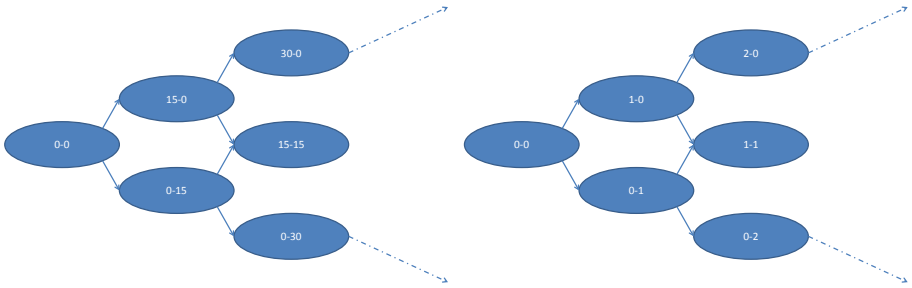


Figure 1: Illustration of a tennis match proceedings regarding on points (left) and games/sets (right).

In a unique game, only one player serves, i.e. hits the ball first in the exchange, until the game is over. His opponent will serve in the following game and so on. The order is defined by tossing a coin at the beginning of the match and will not change until the end. Conversely to other sports, serving is a advantage in tennis and the proportion of service game lost (called "break") is low. For example, in 2013, the percentage of service games won by Rafael Nadal (number one at the end of the season), Novak Djokovic (number two at the end of the season) and Roger Federer (number six at the end of the season) were respectively 88%, 88% and 87%.

This proportion depends on the surface (the fastest the surface, the easier for the server) and on the player characteristics. Indeed, in a season, tournaments propose different types of surfaces which can be summarize in 3 categories: hard court (most common), clay or grass (very limited). Each surface requests different players' abilities and could play a major role in the modeling of their performances.

Even if tennis is a very old sport (Quidet, 1976), the professionalization and the globalization

 ISSN: 2102-6238

of tennis lead, in 1972, to the creation of the Association of Tennis Professionals (ATP, 2014). It supervises the different tournaments all over the globe and builds the ATP ranking which is the official ranking of the professional tennis tour. This ranking has an informative role (e.g. designate the World best player at the end of the season) but also a real impact on the following tournaments (e.g. the seeds of almost every tournaments are chosen according to this ranking).

In this work, we studied the predictive properties of the binomial model representing the event of a victory of one player against another. The Bayesian framework enables the use of an informative prior on the probability of winning (through a Beta distribution) by taking into account the collected information. Different methods for the choice of the prior are discussed and compared in the paper.

This article is composed as follows: Section 2 details the model and the informative prior distributions, Section 4 presents the method comparison and different applications, Section 3 describes different prior settings, and Section 5 deals with the perspectives of the proposed method.

2. Modelling

2.1. Notations

In this article, we are mainly interested in the probability for a player i to win against a player j and the previous respective victories for player i and j. Later in this article, we use the following notations:

$$\begin{aligned} v_{ij} &= \text{Number of previous victories of Player } i \text{ against Player } j \\ R_{ij} &= \text{The random binary variable for the result of the match "Player } i \text{ against Player } j\text{"} \\ &= 1 \quad \text{if Player } i \text{ wins else } 0 \\ \pi_{ij} &= \mathbb{P}(\text{Player } i \text{ wins against Player } j) \\ &= \mathbb{P}(R_{ij} = 1) \end{aligned}$$

Since there draw cannot result in a tennis match, the variable R_{ij} is a binary outcome. As explained in Section 4, the quantities v_{ij} are calculated on 2 successive years.

2.2. Beta-Binomial model

We represent the random variable R_{ij} by a Bernoulli model with the probability π_{ij}. This model is fitted under the Bayesian framework that allows the incorporation of prior information. We use a conjugate Beta distribution for the prior probability π_{ij}, defined by the prior parameters α_{ij} and β_{ij}. We denote $[X]$ the density function of the random variable X.

 ISSN: 2102-6238

$$\begin{aligned}
R_{ij} &\sim \mathscr{B}ernoulli(\pi_{ij}) \\
\Rightarrow [R_{ij}|\pi_{ij}] &\propto \pi_{ij}^{R_{ij}} \times (1-\pi_{ij})^{R_{ij}} \\
\pi_{ij} &\sim \mathscr{B}eta(\alpha_{ij},\beta_{ij}) \\
[\pi_{ij}] &= \frac{\Gamma(\alpha_{ij}+\beta_{ij})}{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})}\pi_{ij}^{\alpha_{ij}-1}(1-\pi_{ij})^{\beta_{ij}-1}I_{[0;1]}(\pi_{ij}) \\
\alpha_{ij} &= p_{ij} \times (1+ess) \\
\beta_{ij} &= (1-p_{ij}) \times (1+ess)
\end{aligned}$$

The variable p_{ij} represent the *a priori* expectation. It can be set by different methods, detailed in Section 3, relying on a complete season of professional tennis tournaments (or several seasons). The variable *ess* is an arbitrary prior parameter. The *ess* interpretation is discussed in Section 2.3 and its value is discussed in Section 4.1. The posterior distribution $[\pi_{ij}|v_{ij},v_{ji}]$ is easily obtained thanks to the Bayes formula (Bayes and Price 1763, Eq. 1) and the conjugate property. We use the *a posteriori* expectation $\mathbb{E}(\pi_{ij}|v_{ij},v_{ji})$ as the prediction of the random variable R_{ij} (Eq. 2).

$$\begin{aligned}
[\pi_{ij}|v_{ij},v_{ji}] &\propto [v_{ij},v_{ji}|\pi_{ij}] \times [\pi_{ij}] \qquad (1)\\
&\propto \pi_{ij}^{v_{ij}}(1-\pi_{ij})^{v_{ji}} \times \frac{\Gamma(\alpha_{ij}+\beta_{ij})}{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})}\pi_{ij}^{\alpha_{ij}-1}(1-\pi_{ij})^{\beta_{ij}-1} \\
&\propto \frac{\Gamma(\alpha_{ij}+\beta_{ij})}{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})}\pi_{ij}^{\alpha_{ij}+v_{ij}-1}(1-\pi_{ij})^{\beta_{ij}+v_{ji}-1} \\
\Rightarrow [\pi_{ij}|v_{ij},v_{ji}] &\propto \mathscr{B}eta(\alpha_{ij}+v_{ij},\beta_{ij}+v_{ji}) \\
\Rightarrow \mathbb{E}(\pi_{ij}|v_{ij},v_{ji}) &= \frac{p_{ij}(1+ess)+v_{ij}}{(1+ess)+v_{ij}+v_{ji}} \qquad (2)
\end{aligned}$$

2.3. *Expectation and variability of the Beta prior distribution*

The expectation and the variance of the prior Beta distribution on π_{ij} can be expressed as follow:

$$\begin{aligned}
[\pi_{ij}] &\sim \mathscr{B}eta(\alpha_{ij},\beta_{ij}) \\
\Rightarrow \mathbb{E}(\pi_{ij}) &= \frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}} \\
&= p_{ij} \\
\Rightarrow \mathbb{V}(\pi_{ij}) &= \frac{\alpha_{ij}\beta_{ij}}{(\alpha_{ij}+\beta_{ij})^2(\alpha_{ij}+\beta_{ij}+1)} \\
&= \frac{p_{ij}(1-p_{ij})}{2+ess}
\end{aligned}$$

We can remark that a greater value of the *ess* parameter leads to a lower *a priori* variance of π_{ij} (and thus a more informative prior distribution). In Morita et al. (2008), this parameter is

 ISSN: 2102-6238

called the *Effective Sample Size*. It corresponds to an equivalent number of previous observations (i.e. number of head-to-head matches between player i and player j) needed to obtain a similar distribution of π_{ij} (according to a non-informative prior distribution). Figure 2 represents the effect of different *ess* values on the relationship between the prior expectation $\mathbb{E}(\pi_{ij})$ and the posterior expectation $\mathbb{E}(\pi_{ij}|v_{ij}, v_{ji})$ for different values of (v_{ij}, v_{ji}): 0-0, 1-0, 0-1, 1-1, 2-0, 2-1, 2-2, 1-2 and 0-2. The higher the *ess* value is, the more informative the prior distribution is. The *ess* value is discussed in the Section 4.1. In following sections, we discussed of different strategies to determine the parameters p_{ij}. The two first methods are drawn from the pair comparison context. The third one is based on a mechanistic representation of a tennis match.

3. Setting the prior

3.1. Prior expectation based on Elo method

The first studied method is the Elo ranking, Elo (1978). This method has been proposed for the ranking and prediction of chess matches. It is based on a trinary outcome for each match, but it can be easily adapted to the binary case of a tennis match. The aim of this method is to update, after each match, an individual score S_i^t (relative to a given player) according to the result of the match. Then the difference Δ_{ij}^t, between two individual scores, is used as a predictor of the expected prior probability, p_{ij}, for Player i to win against Player j (Eq. 3), thanks to a link function similar to the logistic model (Eq. 4).

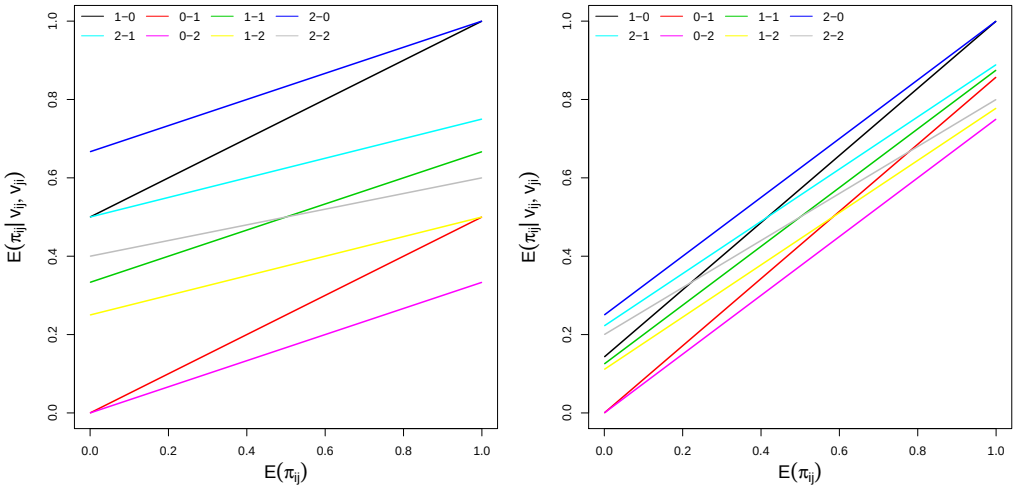
$$\begin{aligned}
\Delta_{ij}^t &= S_i^t - S_j^t && (3)\\
p_{ij}^t &= \left(1 + 10^{-\Delta_{ij}^t/400}\right)^{-1} && (4)\\
S_k^{t+1} &= S_k^t + K \times (W^t - p_{ij}^t)\\
W^t &\in \left\{0 \ \text{(lose)}, 1 \ \text{(win)}\right\}\\
K &= 27 \quad \text{(Other values are possible)}\\
p_{ij} &= p_{ij}^T
\end{aligned}$$

t represents the index of the match used to update the Elo score. It takes value from 1 to T. Since this method provides sequence of probability to win, $(p_{ij}^t)_t$, the last updated probability (p_{ij}^T) as the prior expectation. It is important to notice that the updated value of S_i^t does not correspond to any optimized statistical criterion (likelihood, prediction score...). It is obtained by an algorithmic rule, Elo (1978). This algorithmic rule can be considered as an approximation of the maximum likelihood estimator. This algorithmic method is usually used to update a latent score after each match, without taking into account some others performed matches. All analyses of the Elo method are performed with the R software and the R-package PlayerRatings (R Core Team, 2013; Stephenson, 2012).

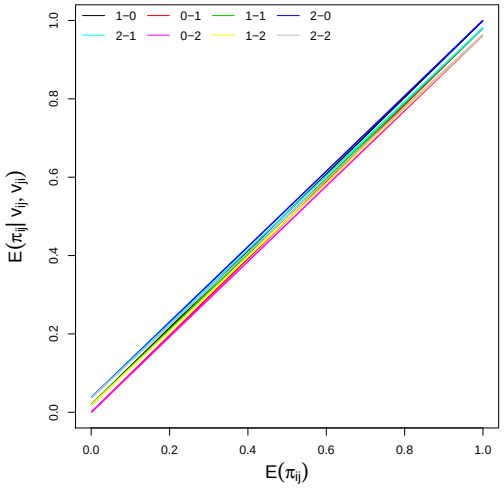
3.2. Prior expectation based on Bradley-Terry method

The Bradley-Terry method, (Bradley and Terry, 1952; Luce, 1959), similar to Elo method, since the expected prior probability p_{ij} is predicted by the difference between two individual scores,

 ISSN: 2102-6238



(a) $ess = 0$ (b) $ess = 5$



(c) $ess = 50$

Figure 2: Effect of the *a priori* expectation on the *a priori* expectation, regarding on the *Effective Sample Size* (*ess*) values.

using a logit link function (Eq. 5). The Elo method can be viewed as an approximation of the Bradley-Terry method. However conversely to the Elo method, these scores are here defined up to a constant and are estimated through a likelihood maximization (Eq. 6). We can notice that the individual scores are considered constant for all head-to-heads between two players, although it was considered as dynamic variables in the Elo method. All analyses of the Bradley-Terry method are performed with the R software and the R-package PlayerRatings (R Core Team, 2013; Turner and Firth, 2012).

 ISSN: 2102-6238

$$\begin{aligned} \delta_{ij} &= \lambda_i - \lambda_j \\ p_{ij} &= \left(1 + e^{-\delta_{ij}}\right)^{-1} \qquad (5) \\ \{\lambda_k\} &= \arg\max_{\lambda_k} \prod_i \prod_{j>i} \left(p_{ij}\right)^{v_{ij}} \left(1 - p_{ij}\right)^{v_{ji}} \qquad (6) \end{aligned}$$

3.3. *Prior expectation based on a Markov process using ATP MatchFacts statistics*

To represent a tennis match performance, we propose a Markov model based on MatchFacts statistics (Table 1) made available by the ATP website. A tennis match can be considered as a Markov process since the value of the score during the match relies only on the previous state of the Markov process (previous value of the score) and the probability of transition between two states (the probability to win a point). If one can simulate a tennis point, we assume that one can simulate a whole tennis match, since these simulations are directly obtained by simulating tennis points and using the corresponding score (according to tennis rules, International Federation of Tennis 2014). We propose to simulate a tennis point as follows:

1. Simulate the 1st or 2nd serve of the server player (the first player to serve is drawn at the beginning of the match from a Bernoulli distribution with a probability 0.5). The variable 1st serve is drawn from a Bernoulli distribution of probability p^k_{s1}.
2. Simulate the event "tennis point is won by the server player" from a Bernoulli distribution with the probability $p^k_{sw1}/(p^k_{sw1} + p^{k'}_{rw1})$ if the first serve is performed and with the probability $p^k_{sw2}/(p^k_{sw2} + p^{k'}_{rw2})$ if not.
3. If the tennis point is a break point, the event "tennis point is won by the server player" is drawn from a Bernoulli distribution with the probability $p^k_{bs}/(p^k_{bs} + p^{k'}_{bc})$.

We can now simulate a large number of tennis matches opposing two players. These simulations are performed through a C++ program. The simulated ratio of victories of Player i is used as the expected prior probability p_{ij}. All Markov process simulations are performed 100.000 times for each tennis match. The probabilities p^k_{xx} are specific to each surface of game.

TABLE 1. *ATP MatchFacts statistics available on ATP website. The variables p^k_{xx} are the probabilities for the player k to perform the event xx.*

Service Record		Return Record	
% 1st Serve	p^k_{s1}	% 1st Serve Return Points Won	p^k_{rw1}
% 1st Serve Points Won	p^k_{sw1}	% 2nd Serve Return Points Won	p^k_{rw2}
% 2nd Serve Points Won	p^k_{sw2}	# Break Points Opportunities	
# Break Points Faced		% Break Points Converted	p^k_{bc}
% Break Points Saved	p^k_{bs}	# Return Games Played	
# Service Games Played		% Return Games Won	
% Service Games Won		% Return Points Won	
% Service Points Won		% Total Points Won	
# Aces			
# Double Faults			

 ISSN: 2102-6238

4. Comparison and application

We now present the comparison of the three detailed methods of prior choice and different applications of the predictive model and the Markov process.

4.1. ESS determination and a validating year

First of all, we defined two separate datasets. The first one corresponds to the tennis matches played in 2011 and 2012 and the second one to the ones played in 2013. All tennis matches corresponding to disqualification, by-pass or withdrawal have been removed from both datasets (cf Table 2). Only the training dataset is used to performed the Bayesian inference. We check the predictive characteristic of the model on the first seen dataset. We do not use tennis matches performed before 2011 since the player characteristics may change by years. We use the training dataset to determine the prior expectation p_{ij} for every match of the first seen dataset, according to the methods detailed in Section 3. Then, after choosing the *ess* parameter, we will be able to predict the tennis matches from the first seen dataset.

TABLE 2. *Datasets used for the ESS determination ann the cross-validation.*

Data	Training dataset	First seen dataset
# of played matches	5947	2866
# of players	378	303
Withdrawal of matches concerned by disqualification, by-pass or withdrawal		
# of retained matches	5326	2529
# of players	377	232
Maximal # of face to face	10	6

Now, we have to determine the value of the Effective Sample Size. This value will be specific to each method of prior elicitation and will be unique for all matches from the first seen dataset. For each ess value from 0 to 50 (integer values), we compare the prediction performance of the binomial model (for each prior distribution) according to the area under a ROC curve (Receiver Operating Characteristic), for which the *a posteriori* expectation is used as a predictor. The results are presented by Figure 3. A similar analysis performed with the score, Brier (1950), leads to similar results (not shown here). We choose, for each method of prior elicitation, the ess value that maximizes the area under the ROC curve. This value provides the ess corresponding to an acceptable trade-off between an informative prior distribution and the assessment of the variability of the observed data (from the training dataset). An infinite value of the ess would imply that the prior expectation is a perfect predictor. It is remarkable that the ess value of the Markov prior distribution is (twice) higher than the ess values of the Elo method and Bradley-Terry method. It means that provides a prior highly more informative than the two other techniques and that should unsurprisingly yield better predictive characteristics of the binomial model based on the Markov prior. The optimal ess value of the Markov prior is equal to 10. This means that the prior information is equivalent to 10 face-to-face matches between any two players, which is the maximum number of face-to-face matches observed on 2011 and 2012 (between Rafael Nadal and Novak Djokovic). Even in this most informative case, the Markov prior provides as much information as the real data.

 ISSN: 2102-6238

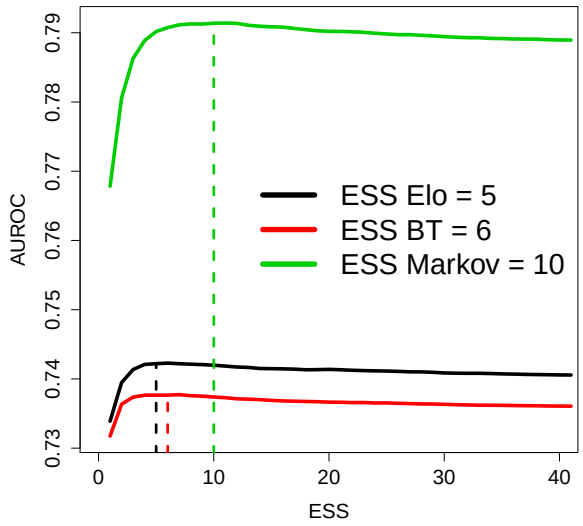


Figure 3: Comparison of the model predictive characteristics for each prior distribution and each ess value. For each prior, the selected ess value is the one that maximize the area under the ROC curve.

Since the training data consists of matches performed in 2011 and 2012, the model parameters may be updated sequentially as 2013 results become available. It may be interesting to evaluate the robustness of the model predictive characteristics according to the increasing delay from the data collection. We used a smooth function to estimate the percentage of correct prediction across the year 2013 (Fig. 4). We can notice that the performance of the Markov method remains the highest one. Elo and Bradley-Terry methods provide a better predictive characteristic for the higher delay than for the medium delay. This is due to the fact that both methods provide average predicted probabilities. The match surface is not used as a predictor variable. Thus Hard surface results eventually make it for this averaging since they correspond to most of the tournaments. Therefore, these hard surface tournaments are performed at the beginning and at the end of the year. This calendar and the specificity of the surface may explain the increasing predictive characteristic of Elo and Bradley-Terry method for a delay higher than 200 days. The Markov method uses the surface as a predictor variable (which increases its predictive capacity). Thus this characteristic is monotonically decreasing with the delay from the data collection. For a gambling practice, the three methods might be easily updated every week to keep a high level of performance.

4.2. A new score for ranking

As mentioned in Section 1, ATP ranking has a major role on the tournaments because the top-seeded players are chosen according to this ranking. This system prevents top players from playing against each others in the first rounds of a tournament. This protection according to the best players directly impacts the tournament and raises some questions about the relevance of the

 ISSN: 2102-6238

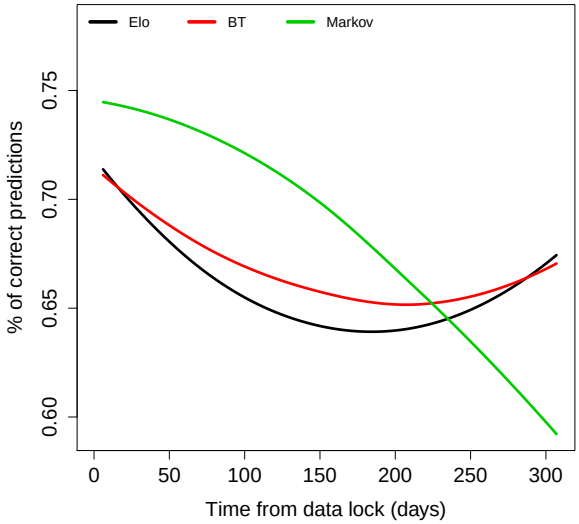


Figure 4: Evolution of the predictive characteristics over time.

ATP ranking. The rank is calculated by taking into account the results of a player in some given tournaments (for the best players essentially Grand Slam, Master 1000, ATP 500 and ATP 250 tournaments) during the last 52 weeks.

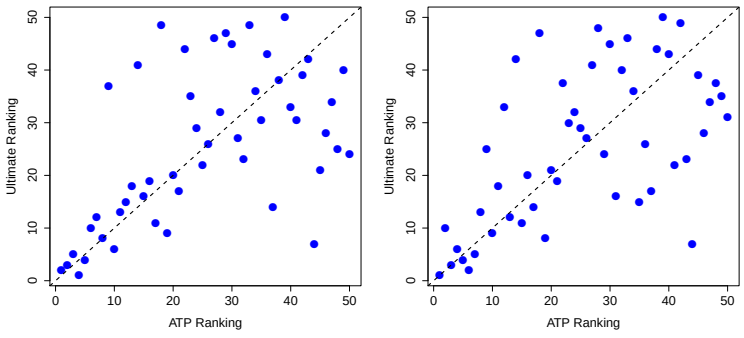
We propose an alternative ranking system based on the Markov methodology presented in this paper and on the notion of "Ultimate Player". Hence, we build the best possible player by fusioning the best statistics (1) from every real player. For example, this "Ultimate Player" will have the best percentage of first serves (or the best return game) among the player in the top 100 ATP and it is the case for each statistic of this "Ultimate Player". Then, by simulating, using the Markov methodology (detailed in section 3.3), a large number of matches between all players and this "Ultimate Player", we estimate the probability of victory of every real players against this "Ultimate Player". By construction, this probability is lower or equal to 50%.Then, this score is used in order to establishing a new ranking among the players. It is also important to note that the characteristics of the "Ultimate Player" change in function of the surface.

The figure 5 displays the difference between the ATP ranking and the "Ultimate Player" ranking (UP ranking) for each surface.

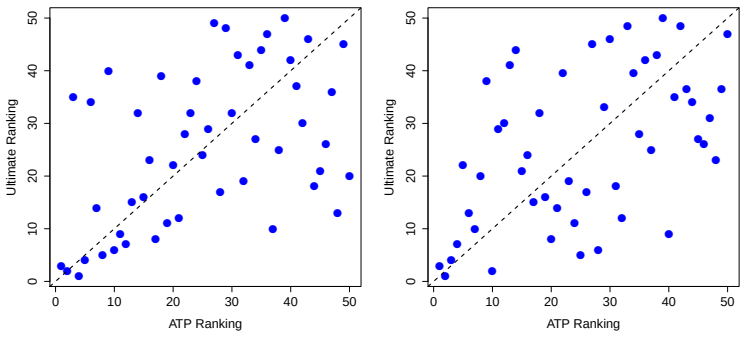
First, we can observe that the two ranking systems present different result (correlation coefficients are respectively equal to 0.55(a), 0.58(b), 0.47(c) and 0.50(d)) even if best players are approximately at the same rank. Then it also appear that the coefficients change from a surface to another which is quite obvious because of the difference of tournament played on the different surfaces. For example, on hard court, on which almost half of the tournaments of the season are played, the coefficient is better than on grass court.

This UP ranking presents a new ranking system which seems relevant because it takes the surface of a tournament into account, which is not the case for the current ATP ranking. For example, Wimbledon which is the only major tournament on grass has the right of changing the

 ISSN: 2102-6238



(a) All (b) Hard



(c) Clay (d) Grass

Figure 5: Ultimate ranking based on the ultimate player compared to the ATP ranking for each court surface.

order of the seeds given by the ATP. This specificity, which comes from the rarity of the surface, raises the issues of using a ranking based on the player results on other surfaces. It can be useful to use the UP ranking for designating the seeds of a tournament, specially for those played on grass and clay.

4.3. Tournament simulations

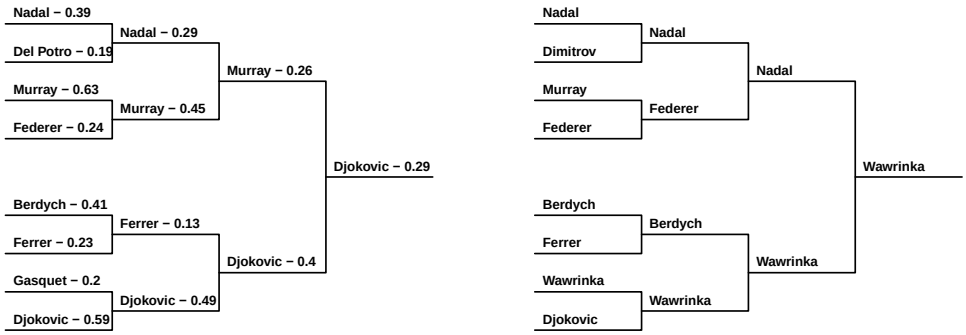
One of the many applications of the match prediction by the binomial model is the possibility of simulating a whole tournament results. For this application, we choose to use the informative prior obtained thanks to the Markov method. By taking into account the draw of the tournament, each match is simulated once until the final. This procedure is repeated a large number of times (50000). Thus, we obtain 50000 different trees, each one starts from the real draw and diverges after. At each step, we take the player which appears the most often.

It has to be noted that this approach does not consists of comparing the probability of victory in each round of the tournament to determine which player wins. Our approach by simulation relying on the complete Bayesian posterior predictive distribution takes into account the various sources of uncertainty and provides many different scenarios and enables us to evaluate the probability of

 ISSN: 2102-6238

each player for reaching a given round. This probability depends on the player level but also on the difficulty of his path within the tournament (so on the draw).

Figure 6 shows simulations of the AUS Open (on hard courts, in January 2014) compared to what happened in the reality (from the quarter-final to the final). It appears that simulations present plausible results (6 on 8 were indeed present in quarter final) and that the top seeded players are massively represented in the final 8 players (7 on 8). It reaffirms the fact that on hard courts, the Markov model gives a hierarchy close to the official ATP one.



(a) Predicted (b) Observed

Figure 6: Tournament simulations of the Australian Open. This graph represents the most probable results from the Quarter-Finals to the Winner (left) and the obsersed results (right).

5. Discussion

In this work, we proposed a Bayesian methodology to predict the outcome of a professional tennis match. The originality of our approach is that the most of the information and the modeling are brought by the prior information. In that purpose, we compare three different methods, respectively based on the ELO ranking, on the Bradley-Terry method and on the vision of the succession of points as a Markov chain. For the three methods, the posterior distribution has been updated by taking into account the direct observed confrontation(s) between the two players. If the elicitation of the ess value is challenging, one can use the prior expectation as the predictive probability. The application of these methodologies on real data points out that the Markov method is the most accurate, partly thanks to its ability of taking the surface into consideration.

Two particular application arise from this method, a whole tournament simulations and the proposition of a new ranking system. In terms of perspectives, the Markov method could be enhanced by different additional information. Likewise the Break points, the TieBreak points could be considered different from the ordinary game points, according to a specific information (e.g. the percentage of TieBreaks won by each player). ATP statistics do not involve the difference between the Indoor hard surface and the Outdoor surface. This distinction could be performed and may improve the predictive characteristics of the proposed model. The Markov method is based on a strong hypothesis: all points of the match are independent, according to the players'

 ISSN: 2102-6238

characteristics. This is not consistent with the persistency that can be observed during of a tennis match since a player may win successive points during a key period to be explained by stress or temporary physical advantage for instance. Finally, it may be interesting to evaluate the predictive characteristics of the Markov approach for sub-events (e.g. number of sets, number of games in the first set, occurrence of TieBreak, ...).

Once the predictive characteristics of a model are checked, a large number of applications can be worth considering. Sport betting is interesting for the gambler point of view as well as from the bookmaker point of view. Such a model could be used to understand some gambling behaviors (bias due to a favorite or an outsider player, bias due to the nationality, ...). The model could be used to support the professional management (tournament selection, recovery management, ...) or to enhance the anti-doping or anti-corruption programs.

Acknowledgements

The authors thank Éric Parent, Jean-Louis Foulley, Anne Gégout-Petit and Jean-François Toussaint for their support and advice.

References

ATP (2014). Atp world tour history, http://www.atpworldtour.com/Corporate/History.aspx.

Bayes, M. and Price, M. (1763). An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions*, 53:370–418.

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, 78(1):1–3.

Elo, A. (1978). *The Rating of Chessplayers, Past and Present*. Arco Pub.

International Federation of Tennis (2014). *ITF Pro Circuit regulations*. http://www.itftennis.com/media/163754/163754.pdf.

Luce, R. D. (1959). *Individual Choice Behaviours: A Theoretical Analysis*. Wiley.

Morita, S., Thall, P. F., and Müller, P. (2008). Determining the Effective Sample Size of a Parametric Prior. *Biometrics*, 64(2):595–602.

Quidet, C. (1976). La fabuleuse histoire du tennis.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Stephenson, A. (2012). Rating australian rules football teams with the playerratings package. *R vignette*.

Turner, H. and Firth, D. (2012). Bradley-terry models in r: The bradleyterry2 package. *Journal of Statistical Software*, 48(9):1–21.

 ISSN: 2102-6238