International Journal of Forecasting 36 (2020) 1329–1341

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Extension of the Elo rating system to margin of victory

Stephanie Kovalchik

Zelus Analytics, Austin, TX, USA

ARTICLE INFO

Keywords:
Paired comparison
Ranking systems
Sports forecasting
State space model
Time series

ABSTRACT

The Elo rating system is one of the most popular methods for estimating the ability of competitors over time in sport. The standard Elo system focuses on predicting wins and losses, but there is often also interest in the margin of victory (MOV) because it reflects the magnitude of a result. There have been few theoretical investigations and comparisons of Elo-based models. In the present study, we propose four model options for an MOV Elo system: linear, joint additive, multiplicative, and logistic. Notations and guidance for tuning each model are provided. The models were applied to men's tennis for several MOV choices. The results showed that all MOV approaches using within-set statistics improved the predictive performance compared with the standard Elo system, but only the joint additive model yielded unbiased ratings with stable variance in the simulation study. This general framework for MOV Elo ratings provide sports modelers with a new set of tools for building systems to rate competitors and forecast outcomes in sport.

1. Introduction

The Elo rating system is one of the most popular methods for estimating the abilities of competitors over time in sport (Langville & Meyer, 2012). One of the most appealing features of the Elo system is its dynamic updating rule, where ratings are adapted after each competitive result based on the residual of the predicted and observed outcomes. This self-correcting characteristic of the algorithm allows it to naturally follow the ups and downs exhibited by a competitor's ability over long periods of time (Glickman, 1995). This feature is particularly suitable for the prediction of game outcomes and its forecasts have been shown to perform well for a range of sports (Barrow, Drayer, Elliott, Gaut, & Osting, 2013).

One sport where the Elo system provides significantly improved predictions compared with alternative methods is tennis (Kovalchik, 2016). The most common approaches for predicting the outcomes of tennis matches belong to the following categories: mathematical models (Newton & Keller, 2005), regression models using player rankings and other predictors (Boulier & Stekler, 1999), or paired comparison models (McHale & Morton, 2011), where the Elo system is a special case. Among these modeling approaches, the Elo system has been shown to obtain the closest performance with respect to bookmakers (Kovalchik, 2016).

The Elo system is one of the most accurate methods for predicting tennis match results, but it still has a modest average prediction accuracy at the tour level of approximately 70%. Picking the winner incorrectly in three out of every 10 matches limits the practical application of this system. One possible strategy for improving the prediction performance of the rating system involves incorporating more information about a player's match performance.

Updates of the standard Elo system are based on the binary outcome comprising the win or loss of a match. However, this outcome makes no distinction between a player who wins in straight sets without dropping a game or a player who just edges out their opponent in a final set tiebreak. These two scenarios contrast extreme differences in the *magnitude* of a performance in tennis. It is an open question whether information about the magnitude

E-mail address: skovalchik@zelusanalytics.com.

https://doi.org/10.1016/j.ijforecast.2020.01.006
0169-2070/

of a match performance might have predictive value in addition to the win result.

Previous studies have incorporated win dominance into rating systems by using the margin of victory (MOV). The MOV is the difference in performance between competitors in terms of some scoring variable that is strongly related to wins, such as the difference in runs scored by each team in a baseball game. An MOV close to zero indicates a close contest, whereas a large MOV reflects an easy win.

Research into the MOV modeling has been conducted for decades (Harville, 1977), but examples of Elo MOV models have only appeared in recent years. One approach shifts the focus of the modeled outcome from the win result to the margin by substituting the estimation step for a linear model of the margin (Carbone, Corke, & Moisiadis, 2016; Constantinou & Fenton, 2013). Other studies have focused on the predictions of the win outcome by models and used the MOV to modulate the system's learning rate, with more frequent updates when more extreme margins occur (Mangan & Collins, 2016). The FIFA rankings of women's soccer are a rare example of the use of an MOV Elo rating system for official rankings.[1] This system employs a plug-in approach where the binary components in the standard Elo are replaced with an "actual match percentage" derived from the MOV in terms of goal (Stefani, 2011).

Little published guidance is available regarding the preferred MOV strategy for modelers who want to implement an MOV Elo rating system. In addition, there have been no theoretical discussions of the range of modeling decisions available in an MOV Elo rating system, including fundamental decisions about whether to modify the estimation rule, update the rule, or both using information related to the MOV. Comparisons of the performance of different MOV rating strategies have also not been reported. Choosing an MOV Elo system for tennis is particularly challenging because the hierarchical scoring system used in this sport means that the best choice of scoring variable is unclear and another layer of decision making is added to the model development process.

In the present study, we developed a framework for MOV Elo rating systems. This framework includes the general structure of four distinct methods for extending the standard Elo model to include MOV information. We provide descriptions of the properties based on simulations and guidance about how to tune each type of model. The models were applied to the outcomes of men's professional tennis matches as a real-world example in order to compare the performance of various MOV rating methods with several choices of MOV variables.

2. Methods

2.1. Margins

Unlike many other sports, the scoring system is hierarchical in tennis. Players win points within games and win games within sets. The number of sets won in a match is the margin that ultimately determines the winner of the match. There is no such guarantee for other match statistics. For example, a player who loses a match 6-1, 5-7, 5-7 has won one fewer set but won one more game overall. The same applies to any other score statistic summed across sets, which is referred to as "Simpson's paradox of tennis" (Wright, Rodenberg, & Sackmann, 2013). This situation is problematic for the design of a tennis MOV rating system because for many possible choices of margin variables, a subset of players with a positive margin will have lost their matches.

The nature of tennis scoring requires greater consideration of the specific properties of an MOV variable. The main goal of a rating system is to provide accurate estimates of the relative abilities of competitors over time.[2] Thus, the first property required by a margin is a strong correlation with the magnitude of a win. We also prefer more informative (high variance) margins because they are expected to help differentiate differences in the ability of competitors. Finally, we prefer score variables with greater coverage in public tennis data sets in order to facilitate the tracking of ability over time.

Among the match statistics available for multiple years of matches played by professional tennis players, five were selected for consideration in the present study, i.e., four count variables comprising sets won, games won, break points won, and total points won, and one percentage variable comprising serve percentage won. This set of variables covers a range of aggregation from the point to set level, as well as skills comprising the serve percentage to focus on the serving ability and break points to focus on the returning ability.

3. Models

3.1. Standard Elo

The Elo rating system is a dynamic algorithm for paired comparisons. There are numerous implementations of the system but the version described in the following is that originally introduced by Elo and it is the most commonly used version (Elo, 1978). We refer to this version as *standard Elo*.

The standard Elo model is a two-step algorithm involving: (1) an estimation step (E-step) and (2) an update step (U-step). The E-step sets the prediction for the outcome of the tth competition given the ratings of competitors i and j at time t:

$$\hat{P}_{ijt} = 1/(1 + 10^{-(R_{it} - R_{jt})/400}). \quad (1)$$

The probability $\hat{P}_{ijt}$ is the point estimate of the probability that the ith competitor will defeat competitor j given competitor ratings R_{it} and R_{jt}. The form of the estimation follows a logistic curve and the denominator of 400 corresponds to two times the standard deviation of the ratings of competitors in Elo's original application to chess.

[1] https://www.fifa.com/fifa-world-ranking/procedure/women.

[2] Rating systems have many possible goals but this is probably the most common (Stern, 2004).

After the result of the competition is known, the rating of each competitor is updated using the following rule for the ith competitor:

$$R_{i(t+1)} = R_{it} + K(W_{ijt} - \hat{P}_{ijt}), \tag{2}$$

In the U-step, W_{ijt} is a zero–one indicator of the actual result and K is a constant learning rate that sets a bound on the maximum gain (or reduction) in a rating after any single result. Elo adapted K to the size and format of tournaments, but he frequently used a value of 32 for K.

3.2. Theoretical foundations

It is important to acknowledge that the Elo system was developed mainly based on statistical heuristics rather than formal theory. However, subsequent studies have shown that these heuristics have direct connections with fully model-based rating systems. The fundamental premise shared by all paired comparison models is an assumption that the expected outcomes are functions of the relative abilities of opponents. In fact, for the binary outcome of a win, the form of the prediction for all paired comparison models is the ratio of the latent ability of competitors:

$$P(W_{ijt} = 1) = \frac{\lambda_{it}}{\lambda_{it} + \lambda_{jt}}. \tag{3}$$

The popular Bradley–Terry model assumes that the abilities are fixed in time, $\lambda_{it} = \lambda_i$, and estimates are generally obtained for each λ_i using likelihood-based techniques according to a binomial likelihood (Cattelan, 2012).

The Elo system assumes that the relationships between abilities and predicted wins are the same but it allows player abilities to change over time. Thus, the Elo system is most closely related to dynamic paired comparison models such as the Glicko system, which is a fully Bayesian state space model for player ratings (Glickman, 2001). In the Glicko system, a competitor's rating at the start of each time period is modeled as a normal distribution, $R_{it} \sim N(\mu, \sigma^2)$. After a single result against opponent j of ability R_{jt} with prior distribution $R_{jt} \sim N(\mu_j, \sigma_j^2)$, the posterior mean for μ is approximately equal to:

$$\mu' = \mu + k(\mu, \sigma^2, \mu_j, \sigma_j) g(\sigma_j^2)(W_{ijt} - \frac{1}{1 + 10^{-g(\sigma_j^2)(\theta - \mu_j)/400}}), \tag{4}$$

where $k(.)$ is a function that is increasing in σ^2, which is the prior uncertainty in rating R_{it}, and $g(\sigma_j^2)$ is decreasing in σ_j^2, which is the prior uncertainty in the opponent's rating. When $\sigma_j^2 = 0$, i.e., when the opponent's ability is assumed to be fixed, then $g(\sigma_j^2) = 1$, and the posterior update in Eq. (4) is identical to the update in the standard Elo system if $K = k(\mu, \sigma^2, \mu_j, 0)$ (Glickman, 1999).

Fahrmeir and Tutz (1994) arrived at the same form for updating the abilities of player in their state space model of dynamic ratings where abilities are regarded as a first-order random walk. When the hyperparameters of the ability parameters were treated as fixed, they derived a generalized Kalman filtering algorithm that sets the updates of abilities as equal to the posterior mode, which is equal in form to the standard Elo update rule. In both the Glicko and Fahrmeir and Tutz (1994) dynamic paired comparison systems, additional smoothing is used to correct all estimates from time zero to time t by using backward smoothing with a standard Kalman filter. If we let $\mu_{s|t}$ denote the player ratings at time s given all information up to time t, then for $s = t - 1, \ldots, 0$, each rating is smoothed as follows:

$$\mu_{s-1|t} = \mu_{s-1} + K_{s-1}(\mu_{s|t} - \mu_{s|s-1}) \tag{5}$$

and K_{s-1} is a function of the information matrix of $\mu_{s|s-1}$ given all information up to $s - 1$ (Fahrmeir & Tutz, 1994).

These connections provide some reassurance that the Elo algorithm may have desirable properties but they do not provide any guarantees regarding its properties. Surprisingly, few studies have addressed the fundamental questions regarding the validity of the system for different interactions among competitors (Aldous et al., 2017). In one of the few studies of the convergence properties of the standard Elo system, Jabin and Junca (2015) proved that for a pool of continuously mixing competitors where there is a positive probability that any two competitors could meet at any time in a continuous kinetic "all meet all" model, the player Elo ratings will converge to their true strengths at an exponential rate (Theorem 2.2, Jabin and Junca (2015)). The continuous kinetic model provides a good description of the interactions in tennis because the pool of competitors is large. Thus, it is probable that competitors with large differences in ratings will meet and the main determinant of the outcomes is the relative difference in the player abilities, and the outcomes are conditionally independent. Importantly, in the same continuous kinetic model of player interactions, the convergence of the player ratings $(R_{it}, R_{jt}) \to (R_i, R_j)$ will be satisfied as $t \to \infty$ for any update rule,

$$\Upsilon(R_{it} - R_{jt}) = Y_{ijt} - \upsilon(R_{it} - R_{jt}), \tag{6}$$

if $E[Y_{ijt}] = \upsilon(R_i - R_j)$ for the score outcome Y_{ijt}, and the function $\upsilon(.)$ is Lipschitz continuous and a strictly increasing function of the difference in ratings. We refer to these conditions for the update rule as the (1) stationarity and (2) Lipschitz conditions, respectively. In order to justify the choice of the update rule for each of the extensions in the following, we evaluate the model's validity under the requirements of the continuous kinetic model.

3.3. MOV models

We proposes four distinct extensions of the standard Elo system to incorporate MOV: linear, joint additive, multiplicative, and logistic models. In the following, we describe each model in detail.

3.3.1. Linear

The linear model is most suitable for applications where the primary interest is predictions about the margin and where (as the model's name suggests) it is reasonable to model the margin as a linear function of the difference in the latent abilities of competitors. Let M_{ijt} be the margin of the ith competitor against the jth competitor in the tth match. The E-step assumes that the expected

margin is proportional to the difference in the ratings of the competitors:

$$\hat{M}_{ijt} = \frac{R_{it} - R_{jt}}{\sigma}, \tag{7}$$

where σ is an unknown tuning parameter. Given the expected margin, the U-step takes the same form as the standard Elo model but with respect to the residual of the MOV:

$$R_{i(t+1)} = R_{it} + K(M_{ijt} - \hat{M}_{ijt}), \tag{8}$$

thereby resulting in two parameters that need to be optimized: σ and K.

In the case where the ratings are stationary at R_i and R_j, the expectation for the margin outcomes is $E[M_{ijt}] = \frac{R_i - R_j}{\sigma} = \hat{M}_{ijt}$. Thus, if the model for the margin expectation is correct, the linear model ensures that player ratings that have reached their true values will have an expected change of zero. Convergence to the true values under the dynamic kinetic model of Jabin and Junca (2015) is guaranteed by the update function, $\upsilon(R_i - R_j) = \frac{R_i - R_j}{\sigma}$, being continuously differentiable and bounded by the positive number σ^{-1}.

3.3.2. Joint additive

The joint additive model is particularly useful for applications where there is an interest in predicting both the MOV and the win result, or where the combined information regarding these outcomes explains more about the differences in the abilities of competitors than either outcome alone. Mathematically, the joint additive model can be considered a combination of the standard Elo system and the linear MOV model. This model involves two calculations in the E-step, where we use the standard estimation calculation for the outcome of the match (Eq. (1)) and the linear calculation for the margin (Eq. (7)):

$$\hat{M}_{ijt} = \frac{R_{it} - R_{jt}}{\sigma_1}, \quad \text{and } \hat{P}_{ijt} = 1/(1 + 10^{-(R_{it} - R_{jt})/\sigma_2}). \tag{9}$$

Deviations from the observed win and the observed margin are both considered in the updating of the player rating using the following additive U-step:

$$R_{i(t+1)} = R_{it} + K_1(M_{ijt} - \hat{M}_{ijt}) + K_2(W_{ijt} - \hat{P}_{ijt}). \tag{10}$$

It should be noted that in Eq. (10), each residual is multiplied by a distinct learning rate, which allows flexibility regarding how much effect the deviations in the expected win and expected margin will have on the rating update.

Both the MOV and win E-steps have scaling factors comprising σ_1 and σ_2, respectively, so they can also be treated as unknowns in the joint additive system. Given the equivalence of the E-steps to the linear model in the case of the MOV and the standard Elo model in the case of the win expectation, the scaling factors are set to the optimal values in these systems. The parameters that need to be optimized are K_1 and K_2.

The update rule for the joint additive model is a linear combination of the rules for the linear and standard Elo model, so the validity of the model will be satisfied by the same conditions as the linear model. As shown above, the key condition for the convergence of the linear model is the correct specification of the expectation for the MOV.

3.3.3. Multiplicative

All of the examples given above use an update rule where the residual of the MOV estimate is applied additively on the player rating. In the multiplicative MOV model, the residual is used to modulate the learning rate, which has a nonlinear relationship with the changes in player ratings. Thus, the multiplicative model is most suitable for applications where the primary aim is predicting the win result but where the magnitude of the MOV residual modulates confidence in the win result.

The E-step for the model is the same as that in the standard Elo model, $\hat{P}_{ijt} = 1/(1 + 10^{-(R_{it} - R_{jt})/\sigma_2})$, but the form of the U-step changes to:

$$R_{i(t+1)} = R_{it} + K(1 + |M_{ijt}/\sigma_1|)^{\alpha}(W_{ijt} - \hat{P}_{ijt}), \tag{11}$$

where $\alpha > 0$ and $\sigma_1 > 0$. When $\sigma_1 = 1$, the multiplicative model is equal to the "goal-based Elo rating" described by Hvattum and Arntzen (2010). A consequence of this formulation is that when the MOV has no theoretical upper limit, the update from a single event can be large. Both σ_1 and α control the size of the multiplier by setting the scale for the MOV and controlling the rate such that $(1+|M_{ijt}/\sigma_1|)$ increases with an increasing margin. In their application to football, Hvattum and Arntzen (2010) found that setting α in a range from 0.2 to 1.6 yielded learning rates with a similar magnitude to the standard Elo model.

If we set $K' = K(1+|M_{ijt}/\sigma_1|)^{\alpha}$, then the change in the rating for the ith competitor can be written as $K'(W_{ijt} - \hat{P}_{ijt})$, which is the identical to the form of the standard Elo model for some dynamic, strictly positive K'. Thus, the stationarity and convergence of the multiplicative model follow directly from the equivalence of the update rule.

3.3.4. Logistic

The final model is the logistic MOV model where the main objective is also predicting win results. The application of the logistic model is distinct from the other win-focused model because when competitor ratings are updated, it replaces the observed win result with a function of the observed margin. Therefore, this model is most suitable for applications where stronger competitors lose frequently or they have close results against lesser competitors. In this approach, the E-step uses the generalized logistic form for the predicted win:

$$\hat{P}_{ijt} = L((R_{it} - R_{jt})/\sigma_2), \tag{12}$$

where $L(x) = 1/(1+\alpha^{-x})$ is a generalized logistic function with base rate α for any $\alpha > 1$. When $\alpha = 10$ and $\sigma_2 = 400$, this function becomes the E-step for the standard Elo model.

The estimated win probability for the logistic model is then compared to the same transformation of the observed margin in the U-step:

$$R_{i(t+1)} = R_{it} + K[L(M_{ijt}/\sigma_1) - L((R_{it} - R_{jt})/\sigma_2)], \tag{13}$$

where the logistic transformation of the MOV places it on a 0-1 scale. If σ_1 and σ_2 are equal to twice the standard

deviation of the MOV and rating difference, respectively, then the difference in the update rule is approximately equal to the comparison of the percentiles of the two standard normal variables. However, whenever the performances in terms of the MOV and the pre-match rating difference are equal in standard deviation units, there will be no change in the ratings, whereas players whose MOV result exceeds their win expectation in standard deviation units will gain in proportion to the percentile difference in their result and win expectation. Thus, setting σ_1 and σ_2 according to the standard deviation of the MOV and rating is intuitively appealing. However, they could also be treated as unknowns together with K and α, and optimized according to the desired objective function.

The reader may have noted that the logistic model is the only model where the "residual" component is not a true residual. The residualized form of the update rule in Elo systems is crucial for their validity, but it is unclear whether the player ratings in the logistic model will converge to their true abilities. First, we can consider the stationarity of the logistic update rule, which is equivalent to determining the expectation of $E[L(M_{ijt}/\sigma_1)]$. If we suppose that $M_{ijt} \sim N(\frac{R_i - R_j}{\delta}, \tau^2)$, such that the mean of the MOV is proportional to the true difference in the player strengths up to some positive constant $1/\delta$, then it can be shown that:

$$E[L(M_{ijt}/\sigma_1)] \approx \frac{1}{1+\alpha^{-h(\tau^2)\frac{(R_i-R_j)}{\delta}}} = L(h(\tau^2)(R_i - R_j)/\delta) \tag{14}$$

where $h(\tau^2)$ a function of the variance of the random variable M_{ijt} (Glickman, 1999). When $h(\tau^2)/\delta = 1/\sigma_2$, this is equivalent to the expected win probability and the stationarity condition will be satisfied. Moreover, for all $\alpha > 1$, the logistic model is continuously differentiable and increasing in the rating difference, thereby satisfying the Lipschitz condition for convergence.

Clearly, the validity of the logistic model is dependent on a stronger set of assumptions than any of the other proposed extensions to the standard Elo system. This is a justifiable reason for questioning this model choice, but it is included because it is the closest generalization to the previous applications of MOV Elo ratings based on the percentage margin, $Y_{it}/(Y_{it} + Y_{jt})$ (where $M_{ijt} = Y_{it} - Y_{jt}$), instead of the observed win result W_{ijt} (Stefani, 2011). Indeed, the official FIFA women's world rankings follow this practice, although with an adjusted percentage margin. Both the percentage margin and logistic model lack a residualized form, but our proposed logistic model has the advantages of demonstrable validity for specific properties of the MOV and a more flexible functional form for transforming the MOV into a 0-1 scale than the percentage margin approach, thereby allowing modelers to tune the model based on its predictive performance.

4. Simulation study

The convergence criteria applied to each model provide some reassurance about their theoretical validity, but the theoretical results are still limited by their reliance on the continuous kinetic model of player interactions and the lack of specific convergence rates for each model type. In the absence of a probability theory for the Elo system to address these limitations, we considered some important practical properties of the models in simulations.

Table 1
Summary of the parameter settings for the simulation study.

Model	Parameters
Linear	$K = 32/3, \sigma = 200$
Joint Additive	$K_1 = 32/6, K_2 = 32/2, \sigma_1 = 200, \sigma_2 = 400$
Multiplicative	$K = 32/2, \alpha = 1, \sigma_1 = 1, \sigma_2 = 400$
Logistic	$K = 32, \alpha = 10, \sigma_1 = 1, \sigma_2 = 400$

Two important properties are the unbiasedness and stable variance of the estimated ratings. We conducted a small simulation study to examine both of these properties in the case of equally matched competitors with ratings that vary randomly with time. Let the difference in the competitor ratings in the ith event be Δ_i. The simulation was performed by drawing $\Delta_i \sim N(0, 50)$ for $N = 1000$ simulated games. Most of the MOV ratings considered in practice are strongly related if not perfectly correlated with the win result, so we first drew the MOV conditional on the difference in the player ratings and then under the condition of the win result based on the observed MOV. In particular, the ith simulated MOV was drawn from the normal distribution, $MOV_i|\Delta_i \sim N(\Delta_i/200, 1)$, and the win result was set equal to $W_i|MOV_i \sim Bernoulli(1/(1 + 10^{-MOV_i/2}))$. The initial ratings were set to 1500 and these settings reflected the standard Elo rating scale.

Each MOV model was applied to the simulated event outcomes using parameter settings consistent with the standard Elo model. To preserve the standard deviation in the 200 ratings, the scale for the linear relationship of the ratings relative to the MOV was set to $200/SD_{MOV} = 200$. The initial values of the learning rates were selected as equivalent to $K = 32$ for all components in the update rule. For example, with the linear model, the choice is $32/3SD_{MOV}$ given that the standard deviation of the win residual is approximately 1/3. All base rates for the logistic curves were set to $\alpha = 10$. For the multiplicative model, we set $\alpha = 1$ and scaled the MOV to the scale of its standard deviation and reduced K by one-half. All of the parameter settings are shown in Table 1.

Based on 10,000 iterations of the equal competitor scenario, only two models exhibited negligible bias, i.e., the joint additive and multiplicative models, whereas the linear and logistic models exhibited positive bias where they overestimated the difference in the ratings of competitors by 7 to 8 points (Table 2). Differences were also observed in the variance properties of the model. After considering the changes in the ratings between consecutive updates, we found that the joint additive model had the lowest variance, whereas the multiplicative model had the highest variance. Analysis of individual iterations, as shown in Fig. 1, suggested that the high-variance models were those with a tendency to exhibit clustering in terms of the ratings, where periods with relative low rating estimates were followed by periods with relatively

Table 2
Summary of simulation study results (N = 10,000)

Model	$\hat{\Delta}$ (95% CI)	Mean SD between games (95% CI)
Linear	7.13 (−4.77, 19.87)	22.57 (21.63, 23.55)
Joint Additive	4.19 (−8.43, 15.99)	19.93 (19.28, 20.53)
Multiplicative	0.29 (−21.91, 22.10)	28.37 (29.63, 29.98)
Logistic	7.93 (−6.33, 22.31)	22.15 (21.90, 22.45)

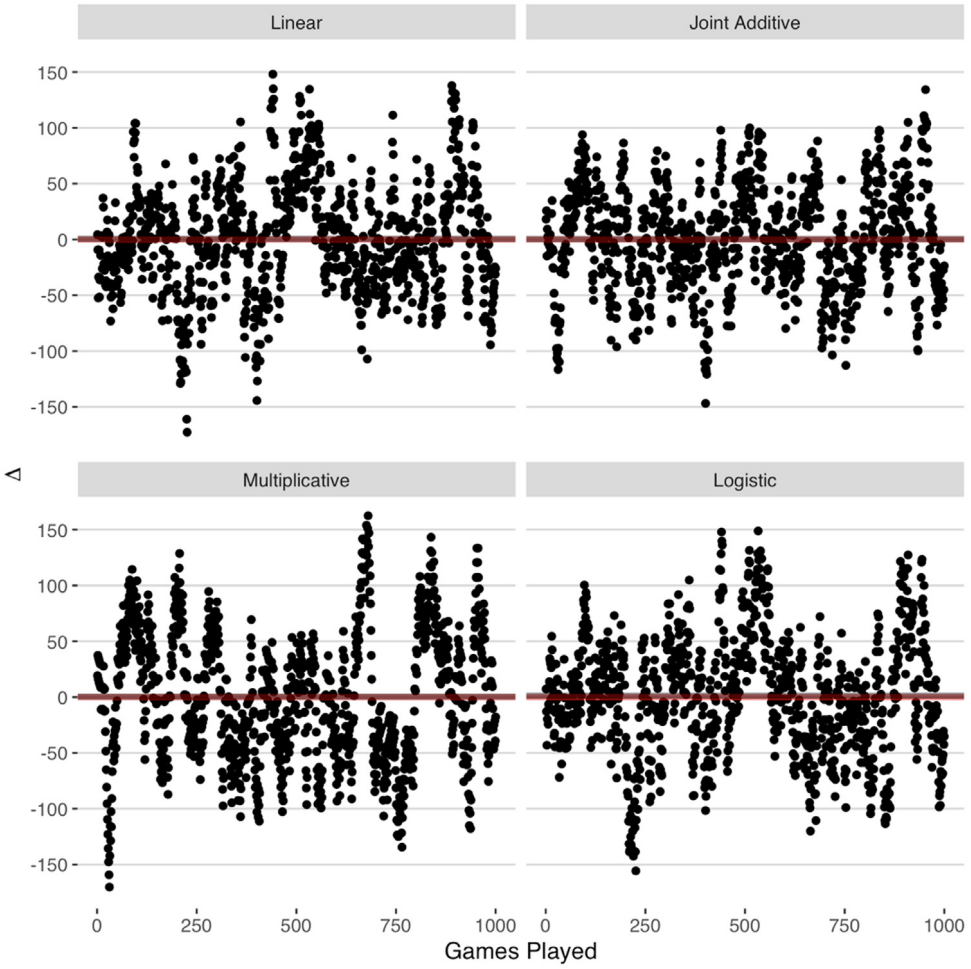

Fig. 1. Difference in competitor ratings for one iteration of the simulation study where the average difference was zero (equal competitor scenario). The horizontal red line denotes zero and the horizontal grey line is the observed average across all of the simulated games.

high ratings. Thus, the consistency of the estimates obtained using these approaches was sensitive to chance streakiness. The joint additive model obtained the most stable estimates among the four methods, thereby indicating that the variance was more constant across the observations.

5. Application

5.1. Model tuning

In real-world applications, each MOV model will require the tuning of two or more parameters. Thus, a multiparameter optimization approach was used to determine the parameter values that minimized the objective function in a training data set. Given N match results with the winning player as the reference player throughout, the form of the objective function was:

$$\mathcal{L}(\theta) = 1/N \left[\frac{\sqrt{\sum_{i,j,t}(\hat{M}_{ijt}(\theta) - M_{ijt})^2}}{3SD} - \sum_{i,j,t} log(\hat{P}_{ijt}(\theta)) \right] \tag{15}$$

with a weighted sum of the root mean squared error (RMSE) for the MOV and negative log-loss for the win prediction. The RMSE is the popular L^2 or least-squared loss for continuous outcomes, and log-loss is the common objective for binary outcomes, where it is equal to the

log-likelihood or cross-entropy of the observed successes and failures. In Eq. (15), $\hat{M}_{ijt}(\theta)$ is the margin of victory estimate and $log(\hat{P}_{ijt}(\theta))$ is the win prediction estimate for the winning player given tuning parameters θ. The sum of these two components allows both the error in the margin and win results to influence the choice of parameters. To understand the rationale behind the specific choice of sum, it should be noted that we can equate the objective function to the weighted sum of the log likelihood for normally distributed MOV with means $\hat{M}_{ijt}(\theta)$ and the win results with probabilities $\hat{P}_{ijt}(\theta)$ if these outcomes are independent. The factor $3SD$ is three times the standard deviation of the MOV variable, which places the RMSE on the same scale as the log-loss, thereby giving both components equal weight in the minimization process. The linear model only estimates the MOV so we set the log-loss component to zero for the objective function. The RMSE component was set to zero for the logistic and multiplicative models where the estimation step only involves the win outcome.

Optimization was performed with a bound-constrained quasi-Newton method (Nash, 2014). The use of bounds ensured the positivity of the parameters and also helped to improve the efficiency of the optimization algorithm. Due to the recursive nature of Elo ratings, for every choice of parameters, the ratings must be updated from the first to the last match in the training data. Without efficient scripts and careful selection of the starting values and bounds, the optimization process can take hours or more to run on data sets with the size considered in this study. Thus, in addition to optimizing the programs for the rating algorithms, we developed a strategy for selecting the initial values and parameter ranges for each MOV variable. The initial values were set according to the same approach described for the simulation study above, and they were largely determined by the size of SD_{MOV}. The lower and upper bounds were set using five times and one-fifth the size of SD_{MOV}, respectively. For example, considering "games won" with $SD = 4$ and the linear model with tuning parameters K and σ, this approach used a starting value 50 for σ with bounds of $(10, 125)$ and 2.7 for K with bounds of $(0.5, 13.3)$. For the multiplicative model parameter α, the bounds were set to $(0.2, 2)$ in a similar manner to the range considered by Hvattum and Arntzen (2010).

The training data set comprised 10 years of ATP singles matches (2005–2015) where the aggregated match statistics were taken from the OnCourt database (www.oncourt.info). Matches were included from all events at the Challenger level and above, which can earn official ranking points. Matches that ended in retirement were excluded. In total, 55,602 matches satisfied these conditions and they comprised the sample for model training. To prevent initial adaptation from influencing the choice of tuning parameters, each iteration used all training data for updating the ratings but only the last three years of the training data were used in the objective function summary. Finally, fivefold cross-validation was conducted in order to prevent overfitting and the selected tuning values were the medians across the folds.

Table 3
Distribution of margin of victory (MOV) variables for the winners of matches (n = 36,471)

MOV variable	Median (IQR)	Positivity rate
Sets Won	2 (1)	100%
Games Won	5 (4)	98%
Break Points Won	2 (2)	97%
Total Points Won	14 (10)	96%
Serve Percentage Won	11% (13%)	93%

5.2. Performance evaluation

The performance of each model was evaluated based on main draw ATP matches at the Challenger level and above held in 2016 and 2018, where the matches in these years were not included in the model training process. Three years of data for out-of-sample testing were assigned to a large data set in order to obtain precise comparisons at both the overall levels and by tournament tier, which helped to avoid seasonal effects. The distributional characteristics of the winners of matches in the test sample of men's professional matches are shown in Table 3.

All of the margins had a positivity rate of 93% or greater in the study test sample, thereby indicating that they had strong relationships with the outcomes of professional matches. In terms of the variation from match to match, the set margin was the least informative with the lowest variance-to-mean ratio, and the information provided by the other margin variables was similar. After examining the pairwise correlations of the margins, we found that the correlations ranged from 0.48 to 0.96, where the strongest correlations were between the break points and games won ($r = 0.96$) and the serve percentage and total points won ($r = 0.89$).

Elo ratings require a period of time to adapt, so the algorithms were applied to matches starting in 2005 and they were run for all matches, including qualifying rounds, until the end of 2018 by using the parameters determined by model training. This process replicated how the Elo forecasts would have performed from 2016 to 2018 in a real-world scenario based on the information available prior to 2016. The predictive performance of MOV was evaluated based on the RMSE and the match win prediction performance was evaluated using accuracy and log-loss metrics. All of the statistics were summarized across tournament categories and within each tournament category. Confidence intervals were obtained for the performance metrics using 1000 bootstrap resamplings of the test data, where an equal number of samples were tested with replacement from the test data in order to account for the variance due to the test sample, while ignoring the uncertainty in terms of the choice of tuning parameters.

Model construction and analysis were performed in the R statistical programming language (R Core Team, 2018). An R package has been written by the author with functions for implementing each model type and with Grand Slam tennis data for testing, which is available as Supplementary Material.

Table 4
Summary of the optimized parameter settings for the margin of victory (MOV) variable models.

MOV	Linear	Joint additive[a]	Multiplicative	Logistic
Sets Won	$\sigma = 380$	$K_1 = 7.5$	$\sigma_1 = 1$	$\sigma_1 = 0.5$
	$K = 15.00$	$K_2 = 34$	$K = 20$	$K = 60$
			$\alpha = 0.8$	$\alpha = 2$
Games Won	$\sigma = 75$	$K_1 = 2.5$	$\sigma_1 = 4$	$\sigma_1 = 2.5$
	$K = 4.75$	$K_2 = 24$	$K = 20$	$K = 100$
			$\alpha = 1$	$\alpha = 2$
Break Points Won	$\sigma = 150$	$K_1 = 4.5$	$\sigma_1 = 2$	$\sigma_1 = 1.5$
	$K = 9.30$	$K_2 = 28$	$K = 22$	$K = 60$
			$\alpha = 1$	$\alpha = 6$
Total Points Won	$\sigma = 25$	$K_1 = 1.0$	$\sigma_1 = 10$	$\sigma_1 = 0.5$
	$K = 1.69$	$K_2 = 26$	$K = 20$	$K = 80$
			$\alpha = 1$	$\alpha = 6$
Serve Percentage Won	$\sigma = 2200$	$K_1 = 76.0$	$\sigma_1 = 0.13$	$\sigma_1 = 0.09$
	$K = 153.00$	$K_2 = 20$	$K = 24$	$K = 60$
			$\alpha = 1$	$\alpha = 6$

Scaling factors σ_1 were set to the same value as that in the linear model; $\sigma_2 = 400$ for all MOV.

6. Results

The parameters that reduced the objective function in the training data for the linear model obtained multiplicative factors and learning rates that had inverse relationships with the magnitude of the MOV variable's standard deviation (Table 4). In general, variables with lower numerical variance had a higher optimal σ and K. The MOV learning rates for the joint additive model were approximately half the magnitude of the corresponding rates in the linear model. Only one MOV, i.e., "sets won", had a greater learning rate for win expectations than the standard rate of 32. All of the other tuning values for K_2 were smaller in magnitude, where the smallest was for "serve percentage won", thereby suggesting that the MOV in this model was given greater relative importance. For the multiplicative model, the scale of σ_1 closely matched the scale of the MOV standard deviation, where the power $\alpha = 1$ was the optimal choice for nearly all of the margin variables. It should also be noted that the magnitude of K for the logistic model was approximately two times the magnitude compared with the standard Elo model, which could be explained by the reduced variance due to the logistic transformation applied in its update rule.

In terms of overall accuracy of the match win predictions, the highest out-of-sample accuracy was 68%, which was achieved by three models comprising the joint additive model with "serve percentage won", the multiplicative model with "serve percentage won", and the logistic model with "break points won" (Fig. 2). By contrast, the accuracy of the standard Elo model was 66% (95% CI 65– 67%). Thus, the best-performing MOV models provided an average two percentage point gain in accuracy compared with the standard ratings. Indeed, although some gains were modest, the average accuracy of all the MOV models was higher than that of the standard Elo method. However, only the bootstrap confidence intervals of the "joint additive" and "logistic" MOV models consistently exceeded the accuracy of the standard Elo model.

There were differences in performance according to the model type and MOV employed, but the differences due to model choice were smaller than those due to MOV. Indeed, when averaged over the MOV, the average prediction accuracy for all models was 67%. In particular, the linear and multiplicative models exhibited more variable performance in terms of MOV compared with the joint additive and multiplicative models, where the average accuracy spread ranged from 1.1% for the logistic, 0.7% for the linear, and 0.3% for the joint additive and multiplicative models (Fig. 2). "Break points won" and "games won" had the two highest average accuracy results across all models (67.4% and 67.2%, respectively), while "serve percentage won" achieved two of the highest accuracy results for two specific model choices comprising the joint additive and multiplicative models. These results suggest that "serve percentage won" was a less robust model choice than "break points won" or "games won" but it still achieved some of the best results in specific models.

In contrast to the accuracy, the log-loss metric assigns more weight to incorrect predictions that greatly favor the wrong competitor. Only one model comprising the joint additive model with "serve percentage won" achieved a log-loss as low as 0.601 (95% CI 0.591–0.610) (Fig. 3). Despite the very similar accuracy across the MOV types, the log-loss with the joint additive model outperformed the multiplicative model for all MOV types except for "sets won". The standard Elo had a log-loss of 0.613 (95% CI 0.605–0.624). The majority of the models performed better than the standard Elo model, but the models that used "sets won" as the MOV performed worse in three of the four models, and significantly worse for the linear and joint additive models.

The performance in terms of the point estimates for the margin differed very little across the two model types that incorporated an MOV estimate in their estimation step (Table 5). Notable differences were found with respect to the choice of MOV variable. In particular, "break points won" and "serve percentage won" were the only score variables with RMSE values of less than two standard deviations in magnitude.

Our comparison of log-loss performance by tournament tier showed that the performance increased with the tournament tier for all models, where the average log-loss for the MOV models was 0.507 at Grand Slams and 0.650 for ATP 250 events, but there was considerable

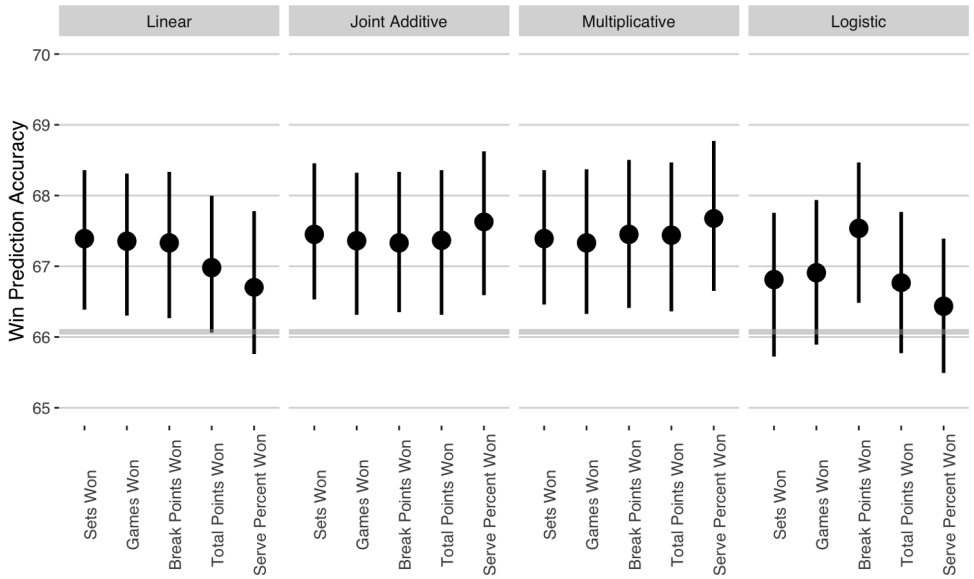

Fig. 2. Overall prediction accuracy in the test sample from 2016 to 2018. Horizontal lines show the median accuracy with the standard Elo. Lines denote the 95% bootstrap confidence interval. Higher values indicate better predictive performance.

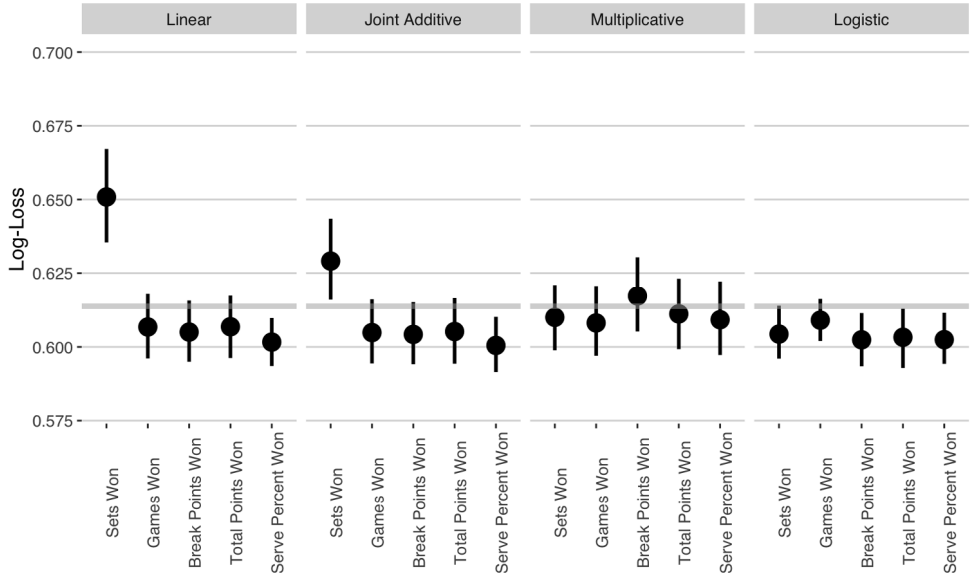

Fig. 3. Overall log-loss in the test sample from 2016 to 2018. Horizontal lines show the median log-loss with the standard Elo. Lines denote the 95% bootstrap confidence interval. Lower values indicate better predictive performance.

Table 5
RMSE (95% CI) results for margin of victory (MOV) expectations in the test sample.

MOV	Linear	Joint additive
Sets Won	1.66 (1.64–1.68)	1.66 (1.65–1.68)
Games Won	4.92 (4.86–4.98)	4.92 (4.85–4.98)
Break Points Won	2.35 (2.32–2.39)	2.36 (2.33–2.39)
Total Points Won	14.16 (13.96–14.33)	14.16 (13.97–14.36)
Serve Percentage Won	0.122 (0.120–0.1245)	0.123 (0.122–0.126)

overlap in terms of the point estimates and confidence intervals for the ATP 500 and Masters results (Fig. 4). Gains over the standard Elo were observed for all tiers but on average, they were highest for the ATP 500 level and above where the percentage improvement in the log-loss ranged from 1.5–2.2%. The joint additive model had

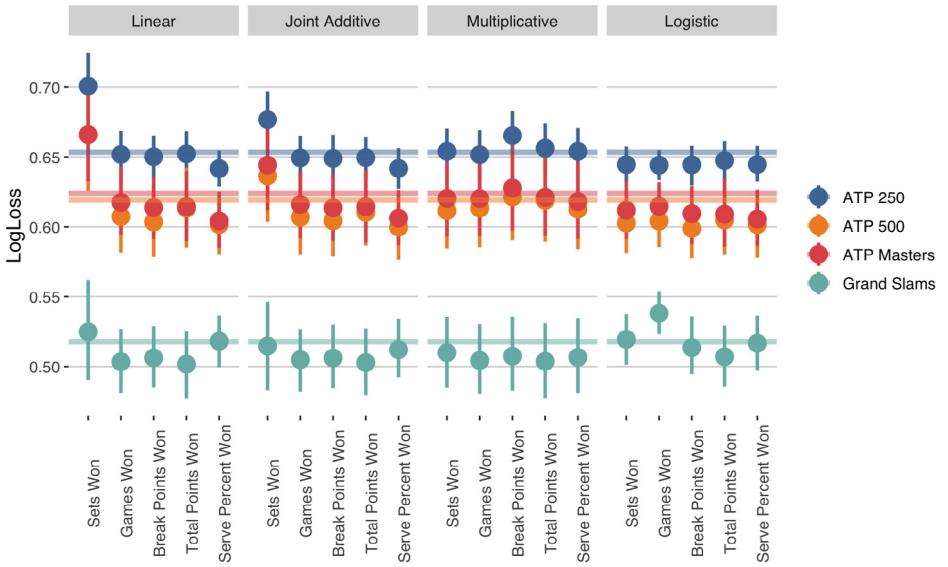

Fig. 4. Prediction log-loss by tournament category for ATP 250 tournaments and above in the test sample from 2016 to 2018. Horizontal lines show the median log-loss with the standard Elo. Lines denote the 95% bootstrap confidence interval. Lower values indicate better predictive performance.

the strongest performance across tournament levels for all MOV types except for "sets won", with a 1.6–2.4% gain compared with the log-loss for the standard Elo at ATP 500 events and higher. Interestingly, "serve percentage won" was the best-performing MOV type for the joint additive model in most tournament tiers, but this was not the cases for Grand Slams where only "games won", "break points won", and "total points won" provided significant improvements in performance compared with the standard Elo.

Summaries across many matches and players do not tell us how the MOV models might affects the rating progression for any one player. As an illustrative example of an MOV Elo system, we considered the ratings progression for one player: Alex De Minaur. De Minaur has been one of the most successful young Australian players in recent times. Before the age of 19, he reached his first tour final in 2018 at the Sydney International and went on to win that event in the following year. His rapid rise over a short period of time makes him an especially interesting case for comparing how different rating systems adapted during his breakout period.

Using the joint additive model and "break points won" as the MOV, Fig. 5 compares how De Minaur's Elo rating progressed from the start of the 2018 season with an MOV Elo rating versus the standard Elo rating (shown in grey). Both systems obtained steady increasing trends, but the ratings clearly diverged in two sections of the time period. The first divergence occurred over the first 20 matches during the 2018 season (January to April). This period included a semifinal result in Brisbane, a runner-up finish in Sydney, and another runner-up finish at the Alicante Challenger event. Throughout this period, the MOV Elo rating was higher for De Minaur and it was 80 points higher than the standard Elo rating at its peak. This divergence occurred because De Minaur's wins during this period had large positive margins, where two matches at the start of the season had margins of +4. In addition, his two losses in Brisbane and Sydney were very close in terms of the break points won. De Minaur won an equal number of breaks as Ryan Harrison, who stopped his run in Brisbane, and he was actually ahead by 1 break point during his loss to Daniil Medvedev in Sydney.

A similar pattern occurred again toward the end of the 2018 season from the US Open until De Minaur's third round loss at the Shanghai Masters (Fig. 5). Similar to the start of the year, De Minaur's MOV Elo rating increased to a much higher level in this period compared with the standard Elo rating due to a succession of wins with dominant break point winning margins punctuated by close losses, where the difference in break points won was minimal.

7. Discussion

In this study, we conducted the first general investigation of the design of MOV Elo rating systems. Four approaches were proposed with the same estimation and update formula as Elo, but they incorporate MOV information using different methods. The linear and joint additive models include MOV as the estimation target, whereas the multiplicative and logistic models maintain the same estimation step as the Elo system but they use MOV information to moderate the learning rate in the case of the multiplicative model and the observed outcome in the case of the logistic model. We provided general guidance for model tuning, which can be used in a range of applications. The information presented in this study will provide modelers with greater flexibility in the design and implementation of rating systems for their sport of interest.

The four models have different targets in the estimation step, where the linear model targets estimates for the MOV, the logistic and multiplicative models target

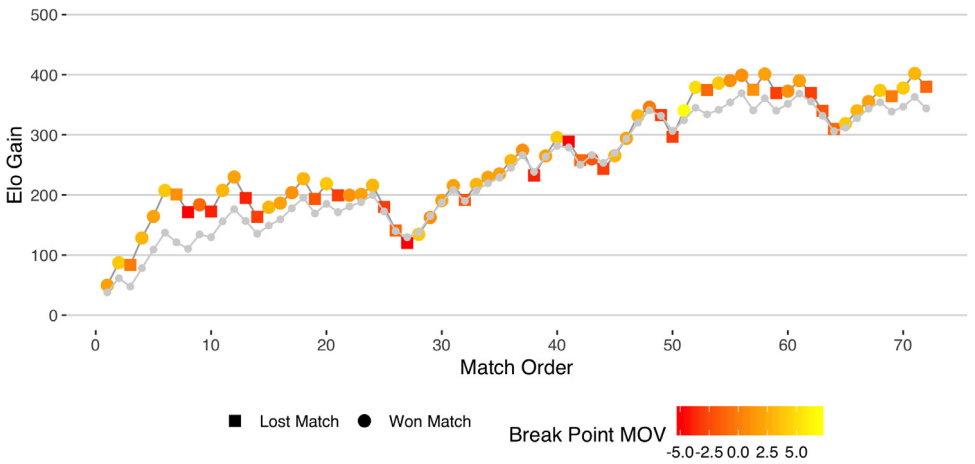

Fig. 5. Comparison of gain in ratings with the standard Elo and joint additive Elo with "break points won" as the margin of victory (MOV). All ratings shown are the cumulative gain for Alex De Minaur from the start of 2018 until early 2019.

estimates for the win result, and the joint additive model targets both, but they are not direct alternatives in terms of the prediction goal. Instead, each provides a strategy for estimating the dynamic latent abilities of competitors. When either the win or the MOV outcomes of a rating system are assumed to be functions of the relative strengths of competitors, the simulation results presented in this study suggest that the joint additive model should be preferred to the alternatives because this method obtained the most stable variance and minimal bias with respect to the differences in the competitor ratings.

For many applications, the predictive performance for win results is the primary objective and the property based on which MOV models are judged (Hvattum & Arntzen, 2010). Using a real-world application in professional tennis, we provided guidance regarding the tuning and testing of each model for forecasting. This is the first application of an MOV rating system to this sport and we obtained several novel conclusions regarding tennis rating systems. First, we found that all of the four types of MOV models improved the performance of the win predictions when the margin variables used within-set information, where the best-performing models added two percentage points to the overall accuracy and a decrease of 0.012 in the log-loss for win forecasts compared with the standard Elo. A previous study compared the performance of 11 different tennis prediction models at predicting men's tour-level win results and determined prediction accuracies of 59% to 70% and log-loss measures from 0.59 to 0.68 (Kovalchik, 2016). The absolute values obtained in previous studies are not directly comparable with the present study due to differences in the years and tournament tiers considered, but their performance ranges provide a basis for assessing the magnitude of improvement and they suggest that the gains observed are meaningful in practice.

Second, using "sets won" alone obtained no improvement compared with the standard Elo system and it was worse in terms of the log-loss for several models. The lack of improvement in the accuracy reflects the limited information a set margin provides regarding the magnitude of the win because most matches are played as the best of three and they will have a margin of either $+2$ or $+1$ for the winning player. Thus, "sets won" is more similar to a binary outcome than a continuous outcome, and it was the most strongly correlated with the win result. Both of these factors could result in over-confident updates, as demonstrated by the poor log-loss performance of the linear and joint additive models using "sets won" as the margin of victory. We hypothesized that "sets won" would not obtain improvements compared with the standard Elo, but the great difference in its performance compared with the other MOV choices suggests that these models are more suitable for continuous MOV, where the MOV is not strongly dependent on the win result in the case of the joint additive model.

Third, for a given MOV variable, the win prediction performance was in a similar range for the four model types, thereby suggesting that the performance gains were fairly robust to the choice of implementation for this application. The joint additive and logistic models that used "break points won" had some of the strongest performance characteristics in terms of both the win predictions and MOV predictions. There is less than one break point per game in professional tennis where one of every three service games has a break of service (Knight & O'donoghue, 2012). These characteristics and the findings obtained in this study suggest that much of the information regarding a player's win ability is contained in a small number of key points.

The joint additive model exhibited strong predictive performance among the MOV models for several margin variable choices. Only two choices of MOV also yielded good estimates for the MOV, i.e., "break points won" and "serve percentage won", and thus the linear expectation model was most reasonable for these margins. Interestingly, these MOV variables also had the strongest predictive performance for the win result, which suggests that a linear relationship with the rating difference is an important property when selecting an MOV.

Setting accurate expectations regarding the difference in serve performance between opponents has important implications for mathematical models in tennis. Assuming that service points are independent and identically distributed, the probability of any event of interest can be derived analytically, where the only input is the expected win probability on serve for one or both players (Newton & Keller, 2005). These formulae may facilitate further research, including the development of models for forecasting outcomes during matches (Kovalchik & Reid, 2018), assessing performance under pressure (Knight & O'donoghue, 2012), and deriving a measure of point importance (Morris, 1977). The discovery that the joint additive MOV Elo system with a "serve percentage won" margin can improve the log-loss for win predictions compared with the standard Elo system and yield good predictions of the difference in the "serve percentage won" means that it is a desirable choice for rating tennis players.

Similar to the MOV extensions proposed in this study, the Elo system is fundamentally an algorithm and it has not received the same theoretical treatment as fully model-based paired comparison methods (Aldous et al., 2017). In this study, we established some general conditions for assessing the validity of the models and their convergence to the true relative strengths when they are applied to a large competitive pool where players of all levels are likely to meet. In this case, the logistic model requires the strongest conditions for convergence to be satisfied because it is the only model that deviates from the standard "residualized" update rule. The properties of MOV model ratings were also studied in non-asymptotic simulations of stationary ratings. Our findings provide additional reassurance regarding some desirable properties of the models, i.e., unbiasedness and stable variance when the ratings are stationary, but further studies of the theoretical properties of the Elo system and its extensions would be valuable.

The development of model-based analogs to the dynamic MOV rating models presented in this study may help to strengthen the theoretical properties of these methods. Model-based dynamic systems, such as the Glicko system (Glickman, 2001) or Kalman filter (Baker & McHale, 2014; Fahrmeir & Tutz, 1994), have crucial advantages for measuring the uncertainty of ratings and allowing updates of the variances of ratings in addition to their mean over time. Most of these models are intended for binary or ordinal outcomes, thereby limiting their application to continuous MOV. Some exceptions are the *moderated* paired comparison model proposed by Stern (2011), which uses penalized likelihood methods to incorporate a continuous MOV outcome into a Bradley–Terry framework, as well as the MOV models proposed by Harville (2003) and Stefani (1980), which are both designed based on the least-squares methodology. Comparisons of model-based approaches have frequently demonstrated no gain in predictive performance compared with the Elo algorithm (Coulom, 2008; Glickman, Hennessy, & Bent, 2018), but further development of these models would be of great interest and the inferential advantages would justify the additional modeling and computational complexity.

Exploiting the connection between dynamic paired comparison models and the Kalman filter (Fahrmeir & Tutz, 1994) may be a promising strategy for developing model-based analogs for the class of MOV Elo systems introduced in this study.

Modelers of sports rating systems often ask whether a model would be better if it considers the magnitude of wins? There is no single answer to this question, but in the present study, we provided new tools that modelers can use to test this question in the their sport of interest. By providing a general framework for MOV Elo systems with accompanying implementation guidelines, we hope that more researchers will further develop and improve the rating systems used in various sports.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2020.01.006.

References

Aldous, D., et al. (2017). Elo ratings and the sports model: A neglected topic in applied probability? *Statistical Science*, *32*(4), 616–629.

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

Barrow, D., Drayer, I., Elliott, P., Gaut, G., & Osting, B. (2013). Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, *9*(2), 187–202.

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

Carbone, J., Corke, T., & Moisiadis, F. (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.

Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 412–433.

Constantinou, A. C., & Fenton, N. E. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, *9*(1), 37–50.

Coulom, R. (2008). Whole-history rating: A Bayesian rating system for players of time-varying strength. In *International conference on computers and games* (pp. 113–124). Springer.

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

Fahrmeir, L., & Tutz, G. (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association*, *89*(428), 1438–1449.

Glickman, M. E. (1995). A comprehensive guide to chess ratings. *American Chess Journal*, *3*, 59–102.

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.

Glickman, M. E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, *28*(6), 673–689.

Glickman, M. E., Hennessy, J., & Bent, A. (2018). A comparison of rating systems for competitive women's beach volleyball. *Statistica Applicata - Italian Journal of Applied Statistics*, *30*(2), 233–254.

Harville, D. (1977). The use of linear-model methodology to rate high school or college football teams. *Journal of the American Statistical Association*, *72*(358), 278–289.

Harville, D. A. (2003). The selection or seeding of college basketball or football teams for postseason competition. *Journal of the American Statistical Association*, *98*(461), 17–27.

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

Jabin, P.-E., & Junca, S. (2015). A continuous model for ratings. *SIAM Journal of Applied Mathematics*, *75*(2), 420–442.

Knight, G., & O'donoghue, P. (2012). The probability of winning break points in Grand Slam men's singles tennis. *European Journal of Sport Science*, *12*(6), 462–468.

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

Kovalchik, S., & Reid, M. (2018). A calibration method with dynamic updates for within-match forecasting of wins in tennis. *International Journal of Forecasting*, *35*(2), 756–766.

Langville, A. N., & Meyer, C. D. (2012). *Who's #1? The science of rating and ranking*. Princeton University Press, URL http://gen.lib.rus.ec/book/index.php?md5=65b69ca76799745881e1de18e03e6f8e.

Mangan, S., & Collins, K. (2016). A rating system for gaelic football teams: Factors that influence success. *International Journal of Computer Science in Sport*, *15*(2), 78–90.

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

Morris, C. (1977). The most important points in tennis. *Optimal Strategies in Sport*, 131–140.

Nash, J. C. (2014). On best practice optimization methods in R. *Journal of Statistical Software*, *60*(2), 1–14.

Newton, P. K., & Keller, J. B. (2005). Probability of winning at tennis I. Theory and data. *Studies in Applied Mathematics*, *114*(3), 241–269.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org/.

Stefani, R. T. (1980). Improved least squares football, basketball, and soccer predictions. *IEEE Transactions on Systems, Man, and Cybernetics*, *10*(2), 116–123.

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, *7*(4).

Stern, H. S. (2004). Statistics and the college football championship. *The American Statistician*, *58*(3), 179–185.

Stern, S. E. (2011). Moderated paired comparisons: a generalized Bradley–Terry model for continuous data using a discontinuous penalized likelihood function. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *60*(3), 397–415.

Wright, B., Rodenberg, M. R., & Sackmann, J. (2013). Incentives in best of n contests: Quasi-Simpson's paradox in tennis. *International Journal of Performance Analysis in Sport*, *13*(3), 790–802.