Weighted Elo rating for tennis match predictions

Published Version:
Weighted Elo rating for tennis match predictions / Angelini, Giovanni; Candila, Vincenzo; De Angelis, Luca. - In: EUROPEAN JOURNAL OF OPERATIONAL RESEARCH. - ISSN 0377-2217. - ELETTRONICO. - 297:1(2022), pp. 120-132. [10.1016/j.ejor.2021.04.011]

Availability:
This version is available at: https://hdl.handle.net/11585/821483 since: 2021-06-04

Published:
DOI: http://doi.org/10.1016/j.ejor.2021.04.011

Weighted Elo rating for tennis match predictions

Giovanni Angelini[a], Vincenzo Candila[b], Luca De Angelis[c,*]

[a]Dept. of Economics, University of Bologna, g.angelini@unibo.it
[b]MEMOTEF, Sapienza University of Rome, Italy, vincenzo.candila@uniroma1.it
[c]Dept. of Economics, University of Bologna, l.deangelis@unibo.it

Abstract

Originally applied to tennis by the data journalists of FiveThirtyEight.com, the Elo rating method estimates the strength of each player based on her/his career as well as the outcome of the last match played. Together with the regression-based, point-based and paired-comparison approaches, the Elo rating is a popular method to predict the probability of winning tennis matches. Notwithstanding its widely recognized merits in terms of ease of reproducibility and good performance, the Elo method does not completely take into account the current form of each player and their recent performances. This paper proposes a new version of the Elo rating method, labelled Weighted Elo (WElo), where the standard Elo updating is additionally weighted according to the scoreline of the players' last match. The proposed method considers not only if a player has won (lost) a match, but also how the victory (defeat) was achieved. In the empirical application, the forecasting performance of the WElo method is evaluated and compared against the most popular forecasting methods in tennis, using a sample of over 60,000 men's and women's professional matches. Overall, the WElo method outperforms all these competing methods. Moreover, it provides meaningfully profitable opportunities, according to a simple betting strategy.

Keywords: Forecasting; Elo rating; Tennis; Betting strategy.

[*] Corresponding author. Postal address: Piazza Scaravilli 2, 40126 - Bologna, Italy. E-mail: l.deangelis@unibo.it

We are grateful to the Editor, Ruud Teunter, and three anonymous referees for their helpful comments on the earlier version of this paper. We also thank the participants at the 12th International Conference of the ERCIM on Computational and Methodological Statistics (CMStatistics) in London (December 2019) for their useful comments.

1. Introduction

In recent years, the academic interest on sports forecasting models has grown significantly. In particular, the number of contributions on the model-based prediction of the outcome of sport events has enormously increased. One of the main purposes of forecasting sport outcomes is to detect and exploit the (informational) inefficiency of sports betting markets which allows better informed bettors to gain at the expense of the less well informed. However, the debate on the possibility of achieving positive returns by using a single or combination of sports forecasting models is still open. Most of the papers in the literature are focused on the forecasting of football match outcomes (e.g., Angelini and De Angelis (2017) and Koopman and Lit (2015)). However, despite its popularity among the betting community, the number of academic contributions on tennis forecasting is rather limited. A survey of these contributions is provided by Kovalchik (2016), where three categories of methods are reviewed and evaluated, namely the regression-based, point-based and paired comparison approaches. Some examples of regression-based contributions are Del Corral and Prieto-Rodriguez (2010), Klaassen and Magnus (2003), Clarke and Dyte (2000) and Boulier and Stekler (1999), among others. Generally, these contributions make use of the probit or logit estimators. For instance, Klaassen and Magnus (2003) propose a logit regression based on the Association of Tennis Professionals (ATP) rankings of two players to predict the winner of the match. Using a probit model, Del Corral and Prieto-Rodriguez (2010) find that the recent performances of a player are the most important factors to forecast match outcomes. Contributions belonging to the point-based approach focus on the prediction of winning a single point. Some examples are Knottenbelt et al. (2012), Barnett et al. (2006) and Barnett and Clarke (2005), among others. Finally, the paired-comparison approach estimates the probability of winning on the basis of the direct evaluation of the latent abilities of the two players. Examples of this approach are the Bradley-Terry (BT) type model (McHale and Morton, 2011), its time-varying version proposed by Gorgi et al. (2019), the dynamic paired comparisons model proposed by Baker and McHale (2014) and Baker and McHale (2017), and the Elo rating system. A recent implementation of many of these approaches into the neural network framework to forecast the tennis winning player is provided by Candila and Palazzo (2020).

The Elo rating method recursively estimates the strength of each contender on the basis of the last match, as well as her/his whole career, in order to predict the probability of winning for the upcoming match. Originally proposed by the Physics Professor Arpad Elo in 1978 (Elo, 1978) for the rating of chess players, the Elo method has been applied in different sports, such as soccer (Hvattum and Arntzen, 2010; Leitner et al., 2010), American football (Ryall and Bedford, 2010), rugby (Carbone et al., 2016), and tennis (Kovalchik, 2016; Kovalchik and Reid, 2019). Notwithstanding its widely recognized merits in terms of ease of reproducibility and good performance, the traditional Elo method does not take into account how the last match outcome of one player (or team) has been achieved, but only if the player (team) has won or lost the match. Operatively, the standard Elo increases (decreases) the Elo rating of a player/team by the same amount, independently of how the player/team has won (lost) the match. In this paper we generalize the Elo rating system in tennis by weighting the matches on the basis of the number of games (or sets) won by each player. The aim of such a generalization is to include the informative content provided by the most recent performance, i.e. *momentum* in financial lingo, of players/teams, which is generally recognized as crucial in sports forecasting. The use of data on recent results has been proved to significantly improve the forecasting performance of a statistical model in many sports; see, for instance, Angelini and De Angelis (2017) for football, and Del Corral and Prieto-Rodriguez (2010) for tennis. In line with this literature, we investigate to what extent this "hot hand" phenomenon − as well as its opposite − can significantly enrich the available information set on which the forecast is based, by comparing the forecasting performance of the standard Elo rating system and the one of a weighted version, which incorporate information from the latest performance of the players. In particular, we consider the number of games (or sets) won and lost by each player in the last matches as proxies for the recent condition and form of the two opponents. Our approach can be easily generalised to many other individual and team sports, in which the point difference can be considered as proxies for the player/team's recent condition. A similar idea was applied by Hvattum and Arntzen (2010) to association football, where a weighted function for the Elo update based on the difference of the goals scored by the two teams is considered.

From a statistical point of view, we demonstrate that the proposed variant of the Elo, labeled Weighted Elo (WElo), is a symmetric rating system, even when the number of games of the opponents are taken into account. Moreover, the WElo confidence intervals are derived using an *ad hoc* bootstrap-based technique. In a very interesting recent paper, Kovalchik (2020) introduces a general framework where the Elo system is extended to account for the margin of victory (MOV) in tennis, measured by either the number of sets won or other within-set statistics, namely the number of games, the break points, the total points, and the serve percentage. Despite our approach can be considered as a special case of her multiplicative model for Elo updating where the number of games won defines the MOV, the WElo system can be computed and updated straightforwardly, hence gaining in terms of reproducibility, as there is no need to estimate or arbitrarily set tuning parameters and, according to our formulation of the MOV, the chance that the update from a single event can be unrealistically large is ruled out. Moreover, with respect to Kovalchik (2020), who compares the predictive performance of the MOV-based approaches against the standard Elo, we extend such comparison to other popular approaches in tennis. In particular, in the empirical analysis, the WElo forecasting performance is extensively evaluated over a sample of over 60,000 men's and women's matches and compared against the performance of four competing models, namely the standard Elo and the BT model for the paired-based approach, and the logit and probit regressions of Klaassen and Magnus (2003) and Del Corral and Prieto-Rodriguez (2010), respectively, for the regression-based approach.

Our results show that the WElo approach allows more accurate predictions, both in terms of minimization of the loss functions considered, namely the Brier score (Brier, 1950) and log-loss function, and maximization of the returns-on-investment (ROI) achieved on the betting markets. In particular, the proposed approach significantly outperforms all the four competing models, independently of the out-of-sample period considered. In terms of ROI, the WElo yields approximately 3.56% (2.93%) of the investment from betting on the men's (women's) matches in the period 2012-2020, if the best odds avaiable on the market are considered. These results are significantly better than what is obtained using the standard Elo or a random betting strategy.

The remainder of the paper is organized as follows. Section 2 outlines the proposed WElo rating system and its properties. Section 3 shows the results of the out-of-sample forecasting performance. Section 4 describes the betting strategy and its performance based on the Elo and WElo methods. Section 5 concludes. Finally, in Appendixes A and B we report results from robustness checks and further analyses. In particular, Appendix A reports the evaluation of the WElo based on sets (instead of games) for the men's matches while Appendix B is devoted to the analysis concerning the data set on women's (WTA) professional matches.

2. Weighted Elo

Let i and j be two opponents in a tennis match and $E_i(t)$ and $E_j(t)$ their Elo ratings for the match at time t. According to the Elo rating system, once $E_i(t)$ and $E_j(t)$ are estimated, the probability that player i wins against player j in match t is defined as:

$$\hat{p}_{i,j}(t) = \frac{1}{1+10^{\left(E_j(t)-E_i(t)\right)/400}}. \tag{1}$$

For the ease of notation, from now on we will only refer to player i. This means that we always refer to the same player i, irrespectively of the match, t, while the opponent player, j, will be (usually) different according to the match. The updating procedure, in a Bayesian sense (Glickman, 1999), for calculating the Elo ratings for player i is:

$$E_i(t+1) = E_i(t) + K_i(t)\left[W_i(t) - \hat{p}_{i,j}(t)\right], \tag{2}$$

where $W_i(t)$ is an indicator function which equals one if player i wins match t and zero otherwise, and $K_i(t)$ is a (positive) scale factor which determines how much the Elo rating should change after match t. The scale factor may also assume different values according to the importance of the match under consideration. For instance, $K_i(t)$ could be larger for matches played in prestigious tournaments, such as the ATP Grand Slams. Another possibility is to make explicit the dependence of the scale factor on the number of matches played by player i, such that the larger the number of matches played by player i, the smaller the scale factor.

Independently of the scale factor $K_i(t)$, the Elo rating for player i in (2) increases if player i won the previous match and decreases in the opposite case, i.e. $E_i(t+1) > E_i(t)$ if i wins match t and $E_i(t+1) < E_i(t)$ if i loses match t. Moreover, from Eq. (2) it is evident that the more player i was likely to win against player j in match t (e.g. in the case of $\hat{p}_{i,j}(t)$ in (1) close to 1), the smaller the Elo rating increases for the upcoming match, and vice versa. But, independently of the scoreline of match t, the Elo rating for player i increases or decreases by the same amount.

The proposed WElo method extends the standard Elo formulation in order to take into account not only if one player won or lost the match but also the scoreline of the match, in order to exploit the information on how the victory (or defeat) has been achieved. Specifically, such a weighted rating system is obtained by incorporating in Eq. (2) an additional function $f(\cdot)$, which depends on the number of games $G_{i,j}(t)$ won by players i and j during match t.

The WElo rating system is then defined as:

$$E_i^*(t+1) = E_i^*(t) + K_i(t)\left[W_i(t) - \hat{p}_{i,j}^*(t)\right] f(G_{i,j}(t)), \tag{3}$$

where $\hat{p}_{i,j}^*(t)$ is estimated using Eq. (1) where $E_i(t)$ and $E_j(t)$ are replaced by the corresponding WElo ratings, $E_i^*(t)$ and $E_j^*(t)$, respectively, and $f(G_{i,j}(t))$ is a function whose values depend on the games played in the previous match. In particular, $f(G_{i,j}(t))$ is defined as:

$$f(G_{i,j}(t)) = \begin{cases} \frac{NG_i(t)}{NG_i(t)+NG_j(t)} & \text{if player } i \text{ has won match } t; \\ \frac{NG_j(t)}{NG_i(t)+NG_j(t)} & \text{if player } i \text{ has lost match } t, \end{cases}$$

where $NG_i(t)$ and $NG_j(t)$ represent the number of games won by player i and player j in match t, respectively. It is worth noting that the structure of the WElo rating system is very flexible and the additional function $f(\cdot)$ can be easily specified using other variables. An example is provided in Appendix A where we report the results for the WElo method in Eq. (3) with a function $f(\cdot)$ based on sets (instead of games).

By construction, the Elo rating system represents an upper bound for the WElo system, that is $E_i^*(t) \leq E_i(t)$. As an illustration, Table 1 reports four cases that highlight the differences between the Elo and WElo ratings.

Table 1: Comparison of WElo and Elo rating systems: examples for player i

	Winner	Score	$\hat{p}_{i,j}(t)$	$f(G_{i,j}(t))$	$\left[W_i(t) - \hat{p}_{i,j}(t)\right] f(G_{i,j}(t))$	$\left[W_i(t) - \hat{p}_{i,j}(t)\right]$
1	i	6-0;6-0	0.60	1.00	+ 0.40	+ 0.40
2	i	7-6;7-6	0.60	0.54	+ 0.22	+ 0.40
3	i	0-6;7-6;7-6	0.95	0.44	+ 0.02	+ 0.05
4	j	1-6;1-6	0.75	0.86	– 0.64	– 0.75

An important feature of the standard Elo rating system is the symmetry of the ratings. This means that the “points” earned by the winning player equal the “points” lost by the defeated player, before the multiplication by the scale factor $K_i(t)$, which could be different between the two players. It is easy to note that, since the additional function $f(G_{i,j}(t))$ is symmetric, the symmetry of the ratings is also maintained in the case of the WElo system. Indeed, it is straightforward to show that the WElo points earned by the winning player, say i, are the WElo points lost by player j, in absolute value:

$$\underbrace{\left[W_i(t) - \hat{p}^*_{i,j}(t)\right] f(G_{i,j}(t))}_{i \text{ wins}} = \underbrace{\left| \left[W_j(t) - \hat{p}^*_{j,i}(t)\right] f(G_{j,i}(t))\right|}_{j \text{ loses}}.$$

This result implies that $\hat{p}^*_{i,j}(t) + \hat{p}^*_{j,i}(t) = 1$, for all t.

2.1. Confidence Intervals

In this section we propose a novel approach for computing the confidence intervals of the WElo rating system. The WElo rating described in (3) can be decomposed as follows:

$$\begin{aligned} E^*_i(t+1) &= E^*_i(t) + K_i(t)\left[W_i(t) - \hat{p}^*_{i,j}(t)\right] f(G_{i,j}(t)) \\ &= E^*_i(t-1) + K_i(t-1)\left[W_i(t-1) - \hat{p}^*_{i,j}(t-1)\right] f(G_{i,j}(t-1)) \\ &\quad + K_i(t)\left[W_i(t) - \hat{p}^*_{i,j}(t)\right] f(G_{i,j}(t)) \\ &= E^*_i(0) + \sum_{s=1}^{t} K_i(s)\left[W_i(s) - \hat{p}^*_{i,j}(s)\right] f(G_{i,j}(s)), \end{aligned} \tag{4}$$

where $W_i(s) \sim \mathscr{B}(\pi_{i,j}(s))$ follows a Bernoulli distribution with true probability $\pi_{i,j}(s)$ that player i defeats player j at time s. Note that the recursion in Eq. (4) highlights the dependence of the WElo rating for the following match $(t+1)$ on all player's career, i.e. from her/his first match $(s = 1)$ to her/his latest match $(s = t)$.

Assumption 1. *Let $W_i(s) \sim \mathscr{B}(\pi_{i,j}(s))$, then it holds that $E(\hat{p}^*_{i,j}(s)) = \pi_{i,j}(s)$ for every $s \in \{1,...,t\}$.*

Under Assumption 1, the confidence intervals can be computed using the following algorithm:

Algorithm 1. (Bootstrap) confidence intervals:

(i) Estimate the WElo in (3) and compute $\hat{p}^*_{i,j}(s)$;

(ii) Generate $\widetilde{W}_i(s)$ from the Bernoulli distribution $W_i(s) \sim \mathscr{B}(\hat{p}^*_{i,j}(s)|\mathscr{F}_{s-1})$ for every $s \in \{1,...,t\}$, where $\mathscr{F}_{s-1}$ is the information set available at time $s-1$;

(iii) Generate $\widetilde{E}_i^*(s)$ for every $s \in \{1,...,t\}$, recursively from (4), by replacing $W_i(s)$ with $\widetilde{W}_i(s)$;

(iv) Repeat (i) and (ii) B times in order to obtain the sequences $\widetilde{E}_i^*(s)^{b=1}, \cdots, \widetilde{E}_i^*(s)^{b=B}$, for every $s \in \{1,...,t\}$;

(v) For every $s \in \{1,...,t\}$, the $(1-\alpha)\%$ confidence intervals are the α percentile and the $1-\alpha$ percentile of $\widetilde{E}_i^*(s)^b$.

This algorithm can be adapted for the Elo by setting $f(G_{i,j}(s)) = 1$.

3. Empirical analysis

All the empirical analysis presented in this section considers the data set on men's professional matches. In Appendix B we report the same analysis applied to women's matches. The dataset used in the empirical analysis is taken from the historical archive of the website `www.tennis-data.co.uk`. This archive includes the match results of all the most important tennis tournaments of the ATP Tour (Masters 250, 500 and 1000, and ATP Finals) and Grand Slams, closing betting odds of different professional bookmakers, as well as the ranking and the ATP points for each player.[1] The period under consideration spans from July 4, 2005 to November 22, 2020. All the results in this paper are obtained using the R package 'welo'. The initial number of matches is 38,868. After the data cleaning process, the final dataset reduces to 33,976 matches.[2]

It is worth noting that the starting value for both the Elo and WElo rating systems is 1,500 points for each player at the beginning of the estimation sample. The scale factor $K_i(t)$ in (2) and (3) for the Elo and WElo rating systems, respectively, is set according to Kovalchik (2016), that is $K_i(t) = 250/(N_i(t)+5)^{0.4}$, where $N_i(t)$ is the number of matches of player i at time t. This setting amplifies the variation of the Elo and WElo ratings if player i has played a small number of matches, while the rate variation tends to decrease as the number of matches

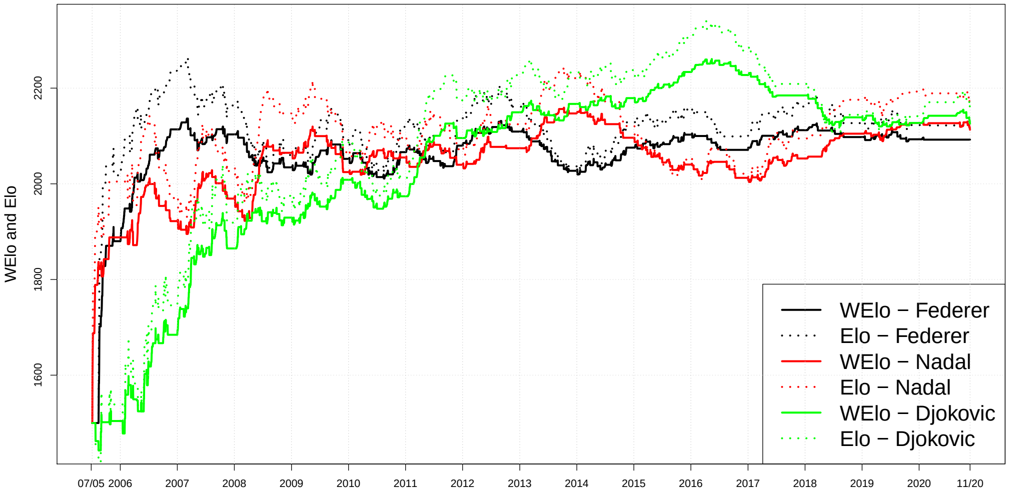
[1] The bookmakers considered are: Bet365, Bet&Win, Centrebet, Expert, Ladbrokes, Gamebookers, Interwetten, Pinnacles, Sportingbet, Stan James, and Unibet.

[2] From the initial data set, we have removed the uncompleted matches and the matches with missing values in the variables used in at least one of the competing approaches listed in Table 2. The cleaning process makes use of the *clean* function of the R package 'welo' to which we refer the reader for all the details.

played by player i increases.[3]

As an example, Figure 1 reports the Elo (dotted lines) and the WElo (solid lines) for three top ATP players: Roger Federer (black lines), Rafael Nadal (red lines) and Novak Djokovic (green lines). Figure 1 shows that substantial differences between the player's ratings arise if one considers the WElo instead of the Elo system to compute the probabilities in (1).

Figure 1: WElo and Elo ratings for Federer, Nadal and Djokovic (July 2005–November 2020)

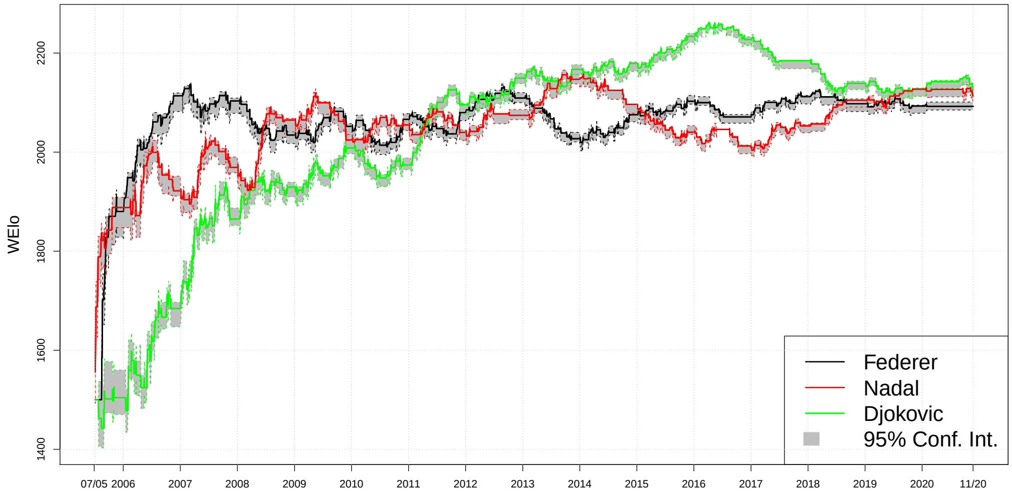


Using Algorithm 1, Figure 2 reports the 95% WElo confidence intervals computed with $B = 1,000$ replicates for the three top players considered in Figure 1. Comparing the WElo ratings with the official ATP rankings, we can observe that, for instance, in September 2018, the ATP rankings were 1, 2 and 6 for Nadal, Federer and Djokovic, respectively. According to the WElo rating system, the ranking of these three players was the opposite of the official one. However, the confidence intervals reported in Figure 2 show that the WElo ratings for Nadal and Federer were not statistically different from each other at that time. Moreover, Djokovic's WElo rating was significantly larger than the one of the other two players. It is worth noting that, since they are computed using the results of a player in the last year, the official ATP rankings tend to penalise injured players, as a player's inactivity does not allow them to add any points and thus

[3] We also repeated our empirical analysis using the other formulations of the scale factor $K_i(t)$ that are used in the literature (e.g. see Kovalchik (2016)). Our results are robust such scale factors. These results are available upon request and can be replicated using the R package 'welo'.

improve or maintain their ranking. For instance, the fact that Djokovic ranked sixth in September 2018 was mainly due to his inactive period in January-April 2018. Conversely, WElo and Elo rating systems do not apply such penalty and their value remains constant if the player is inactive; see, e.g. Djokovic's WElo between January and April 2018 depicted in Figure 1. In principle, this property should provide enhanced forecasting ability with respect to the official ATP rankings. As a matter of fact, Djokovic reached #1 of the ATP ranking a couple of months later (November 2018), as was anticipated by the WElo rating system.

Figure 2: WElo ratings and their 95% confidence intervals for Federer, Nadal and Djokovic (July 2005–November 2020)



3.1. Out-of-sample analysis

In this section, we provide an out-of-sample analysis of the WElo rating system and we compare its forecasting performance with a set of competing methods. This set of competing models consists of the standard Elo, the BT-type model of McHale and Morton (2011, BTM), and the regression-based models of Klaassen and Magnus (2003, KMR) and Del Corral and Prieto-Rodriguez (2010, DCPR). Table 2 summarises the characteristics of such models and the variables used.

The Elo and WElo ratings are set to 1,500 for each player at the beginning of the sample. The initial period from 2005 to 2011 is used as burn-in to predict (one-step-ahead) the winning probabilities for the matches played in the first competing day of 2012 ATP tour. For each of the following days, the information set is updated with the results of the matches played in the previous day. Then,

Table 2: Competing models

Variable	Elo	KMR	DCPR	BTM
Player rank		✓	✓	
Former Top-10			✓	
Height			✓	
Squared Height			✓	
Age			✓	
Squared Age			✓	
Handedness			✓	
Estimator	Bayesian upd. alg.	Logit	Probit	Bayesian upd. alg.

Notes: The table shows the set of variables included in each model. KMR stands for the logit regression of Klaassen and Magnus (2003); DCPR for the probit regression of Del Corral and Prieto-Rodriguez (2010); BTM stands for the BT-type model of McHale and Morton (2011). "upd. alg." stands for updating algorithm.

each model is re-estimated to obtain the winning probabilities for the matches scheduled for the following day. This procedure is repeated until the end of the sample. In other words, we achieve the one-step-ahead forecasts for the matches played at time $t+1$ conditional on the information set up to day t, for the out-of-sample period from 2012 to 2020. For comparison purposes, we evaluate the forecasting performances year-by-year and for the full out-of-sample period.

Table 3 reports the accuracy rates of the WElo approach against the competing models. Specifically, the accuracy rate is computed as the percentage of correctly predicted winners. The results show that the WElo outperforms all the competing models, expect for 2018.

The paired comparison for all the models in Table 2 against the proposed WElo method is done through the Diebold and Mariano (1995, DM) test. The results for the DM test are reported in Table 4. Top and bottom panels respectively illustrate the results under the Brier score and log-loss function. Independently of the scoring function adopted, the null hypothesis of equal predictive ability is largely rejected and, since the test statistic is always negative, we can conclude that overall the proposed WElo rating system significantly outperforms all the competing models. This holds irrespective of the out-of-sample period considered.

Table 3: Accuracy rates (%) across models

	# Matches	WElo	Elo	KMR	DCPR	BTM
2012	2,324	81.583	81.497	78.916	77.969	80.769
2013	2,316	78.843	78.670	76.511	75.648	78.207
2014	2,250	79.467	78.667	75.822	76.533	78.785
2015	2,285	81.961	81.786	79.685	80.210	80.731
2016	2,297	79.094	78.397	77.178	76.045	77.399
2017	2,307	76.843	76.409	73.849	72.940	75.750
2018	2,008	73.406	73.606	73.108	71.016	72.782
2019	2,303	74.109	73.936	69.505	70.200	72.169
2020	1,033	75.581	75.581	73.643	71.512	74.419
2012–2020	19,123	78.234	77.910	75.552	75.065	76.946

Notes: The table reports the percentage of correctly predicted matches by each model in column. The WElo model here considered uses the games in the function $f(\cdot)$. Models' definitions are in Table 2. Shades of grey indicate superior models.

4. Betting strategy

As a further analysis of the out-of-sample forecasting performance, we evaluate the monetary benefits of using the WElo ratings with respect to the standard Elo system through a simple betting framework.

Let $o_{i,j,h}(t)$ and $o_{j,i,h}(t)$ be the odds provided by bookmaker h, with $h = 1,\dots,H$, for the two opponents, players i and j, respectively, of match t. Then, let us define the implied winning probabilities for players i and j as $q_{i,j,h}(t)$ and $q_{j,i,h}(t)$, respectively, obtained as the reciprocal of the odds provided by bookmaker h, that is $q_{i,j,h}(t) = \left(o_{i,j,h}(t)\right)^{-1}$ and $q_{j,i,h}(t) = \left(o_{j,i,h}(t)\right)^{-1}$. Alternatively, one may use the normalized probabilities, with techniques discussed in Štrumbelj (2014) or proposed in Candila and Scognamillo (2018). Moreover, let the best odds among the H bookmakers for player i (and analogously for player j) be defined as:

$$o^{B}_{i,j}(t) = \max_{h=1,\dots,H} o_{i,j,h}(t). \tag{5}$$

As done in other contributions, e.g. McHale and Morton (2011) and Dixon and Coles (1997), we adopt a betting strategy based on a threshold r. In particular, a bet is placed only if this threshold is exceeded. Specifically, the betting strategy adopted here is:

Definition 1. Betting strategy: *The amount of* \$1 *is placed on the best odds*

Table 4: WElo evaluation against competing models by Diebold–Mariano test

		# Matches	Elo	KMR	DCPR	BTM
Brier Score	2012	2,324	-2.116^{**}	-4.341^{***}	-3.403^{***}	-4.062^{***}
	2013	2,316	-3.715^{***}	-2.988^{***}	-2.509^{**}	-3.091^{***}
	2014	2,250	-1.937^{*}	-3.369^{***}	-3.691^{***}	-2.734^{***}
	2015	2,285	-2.051^{**}	-1.377	-1.256	-3.148^{***}
	2016	2,297	-1.683^{*}	-2.059^{**}	-1.387	-3.666^{***}
	2017	2,307	-2.345^{**}	-2.865^{***}	-3.031^{***}	-3.476^{***}
	2018	2,008	-2.347^{**}	-0.365	1.004	-3.803^{***}
	2019	2,303	-2.020^{**}	-2.641^{***}	-2.406^{**}	-5.241^{***}
	2020	1,033	-2.877^{***}	0.243	1.165	-2.309^{**}
	2012–2020	19,123	-6.922^{***}	-6.71^{***}	-5.306^{***}	-10.656^{***}
Log-loss	2012	2,324	-2.371^{**}	-4.436^{***}	-3.428^{***}	-3.735^{***}
	2013	2,316	-3.836^{***}	-2.762^{***}	-2.318^{**}	-3.487^{***}
	2014	2,250	-1.918^{*}	-3.423^{***}	-3.630^{***}	-2.812^{***}
	2015	2,285	-2.198^{**}	-1.423	-1.185	-3.540^{***}
	2016	2,297	-2.400^{**}	-1.807^{*}	-1.140	-3.598^{***}
	2017	2,307	-3.100^{***}	-2.306^{**}	-2.709^{***}	-3.026^{***}
	2018	2,008	-2.269^{**}	0.242	1.310	-4.169^{***}
	2019	2,303	-2.391^{**}	-2.503^{**}	-2.406^{**}	-5.401^{***}
	2020	1,033	-2.964^{***}	0.408	1.256	-2.414^{**}
	2012–2020	19,123	-7.745^{***}	-6.019^{***}	-4.783^{***}	-10.748^{***}

Notes: The table reports the Diebold–Mariano test statistic. Negative values imply that the WElo model outperforms the model in column and vice versa. The WElo model here considered uses the games in the function $f(\cdot)$. Models' definitions are in Table 2. *, ** and *** denote significance at the 10%, 5% and 1% levels, respectively.

$o^{B}_{i,j}(t)$ in (5) for player i for all the matches where it holds that

$$\frac{\hat{P}_{i,j}(t)}{q_{i,j,h}(t)} > r,$$

where $\hat{P}_{i,j}(t) = \left\{\hat{p}_{i,j}(t), \hat{p}^{*}_{i,j}(t)\right\}$ and $q_{i,j,h}(t) > q$.

In general, higher thresholds r imply less betting opportunities (and usually smaller payoffs), but also higher chances of winning. The additional requirement that we make in our betting strategy is that $q_{i,j,h}(t) > q$. This implies that the betting strategy tends to exclude underdogs according to the value of q. In particular, the larger the q, the heavier the underdog. Albeit this condition may appear somewhat arbitrary, the intuition of avoiding longshots is consistent with the well-known favourite-longshot bias found in tennis betting markets and discussed, among others, in Forrest and McHale (2007) and Candila and Scognamillo (2018). Nevertheless, in the empirical analysis below we also report the case in which the longshots are not excluded ($q = 0$).

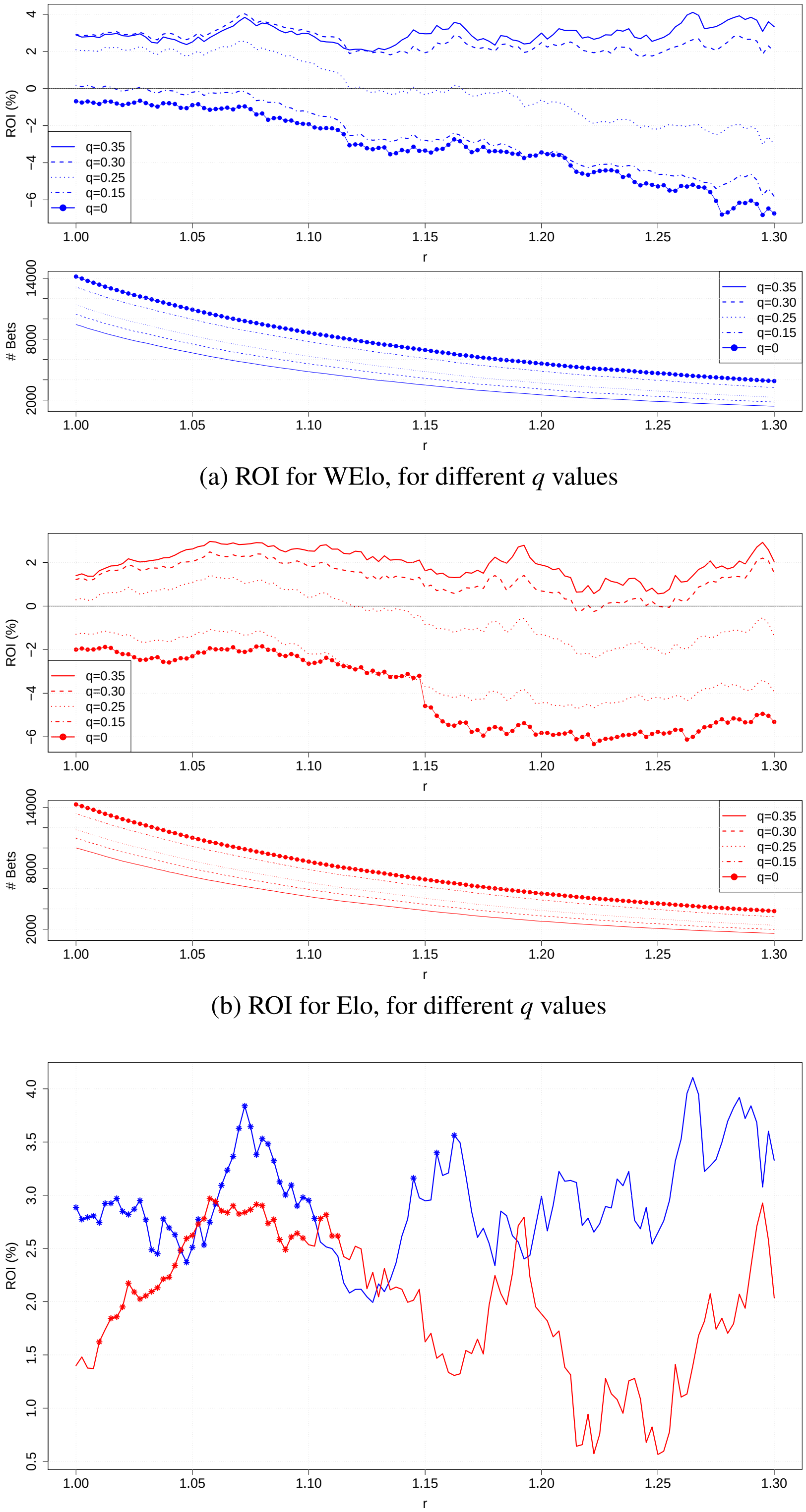
In our analysis we refer to the odds provided by Bet365 bookmaker, i.e. $h = \text{Bet365}$, to define $q_{i,j,h}(t)$.[4] This is because this professional bookmaker has the largest coverage (over 98%) of the matches included in our initial data set. However, the betting strategy in Def. 1 involves wagering on the best odds among the $H = 11$ bookmakers listed in footnote 1. It is indeed common practice among experienced bettors to look for the best possible odds on the market to place their bets. Needless to say, the performance of the betting strategy considering either the average odds or the odds from one single bookmaker is worse than considering the best possible odds on the betting markets (see Table A2 and Figure A1 in Appendix A for comparison). The betting strategies based on the Elo and WElo ratings, i.e. when $\hat{P}_{i,j}(t) = \hat{p}_{i,j}(t)$ and $\hat{P}_{i,j}(t) = \hat{p}^{*}_{i,j}(t)$ in Def. 1, respectively, are compared in terms of return-on-investment (ROI). The ROI defined as the ratio of the net profit (or loss) and the cost of the investment (which, according to Def. 1, in our case coincides with the number of placed bets) in percentage.

The ROI for the WElo and Elo rating systems achieved in our out-of-sample exercise are reported in panels (a) and (b) of Figure 3, respectively with $q \in \{0, 0.15, 0.25, 0.30, 0.35\}$. As expected, as long as the threshold r increases, the number of matches to bet on decreases, as reported in the bottom panels of Figures 3(a) and 3(b). From Figures 3(a) and 3(b), it clearly emerges that the betting strategy based on the WElo rating system dominates the one based on the Elo, both in terms of average ROI and the number of significant values for different r thresholds. This result is even more evident from Figure 3(c), where the ROIs of the two strategies are plotted jointly. In this figure, asterisks denote significant values according to the 90% confidence level computed by i.i.d. bootstrap. As expected and consistently with the favourite-longshot bias found in the literature on tennis betting markets, the ROI tends to decrease (and become strongly negative) for smaller values of q, namely when $q = 0$ and 0.15, i.e. including underdogs and heavy underdogs in the betting strategy, for both WElo and Elo methods. It is interesting to note that the results in Figures 3 show that the ROI is not uniformly increasing with higher values of r.

Table 5 summarises the performance of the two betting strategies for the different years that span our out-of-sample period for the specific case of $q = 0.35$

[4] We also repeated the analysis using the mean of the odds provided by the bookmakers reported in footnote 1 and the results (available upon request) are very similar.

Figure 3: ROI for WElo and Elo value bets using best odds



(a) ROI for WElo, for different q values

(b) ROI for Elo, for different q values

(c) ROI for WElo (blue line) and Elo (red line). * denotes stat. significance computed by bootstrap and $r = 1.1625$.[5] These results show that overall the ROIs and the absolute returns for the WElo-based betting strategy are larger than the ROIs and the absolute returns for the Elo-based strategy (3.563% against 1.307% and \$113.6 against \$46.3, respectively), the former being systematically more profitable than the latter for all the years, except for 2014. Moreover, in Figure A1 of Appendix A we also report a comparison with a random betting strategy replicated for $10,000$ times. The results show that the ROI achieved using the WElo-based strategy is placed towards the right tail of the empirical distribution for all of the cases considered (best, average and Bet365 odds), while the ROI from the Elo-based strategy is very close to the mean and median of the distribution. Finally, Figure 4 compares the ROIs for the WElo and Elo methods with $0.25 \leq q \leq 0.35$ and $1 \leq r \leq 1.3$. Overall, the ROI generated by the WElo-based betting strategy is higher than the ROI generated by the Elo-based strategy for almost all the combinations of q and r, as the blue surface (WElo) appears to envelop the red one (Elo).

In summary, we observe that the forecasting performance of the proposed WElo rating system is overall superior to the standard Elo approach as discussed in Section 3.1. This dominance is further highlighted by the simple betting framework adopted here, where for some combinations of r and q and considering the best odds on the betting market, we also find evidence of profitable opportunities for bettors.[6]

5. Conclusions

Recently, the academic interest in modelling and forecasting sport outcomes has enormously increased. Among a variety of forecasting approaches to predict the probability of winning in tennis matches, the Elo method plays a prominent role. The Elo method recursively estimates the rating of each player based on

[5] The thresholds r and q can be dynamically estimated using the information set available to bettors before the beginning of the match. However, the results are overall robust to such choices (see Figure 4).

[6] It must be acknowledged, however, that, according to the results reported in Table A2 and Figure A1 of Appendix A, adopting the betting strategy outlined in Def. 1 using the average odds as well as the odds from a single professional bookmaker as Bet365, the ROI becomes negative. Therefore, the chance to ‘beat the market’ and achieve positive returns from the betting strategy is mainly related to the identification of the best odds across many different bookmakers. However, this does not undermine the fact that the proposed extension to the Elo rating system overall provides a superior forecasting performance than the standard approach.

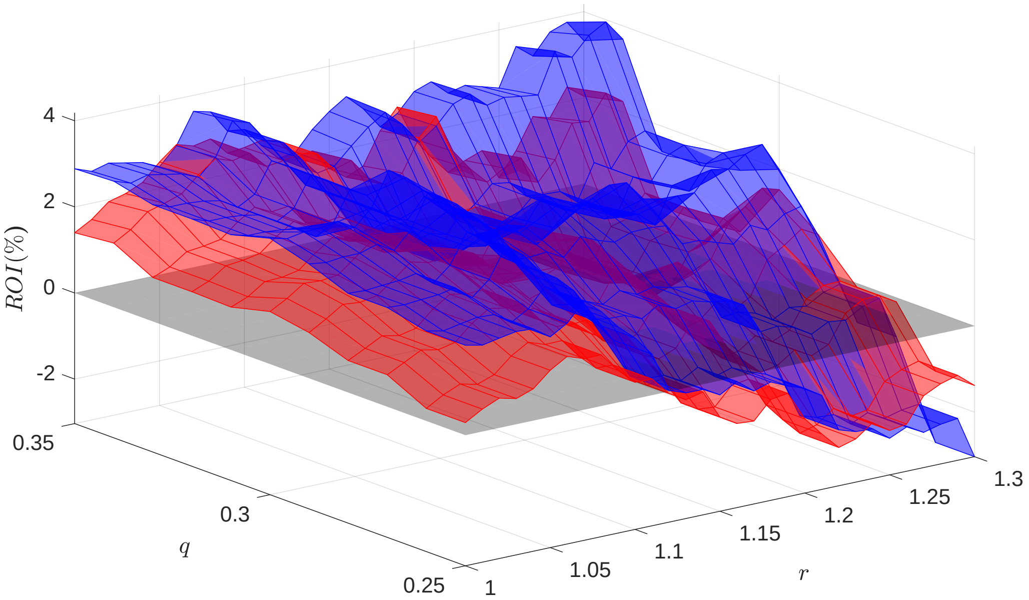
Table 5: ROI of the WElo and Elo models

	# Bets	ROI(%)	Abs. return	# Bets	ROI(%)	Abs. return
		WElo			Elo	
2012	337	8.068	27.189	366	4.825	17.660
2013	367	6.136	22.519	401	2.147	8.609
2014	328	9.131	29.950	387	9.375	36.281
2015	340	4.941	16.799	390	2.433	9.489
2016	322	11.637	37.471	377	6.008	22.650
2017	368	9.201	33.860	419	7.551	31.639
2018	431	−6.329	−27.278	446	−6.578	−29.338
2019	472	2.456	11.592	512	1.795	9.190
2020	224	−17.170	−38.461	244	−24.541	−59.880
2012–2020	3,189	3.563	113.624	3,542	1.307	46.294

Notes: The table reports the ROI(%) and the Absolute return coming from the WElo and Elo estimated probabilities, according to the betting strategy illustrated in Definition 1, using the thresholds $r = 1.1625$ and $q = 0.35$. The WElo model here considered uses the games in the function $f(\cdot)$.

the outcome of the last match played. Afterwards, the probability of winning for the upcoming match is calculated by comparing the Elo ratings for the two players. Despite its popularity and good performance, the main limitation of the Elo method is that it does not take into account how the victory or defeat of a player has been achieved, but only if they win or lose a match. This paper explored the possibility of overcoming this limitation. In particular, we proposed a novel approach to predict tennis match outcomes, generalizing the standard Elo rating system. The proposed weighted version of the Elo (WElo) is able to take into account the information contained in the number of games won by each player in the last match, and not only the match outcome. By construction, the standard Elo ratings are the upper bounds for the WElo and, for a given match, they coincide only in the case of a perfect score, i.e. if one player defeats the opponent by winning all the games in the match. Therefore, the difference between the Elo and WElo ratings represents a proxy of the state of form of each player. From a forecasting viewpoint, this information significantly enriches the information set on which the probability prediction for the upcoming match is based. From a statistical viewpoint, we demonstrate that the proposed WElo rating system is symmetric, even though it takes into account the different number of games won by each player. Moreover, we derive a bootstrap-based procedure for the compu-

Figure 4: ROI(%) for the WElo (games-based version) and Elo rates



Notes: The figure reports the ROI(%) for the WElo (blue surface) and Elo (red surface), according to according to different *r* and *q* values, for the out-of-sample period (2012–2020).

tation of the confidence intervals around the WElo ratings. This bootstrap-based procedure can also be applied to the standard Elo method. In the empirical analysis, we investigate the forecasting performance of the proposed WElo method over a large dataset of over 60,000 men's and women's tennis matches. We find that the WElo statistically outperforms some popular models for forecasting the winning probability in tennis, irrespective of the out-of-sample period considered. Finally, the WElo method provides profitable and statistically meaningful betting opportunities.

Future research may concern some interesting extensions as the inclusion of additional variables, as done by Vaughan Williams et al. (2020) in the context of Elo rating for tennis matches, and the application of the WElo method to other sports.

References

Angelini, G. and L. De Angelis (2017). PARX model for football match predictions. *Journal of Forecasting 36*(7), 795–807.

Baker, R. D. and I. G. McHale (2014). A dynamic paired comparisons model:

Who is the greatest tennis player? *European Journal of Operational Research 236*(2), 677–684.

Baker, R. D. and I. G. McHale (2017). An empirical bayes model for time-varying paired comparisons ratings: Who is the greatest women's tennis player? *European Journal of Operational Research 258*(1), 328–333.

Barnett, T., A. Brown, and S. Clarke (2006). Developing a model that reflects outcomes of tennis matches. In *Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport, Coolangatta, Queensland*, pp. 178–188.

Barnett, T. and S. R. Clarke (2005). Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics 16*(2), 113–120.

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review 78*(1), 1–3.

Candila, V. and L. Palazzo (2020). Neural networks and betting strategies for tennis. *Risks 8*(3), 1–19.

Candila, V. and A. Scognamillo (2018). Estimating the implied probabilities in the tennis betting market: A new normalization procedure. *International Journal of Sport Finance 13*, 225–242.

Carbone, J., T. Corke, and F. Moisiadis (2016). The rugby league prediction model: Using an Elo-based approach to predict the outcome of National Rugby League (NRL) matches. *International Educational Scientific Research Journal 2*(5), 26–30.

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

Del Corral, J. and J. Prieto-Rodriguez (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting 26*(3), 551–563.

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–263.

Dixon, M. J. and S. G. Coles (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 46*(2), 265–280.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Forrest, D. and I. McHale (2007). Anyone for tennis (betting)? *The European Journal of Finance 13*(8), 751–768.

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

Gorgi, P., S. J. Koopman, and R. Lit (2019). The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 182*(4), 1393–1409.

Hvattum, L. M. and H. Arntzen (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting 26*(3), 460–470.

Klaassen, F. J. and J. R. Magnus (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research 148*(2), 257–267.

Knottenbelt, W. J., D. Spanias, and A. M. Madurska (2012). A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications 64*(12), 3820–3827.

Koopman, S. J. and R. Lit (2015). A dynamic bivariate poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*(1), 167–186.

Kovalchik, S. (2020). Extension of the elo rating system to margin of victory. *International Journal of Forecasting 36*, 1329–1341.

Kovalchik, S. and M. Reid (2019). A calibration method with dynamic updates for within-match forecasting of wins in tennis. *International Journal of Forecasting 35*(2), 756–766.

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports 12*(3), 127–138.

Leitner, C., A. Zeileis, and K. Hornik (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. *International Journal of Forecasting 26*(3), 471–481.

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

Ryall, R. and A. Bedford (2010). An optimized ratings-based model for forecasting Australian Rules football. *International Journal of Forecasting 26*(3), 511–517.

Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International Journal of Forecasting 30*(4), 934–943.

Vaughan Williams, L., C. Liu, L. Dixon, and H. Gerrard (2020). How well do Elo-based ratings predict professional tennis matches? *Journal of Quantitative Analysis in Sports*.

Appendix A. ATP matches

Appendix A.1. Different weighting function $f(\cdot)$

As a robustness check and to evaluate the performance of a different weighting function $f(\cdot)$, we exploit the flexibility of the WElo appraoch in (3) by specifying $f(\cdot)$ with the number of sets, instead of the number of games as considered in the paper. Formally, we define $f\left(S_{i,t}(t)\right)$ as:

$$f(S_{i,j}(t)) = \begin{cases} \frac{NS_i(t)}{NS_i(t)+NS_j(t)} & \text{if player } i \text{ has won match } t; \\ \frac{NS_j(t)}{NS_i(t)+NS_j(t)} & \text{if player } i \text{ has lost match } t, \end{cases}$$

where $NS_i(t)$ and $NS_j(t)$ represent the number of sets won by player i and player j in match t, respectively.

Table A1: WElo (sets-based version) evaluation against competing models by Diebold–Mariano test

		# Matches	Elo	KMR	DCPR	BTM
Brier Score	2012	2,324	-1.340	-3.704^{***}	-2.791^{***}	-3.069^{***}
	2013	2,316	-1.585	-1.893^{*}	-1.432	-1.819^{*}
	2014	2,250	-0.451	-2.514^{**}	-2.782^{***}	-1.790^{*}
	2015	2,285	-2.356^{**}	-1.127	-0.989	-2.545^{**}
	2016	2,297	-1.762^{*}	-1.764^{*}	-1.102	-3.030^{***}
	2017	2,307	-1.744^{*}	-2.312^{**}	-2.478^{**}	-2.558^{**}
	2018	2,008	-0.834	0.201	1.528	-2.722^{***}
	2019	2,303	-0.773	-2.043^{**}	-1.783^{*}	-3.910^{***}
	2020	1,033	-1.450	0.903	1.761^{*}	-1.179
	2012–2020	19,123	-4.009^{***}	-4.933^{***}	-3.539^{***}	-7.707^{***}
Log-loss	2012	2,324	-1.567	-3.590^{***}	-2.651^{***}	-3.122^{***}
	2013	2,316	-1.672^{*}	-1.555	-1.129	-2.220^{**}
	2014	2,250	-0.291	-2.444^{**}	-2.610^{***}	-1.855^{*}
	2015	2,285	-2.506^{**}	-1.110	-0.868	-2.965^{***}
	2016	2,297	-1.854^{*}	-1.248	-0.598	-2.861^{***}
	2017	2,307	-1.718^{*}	-1.434	-1.827^{*}	-2.215^{**}
	2018	2,008	-0.208	0.877	1.912^{*}	-3.287^{***}
	2019	2,303	-0.634	-1.706^{*}	-1.588	-4.115^{***}
	2020	1,033	-1.321	1.126	1.901^{*}	-1.295
	2012–2020	19,123	-3.805^{***}	-3.795^{***}	-2.588^{***}	-8.108^{***}

Notes: The table reports the Diebold–Mariano test statistic. Negative values imply that the Weighted Elo outperforms the model in column and vice versa. The WElo model here considered uses the sets in the function $f(\cdot)$. Models' definitions are in Table 2. *, ** and *** denote significance at the 10%, 5% and 1% levels, respectively.

In Table A1 we show the out-of-sample performance of the WElo method in comparison with the competing models detailed in Section 3. Despite such performance is not always found significant on a yearly basis, the results from the DM tests reported in Table A1 show that, for the whole sample (2012-2020), the proposed WElo method significantly outperforms all the competing models. Therefore, overall by comparing the results in Table A1 with those in Table 4, we find evidence of a larger predictive content by using games than using sets in the weighting function $f(\cdot)$ in (3).

Appendix A.2. Different betting strategy

Here we consider two alternative betting strategies. In particular, we replace the best odds $o^B_{i,j}(t)$ in (5) considered in Def. 1 with (i) the average odds across the $H = 11$ bookmakers reported in Footnote 1, i.e. $o^{avg}_{i,j}(t) = \frac{1}{H}\sum_{h=1}^{H} o_{i,j,h}(t)$, and (ii) the odds from Bet365 bookmaker.

Table A2: ROI of WElo and Elo methods for bets placed on average and Bet365 odds

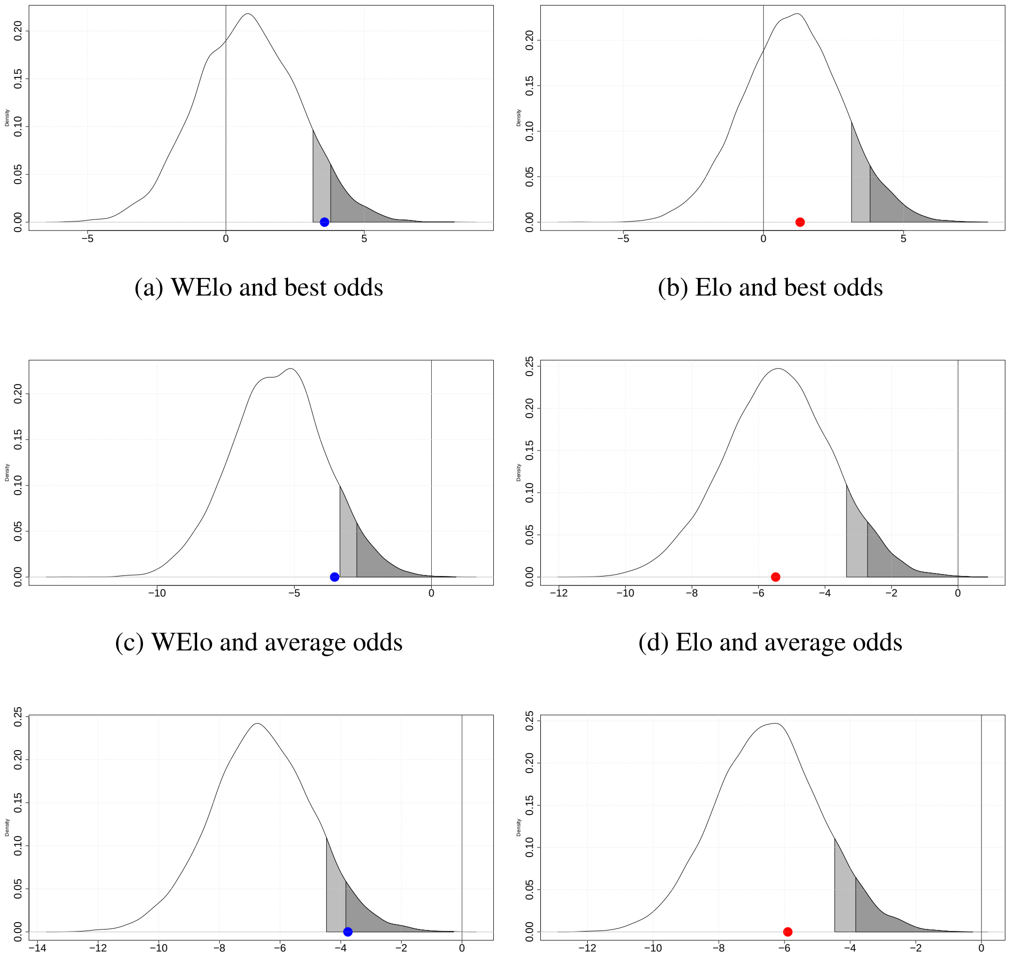
	# Bets	ROI(%)	Abs. return	# Bets	ROI(%)	Abs. return
	Panel A: Bets placed using the average odds					
		WElo			Elo	
2012	337	−0.769	−2.592	366	−3.388	−12.400
2013	367	−2.542	−9.329	401	−6.027	−24.168
2014	328	1.195	3.920	387	1.693	6.552
2015	340	−2.709	−9.211	390	−4.710	−18.369
2016	322	4.705	15.150	377	−0.366	−1.380
2017	368	2.554	9.399	419	1.191	4.990
2018	431	−11.870	−51.160	446	−12.112	−54.020
2019	472	−3.898	−18.399	512	−4.668	−23.900
2020	224	−22.473	−50.340	244	−29.316	−71.531
2012–2020	3,189	−3.530	−112.572	3,542	−5.484	−194.243
	Panel B: Bets placed using the Bet365 odds					
		WElo			Elo	
2012	337	0.223	0.752	366	−2.959	−10.830
2013	367	−2.193	−8.048	401	−5.903	−23.671
2014	328	1.628	5.340	387	1.664	6.440
2015	340	−2.585	−8.789	390	−4.726	−18.431
2016	322	4.236	13.640	377	−0.841	−3.171
2017	368	2.022	7.441	419	0.568	2.380
2018	431	−12.805	−55.190	446	−13.206	−58.899
2019	472	−4.975	−23.482	512	−5.846	−29.932
2020	224	−23.067	−51.670	244	−29.836	−72.800
2012–2020	3,189	−3.763	−120.002	3,542	−5.898	−208.907

Notes: The table reports the ROI(%) and the Absolute return coming from the WElo and Elo estimated probabilities, according to the betting strategy illustrated in Definition 1, using the thresholds $r = 1.1625$ and $q = 0.35$. The WElo method considers the number of games, i.e. $f(G_{i,j}(t))$.

As shown in Table A2 the WElo approach still systematically outperforms the standard Elo in terms of ROI and absolute returns. Unsurprisingly, the returns decrease and become negative when considering the average or Bet365 odds instead of the best possible odds on the betting market.

We further complement the evidence from the betting strategy by constructing the empirical distributions of ROIs obtained from a random betting strategy based on 10,000 replications of random selections of the players on which we bet. In Figure A1 we depict the empirical distributions from the random betting strategy and the ROI obtained from the betting strategy defined in Definition 1, as well as the ROIs from the two strategies outlined in this Section. The results show that the ROI achieved using the WElo-based strategy is placed towards the right tail of the empirical distribution for all of the cases considered (best, average and Bet365 odds). Remarkably, the ROI from the WElo-based betting strategy is found significantly different from the mean of the empirical distribution at a 10% level for the best odds and 5% level when considering Bet365 odds (see panels (a) and (e) in Figure A1). Conversely, the ROIs from the Elo-based strategy are very close to the mean and median of the empirical distributions of the random betting strategies.

Figure A1: WElo and Elo ROI(%) upon empirical distribution of the random betting strategy



(a) WElo and best odds

(b) Elo and best odds

(c) WElo and average odds

(d) Elo and average odds

(e) WElo and Bet365 odds

(f) Elo and Bet365 odds

Notes: The figures depict the empirical distributions of the random betting strategies using the best, average and Bet365 odds. Blue and red points denote the ROI achieved by the WElo and Elo-based betting strategies, respectively. Light and dark shades of grey denote the right tail of the empirical distribution at the 10% and 5% significance level, respectively.

Appendix B. WTA matches

Here we repeat the analyses reported in the paper by considering the data set on women's (WTA) professional matches. We do so to check the robustness of all the results present in the paper.

Table B1 reports the DM test under the Brier score and log-loss function. As in the leading case presented in the paper, the results confirm that the WElo rating system significantly outperforms all the other competing methods.

Table B1: WElo evaluation against competing models by Diebold–Mariano test. WTA matches

		# Matches	Elo	KMR	DCPR	BTM
Brier Score	2013	2,231	-2.413^{**}	-1.769^{*}	-1.984^{**}	-6.327^{***}
	2014	2,023	-2.374^{**}	-1.340	-1.256	-4.096^{***}
	2015	2,129	-2.661^{***}	-0.973	-0.655	-5.611^{***}
	2016	2,145	-3.241^{***}	-1.253	-1.307	-4.118^{***}
	2017	2,077	-1.239	-4.001^{***}	-3.492^{***}	-5.663^{***}
	2018	2,148	-2.050^{**}	-1.940^{*}	-1.451	-7.116^{***}
	2019	2,095	-0.252	-1.456	-0.853	-6.676^{***}
	2020	836	-1.229	0.039	0.661	-4.416^{***}
	2013–2020	15,684	-5.476^{***}	-4.694^{***}	-3.812^{***}	-15.671^{***}
Log-loss	2013	2,231	-2.514^{**}	-1.882^{*}	-2.081^{**}	-5.575^{***}
	2014	2,023	-2.766^{***}	-1.128	-1.093	-4.593^{***}
	2015	2,129	-3.113^{***}	-1.095	-0.880	-5.884^{***}
	2016	2,145	-3.382^{***}	-1.232	-1.258	-5.144^{***}
	2017	2,077	-1.913^{*}	-3.718^{***}	-3.075^{***}	-5.678^{***}
	2018	2,148	-2.182^{**}	-1.794^{*}	-1.330	-5.982^{***}
	2019	2,095	-0.805	-1.814^{*}	-1.096	-6.622^{***}
	2020	836	-1.290	0.467	0.994	-2.908^{***}
	2013–2020	15,684	-6.388^{***}	-4.554^{***}	-3.649^{***}	-13.873^{***}

Notes: The table reports the Diebold–Mariano test statistic. Negative values imply that the WElo outperforms the model in column and vice versa. The WElo model here considered uses the games in the function $f(\cdot)$. Models' definitions are in Table 2. *, ** and *** denote significance at the 10%, 5% and 1% levels, respectively.

Figure B1 shows the results achieved in our out-of-sample exercise. Panels (a) and (b) of Figure B1 depict the ROI for the WElo-based and Elo-based betting strategies, respectively, with $q \in \{0, 0.15, 0.25, 0.30, 0.35\}$. Panel (c) compares the out-of-sample performances of the WElo and the Elo for the specific case with $q = 0.35$. From these figures, as in the case of ATP matches reported in the paper, it is evident that the WElo-based betting strategy dominates the one based on Elo, both in terms of average ROI and the number of significant values for different r thresholds.

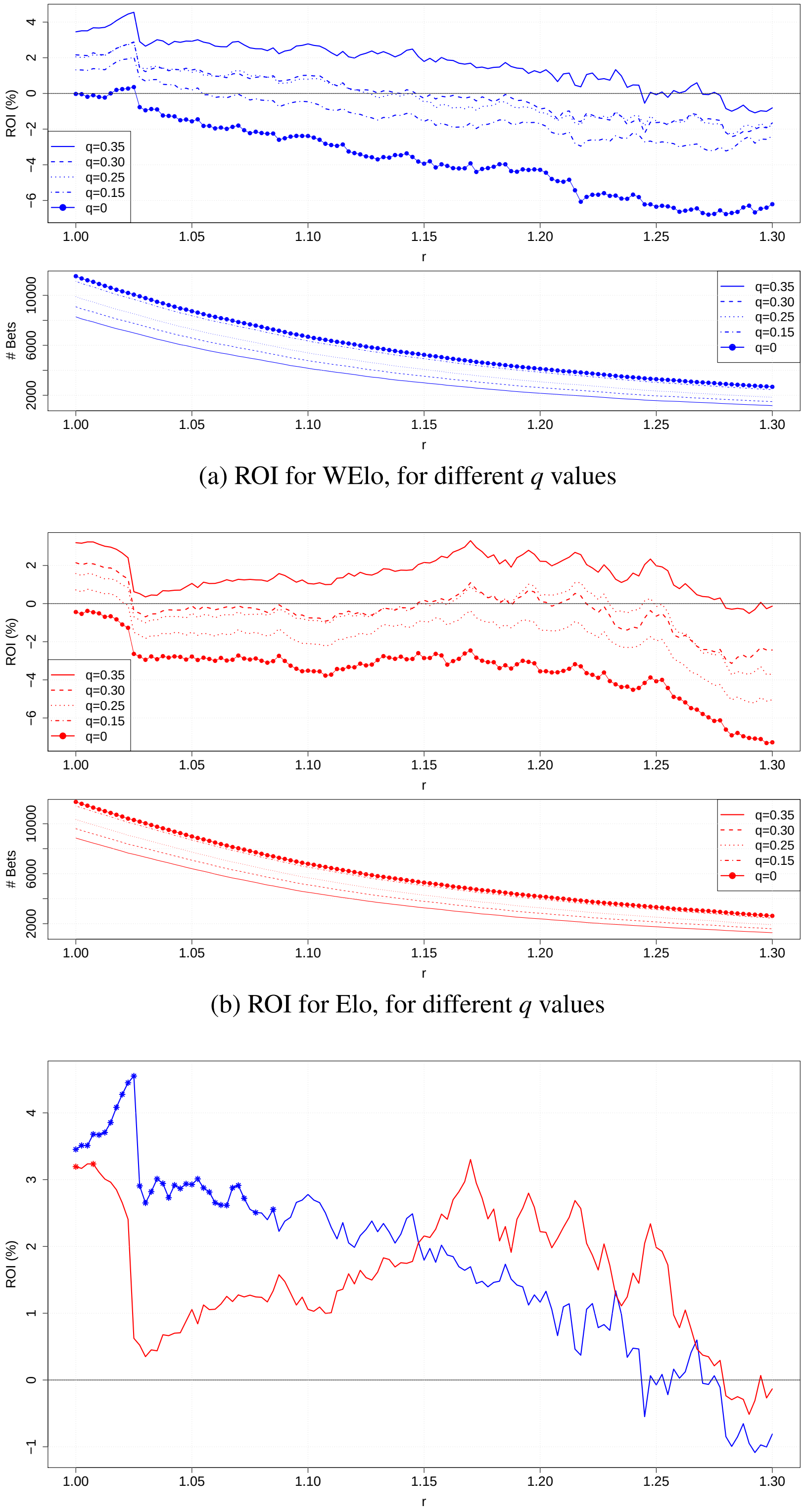
Finally, Table B2 summarises the performance of two rating systems for the specific case with $q = 0.35$ and $r = 1.05$. The results in Table B2 display an overall superior performance of the WElo rating system compared to the Elo in terms of ROI and absolute return also for women's (WTA) professional matches.

Table B2: ROI of the WElo (based on games) and Elo models. WTA matches

	# Bets	ROI(%)	Abs. return	# Bets	ROI(%)	Abs. return
		WElo			Elo	
2013	714	4.499	32.123	794	3.209	25.479
2014	661	−1.254	−8.289	724	−3.790	−27.440
2015	777	8.925	69.347	847	4.815	40.783
2016	726	1.567	11.376	794	−4.563	−36.230
2017	790	−1.978	−15.626	862	−1.036	−8.930
2018	900	5.980	53.820	995	5.551	55.232
2019	927	0.661	6.127	984	1.191	11.719
2020	369	6.173	22.778	404	1.743	7.042
2013–2020	5,864	2.927	171.639	6,404	1.056	67.626

Notes: The table reports the ROI(%) and the absolute return from the WElo and Elo estimated probabilities, according to the betting strategy in Definition 1, using the thresholds $r = 1.05$ and $q = 0.35$. The WElo method considered here uses the games in the function $f(\cdot)$.

Figure B1: ROI for WElo and Elo value bets using best odds. WTA matches



(a) ROI for WElo, for different q values

(b) ROI for Elo, for different q values

(c) ROI for WElo (blue line) and Elo (red line). * denotes stat. significance