Tennis betting: Can statistics beat bookmakers?

Article *in* Electronic Journal of Applied Statistical Analysis · September 2013

DOI: 10.1285/i20705948v10n3p790

2 authors, including:

Francesco Lisi
University of Padova
57 PUBLICATIONS **588** CITATIONS

SEE PROFILE

Electronic Journal of Applied Statistical Analysis
Vol. 00, Issue 00, August 2013, 1-35
DOI: DOI–xxxx

Tennis betting: can statistics beat bookmakers?

Francesco Lisi*and Germano Zanella

University of Padua, Department of Statistical Sciences, Via Battisti, 241 - 35121 Padua, Italy

March 5, 2017

We propose a logistic regression model to predict the win probability in a tennis match in the context of the betting market. The variables included in the model are ATP points and rankings, the players' ages, the home factor and the information derived from bookmaker odds. The model is estimated using data related to 2012 tournaments, and it is then used in an out-of-sample betting experiment where the odds implied by the model are used, following a specific procedure, for betting against bookmakers. The algorithm is applied to all matches of the four Grand Slam Championships 2013, and the whole procedure is evaluated with respect to the global return of the strategy. After 501 matches, the total cumulative return is 16.3%.

keywords: Tennis models, betting market, odds, logistic regression.

1 Introduction

The use of sports analytics has quickly increased in recent years. As a result, several authors have proposed models and methods to analyse and forecast sporting events, as published in the many papers of the Journal of Quantitative Analysis and Sports and, more specific to tennis, in Klaassen and Magnus (2014), who summarise the results they obtained in many previous papers, Chitnis and Vaidya (2014), Koning (2011), Clarke and Dyte (2000), Forrest and McHale (2007), Nadimpalli and Hasenbein (2013), Knight and O'Donoghue (2012).
Two main approaches were used in the literature to forecast the outcome of a match: one based on expert evaluations and one based on more formalised, i.e. statistical, models. The number of works considering models for tennis matches and forecasting outcomes

*Corresponding author: Francesco Lisi: francesco.lisi@unipd.it

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

is few with respect to other sports, and works showing applications of these models to the betting market are even rarer.

In the context of tennis win predictions, according to Kovalchik (2016), three categories of statistical models can be identified: regression, point-based and paired comparison models. Regression models directly model the winner of the match; most are logit or probit models (del Corral and Prieto-Rodríguez, 2010; Gilsdorf and Sukhatme, 2008; Klaassen and Magnus, 2003; Boulier and Stekler, 2003). Point-based approaches model the win probability as the probability of winning a single point, often on serve (Knottenbelt et al., 2012; Spanias and Knottenbelt, 2012; Barnett and Clarke, 2005). Paired comparison models account for the players' abilities (McHale and Morton, 2011). Different approaches not based on statistical models include that of Scheibehenne and Bröder (2007), which predicts the outcomes of matches by mere player name recognition, that of Leitner et al. (2009), which is based on bookmaker consensus, and that of Easton and Uylangco (2010), which uses probabilities implied by betting odds.

In this paper, we build a model for predicting the probability that a player wins a match using several sources of information. The focus, however, in not on the prediction of the final outcome itself but on the possibility to use it in a rewarding way in the betting market. Our approach is based on a logistic regression model, which is not new. The distinguishing features of this work are i) the variables entering the model and ii) the use of the model in a betting procedure and the evaluation of its economical impact.

In the first part of the work we will build the model step-by-step considering different variables: ATP points and ranking, the players' ages, the surface of the courts, the home factor and the information derived from the bookmaker odds.

In the second part of the paper, the identified model will be used in an out-of-sample betting experiment, where the odds implied by the model enter an algorithm for betting against bookmakers. We apply our algorithm to the four Grand Slam 2013 matches and the whole procedure is then evaluated with respect to the global return of the strategy. Thus, the emphasis here is not on the percentage of correct predictions of match outcomes but rather on the return obtained from betting using the model. Of course, these two aspects are not independent; however, because in betting there is often an asymmetry between the amount of money one can win or lose, a crucial issue is to avoid betting on matches that are too uncertain. Similarly, it is important to find a strategy requiring the investment of as little money as possible, that is, to bet on a few matches as possible.

The rest of paper is organised as follows: Section 2 describes the odds and the logistic regression model for odds computation. Section 3 analyses data and regressors variables and identifies the model that will be successively used. Section 4 compares the probabilities and the odds produced by the model with those quoted in the betting market. Section 5 contains the betting procedure will be applied to the Grand Slam 2013 matches and analyses the obtained results. Section 6 provides the conclusions.

2 The odds and the model for odds

An odd o is the amount of money paid for a winning unitary bet. It is clear that the lower the ex-ante probability of winning the bet, the higher the odd itself. The implied probability associated with an odd o is simply $p=1/o$. On the other hand, if one starts from the ex-ante probability p of winning the bet, the fair odd is $o=1/p$. Thus, when two players are equally likely to win ($p = (1-p) = 0.5$), the fair odds would be $o = 2.00$ for both players. However, to make an overall profit regardless of the result of the match, bookmakers usually overround the win probabilities for both players, and this translates to odds that are not completely fair, such that their sum is greater than one. If, in the previous example, the overround is 8%, the probability $p = 0.5$ would become $p = 0.54$ and the odds $o = 1.85$. When the probability of a player winning is high, i.e $p > 0.92$, an 8% overround would imply an odd less than one, which is not reasonable, because a win would receive less than the invested money. In these cases, bookmakers quote a small and unattractive odd, such as $o = 1.01$. To be cautious, they usually also put an upper threshold on odds. Indeed, even if a player has a low win probability, say $p = 0.05$, the implied odd ($o = 20$) would be unsafe for bookmakers, because the possibility of a defeat or an injury to the favorite player always exists. Therefore, they quote an interesting but not too high odd (Graham and Stott, 2008; Skiena, 2004).

The models used for forecasting the win probability, as described in the Introduction, are also used to fix odds. Here, we consider a simple logistic regression model that allows us to estimate the win probability as a function of a set of variables related to the features of a match.

Given a match, we define as 'favorite' the player with the highest ATP ranking and as 'unfavorite' the other player, and we denote p as the win probability of the favorite player and $q = (1-p)$ for the unfavorite player. Thus, the random variable describing the outcome of the favorite player is distributed as a Bernoulli, with probability p. To model p conditionally to a (sub)set $\boldsymbol{x} = (x_1, ..., x_k)'$ of k explicative variables, we refer to a logistic regression model, implying a linear logit

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i, \tag{1}$$

where $p(\boldsymbol{x})$, the conditional probability, is such that

$$p(\boldsymbol{x}) = \frac{exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}{1 + exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)} \tag{2}$$

with $\boldsymbol{\beta} = \beta_0, \beta_1, ..., \beta_k{}'$ vector of constant parameters that can be estimated by a maximum likelihood procedure.

Given $p(\boldsymbol{x})$, the implied fair conditional odd is $o(\boldsymbol{x}) = p(\boldsymbol{x})^{-1}$. However, in our experiment, to make the odds comparable to those of the bookmakers, we apply to the model's probability a 6%-overround[1], so the odd actually used becomes $o(\boldsymbol{x}) = (1.06 * p(\boldsymbol{x}))^{-1}$.

[1]This value is based on the mean overround of some bookmakers. For example, in 2012 the mean overround was 6.2% for Bet365, 6.6% for Expekt, 6.7% for Ladbrokers, 2.3% for Pinnacles Sports

3 Building the models

To find the most appropriate model for quoting odds, we consider a data set consisting of 1081 matches played in 2012 and referring to four Grand Slam championships (Australian Open, Roland Garros, Wimbledon and US Open), nine ATP World Tour Masters 1000 (Indian Wells, Miami, Monte Carlo, Madrid, Rome, Toronto, Cincinnati, Shanghai and Paris-Bercy) and the ATP World Tour Final, also known as the Masters Cup. The ATP 500 and 250 tournaments, as well as Challenges, the Davis Cup and the Olympic Games, were not considered.
Data, provided by www.tennis-data.co.uk[2] include the name and the location of the tournament, the game surface, the ATP ranking and the ATP points of both players the week before the match and the ages of the players for each match. They also include the pre-match odds of seven bookmakers.
These pieces of information can be thought of as primitive variables, and they will be considered individually to build the variables to be used as regressors in model (1).

3.1 ATP points

The ATP ranking includes the players that participated in at least one international tournament in the last year, and it is updated weekly to include the most recent outcomes. Indeed, each tournament assigns points to players depending on the qualification round reached and on the kind of tournament[3]. Points last for a year, so that in each moment the ATP points of a player are the sum of the points won in the last 52 weeks. ATP scores then produce the ATP ranking.
Although several models in the literature consider the ATP raking the most important covariate (see references in the next subsection), in our analyses we follow the approach of Clarke and Dyte (2000), who use ATP points instead of ranking. This variable was also considered in McHale and Morton (2011). As ATP points define the ranking, they include all the information contained in the ranking, but they also establish a sort of distance among players that is more accurate than ranks.
Actually, we do not use ATP points directly, but we consider the variable $\Delta Points$ given by the difference[4] between the points of the favorite player (PF) and those of the unfavorite player (PU) the week before the match, namely $\Delta Points = PF - PU$. Row 'Mod1' of Table 2 lists the output of the estimation step. Both the constant and ΔP are highly significant. As expected, the coefficient of $\Delta Points$ is positive.

and 6.7% for Stan James

[2] It is the same source of data of several other papers, such as Kovalchik (2016), McHale and Morton (2011) and Scheibehenne and Bröder (2007), among others.

[3] Grand Slam tournaments assign 2000 points to the winner, the so-called Master 1000 tournaments assign 1000 points, the 500 Series 500 points and the 250 Series 250 points

[4] We also considered the absolute values of the two ATP scores, but this proved less efficient.

3.2 The ATP ranking intervals

In the literature, almost all models consider, directly or indirectly, the ATP rankings of the two players, and this appears quite natural (del Corral and Prieto-Rodríguez, 2010; Klaassen and Magnus, 2003; Clarke and Dyte, 2000; Boulier and Stekler, 2003, among others). This variable is highly correlated with ATP points, but the variables given by the difference in points and in ranking are much less correlated. To further reduce the correlation, we suppose the existence of zones or intervals in the ATP ranking within which there is some kind of homogeneity among the players. The underlying idea is that the technical difference is larger among high-ranked players than among medium-ranked, or even more low-ranked, players. This issue was considered, in a different way, also by Klaassen and Magnus (2003).
This raises the issue of how to choose intervals. One possibility is to choose them arbitrarily, exploiting the considerations of sports analysts and the knowledge of tennis. In this context, a reasonable classification could be given by the following five intervals of the ranking: 1–10, 11–30, 31–100, 101–300, >300.
This kind of classification, however, could be questionable and it is completely independent of the data. To have a classification based on sample data we analyse the empirical distribution of ATP points. The left panel of Figure 1 shows the histogram of the ATP points for the first 200 players in January 2014. It can be seen that only a few players (15) have more than 2000 points, whereas most (185) have less than this threshold. The right panel shows the histogram limited to players with less than 2000 points. A reasonable

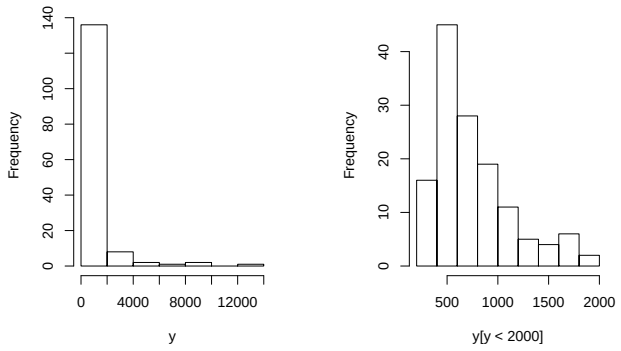

Figure 1: Left: distribution of ATP points for the first 200 players of the ATP ranking, in January 2014; Right: the same for points< 2000.

classification, obtained by looking at Figure 1, is shown in the 'histogram' row of Table 1. Despite its good sense for people knowing tennis, it is yet a subjective clustering. To find a more objective classification, we used the cluster analysis to fix the intervals. In particular, we applied a hierarchical cluster analysis to the data, where the classification variable was the number of points, and we fixed the number of clusters equal to five, maintaining the threshold of 2000 points for the last class. The ranking intervals we

Table 1: Ranking classifications, based on the histogram and on the cluster analysis.

Method	I interval	II interval	III interval	IV interval	V interval
Histogram	0-400pt	400-800pt	800-1200pt	1200-2000pt	>2000pt
Clusters	0-560pt	560-920pt	9200-1460pt	1460-2000pt	>2000pt

derived are shown in the 'Clusters' row of Table 1: one can see they are not so different from those obtained by a subjective interpretation of the histogram. Once the ranking intervals have been defined, and denoting by FI and UI the intervals corresponding to the favorite and unfavorite players, we considered the variable $\Delta Int = FI - UI$. In our case, ΔInt assumes values $0, 1, 2, 3$ and 4 and, when included in the model, it has a positive coefficient and is largely significant (column 'Mod2' of Tab. 2). A comparison of this model with the previous one by means of the Akaike criterion (AIC) suggests the inclusion of ΔInt improves performance.

3.3 Age

A professional player's career usually begins around age 10–12, typically entering the international ranking around age 17–20, and it usually ends around age 34–35[5]. Especially in singles, age can be an important factor. A younger player usually has a better physical performance and more resistance, but more mature players often are more experienced and have better strategic views. To try to understand the relationship between age and performance (measured by the ranking), we analysed the ranking positions of the first 150 players and their (rounded) ages. Figure 2 shows the scatterplot of these two variables. The full line represents the fitted curve of a quadratic regression[6] of position over the age, and it shows that players have, on average, the best rank at around age 27. The (difference of) age was considered also by del Corral and Prieto-Rodríguez (2010) who also find a quadratic relationship. Accounting for age, however, is not so simple. Indeed, using several forms and transformations of this variable was not significant or led to very marginal improvements. In particular, we considered the ages of both players (AIC: 1182.8), the difference between their ages (AIC: 1183.4), the squared ages of both players (AIC: 1185.3), both ages and squared ages (AIC: 1183.1), the age of the favorite player and its square (AIC: 1182.8), the age of the unfavorite player and its square (AIC: 1182.2), the age of the favorite player (AIC:1181.1). Apart from the last case, all these attempts showed the lack of significance of the variable.
Row 'Mod3' of Table 2 shows the estimated model when the age of the favorite player ($FAge$) is included. The estimated coefficient of $FAge$ is negative, meaning that in-

[5] Of course, there are some exceptions; for example Rafael Nadal entered the international ranking at 15 and Jimmy Connors played a semi-final of the US Open at 41.
[6] A nonparametric regression led basically to the same parabolic relation.

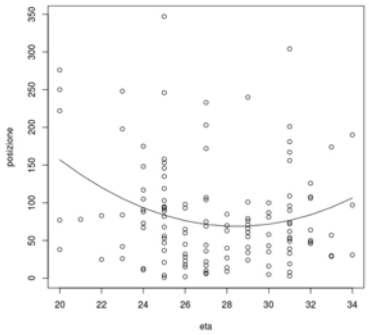

Figure 2: Scatterplot of the age *vs* the ranking position. The full line represents the estimated relation, using a quadratic polynomial of the age.

creasing the age reduces the win probability. Also the AIC improves.

3.4 The surface

Tennis can be played on different courts. As each surface has specific characteristics, tournaments can also be classified according to the surfaces of their courts. With some level of approximation, we can say there are three main kinds of surfaces: clay, hard and grass. Each of these surfaces tends to favor players with specific characteristics, and, usually, each player has a preferred surface, that exalts his own technical characteristics. Although the idea that the surface type can be an important factor in the final outcome, finding a data-based criterion to identify the best surface for each player is not so simply. We could count the number of successful matches on different surfaces, but this is influenced also by the number of tournaments on a specific surface. For example, there are fewer championships on grass than on clay and fewer on clay than on hard surfaces. Some authors indirectly consider the surface, building a model for each surface. In this work, we considered only two surfaces: 'fast' including hard, grass, carpet and acrylic, and 'slow', referring to clay. Then, for each player, we defined the preferred surface as that with the higher percentage of winning matches along the entire career. This allows us to consider the dummy variable S_x, assuming value 1 when the surface on which the match is played is preferred by the player x and 0 otherwise. To account for the preferences of both players, we then build the variable $Sur = S_F - S_U$, where the subscripts F and U refer to the favorite and unfavorite players, respectively. Sur assumes only three values: 1 if only the favorite player plays on his favorite surface, -1 if the same is true for the unfavorite player, and 0 when the surface is or is not preferred by both players. As we are modeling the win probability of the favorite player, the expected coefficient is positive. However, the variable Sur was not significant, and it did not lead to any improvement (AIC=1186.1).
A possible explanation is that the surface might be important only when the technical

difference between the two players is not too big. To verify this hypothesis, we considered a new variable $Sur2$, such that $Sur2 = Sur$ if the difference between the two players in terms of ATP points is less than 400[7] and 0 if otherwise. This new variable is significant at 5% of significance level, but the AIC criterion does not improve (see column 'Mod4' in Tab. 2).

3.5 The home factor

In the context of international sports, home advantage assumes that competitors will perform above their expected level when competing in events held in their own country. For several sports, particularly for team sports, to play 'at home' represents a well-known advantage (Dixon and Coles, 1997). In the field of tennis, this issue has been considered by Holder and Nevill (1997), who find that the home factor is not important, and by Koning (2011) who instead finds that a home advantage does exist for men but not for women.
We tried to account also for home effect by considering dummy variables H_x assuming value 1 if player x plays in his country and 0 otherwise. As for surface, the variable $Home$ is built as the difference of the two dummies $Home = H_F - H_U$, thus it assumes only value 1, 0 and -1. The expected coefficient of $Home$ is positive.
Unexpectedly, the home factor turns out to be highly significant and renders the variable $Sur2$ barely significant (see column 'Mod5' of Tab. 2). The last point might be the result of a collinearity between home and surface factors. The inclusion of the $Home$ variable improves the Akaike criterion (see Tab. 2), suggesting that the home factor is not as negligible as previously thought. Furthermore, the exclusion of the $Sur2$ variable improves the Akaike criterion (AIC=1166.3) further. Thus, $Sur2$ was discarded.

3.6 External information

All the previous variables measure specific aspects describing the balance of power between two players. It is clear there are many other factors that, in principle, could affect the outcome. Examples include: the physical conditions, the psychological conditions in a particular period, a particular sequence of wins or defeats, etc.
Whether to consider or not these factors is surely important, but it presents the problem of how to measure and include them in the model, because they generally require a detailed and informed analysis of each match. We thought to derive this kind of information indirectly from the odds of bookmakers. Indeed, in defining the odds for a match, they account for all available information and, for their job, they pay particular attention to any news. An example of this situation is the match played by Fognini and Bhambri at the ATP Chennay Tournament in 2014. In that moment, Fognini was ranked 16th and Bhambri 195th, but the Italian player was returning from a physical injury. Bookmakers, who were aware of this injury, quoted for Fognini an odd around 1.55 and for Bhambri around 2.2. It is probable that, without this piece of information,

[7] This threshold was fixed by considering a grid of values and choosing the value that maximises the impact of the variable.

the odds would have been quite different. The model described in column 'Mod5' of Tab.2, for example, quotes 1.32 for Fognini and 2.94 for Bhambri. The match ended with the withdrawl of Fognini in the second set.
To account for these situations and for others requiring updated information, we build the variable $Info$. The idea is to exploit the bookmakers' odds as they contain 'non-sample' information. In particular, $Info = 0$ if $o_F \leq 2$, with o_F being the odd offered by the bookmakers for the favorite player, and $Info = o_F$ if $o_F > 2$. This means that $Info$ activates when the bookmakers assigns to the favorite (in the sense of ranking) player a win probability of less than 0.46. The greater the odd, the more the variable weighs. The rationale is that if bookmakers give to the player that should be the favorite a win probability of less than 0.5, they might know something that the model does not see. When included in the model, the $Info$ variable is largely significant (see column 'Mod6', Tab.2), and, as expected, its coefficient is negative. In addition, it improves the AIC criterion that becomes AIC=1139.6.
In our analyses, over the 1081 matches that have been considered, the $info$ variable is not null for 132 matches, that is, 12.86% of times.

4 The final model for odds

The in-sample selected model for the win probability of the favorite player is Mod6, corresponding to the logistic model (2) with

$$\boldsymbol{x}'\boldsymbol{\beta} = 2.30 + 0.0002\Delta Points + 0.187\Delta Int - 0.064 FAge + 0.497 Home - 0.446 Info. \quad (3)$$

For this model, Table 2 lists the pseduo-R^2 and the Brier score[8] both for the model and for bookmakers[9].
This model is simple and uses only easily available information. Moreover, through expression (2), it allows us to estimate the win probability of the favorite and the unfavorite players p and $q = 1 - p$, respectively. Starting from p and q, we built the odds by applying a 6%-overround, so that the odds for the favorite and the unfavorite players are $o_f = (1.06p)^{-1}$ and $o_U = (1.06q)^{-1}$, respectively.
It is not strange that between the odds quoted by bookmakers and those implied by the model, there is a spread. There are several possible motivations for this difference. For example, bookmakers fix the odds match by match, while the model fixes them strictly following some rules; once the odds have been quoted, bookmakers can adjust them to follow the betting flows while the model cannot; finally, among the possible reasons, there is also the possibility that the model and the bookmakers estimate the odds with different views and accuracy.
To understand better the relationship between quoted odds and those implied by the model, we first compare the win probabilities produced by the model with those implied

[8]The Bries score is defined as $B = \sum_{i=1}^{N}(P_i - X_i)^2/N$, where P is the predicted probability that the favorite (higher-ranked) player wins, while $X = 1$ if the favorite player wins and 0 otherwise. $0 \leq B \leq 1$ and small values of B indicate good prediction accuracy.
[9]For bookmakers we used normalised implied probabilities

Table 2: Estimated model's coefficients and statistics of (in-sample) predictive accuracy. In brackets the corresponding $p-$values.

Model	Mod1	Mod2	Mod3	Mod4	Mod5	Mod6
Const	0.438	1.469	1.805	1.876	1.929	2.297
	(4.9e-07)	(4.2e-05)	(7.0e-03)	(6.9e-03)	(4.3e-03)	(9.6e-04)
$\Delta Points$	2.9e-04	2.4e-04	2.3e-04	2.1e-04	2.34e-04	2.01e-04
	(1.1e-12)	(6.6e-10)	(1.3e-09)	(2.9e-09)	(3.02e-09)	(3.03e-07)
ΔInt	–	0.280	0.3454	0.273	0.288	0.187
	–	(3.7e-06)	(4.5e-06)	(9.0e-05)	(4.0e-05)	(0.010)
$FAge$	–	–	-0.062	-0.059	-0.064	-0.064
	–	–	(0.009)	(0.013)	(0.008)	(0.009)
$Sur2$	–	–	–	0.405	0.352	–
	–	–	–	(0.022)	(0.049)	–
$Home$	–	–	–	–	-0.561	0.497
	–	–	–	–	(4.5e-04)	(0.002)
$Info$	–	–	–	–	–	-0.446
	–	–	–	–	–	(5.3e-09)
AIC	1206.3	1191.0	1181.1	1183.8	1173.3	1139.6
Pseudo–R^2	0.064	0.077	0.082	0.084	0.093	0.122
Number of observations			1081			
Brier Score (Mod6)			0.172			
Brier Score (Bookmakers)			0.166			

by the bookmakers' odds. We consider probabilities instead of odds because the bookmakers' overround is not constant over the matches and, thus, comparing probabilities is more appropriate. However, as bookmakers do not offer fair odds, simple inverse odds can not be used because, usually, their sum is not equal to one. In order to use inverse odds as probability estimates, a normalization is required. Here, following Štrumbelj (2014) we normalized the inverse odds according to Shin (1993, 1991) and, in this section we will always refer to normalized (implied) probabilities.

Figure 3 shows the scatterplots of the model's probabilities versus the bookmaker's win probabilities for the favorite (left) and for the unfavorite (right) player, over the 1081 matches of 2012 that represent our in-sample dataset. As expected, in both cases the favorite player's probability tends to assume high values and, on the contrary, the unfavorite player's probability lower values. There are some exceptions on which, however,

both bookmakers and the model seem to agree. The points tend also to place themselves around the bisecting line, meaning that, on average, the two probabilities are similar and, indirectly, supporting the estimates of the model. Figure 4, shows the distributions of the difference between the bookmaker's normalized and the model's probability for the favorite player, $d1$ (left panel), and for the unfavorite player, $d2$ (right panel). The two distributions have means equal to -0.0031, for $d1$, and 0.0031, for $d2$, while both of them have standard deviation equal to 0.1. Jointly, Figures 3 and 4 point out that there are matches on which the model and the bookmakers do not completely agree, at least in terms of probability. Moreover, as $\hat{P}(d1 > 0) \simeq 0.53$ and $\hat{P}(d2 > 0) \simeq 0.47$, bookmakers tend to be slightly more conservative and to evaluate as more likely the chances of the player with the higher ATP ranking to win. Symmetrically, when unfavorite players are involved, the model tends to assign higher win probability than the market. This propensity is more pronounced for low levels of probability for the favorite player and high levels of probability for the unfavorite player (see Figure 3). In terms of odds, this translates in model-implied odds relatively higher that those of bookmakers for the favorite player and relatively lower for the unfavorite player (see Fig.5).
Globally, these analyses point out that model-implied odds are in line with those quoted in the market, but they are not mere replications of bookmakers' odds.
Table 3 gives the percentages of correct and uncorrected winner's predictions for the model and for bookmakers. Also in this case, the final behaviour is very similar: over the 1081 matches of the in-sample dataset, the model correctly predicts the winner 75.8% of times while bookmakers guess the outcome 77.5% of times. Nevertheless, we want to stress that, as noted also by Mchale and Morton (2011), obtaining a positive reward from betting does not simply coincide with a good prediction of the final outcome. In Section 5, indeed, we will use the spreads between model's and the bookmakers' odds to identify situations that are worthy to be exploited.

Table 3: Percentage of in-sample correct and uncorrected predictions for model and bookmakers. Fav.=favorite (higher rank) player, Unfav.=unfavorite palyer.

In-sample	Model's bet		Bookmakers' bet	
	Fav.	Unfav.	Fav.	Unfav.
Win Fav.	66.3	5.6	66.0	5.9
Win Unfav.	18.5	9.6	18.6	9.5

5 Out-of-sample prediction of the model for betting

In this section, we discuss the out-of-sample performance of our model. To this purpose, we consider, as out-of-sample dataset, all the matches played in the four Grand

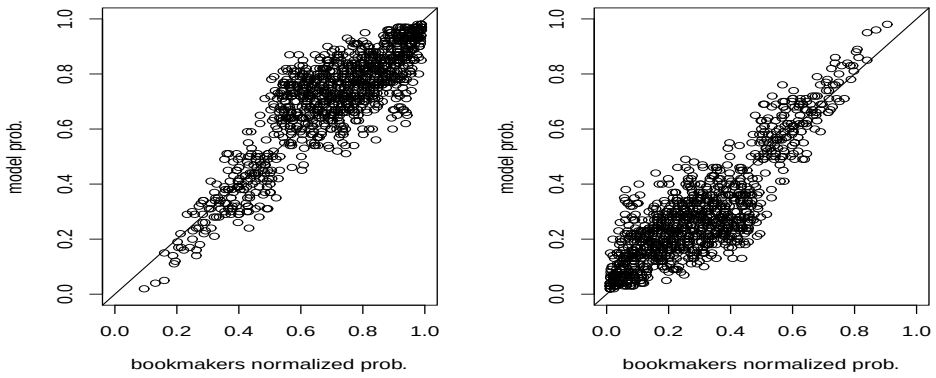

Figure 3: Left: scatterplot of the bookmaker normalized (implied) probabilities vs the model probabilities, for the favorite player and for the 1081 matches of 2012 . Right: the same for unfavorite players.

Slam Championships in 2013: the Australian Open 2013, Roland Garros 2013, Wimbledon 2013 and US Open 2013. The Australian Open and the US Open are played on hard courts, the Roland Garros on clay and Wimbledon on grass. The Grand Slam tournaments have draws composed of 128 players and 127 matches. On the whole, we considered 501 matches[10].

Let us first consider the predictive performance, in terms of percentage of correct predicted outcomes. Table 4 gives the percentages of correct and uncorrected out-of-sample predictions for the model and for bookmakers, as well as the out-of-sample Bries scores over all the 501 matches of the out-of-sample dataset. The results are very similar: the model correctly guesses the final outcome 77.2% of times against the 78% of bookmakers. Also the Brier score is only slightly higher than that for the market.

However, in this work the the emphasis is not on the ability on predicting the final outcome but on the economic performance, when the model is embedded in a strategy for betting on matches of the ATP tour. Indeed, as noted also by (McHale and Morton, 2011), obtaining positive reward from betting has more profound than the simple prediction of the final outcome. The idea is to bet 'against' the betting market, that is, to bet that our statistical model will manage the information more efficiently than the bookmakers will. To this end, once the models' odds have been computed, we try to find in the market the quoted odds that, with respect to the models' odds, over- or under-estimate the probability of a specific outcome.

As the market's and model's odds rarely coincide, the problem is to understand when is worthy to exploit their difference. This, in turn, implies the definition of a strategy or set of rules suggesting whether to bet on each match.

[10]For the four tournaments, the total number of matches should be 508 but 7 of them were not played.

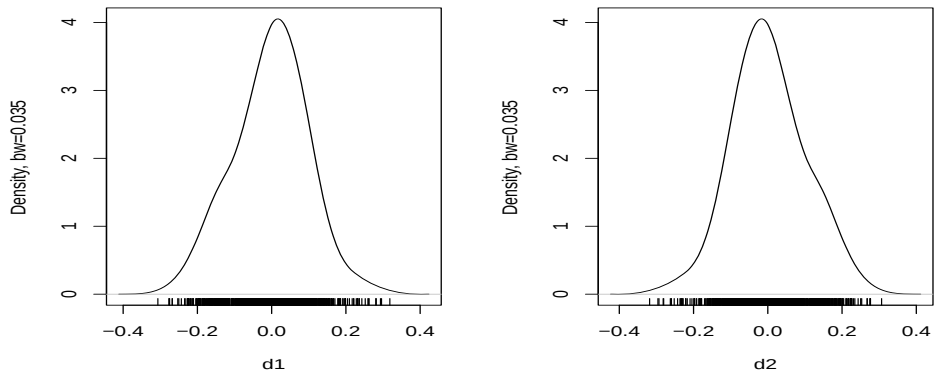

Figure 4: Left: kernel density of the difference between bookmakers and model probability for the favorite player ($d1$) for the 1081 matches of 2012 ; right: the same for the unfavorite player ($d2$).

The theory behind this approach refers to the large numbers law, stating that when we have an outcome with happening probability p, if the success is rewarded by $w > 1/p$, then the expected reward of the bet is positive. We make use of model (3) to estimate, as best as possible, the probability p. Clearly, as this law is asymptotic, to produce significant results we must consider a large number of matches.
In our case, for each of the 501 matches of the four Grand Slam tournaments, we computed the odds using model (3) and we compared them with the best-quoted odd among the seven reported on www.tennis-data.co.uk. Usually, there is not a large difference among the quoted odds; nevertheless, we selected the higher one to try to exploit any kind of opportunity.
Let us define by $o_{m,f}$ and $o_{m,u}$ the odd implied by the model for the favorite and the unfavorite player and by $o_{b,f}$ and $o_{b,u}$ the analogous quantities for the best odds quoted by bookmakers. To choose whether to bet we propose a betting strategy based on these three conditions:
1. the model and the bookmakers must agree on the (un)favorite player:
2. the bookmakers' odds must be such that:
- $o_{b,f} > o_{min,f}$ and $o_{b,u} > o_{min,u}$;
3. the bookmakers and model's odds must be such that:
- $t_{inf,f} \leq (o_{b,f}/o_{m,f}) \leq t_{sup,f}$ if we are betting on the favorite player or
- $t_{inf,u} \leq (o_{b,u}/o_{m,u}) \leq t_{sup,u}$ if we are betting on the unfavorite player.

Condition 2 is independent of the model's results, whereas conditions 1 and 3 are model-related. More specifically, condition 1 is intended to prevent betting on matches that are too uncertain, typically when both p_f and p_u belong to the range $0.45 - 0.55$ or when there are conditions or news able to change the natural favorite. These cases are uncommon, because the variable $Info$ usually prevents them.

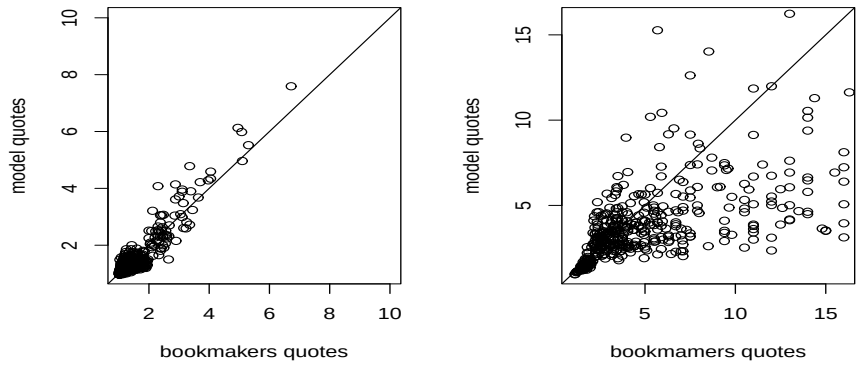

Figure 5: Left: scatterplot of the bookmaker vs the model-implied odds (6%-overround), for the favorite player and for the 1081 matches of 2012 . Right: the same for unfavorite players.

Condition 2. prevents betting on matches with quoted odds less than the minimum thresholds $o_{min,f}$ and $o_{min,u}$. Note that, these thresholds are absolute limits, that are independent of the model's odds. In practice, they ensure we wager only when the favorite and unfavorite's odds belong to a moderately sized interval of values. For example, with the thresholds chosen in this paper, we bet only on matches where $o_{b,f}$ belongs to the range (1.30–1.70) and $o_{b,u}$ to the range (2.40–4.30). These ranges discard stakes with a too low margin on the favorite or a too risky margin on the unfavorite. Avoiding too high odds for the unfavorite player allows us to account for situations where it is highly unlikely that he could win because of the technical difference between the two players. For example, in 2012 over 220 matches for which the odds belonged to this category, only in 11 (5% of times) cases the favorite player was defeated.
Finally, condition 3. fixes some thresholds, $t_{inf,.}$ and $t_{sup,.}$, beyond which we do not bet. They are not based on the absolute value of the odds but on the relative value of the bookmakers' odds with respect to the model's odds and are particularly important to establish when it is worth to bet. As $(o_{b,f}/o_{m,f}) = (p_{m,f}/p_{b,f})$, in practice, the limits refer to the differential of the probabilities assigned by the model and by bookmakers. The lower threshold selects matches where the win probability assigned by the model is greater than that assigned by bookmakers of a minimum level $t_{inf,f}$. The greater the difference, the more convenient it is to bet. However, when the difference it is too large, it is possible that bookmakers are accounting for some factor not included in the model. Thus, we fix and upper threshold, $t_{sup,f}$, beyond which, to be safe, we prefer not to bet. Analogously, when we bet on the unfavorite player.
The quantities $o_{min,f}$, $o_{min,u}$, $t_{inf,f}$, $t_{sup,f}$, $t_{inf,u}$ and $t_{sup,u}$ are the parameters of the procedure and were calibrated to maximise the in-sample cumulative return. In practice, we ran the procedure in-sample over a grid of combinations of these parameters and we

Table 4: Percentage of correct and uncorrected out-of-sample predictions for model and bookmakers and Bries scores. Fav.=favorite (higher rank) player, Unfav.=unfavorite palyer.
Note that these results do not coincide with those of Tab.5 as they refer to all the 501 matches of the out-of-sample dataset, while the latter refer only to the matches on which we bet.

Out-of-sample	Model's bet		Bookmakers' bet	
	Fav.	Unfav.	Fav.	Unfav.
Win Fav.	69.2	5.8	69.6	5.4
Win Unfav.	17.0	8.0	16.6	8.4
Number of observations			501	
Brier Score (Mod6)			0.165	
Brier Score (Bookmakers)			0.150	

chose the values that led to the higher cumulative return.
These three conditions limit considerably the number of matches on which to bet. This is a strong point of the proposed procedure. Indeed, it limits the needed amount of money to invest and focuses on the matches whose outcomes are clearer. However, we do not want to reduce too much the number of bets because on one hand, they represent a risk and on the other, they are opportunities. Therefore, in the in-sample optimisation we discarded the combination of parameters leading to a percentage of bets less than 10%. For the 2012 data, the optimised parameters were $o_{min,f} = 1.3$, $o_{min,u} = 2.4$, $t_{inf,f} = 1.05$, $t_{sup,f} = 1.20$, $t_{inf,u} = 1.10$ and $t_{sup,u} = 1.25$
First of all, we consider the predictive performance of the model over the entire out-of-sample dataset and we compare it with the analogous performance of bookmakers. Table 4 lists the percentages of correct and uncorrected predictions over the 501 matches of the four Grand Slam Championship 2013. The model guesses the final outcome 77.2% of times whereas bookmakers 78% of times. Also the Brier scores are similar: 0.165 for the model and 0.150 for bookmakers. Thus, in terms of outcome's prediction, the model and bookmakers behave similarly.
We then consider the results connected to the previous described strategy. It led us to bet on 72 matches over 501. This means that we actually bet only on 14.4% of matches. Of the 72 bets, 53 were successful, corresponding to 73.6% of times, a figure slightly higher than that reported in Table 4 in Kovalchik (2016). Figure 6 shows the pattern of the cumulative return (in Euro) of the betting sequence in chronological order. The maximum drawdown is 30%, but it is important to note that this figure was reached after only two stakes, meaning that in monetary terms it is not particularly relevant. After ten stakes, the maximum drawdown is 0 and the cumulative return is always positive Table

5 lists the results for the whole sequence of bets and singularly for each tournament. We can see the percentage of matches on which we bet ranges from 11.0% to 17.7%: these percentages largely reduce the amount of money required for the global investment. The percentage of winning bets instead ranges from 64.3% to 81.8%, but it is interesting to note that, because of the different odds, a higher percentage of winning bets does not always translate to a larger reward. Indeed the overall profit heavily depends on winning a few bets at long odds. It is these wins at long odds which make the difference in the final return. Finally, Table 5 includes the percentage cumulative return of each championship: it ranges from 6.5% to 31.5%. Thus, even if the reward for the US Open 2013 was moderate, all four tournaments have positive cumulative return.
To have some benchmark, we computed the cumulative return of a strategy based on betting i) always on the favorite player and ii) randomly. In the first case, over the 501 matches, and with respect to the quoted bookmakers' odds, the percentage cumulative return was -14.6%. In the second case, we generated $M = 10000$ binary sequences (1 =bet on the favorite, 0 =bet on the unfavorite) of length $n = 501$, each of which is distributed as a Binomial$(n = 501, p = 0.5)$. Thus, each sequence defines (randomly), for each match, the player on which to bet. For each sequence, we computed the percentage cumulative return, with respect to the quoted oddmandelP4, s. Finally, we took the average of the 10000 cumulative returns. This (unwise) strategy led to a mean percentage cumulative return of -39.6%.

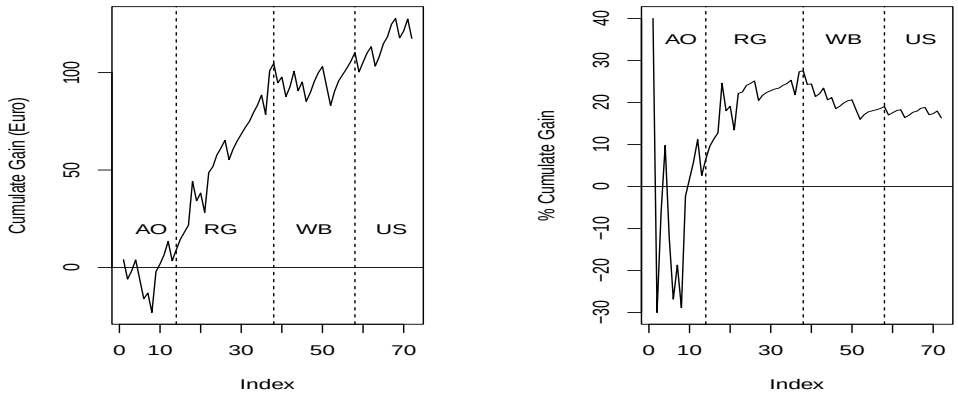

Figure 6: Cumulative gain in Euro (left) an percentage (right). AO=Australian Open, RG=Roland Garros, WB=Wimbledon, US=US Open. The dashed vertical lines specify the end of each tournament. Amount of money invested in each bet=10 Euro

Table 5: Out-of-sample results. n.B=number of bets; %B=% of matches on which we bet; %W=% of wins over the bets; %Cum ret=% cumulative return over the amount of invested money; %MaxD=% maximum drawdown.

	AO13	RG13	WB13	US13	TOT
n.matches	127	124	125	125	501
n.B	14	22	20	16	72
%B	11.0	17.7	16.0	12.8	14.4
%W	64.3	81.8	70.0	75.0	73.6
%Cum ret	6.5	31.5	11.8	9.5	16.3
%MaxD	30.0	0.0	0.0	6.3	30.0

6 Conclusions

We proposed a strategy to bet against bookmakers on the tennis market. First, we built a model producing the win probability of a tennis match, then the outcome of the model is used to fix the odds. Finally, we suggested some empirical rules, to choose whether to bet. We calibrated all the parameters of the strategy in 1081 matches representing our in-sample dataset, and we used it in an out-of-sample context over the 501 matches of the four Grand Slam tournaments in 2013. After 501 matches we got a cumulative return of 16.3%.

As the value of any empirical analysis is, in some way, limited to the data have been considered, this result can not be conclusive and requires further evidences involving a larger out-of-sample dataset, different periods and more in-depth analyses. In particular, future research should include also the 500 and 250 ATP and/or WTA championships as well as a larger number of matches. A rolling calibration of the model, and consequent analyses of predictive results, are also important to study the stability of the estimated procedure parameters and of the results.

Acknowledgement

The authors wish to thank two anonymous referees, whose comments greatly contributed to improve the quality of this work.

References

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

Boulier, B. and Stekler, H. O. (2003). Predicting the outcomes of national football league games. *International Journal of Forecasting*, 19:257–270.

Chitnis, A. and Vaidya, O. (2014). Performance assessment of tennis players: Application of dea. *Procedia - Social and Behavioral Sciences*, 133:74–83.

Clarke, S. and Dyte, D. (2000). Using offcial ratings to simulate major tennis tournaments. *International Transaction in Operational Research*, 7:585–594.

del Corral, J. and Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for grand slam tennis matches? *International Journal of Forecasting*, 26:551–563.

Dixon, M. and Coles, S. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statitics*, 46:265–280.

Easton, S. and Uylangco, K. (2010). Forecasting outcoms in tennis matches using within-match betting markets. *International Journal of Forecasting*, 26:554–575.

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

Gilsdorf, K. F. and Sukhatme, V. A. (2008). Testing rosen's sequential elimination in tournamento model incentives and player performance in professional tennis. *Journal of Sports Economics*, 9:287–303.

Graham, I. and Stott, H. (2008). Predicting bookmaker odds and efficiency for uk football. *Applied economics*, 40:99–109.

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

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

Klaassen, F. and Magnus, J. (2014). *Analyzing Wimbledon. The Power of Statistics*. Oxford University Press.

Knight, G. and O'Donoghue, P. (2012). The probability of winnning break points in grand slam men's singles tennis. *European Journal of Sport Science*, 12:462–468.

Knottenbelt, W. J., Spanias, D., and Madurska, A. M. (2012). A common- opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications*, 64:3820–3827.

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

Kovalchik, S. A. (2016). Searching for the goat of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12:127–138.

Leitner, C. A., Zeileis, A., and Hornik, K. (2009). Is Federer stronger in a tournament without Nadal? an evaluation of odds and seedings for Wimbledon 2009. *Research Report Series/Department of Statistics and Mathematics 94*.

McHale, I. and Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27:619–230.

Nadimpalli, V. K. and Hasenbein, J. J. (2013). When to challenge a call in tennis: a Markov decision process approach. *Journal of Quantitative Analysis in Sports*, 9:229–238.

Scheibehenne, B. and Bröder, A. (2007). Predicting Wimbledon 2005 tennis results by

mere player name recognition. *International Journal of Forecasting*, 23:415–426.

Shin, H. S. (1991). Optimal betting odds against insider traders. *The Economic Journal*, 101:1179–1185.

Shin, H. S. (1993). Measuring the incidence of insider trading in a market for state-contingent claims. *The Economic Journal*, 103:1141–1153.

Skiena, S. (2004). *Calculated bets. Computers, gambling and mathematical modeling to win*. Cambridge University Press.

Spanias, D. and Knottenbelt, W. J. (2012). Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*, 24:311–320.

Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International Journal of Forecasting*, 30:934–943.