Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 26 (2010) 551–563
international journal of forecasting
www.elsevier.com/locate/ijforecast

Are differences in ranks good predictors for Grand Slam tennis matches?

Julio del Corral[a,*], Juan Prieto-Rodríguez[b,1]

[a] *Fundación Observatorio Económico del Deporte—Department of Economics, University of Castilla la Mancha, Ronda de Toledo s/n, 13071, Ciudad Real, Spain*

[b] *Fundación Observatorio Económico del Deporte—Department of Economics, University of Oviedo, Avda. del Cristo s/n, 33006, Oviedo, Spain*

Abstract

This paper tests whether the differences in rankings between individual players are good predictors for Grand Slam tennis outcomes. We estimate separate probit models for men and women using Grand Slam tennis match data from 2005 to 2008. The explanatory variables are divided into three groups: a player's past performance, a player's physical characteristics, and match characteristics. We estimate three alternative probit models. In the first model, all of the explanatory variables are included, whereas in the other two specifications, either the player's physical characteristics or the player's past performances are not considered. The accuracies of the different models are evaluated both in-sample and out-of-sample by computing Brier scores and comparing the predicted probabilities with the actual outcomes from the Grand Slam tennis matches from 2005 to 2008 and from the 2009 Australian Open. In addition, using bootstrapping techniques, we also evaluate the out-of-sample Brier scores for the 2005–2008 data.

Keywords: Sports forecasting; Probit models; Prediction intervals; Tennis; Brier scores; Bootstrapping

1. Introduction

Several papers have studied sports data predictions, given that sports events are quite distinct from completely randomized events such as lotteries. The previous literature on sports forecasting can be divided into several groups, according to the type of forecasts made. For example, several papers have aimed to predict the result of a particular match between two contestants (Boulier & Stekler, 2003; Caudill, 2003; Klaassen & Magnus, 2003). Other papers have aimed to predict the point spread between two contestants (Smith & Schwertman, 1999), and still other papers have aimed to predict the winner of sports events involving several contestants, such as tournaments (Anderson, Edman, & Ekman, 2005; Clarke & Dyte,

* Corresponding author. Tel.: +34 926295300x3521; fax: +34 926 295 211.
E-mail addresses: julio.corral@uclm.es (J. del Corral), juanprieto@uniovi.es (J. Prieto-Rodríguez).
[1] Tel.: +34 985103768; fax: +34 985104871.

0169-2070/$ - see front matter © 2010 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2009.12.006

2000), leagues (Rue & Salvesen, 2000) or races (Bolton & Chapman, 1986).

Two main methods for predicting the outcome of a sports event exist, namely statistical models and expert evaluations. Thus, some scholars have compared the accuracies of these competing methods (Anderson et al., 2005; Boulier & Stekler, 2003; Forrest, Goddard, & Simmons, 2005). With regard to statistical models, a myriad of different models have been used. For instance, if the objective is to construct winning probabilities for contestants in a match, the most prominent approach is to use either logit (Clarke & Dyte, 2000; Klaassen & Magnus, 2003) or probit (Abrevaya, 2002; Boulier & Stekler, 1999) models. Another alternative is to use the maximum score estimator (Caudill, 2003). If a tie is possible (e.g., in soccer), either ordered probit models (Goddard & Asimakopoulos, 2004) or multinomial logit models (Forrest & Simmons, 2000) can be used. However, if the probabilities for a match are based on the number of points, goals or run probabilities, Poisson regression (Dixon & Coles, 1997) and negative binomial (Cain, Law, & Peel, 2000) models should be used to take the discrete nature of the data into account.

With regard to the variables that enter these statistical models, Goddard and Asimakopoulos (2004) and Goddard (2005) used several variables related to past results, as well as information about the number of goals scored and conceded. Forrest and Simmons (2000) used the performance in previous matches as an explanatory factor in their logit models of the English national soccer league. Dyte and Clarke (2000) used *Fédération Internationale de Football Association* (FIFA) ratings to predict the numbers of goals scored by national teams competing in the 1998 FIFA World Cup. Similar to FIFA, other sports-governing bodies also produce rankings based on the past performances of contestants. These rankings have been used as predictors of victory in several different settings. For instance, Boulier and Stekler (1999) found that the ranking difference between contestants is a good predictor in professional tennis and collegiate basketball. Lebovic and Sigelman (2001) demonstrated the accuracy of collegiate football rankings in predicting match outcomes. Smith and Schwertman (1999) showed that the difference in rankings is a good predictor of the victory margin in collegiate basketball. Caudill and Godwin (2002) developed a heterogeneous skewness model that takes into account not only differences in rankings but also their degree.

In tennis in particular, Klaassen and Magnus (2003) proposed a method of forecasting the winner of a match at the beginning of the match, as well as during it. For this purpose, they used a measure based on nonlinear differences in rankings, similar to that used by Caudill and Godwin (2002). Clarke and Dyte (2000) used tennis rankings to estimate the chance of winning as a function of the difference in rating points, and were able to estimate a player's chance of a tournament victory once the draw for the tournament became available.

The aim of the present paper is to extend the previous literature in the following ways. First, we test whether the ranking difference is a good predictor of tennis victories, using an approach which is different from those used in previous studies. Specifically, we classify our variables into three groups, namely a player's past performance, a player's physical characteristics and the match characteristics. We then estimate three alternative probit models for men and women separately using Grand Slam tennis match data from 2005 to 2008. In the first model, all three explanatory variables are included, whereas in the other two specifications, either the player's physical characteristics or the player's past performances are not considered. Subsequently, the forecasting accuracies of the different models are evaluated by computing the Brier scores and comparing the predicted probabilities with the actual outcomes from the 2005 to 2008 Grand Slam matches, as well as from the 2009 Australian Open matches. Moreover, using bootstrapping techniques, we evaluate the out-of-sample Brier scores from 2005 to 2008. Using probit estimates, we also study the effect of ranking differences on predicting Grand Slam tennis outcomes, and in particular we analyze whether this effect varies by gender.

Section 2 presents the empirical model specifications. The data are presented in Section 3, followed by the probit results in Section 4. In Section 5, the predictive accuracies of the models are analyzed. Our conclusion is given in Section 6.

2. Empirical model specifications

In order to obtain the determinants of match outcomes, we estimate probit models in which the

dependent variable (HIGHER-RANKED VICTORY) takes the value of one when the higher-ranked player wins. It is noteworthy that the unit of analysis is the match. Two kinds of covariates are used. The first type is the match characteristics, and the second is the player's characteristics. To control for the match characteristics, we use dummies for the tournament (AUSTRALIA, FRENCH OPEN, WIMBLEDON and US) and a set of dummies to control for the round of the match. The round dummies are expected to increase when advancing rounds. This expectation is based on the analysis of Gilsdorf and Sukhatme (2007), who found that the larger the difference in prizes between the winner and the loser, the less likely an upset was.

With respect to the player's characteristics, we used two kinds of variables. The first refers to the previous results of the player, whereas the second refers to the player's physical characteristics. Thus, we use three kinds of covariates: match characteristics, previous results for a player and the player's physical characteristics. This classification is necessary to ascertain the relevance of past performance variables in predicting tennis victories. In making these predictions, we estimate three alternative models for each gender. In the first model (M1 for men and F1 for women) we use all available covariates. In the second model (M2 for men and F2 for women) we remove the variables related to past performance, and in the last model (M3 for men and F3 for women) we remove the physical characteristics of the players. With regard to a player's past performance, we use the difference between the natural logarithms of the rankings of the lower-ranked player and the higher-ranked player (i.e., *log*(*lower-ranked player's ranking*) – *log* (*higher-ranked player's ranking*)) at the beginning of the tournament (DIFRANKING).[2] The main advantage of this measure is that the differences in player quality are not linear; instead, they grow at an increasing rate as we move up the ranking. This implies that a difference of one position in tennis rankings corresponds to an almost insignificant difference in quality if the players are at the bottom of the ranking, but corresponds to a more substantial difference when we compare the top ten players.[3] The expected value of the coefficient of this variable is positive.

To control for tournament-specific effects, we use the difference between the rounds achieved by the higher- and lower-ranked players for the previous year at the same tournament (DIFROTOUR). We expect this variable to have a positive coefficient, especially for men, whose skills are more surface-biased than those of women.[4]

Lastly, we created a dummy variable that takes the value of one if the player has been a top-ten player at some point over the past five years. The specification also includes dummies that take the value of one if a former top-ten player is the lower-ranked player (EXTOP10L) or the higher-ranked player (EXTOP10H). The expected signs of the coefficients are positive for EXTOP10H and negative for EXTOP10L. This is based on the expectation that former top-ten players will play well if they are sufficiently extra-motivated in a particular match, and thus, they are more likely to outperform the expected result based only on their actual ranking. For instance, both Williams sisters, who had previously been top-ranked players, won a Gram Slam tournament, even though at the time they were ranked outside the top ten.[5]

With regard to a player's physical characteristics, we include the difference in age between the higher- and lower-ranked players (DIFAGE) and its square (DIFAGE2). With a large dataset, the younger the higher-ranked player is relative to the lower-ranked player, the greater the probability that the higher-ranked player will win. Therefore, a negative joint effect on these variables is expected; that is, if the higher-ranked player plays against

[2] Klaassen and Magnus (2001) were the first to use this measure.

[3] The absolute difference in the ranking was also used, but the results were worse than those obtained using a log difference. This could imply that absolute differences do not capture quality differences adequately. The results are not reported here but are available on request. Table A.1 provides some examples of DIFRANKING.

[4] Wimbledon developed a complex yet objective way of assigning seeds for men by taking into account not only the rankings but also previous performances in grass tournaments; however, for women, the WTA rankings are followed in assigning seeds, unless a committee determines that some seedings should be changed. This demonstrates, in some ways, that skills are more surface-biased in men than in women.

[5] Serena Williams won the 2007 Australian Open when she was ranked 81st; Venus won the 2007 Wimbledon when she was ranked 31st.

a younger contestant, the higher-ranked player has a lower probability of winning. We also include the height difference (DIFHEIGHT) and its square (DIFHEIGHT2) between higher- and lower-ranked players. Since player height is positively correlated with service skills (i.e., a taller player is more likely to serve faster), we expect a positive joint effect for these variables. Finally, to control for right- versus left-handedness, we include a set of dummy variables reflecting the four possibilities, that is, when both players are right-handed (BOTHRIGHT), which is the reference category; when both players are left-handed (BOTHLEFT); when the lower-ranked player is left-handed and the higher-ranked player is right-handed (LEFTL); and when the lower-ranked player is right-handed and the higher-ranked player is left-handed (LEFTH).

3. Data

The tennis Grand Slam is composed of four tournaments: the Australian Open, the French Open, Wimbledon and the US Open. Even though these tournaments are similar in prestige and prize money, they differ in terms of court surfaces. The Australian Open and the US Open are played on hard courts, the French Open is played on clay, and Wimbledon is played on grass. The results of the matches were collected using the individual draws for the tournaments from 2005 to 2008. Each draw is composed of 128 players. Thus, 127 matches are played in each tournament for both men and women. In total, we gathered data from 4064 matches.

In addition, we collected data regarding individual characteristics, such as a player's ranking at the time of the tournament, his/her best previous ranking and its date, and the height and date of birth of the players. Most of the data were gathered from the ATP (www.atptennis.com) and WTA (www.sonyericssonwtatour.com) websites, but some data were gathered from the website www.tennis-data.co.uk.[6] Table 1 provides some descriptive statistics of the sample.

The higher-ranked players win their matches with almost the same frequency (roughly 71.5%) across the genders. In fact, the difference is not significant according to a mean differences t-test ($t = 0.14$). An interesting gender difference is related to past performance. For men, around 28.6% of the players were ranked in the top ten at some point during the five years prior to the tournament under consideration (EXTOP10); in contrast, only 15.1% of women players had been ranked in the top ten within the same time period. This difference is significant according to a mean differences t-test ($t = 6.9$), and can be related to the fact that men generally have longer careers. Therefore, it is quite common to find players like Andre Agassi, Lleyton Hewitt, Carlos Moyá or Marat Safin, who were top-ranked players in the ATP ranking in the past but were still active within our sample.

4. Probit results

Tables 2 and 3 present the coefficients, standard errors and marginal effects based on the probit models for men and women, respectively.[7]

Regarding player characteristics, the most relevant variable is the difference in ATP or WTA rankings, since it is the only variable that was significant across all models. We also found that, as expected, the larger the ranking difference between the two contestants, the higher the probability of victory for the higher-ranked player. Furthermore, since we used a logarithmic transformation, these differences are more important as we move toward the top of the distribution. Table 4 illustrates this result by showing the predicted probabilities of victory for the 1st, 11th, 51st and 61st ranked players against an average ranked player in the sample (i.e., 79 for men and 73 for women), assuming that the other variables take the mean values across observations observed for M1 and F1.

As Table 4 shows, the predicted probability of victory for the higher-ranked player increases for the 1st player, relative to the 11th player, by 15% for men and 16% for women, but the predicted probability only increases by 2% if we compare the 51st and 61st players in the ATP, and only 3% in the WTA. Thus, the same difference in ranking positions can have a very different effect, depending on where we are in the ranking distribution.

[6] These data were previously used by Forrest and McHale (2007) to estimate the relationship between returns and odds.

[7] It is important to note that in the M1, F1, M2 and F2 models, it was not possible to use all observations due to missing data.

Table 1
Descriptive statistics.

	Male		Female	
	Mean	Std. dev.	Mean	Std. dev.
HIGHER-RANKED VICTORY	0.712	0.453	0.715	0.452
DIFRANKING	1.440	1.081	1.399	1.051
EXTOP10H	0.262	0.440	0.245	0.430
EXTOP10L	0.115	0.319	0.065	0.247
DIFHEIGHT (m.)	−0.002	0.091	0.025	0.091
DIFHEIGHT2 (m.)	0.008	0.014	0.009	0.012
DIFAGE	−0.112	4.897	0.070	5.767
DIFAGE2	23.978	34.341	33.246	45.016
LEFTL	0.110	0.313	0.077	0.267
LEFTH	0.094	0.293	0.064	0.244
BOTHLEFT	0.021	0.143	0.004	0.064
BOTHRIGHT	0.774	0.418	0.855	0.352
Number of observations	2022		1930	
EXTOP10	0.286	0.503	0.151	0.369
RANKING	87.191	62.809	83.349	58.866
AGE	25.869	2.633	23.821	2.864
HEIGHT	1.842	0.045	1.721	0.045
RIGHT-HANDED	0.874	0.244	0.882	0.226

The statistics for the bold variables refer to the first round of all tournaments.

To demonstrate the effects of ranking differences on the predicted probabilities, Fig. 1 displays these effects for both men and women according to M1 and F1, including both plots and confidence intervals. The figure demonstrates the similarity of these effects for the two groups, with solid lines representing the effect for men and dashed lines representing the effect for women. Note that the estimated probabilities are always within the other gender's confidence interval bounds, i.e., the rank-difference effect is not statistically different between men and women. Furthermore, as the ranking differences (in log terms) move toward zero, the probability of a higher-ranked player victory tends toward 0.5. However, as the ranking difference becomes larger, the probability of a lower-ranked player victory almost vanishes.

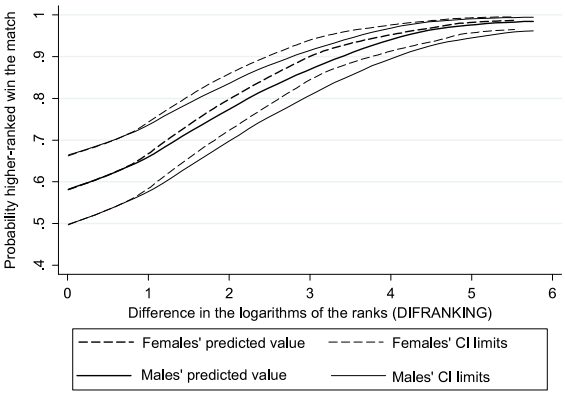

Fig. 1. The rank-difference effect on predicted probabilities in M1 and F1.

With regard to the rest of the variables, DIFROTOUR has a positive and significant coefficient for men, but the coefficient is insignificant for women, implying that previous outcomes in the same tournament are correlated with better results for men but not for women. Therefore, there is an individual-tournament effect among men that does not exist among women. A plausible explanation of this result is that tennis skills are much more surface-biased in men than in women. However, the question of whether a player has previously been ranked in the top ten appears to be much more important for women than for men.

Height differences only have a significant coefficient in F2. In addition, differences in age between the higher-ranked player and the lower-ranked player display a significant, monotonically decreasing pattern for men and an inverted U-shaped effect for females, as is shown in Fig. 2. Consequently, we can estab-

Table 2
Probit results for men.

	M1			M2			M3		
	Coeff.	St. dev.	ME	Coeff.	St. dev.	ME	Coeff.	St. dev.	ME
DIFRANKING	0.321***	0.039	0.104				0.342***	0.038	0.112
EXTOP10H	0.129	0.080	0.041				−0.005	0.075	−0.002
EXTOP10L	−0.373***	0.106	−0.130				−0.113	0.098	−0.038
DIFROTOUR	0.081***	0.017	0.026				0.071***	0.017	0.023
DIFHEIGHT	0.314	0.340	0.102	0.052	0.326	0.018			
DIFHEIGHT2	−2.364	2.189	−0.764	−3.955*	2.110	−1.338			
DIFAGE	−0.050***	0.007	−0.016	−0.037	0.006	−0.013			
DIFAGE2	0.002**	0.001	0.001	0.001	0.001	0.000			
LEFTL	−0.176*	0.100	−0.059	−0.093	0.095	−0.032			
LEFTH	−0.035	0.111	−0.012	−0.051	0.104	−0.017			
BOTHLEFT	0.023	0.228	0.007	0.058	0.214	0.019			
2ND ROUND	0.029	0.078	0.009	0.183**	0.074	0.060	0.044	0.077	0.014
3RD ROUND	−0.058	0.098	−0.019	0.039	0.093	0.013	−0.075	0.097	−0.025
4TH ROUND	−0.068	0.132	−0.022	0.064	0.125	0.021	−0.061	0.131	−0.020
QUARTERFINAL	0.118	0.198	0.037	0.290	0.185	0.090	0.128	0.194	0.040
SEMIFINAL	−0.124	0.253	−0.042	0.069	0.242	0.023	−0.156	0.246	−0.053
FINAL	0.048	0.375	0.015	0.061	0.340	0.020	−0.053	0.359	−0.018
AUSTRALIA	0.110	0.089	0.035	0.107	0.085	0.036	0.123	0.088	0.039
FRENCH OPEN	−0.043	0.087	−0.014	−0.049	0.084	−0.017	−0.042	0.086	−0.014
WIMBLEDON	−0.010	0.088	−0.003	−0.037	0.084	−0.013	0.003	0.086	0.001
CONSTANT	0.056	0.088		0.514***	0.075		0.027	0.078	
Number of observations		2022			2022			2032	
Likelihood ratio test		253			57			197	
Pseudo-R^2		0.104			0.023			0.081	
Log-likelihood		−1088			−1186			−1120	

Note: ME denotes marginal effect.
* Indicates significance at the 10% level.
** Indicates significance at the 5% level.
*** Indicates significance at the 1% level.

lish that the probability of a higher-ranked player victory decreases as this player competes against younger players. Among men, if the higher-ranked player is 15 years younger than the lower-ranked player, his winning probability is over 20% higher than when he plays against an opponent his own age. However, for women, this effect is close to zero.

Finally, the results for left-handed, lower-ranked players in men's matches are somewhat remarkable. Once quality differences are controlled for, left-handed, lower-ranked players are more likely to defeat right-handed, higher-ranked players. According to the estimated marginal effect, on average, higher-ranked players have a 5.9% lower probability of winning a match when they face a left-handed player. An intuitive explanation of this result is that left-handed players are scarce, and so right-handed players are not as accustomed to playing against them; left-handed players, however, are accustomed to playing

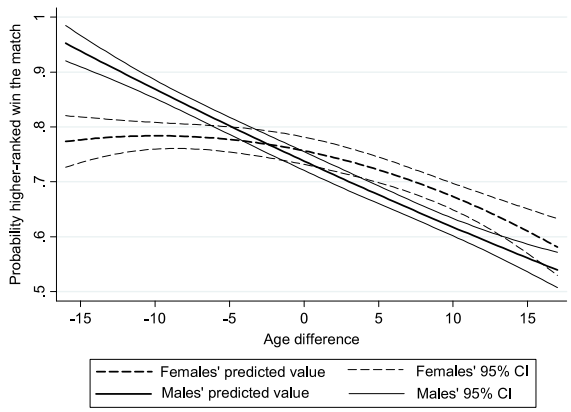

Fig. 2. The age-difference effect on predicted probabilities in M1 and F1.

Table 3
Probit results for women.

	M1			M2			M3		
	Coeff.	St. dev.	ME	Coeff.	St. dev.	ME	Coeff.	St. dev.	ME
DIFRANKING	0.384***	0.044	0.123				0.410***	0.041	0.132
EXTOP10H	0.453***	0.095	0.133				0.415***	0.090	0.124
EXTOP10L	−0.562***	0.146	−0.203				−0.498***	0.143	−0.179
DIFROTOUR	0.003	0.018	0.001				−0.006	0.017	−0.002
DIFHEIGHT	0.646	0.411	0.207	1.799***	0.385	0.604			
DIFHEIGHT2	2.463	3.269	0.790	4.661	3.087	1.564			
DIFAGE	−0.019***	0.006	−0.006	−0.003	0.006	−0.001			
DIFAGE2	0.000	0.001	0.000	−0.001	0.001	0.000			
LEFTL	0.023	0.121	0.007	0.023	0.117	0.008			
LEFTH	−0.074	0.126	−0.024	−0.182	0.122	−0.064			
BOTHLEFT	−0.450	0.463	−0.162	−0.172	0.450	−0.061			
2ND ROUND	−0.007	0.081	−0.002	0.152**	0.077	0.050	0.008	0.078	0.003
3RD ROUND	−0.172*	0.104	−0.057	0.045	0.096	0.015	−0.157	0.102	−0.052
4TH ROUND	−0.170	0.139	−0.057	0.030	0.127	0.010	−0.156	0.137	−0.052
QUARTERFINAL	−0.008	0.195	−0.003	0.060	0.176	0.020	0.010	0.193	0.003
SEMIFINAL	−0.743***	0.256	−0.278	−0.798***	0.229	−0.304	−0.724***	0.256	−0.270
FINAL	−0.544	0.350	−0.199	−0.495	0.318	−0.184	−0.547	0.347	−0.201
AUSTRALIA	−0.259***	0.092	−0.086	−0.251***	0.088	−0.087	−0.263***	0.089	−0.088
FRENCH OPEN	−0.150	0.091	−0.049	−0.136	0.088	−0.047	−0.150*	0.089	−0.050
WIMBLEDON	−0.266***	0.091	−0.089	−0.224**	0.088	−0.078	−0.253***	0.089	−0.085
CONSTANT	0.249***	0.089		0.652***	0.078		0.222***	0.078	
Number of observations		1930			1930			2032	
Likelihood ratio test		257			74			252	
Pseudo-R^2		0.111			0.031			0.104	
Log-likelihood		−1025			−1117			−1089	

Note: ME denotes marginal effect.
* Indicates significance at the 10% level.
** Indicates significance at the 5% level.
*** Indicates significance at the 1% level.

Table 4
Predicted probabilities for specific rankings evaluated at the mean of the other variables for the M1 and F1 models.

Ranking	Men	Women
1	0.944	0.961
11	0.794	0.802
51	0.628	0.602
61	0.606	0.576

against right-handed opponents. As a result, left-handed players can often outperform their ATP ranks.

Despite the fact that some players feel more comfortable playing on certain surfaces, we only found a significant court effect regarding higher-ranked player victories in Wimbledon's women's tournament and Australia's women's tournament. Furthermore, none of the round dummies were significant for men, but the semifinal dummy was significant for women.

5. Prediction accuracy of the models

In this section, we evaluate the predictive (in-sample) and forecasting (out-of-sample) accuracies of the different models. The first evaluation of the predictive accuracy relates to the Pseudo-R^2.[8]

[8] The Pseudo-R^2 is calculated as $1 - \ln L / \ln L_0$, where $\ln L$ is the log likelihood of the estimated model, which must include a constant term, and $\ln L_0$ is the log likelihood function for a model that only has a constant (Greene, 2008). Since the log likelihood always takes negative values in probit models, the higher (or lower in absolute terms) the log likelihood of the estimated model, the higher the Pseudo-R^2.

Table 5
The in-sample predictive accuracy for men (2005–2008).

Predicted interval	M1		M2		M3	
	Actual outcome		Actual outcome		Actual outcome	
	Und. victory	Fav. victory	Und. victory	Fav. victory	Und. victory	Fav. victory
0–0.5	**94**	98	**2**	3	**46**	42
	(49%)	(51%)	**(40%)**	(60%)	**(52%)**	(48%)
0.5–1	489	**1341**	581	**1436**	538	**1406**
	(27%)	**(73%)**	(29%)	**(71%)**	(28%)	**(72%)**
0.2–0.3	**1**	0				
	(100%)	(0%)				
0.3–0.4	**18**	16			**2**	2
	(53%)	(47%)			**(50%)**	(50%)
0.4–0.5	**75**	82	**2**	3	**44**	40
	(48%)	(52%)	**(40%)**	(60%)	**(52%)**	(48%)
0.5–0.6	156	**181**	47	**62**	167	**210**
	(46%)	**(54%)**	(43%)	**(57%)**	(44%)	**(56%)**
0.6–0.7	154	**262**	273	**534**	189	**337**
	(37%)	**(63%)**	(34%)	**(66%)**	(36%)	**(64%)**
0.7–0.8	111	**316**	226	**619**	117	**366**
	(26%)	**(74%)**	(27%)	**(73%)**	(24%)	**(76%)**
0.8–0.9	55	**325**	34	**214**	50	**298**
	(14%)	**(86%)**	(14%)	**(86%)**	(14%)	**(86%)**
0.9–1	13	**257**	1	**7**	15	**195**
	(5%)	**(95%)**	(13%)	**(88%)**	(7%)	**(93%)**
Pearson χ^2	219			51		169
Brier score	0.183			0.199		0.187

Und. Victory refers to a lower-ranked player victory, and Fav. Victory refers to a higher-ranked player victory. The predicted interval refers to the probability interval that contains the probability of a higher-ranked player victory. The correctly-predicted matches are shown in bold.

It is clear that the models that use players' past performances outperform those that do not. The same conclusion can be reached by evaluating the likelihood ratio test values; even though all of the models are statistically significant, M1 and M3 are statistically better than M2, and the same thing applies to F1, F3 and F2.

Tables 5 and 6 show contingency tables for the predicted values (i.e., a higher-ranked player victory if the predicted probability is greater than 0.5 and a lower-ranked player victory if the predicted probability is lower than 0.5) and the actual outcomes for men and women, respectively. In addition, in order to analyze the predictions in detail across the entire distribution, the intervals (i.e., 0 to 0.5 and 0.5 to 1) are split into decimal intervals.

Note that 82% and 78% of the observations are predicted in the interval between 0.6 and 0.8 in M2 and F2, respectively. However, this percentage is lower than 50% for the models that use ranking information. That is, most of the predictions generated by the models that do not use ranking information are close to the mean probability (0.71), whereas the predictions of the models that use ranking information are not as concentrated around that value. In addition, it is important to note that the accuracy of the predictions at the tails of M1, F1, M3 and F3 is rather high: it is close to 95% at the upper tail (i.e., predictions above 0.9) and over 50% at the lower tail (i.e., predictions below 0.4).

Another indicator of predictive accuracy is the Brier score.[9] The lowest values of the Brier score were generated for M1 and F1, whereas the highest values

[9] The Brier score is defined as:

$$B = \frac{\sum_{i=1}^{N} (P_i - X_i)^2}{N},$$

where P is the predicted probability that the higher-ranked player wins, and X takes the value of one if the higher-ranked player wins

Table 6
The in-sample predictive accuracy for women (2005–2008).

Predicted interval	F1		F2		F3	
	Actual outcome		Actual outcome		Actual outcome	
	Und. victory	Fav. victory	Und. victory	Fav. victory	Und. victory	Fav. victory
0–0.5	**91**	57	**22**	16	**72**	46
	(61%)	(39%)	**(58%)**	(42%)	**(61%)**	(39%)
0.5–1	459	**1323**	528	**1364**	508	**1406**
	(26%)	**(74%)**	(28%)	**(72%)**	(27%)	**(73%)**
0–0.1	**1**	0				
	(100%)	(0%)				
0.1–0.2	**5**	0			**7**	0
	(100%)	(0%)			**(100%)**	(0%)
0.2–0.3	**13**	4	**1**	1	**9**	2
	(76%)	(24%)	**(50%)**	(50%)	**(82%)**	(18%)
0.3–0.4	**13**	10	**8**	6	**14**	11
	(57%)	(43%)	**(57%)**	(43%)	**(56%)**	(44%)
0.4–0.5	**59**	43	**13**	9	**42**	33
	(58%)	(42%)	**(59%)**	(41%)	**(56%)**	(44%)
0.5–0.6	133	**189**	37	**52**	162	**201**
	(41%)	**(59%)**	(42%)	**(58%)**	(45%)	**(55%)**
0.6–0.7	164	**267**	236	**457**	178	**316**
	(38%)	**(62%)**	(34%)	**(66%)**	(36%)	**(64%)**
0.7–0.8	99	**305**	207	**598**	100	**328**
	(25%)	**(76%)**	(26%)	**(74%)**	(23%)	**(77%)**
0.8–0.9	44	**282**	48	**237**	50	**299**
	(14%)	**(87%)**	(17%)	**(83%)**	(14%)	**(86%)**
0.9–1	19	**280**	0	**20**	18	**262**
	(6%)	**(94%)**	(0%)	**(100%)**	(6%)	**(94%)**
Pearson χ^2	243			64		237
Brier score	0.178			0.196		0.180

Und. Victory refers to a lower-ranked player victory, while Fav. Victory refers to a higher-ranked player victory. The predicted interval refers to the probability interval that contains the probability of a higher-ranked player victory. The correctly-predicted matches are shown in bold.

were generated for the models that do not include ranking information (i.e., M2 and F2). Moreover, the Brier scores for M3 and F3 are closer to those of M1 and F1 than to those of M2 and F2. Therefore, it seems that the variables pertaining to the past performance of the player are the most important for increasing the accuracy of the predictions.

Table 7 shows the in-sample Brier scores for the different models organized by tournament. For all of the tournaments, the highest Brier scores belong to the models that do not account for ranking differences (i.e., M2 and F2). Moreover, the highest Brier scores are obtained for the French Open and Wimbledon (for men) and for the Australian Open (for women).

and zero otherwise. Smaller values of B indicate more accurate predictions.

We now analyze the out-of-sample forecasting accuracy. We collected data from the 2009 Australian Open (see Tables 8 and 9). To analyze the out-of-sample forecasting accuracy in this tournament, we use the estimates from Tables 2 and 3 to obtain predicted values, and subsequently compare the predicted values with the actual outcomes. Once again, M2 and F2 exhibit the poorest performance. Therefore, the most important variables for improving the forecasting accuracy are those related to the past performances of the players.

In addition to analyzing the 2009 Australian Open, we also checked the out-of-sample forecasting accuracy of the models using bootstrapping techniques. In each trial, we randomly selected 25% of the 2005–2008 sample in order to estimate the different models. We subsequently calculated the Brier score

Table 7
The in-sample predictive accuracy by tournament.

Tournament	Male			Female		
	M1	M2	M3	F1	F2	F3
Australian Open	0.175	0.188	0.175	0.180	0.206	0.185
French Open	0.191	0.205	0.196	0.178	0.196	0.182
Wimbledon	0.180	0.205	0.188	0.188	0.203	0.186
US Open	0.185	0.200	0.190	0.167	0.180	0.168

Table 8
The out-of-sample forecasting accuracy for men (2009 Australian Open).

Predicted interval	M1		M2		M3	
	Actual outcome		Actual outcome		Actual outcome	
	Und. victory	Fav. victory	Und. victory	Fav. victory	Und. victory	Fav. victory
0.2–0.3	**0**	1				
	(0%)	(100%)				
0.3–0.4	**2**	1				
	(67%)	(33%)				
0.4–0.5	**6**	1			**2**	1
	(86%)	(14%)			**(67%)**	(33%)
0.5–0.6	6	**7**	1	**1**	5	**8**
	(46%)	**(54%)**	(50%)	**(50%)**	(38%)	**(62%)**
0.6–0.7	6	**19**	10	**14**	14	**21**
	(24%)	**(76%)**	(42%)	**(58%)**	(40%)	**(60%)**
0.7–0.8	5	**14**	14	**47**	8	**22**
	(26%)	**(74%)**	(23%)	**(77%)**	(27%)	**(73%)**
0.8–0.9	4	**28**	4	**25**	4	**24**
	(13%)	**(88%)**	(14%)	**(86%)**	(14%)	**(86%)**
0.9–1	1	**17**	1	**1**	0	**18**
	(6%)	**(94%)**	(50%)	**(50%)**	(0%)	**(100%)**
Pearson χ^2	26		7		15	
Brier score	0.158		0.182		0.167	

Und. Victory refers to a lower-ranked player victory, while Fav. Victory refers to a higher-ranked player victory. The predicted interval refers to the probability interval that contains the probability of a higher-ranked player victory. The correctly-predicted matches are shown in bold.

of the estimated models for the remaining 75% of the sample and repeated these steps 12,500 times. As a result, we obtained a distribution of Brier scores. In Figs. 3 and 4, we show the kernel densities of the bootstrapped Brier scores for the three alternative models for men and women.

Both figures show that the models that do not use the ranking information (M2 and F2) have the largest out-of-sample Brier scores, that is, the worst forecasting accuracies. It is interesting to note that the distributions of the Brier scores for M2 and F2 do not collapse with the other distributions. In addition, the player's physical characteristic variables are more important for men than for women, since the Brier score distributions for F1 and F3 are more similar to each other than those of M1 and M3.

6. Conclusions

This study has tested whether differences in rankings between individual players are good predictors for Grand Slam tennis outcomes. In so doing, we have separately estimated three alternative probit models for men and women. The most relevant explanatory variable appears to be the difference in the ATP or WTA rankings, since it is the only significant variable across all of the models. We also found that this difference in ranking effect is not statistically different be-

Table 9
The out-of-sample forecasting accuracy for women (2009 Australian Open).

Predicted interval	F1		F2		F3	
	Actual outcome		Actual outcome		Actual outcome	
	Und. victory	Fav. victory	Und. victory	Fav. victory	Und. victory	Fav. victory
0.1–0.2	**0**	1			**0**	1
	(0%)	(100%)			**(0%)**	(100%)
0.2–0.3	**0**	2			**0**	2
	(0%)	(100%)			**(0%)**	(100%)
0.3–0.4	**1**	1	**0**	1	**0**	1
	(50%)	(50%)	**(0%)**	(100%)	**(0%)**	(100%)
0.4–0.5	**4**	5	**0**	2	**2**	3
	(44%)	(56%)	**(0%)**	(100%)	**(40%)**	(60%)
0.5–0.6	8	**11**	3	**3**	14	**16**
	(42%)	**(58%)**	(50%)	**(50%)**	(47%)	**(53%)**
0.6–0.7	7	**16**	18	**39**	11	**18**
	(30%)	**(70%)**	(32%)	**(68%)**	(38%)	**(62%)**
0.7–0.8	8	**6**	9	**27**	4	**17**
	(57%)	**(43%)**	(25%)	**(75%)**	(19%)	**(81%)**
0.8–0.9	2	**19**	1	**3**	4	**20**
	(10%)	**(90%)**	(25%)	**(75%)**	(17%)	**(83%)**
0.9–1	2	**14**	1	**0**	0	**14**
	(13%)	**(88%)**	(100%)	**(0%)**	(0%)	**(100%)**
Pearson χ^2	15		5		16	
Brier score	0.175		0.219		0.193	

Und. Victory refers to a lower-ranked player victory, while Fav. Victory refers to a higher-ranked player victory. The predicted interval refers to the probability interval that contains the probability of a higher-ranked player victory. The correctly-predicted matches are shown in bold.

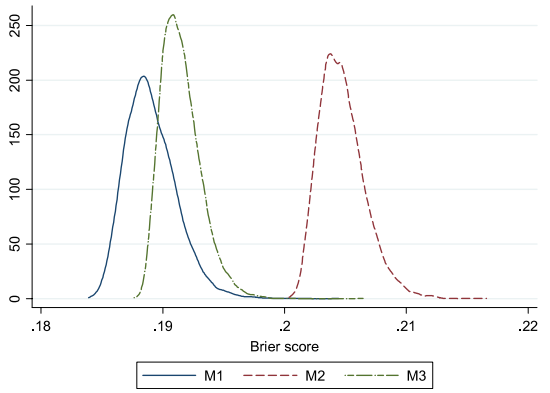

Fig. 3. The kernel densities of the bootstrapped Brier scores for out-of-sample accuracy with regard to M1, M2 and M3 (2005–2008).

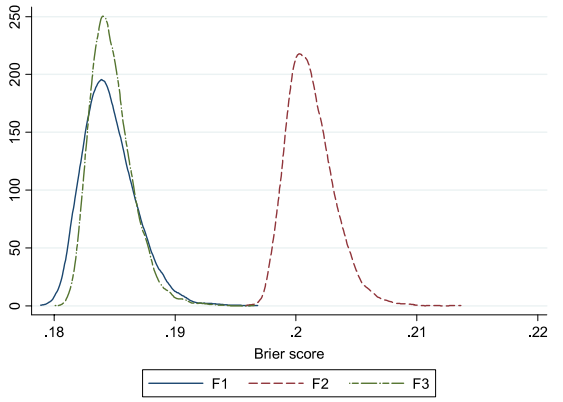

Fig. 4. The kernel densities of the bootstrapped Brier scores for out-of-sample accuracy with regard to F1, F2 and F3 (2005–2008).

tween men and women, and that rank differences are more important as we move to the top of the distribution of players for both men and women. Another significant effect relating to a player's past performance is the individual-tournament effect, but this is only significant for men. On the other hand, the variable related to being a previous top-ten player is a more relevant predictor of victory among women than men.

With regard to personal characteristics, age differences have a significant effect for both men and women. In particular, we found that the probability of a higher-ranked player victory decreases as s/he plays

Table A.1
DIFRANKING correspondence for specific rank positions.

Higher ranked	Lower ranked																
	2	3	4	5	6	7	8	9	10	20	30	40	50	75	100	150	200
1	0.69	1.10	1.39	1.61	1.79	1.95	2.08	2.20	2.30	3.00	3.40	3.69	3.91	4.32	4.61	5.01	5.30
2		0.41	0.69	0.92	1.10	1.25	1.39	1.50	1.61	2.30	2.71	3.00	3.22	3.62	3.91	4.32	4.61
3			0.29	0.51	0.69	0.85	0.98	1.10	1.20	1.90	2.30	2.59	2.81	3.22	3.51	3.91	4.20
4				0.22	0.41	0.56	0.69	0.81	0.92	1.61	2.01	2.30	2.53	2.93	3.22	3.62	3.91
5					0.18	0.34	0.47	0.59	0.69	1.39	1.79	2.08	2.30	2.71	3.00	3.40	3.69
6						0.15	0.29	0.41	0.51	1.20	1.61	1.90	2.12	2.53	2.81	3.22	3.51
7							0.13	0.25	0.36	1.05	1.46	1.74	1.97	2.37	2.66	3.06	3.35
8								0.12	0.22	0.92	1.32	1.61	1.83	2.24	2.53	2.93	3.22
9									0.11	0.80	1.20	1.49	1.71	2.12	2.41	2.81	3.10
10										0.69	1.10	1.39	1.61	2.01	2.30	2.71	3.00
20											0.41	0.69	0.92	1.32	1.61	2.01	2.30
30												0.29	0.51	0.92	1.20	1.61	1.90
40													0.22	0.63	0.92	1.32	1.61
50														0.41	0.69	1.10	1.39
75															0.29	0.69	0.98
100																0.41	0.69
150																	0.29

against younger players. However, the pattern of this effect is very different across genders: it decreases monotonically for men but has an inverted U-shaped pattern for females.

After estimation, the in-sample and out-of-sample accuracies of the models were evaluated. The main result is that past performance variables, including ranking information, are the most relevant variables for increasing the prediction accuracy. In fact, most of the predictions from the models that do not use ranking information are close to the mean probability, whereas the predictions from the models that include ranking information are not as concentrated around that value. Moreover, forecasting differences are especially relevant at the tails of the probability distribution. For instance, models that account for ranking differences predict a considerable number of victories for lower-ranked players, whereas those without ranking information do not. Therefore, the models that do not include ranking variables perform much worse than those that do.

With regard to the out-of-sample forecasting accuracy, we collected data from the 2009 Australian Open. Importantly, these data were not used in the estimations; rather, we computed the predicted values for this tournament according to the estimated models and then compared these estimates with the actual, observed outcomes. Once again, the models that did not include past performance information had the poorest performances.

In order to avoid a tournament-specific bias, we verified the out-of-sample forecasting accuracy using bootstrapping techniques for the 2005–2008 sample. In each trial, we estimated the probit models by randomly sampling 25% of the original sample, then calculating the Brier scores for the other 75% of the sample. Once again, the results demonstrated that models that do not use ranking information have the largest out-of-sample Brier scores, that is, the worst forecasting accuracies.

Acknowledgements

This research has benefited from the Spanish Ministry of Science and Technology Projects ECO2008-04659 and SEJ2007-64700/ECON. Julio del Corral also acknowledges the Spanish Ministry of Education for his FPU fellowship. This paper has benefited from helpful comments by Markus Lang, two anonymous referees and the editor Herman Stekler. Finally, the authors would also like to thank Julio del Corral Sr. for his research assistance. Any remaining errors are the sole responsibility of the authors.

Appendix

See Table A.1.

References

Abrevaya, J. (2002). Ladder tournaments and underdogs: Lessons from professional bowling. *Journal of Economic Behavior and Organization*, *47*, 87–101.

Anderson, P., Edman, J., & Ekman, M. (2005). Predicting the World Cup 2002 in soccer: Performance and confidence of experts and non-experts. *International Journal of Forecasting*, *21*, 565–576.

Bolton, R., & Chapman, R. (1986). Searching for positive returns at the track: A multinomial logit model for handicapping horse races. *Management Science*, *32*, 1040–1060.

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

Boulier, B., & Stekler, H. (2003). Predicting the outcomes of National Football League games. *International Journal of Forecasting*, *19*, 257–270.

Cain, M., Law, D., & Peel, D. (2000). The favourite-longshot bias and market efficiency in UK football betting. *Scottish Journal of Political Economy*, *47*, 25–36.

Caudill, S. (2003). Predicting discrete outcomes with the maximum score estimator: The case of the NCAA men's basketball tournament. *International Journal of Forecasting*, *19*, 313–317.

Caudill, S., & Godwin, N. (2002). Heterogeneous skewness in binary choice models: Predicting outcomes in the men's NCAA basketball tournament. *Journal of Applied Statistics*, *29*, 991–1001.

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

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

Dyte, D., & Clarke, S. (2000). A ratings based Poisson model for World Cup soccer simulation. *The Journal of the Operational Research Society*, *51*(8), 993–998.

Forrest, D., Goddard, J., & Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, *21*, 551–564.

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

Forrest, D., & Simmons, R. (2000). Forecasting sport: The behavior and performance of football tipsters. *International Journal of Forecasting*, *16*, 317–331.

Gilsdorf, K., & Sukhatme, V. (2007). Testing Rosen's sequential elimination tournament model. Incentives and player performance in professional tennis. *Journal of Sports Economics*, *9*, 287–303.

Goddard, J. (2005). Regression models for forecasting goals and match results in Association Football. *International Journal of Forecasting*, *21*, 331–340.

Goddard, J., & Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, *23*, 51–66.

Greene, W. (2008). Discrete choice modeling. In T. Mills, & K. Patterson (Eds.), *The handbook of econometrics: Vol. 2, Applied econometrics* (Part 4.2). London: Palgrave.

Klaassen, F., & Magnus, J. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, *96*, 500–509.

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

Lebovic, J., & Sigelman, L. (2001). The forecasting accuracy and determinants of football rankings. *International Journal of Forecasting*, *17*, 105–120.

Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *The Statistician*, *49*, 399–418.

Smith, T., & Schwertman, N. (1999). Can the NCAA basketball tournament seeding be used to predict margin of victory? *The American Statistician*, *53*(2), 94–98.

Julio del Corral is a visiting professor of Economics at the University of Castilla la Mancha. He received his Ph.D. in Economics from the University of Oviedo. His fields of specialization are efficiency and productivity analysis and sports economics. He has already published articles in the *Journal of Dairy Science*, *Journal of Sports Economics* and *Revista de Economía Aplicada*.

Juan Prieto-Rodríguez is an assistant professor of Economics at the University of Oviedo. He received his Ph.D. in Economics from the University of Oviedo. His fields of specialization are cultural, leisure and labor economics. He has published articles in *Economics Letters*, *Fiscal Studies*, *Journal of Economic Psychology* and *Journal of Sports Economics*.