The predictive power of ranking systems in association football

Jan Lasek*

Faculty of Sciences,
VU University Amsterdam,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
and
Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw,
Banacha 2, 02-097 Warsaw, Poland
E-mail: jan.lasek@student.uw.edu.pl
E-mail: janek.lasek@gmail.com
*Corresponding author

Zoltán Szlávik

Department of Computer Science,
Computational Intelligence Group,
VU University Amsterdam, De Boelelaan 1081a,
1081 HV Amsterdam, The Netherlands,
E-mail: z.szlavik@vu.nl

Sandjai Bhulai

Department of Mathematics,
Stochastic Operations Research,
VU University Amsterdam, De Boelelaan 1081a,
1081 HV Amsterdam, The Netherlands,
E-mail: s.bhulai@vu.nl

Abstract: We provide an overview and comparison of predictive capabilities of several methods for ranking association football teams. The main benchmark used is the official FIFA ranking for national teams. The ranking points of teams are turned into predictions that are next evaluated based on their accuracy. This enables us to determine which ranking method is more accurate.

The best performing algorithm is a version of the famous Elo rating system that originates from chess player ratings, but several other methods (and method versions) provide better predictive performance than the official ranking method. Being able to predict match outcomes better than the official method might have implications for, e.g., a team's strategy to schedule friendly games.

Keywords: FIFA ranking; predictive capabilities; predictive power; rankings; ratings; team strength; football predictions.

Reference to this paper should be made as follows: Lasek, J., Szlávik, Z. and Bhulai, S. (2013) 'The predictive power of ranking systems in association football', *Int. J. Applied Pattern Recognition*, Vol. 1, No. 1, pp.27–46.

Biographical notes: Jan Lasek is a student of the final years of Master programmes in Mathematics and Economics at the University of Warsaw, Poland. He completed his Master's Thesis in Mathematics at the VU University Amsterdam during one-year exchange programme. This article is a summary of his thesis. His research interests include statistics, econometrics and data analysis techniques from related fields of data mining and machine learning, particularly with applications in finance and football predictions.

Zoltán Szlávik graduated at the University of Pannonia, having studied computer science/informatics. He received his PhD from Queen Mary, University of London in 2008 where his thesis was on 'Content and structure summarisation for accessing XML documents'. Shifting his focus from information retrieval, in 2009, he joined the Computational Intelligence Research Group of the VU University Amsterdam as a Post-doctoral Researcher, where his main focus has been data mining. He also developed (or redesigned) and delivered several courses there. He is particularly interested in data analysis and research with strong social impact.

Sandjai Bhulai received his MSc degrees in Mathematics and in Business Mathematics and Informatics, both from the VU University Amsterdam, The Netherlands. He carried out his PhD research on 'Markov decision processes: the control of high-dimensional systems' at the same university for which he received his PhD in 2002. After that, he has been a Post-doctoral Researcher at Lucent Technologies, Bell Laboratories as NWO Talent Stipend Fellow. In 2003, he joined the Mathematics Department at the VU University Amsterdam, where he is an Associate Professor in Applied Probability and Operations Research. His primary research interests are in the general area of stochastic modelling and optimisation, in particular, the theory and applications of Markov decision processes. His favourite application areas include telecommunication networks and call centres. He is currently involved in the control of time-varying systems, partial information models, dynamic programming value functions, and reinforcement learning.

1 Introduction

Rankings in sports have a number of important applications. One of the roles of ranking tables is to provide an objective indication of the strength of individual competitors, based on their previous performance. In this way, rankings provide information about the actual level and current progress for competing parties, and encourage competition. Accurate rankings can be used in scheduling match-ups by pairing teams or players of similar strength. This is strongly connected to perhaps the most important application of rankings in scheduling competitions: when tournaments are preceded by a draw, teams or players are seeded according to official rankings. It is common to pair higher and lower ranked rivals to prevent the strongest opponents (those ranked highest) to meet in an early stage of the competition. Therefore, rankings have a crucial impact on the competition.

Another use of rankings, in association football, is that the UK Government uses national teams' rankings for granting work permits for players outside the European

Union. According to the rules, a player is eligible to play in England when his country has a ranking position higher than the 70th rank, averaged over a period of two years (Internationalworkpermits.com, 2012). Hence, the official rankings influence player careers as well.

In this paper, we focus on international football, and rankings of national teams, whose corresponding official ranking is maintained by *Fédération Internationale de Football Association* (FIFA) the international governing body of football. We provide an overview of existing ranking methods in sports, including that by FIFA, and compare them using two evaluation measures described later in this paper. The goal is to assess the predictive capabilities of these ranking systems.

The remainder of the paper is structured as follows. After presenting related work, we describe and briefly discuss several ranking methods used in our experiments. Afterwards, we present our experiments comparing these ranking methods. Finally, we discuss the results and conclude with future work.

2 Related work

The FIFA ranking method is often subject to criticism. A constructive judgment of the ranking was done by McHale and Davies (2007). By building and analysing several statistical models for predicting match results, the authors conclude that the ranking does not use the information on past results efficiently and it does not react quickly enough to recent changes in team performance. A suggestion is made to look for another ranking system or improving the current one.

Several authors studied the efficacy of predictions in terms of agreement between the ranking and results of major football competitions. Suzuki and Ohmori (2008) evaluate the accuracy of predictions based on the official ranking with respect to the results of four World Cup tournaments between 1994 and 2006. The authors conclude that ranking-based predictions are reasonably accurate. Luckner et al. (2008) compare predictions based on the FIFA ranking against forecasts derived from a market for football teams specifically created for this purpose. The predictions are evaluated against the final standings of the World Cup tournament in 2006 and the market forecasts turn out to be more accurate than those based on the FIFA ranking. Leitner et al. (2010) compare the accuracy of the FIFA ranking and bookmakers' predictions of the results of the 2008 European Championships. They measure accuracy using Spearman's rank correlation between the final tournament standings and ranking tables. They show that bookmakers are more accurate than the FIFA ranking in their predictions.

In the related work described above, evaluation of the FIFA ranking is based solely on the position of a team in the table. Predictions are made by indicating that the higher ranked team will win the game. In our work, we examine the accuracy of predictions based on the rating points rather than only the ranking position. We treat the official rating method as a benchmark and discuss several methods for measuring team strength. Based on a day's rating points, we make predictions for games played the next day.

It is worth mentioning the work spent on rating of the chess players, which has a long history (Glickman, 1995). These rating systems receive much attention and serve as basis for rating methods in other sports. They are also constantly improved. For example, in recent years the website *Kaggle.com* (2010, 2011) hosted two competitions where the

goal was to improve accuracy of current chess player ratings methods. We also used chess-based ranking methods in the work described in this paper.

3 Overview of ranking systems

In this section, we describe several rating systems whose performance we will later compare. Though the list of methods described here is not exhaustive, we believe our sample contains the main ranking methods used in sports, and it is sufficiently diverse to provide a meaningful comparison of ranking method types. Throughout the text, when we refer to a *ranking system*, we are in fact interested in the *rating points* provided by the described ranking methods, which are then used to determine actual rankings by each of these methods.

We begin with discussing 'earned rating' methods, where teams accumulate points after each game. Two examples here are the official FIFA ranking and the Elo rating system. Then we present two methods that estimate strengths from the global look at match result data rather than by an iterative updating of ratings after each game. The methods of this kind are the Elo++ rating system – the winner of the first *Kaggle* competition on rating chess players – and the least squares ratings. Finally, we discuss graph-based ranking methods.

3.1 The official FIFA ranking

The current FIFA World Ranking methodology was introduced after the World Cup in Germany in 2006. Its original description is available via the official FIFA website (FIFA.com, 2012b).

To calculate ranking points for teams, four years of play are considered. During that period of time a weighted average of points is computed that results in a team's rating points.

For a chosen team the formula for the calculation of points P awarded after a single game is as follows

$$P = M \times I \times T \times C, \tag{1}$$

where the letters stand for:

- the outcome of the game (M points)
- the importance of the game (I)
- the strength of the opposing team (T)
- the average of confederation strengths (C) of participating teams.

For the outcome of the game a standard convention is applied. For a win three points are awarded, one for a draw, and zero for a loss. The matches that ended after a penalty shoot-out are treated differently – a winning team receives two points while a losing team gets one point. The multiplier for the importance of a game, I, assumes values between one and four. The World Cup games are considered the most important, while friendly games the least important, with I set to four and one, respectively. The confederation strength is a number between [0.85, 1]. Currently, European (UEFA) and South

American (CONMEBOL) football confederations are assigned the maximal value of confederation strength. The lowest rated confederation is the Oceania Football Confederation (OFC). When two teams play, C is computed as the average of the confederation strengths to which they belong. Finally, the strength of opposition is computed as 200 minus the position of the opponent in the current ranking release. As an exception to the formula, the team at the top is assigned the maximum strength of 200 and teams ranked at position 150 or lower get the minimal strength of 50.

Once we calculate the points that a team earned over a period of four years, we compute a weighted average of points in the following manner. In the consecutive years the mean of the accumulated points is computed. In case a team played less than five games in a chosen year, instead of calculating the average, we divide its total number of points by five. Then the four yearly averages are summed up with weights 0.2, 0.3, 0.5, and 1, where more recent results are assigned a higher number.

The FIFA ranking is released on an approximately monthly basis. To get better insight into the capabilities of the official rating system we implement the algorithm to obtain team ratings on a daily basis.

3.2 The Elo rating system

The Elo rating system was created by the Hungarian physicist and chess master Arpad Emrick Elo. It is one of the most prominent systems for rating skills in two-player games. Due to its general merits, it is the first system we introduce after the official FIFA ranking. It has several generalisations including Glicko (Glickman, 1999) or TrueSkill (Herbrich et al., 2007) rating systems. Primarily, it was used for rating chess players. For a more detailed discussion of the Elo rating system we refer to the work by Glickman (1995).

Similar to the official FIFA ranking, the Elo model is an earned rating system. The ranking points are updated iteratively after every match. The main idea is that the update rule can be seen as a correction to the teams' rating points subject to actual results and what we expected from the ratings prior to the match.

The update formula for rating points for a team A against an opponent B is as follows:

$$r'_A = r_A + K\left(s_A - p_A\right), \tag{2}$$

where

- r_A and r'_A are the old and updated mean rating (performance) values for team A respectively
- s_A is the actual result of the match from the perspective of team A against its opponent B
- p_A is the expected score of team A against B, derived from the values r_A, r_B prior to the mutual game between A and B
- K, called a K-factor, is a positive constant.

If we follow the convention from chess player ratings, the actual result of the match s_A is mapped to the value of 0, 0.5, or 1 in case team A loses, draws or wins the game, respectively (accordingly for team B). The K-factor governs the magnitude to the changes in ratings after a single game. It can be modelled according to discipline characteristics.

The original Elo model assumes a normal distribution for player A's performance in a match around the mean value of r_A. A simplifying assumption is that the variance is homogeneous among all players, $\sigma_A = \sigma$ for every player A. When two players, A and B, meet in an encounter we are comparing two performance distributions P_A and P_B with $P_A \sim \mathcal{N}(r_A, \sigma^2)$ and $P_B \sim \mathcal{N}(r_B, \sigma^2)$. The probability that player A wins the game is equal to the probability of the event that it draws a higher value from its performance distribution. From the properties of the normal distribution we have that $P_A - P_B \sim \mathcal{N}(r_A - r_B, 2\sigma^2)$ under assumption of independence. In this manner we compute the expected result of the game from the perspective of player A:

$$p_A = \mathbb{P}(P_A > P_B) = \Phi\left(\frac{r_A - r_B}{\sigma\sqrt{2}}\right), \tag{3}$$

where Φ denotes the cumulative distribution function for a standard normal variable $\mathcal{N}(0, 1)$. The draws are disregarded. If the computed value of p_A is around 0.5 then we would expect a draw. Possible extensions to the prediction model (3) to express the probability of a draw are discussed by Rao and Kupper (1967) or Davidson (1969).

In applications, it is common to use a logistic distribution for the players' performance distribution difference and compute p_A as

$$p_A = \frac{1}{1 + e^{-a(r_A - r_B)}}, \tag{4}$$

where a is an appropriate scaling factor. Formula (4) derived from the logistic distribution seems to be more tractable.

The main idea behind the model is that if a team performed better than expected against its opponent B, i.e., $s_A > p_A$ we shall increase its rating accordingly and decrease the rating of the opponent. The rating points in Elo model are self-correcting. Based on current ratings we perform prediction of the future game. The bigger discrepancy between the observed result and our expectations the bigger magnitude of changes to the performance rating estimates for both teams.

An important part of the Elo model is the choice of prior ratings, i.e., initial values for the r_A in the rating period. To reduce the influence of the prior in accurately determining the teams' strength estimates it is necessary to have many games played by every team in the dataset. Otherwise, we cannot see the rating as a reliable estimate of a team's strength.

There are several choices for prior ratings. One possibility is to set ratings for every team to a fixed number, e.g., 1,500. With such choice the ratings would approximately distribute between 1,000 and 2,000 points. However, what is important from the ranking point of view is only the difference in teams' ratings. Another option is to pose a question: what if FIFA would have changed its rating system to that applied in FIFA Women's World Rankings? To answer this, we set prior ratings for the teams to the FIFA ranking points from the 12 July 2006 release. We expect that this prior is better-informed than initialising ratings equally. We compare the predictive powers of these ranking system in our experiments, along with other methods described in this section.

3.2.1 FIFA women's world rankings

FIFA uses a different algorithm to rate and rank women's national teams. In fact, we can recognise an Elo version of the model behind this rating method. The description of the algorithm can be found on the official FIFA website (FIFA.com, 2012a).

In the official FIFA women's ranking after each game the ratings are updated according to the formula (for chosen team A)

$$r'_A = r_A + K \times I \times (s_A - p_A). \tag{5}$$

Next to the basic K-factor, which is set to the constant value of 15, we have an additional multiplier associated with the importance of the game. Analogously to the FIFA men's ranking, its values are tabularised. Table 1 presents possible values that the multiplier I can assume, extracted from the analogous table for women's competition.

Table 1 Match importance multipliers

Competition	*Multiplier*
Friendly match (including small competitions)	1
Confederation-level qualifier	2
FIFA Confederations Cup	3
FIFA World Cup qualifier	3
Confederation-level final competition	3
FIFA World Cup final competition	4

Yet another modification to the original formulation is included in mapping of the actual result to the number s_A. It is assumed that a team losing, for example 1-0, receives the actual score equal to 0.15 rather than 0. The winner is awarded the remainder of the points. The idea is that a team after scoring many goals and losing with a small margin is awarded a small positive value for its actual performance. A complementary argument applies to the winning team. The values of the actual results are presented in Table 2.

Table 2 Actual result of the game in FIFA WOMEN WORLD ranking methodology from the losing team perspective

Goals scored	*Goal difference*						
	0	*1*	*2*	*3*	*4*	*5*	*6+*
0	0.5	0.15	0.08	0.04	0.03	0.02	0.01
1	0.5	0.16	0.089	0.048	0.037	0.026	0.015
2	0.5	0.17	0.098	0.056	0.044	0.032	0.02
3	0.5	0.18	0.107	0.064	0.051	0.038	0.025
4	0.5	0.19	0.116	0.072	0.058	0.044	0.03
5+	0.5	0.20	0.125	0.080	0.065	0.05	0.035

Finally, the expected result of the game is computed with the use of the logistic distribution function. Moreover, a correction is made to incorporate the advantage of the home team. Namely, the probability that team A wins the game is calculated as

$$p_A = \frac{1}{1+10^{-(r_A+100-r_B)}}, \tag{6}$$

where 100 additional points are credited for the host of the game (here by default set to team A).

3.2.2 EloRatings.net

In this subsection we describe another version of the Elo model maintained on the website EloRatings.net (2012). The update formula in the EloRatings.net model is as follows:

$$r'_A = r_A + K \times G \times (s_A - p_A). \tag{7}$$

In this model, s_A is mapped to one of the three possibilities from the set {0, 0.5, 1} and the prediction function is the same as above (6). The K-factor is again determined by the relative importance of the game. One may read possible values it may assume from Table 3. The magnitude of K is modified by the goal difference G. In case the absolute value of the difference of goals scored by both teams is equal to N, K is multiplied by an additional factor G set to

- 1 if $N \leq 1$
- 1.5 if $N = 2$
- $\frac{N+11}{8}$ if $N \geq 3$.

For the EloRatings.net method we managed to obtain historical tables on the website Football-rankings.info (2012) from 9 July 2010. Thus, in addition to the choice of the priors discussed above (uniform and the FIFA ranking release) we can initialise the ratings with retrieved ones.

Table 3 Match importance in the EloRatings.net model.

Competition	*Multiplier*
Friendly match	20
All minor tournaments	30
World Cup and continental qualifiers and major tournaments	40
Continental championship finals and major intercontinental tournaments	50
FIFA World Cup finals	60

The next method we discuss is the winning solution of the first Kaggle competition on chess ratings (Kaggle.com, 2010).

3.3 Elo++ model

Kaggle's competition on chess player ratings was an exciting event with over 250 active participants. In this section, we briefly introduce the winning model, which was proposed by Sismanis (2011).

Before going further we note how the competition's solutions were assessed. For each game in the test set the participants were supposed to provide a single number that expressed the probability of the event that the first player from an ordered pair wins the game. As in the Elo model, draws were disregarded from the analysis. The accuracy of predictions was measured by a monthly aggregated mean squared error. This measure differs from the standard mean squared error only by a minor modification that a player's actual and predicted results are summed in each month. The mean squared error is then calculated on aggregated results and predictions rather than on individual matches. If in every month each player takes part in at most one game, this error measure is exactly equal to the mean squared error.

Next in this subsection, we describe the model, automatically adapting appropriate terminology to football.

3.3.1 Outcome prediction function

From the dataset of the results, we want to estimate rating r_i for every team i. For two teams i and j, that are rated with r_i and r_j, respectively, the probability of team i winning the match is calculated with the logistic cumulative distribution function

$$p_{ij} = \frac{1}{1+e^{-(r_i+h-r_j)}}, \tag{8}$$

where h is a parameter for modelling the advantage of the home team. If the game is played on neutral ground we set $h = 0$. The probability of the opposite team's win is calculated as $p_{ji} = 1 - p_{ij}$:

3.3.2 Time scaling

Each match in the database is assigned a weight that depends on how long ago it was played. Let $t_{\min}$ and $t_{\max}$ denote the minimal and the maximal month number in the data. Then a game between two teams i and j taking place in the t^{th} month is associated with the weight

$$w_{ij} = \left(\frac{1+t-t_{\min}}{1+t_{\max}-t_{\min}} \right)^2. \tag{9}$$

In this way the weighting factor assumes values in the interval (0, 1] for all games in the database and increases monotonically in t.

3.3.3 Neighbours

Another idea in the Elo++ model is that team strength should not deviate much from the ratings of teams that it competes against. It seems to be a reasonable assumption not only in chess or football but in general in sports. Incorporation of the schedule of games in rating computations may be concisely summarised by the saying "you are known by the company you keep".

Let us define N_i as the multi-set of opponents that a chosen team i played against in mutual games, with $| N_i |$ the size of this multi-set (possibly it includes the same team a

few times in case of multiple matches). We would expect that the average rating of rivals of team i should be close to the team i rating itself. The weighted average is computed as

$$a_i = \frac{\sum_{k \in N_i} w_{ik} r_k}{\sum_{k \in N_i} w_{ik}}, \tag{10}$$

where we sum over all the opponents of team i and weight the corresponding ratings with the previously introduced time factor.

3.3.4 Calculation of the ratings

The ratings r are computed by finding the minimum of *the loss function*

$$L(r_1, r_2, \ldots, r_k) = \sum_{\text{games}} w_{ij} \left(s_{ij} - p_{ij}\right)^2 + \lambda \sum_{\text{teams}} \left(r_i - a_i\right)^2, \tag{11}$$

where λ is the weight we assign to the regularisation component. With application of numerical methods we hope to find the minimum of the loss function on the training set. In this setting there are two parameters that need to be optimised: h that stands for the home team advantage and λ which governs the importance of regularisation component.

To focus our attention on the comparison rather than optimisation of individual models we will apply a stochastic gradient descent for the problem of minimisation of the error function (11). This algorithm was suggested by the author in his original description of the method.

In stochastic gradient descent, the database of results is scanned for a fixed number of P iterations. Initially, we set $r_i = 0$ for every team. During each iteration, we scan the entire database of results. We perform the following updates for every game (here between teams i and j) in random order:

$$r_i \leftarrow r_i - \eta \left[w_{ij} \left(s_{ij} - p_{ij}\right) p_{ij} \left(1 - p_{ij}\right) + \lambda \frac{1}{|N_i|} \left(r_i - a_i\right) \right],$$

$$r_j \leftarrow r_j - \eta \left[-w_{ij} \left(s_{ij} - p_{ij}\right) p_{ij} \left(1 - p_{ij}\right) + \lambda \frac{1}{|N_j|} \left(r_j - a_j\right) \right],$$

where η is the learning rate set to

$$\eta = \left(\frac{1 + 0.1 \cdot P}{p + 0.1 \cdot P} \right)^{0.5}, \tag{12}$$

with p being the number of the current iteration. The averages a_i are recomputed only after each iteration.

In this setting, we have two parameters to optimise: λ for the regularisation component and h that measures the impact of home advantage. The choice of the parameters is experimental based on the accuracy of predictions on the validation set.

Elo++ in its primary application in the rating of the chess players performed very well. We hope for similar results when applied to rating football teams.

3.4 Least squares ratings

The next model can be summarised as *least squares ratings*. We sketch the main idea of the model based on the work by Stefani (1977, 1980). A detailed analysis of the least squares ratings may be found in the work by Massey (1997).

The model assumes that the margin of victory of team A against the other team B, denoted as y, is proportional to the difference in both team ratings r_A, r_B:

$$y = r_A - r_B + \varepsilon, \tag{13}$$

where ε is an error in the measurement. The model can be estimated by minimising the sum of squared errors across all the games. In this setting it is not possible to identify the parameters. We shall impose a sum-to-zero constraint or agree on some reference state and set the rating for a chosen team i to a default level, say, $r_i = 0$.

A simple modification to the method may be applied by introducing the home advantage parameter. Because home teams tend to score usually more goals we may capture this by setting

$$y = r_A - r_B + h + \varepsilon. \tag{14}$$

Again, after imposing a proper constraint to identify parameters, the model is estimated by least squares.

The ratings are computed on daily basis in a sliding window approach. On a current day, we include four last years of play to compute the ratings.

3.5 Network-based rating system

The following two methods are derived from graph analysis with the teams represented as nodes and the edges corresponding to the games played between them. The first method we mention originates from social network analysis and can be viewed as a version of Katz (1953) centrality measure of a graph, which was introduced to determine the relative importance of individuals in a network of actors. This was done by counting direct and indirect acquaintances of an actor. If person A knows B one point is awarded. If person B knows person C the indirect connection between A and C is counted with a discount factor $\alpha \in (0, 1)$. More generally, if there is a path of length k in a network of actors between two persons, it is computed as one point discounted with α^{k-1}.

The adaptation of the method to rating sport teams was done by Park and Newman (2005). Let A be a following modification to an adjacency matrix of a graph. The (i, j) entry of the matrix A, a_{ij}, $i, j = 1, 2,\ldots,n$, where n is the number of teams, corresponds to the number of victories of team j over the team i. We assume that a draw corresponds to a half loss and a half win. In analogy to social network ratings we may define *a win* and *a loss score* for the teams. The win score counts the total number of direct and indirect victories of a team, where indirect matches are discounted by an appropriate power of the discount factor α. For a chosen team i, we have that all direct wins of the team can be written as

$$\text{direct wins for team } i = \sum\nolimits_j a_{ji},$$

where the summation is over all indexes $j \in \{1, 2,\ldots,n\}$. The number of indirect wins at distance 2 is given by

$$\text{indirect wins of distance 2 for the team } i = \sum\nolimits_{j,k} a_{kj} a_{ji},$$

and so forth. We can compute the win score w_i for team i weighted with discount factor α^{k-1} for the wins at distance k as

$$\begin{aligned} w_i &= \sum_j a_{ji} + \alpha \sum_{j,k} a_{kj} a_{ji} + \alpha^2 \sum_{j,k,l} a_{lk} a_{kj} a_{ji} + \cdots \\ &= \sum_j \left(1 + \alpha \sum_k a_{kj} + \alpha^2 \sum_{l,k} a_{lk} a_{kj} + \cdots \right) a_{ji} \\ &= \sum_j \left(1 + \alpha w_j \right) a_{ji} = d_i^{in} + \alpha \sum_j \left(A^T \right)_{ij} w_j, \end{aligned} \tag{15}$$

where d_i^{in} is the number of edges pointing to the vertex i, i.e., the number of direct wins of team i and A^T is the transpose of matrix A. From the above, we see that the win score for team i is the sum of the number of teams that i beat in direct encounters and these teams' win score. Analogously we define the loss score l.

The power series (15) converges whenever $\alpha < \lambda_{\max}^{-1}$, where $\lambda_{\max}$ denotes the largest eigenvalue of matrix A. In case $\lambda_{\max} = 0$, then in fact all eigenvalues of matrix A are zero and there are no restrictions on the choice of the parameter α.

Working out equation (15) in matrix notation we arrive at the following formula for the vector of win scores w

$$w = \left(I - \alpha A^T \right)^{-1} d^{in}, \tag{16}$$

and an analogous expression for the vector of loss scores

$$l = (I - \alpha A)^{-1} d^{out}, \tag{17}$$

where d^{out} is a vector of length n in which the i^{th} coordinate stands for the number of edges pointing out of the vertex i, i.e., the number of direct losses by team i. As the ratings set $r = w - l$.

For social network ratings we need to optimise parameter α. To this purpose, we express the parameter α as the percentage of the largest possible value it can assume, i.e., $\lambda_{\max}^{-1}$. We search from 0 to 95% with the step size of 5% and calculate the ratings. Next, we make predictions for the games in the validation set (in a four-year sliding window approach) and measure their accuracy. Our final choice for the parameter α is the one yielding the best accuracy of predictions.

3.6 Markovian ratings

The method described in this subsection is derived from the analysis of an appropriate Markov chain. Application for rating sport teams was considered by several authors

including Callaghan et al. (2007) or Mattingly and Murphy (2010). Perhaps one of the most spectacular applications of Markovian ratings in other domains is the Google's PageRank algorithm for rating web pages (Brin et al., 1999). The approach described below resembles mostly the ideas incorporated in the works Mattingly and Murphy (2010) and Kenner (1993) for estimating the probabilities of transitions.

We construct a simple Markov chain that models the behaviour of a football fan, which is not stable in his feelings. When supporting a particular team, the fan looks for all the opponents that his team has played and either remains with his current team or switches his support in favour of another team. The better a team performs, the bigger chance for the supporter to choose it. With an appealing assumption that the fan is memoryless we can analyse an appropriate Markov chain with the states corresponding to teams. By calculating the probability distribution of which teams the fan is going to support in the long run we obtain the ranking for the teams.

We formalise the discussion as follows. Let i, j be two teams which played a certain number of matches in the past with team i scoring G_i goals in total. The probability that the supporter prefers team j over i, p_{ij}, is proportional to the expression

$$\hat{p}_{ij} = \frac{G_j + 1}{G_i + G_j + 2}, \tag{18}$$

where the corrections made by adding 1 to the nominator and 2 to the denominator aim to prevent division by zero and zero transition probabilities as well. Another possibility is to set

$$\hat{p}_{ij} = \frac{W_j + 1}{W_i + W_j + 2}, \tag{19}$$

where W_i counts the number of victories of team i over team j (we treat draws as a half win and a half loss). The probability p_{ii} of the event that the fan remains with his current team i is proportional to the value

$$\hat{p}_{ii} = \sum_j \left(1 - \hat{p}_{ij}\right). \tag{20}$$

If two teams have not played against each other, then it is not possible to make a transition between them. We plug the computed values to a square matrix $(\hat{p}_{ij})_{i,j=1,2,\ldots,n}$ and normalise its rows by dividing by n to obtain stochastic matrix M with entries $(p_{ij})_{i,j=1;2,\ldots,n}$. Matrix M is a model of the fan's behaviour.

To assure existence of a stationary distribution we may use similar idea as in the PageRank algorithm. Let E be a $n \times n$ matrix with all entries equal to $\frac{1}{n}$ and consider the convex combination of the matrices

$$\tilde{M} = \alpha M + (1 - \alpha)E, \tag{21}$$

where $\alpha \in [0, 1]$. The modified matrix $\tilde{M}$ is also a stochastic matrix and corresponds to an irreducible and aperiodic Markov chain for any $\alpha \in (0, 1]$ (possibly for $\alpha = 0$, if the original Markov chain has both properties itself). For the team ratings we compute stationary distribution π of the chain

$$\pi = \pi \tilde{M}, \tag{22}$$

which gives us team ratings, $r_i = \pi_i$ for the i^{th} coordinate of the vector π corresponding to the i^{th} team.

The optimisation of parameter α is done in analogous manner as in the case of the discount factor in the network-based system above. In our work, we compute different ratings by varying the parameter from 90% to 99% with the step of 1% again with the use of four-year sliding window. Next, we measure the accuracy of derived predictions on the validation set (see Section 4.2), and set α to its optimal value.

3.7 The Power Rank

The Power Rank rating system is the last algorithm under our consideration. The ratings produced by this method for different sports are maintained on the website ThePowerRank.com. Predictions for comparison were provided by Dr. Edward Feng, inventor of the method. The author does not publish the details on how exactly his algorithm works. We only know that it is a combination of the PageRank algorithm with certain techniques applied in statistical physics.

Having described several ranking methods now we describe how their predictive powers were compared.

4 Experimental setup

In this section, we describe the dataset used for our experiments, validation and test sets, how we turned ratings into predictions, and what evaluation measures were used to assess the methods' performance.

4.1 Dataset

The data used for experiments described in this paper is concerned with international football matches, and it was obtained via the official FIFA website. The games were played between 15 July 2006, when the new version of the FIFA ranking was introduced, and 2 May 2012. For each game we have information on the outcome, possible extra time or penalty shoot-out, the date it was played, its location and the type of the game (friendly, World Cup match, etc.). For the purposes of the analysis below we are interested only in the final score of the match with no special treatment of the games ended after extra time or penalty shootout. In a few cases the victory has been awarded to either of the teams, or the game was suspended. We deleted such matches from the dataset.

Using the location of matches, we derived an additional attribute indicating the host of a match. It is a well-known phenomenon in sports, and particularly in football, that the team playing at home has some advantage over the opposition. Its sources and variations have been studied extensively (see, e.g., Pollard, 2008; Seckin and Pollard, 2008; Pollard et al., 2008). A corresponding phenomenon in chess is the advantage due to playing white. The information about the home team is used explicitly by some methods under our consideration – Elo models and the least squares ratings.

4.2 Validation

Calculation ratings is done on a daily basis, i.e., ratings calculated for one day might be different from those a day before or after. The matches taken into account are either those having been played in a four-year period before the day in question (as in the FIFA ranking – we employ the same convention in the least squares ratings, network-based and Markovian ratings), or from the first date in the dataset (Elo methods, The Power Rank). Note that because of our particular dataset, the latter also covers a period not significantly longer than four years.

The performance reported in the next section is calculated based on 979 games played between 1 April 2011 and 2 May 2012.

As some rating methods require parameter tuning, the introduction of a validation set was also necessary. The validation set covers 726 games played between 15 July 2010 and 31 March 2011.

4.3 Prediction function

In this section, we describe how we turn ratings into predictions of the outcome of individual games.

Some of the methods we consider (Elo, Elo++) are self-contained in the sense that they not only estimate the ratings but also provide a way to predict future games. In fact, the prediction function is the core of the Elo model and its main driving force. In case of other methods, we need to transform the ratings produced by them to predictions.

Given two teams A and B with ratings r_A and r_B, respectively, the prediction of the match outcome is given by the model

$$\mathbb{P}\left(s \mid r_A, r_B\right) = \frac{\left(e^{a(r_A - r_B)+h\cdot 1_{\{A \text{ at home}\}}}\right)^s}{1+e^{a(r_A - r_B)+h\cdot 1_{\{A \text{ at home}\}}}}, \tag{23}$$

where the component associated with the parameter h aims to capture the advantage of the home team and $1_{\{A \text{ at home}\}}$ is equal to 1 if team A is the host of the game and 0 otherwise (we assume such an ordering that team B is always a guest of the game). As usual, s stands for the actual result of the game, i.e., $s = 1$ and $s = 0$ corresponds to the win of team A and team B, respectively. To incorporate draws we follow the convention adopted in Glickman (1999). We assume that a single draw yields the same likelihood as a win followed by a loss. The probability of team A's win over B followed by a loss against the same team (or the other way around) is equal to (under assumption of independence between these events)

$$\frac{e^{a(r_A - r_B)+h}}{1+e^{a(r_A - r_B)+h}} \cdot \frac{1}{1+e^{a(r_A - r_B)+h}}.$$

For modelling a single draw we take the square root of the expression above to arrive at

$$\frac{\left(e^{a(r_A - r_B)+h}\right)^{0.5}}{1+e^{a(r_A - r_B)+h}}.$$

Therefore, draws are included in the model by setting $s = 0.5$ in (23).

The likelihood function for the observed match results as the function of the parameters a and h is as follows (provided that outcomes of the matches are independent events):

$$\mathcal{L}(a, h) = \prod_{i^{\text{th}}\ \text{game}} \left(\frac{\left(e^{a\left(r_A^{(i)} - r_B^{(i)}\right) + h \cdot 1_{\{A\ \text{at home}\}}} \right)^{s^{(i)}}}{1 + e^{a\left(r_A^{(i)} - r_B^{(i)}\right) + h \cdot 1_{\{A\ \text{at home}\}}}} \right). \tag{24}$$

The parameters (a, h) are set to their maximum likelihood estimates. To perform predictions on ratings for a match at a given day we estimate the prediction function with the use of games in the period prior to that day. For the first game in the test set, these parameters are obtained by estimation on the games in the validation set. The ratings r_i that we plug to the likelihood function (24) are the most recent (daily updated) estimates derived by a given method. We make one exception to this rule by providing accuracy measurements for monthly FIFA ranking releases that are published on the official website.

4.4 Evaluation measures

When predicting the result of a future match we follow the same convention as in the Elo model. In other words, we attempt to calculate the probability that a chosen team from an ordered pair wins the game.

In our experiments, we consider two evaluation measures, i.e., the binomial deviance and the squared error of the predictions.

Our main accuracy measure is the statistic of the binomial deviance which for prediction p_i for the i^{th} game is equal to

$$-\left(s_i \log_{10} p_i + (1 - s_i) \log_{10} (1 - p_i)\right),$$

where $s_i \in \{0, 0.5, 1\}$ is the actual result of the game from the perspective of a chosen team. The binomial deviance is undefined (infinite) in case of sure predictions $p_i = 0$ or $p_i = 1$ when $y_i = 1$ or $y_i = 0$, respectively. In the computations of this statistic, we round values of p_i lower than 0.01 to 0.01 and for the values of p_i greater 0.99 we set 0.99.

Since we estimate the prediction function by maximum likelihood we consider the binomial deviance as our main quality indicator.

Nevertheless, we also calculate the squared error of the prediction, as it is a common measure used in evaluating predictions:

$$\left(s_i - p_i\right)^2.$$

The accuracy of a given method is measured by averaging prediction errors for individual games.

Similar accuracy measures were used for assessment of proposed solutions in both Kaggle chess players rating competitions.

5 Results and discussion

Table 4 shows the accuracies of various ranking methods and their versions, as measured using binomial deviance and mean squared error. The 90% confidence intervals derived by normal approximation are also reported.

In addition to the methods described in Section 3, we report performance on three additional methods. Two of these are considered baselines aimed at providing context to performance values. One of these two methods always predicts draws, while the other always predicts the home team to win. The third method is an ensemble, i.e., its predictions are formed by combining the predictions individual methods. Combination is done by using the best performing four individual methods (*Elo WWR*, *EloRatings.net*, *least squares* and *The Power Rank*), by averaging individual predictions. The introduction of this method is motivated by the fact that ensembles often work better than individual predictors (see, e.g., Dietterich, 2000).

5.1 Individual methods

As shown in Table 4, all described methods outperform the two baselines, i.e., *all draws* and *home team*, significantly.

Table 4 Accuracy of predictions

Ranking system	*Binomial deviance*		*Mean squared error*	
FIFA ranking *daily*	1.3681	(1.3481, 1.388)	0.1443	(0.1244, 0.1643)
FIFA ranking *release*	1.3705	(1.3504, 1.3905)	0.145	(0.125, 0.1651)
Elo WWR *1500*	1.3698	(1.3498, 1.3898)	0.1447	(0.1246, 0.1647)
Elo WWR *FIFA06*	1.2674	(1.2489, 1.2861)	**0.1268**	(0.1081, 0.1455)
Elo WWR *FIFA06 WDL*	1.2934	(1.2744, 1.3123)	0.1302	(0.1113, 0.1492)
EloRatings.net	**1.2634**	(1.2446, 1.2821)	0.1271	(0.1084, 0.1458)
EloRatings.net *1500*	1.3265	(1.307, 1.346)	0.137	(0.1176, 0.1565)
EloRatings.net *FIFA06*	1.2811	(1.2624, 1.2999)	0.128	(0.1092, 0.1468)
Elo++ (λ, h) = (0.05, 0.4)	1.3062	(1.2871, 1.3254)	0.1336	(0.1144, 0.1527)
Least squares	1.2786	(1.2597, 1.2975)	0.1288	(0.11, 0.1477)
Least squares *home team*	1.2681	(1.2493, 1.2869)	0.1272	(0.1085, 0.146)
Network-based ratings	1.4268	(1.4061, 1.4476)	0.1556	(0.1348, 0.1763)
Markovian ratings *wins*	1.3605	(1.3407, 1.3803)	0.1406	(0.1208, 0.1604)
Markovian ratings *goals*	1.3557	(1.336, 1.3754)	0.1402	(0.1205, 0.1599)
The Power Rank	1.2735	(1.2546, 1.2924)	0.1286	(0.1096, 0.1475)
Ensemble	**1.2358**	(1.2174, 1.2543)	**0.1223**	(0.1038, 0.1407)
All draws	1.5960	-	0.1902	-
Home team	4.1733	-	0.3325	-

Among the single methods, the best accuracy is achieved by two Elo models: the EloRatings.net system is the most accurate with respect to binomial deviance, and the Elo model applied by FIFA in ranking women's teams, when we look at the mean squared

error. The difference in accuracy between the FIFA ranking and two versions of the Elo model, Elo++, the least squares and The Power Rank predictions are significant with respect to binomial deviance, based on the 90% confidence intervals. Slightly better performance is achieved by the two versions of the Markovian ratings. This shows that the FIFA ranking method can be outperformed by several alternative methods.

5.2 Method versions

Looking at the performance of different versions of Elo models, we see that the choice of the prior has a major impact on accuracy. For instance, uniform priors (see '1500' versions) are outperformed by better-informed priors.

Experimenting with the parameters (λ, h) in Elo++ we obtain best results when setting the values to (0.05, 0.4). Despite Elo++ winning the Kaggle competition, even its optimised version does not give best performance for football. However, this might be because of the fact that it uses less information, i.e., it does not incorporate either information on goals scored or match type. The importance of goals scored is shown by our results on the Elo WWR WDL model version in which the actual result is mapped to an appropriate value from the set {0, 0.5, 1} rather than to the values in Table 2.

The significance of the information on margin of victory is also stressed by good performance of the simple least squares model. The least squares method can further be improved by taking into account home team advantage (already in Elo++).

Regarding the Network-based ratings model, we could not achieve good performance even by tuning its parameter (the best performance, reported in Table 4, was achieved using α set to 20% of the bound $\lambda_{\max}^{-1}$). This low performance might be due to the issue of regional grouping of games (confederations), the lack of the time dimension and possible 'loops' in matches (e.g., team A beat team B, B beat C and C beat A).

Concerning Markovian ratings, the optimal values of parameter α for the chain with transitions computed on goals is $\alpha = 0.96$, and for the transitions calculated solely on win/draw/loss information, it is $\alpha = 0.99$. In both of these versions, the same graph structure of the teams is explored, which results in virtually the same performance.

5.3 Method combination

As Table 4 also shows, a simple combination of predictions (*ensemble*) based on several rating methods produces superior performance to any single method described in this paper. However, if one is to create rankings based on such combination, it is not straightforward how to do this, and also, it is often desirable that the ranking algorithm is transparent and easy to understand. Based on this, we recommend that if the goal is to make predictions, several methods need to be combined, and if rankings are to be created, a well performing individual method might be more desirable, so as to balance predictive performance and method transparency.

6 Conclusions and future work

The aim of the paper was to provide an overview of, and investigate the predictive capabilities of different ranking systems for national football teams. The main benchmark

was the FIFA ranking. Our experiments has shown that it is possible to outperform the official ranking procedure by relatively simple algorithms, which is surprising given the high influence of this ranking on football competitions. On the other hand, the FIFA methodology used for ranking women's teams, based on the Elo rating system, is indeed a very competitive rating method. Applying an analogous procedure in ranking men's national teams might be worth taking into consideration.

We see two possible directions of future work on the topic of ranking football teams. First, we may develop better performing ranking methods, which we would base on one of the two discussed Elo rating system. Their performance is high, they are not overly complicated, and perhaps they can be improved further for even better predictive performance.

Second, it may be worth investigating how possible inefficiencies in the FIFA ranking can be exploited by national football associations. We have seen in this paper that the official ranking system does not award points in the most accurate manner. Using this information, and a better model, a team might be able to advance in the current rankings by choosing opponents for friendly games that they are likely to gain the most ranking points against. Hence, we may want to seek an optimal strategy for scheduling friendly matches, or to identify if some teams apply such strategy already.

Acknowledgements

The authors would like to thank Dr. Edward Feng for providing predictions which enabled to include his rating method into comparison.

References

Brin, S., Page, L., Motwami, R. and Winograd, T. (1999) *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report 1999-0120, Computer Science, Stanford University.

Callaghan, T., Mucha, P. and Porter, M. (2007) 'Random walker ranking for NCAA division I-A football', *American Mathematical Monthly*, Vol. 114, No. 9, pp.761–777.

Davidson, R. (1969) *On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments*, Technical Report No. 37, FSU Statistics Report Ml69 ONR.

Dietterich, T. (2000) 'Ensemble methods in machine learning', in Kittler, J. and Roli, F. (Eds.): *Multiple Classifier Systems*, pp.1–15, Springer-Verlag, New York.

EloRatings.net (2012) *The World Football Elo Rating System*, [online] http://www.eloratings.net/system.html (accessed 3 March 2012).

FIFA.com (2012a) *FIFA Women's World Ranking Methodology*, [online] http://www.fifa.com/worldranking/procedureandschedule/womenprocedure/index.html (accessed 11 February 2012).

FIFA.com (2012b) *FIFA/Coca-Cola World Ranking Procedure*, [online] http://www.fifa.com/worldranking/procedureandschedule/menprocedure/index.html (accessed 29 January 2012).

Football-rankings.info (2012) *Elo Ratings Update*, 9 July 2010, [online] http://www.football-rankings.info/2010/07/elo-ratings-update-9-july-2010.html (accessed 17 May 2012).

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

Glickman, M. (1999) 'Parameter estimation in large dynamic paired comparison experiments', *Applied Statistics*, Vol. 48, No. 3, pp.377–394.

Herbrich, R., Minka, T. and Graepel, T. (2007) 'TrueSkill(tm): a Bayesian skill rating system', in Schölkopf, B., Platt, J. and Hoffman, T. (Eds.): *Advances in Neural Information Processing Systems 19*, pp.569–576, MIT Press, Cambridge, MA.

Internationalworkpermits.com (2012) *Football Players Work Permits*, [online] http://www.internationalworkpermits.com/football-players-work-permits.html (accessed 10 August 2012).

Kaggle.com (2010) *Chess Ratings – Elo versus the Rest of the World*, [online] http://www.kaggle.com/c/chess/details/ (accessed 25 January 2012).

Kaggle.com (2011) *Deloitte/Fide Chess Rating Challenge*, [online] http://www.kaggle.com/c/ChessRatings2/details/ (accessed 25 January 2012).

Katz, L. (1953) 'A new status index derived from sociometric analysis', *Psychometrika*, Vol. 18, No. 1, pp.39–43.

Kenner, J. (1993) 'The Perron-Frobenius theorem and the ranking of football teams', *SIAM Review*, Vol. 35, No. 1, pp.80–93.

Leitner, C., Zeileis, A. and Hornik, K. (2010) 'Forecasting sports tournaments by ratings of (prob)abilities: a comparison for the EURO 2008', *International Journal of Forecasting*, Vol. 26, No. 3, pp.471–481.

Luckner, S., Schröder, J. and Slamka, C. (2008) 'On the forecast accuracy of sports prediction markets', in Gimpel, H., Jennings, N.R., Kersten, G.E., Ockenfels, A. and Weinhardt, C. (Eds.): *Negotiation, Auctions and Market Engineering*, Vol. 2, pp.227–234, Springer-Verlag, Berlin Heidelberg.

Massey, K. (1997) *Statistical Models Applied to the Rating of Sports Teams*, Master's thesis, Bluefield College.

Mattingly, R. and Murphy, A. (2010) *A Markov Method for Ranking College Football Conferences*, [online] http://www.mathaware.org/mam/2010/essays/ (accessed 26 March 2012).

McHale, I. and Davies, S. (2007) 'Statistical analysis of the effectiveness of the FIFA world rankings', in Albert, J. and Koning, R. (Eds.): *Statistical Thinking in Sports*, pp.77–90, Chapman & Hall/CRC, Boca Raton, Florida.

Park, J. and Newman, M. (2005) 'A network-based ranking system for US college football', *Journal of Statistical Mechanics: Theory and Experiment*, No. 10, p.P10014.

Pollard, R. (2008) 'Home advantage in football: a current review of an unsolved puzzle', *The Open Sports Sciences Journal*, Vol. 1, No. 1, pp.12–14.

Pollard, R., da Silva, C. and Nisio, C. (2008) 'Home advantage in football in Brazil: differences between teams and the effects of distance traveled', *The Brazilian Journal of Soccer Science*, Vol. 1, No. 1, pp.3–10.

Rao, P. and Kupper, L. (1967) 'Ties in paired-comparison experiments: a generalization of the Bradley-Terry model', *Journal of the American Statistical Association*, Vol. 62, No. 317, pp.194–204.

Seckin, A. and Pollard, R. (2008) 'Home advantage in Turkish professional soccer', *Perceptual and Motor Skills*, Vol. 107, No. 1, pp.51–54.

Sismanis, Y. (2011) *How I Won the 'Chess Ratings: Elo vs the Rest of the World' Competition*, [online] http://blog.kaggle.com/2011/02/08/how-i-did-it-yannis-sismanis-on-winning-the-elo-chess-ratings-competition/Kaggle.com blog (accessed 25 January 2012).

Stefani, R. (1977) 'Football and basketball predictions using least squares', *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 7, No. 2, pp.117–121.

Stefani, R. (1980) 'Improved least squares football, basketball, and soccer predictions', *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 10, No. 2, pp.116–123.

Suzuki, K. and Ohmori, K. (2008) 'Effectiveness of FIFA/Coca-Cola world ranking in predicting the results of FIFA World Cup finals', *Football Science*, Vol. 5, pp.18–25.