VU University Amsterdam
Faculty of Sciences

University of Warsaw
Faculty of Mathematics, Computer Science and Mechanics

Joint Master of Science Programme

Jan Lasek

Student no. 266756 (UW), 2029944 (VU)

Football team rankings

Master's thesis
in MATHEMATICS

Supervisors:

Dr. Sandjai Bhulai
Dr. Zoltán Szlávik

March 2012

Supervisor's statement

Hereby I confirm that the present thesis was prepared under my supervision and that it fulfils the requirements for the degree of Master of Mathematics.

Date Supervisor's signature

Author's statement

Hereby I declare that the present thesis was prepared by me and none of its contents was obtained by means that are against the law.

I also declare that the present thesis is a part of the Joint Master of Science Programme of the University of Warsaw and the VU University Amsterdam. The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

We present different models for rating football teams and compare them with the official FIFA ranking for national teams. We define an evaluation measure that is based on predictive capabilities of a ranking system. This allows us to say which ranking method is better. Finally, we address the issue of scheduling friendly games so that a team can make largest progress in the table by appropriate choice of opponent.

Keywords

Rankings, ratings, team strength estimation, football.

Thesis domain (Socrates-Erasmus subject area codes)

11.2 Statistics

Subject classification

62-07. Statistics: data analysis.

Acknowledgements

I would especially like to thank my supervisors Dr. Sandjai Bhulai and Dr. Zoltán Szlávik. Their support, comments, suggestions and general help have been invaluable to me.

I would also like to thank Dr. Edward Feng for his precious comments and predictions from his exciting website ThePowerRank.com.

Contents

Chapter 1

Introduction

1.1. Problem statement

The topic of ranking concerns a broad list of fields with a variety of different applications. It is present in our everyday life although sometimes we are not aware of it. For instance, think of your to do list with set priorities for different tasks - after ordering the issues with respect to their relative importance you obtain a ranking of what you need to do. Of course, this example is quite elementary but it shows that ranking methods are omnipresent. If we combine the ranking methodologies with football, probably the most popular sport in the world, we obtain a fascinating mixture for studying and investigation.

There are many different kinds of football competitions at diverse levels. We focus on football at the international stage. To provide a little background we shall briefly recall the rules of the game. Two teams, each consisting of 11 players, are competing against each other for two 45 minute halves. The objective is to defend firmly and score more goals than the opponent. The game can end in a tie. Depending on the level of competition, sometimes extra time is needed to decide a winner. It consists of two additional halves, 15 minutes each. If the game is still undecided after 120 minutes of play, a penalty shootout contest is played, where teams compete until deciding the winner. In general, these rules are applicable worldwide now[1]. Historically, they were changing constantly. For example, in the past if the game was drawn after regular time a coin flip decided the winner. One of the most famous examples of such a situation was the semifinal match of the European Championships 1968 - Italy against Soviet Union. After 120 minutes of play and a goalless draw the game was decided by a coin toss - Italy "won" the match and ultimately became European Champion that year[2]. However, during the years under consideration, the rules are not changing.

Currently there are 208 national football associations united under FIFA - *Fédération Internationale de Football Association* - the international governing body of football. Besides the organization of the World Cup - the most prestigious international football tournament - each month FIFA releases the ranking of teams around the world. By taking part in different tournaments or playing friendly matches, the teams earn points in the official FIFA World Ranking according to a specified algorithm. It may seem that it is not important to rank teams since the World Cup decides which team is the best. However, only a limited number of teams take part in the greatest tournaments. Moreover, the final standings of major competitions are

[1] For current rules see `www.fifa.com/mm/document/affederation/generic/81/42/36/lawsofthegame_2011_12_en.pdf`, accessed 18 April 2012

[2] More examples under `www.guardian.co.uk/football/2002/aug/08/theknowledge.sport`, accessed 10 April 2012

sometimes surprising, which accounts for the beauty of the sport and makes it so interesting. The ranking provides information on the overall strength for each team.

Probably the most important application of the official FIFA ranking is the fact that the ranks are used for scheduling competitions. Usually each major tournament in international football is preceded by qualification rounds. The teams are divided into groups. To balance the level of teams in the groups, the teams are seeded in different pots and next a draw take place. Seeding is done according to teams' position in the ranking. Therefore it is an important task to have a rating procedure that is relevant to the true teams' abilities. There is a need for an algorithm that produces accurate team rankings. Its influence on sport is crucial.

Last but not least, according to the official FIFA ranks, work permits are granted by UK government for non-European players. In this way the ranking influences players' careers. We also conclude that the authorities believe that it is as accurate as possible. In the past, there were cases when a player was rejected to play in English Premiership, because his country was ranked outside the top 70 of the FIFA ranking[3]

The official ranking is often subject to criticism in the media. A constructive judgment of the ranking was done by McHale and Davies [21]. By building and analyzing several statistical models for predicting match results, the authors conclude that the ranking does not use the information on past results efficiently. A suggestion was made to look for another ranking system or improving the current one.

The main focus of our project is to *provide a rating scheme that under a specified evaluation measure performs better than the official World Ranking provided by FIFA*. The *evaluation measure* we define is based on predictive capabilities of a ranking.

On the other hand, a team might try to schedule its friendly matches to maximize its position in the official ranking. The team therefore seeks for *an optimal strategy of choosing its opponents to play against*. Next to the problem of creating a rating procedure with the highest accuracy we will focus on the topic of *scheduling games by a chosen team in order to maximize its expected position in the ranking*. With application of the optimal strategy, the team can climb up the ranking. This may allow to gain a seeded rank during the draw and avoid the strongest opponents.

1.2. Approach to solve the task

Because of the broad spectrum of the domain there is a number of ways to tackle the problem. The solutions we are going to present are mainly inspired by associated problems in other sport disciplines. Nevertheless, some of the methods we are going to discuss are derived from different domains with no obvious connection to sport.

Usually methods for rating provide a single value representing a team's strength. Some of the models give s more detailed characterization of the skills, for example, standard deviation of the rating value or decomposition of the overall strength into offensive and defensive capabilities. In any case, from the rating vector obtained by the specific algorithm we shall always be able to extract a single real-valued number that exhibits the strength of a team. The linear ordering of the rating points results in the ranking that can be represented as a table with the best team appearing at the top.

We shall emphasize an important methodology issue applied in the thesis that imposes some restrictions on the choice of a model. Since we want to make direct comparison to the FIFA ranking, for each method we want to obtain the ratings that correspond to the

[3] http://www.workpermit.com/news/2005_08_30/uk/sports_stars.htm, accessed 2 June 2012.

points accumulated by the teams in the official FIFA ranking. Now, in order to be able to say what a better ranking system means we need to define an evaluation measure that will serve for assessment of the model accuracy. To this end, we need a method of confronting the rating points with actual results of the matches. This will be done by making predictions with the use of ratings. In this way we will be able to say something about the accuracy of a particular algorithm. The diagram in Figure (1.1) presents the scheme of methodology employed throughout the thesis.

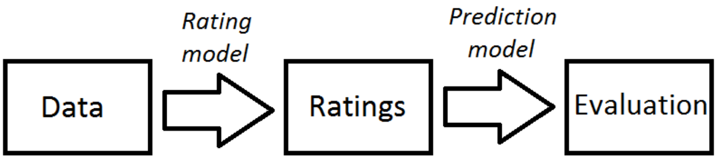

Figure 1.1: The modeling scheme in the thesis.

It should be stressed that we will make inference about teams' skills based solely on the results of the matches regarded by FIFA as the official ones (available on the website `www.fifa.com`). If we were to focus only on predicting the outcome of a particular game we would like to go with the analysis more deeply - perhaps one would like to consider the lineups (presence/absence of key players), recent shape of the team, or even GDP of a country. There is a number of different attributes that one can derive in order to build a predictive model for future games. In this case, also different methods can be applied, which do not fit in the framework presented in the diagram above. This observation is clear. Nevertheless it is important to emphasize it in the general scope of the work.

1.3. Structure of the thesis

To complete the introductory chapter, we outline the structure of the thesis. In the second chapter we give an overview of the possible approaches for rating. This is in no way an exhaustive treatment of the topic. Nevertheless, according to the author, it should provide a small but representative sample of possible methods. Next, in Chapter 3 we describe the data that we are going to work with. In Chapter 4 we turn to the implementation and evaluation of the methods. In this part we also formulate the exact definition of a better ranking, based on its predictive power. Ranking tables produced by individual methods are provided in the Appendix. In Chapter 5 we discuss the issue of scheduling friendlies. Finally, Chapter 6 concludes the thesis along with suggestions for future work.

All analyses and computations are performed with the use of R, an environment for statistical computing.

Chapter 2

Overview of rating systems

The variety of applications of ranking methods is broad. Apart from sport, which is our primary interest, we consider rating models from diverse fields. In general, we shall present a few typical approaches that can be categorized under different headings. First, we present earned ranking methods which from the practical point of view are most widely applied in rating sport teams or individual players. The official FIFA ranking is that kind of model. Similar methods are applied in ATP rankings for tennis players around the world or in rating on-line gamers. Accurate rankings can be used for pairing players for even match-ups. In addition, thanks to the ratings, individual players can monitor their progress. In this way, rankings also encourage competition.

The next type of methods can be considered as statistical approaches or error minimization methods. The methodology here is quite different from the earned ranking systems. The ratings are not updated iteratively after each match (however, it is obviously possible). Usually those methods look at the dataset as a whole and try to infer about the strength of the teams by minimizing a predefined cost function, that links the results with teams' rating points.

The last group of models we discuss is derived from graph analysis. The nodes in the graph represent the set of teams and the edges reflect mutual encounters between the teams. One of the methods discussed here is analogous to the PageRank algorithm which is the core of Google's search engine. Another approach is inspired by social network analysis. The graph structure of the society of actors was investigated to determine the relative importance of the people in the community. This is connected with the Katz centrality measure [15]. Yet another application of the discussed method was applied in an investigation of the dominance hierarchy among animals. Within a population of American bison a relation between clashes of the bison and their breeding success were examined. It turned out that there is high correlation between these two phenomena - the most successful bison, that took part in aggressive interactions, establishes a hierarchy in the herd. It was quantified by the ranking method. The higher ranked bison turned out to be more successful in mating [23].

A valuable source of the rating models is the documentation on two **Kaggle.com** competitions on chess ratings [13], [14]. **Kaggle.com** is a website hosting data mining contests open to every audience. Although chess and football seem to be quite distinct disciplines - certainly chess is less spectacular than football - there are a few similarities between them. Not meaningless, in both sports a tie is possible. Also, a player with white chess pieces has a little advantage its opponent by making the initial move. The corresponding phenomenon in football is *home team advantage*. Because of rich resources and the tradition of ratings chess players, it is helpful to consider these models in the context of football [10].

Perhaps it is needles to ask a football fan which team is the best - most probably the person will indicate his (or her!) favorite team. Although it is quite a loose remark, undoubtedly there is a great dose of subjectivity when it comes to assessment of teams by people. The algorithms implemented with the use of a computer try to overcome this subjective view. It is reasonable to assume that the computer will not favor any team. The ratings provided by an algorithm should be objective. However, there is still subjectivity in the choice of the model which unavoidably has its own flaws. Possible disadvantages induce that still some of the methods will rate the teams in a somehow subjective manner. It is a consequence of a variety of approaches and differences between them. In any case, we should strive to find the most accurate model.

We turn to the description of the methods. We shall start with our main benchmark which is the official FIFA ranking [9]. The consecutive parts concerning distinct methods include different structure of the paper as a consequence of their diversity. We conclude this chapter with a summary and a table presenting the features incorporated by individual models. The effort is focused on indicating possible advantages and disadvantages of a particular approach.

2.1. The official FIFA ranking

The official FIFA ranking can be seen as an earned rating system. It is released on approximately monthly basis. The current procedure of awarding points is applicable since 12 July 2006, following World Cup Finals in Germany. The previous algorithm was of similar kind with a few differences [32].

The results of the matches from the last four years are used to calculate ranking points. In each year, the average of the accumulated points by a team is computed. Next, these four averages are summed up with weights that depend on time.

During the period of consecutive issues of the ranking teams compete and gain points. The ranking is recalculated accordingly. Let us first describe how points per single game are awarded. This depends on several factors:

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

The number of points P gained by the team is computed according to the formula

$$P = M \times I \times T \times C. \tag{2.1}$$

Let us explain how the values M, I, T, C are obtained.

The points for the result of the game

Teams gain 3 points for a victory, 1 point for a draw and 0 points for a defeat. In a penalty shoot-out, the winning team gains 2 points and the losing team gains 1 point.

The importance of the game

Depending on the type of the game the factor I can assume four values as presented in Table (2.1). In this way, we distinguish several types of games and assume that particular kinds of matches (e.g., World Cup games) are more important indicators for a team's strength. It is

Competition	**Multiplier**
Friendly match (including small competitions)	1.0
FIFA World Cup qualifier or confederation-level qualifier	2.5
Confederation-level final competition or FIFA Confederations Cup:	3.0
FIFA World Cup final competition:	4.0

Table 2.1: Match importance multiplier in the FIFA ranking calculation.

reasonable to assume that the results of friendly games should be considered as less important. Often in those games the managers experiment with the team. In particular, usually those matches do not involve the strongest "eleven". However, the way to quantify game importance is not clear. One may argue on the particular choice of the values for multipliers. It is based on the subjective view on match importance.

Strength of opposition

The strength of the opposing team is calculated as 200 minus the ranking position of that team (from the latest release of the ranking). As an exception, the team at the top of the ranking is assigned a maximum value of 200 and the teams ranked at 150th or lower place are assigned the minimum value of 50.

Strength of confederation

There are six confederations recognized by FIFA that oversee the game in different parts of the world. Each confederation is assigned a number in the range $[0.85, 1]$ that indicates its overall strength. These values are calculated based on the result of the last three World Cup tournaments[1].

For both teams the value of C in Formula (2.1) is computed as the mean value of the strength of confederations that the two competing teams belong to. Table (2.2) presents these values for all six confederations. For more details about confederations and their regional boundaries we refer to Chapter 3.

Confederation	**Continent**	**After WC 2006**	**After WC 2010**
AFC	Asia and Australia	0.85	0.86
CAF	Africa	0.85	0.86
CONCACAF	North and Central America	0.85	0.88
CONMEBOL	South America	0.98	1
OFC	Oceania	0.85	0.85
UEFA	Europe	1	1

Table 2.2: Strength of confederations used in ranking points calculation. Following World Cup (WC) Finals in South Africa the values were updated.

An example of calculation of the points is presented in Table (2.3).

[1] For more detailed information how these values are derived see `www.fifa.com/mm/document/fifafacts/r&a-wr/52/00/97/fs-590_10e_wrpoints.pdf` (accessed 11 February 2012)

	Team	Result	Match	Opposition	Confederation	Ranking
		points	importance	strength	strength	points gain
🇳🇱	Netherlands	3	4	200	1	**2400**
🇧🇷	Brazil	0		196		**0**

Table 2.3: Calculation of the ranking points after Netherlands 2:1 victory over Brazil in the FIFA World Cup Final competition. On the day the match was played the Netherlands were placed 4th and Brazil 1st.

Time factor

The ranking scheme also incorporates a time factor. It is used for weighting yearly averages. Only the results from the matches played within last four years are used for points calculation. Depending on how many years a game was played, the averages are discounted by the factor as shown in the following table. Only the results from the last year of play receive full weight.

Date of match	**Multiplier**
past 12 months	1.0
12-24 months ago	0.5
24-36 months ago	0.3
36-48 months ago	0.2

Frequency of matches played

If the average number of points calculated for each of 4 years in the FIFA ranking is based on a smaller number than 5 games, the it is additionally discounted according to the table below.

Number of games used	**Discount**
to compute yearly mean	**factor**
4	0.8
3	0.6
2	0.4
1	0.2

In this way, the teams that compete rarely are penalized. To obtain the full average of points in a particular year, a team needs to play at least 5 international games in that year. To put it in an alternative equivalent way, in case a team had played less than five games, we add fake loses to its tally so that the total number of games in each year is 5. Next we compute yearly averages and add them up with appropriate time weights.

Summing up, in the official FIFA ranking teams accumulate points by playing matches. The number of points awarded after single match depends on the outcome of the game, its importance, opposition and confederation strength. Most recent matches have counts more towards a team ranking position.

2.2. 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 famous systems for rating skills in two player games. Due to its general merits in players' ranking, it is the first system we describe after the official FIFA ranking. It can serve as a basis for other rating systems and has several generalizations. Primarily it was used for chess player ratings. We will discuss it in terms of "teams" rather than "players" for adaptation to football.

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

Let us formalize the discussion. The assumption of the model is that a team's performance in a match is a normally distributed random variable. It is characterized by the mean team's performance r and standard deviation $\sigma > 0$. The variance of a team's rating measures uncertainty about its mean performance. A simplifying assumption of the model is that the standard deviation is homogeneous among all teams. From the data we aim to estimate the mean performance parameter r for every team. In the basic formulation, there is one more parameter in the model $K > 0$, called a *K-factor* that governs the magnitude of changes to estimates r when evidence is obtained (results of the game). It also provides an upper bound for the maximal adjustment for a team's rating per game (in absolute value). The K-factor may be modeled according to the specific discipline's characteristics.
The formula updating rating points for a team A against an opponent B is as follows:

$$r'_A = r_A + K(s_A - p_A), \tag{2.2}$$

where

- r_A and r'_A are old and updated mean rating (performance) values respectively for the team A
- s_A is the actual match result from the perspective of the player 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.

We describe below how the values for the variables in Equation (2.2) are derived.

Actual and expected score of the match

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 the team B). The idea is that p_A provides an estimate for the fraction of games that the team A is likely to win against the team B when a large number of games are played. For both teams, their performance in a particular game is modeled as a random variable. When they meet in a match, we can imagine that the teams draw a number from their performance distribution and compare them. The team which draws a higher number wins the game. Note that we do not estimate the probability of a tied game but only a binary win/loss outcome. If the expected score of the game would be around 0.5 then we say that team A is believed to win on average half of the encounters against its opponent B (likewise for the second team). We may interpret this value as the fact that the teams are most likely to draw a game. Therefore draws are not explicitly modeled but can be viewed as halfway between win and loss.

In the original Elo model we assume a normal distribution with equal variances across all teams for their performance distribution, denoted as P_A (for team A). If two teams A and B meet in a game, we are comparing two performance distributions $P_A \sim \mathcal{N}(r_A, \sigma^2)$,

$P_B \sim \mathcal{N}(r_B, \sigma^2)$ which are assumed to be independent. The probability that team A wins the game is equal to the probability of the event that it draws a higher value from its performance distribution, $\{\omega : P_A(\omega) > P_B(\omega)\}$. From the properties of the normal distribution we have that $P_A - P_B \sim \mathcal{N}(r_A - r_B, 2\sigma^2)$. Now we may compute p_A as

$$p_A = \mathbb{P}(P_A > P_B) = \mathbb{P}(P_B - P_A < 0) = \mathbb{P}\left(\frac{P_B - P_A - (r_B - r_A)}{\sigma\sqrt{2}} < \frac{r_A - r_B}{\sigma\sqrt{2}}\right) = \Phi\left(\frac{r_A - r_B}{\sigma\sqrt{2}}\right),$$

where Φ denotes the cumulative distribution function for a standard normal variable $\mathcal{N}(0,1)$. Analogously we obtain

$$p_B = \mathbb{P}(P_B > P_A) = 1 - \mathbb{P}(P_A > P_B) = 1 - \Phi\left(\frac{r_A - r_B}{\sigma\sqrt{2}}\right).$$

Now we are ready to provide an interpretation to Formula (2.2) and explain the main idea of the model. Since $K > 0$, the formula implies that if a team performed better than expected against its opponent B, i.e., $s_A > p_A$ we shall increase its rating accordingly. On the other hand, if the observed result of the game is worse that we would expect for team A, $s_A < p_A$, then this team has its rating decreased as we conclude that perhaps we overestimated the value of r_A. The bigger discrepancy between the observed result and our expectations the bigger magnitude of changes to the performance rating estimates for both teams. When a team with expected score close to zero actually won the game ($s_A = 1$ and $p_A \approx 0$) it receives a substantial update to its mean performance r_A. If the outcome of the game is as expected from our model ($s_A \approx p_A$) the changes to the estimates r_A are minor as we believe that those are accurate estimates of the performance ratings. K-factor also provides a bound on the maximal change in rating points $r_A - r'A$ since for $p_A, s_A \in [0,1]$ we have $|r'_A - r_A| = K|s_A - p_A| < K$.

Often in applications the probability of team A winning against team B as a function of the rating difference $r_A - r_B$ is computed with the use of the logistic cumulative distribution function

$$p_A = \frac{1}{1 + e^{-a(r_A - r_B)}}, \tag{2.3}$$

where $a > 0$ is the scaling factor. It has been argued if the logistic distribution provides a better fit to the data than the normal distribution. From a practical point of view there is no difference in the choice between these distributions [31]. The logistic distribution has "thicker tails" than the normal distribution. Thus under this distribution rare events are slightly more likely, but virtually there is no difference between them. Perhaps Formula (2.3) is simpler to work with than the corresponding expression for the normal distribution.

It is also worth noting that the logistic distribution is related to the Bradley-Terry[2] model, which is one of the first in the field of *pairwise comparison* in statistics. The model assumes that in the set of n objects, the i-th item is preferred with probability $\pi_i > 0$, $i = 1, 2, ..., n$ with $\sum_{i=1}^{n} \pi_i = 1$. When two items i, j, are being compared, the probability $\mathbb{P}(i)$ that the object i is preferred over the object j is expressed as

$$\mathbb{P}(i) = \frac{\pi_i}{\pi_i + \pi_j}$$

and analogously for item j. If we substitute $\pi_i = e^{r_i}$, $r_i \in \mathbb{R}$, then we can write

$$\mathbb{P}(i) = \frac{e^{r_i}}{e^{r_i} + e^{r_j}} = \frac{1}{1 + e^{r_j - r_i}},$$

[2] However, the model was probably first formulated by Zermelo in 1928. For a brief history of the pairwise comparison we refer to the paper *"Introductory note to 1928"* by Glickman.

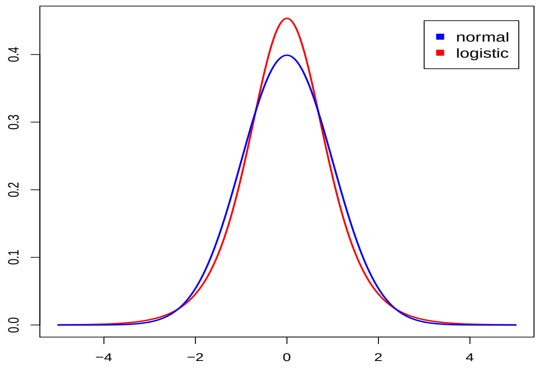

Figure 2.1: The standard normal and the logistic probability density functions with equal, variances centered at 0.

that is a more familiar expression for the cumulative distribution function of the logistic distribution. In this setting we interpret "preference" as the probability that a team wins the game and the values π_i (or their non-linear transformations r_i) as the teams' ratings.

Choice of the K-factor and prior ratings

The parameter K in Equation (2.2) governs the magnitude of changes in the rating values after the result of the game is observed. If it is chosen to be relatively small, then the players' rating estimates may converge slowly to their actual skills. On the other side, if we choose it too big then the ratings may be too sensitive for the recent results. The choice of a constant K-factor is simple, but it can also be dependent on the discipline's specifics. We can also consider it to be a function of the team ratings. For example, in chess it is common to decrease K as the player's rating increases - it follows from the observation that the highest rated played have usually consistent results while younger competitors or players that have never been rated are assigned bigger value of K. It allows their rating to adjust more quickly to their actual skills.

Another question is the choice of the prior rating for a team's performance. Several approaches may be applied. The easiest one is perhaps to assign every team a prior mean performance of, say, 1500. The K-factor also may be seen as the weight we assign to results rather that the prior ratings. The bigger value of K, the more important are the results towards determining the player's rating.

Current implementations

We discuss below two Elo-type models for rating football teams around the world. The first of them, is applied to rank national women football teams in the official FIFA Women's World ranking [8]. The second, called The World Football Elo Rating System, is implemented and maintained on the website EloRatings.net [7].

The FIFA Women's World Ranking

In the official FIFA Women's ranking[3] after each game the ratings are updated according to the formula (for chosen team A)

$$r_A' = r_A + K \times I \times (s_A - p_A). \tag{2.4}$$

Let us explain the ingredients K, I of this formula in steps.

K-factor and match importance

The value K is called the basis factor and it is set to 15. It is multiplied by the match importance factor I. An analogous table as in the official FIFA ranking is presented below. We extract only the type of the games in our interest in men's football[4]:

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

Actual result of the game

In the FIFA Women's World ranking methodology, the actual result of the game is modified with respect to the goal difference. Not only is the fact of win or loss taken into account but also the margin of victory. Table (2.4) shows the points awarded as actual result of the game from the perspective of the losing team. The winning team receives the remainder of this fraction (1 - *points for losing team*). The intuition behind the table is that the team losing after scoring a higher number of goals is awarded a higher value as their actual result. Also the margin of defeat is taken into account. In case of a narrow loss, a team is believed to fight for victory. Therefore its performance is assessed with a small positive value.

	Goal difference						
Goals scored	**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

Table 2.4: Actual result of the game in FIFA Women World ranking methodology from the losing team perspective.

[3] www.fifa.com/worldranking/procedureandschedule/womenprocedure/index.html

[4] Confederations Cup is not played in women's football. We assign the weight from the corresponding importance table in men's World Ranking calculation

Expected result of the game

Predictions of the outcome of games p_A in Formula (2.4) are made with the logistic distribution function corrected for *the advantage of the home team*:

$$p_A = \frac{1}{1 + 10^{-(r_A + 100 - r_B)}} = \frac{1}{1 + e^{-a(r_A + 100 - r_B)}}, \tag{2.5}$$

with $a = \frac{\log_e 10}{400}$. It is observed in sport that the team playing at home has an advantage over its opponent (we discuss it in more detail in Chapter 3). In the Elo Women World ranking model a team playing at home automatically receives additional 100 rating points to its overall strength r. The scaling factor a governs sensitivity of outcome prediction with respect to the rating difference between the teams. Here it is chosen such that in case of equal ratings, $r_A = r_B$, the probability of team A winning the game (that is assumed to play at home) is equal to 0.64. There are two concerns that arise here - the additive impact of home advantage towards the result of the game and its magnitude. The first one is just a modeling assumption. However, there is no clear motivation for the particular choice of home team advantage correction. A possible explanation is that on average, home teams gain around 64% of all possible points (as indicated by Table (3) in Chapter 3). □

Eloratings.net

The update formula in the Elo ratings.net is the following:

$$r_A{}' = r_A + K \times G \times (s_A - p_A). \tag{2.6}$$

We explain the definitions for K and G in below.

K-factor corrected for goal difference

In the model the K-factor is again determined by the relative importance of the game. One may read possible values it may assume from the table:

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 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, the K-factor is multiplied by G equal to

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

Actual and expected result of the game

Analogously to the original formulation of the Elo model the actual outcome of the game is mapped to 0, 0.5 or 1 in case of a team's loss, draw or win. The outcome prediction function is the same as in the women's ranking above (2.5). □

Extensions to the Elo model

The Elo model assumes that the standard deviation of the performance distribution among all teams is homogenous. There are models for rating players that relax this assumption including TrueSkill [12] or the Glicko rating system [11]. Both of the approaches assume that if a player competes more often then we become more certain about the estimates of his skill. This is incorporated into the model as variance reduction. In the Glicko system the variance also depends on time. When a player has a stagnation period and does not participate in the competition, then the variance of his strength estimate increases accordingly since we become less and less certain about this value. Yet another alternative for the Elo model is the Chessmetrics rating system, which additionally incorporates time depreciation of the results [3].

In the next section we discuss a model that aims to outperform the Elo approach.

2.3. Elo++

The first **Kaggle.com** competition on chess player ratings, under the name "Chess ratings - Elo versus the Rest of the World", was an exciting event with 257 active participants. In this section we present the model that won the competition. The winning solution was proposed by Yannis Sismanis. His approach, called Elo++, outperformed any other model in predicting future outcomes of chess games.

Before going further we note how proposed solutions in the competition were assessed. The accuracy of the models was measured by monthly aggregated mean squared errors. This measure differs from the standard mean squared error only by a minor modification that a player's actual and predicted results are summed (aggregated) 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 only one game, this error measure is exactly equal to the mean squared error.

A natural idea that comes up when the model is evaluated with the use of the mean squared error is to apply a numerical method to find rating values that minimize a predefined error measure. However, when done in this way, there is a threat of possible overfitting. The Elo++ approach addresses this problem in an elegant way by introducing *a regularization* component to *the cost function* that also aims for better generalization and improved predictions of future games.

We proceed to the description of the model. We automatically adapt appropriate terminology to football. As above, we discuss the model in terms of "teams" rather than "players" unlike in the original description of the method [28].

The formulation of the Elo++ approach has a few resemblances to the original Elo model. The main modeling assumption is again that each team's capabilities can be described by a single real number r, but no assumption is made towards the distribution of a team's performance. There are also similarities between the outcome prediction function and the update equations in both models.

Outcome prediction function

For two teams i and j, that are rated with r_i and r_j respectively, the probability of team i winning the encounter is calculated with the logistic cumulative distribution function

$$p_{ij} = \frac{1}{1 + e^{-(r_i - r_j)}}. \tag{2.7}$$

If we have an additional information about which team plays at home, say team i, we may assume a slight advantage in favor of that team. An analogous idea is applied to the two Elo-type rating systems discussed earlier. The outcome prediction function adjusted for home team advantage is of the form

$$p_{ij} = \frac{1}{1 + e^{-(r_i + h - r_j)}}. \tag{2.8}$$

For the victory of the opposite team we set

$$p_{ji} = 1 - p_i = \frac{e^{-(r_i + h - r_j)}}{1 + e^{-(r_i + h - r_j)}}.$$

Draws are not explicitly modeled, but again they correspond to the predictions p_{ij} around the value of 0.5.

Time scaling

Elo++ incorporates time in the ratings computation. In general, a team's performance in time can exhibit substantial inconsistencies. It is decided by many factors like the team squad, new manager and also recent shape of a team. To estimate the current strength of teams more accurately, the most recent matches are regarded as more important. This is incorporated by assigning each game a weight, which is a monotonically increasing function of time (towards most recent results).

For each game in the data we assign a weight depending on the time the game takes place. To this end, time needs to be discretized. We may choose arbitrary unita for time discretization e.g., day, month or year. It is reasonable to believe that a team's shape remains constant over a month of time.

The Elo++ model weights the games on monthly basis (somehow there was no thinner possibility as the data provided in the competition contained only information up to each month). Let t_{min} and t_{max} denote the minimal and the maximal month index number in the data. Then a game between two teams i and j taking place in t-th month is associated with the weight

$$w_{ij} = \left(\frac{1 + t - t_{min}}{1 + t_{max} - t_{min}} \right)^{\alpha}, \tag{2.9}$$

where $\alpha > 0$ is a parameter ($\alpha = 2$ in the original formulation of the model). In this way the weighting factor assumes values in the interval $(0, 1]$ for all games in the database and increases monotonically in t.

Neighbors

The regularization component in the Elo++ model aims to prevent from overfitting. The main idea is that team strength should not deviate by 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 sports general. The teams participating in the tournaments are usually selected by qualification rounds that should emerge a group of teams being at a similar level. It is also natural to choose friendly match opponents that are comparable in terms of their strength. Incorporation of the schedule of games in rating computations may be concisely summarized by saying that "you are known by the company you keep".

Let us define N_i as the mulitiset of opponents that a chosen team i played against in mutual games, with $|N_i|$ the size of this multiset (possibly it includes the same team a few times in case of multiple matches between the teams). We would expect that the average rating of rivals of team i should be close to the team i rating itself. Let us define the weighted average as

$$a_i = \frac{\sum_{k \in N_i} w_{ik} r_k}{\sum_{k \in N_i} w_{ik}}, \tag{2.10}$$

where we sum over all the opponents of team i and weight the corresponding ratings with the time factor depending on when the game was played. We assign a bigger weight to the ratings of teams that team i recently played against. The restriction that r_i and a_i should be close is the major component in the Elo++ approach that allowed this model to win the **Kaggle.com**'s competition.

Calculation of the ratings

Recall that for a game between teams i and j, with the former playing at home we estimate the probability of team i winning the game with the use of the logistic function

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

which aims to match as close as possible to the observed results s_i of the games. The ratings r are computed by finding the minimum of *the loss function*

$$L(r_1, r_2, ..., r_k) = \sum_{games} w_{ij}(s_{ij} - p_{ij})^2 + \lambda \sum_{teams} (r_i - a_i)^2, \tag{2.11}$$

where λ is the weight we assign to the regularization component.

The technique of a stochastic gradient descent was used to minimize the loss function (2.11). The main idea of the algorithm is to use a noisy estimate of the gradient of the loss function by computing it with respect to a single match in the database [29].

In the stochastic gradient descend algorithm the database of results is scanned for a fixed number of P iterations. In each iteration, we (hopefully) make a step towards values of r's that minimize the loss function. The magnitude of the step is governed by a learning rate parameter η. This parameter in its simplest setting may be set to a constant value. We may also decrease it monotonically depending on the number of current iteration. The choice of the learning rate should be motivated by monitoring the convergence that the algorithm shall hopefully exhibit.

In each iteration, the entire database of results is scanned. We perform following updates for every game (here between teams i and j) in the database, in random order:

$$r_i \leftarrow r_i - \eta \left[w_{ij}(s_{ij} - p_{ij}) p_{ij}(1 - p_{ij}) + \lambda \frac{1}{|N_i|}(r_i - a_i) \right], \tag{2.12}$$

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

The averages a_i are recomputed only after each iteration. We note that the gradient of the loss function (2.11), computed with the use of a single match, differs from the expressions included in update rules (2.12). The regularization part is additionally divided by the size of the multiset of opponents $|N_i|$ of team i. This is a heuristic argument - the more matches are played, the bigger weight we assign to the results rather than the regularization component in the loss function (2.11).

The term "stochastic" is motivated by two sources of randomness in the algorithm. Firstly, during each iteration the order of games for update equations (2.12) is random. Secondly, we do not directly compute the gradient of the loss function (2.11) but its estimate computed with the use of a single game per update.

Equations (2.12) are analogues of the corresponding update formulas for the Elo model (2.2). We read out version of the K-factor in this formulation:

$$K = w_{ij} p_{ij} (1 - p_{ij}).$$

The other part of the update equations (2.12) is associated with the regularization component that aims to keep the rating of each team close to its average opponents' ratings.

Choice of global parameters

The choice for the parameters h and λ is experimental: we set these parameters to some values, calculate the ratings on a training set and test their prediction accuracy on an independent test set. The pair (h, λ) yielding the best results shall be our choice for final model estimation.

Although our talk about Elo and Elo++ is quite different, both methods exhibit a number of resemblances. The most interesting are perhaps the update formulas and correspondence of K-factors. Elo++ in its primary application to rating chess players performed very well. We hope for similar results in application to rating football teams.

2.4. Least squares ratings

The model we present in this section can be concisely summarized under the name *least squares ratings*. The discussion is based on the work by Stefani [30] and Massey [18], where the application of the least squares method in US college football is discussed in detail.

In the model we perform regression on the goal difference in a match between two teams. The independent variables explaining the goal difference are indicators for two teams participating in a match. Again, as a rating value for a team we will obtain a single number that is the coefficient for the indicator variables in the regression model. After an introductory part and the example of calculation of this rating method we will present some possible extensions to the model.

Let us formalize the ideas described above. The main assumption of this method is that the score difference y between two teams A and B meeting in an encounter is proportional to the participating teams' ratings r_A, r_B difference, i.e., we assume that (for simplicity the proportionality constant is equal 1):

$$y = r_A - r_B + \varepsilon, \tag{2.13}$$

where ε is an error in the measurement. For a set of n teams playing a total number of k matches it can be written in matrix notation

$$\mathbf{y} = \mathbf{X}\mathbf{r} + \mathbf{e}, \tag{2.14}$$

where $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \dots & y_k \end{pmatrix}$ is the column vector of goal differences in consecutive matches, $\mathbf{y} = \begin{pmatrix} r_1 & r_2 & \dots & r_n \end{pmatrix}$ are unknown rating values we want to estimate (column vector), $\mathbf{X}$ is a $k \times n$ matrix that in each row all entries are zero but for the two teams participating in a game (indicator variables) where we set 1 and -1 depending on which rating difference is appropriate and finally $\mathbf{e}$ is the residuals vector. For instance, in a game between the teams indexed i and j with $i < j$ the corresponding (say m-th) row of the matrix $\mathbf{X}$ is given

$$(\mathbf{X})_{m,\cdot} = \begin{pmatrix} 0 & 0 & \dots & 1 & \dots & -1 & \dots & 0 \end{pmatrix} \tag{2.15}$$

with the only non-zero elements at the i-th and j-th entry with 1 and -1 respectively. In order to estimate the ratings (coefficients in the regression model) we look for a vector $\mathbf{r} \in \mathbb{R}^n$ that minimizes the sum of squared errors across all the games

$$\sum_{i=1}^{k} \varepsilon_i^2 = \sum_{i=1}^{k} \left(y_i - (r_{A(i)} - r_{B(i)})\right)^2 = (\mathbf{y} - \mathbf{X}\mathbf{r})^T(\mathbf{y} - \mathbf{X}\mathbf{r}), \tag{2.16}$$

where $A(i)$, $B(i)$ are the two teams associated with the i-th game in appropriate order and M^T denotes the transpose of matrix M (in our case this is a vector).

To find candidates solutions $\mathbf{r}$ that minimize (2.16) we regard this expression as a function F of $\mathbf{r}$ and compute its first order derivatives w.r.t. $\mathbf{r}$ denoted as $D_{\mathbf{r}}F(\mathbf{r})$ in a row vector:

$$\begin{aligned} D_{\mathbf{r}}F(\mathbf{r}) &= D_{\mathbf{r}} \left(\mathbf{y} - \mathbf{X}\mathbf{r})^T(\mathbf{y} - \mathbf{X}\mathbf{r}\right) = \\ &= D_{\mathbf{r}}(\mathbf{y}^T - \mathbf{r}^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\mathbf{r}) = \\ &= D_{\mathbf{r}}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{r} - \mathbf{r}^T\mathbf{X}^T\mathbf{y} + \mathbf{r}^T\mathbf{X}^T\mathbf{X}\mathbf{r}) = \\ &= -2\mathbf{y}^T\mathbf{X} + 2\mathbf{r}^T\mathbf{X}^T\mathbf{X} \end{aligned} \tag{2.17}$$

Now we find critical points of F, i.e., points where the first order derivatives are 0. We set (2.17) equal to 0 and, after transposing and dividing by 2, we obtain following system of equations that $\mathbf{r}$ must satisfy, called *normal equations*:

$$\mathbf{X}^T\mathbf{X}\mathbf{r} = \mathbf{X}^T\mathbf{y}. \tag{2.18}$$

Let us assume for now that the matrix $\mathbf{X}$ has full column rank, i.e., if $\mathbf{X}\mathbf{v} = 0$ for a vector $\mathbf{v} \in \mathbb{R}^n$ then necessarily $\mathbf{v} = 0$. If this is the case, a vector $\mathbf{r} \in \mathbb{R}^n$ solving the system (2.18) is in fact corresponding to the minimum of the error function (2.16). Indeed, differentiating once again w.r.t. to $\mathbf{r}$ the expression (2.17) for $D_{\mathbf{r}}F(\mathbf{r})$ we obtain that the Hessian matrix of the error function is of the form

$$D^2_{\mathbf{r}}F(\mathbf{r}) = 2\mathbf{X}^T\mathbf{X}.$$

Now, $\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} = (\mathbf{X}\mathbf{v})^T\mathbf{X}\mathbf{v} = \mathbf{w}^T\mathbf{w} \geq 0$ by the properties of the inner product with the equality $\mathbf{w}^T\mathbf{w} = 0 \Leftrightarrow \mathbf{w} = 0 \Leftrightarrow \mathbf{v} = 0$. Hence the Hessian matrix is positive definite and consequently the critical point we find by solving (2.18) corresponds to a minimum.

The system (2.18) has a unique solution if the square matrix $\mathbf{X}^T\mathbf{X}$ has full rank (in our case it should be equal to the number of teams n). In this case it is invertible and the solution is given by $\mathbf{r} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. However, in our formulation of the model (2.14) the parameters in the vector $\mathbf{r}$ are *not identified* - whenever $\mathbf{r}$ is a solution to the error minimization problem (2.16) then so is $\mathbf{r} + \mathbf{c}$ for a constant vector $\mathbf{c}$ (with all entries equal to some $c \in \mathbb{R}$). From (2.13) it is clear that the team ratings r_A and r_B will yield the same value of the goal difference as $r_A + c$ and $r_B + c$. Consequently, the system of normal equations (2.18) does not have a unique solution.

Perhaps the easiest way to see that the matrix $\mathbf{X}^T\mathbf{X}$ does not have a full rank is the fact that in each row of the matrix $\mathbf{X}$ the sum of its elements is zero - we have only two non-zero elements which are 1 and -1. In this case the columns of the matrix $\mathbf{X}$ are linearly dependent and therefore the rank of this matrix is $\leq n-1$. Consequently, because $rank(\mathbf{X}) = rank(\mathbf{X}^T\mathbf{X})$, the latter matrix also cannot have full rank. In this case the system of normal equations has either no solution or an infinite number of these.

To give yet another argument for the problem of finding a unique solution to the system of normal equations (2.18) let us have a closer look at its very form:

$$\begin{pmatrix} G_1 & -g_{12} & -g_{13} & \cdots & -g_{1n} \\ -g_{21} & G_2 & -g_{23} & \cdots & -g_{2n} \\ -g_{31} & -g_{32} & G_3 & \cdots & -g_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -g_{n1} & -g_{n2} & -g_{n3} & \cdots & G_n \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_n \end{pmatrix}. \tag{2.19}$$

We explain here what the notation means. The (i, j) entry of the matrix $\mathbf{X}^T\mathbf{X}$ is the inner product of the i-th and the j-th column of matrix $\mathbf{X}$. If $i = j$ its value will be equal to the number of games played by the i-th team, G_i. Whenever we have a non-zero element of the i-th column of the matrix $\mathbf{X}$, 1 or -1, it will be squared and summed over all games in the database, resulting in the total number of games played by the i-th team. In the case $i \neq j$ the inner product of the corresponding columns will be equal to the number of games played between the i-th and the j-th team times (-1). Indeed, the corresponding (i.e., lying in the same row) entries of the columns of the matrix $\mathbf{X}$ are both non-zero if and only if the two teams meet in one of the encounters (corresponding to that particular row). Then the product of these elements is -1 as one of them is equal to 1 and another -1. Summing over all matches we obtain the number of games played between the teams $g_{ij} = g_{ji}$, with the minus sign. This explains the form of the matrix $\mathbf{X}^T\mathbf{X}$ in (2.19). Now, the i-th entry of the vector on the right hand side of Equation (2.19) is the goal difference for the i-th team, denoted as d_i. It is equal to the inner product of the i-th column of the matrix $\mathbf{X}$, that in fact indicates in which games the i-th team took part, with the vector of goal differences $\mathbf{y}$. We add up all the goals scored or conceded by a particular team multiplied by 1 or -1 accordingly.

From this discussion we see that $G_i = \sum_{j \neq i}^{n} g_{ij} = \sum_{j \neq i}^{n} g_{ji}$ and also $\sum_{i=1}^{n} d_i = 0$ so the entries in each row (and column) of matrix $\mathbf{X}^T\mathbf{X}$ sum to 0 as well as the goal differences. Hence matrix $\mathbf{X}^T\mathbf{X}$ does not possess inverse so the system (2.18) does not have a unique solution.

We shall suggest a few ways to overcome the drawback of singularity of our matrix. First, we can decide for a chosen team i to have a zero (or some other chosen number) rating r_i and rate all other teams relative to team i. Another possibility would be to impose a constraint on the sum of ratings

$$r_1 + r_2 + \cdots + r_n = c. \tag{2.20}$$

for some constant c, which we may choose to be 0 for convenience. There are two possibilities for incorporating this restriction in our model: we may introduce it as *a soft constraint* by adding another component to the sum of squared errors in (2.16)

$$\lambda \left(r_1 + r_2 + ... + r_n\right)^2$$

where $\lambda > 0$ is a regularization parameter (analogously as in the Elo++ model (2.11)). We may also decide to make it *a hard constraint*, that is matched exactly. By the preceding discussion the rows (columns) of the matrix $\mathbf{X}^T\mathbf{X}$ are linearly dependent. Therefore any of n

equations in the system (2.18) is redundant. We may replace it by Equation (2.20) to arrive at the modified system (2.19):

$$\begin{pmatrix} G_1 & -g_{12} & -g_{13} & \dots & -g_{1n} \\ -g_{21} & G_2 & -g_{23} & \dots & -g_{2n} \\ -g_{31} & -g_{32} & G_3 & \dots & -g_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -g_{(n-1)1} & -g_{(n-1)2} & -g_{(n-3)3} & \dots & -g_{(n-1)n} \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_{n-1} \\ r_n \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{n-1} \\ 0 \end{pmatrix}, \tag{2.21}$$

where we replaced the last row in the original system by the restriction on r's. In this case introducing a strong restriction of the form (2.20) will force the ratings to sum precisely to 0. If we incorporate it as in the latter case of a soft constraint, the sum shall be close to 0. However we should not expect that it would be met precisely.

To focus our attention in the further discussion we should decide on introducing the constraint in the strong form as in the system above. This approach is equivalent to minimizing the sum of squared errors (2.16) subject to constraint (2.20).

To complete the mathematical details of our model we should note that our method of fixing the system (2.18) in fact solves the problem of singularity of the matrix $\mathbf{X}^T\mathbf{X}$. Later, when applying the model to the data we shall check that $rank(\mathbf{X}^T\mathbf{X}) = n - 1$ so only one additional equation is needed for unique solvability of the system (2.18). Otherwise, additional constraints are needed. For example, it will be the case if the set of teams under consideration can be decomposed into two or more nonempty subsets such that each subset consists only of the teams that play against each other while playing against no teams from other subsets. In this situation, the comparison of the teams in the global scale with available data is impossible as we do not have any evidence how the teams from different subsets would perform against each other. This is one of the problems we should discuss in Chapter 3 on data description.

Now we proceed to a small example of calculation of the method.

Example

Let us consider a small tournament with the results of games given in Table (2.5) We want

Team 1	Result	Team 2
C	**3:1**	D
B	**1:2**	A
D	**2:1**	A
C	**4:2**	A
C	**1:0**	B

Table 2.5: Sample dataset.

to obtain ranks $r = (r_A,\ r_B,\ r_C,\ r_D)$ for the teams A, B, C and D respectively. The matrix $\mathbf{X}$ indicates teams that take part in consecutive matches and it is of the form

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 1 & -1 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \end{pmatrix}$$

and the goal differences vector for consecutive matches is $\mathbf{y} = (2, -1, 1, 2, 1)$. The system of normal equations (2.18) becomes

$$\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} r_A \\ r_B \\ r_C \\ r_D \end{pmatrix} = \begin{pmatrix} -2 \\ -2 \\ 5 \\ 1 \end{pmatrix}$$

We note that the above system of equations has the special form given in (2.18). The rank of the matrix $\mathbf{X}^T\mathbf{X}$ is 3 so one equation is superfluous. We replace the last row by imposing the constraint $r_A + r_B + r_C + r_D = 0$:

$$\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} r_A \\ r_B \\ r_C \\ r_D \end{pmatrix} = \begin{pmatrix} -2 \\ -2 \\ 5 \\ 0 \end{pmatrix}$$

Solving this system we obtain that team C has the highest ranking of 1.25 followed by team D (-0.125), A (-0.5) and B (-0.625). □

Interpretation of the ratings

To give more insight into the resulting ratings from the least square rating method we shall now provide the interpretation of the ratings. Let us consider rating r_i for the i-th team under consideration and the associated i-th normal equation in the system (2.19):

$$r_i G_i - \sum_{i \neq j} r_j g_{ij} = d_i.$$

After rearrangement this becomes

$$r_i = \frac{d_i + \sum_{i \neq j} r_j g_{ij}}{G_i} \tag{2.22}$$

We can see that the ratings are obtained as weighted average of the goal difference and the strength schedule for a team i. In this manner, a team scoring many goals will tend to have a high rating, which is also clear from the form of the equation

$$y = r_i - r_j + \varepsilon,$$

describing the match against opponent j. The other part of the weighted average is concerned with the rivals the chosen team i plays against. If team i plays against high rated opponents it will be rated high itself. Ratings should be considered in relative way, i.e., usually the magnitude of r_i is irrelevant, but the magnitude of the difference $r_i - r_j$ tells us about the strength of the teams relative to each other.

By scoring many goals the team can get its rating go up. It can be considered as a drawback of this method: a relatively big component of the goal difference d_i in the average (2.22) may give a team an exceedingly high rating[5]. To reduce the role of the extraordinary

[5] An example of overrating/underrating teams by the least square method may be the case of Australia prior to 2006, when they still used to be a member of the Oceania Football Confederation (OFC). Australia used to win the games in its zone with extremely high margins (for instance 31:0 against American Samoa). Meanwhile, Italy became World Champion in 2006 and the team was known for the fans as the one with narrow victories and solid defense, thus having a relatively low goal difference comparing to that of Australia, but was clearly considered to be a better team. Moreover, on the road to winning the World Cup tournament they beat Australia in last sixteen (with the lowest margin 1:0 after an injury-time penalty kick).

high victories (losses) in the model one can consider the vector $\sqrt{|\mathbf{y}|}$ with $+/-$ at appropriate entries instead of the original the goal differences vector $\mathbf{y}$.

Extensions to the model

Several improvements to the original model can be suggested to make it more reliable and to facilitate the interpretation given to the weights. Among many possibilities we present in particular 3 concerning incorporation of home team advantage, time weighting and game importance weighting.

First of all, if we have information on which team plays at home we may include it in the model by introducing another explanatory variable that indicates where the appropriate game took place. In this case we arrive at the following equation describing the game against teams A and B

$$y = r_A - r_B + h + \varepsilon, \tag{2.23}$$

where team A is assumed to be the host of the game. Here we make again an assumption about the influence of home advantage - we assume that it has an additive effect on the goals scored by the team and is equal for all teams under consideration.

Another modification would be to assign weights relative to the importance of the game $I_m \in \{i_1,\ i_2,\ ...,\ i_l\}$, where l is the number of matches with different prestige and i's are importance weights (compare to the original FIFA ranking). We may also introduce time weighting $w_m(t)$ by multiplying Equation (2.23) for the particular match m by these numbers:

$$w_m(t) \times I_m \times y = w_m(t) \times I_m \times (r_A - r_B + h) + \varepsilon. \tag{2.24}$$

The possible choices for the function w can be $w(t) = (T-t)^\alpha$ or $w(t) = exp(-\alpha t)$ where $t \leq T$ denotes the time elapsed between the date of the match and the date of ratings the teams (in appropriate units), $\alpha > 0$ is a single parameter and T is the time horizon.

If we denote as $\mathbf{W}$ a diagonal $k \times k$ matrix (where k equals to the number of games), with weights for the m-th match at the m-th entry on the diagonal we can reformulate the initial least square model (2.14) to the weighted model

$$\mathbf{Wy} = \mathbf{WXr} + \mathbf{e}, \tag{2.25}$$

for which we find the least square solution in the same manner as above from the system of normal equations

$$\mathbf{X}^T\mathbf{W}^2\mathbf{Xr} = \mathbf{X}^T\mathbf{W}^2\mathbf{y}.$$

Again we shall impose constraint (2.20) for the above system to have a unique solution in case the matrix at the left hand side of the equation is singular.

The interpretation of the weights still remains similar, but now each component of the goal difference d and particular games in the sum $\sum r_i g_{ij}$ in (2.22) are weighted according to the time and importance. Obviously, in this setting the ratings $\mathbf{r}$ are determined by the choice of the time weighting function and the importance factors.

Strength estimation on goals scored - related work

We presented the linear regression approach for explaining the goal difference between two teams. We mention below a few related models from the literature.

Moroney [22] suggests the Poisson and the negative binomial distribution for modeling goals in a single match. Maher [17] investigates further the Poisson distribution. He assumes that goals scored by two teams in a game are independent Poisson variables and also proposes a bivariate Poisson distribution, which allows for the correlation between the scores. This partially relaxes the assumption of independence. Dixon and Coles' [6] approach aims to develop further modeling of scores by Poisson variables. In literature, it is common to divide the estimate of strength r into two components that indicate offensive and defensive skills of the teams. This is also possible in the least squares approach [18].

2.5. Network-based rating system

The last two algorithms we present are based on intuitive principles governing football fans' behavior. The methods we are going to discuss are derived from the analysis of a graph with the teams represented as nodes and edges corresponding to games played between them. The method we discuss was applied for rating of US college football teams but primarily originates from social network analysis [23].

An intuitive idea behind the algorithm is based on a conversation between two football fans. One of them supports team A and the second is a fan of team C. They argue which of the teams is the better one. Their teams have not played each other, however, each of them played against a third team B: team A winning its encounter and team C losing to team B. The fan of team A argues therefore that his team is the better one because it has beaten team B, which in turn beat team C. We can say that the team A has *an indirect win* over team C.

Let us represent the considered situation with the use of a graph, where nodes represent teams which are connected by an edge if they met in a match. We make the graph directed with the edges pointing to the winners (by convention). The situation described above corresponds to a directed path of length two, $C \to B \to A$. We can further lengthen the paths to find more indirect wins of a chosen team. We can apply the same procedure for counting losses. It is natural to assume that indirect wins/losses count for less. We introduce a discount factor $\alpha \in (0, 1)$. A win (loss) which corresponds to a directed path of length $k \geq 1$ is weighted by α^{k-1}. Thus the most valuable are wins in direct matches with a weight equal to 1.

Suppose that we are given a set of n teams. Let us define the total *win score* for the i-th team w_i as the sum of all direct and indirect wins of all distances. To deal with draws we treat them as half win half loss. If a tie occurs between two teams we credit 0.5 win and 0.5 loss for each of them. Let A be a version of the adjacency matrix for the network of teams, a square $n \times n$ matrix with entries 0, 0.5, 1, 1.5, 2, 2.5, ... such that entry $(A)_{ij} = a_{ij}$ is equal to the number of victories of the j-th over the i-th team (halves are possible as they correspond to draws). Now, all direct wins of the j-th team can be written as

$$\text{direct wins for team } j = \textstyle\sum_i a_{ij},$$

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

$$\text{indirect wins of distance 2 for the team } j = \textstyle\sum_{i,k} a_{ki} a_{ij},$$

and so forth. We can compute the number of all wins w_i by the i-th team weighted with discount factor α^{k-1} for the wins at distance k as

$$w_i = \sum_j a_{ji} + \alpha \sum_{j,k} a_{kj} a_{ji} + \alpha^2 \sum_{j,k,l} a_{lk} a_{kj} a_{ji} + ... = \tag{2.26}$$

$$= \sum_j (1 + \alpha \sum_k a_{kj} + \alpha^2 \sum_{l,k} a_{lk} a_{kj} + ...) a_{ji} = \sum_j (1 + \alpha w_j) a_{ji} = d_i^{in} + \alpha \sum_j (A^T)_{ij} w_j,$$

where d_j^{in} is the number of edges pointing to the vertex j, i.e., the number of direct wins of team j 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. In a similar manner we can compute *the loss score* l_i for team i

$$l_i = \sum_j a_{ij} + \alpha \sum_{j,k} a_{ij}a_{jk} + \alpha^2 \sum_{j,k,l} a_{ij}a_{jk}a_{kl} + ... = \tag{2.27}$$

$$= \sum_j a_{ij}(1 + \alpha \sum_k a_{jk} + \alpha^2 \sum_{k,l} a_{jk}a_{kl} + ...) = \sum_j a_{ij}(1 + \alpha l_j) = d_i^{out} + \alpha \sum_j (A)_{ij} l_j,$$

where d_i^{out} is the number of edges pointing out of the chosen node j. Let us finally define the ranks for the teams i as $r_i = w_i - l_i$. We can see that if a team beats a strong team, which is regarded as the one with a high value of w_i, then this team is rewarded with a big increase in its rating. Analogously, a loss to a team with a high value of a loss measure l results in a considerable decrease in ranking for that team.

Let us put Formulas (2.26) and (2.27) in matrix notation with (column) vectors $w = \begin{pmatrix} w_1 & w_2 & ... & w_n \end{pmatrix}$, $l = \begin{pmatrix} l_1 & l_2 & ... & l_n \end{pmatrix}$, $d^{in} = \begin{pmatrix} d_1^{in} & d_2^{in} & ... & d_n^{in} \end{pmatrix}$ and $d^{out} = \begin{pmatrix} d_1^{out} & d_2^{out} & ... & d_n^{out} \end{pmatrix}$:

$$w = d^{in} + \alpha A^T w,$$

$$l = d^{out} + \alpha A l.$$

If the matrix $(I - \alpha A)$ is invertible i.e. whenever α^{-1} is not an eigenvalue of the matrix A we can put the formulas above in the form

$$w = (I - \alpha A^T)^{-1} d^{in}, \tag{2.28}$$

$$l = (I - \alpha A)^{-1} d^{out}. \tag{2.29}$$

This gives us the procedure to obtain the rating vector $r = w - l$ for the teams under consideration.

The power series (2.26) and (2.27) converge whenever $\alpha < \lambda_{max}^{-1}$, where λ_{max} is the largest eigenvalue of the adjacency matrix A. Therefore λ_{max}^{-1} is an upper bound on the possible choices for the single parameter of the algorithm α. Moreover, if the graph has no loops of the form team A beat B, B beat C and C beats A (or longer) then all eigenvalues of the adjacency are zero. In this case no restriction is imposed on the choice of parameter α.

Example

Let us again consider the sample dataset given in Table (2.5). The graph structure of the competition is presented in Figure (2.2). The associated matrix A has the form

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Consecutive rows (columns) of this matrix correspond to the teams A, B, C, D and a column standing for a particular team indicates its victories over other teams. The characteristic polynomial of matrix A is equal to

$$det(A - \lambda I) = \lambda^4$$

and hence all eigenvalues of this matrix are zero. Therefore there is no restriction on the choice of parameter α (there are no loops in the network (2.2)). For $\alpha = 0.5$ the solutions (2.28) and (2.29) are given by

$$w = \begin{pmatrix} 1 & 0 & 4.25 & 1.5 \end{pmatrix}$$

and

$$l = \begin{pmatrix} 2.5 & 3.25 & 0 & 1 \end{pmatrix}$$

Hence we obtain ratings

$$r = w - l = \begin{pmatrix} -1.5 & -3.25 & 4.25 & 0.5 \end{pmatrix}.$$

We arrive with the ordering of the teams C, D, A, B with team C rated the highest (the same as in least square model). □

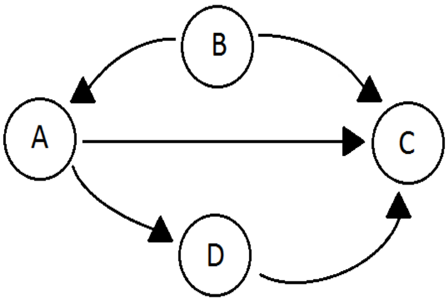

Figure 2.2: Visualization of the small competition (2.5) with the use of a graph with the edges pointing towards the winners.

2.6. Markovian ratings

The last algorithm we discuss is also based on an intuitive principle. It describes behavior of a football fan, which is not stable in his feelings. The main idea behind this rating method is to consider a glory supporter who always prefers the winning teams. By studying his/her preferences in the long run, we may answer the question which teams with what probability the fan is going to support. The most favorite teams will get the highest ratings. The model can be formulated in terms of Markov chains.

The idea of the model is based on the PageRank algorithm that aims to rate web pages in accordance to their importance [1]. Application of this algorithm to rating sport teams is not novel [2], [19]. Also Markovian ratings for different sport disciplines are implemented and maintained on the website `www.thepowerrank.com`. We present a method for modeling head-to-head transition probabilities for the network of teams that allows for multiple matches and incorporates draws.

Similarly to the previous setting, we can represent our situation with the use of a graph (2.2). A football fan "jumps" over the graph and changes his affinity in favor of a winning team. When the supporter considers a chosen team, he looks at all games it played against other teams and either sticks with supporting his current team or jumps to another one. By constructing an appropriate transition matrix we can find the stationary distribution over the supporter's preference towards each team.

Let us explain how the transition probabilities are modeled. First, we focus on estimating the supporter's preference towards two chosen teams A and B. We assume that the hypothetical fan preference can be described by a single real number p that assumes value in the interval $[0,1]$ according to some distribution π. A situation when $p = 0$ corresponds to the total preference of the team A over B and $p = 1$ the other way around in favor of the second team B. The intermediate values of p correspond to the situation that the fan is not fully decided which team to prefer. The particular case $p = 0.5$ is interpreted as indifference towards both teams (in the general setting, p and $1 - p$ express the preference towards team A and B respectively). We describe how the distribution π over p is estimated from data. Prior to the game, the fan chooses his preference parameter p randomly. This corresponds to drawing p from the uniform distribution on the unit interval, $U([0,1])$. The uniform distribution is a particular case from a more general class of *beta distributions*. This family is described by two parameters α, β, and has the density function

$$B_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ x \in [0,1],$$

where $\Gamma(x)$ is the *gamma function*, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-x}dx$. The uniform distribution corresponds to beta distribution with $\alpha = 1$ and $\beta = 1$. When evidence E about the outcome of the game is obtained we can compute *the posterior* distribution over the parameter p according to the Bayes' rule:

$$\pi(p|E) = \frac{\mathbb{P}(E|p) \cdot \pi(p)}{\mathbb{P}(E)}. \tag{2.30}$$

For now let us assume two way win/loss outcome for a single match (later we explain how we deal with the draws). The fan watched total number of n games with k wins of team A and accordingly $n - k$ wins of the second team B. If we assume independence between the consecutive games, the likelihood of observing the corresponding evidence as the function of p is equal to

$$\mathcal{L}(p) = \mathbb{P}(E|p) = \binom{n}{k} p^k(1-p)^{n-k}.$$

The formula above expresses the probability of observing k heads and $n - k$ tails in n independent coin flips, where as the single outcome we obtain 'head'with probability p. In our setting, this can be interpreted as *the total utility* of the supporter, when his preference towards team A is governed by the number $p \in (0,1)$. With the use of Bayes' formula (2.30) we may compute *a posteriori* distribution over the parameter p:

$$\pi(p|E) \propto p^k(1-p)^{n-k}p^{\alpha-1}(1-p)^{\beta-1} \propto p^{k+\alpha-1}(1-p)^{n-k+\beta-1}, \tag{2.31}$$

where the proportionality symbol $\propto$ allows us to skip the constant term that does not depend on p. We read off the posterior distribution to be again a Beta density with parameters $(\alpha+k)$ and $(\beta + n - k)$. This distribution exhibits how the preference of the fan has changed after watching n matches between the two teams A and B under consideration.

The draws in this formulation can be treated as a halfway between a victory and a loss [11]. We just include them to the model by adding 0.5 to the both parameters of the beta density. Justification for doing so is by assuming that a win followed by a loss (or the other way around) yields the same likelihood as a draw. This likelihood is equal to $p(1-p)$ so the likelihood from a single draw should be $\sqrt{p(1-p)} = p^{0.5}(1-p)^{0.5}$. Incorporating a single draw to Expression (2.31) results in the posterior beta density with both parameters increased by a half.

The discussion above is based on the analysis of a series of coin flips that result in 'head' with probability p under the assumption of *the prior* beta distribution on parameter p. The beta distribution is called *a conjugate prior distribution* to the Bernoulli trials, that enables for easy computation of *the posterior* distribution over the parameter p, with the use of Bayes

Formula (2.30). Our adaptation of this situation for modeling a football fan's behavior is slightly different, but it may correspond to the 'preference' of the coin to produce 'head' or 'tail' result. The preference parameter is adjusted according to the observed results of an experiment (coin flip, match).

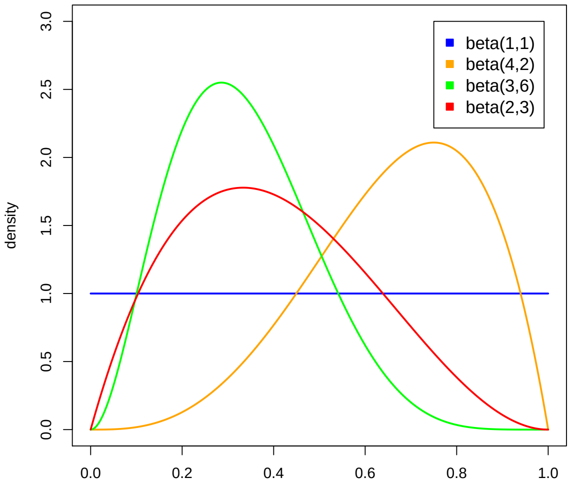

Figure 2.3: Beta densities for different pairs(α, β).

So far we discussed the modeling transition between two chosen teams A and B. Now we turn to the description of the whole system. If additionally we assume that the fan is memoryless, i.e., when he sticks with one of the teams and he forgets about all the teams he supported in the past, then his behavior can be modeled by a Markov chain. Given the number of N teams under consideration we create $N \times N$ transition matrix M that describes how the supporter changes his preferred team. Each row of the matrix corresponds to one of the teams and its entries are the transition probabilities in favor of other teams. For the i-th and the j-th team the distribution of the 'preference' parameter p towards the i-th versus the j-th team is modeled as above. However, to fill in the entries of the matrix M we need a single number. To this end, we may use maximum posterior estimate $\hat{p}_{MP}$, which is the value that maximizes the posterior density. The posterior beta distribution with parameters $(\alpha + k)$ and $(\beta + n - k)$ assumes the maximum at the point

$$\hat{p}_{MP} = \frac{\alpha + k - 1}{\alpha + \beta + n - 2}. \tag{2.32}$$

Another possibility would be to take the expected value of the posterior distribution, $\hat{p}_{EP}$, which in the case of the discussed distribution is equal to

$$\hat{p}_{EP} = \frac{\alpha + k}{\alpha + \beta + n}. \tag{2.33}$$

The corresponding (i, j) entry of the matrix M is proportional to the estimated value $\hat{p}$ and accordingly the entry (j, i) to the value $1 - \hat{p}$ (we say proportional rather than equal since the matrix M still needs to be scaled in order to become a stochastic matrix). Finally, the diagonal entry is proportional to the value

$$\hat{p}_{ii} = \sum_j (1 - \hat{p}_{ij}),$$

where the summation goes over all the indexes j of teams that the i-th team played against. This expression is proportional to the probability of the event that the supporter stays with his current team (denoted as S). Indeed, from the law of total probability we have that

$$\mathbb{P}(S) = \sum_{j=1}^{N} \mathbb{P}(S|team\ i\ plays\ team\ j) \cdot \mathbb{P}(team\ i\ plays\ team\ j) = \frac{1}{N}\sum_{j=1}^{N}(1 - p_{ij}),$$

assuming that the choice of the match for each opponent is done with equal probability $\frac{1}{N}$.

We should normalize the rows of the matrix M, so that it becomes a stochastic matrix, where the entries in each row add up to 1.

Now it remains to compute the stationary distribution of the described Markov chain. This is defined as a non-zero (row) vector π solving the equation:

$$\pi = \pi M.$$

The stationary distribution exists if the corresponding Markov chain is *irreducible* and *aperiodic.* The irreducibility assumption is satisfied if there exists a *recurrent* state, i.e. a state from which there is a path to any other state[6]. The aperiodicity of the chain is usually harder to check. However, both of the assumptions may be satisfied by introducing a modification[7] to the original matrix M

$$\widetilde{M} = \alpha M + (1 - \alpha)E, \tag{2.34}$$

where E is $N \times N$ matrix with all entries equal to $\frac{1}{N}$ and $\alpha \in (0, 1)$. By a such construction the matrix $\widetilde{M}$ is also a stochastic matrix and all necessary assumptions for existence of a stationary distribution are clearly satisfied.

Example

Let us again consider the set of results given in the Table (2.5). We use the mean posterior estimator (2.33) for the entries of the matrix M. In case we decide to use the maximum posterior estimator (2.32) we would arrive with 0 or 1 results as in this case the maximum of the posterior beta density is assumed at the boundary of interval $[0, 1]$. Each of the five games ended in a win so the mean value of posterior distribution gives

$$\hat{p}_{EP} = \frac{\alpha + k}{\alpha + \beta + n} = \frac{1 + 1}{1 + 1 + 1} = \frac{2}{3},$$

towards the winning team. The complete transition matrix M is of the form

$$M = \begin{pmatrix} \frac{4}{9} & \frac{1}{9} & \frac{2}{9} & \frac{2}{9} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{9} & \frac{1}{9} & \frac{2}{3} & \frac{1}{9} \\ \frac{1}{6} & 0 & \frac{1}{3} & \frac{1}{2} \end{pmatrix}.$$

[6] Australia is a good example of a recurrent state - it is enough that it plays a single game against one of the countries from each of 6 football confederations, which is usually the case.

[7] This idea is applied in the PageRank algorithm as for the adjacency matrix of web pages, which is usually sparse - it is hard to satisfy necessary assumptions for existence of stationary distribution.

We compute the stationary distribution as a probabilistic vector $\pi = (\pi_A, \pi_B, \pi_C, \pi_D)$, $\sum_{i\in\{A,B,C,D\}} \pi_i = 1$ that solves

$$\pi = \pi M.$$

After computations we arrive at

$$\pi = \begin{pmatrix} \frac{33}{149} & \frac{17}{149} & \frac{69}{149} & \frac{30}{149} \end{pmatrix} \approx \begin{pmatrix} 0.22 & 0.11 & 0.46 & 0.20 \end{pmatrix}.$$

We obtain that the highest rated team is C followed by A, D and B on further positions. □

Variations to the method

The approach described here is similar to the work by Mattingly and Murphy [19] for rating College Football in USA. The authors assume that each team plays one another not more than once and they introduce a single parameter p that governs the fan preference towards winning teams. The setting above allows for multiple games between the same teams and also incorporates draws.

We may introduce a modification to the method, that the fan changes his preference whenever a single goal is scored. Then the prior distribution is updated according to the goals scored by each team. For example, in case of a game between A and B teams that ended with the score s_A to s_B, the posterior probability for preference of team A is beta$(1 + s_A,\ 1 + s_B)$ with mean value

$$\hat{p} = \frac{s_A + 1}{s_A + s_B + 2}.$$

Considering exact scores gives us alternative model analogous to the one described by Keener [16].

Remark - Unofficial Football World Champions

A somewhat radical version of the algorithm above is applied for determining so called unofficial world champion. The idea appeared when Scotland beat England in 1967, which won the World Cup the year before. Scottish fans claimed that, having beaten world cup holders, they had become unofficial world champions. Since then defenses of the title are scheduled in a boxing-style competition with the winner becoming the next unofficial world champion[8].

2.7. Summary

In this section we provide a summary of features incorporated in individual ratings systems. Most of the entries of Table (2.7) are self-explanatory. We discuss in more details the match importance and the time factor.

Match importance

We note that it is virtually possible to incorporate match importance factor to any model we discussed. This is included in both Elo models we presented at the beginning. In Elo++ or least squares approaches we may attach bigger weights to the components in the cost function that are believed to be more important in determining the true strength of the teams. In graph analysis models we may also introduce multipliers for different type of matches. It is natural to regard friendlies as less informative for team strength, as the teams may not play their first

[8] For more details about this not so serious football competition visit www.ufwc.co.uk.

teams. However, the importance of the games is hard to quantify and amounts for another set of parameters in the model. It is based on the subjective view of the inventor of a rating method.

Ranking	Time factor	Goal scored	Match importance	Home team advantage
FIFA ranking	✓	×	✓	×
Elo WWR	×	✓	✓	✓
Elo ratings.net	×	✓	✓	✓
Elo ++	✓	×	×	✓
Least squares	×	✓	×	✓/×
Network based system	×	×	×	×
Markovian ratings	×	✓/×	×	×

Table 2.6: The summary of the features included in individual rating systems. WWR states for Women's World Ranking.

Time factor

Incorporating time into the algorithms is somehow optional, but as the shape of the teams is not constant in time it seems reasonable to include this factor in the rating method. It is typical that due to the changes of the players as well as managers the teams are constantly changing in time, so the estimates of the strength may not be accurate half a year later. We note that the time is not included in the Elo model explicitly, however the ratings obtained by this method are dependent on the order of updates - the most impact on the current values of ratings are dependent on the most recent games.

In the Elo++ model the most recent games have more impact on ratings. A similar idea may be applied in the least squares model by attaching time weights to the sum of squares cost function.

Incorporating both time and importance weights in graph algorithms is possible but at the expense of intuitive properties of these models. In their very basic form the ratings provided by them have a direct interpretation. We will keep these methods simple. Introducing modifications gives us additional parameters to optimize. Obviously we can introduce time or match prestige to every model under consideration but we will not pursue the extension here.

Figure (2.4) presents different methods for game weighting with the dependence on the time it took place.

Home team advantage

Incorporation of advantage of the home team into the model is made under rather strong assumptions. First of all, in every model it is considered to have an additive impact on rating points for the team playing at home. Moreover, we assume that this impact is homogenous among all teams. The advantage in favor of a team playing at home is a well-known fact in football and it seems reasonable to include it into consideration. We provide some more details on this phenomenon in the next chapter.

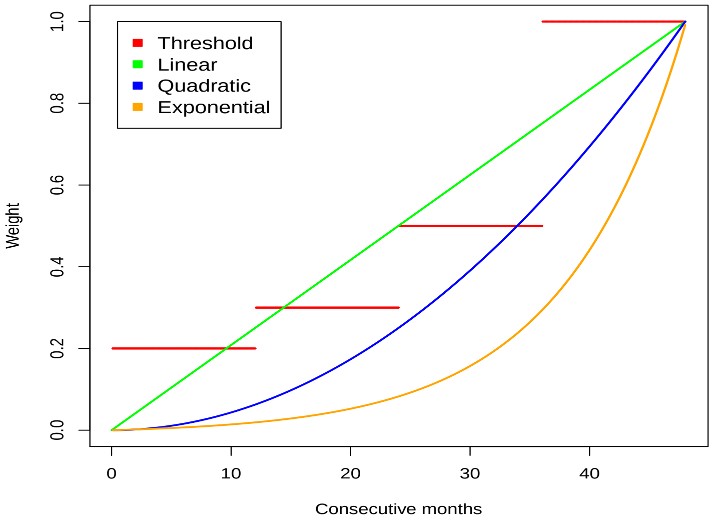

Figure 2.4: Different choices for weighting games in time. The most recent of the results obtain maximal weight of 1. Threshold weighting is applied in the official FIFA ranking, quadratic in the Elo++ model.

Impact of frequency of games played

The algorithm used in the calculation of FIFA rankings has the intuitive property that less active countries are penalized by taking only a fraction of their yearly point averages. Frequency of games played enters in our models in different ways.

We note that a team that does not play for long time in the Elo rating model it does not lose any points. This is considered a drawback of the Elo rating system. Often, once a chess player gets a high rating he does not have incentive to play more games.

There is no particular reason to favor or penalize less active teams by Elo++ or the least squares methods. However, in those approaches, if a team performs well in relatively few games it would be rated high. In those models (as well as any other) to obtain a reliable estimate of the team strength we need a reasonable number of games played by every team.

In social network ratings consideration of the number of games played is not clear. It may well happen that a team with one game played which was a victory over a top ranked team may become the new ranking leader. It depends on the discount factor. If it is chosen to be large enough, this team may receive substantial number of points to its win score and in turn may overtake the leader in the table. Obviously the loss score for that team is 0. When we discount indirect wins by a smaller fraction, the team may not gain much for beating the leader. In particular, setting the discount parameter α to 0 corresponds to the situation when we count only direct wins and losses for each team.

In Markovian ratings the schedule of the games has an important impact on the ratings. First of all, a game between two teams opens a connection between them. When a team plays often and has many connections it gives it a substantial rating gain. This is particularly rewarding when the team plays against strong opponents. This idea fits particularly well in the framework of the web surfer that finds relevant website by links which point to it.

Graph algorithms primarily were applied to ratings in a different domain than sport. It makes it harder to consider them in football. It is possible to tailor them for rating sport teams and extend by time factor, match importance or home team advantage. However, the simplicity of those methods is particularly appealing.

Before we compare the algorithms in experimental results, in the next chapter we describe the dataset that we are working with.

Chapter 3

Data description

The data for our analysis are the results of international matches played between 15 July 2006 and 2 May 2012 recognized by FIFA as the official ones. The dataset was obtained with the use of information available at the official FIFA website[1]. Along with results of the games, we extracted attributes describing the type of the match, the date, the venue (city) where it took place and information about possible extra time or penalty shoot-out. There are a few exceptions when the match was abandoned or a victory was awarded due to exceptional circumstances. We skip those games in the analysis and later in the model estimation as often those games do not have sport background. After deletion of those matches we are left with a total number of 5297 international games. In this part we provide characterization of the dataset. The discussion is supported with the Wikipedia website on FIFA ranking [32].

Data preparation requires substantial amount of time. It is a costly and sometimes long process to obtain good quality data. Usually running an algorithm on a ready dataset takes a while (main task), but the whole setup before doing so is laborious.

Type of the match

We can distinguish 6 different kind of games that a country can be involved in: FIFA World Cup match, FIFA World Cup Qualifier, Continental Final (the major competition played within each confederation), Continental Qualifier, FIFA Confederations Cup and a friendly game. The number of different kind of matches in the dataset is given in the table (3). In general, the importance of the game is determined by the prestige of the competition that a country is taking part in. It seems reasonable to assume that the World Cup games or continental finals are the best indicators for a team's true strength. Often, the purpose of a friendly game is to experiment with the squad. The result of such games can be sometimes surprising for a football fan. Presumably it is more likely that a potentially weaker team wins a friendly match rather than a high stake game at a major tournament or its qualification rounds. However, the most common type of game in the data is a friendly match which amounts for over 50% of all games.

Teams and frequency of matches played

On the release of the official FIFA ranking in May 2012 there are 207 countries all over the world included in the ranking. The number of ranked teams varies between periods as a consequence of, for example, the change of political status of a country. The data between the years 2006-2012 consists of the matches played by a total number of 208 national teams.

[1] www.fifa.com/worldfootball/results/index.html

Type of the game	**Matches played**	**Percentage**
FIFA World Cup Final	64	1 %
FIFA World Cup Qualifier	1064	20%
Continental Final	303	6 %
Continental Qualifier	1120	21%
FIFA Confederations Cup	16	< 1%
Friendly	2730	52%
Total:	**5297**	

Table 3.1: Proportions of different type of matches in the dataset

Since the dissolution of the Netherlands Antilles in October 2010, this country is no longer included in the ranking. Since March 2011, Curaçao is a successor of Netherlands Antilles. Such issues are of concern in data integration and preparation for the analysis. Although Caribbean countries are rather not expected to shake world ranking table, one needs to be careful about such details to prepare a reliable source of data.

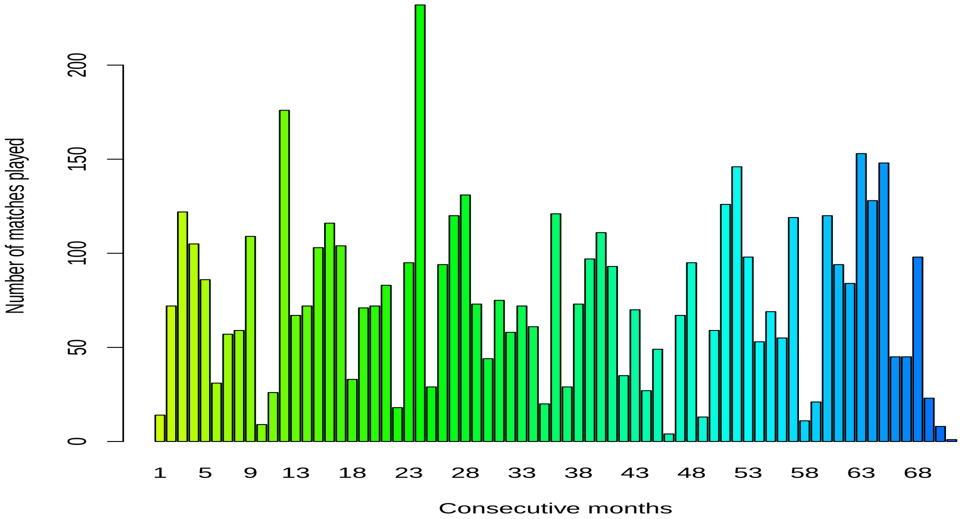

Figure 3.1: Number of games played in consecutive months. The greatest number of matches (232) were played in June 2008, the least number in April 2010. In FIFA ranking methodology, that weights the game by a threshold parameter associated with the time, the ranking can suddenly change when games get lower weight upon new ranking release.

During the investigated period of five years the number of matches played by a single team varies considerably. With 5297 matches and 208 teams we have the average of approximately 51 matches played by a single team. However, some of the countries played very few official matches, e.g., Democratic Republic of Sao Tome and Principe (5 matches), Papua New Guinea (4 matches), while others many of them, e.g., Bahrain (100), Oman (98) or Mexico and Saudi Arabia (97). Small number of games played by a single team makes it more difficult to infer about its actual strength.

If the teams around the world were to compete in a single round-robin tournament (if such event could be organized) it would require $\binom{208}{2} = 21528$ matches. Across the dataset, some of the teams compete frequently in mutual matches while others (majority) have never played each other. The largest number of matches played was 11 games between Honduras and El Salvador (8/2/1) and 10 matches between Costa Rica and El Salvador (6/2/2). If we take into account only matches between different opponents we are left with 2657 games (A vs. B is considered to be the same game as B vs. A). This is merely a fraction of all 21528 possibilities (12%).

The distribution of the number of matches in time depends mainly on the official time scheduled by FIFA for qualifiers, friendly matches and on the dates of competitions (which also encourages playing friendly games). During the observed period of time the peak number of matches was in June 2008 with the European Championship taking place along with a number of FIFA World Cup qualifiers on other continents and friendly games. The number of games in consecutive months in the dataset is presented in figure (3.1).

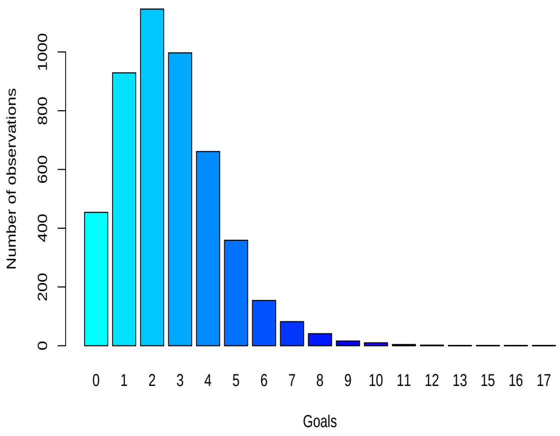

Figure 3.2: The number matches with particular number of goals scored with the mean around 2.64. The shape of the plot may suggest a Poisson distribution for the total number of goals scored in a single game. However, application of χ^2 goodness of fit test rejects this hypothesis. In particular, there are more observations without a single goal scored than one would expect - 455 vs. 346 expected under Poisson distribution with mean 2.64. This does not mean that football is a boring game.

Regional grouping of matches

FIFA recognizes 6 confederations that supervise football competitions in different parts of the world: AFC (currently unites 46 national football associations), CAF (53), CONCACAF (35), CONMEBOL (10), OFC (11) and UEFA (53). The boundaries of confederations are roughly determined by the borders of the continents. However, there are some exceptions. For example, countries such as Israel, Kazakhstan or Russia are united under the Union of European Football Associations (UEFA) although geographically either they are not located in Europe or do so only partially. The membership of a country can also change in time. Australia is member of Asian Football Confederation (AFC) since 2006, previously belonging to the Oceania Football Confederation (OFC).

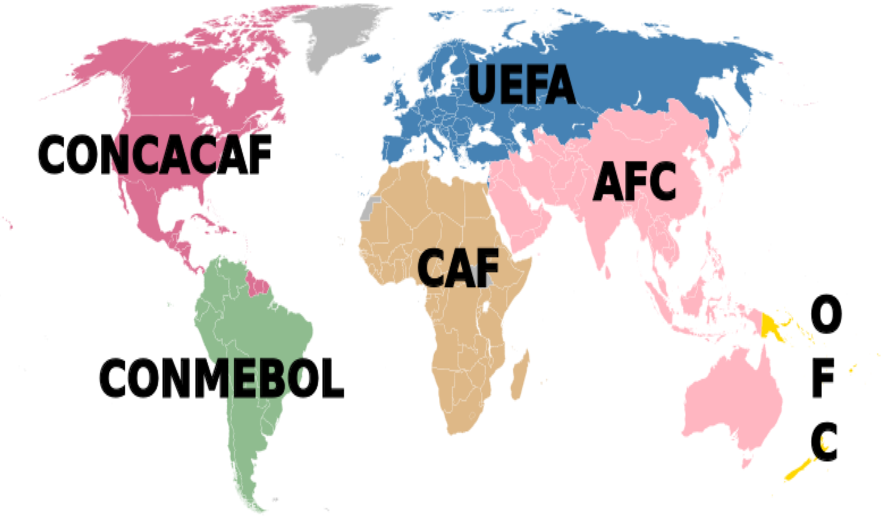

Figure 3.3: The membership of national football associations around the world to different confederations (source [32]).

The issue that comes with the data is the fact that around 85% of all games are the matches played between the teams from the same confederation, which is a significant fraction. The continental finals as well as their qualification rounds amount for a huge number of matches played within each confederation separately. The FIFA World Cup, held every 4 years, is the only major intercontinental competition (the second is FIFA Confederations Cup, but undoubtedly less important). However, qualifications for the World Cup are played again within each continent independently (with very few exceptions). What is more, usually weaker teams rarely qualify for major tournaments so seldom have they an opportunity to play against a team from a different region. The only intercontinental matches are played during the two FIFA tournaments and friendly games (including small competitions). This cluster nature of the data makes it harder to compare teams on the intercontinental level. In an extreme case, it may well happen that different teams within two separate regions of the world have never met in a match. In this situation the comparison on global scale is impossible.

The structure of the matches in the confederation vs. confederation form is presented in Table (3). There were only 3 encounters between teams from Oceania zone and Africa. For visualization of the phenomenon of regional grouping of matches see Figure (3.4).

	AFC	CAF	CONCACAF	CONMEBOL	OFC	UEFA
AFC	1138	122	21	38	12	117
CAF	122	1136	16	37	3	78
CONCACAF	21	16	627	146	5	67
CONMEBOL	38	37	146	204	3	103
OFC	12	3	5	3	73	7
UEFA	117	78	67	103	7	1344

Table 3.2: The number matches played on confederation level. High numbers on the diagonal indicates that the vast majority of the games are played within 6 world confederations.

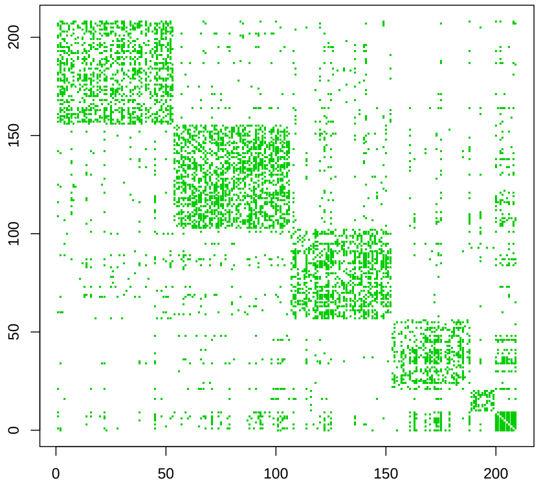

Figure 3.4: Visualization of the adjacency matrix for the teams as nodes and an edge between them if a match was played. The consecutive rows (columns) are arranged in order of confederation membership. The block nature of the matrix indicates that most of the games are played within the confederations with much fewer games played intercontinentally (entries besides main blocks).

Home team advantage

The topic of the advantage due to playing at home ground is well studied in different sports and particularly in football [25], [27]. A concise summary of the factors contributing to the advantage of the host of the game is given by Pollard in his work [24]. The author provides a summary of research that has been done so far in this area. The main factors are associated with crowd effect, travel effects, familiarity, referee bias and territoriality.

Certainly the crowd support contributes for the better results obtained by the home team. Travel effects and associated fatigue was also studied as another factor. It was also argued that referees tend to favor home teams which is apparent by the number of disciplinary cards and other decisions. The familiarity with the avenue is also considered to contribute to the advantage of the home team. Finally, the territoriality is associated with the fact that human and animals are known to respond for invasion of their home territory.

In order to analyze the phenomenon of home advantage we need a method to quantify it. In research, usually it is done by calculating the percentage of total available points gained by the teams playing at home ground (with the convention of awarding 3 points for a victory, 1 point for a draw and 0 for a loss) [24]. This statistic is equal 50% if and only if home and away teams win equal number of matches, regardless of the number of draws. The greater it is than 50%, the more it indicates for advantage of teams playing at home ground. We present calculations of this statistic for our data in Table (3). For chosen tournaments in this table, the teams play one another twice. The schedule is balanced and hence the computations should be unbiased as no trend that stronger teams are usually hosting the game is included (however, it does not apply to friendly and overall stats).

In order to incorporate information about the host of particular game an additional attribute was created that indicates which team plays at home. The locations of the games were

Competition	Home wins	Draws	Away wins	% home wins	% Draws	% Away wins	HTA (%)
Euro 2012 qualifications	144	44	81	47.7	18.41	33.89	**57.36**
World Cup 2010 Qual. AFC	60	30	39	46.51	23.26	30.23	**58.82**
World Cup 2010 Qual. UEFA	122	56	80	47.29	21.71	31.01	**58.77**
World Cup 2010 Qual. CONMEBOL	47	22	21	52.22	24.44	23.33	**65.73**
World Cup 2010 Qual. CAF	115	33	43	60.21	17.28	22.51	**70.00**
Friendlies	947	494	464	49.71	25.93	24.36	**63.88**
Overall	2011	922	1050	50.49	23.15	26.36	**63.01**

Table 3.3: Statistics regarding home team advantage in football in different international competitions.

assigned to particular countries which allowed identifying home teams. In total, for around 75% games in the database a host were indicated and the rest of games were considered to be played at a neutral ground. We hope that inclusion of this attribute will support the analysis of strength estimation and improve predictions.

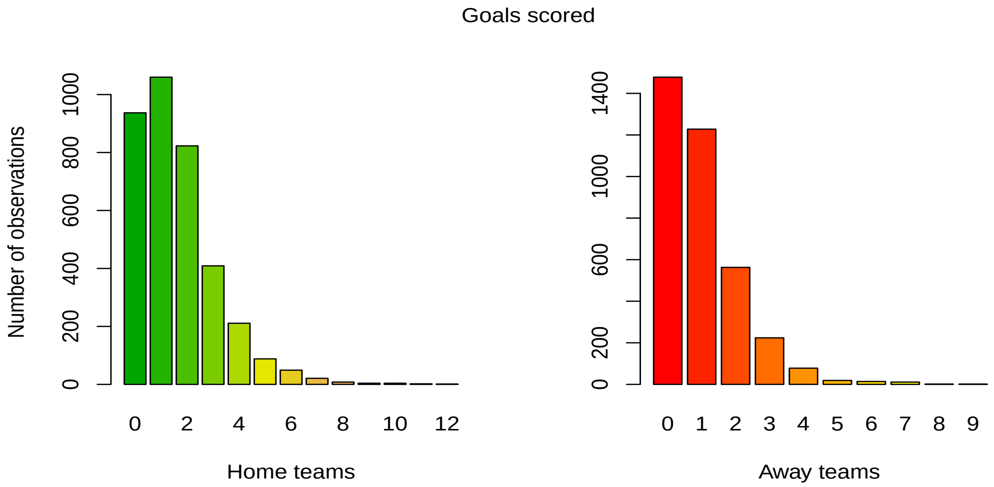

Figure 3.5: Distribution of the number of goals scored by teams playing at home ground and by the teams playing away (the games played at neutral ground are skipped). Mean values are 1.62 and 1.00 for home at away teams respectively.

Chapter 4

Implementation and evaluation of rating algorithms

In order to say what a good rating system is we need to define a method that assesses its quality. In this chapter we present two evaluation measures of the quality of a ranking. Both of them are based on predictive capabilities of a ranking. Based on predefined evaluation measures, we present comparison between rating systems discussed in Chapter 2.

Some of the models we present in Chapter 2 are equipped with a method for predicting outcome of games given teams' ratings. In fact, the prediction function is the core of the Elo approach and its main driving force. However, we need a method for calculating the probability of the match outcome with the use of the teams' ratings derived from the algorithms that do not provide these estimates themselves. In the next section we discuss methods for turning ratings into predictions. This will enable us to evaluate rating algorithms according to their predictive power.

4.1. Match outcome prediction

Let two teams A and B be rated with values r_A and r_B respectively. Given those numbers we make a comparison between them. In case $r_A > r_B$ we would say that team A has bigger chances to win the encounter against team B. Intuitively, the bigger difference (or ratio) in the rating points the more probable is the victory of the higher ranked team. Let us denote by $\mathbb{P}(A|r_A, r_B)$ the probability that team A wins the match against team B given their strength estimates. Analogously as in the Elo model, we disregard the possibility of a draw and model only binary win/loss match outcomes. We can estimate probabilities of these events with a function of ratings values of the teams

$$\mathbb{P}(A|r_A, r_B) = f(r_A, r_B)$$

and accordingly

$$\mathbb{P}(B|r_A, r_B) = 1 - f(r_A, r_B),$$

where the function f assumes its values in the interval $[0, 1]$ in accordance with probability axioms. We can interpret the estimated probabilities close to 0.5 as a draw to be the most probable result to happen. It is also reasonable to demand that the function f is such that in case of equal ratings, neither of the teams is expected to win the game, $f(r, r) = 0.5$.

A simple choice of the function f would be to pick the higher ranked team to win the game:

$$f(r_A, r_B) = \begin{cases} 1 & \text{if } r_A > r_B, \\ 0.5 & \text{if } r_A = r_B, \\ 0 & \text{if } r_A < r_B. \end{cases} \tag{4.1}$$

However, we would expect that if the teams are closely ranked then victory of the higher ranked team value is less likely than in case of the bigger difference in ratings. That method seems to be quite rough and we could instead choose a continuous link function f.

Before we proceed, we will make an additional assumption on the form of the prediction function. When comparing skills of two teams, r_A and r_B, we will be interested only in the difference $r_A - r_B$ in their rating points. This seems to be a reasonable assumption as we may think that when two teams meet in an encounter, the absolute value of their ratings cancels off and what really matters is the difference between them. However, it was observed that, for example, in chess there is a bigger fraction of draws at the highest level of competition and hence the result is somehow dependent on the magnitude of skills [10]. This assumption is a simplification.

In the following, we consider $f(r_A, r_B) = g(r_A - r_B)$ with g a non-decreasing function (the bigger difference in ranking the higher probability of victory of team A) such that $g(0) = 0.5$. Taking into account the consideration above, we can assume a linear probability model between the difference in rating points and expected result of the game

$$\mathbb{P}(A|r_A, r_B) = \frac{1}{2} + a(r_A - r_B),$$

where $a > 0$ is a single parameter. In this case we cannot assure that the function models a probability distribution as it is clear that we can obtain values outside interval $[0, 1]$ for sufficiently big values of $|r_A - r_B|$. We can overcome this drawback by truncating the values of the function outside the interval $[0, 1]$:

$$\mathbb{P}(A|r_A, r_B) = \begin{cases} 1 & \text{if } \frac{1}{2} + a(r_A - r_B) > 1, \\ \frac{1}{2} + a(r_A - r_B) & \text{if } \frac{1}{2} + a(r_A - r_B) \in [0, 1], \\ 0 & \text{if } \frac{1}{2} + a(r_A - r_B) < 0. \end{cases}$$

to arrive at correct probability values.

A standard method would be to use the link functions between rating differences and the probability of result from logit or probit regression models and consider the following functions [20] (which are also applied for outcome prediction in the models discussed in Chapter 2)

$$\mathbb{P}(A|r_A, r_B) = \frac{1}{1 + e^{-a(r_A - r_B)}}, \tag{4.2}$$

or

$$\mathbb{P}(A|r_A, r_B) = \Phi(a(r_A - r_B)), \tag{4.3}$$

where Φ is a cumulative distribution function of a standard normal variable $\mathcal{N}(0, 1)$.

The models considered above can be extended for predicting a probability of a drawn game [26], [4]. Basic modification would be to include the information on which team is the host of the game.

Advantage due to playing at home ground

As we already mentioned several times it is a well-known fact in sport the team playing at home has an advantage over its rival. One might expect that if two teams are approximately equally ranked then it is slightly more probable that the home team would be the winner. This feature is incorporated in some of the models that we discussed.

When making predictions on rating points we may introduce home advantage in the chosen model function (4.2). We automatically confer h additional ratings points to the team A playing to arrive with a model

$$\mathbb{P}(A|r_A, r_B) = \frac{e^{a(r_A+h)}}{e^{a(r_A+h)} + e^{ar_B}} = \frac{1}{1 + e^{-a(r_A+h-r_B)}} \tag{4.4}$$

and accordingly for the away team victory

$$\mathbb{P}(B|r_A, r_B) = 1 - \mathbb{P}(B|r_A, r_B) = \frac{e^{-a(r_A+h-r_B)}}{1 + e^{-a(r_A+h-r_B)}}.$$

In this manner, if the teams are equally ranked we obtain that the probability $\mathbb{P}(A|r_A, r_B) > 0.5$.

4.2. Accuracy of predictions

With a model for the estimation of the probability of the result given rating values we can define an evaluation measure of a rating system. We present two standard statistics for assessment of accuracy of predictions.

A sample dataset is shown in the table (4.1). The numbers in the columns with actual and predicted result (denoted as s_i and p_i for i-th game) represents probabilities that the team in the first column team wins a match. As usual, if this team in fact won the match then the actual result is mapped to 1 (and 0 in case of loss). A draw is mapped to 0.5. Our main accuracy measure will be the statistic of binomial deviance (4.2) - logarithm of likelihood - which for the i-th game is equal to

Match					Actual	Predicted	Binomial	Squared
					result	result	deviance	error
🇵🇱	Poland	**2:2**	Germany	🇩🇪	0.5	0.3	1.797	0.04
🇨🇷	Costa Rica	**0:1**	Brazil	🇧🇷	0	0.12	0.294	0.014
🇺🇦	Ukraine	**3:0**	Bulgaria	🇧🇬	1	0.72	0.756	0.078
🇪🇸	Spain	**3:1**	Scotland	🏴󠁧󠁢󠁳󠁣󠁴󠁿	1	0.93	0.167	0.005
🇵🇱	Poland	**2:1**	Hungary	🇭🇺	1	0.65	0.992	0.123
						Mean:	**0.801**	**0.052**

Table 4.1: Sample dataset.

$$-\left(s_i \log_{10} p_i + (1 - s_i) \log_{10}(1 - p_i)\right). \tag{4.5}$$

Averaging it over all games in the testing period gives us accuracy of the predictions based on a given rating method.

We can see that binomial deviance is infinite (undefined) in case of sure predictions $p_i = 0$ or $p_i = 1$ in case $y_i = 1$ and $y_i = 0$, respectively. This means a heavy penalty for prediction the events that have not happened. In computations of this statistic, we round estimated values of p_i that are lower than 0.01 to 0.01. Analogously, in case the predicted value of $p_i > 0.99$ we set $p_i = 0.99$. In case of predicting home team victory with probability one ($p_i = 1$) when it actually happened ($s_i = 1$) the binomial deviance should be zero as function $x \log x$ can be continuously extended at 0, $\lim_{x \to 0} x \log x = 0$. With the restriction for the values of p_i to the interval $[0.01, 0.99]$ we make the error smaller than $0.03 > -\log_{10}(0.99)$. If the game ended in a draw, the value of p_i minimizing (4.5) is equal to 0.5.

Another possibility would be to calculate the squared error (4.2) of the prediction for the i-th game

$$(s_i - p_i)^2.$$

We report both measures in experimental results below. We note that when making predictions on rating points below we are in fact maximizing the logarithm of the likelihood of the predictions (see below). Therefore this measure is our main benchmark towards assessment of the accuracy. The mean squared error is an alternative, but it is not the main focus of optimization techniques to keep it low.

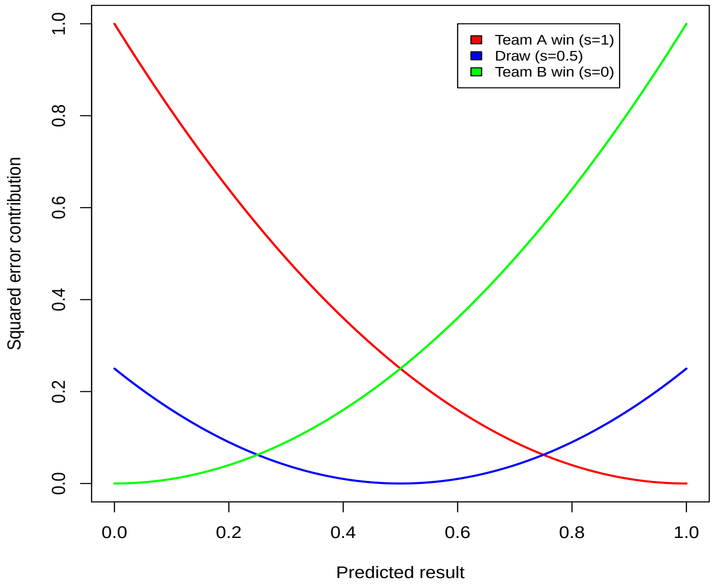

Figure 4.1: Contribution towards squared error of prediction of a single match.

Similar accuracy measures were used for assessment of the predictions in both **Kaggle.com** chess ratings competitions [13], [14].

4.3. Experimental results

We have seen how rating points can be used for predicting game results and next how we can assess their accuracy. This section presents the performance of the ranking algorithms under consideration. Firstly, we describe the training and the test set choice.

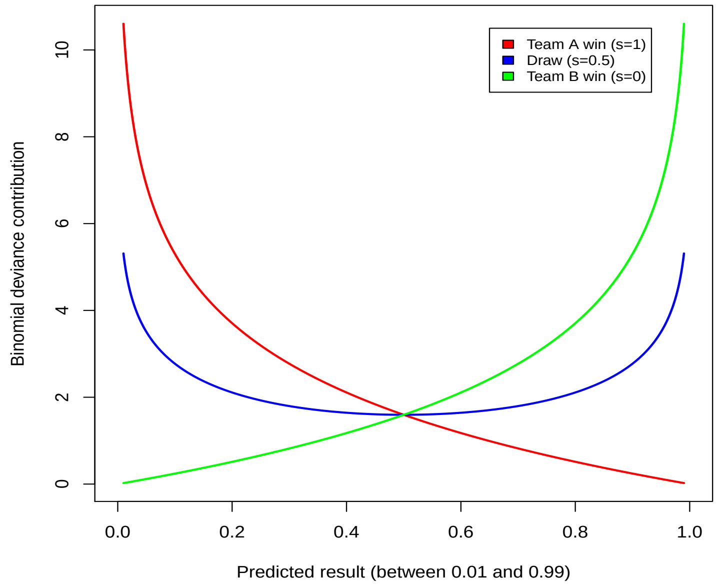

Figure 4.2: Contribution towards the binomial deviance (logarithm of likelihood) statistic.

4.3.1. Training and test set

Each release of the FIFA ranking takes into account the last four years of international games. To make a comparison between the benchmark model, initially we shall choose the training set for the models under consideration that includes games played during a 4-year period dating from the rating day, similarly as the FIFA does. However, we experiment with the time span for the training period to see if the suggested models, with a restricted training period, yields better accuracy. The final models will be evaluated on an independent test set.

Some of the models we discuss involve parameters that need to be optimized. To this end, we shall choose a validation set, independent of the final test set. Table (4.2) presents our choice for the final training set, validation set and test set.

	Begin date	End date	Number of games
Training set	15/07/2006	31/03/2011	4318
Validation set	15/07/2010	31/03/2011	726
Test set	1/04/2011	2/05/2012	979

Table 4.2: The choice of training, validation and final test set for the models.

The structure of the test set is presented in Table (4.3). As we already saw in the previous chapter, most of the games are friendlies.

We evaluate models on a daily basis (remember that the FIFA ranking is released on a monthly basis). We note in the FIFA rating procedure a team's rating points are determined solely by the results achieved by that team. This is also a property of the Elo rating system. However, in Elo++, Least squares and graph algorithms estimation of the strength is strongly connected with the schedule for a chosen team. Because ratings are partially determined by the ratings of all opponents a team play with within the training period, they are more

Test set	
Friendly	461
FIFA World Cup Qualifier	219
Continental Qualifier	219
Continental Final	80

Table 4.3: Different type of games in test set.

dynamic and are changing with respect to the opposition performance. This is particularly evident in graph algorithms. A single match influences the whole network of teams. However, impact on ratings of the teams far apart in the ratings is negligible.

The table below presents the FIFA ranking release dates, in the validation and test set. We will present the FIFA algorithm performance on a daily basis and also with respect to monthly releases that are published on the official FIFA website.

FIFA ranking release	No of games until the next release	No of games in last 4 years
14 Jul 10	5	3592
11 Aug 10	156	3575
15 Sep 10	158	3563
20 Oct 10	38	3599
17 Nov 10	103	3575
15 Dec 10	46	3637
12 Jan 11	46	3662
02 Feb 11	55	3657
09 Mar 11	128	3648
13 Apr 11	3	3671
18 May 11	131	3662
29 Jun 11	87	3600
27 Jul 11	83	3622
24 Aug 11	165	3640
21 Sep 11	132	3691
19 Oct 11	133	3726
23 Nov 11	56	3736
21 Dec 11	19	3762
18 Jan 12	41	3752
15 Feb 12	90	3684
07 Mar 12	22	3762
11 Apr 12	8	3694
Σ	1705	

4.3.2. Estimating prediction function

We note that some of the models under consideration (Elo, Elo++) are self-contained in the sense that they not only estimate the strength of the individual teams but also provide a method to predict future match outcomes. However, for the FIFA rating algorithm (and the rest) we need a method for predicting match results. Among the possibilities discussed at the beginning of this chapter we choose a logistic function as a link between team ratings and the probability of match result. Given two teams A and B with ratings r_A and r_B, respectively, the prediction of match outcome is given by the model

$$\mathbb{P}(s|r_A, r_B) = \frac{(e^{a(r_A - r_B)+h})^s}{1 + e^{a(r_A - r_B)+h}}, \tag{4.6}$$

where parameter h denotes credited points for home team advantage in favor of team A (h equals 0 if the game is played on a neutral ground) and a is a parameter that governs sensitivity of the probability of an outcome with respect to the difference in ratings. Moreover, s stands for the actual score of the game - $s = 1$ and $s = 0$ corresponds to the win of team A and team B, respectively. For incorporating draws we follow the convention adopted by Glickman [11]. 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 modeling a single draw we take the square root of the expression above to arrive at

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

Hence we include the draws in the analysis by setting $s = 0.5$ in (4.6). Now, we construct the likelihood of the observed match results as the function of the parameters a and h. If we assume that outcomes of the matches are independent events (which is rather not true in some cases, but nevertheless something needs to be assumed) the likelihood is given by

$$\mathcal{L}(a, h) = \prod_{i-th\ game} \left(\frac{(e^{a(r_A^{(i)} - r_B^{(i)})+h \cdot \mathbb{1}_{\{A\ at\ home\}}})^{s^{(i)}}}{1 + e^{a(r_A^{(i)} - r_B^{(i)})+h \cdot \mathbb{1}_{\{A\ at\ home\}}}} \right), \tag{4.7}$$

where $\mathbb{1}_{\{A\ at\ home\}}$ indicates whether the i-th match is played at team A's home ground, $r_A^{(i)}$ and $r_B^{(i)}$ stand for the given rating values of teams A and B involved in the i-th game and $s^{(i)}$ is the actual result of the game. To evaluate ratings at a chosen day we derive parameters (a, h) by estimation of the prediction function with the use of games in the period prior to the day of evaluation. Hence the ratings r_i that we plug to the likelihood function (4.7) are the most recent estimates provided by a given method. We make one exception to this rule by providing accuracy measurements for monthly FIFA ranking releases (available on the official website). Finally, the likelihood function (4.7) is maximized w.r.t. parameters a and h.

If rating points for some team are not available this game is disregarded in the estimation. This happens when, for example, a team has not played single game in the last 4 years.

Table (4.4) presents the result of estimation of the parameters (a, h) on the validation and test set consisting roughly of 1700 games. Note that these estimates are constantly changing. Once we have made a prediction for a chosen day, we include that day into the training set in a sliding window approach. Nevertheless, the changes are minor.

The estimates (a, h) depend on the scale in which rating points are expressed. For better insight, the ranking tables are provided in the Appendix.

Model	estimate a	estimate h
FIFA ranking daily releases	0.003	0.552
FIFA ranking monthly releases	0.003	0.554
Least squares	0.845	0.503
Least squares home team	0.886	0.531
Network-based ratings	0.035	0.514
Markov Wins	2.256	0.504
Markov Goals	1.981	0.511

Table 4.4: Estimated parameters for logistic prediction function.

4.3.3. Comparison between the models

Table (4.5) presents accuracy measures of the models. The assessment is based on the binomial deviance statistic (*BinDev*) and the mean squared error (*MSE*).

Before we discuss the setup of the chosen algorithms let us explain how confidence intervals for error measures are derived.

Let X_1, X_2, ... X_n be an i.i.d. sample from distribution $\mathbb{F}$ with unknown mean μ and variance σ^2. From the Central Limit Theorem we know that the statistics

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \tag{4.8}$$

converges in distribution as $n \to \infty$ to a random variable Z with the standard normal distribution, $Z \sim \mathcal{N}(0,1)$. Let z_α denote the quantile of order α for Z, $\mathbb{P}(Z \leq z_\alpha) = \alpha$. If n is large, the distribution of (4.8) should be approximately standard normal and we may write (informally) that

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha.$$

We work out the expression in brackets w.r.t the (unknown) value of the mean μ to arrive at

$$\mathbb{P}\left(\overline{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} \ \leq\ \mu\ \leq\ \overline{X} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}} \right),$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. For a standard normal variable we have $z_{\alpha/2} = -z_{1-\alpha/2}$. We obtain a confidence interval for μ at the level of α of the form

$$\left(\overline{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}},\ \overline{X} + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} \right). \tag{4.9}$$

In our case we are interested in the confidence intervals for error measures *MSE* and *BinDev*, where X_1, X_2, ..., X_n denote consecutive error measurements. To this end, we use the derived interval (4.9) with σ replaced by a sample standard deviation $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$. If for two chosen rating methods confidence intervals of their error measures overlap we will say that they do not differ significantly in accuracy.

There are two sources of approximation when we derive the confidence intervals in this way. Firstly, we use approximate by a normal distribution for a big sample. Secondly, we use the estimate $\hat{\sigma}$ of the standard deviation rather than σ itself. Moreover, it is arguable if consecutive error measurements can be regarded as an i.i.d. sample. Anyway, the confidence intervals give us some idea about the true accuracy of the models.

Let us proceed to the discussion of the setup for algorithms under consecutive headings.

Rating method	BinDev	90% confidence interval	MSE	90% confidence interval
FIFA ranking daily releases	1.3681	(1.3481, 1.388)	0.1443	(0.1244, 0.1643)
FIFA ranking monthly releases	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)
Elo ratings.net 1500	1.3265	(1.307, 1.346)	0.137	(0.1176, 0.1565)
Elo ratings.net	**1.2634**	**(1.2446, 1.2821)**	**0.1271**	**(0.1084, 0.1458)**
Elo ratings.net FIFA06	1.2811	(1.2624, 1.2999)	0.128	(0.1092, 0.1468)
Elo++ $\lambda = 0.1$	1.2997	(1.2806, 1.3188)	0.132	(0.1129, 0.1512)
Elo++ $\lambda = 0.05$	1.288	(1.2690, 1.3069)	0.1305	(0.1115, 0.1494)
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	1.4224	(1.4018, 1.4431)	0.154	(0.1333, 0.1746)
Markovian ratings Wins	1.3588	(1.3391, 1.3786)	0.1406	(0.1209, 0.1604)
Markovian ratings Goals	1.3541	(1.3344, 1.3738)	0.1399	(0.1202, 0.1595)
PowerRank.com	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	-

Table 4.5: Accuracy of predictions. In two benchmarks, *All draws* and *Home team* we always predict a draw (0.5) and home team victory.

The official FIFA ranking

Table (4.5) presents two versions of FIFA algorithm - updated on a monthly and a daily basis. We implemented the algorithm for daily updates to provide better insight in capabilities of this rating method.

To implement the FIFA algorithm to daily updates the following steps were taken. We retrieved all rating points gained by the teams during a 4 year-period between the dates 15/07/2006 and 14/07/2010. To this end, it was necessary to record all ranking released during that time since the points are calculated based on the teams' positions in the ranking. Having done this, we may run the official FIFA algorithm described in Chapter 2 on daily basis. We note that during the days when no games are played the ranking can change due to the time factor. To calculate points earned by teams on a particular matchday we used the ranking table from the day before.

FIFA Elo Women's World Rankings

An important part of the Elo model is the choice of prior ratings. To reduce the influence of prior ratings in determining accurately the strength estimation it is necessary to have many games played by each team. For example, in chess player ratings, estimates of capabilities that are calculated with the use of 20-30 games should be considered provisional. Only after the ratings are calculated with more games per player, they become reliable estimates of the player's strength.

There are a few choices of prior ratings. One possibility is to set ratings for every team to, say, 1500. It is a simple choice. However, with very few games played by some teams the ratings may not be reliable with such a rough choice of the prior. Another option would be to pose a question: what if FIFA would have changed its rating system to that applied in FIFA Women's World Rankings? To answer it, we set prior ratings for the teams to the FIFA ranking points from the 12 July 2006 release[1].

The results for different choices of prior are presented in Table (4.5): in **Elo WWR** 1500 each team is assigned 1500 points as prior, FIFA06 stands for the prior set to the mentioned FIFA ranking release. To get better insight into the importance of margin of victory, incorporated in FIFA Women World ranking methodology, we present also performance of another version denoted as FIFA06 WDL. In this model we map actual result of the game to 0, 0.5 or 1, like in chess, instead of using the values in the table (2.4).

As we already discussed in the summary of the modeling chapter, the importance weighting (included in the K factor in Elo WWR) is an arguable part of any model. However, we may experiment with the value of scaling parameter a in the prediction function, originally set to $a = \frac{\log_e 10}{400}$:

$$\frac{1}{1 + e^{-a(r_A + h - r_B)}} = \frac{1}{1 + 10^{-\frac{1}{400}(r_A + h - r_B)}}, \tag{4.10}$$

to see if we can achieve lower error measurements. This parameter governs the sensitivity of predictions with respect to rating differences. It turns out that modification yields some improvement. We vary the denominator in

$$\frac{r_A + h - r_B}{400}$$

by ±10 and report the error measurements on the validation set in Figure (4.3).

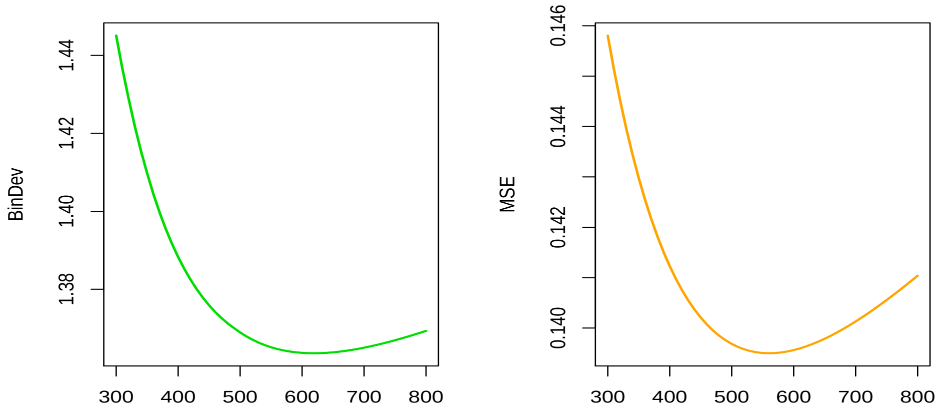

Figure 4.3: Impact on error measures with different choices of scaling in the Elo Women's World Ranking model (4.10). The lowest error measurements are obtained for $s = 620$ and $s = 560$ with respect to binomial deviance (1.3636) and mean squared error (0.1395) respectively.

[1] Note that from this day on, following the World Cup tournament in Germany, the official FIFA ranking calculation methodology has changed.

Elo ratings.net

For the second Elo model we obtained historical ratings from the website from 9 July 2010 available under `www.football-rankings.info`[2]. In this way we make direct comparison of **Eloratings.net** to other methods (there can be minor differences because of rounding to integers of the ratings maintained on the website). We also report the results for uniform 1500 and FIFA July 2006 ranking release priors.

Elo++

For the chess rating competition in the winning approach we have three parameters to optimize: λ for the regularization component, h that measures the impact of home advantage and $\alpha > 0$ that determines the choice of time weights in the model (2.9):

$$\left(\frac{1 + t - t_{min}}{1 + t_{max} - t_{min}}\right)^{\alpha}.$$

Table (4.4), presenting the values of estimated parameters (a, h) in the logistic function for predicting match outcome, may be helpful in setting appropriate values of the home advantage parameter. The choice of the parameters is experimental. It turns out that we obtain good results with the setting $(\lambda, h, \alpha) = (0.05, 0.5, 1.5)$.

Another issue that comes with the algorithm is its adaptation to the sliding window framework. Note that for prediction of games on a particular day we use all games played before that day up to four years. Here we experimented with daily, weekly and monthly and a few other prediction periods. It turns out that predictions day by day do not yield the best accuracy. Table (4.2) presents the results for monthly predictions as the ones achieving the best score.

The Elo++ method is the only algorithm that does not produce unique ratings. To minimize the cost function the stochastic gradient descent algorithm is used. In every iteration the gradient of the cost function is computed with the use of a single match that is chosen at random. To be able to reproduce results we set the seed for the pseudorandom number generator.

Least squares ratings

We consider two different least squares ratings models. The first of them is based on estimating team strengths without considering home team advantage. For the goal difference y between two teams i and j we assume the model

$$y = r_i - r_j + \varepsilon.$$

The second model incorporates additional information whether team i is the host of the game

$$y = r_i - r_j + h \cdot \mathbb{1}_{\{team\ i\ plays\ home\}} + \varepsilon.$$

Network-based rating system

For social network ratings we need to optimize the discount parameter α for indirect wins and losses. To this end, we use the validation set. The choice of this parameter is restricted to interval $(0, \lambda_{max}^{-1})$, where λ_{max} is the largest eigenvalue of the (modified) adjacency matrix A.

[2] `www.football-rankings.info/2010/07/elo-ratings-update-9-july-2010.html`, accessed 17 May 2012

We express the parameter α as the percentage of the largest possible value it can assume, i.e. λ_{max}^{-1}. Figure (4.4) presents the accuracy measures for different choices of the parameter. We achieve the best results with setting α to 30% of λ_{max}^{-1} with respect to the binomial deviance, which is our main accuracy indicator.

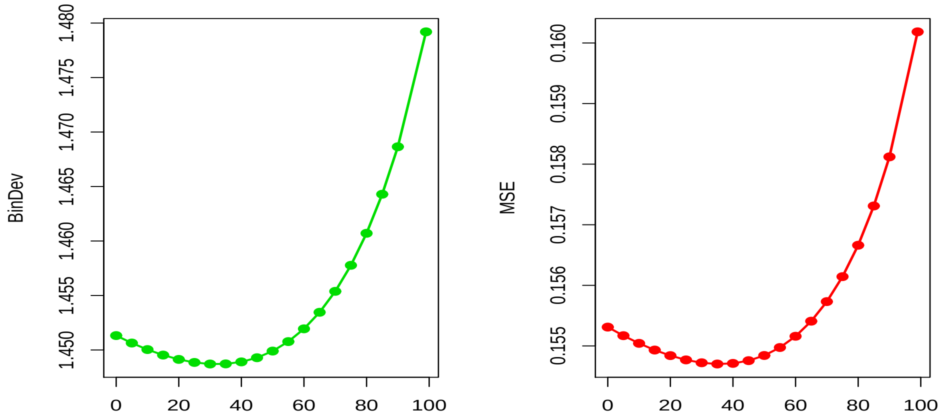

Figure 4.4: The impact of the choice of discount parameter α on error measures. Estimation is based on the total number of 394 matches. The rest of the games in the validation set (332) are used for initial estimation of the logistic prediction function parameters. These parameters are re-estimated in accordance with the sliding window approach.

Markovian ratings

Markovian ratings involve a single parameter $\alpha \in (0, 1)$ that allows for transitions between every country in the network of teams with a certain probability. Modification of the original matrix M by introducing perturbation in Equation (2.34)

$$\widetilde{M} = \alpha M + (1 - \alpha)E,$$

assures that corresponding Markov chain is irreducible and aperiodic. We investigate experimentally what impact on predictive capabilities of the Markovian ratings the parameter α has. To this end, we calculate ratings with different choices of that parameter and check accuracy of predictions on the validation set. The impact of small transitions on the error measures is presented in Figure (4.5).

We conclude that allowance for transitions between every team has negative influence on the model accuracy. In final computations we set α to a value close to 1, say $\alpha = 1 - 10^{-6}$. In this way we assure that all necessary assumptions for existence of the stationary distribution are satisfied and we restrict the impact of the parameter on predictions to a minimum.

PowerRank.com

The last predictions included into our comparison are obtained from the PowerRank ratings, maintained on the website `thepowerrank.com`. They 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.

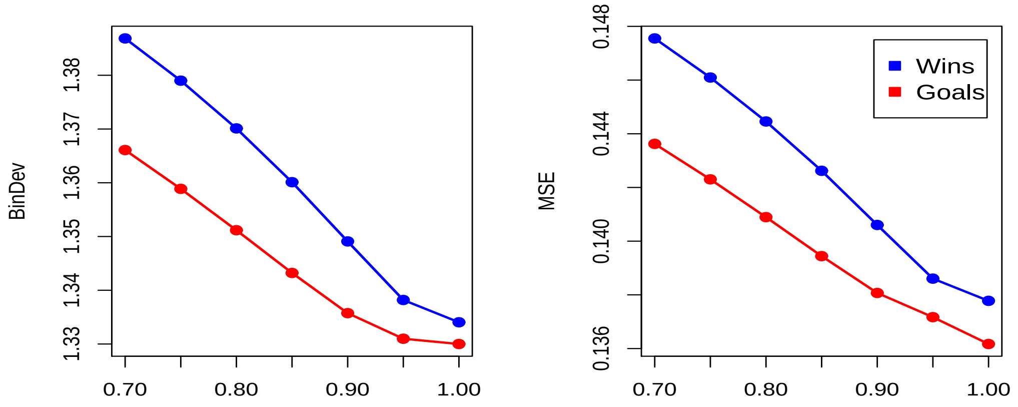

Figure 4.5: Impact on accuracy measures with various choices of the parameter α.

Ensemble approach

A popular technique for increasing accuracy is to combine several models into a single one. In this way we benefit from advantages of individual methods [5].

We suggest a simple ensemble approach that averages predictions from four of the best performing models: **ELO** WWR, **Elo ratings.net** and **Least squares** with home advantage and the **PowerRank.com** method. The combined model achieves the best score among all methods.

4.4. Performance comparison

In the previous section we presented the performance of different rating systems and their variations. The best accuracy is achieved by two Elo models. Eloratings.net system turns out to be the most accurate with respect to binomial deviance and the Elo model applied by FIFA in ranking women's team when we look at the mean squared error. The difference between FIFA ranking and Elo-type models, least squares ratings and the PowerRank.com ratings is significant when we compare their performance with respect to binomial deviance. Slightly better performance is achieved by the two versions of the Markovian ratings. When we compare the performance of the algorithm with respect to the mean squared error, the differences are not significant. However, we are mainly interested in the binomial deviance since the models are estimated using the maximum likelihood method.

By comparison between different versions of the two Elo models, we see that the choice of the prior has a major impact on the model performance. When we set the prior ratings to FIFA ranking release from July 2006 we see that we obtain more accurate estimates of team strength than the ones derived by the official FIFA algorithm.

The performance of the Elo++ model is somehow disappointing. It has strong generalization capabilities, however, this model neither incorporates the information on goals scored nor the match importance. By looking at the improvement of predictions in the Elo WWR model, when the actual result is mapped according to Table (2.4) rather than 0, 0.5 and 1, we see that it is reasonable to include the information about the margin of victory. Moreover, the match importance is not incorporated in this model, which also potentially can improve predictions.

The importance of analysis of goals scored is also stressed by good performance of the simple least squares model. In case of this rating method we see that incorporating information about team playing at home improves the model. By extending the model we remove the bias included by the fact that the host of the game usually has an advantage.

In the comparison social network ratings has the lowest accuracy of predications. A possible explanation is that the network of the teams is dense within each of the six world football confederations but there are relatively few connections between different continents. Efforts were made to improve the accuracy of this model.

We note that application of the algorithm in its basic form has some less intuitive properties. When team A beat team B about four years ago and team B beat team C yesterday we count it as an indirect win of team A over C despite the fact that both games are separated with a large time span. Moreover, during four years of play many loops appear in the network, which impose further constraints on the parameter α as λ_{max} tends to be large. Usually the choice of the discount parameter is restricted to a narrow interval $(0, 0.045 \pm 0.002)$. We suggested following modification to the algorithm. We calculate ratings on a monthly basis. More precisely, in each month teams that compete create a network. This network is sparse comparing to that in a 4 year window and only some of the teams are competing. Teams that are not playing receive both a win and a loss score equal 0, and for other teams we compute the ratings as in the basic algorithm. We sum up all the monthly ratings at the span of four years to receive the final ranking of the teams. In this way the algorithm employs its basic intuitions and the range for possible values of the parameter α is bigger. Usually tournaments are played within a short period of time so computations on a monthly basis can capture the outcome of competition more exactly. However, the modification does not help to improve accuracy of the predictions.

The two Markovian ratings are more accurate than the FIFA ranking, albeit not significantly. We also note that there is little difference between calculation of the ratings based solely on win, draw, loss information and goals scored. A possible reason is that in both versions the same graph structure of the teams playing is analyzed. The only difference is in estimation of head-to-head probabilities. The structure of the network is the most important in determination of the stationary distribution of the appropriate Markov chain.

The evaluation Table (4.5) indicates that presumably the ranking points are not awarded in an efficient manner in the official FIFA ranking calculation. We may capitalize on that by choosing appropriate friendly game opponents. This should allow us to climb up the ranking table with limited effort. We turn to this issue in the next chapter.

Chapter 5

Exploiting inefficiencies in FIFA ranking

In previous chapter we saw that the official FIFA ranking is not the best performing rating system. We made predictions on the team points in the ranking and it turned out that other rating systems outperform the official ranking for the teams around the globe. This inefficiency can be used for making jumps by individual teams by appropriate choice of an opponent in a friendly game that we are likely to beat. That kind of strategy might turn out to be profitable. As the official ranks are used for seeding teams in competition draws, by climbing up the ranking we may avoid stronger opposition in preliminary tournament rounds. In this chapter we introduce the approach that allows maximizing the position in the ranking. Before we do so, let us turn to an example, when potentially a team seems to be underrated in the FIFA ranking.

Example - Ukraine

In the May 9th 2012 release Ukraine is placed in the official FIFA ranking at the 50th position (see Table (5.1)). By other methods it is rated higher: for example in Elo WWR (with the FIFA06 prior, where it has the 15th rank) Ukraine occupies the 13th position, in Eloratings.net 26th, by the least squares method it is rated at 26th place and the Elo++ model places Ukraine at the 27th position. What is more, Ukraine met with Estonia in a friendly game on the 28th of May 2012. Estonia, ranked only 4 positions below Ukraine, should be of comparable strength. However, their match ended in a convincing 4:0 victory for Ukraine. Looking at its neighbors in the ranking and their recent results, it seems that their 50th position is not reflecting their actual strength.

One may argue that a low position of both Poland (65th) and Ukraine is the result of lack of high stake games played by those teams in recent time. Because those teams are host nations of the European Championships 2012, they did not participate in qualification rounds to the tournament. This means that recently both teams played mainly friendly matches, which enabled them to gain only few points because of the match importance factor in the official FIFA ranking calculation. Anyway, this example partially explains poor performance of the FIFA ranking when compared to other methods. □

Choice of an appropriate opponent

In this section we propose a few strategies for the choice of an opponent for a friendly match based on different gain functions. We focus for strategy for a chosen team A. Moreover, we

Pos	Team	Points
	...	
45	Romania	603
46	Libya	602
47	Armenia	598
48	Scotland	596
49	El Salvador	591
50	Ukraine	589
51	Jamaica	576
52	Iran	575
52	Panama	575
54	Estonia	574
55	Montenegro	569
	...	

Table 5.1: Places 45-55 in the FIFA May 2012 ranking release.

assume that the ranking is static: the only game that takes place is our match against some chosen opponent T. This assumption is not realistic. In practice, there are scheduled dates in which teams can play friendly games. Usually a number of games are played on that day. Later, we relax this assumption. We also assume that a model for calculating probabilities of a win, draw and loss is given.

We may define the following random variables that describe possible gains for team A when it plays against team T:

- P_T for expressing the gain in ranking points,
- R_T for gain in ranking positions.

The random variable R_T tells us what is the possible jump in the table for a particular choice of team T to play against. For the set of teams in the ranking S, we obtain two families of random variables $\{P_T\}_{T \in S}$ and $\{R_T\}_{T \in S}$. We note that both P_T and R_T can assume negative values, also in case we win the game. This can happen if according to the new ranking release our past games become weighted with a smaller fraction. If we had a very good record exactly two years ago, then in the next ranking issue those games became weighted with 0.3 rather than 0.5, because of the time factor incorporated in the ranking calculation.

We may compute the expectation of the defined random variables. This leads to simple strategies for the choice of opponent. Let F denote a team to play against in a friendly game. Then our choice can be such that $\mathbb{E}P_F = \max_{T \in S} \mathbb{E}P_T$. This means that we motivate the choice by the expected point gain. The other option would be to choose a friendly game opponent such that $\mathbb{E}R_F = \max_{T \in S} \mathbb{E}R_T$. In this case, we wish to play a team that gives us the possibility of the highest expected jump in the table.

We note that averaging can be sometimes misleading. We illustrate it by a simplistic example: suppose it is possible to jump 100 positions up in the ranking with probability 0.01 when playing team A. On the other hand, we may move 1 position with probability 1 when playing against team B. In both cases, the average gain is the same but it is clearly better to choose team B to play against.

Another possibility would be to make our choice dependent on two parameters (r, p). The first parameter specifies the minimal jump in the ranking we want to achieve. The second parameter p denotes the minimal level of probability, that allows us to progress r positions in the ranking. Given the set of teams to play against, our choice of the opponent T is such that it is possible to advance at least r positions in the ranking with probability at least equal to p.

Extension to full schedule of games

So far we assumed that the ranking is static, i.e., we did not take into account any other games. As we already mentioned this is a simplistic assumption. We may relax it by looking at all games that take place on a particular day.

Let us assume that the results of the matches are independent random variables. The ranking table, now dependent on the outcomes of all games, becomes random. Our expected gains in points and rank, $\mathbb{E}P_T$ and $\mathbb{E}R_T$ may be computed as

$$\mathbb{E}P_T = \mathbb{E}(\mathbb{E}(P_T|Ranking)),$$

where we compute the expectation by conditioning on the random ranking table. In similar manner we calculate $\mathbb{E}R_T$.

When we want to plan our strategy dependent on the parameters (r, p), the probabilities p for each team can be computed using the law of total probability. More precisely, when we compute the probability p of advancing 2 positions when playing team T, $\{R_T = 2\}$, we shall take into account how this probability changes depending on the outcomes of other games. This can be done by appropriate conditioning.

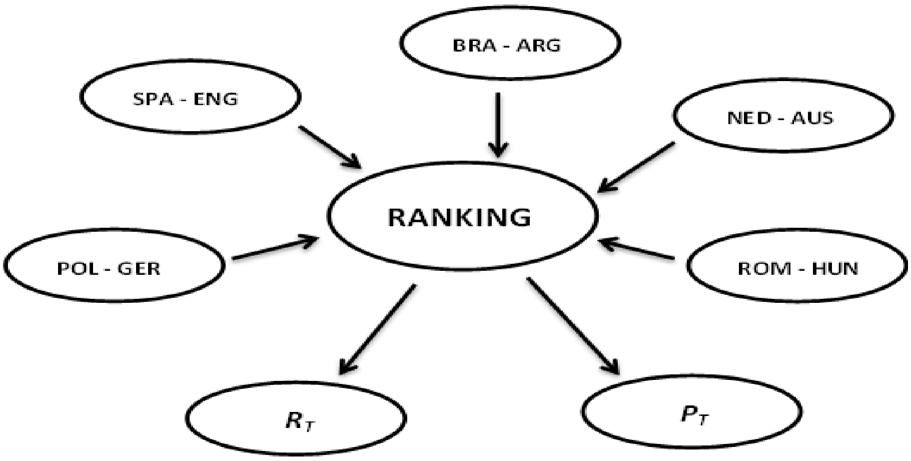

Figure 5.1: Fixture introduces randomness in the ranking. This in turn affects the variables P_R and R_T, that expresses our possible gains. The situation described can be analyzed with the use of a Bayesian network.

Chapter 6

Conclusions

The aim of the work was to present different rating systems and apply them for ranking national teams around the world. We put emphasis on the diversity of the models, their general application in rating and performance. We kept the setting for the models simple. It is natural to extend the models, but the simplicity is also appealing.

In case of graph algorithms it was harder to tailor them to ranking sport teams. The two methods we discussed primarily originate from different domains.

As output of each algorithm, including the official FIFA ranking, for each team we have a single number that should reflect its overall strength. With the use of ratings we are able to make a prediction for a match outcome. Every method is wrong in predicting match results but some of them make considerably less mistakes. It turns out that almost every alternative rating method outperforms the official FIFA ranking in terms of predictive capabilities.

Given the importance of the FIFA ranking and its direct influence on football competition it is surprising that it achieves relatively poor performance. On the other hand, the FIFA methodology used in rating women teams, based on the famous Elo rating system for chess players, turned out to be a very competitive rating method. A possible suggestion would be to apply an analogous procedure in ranking men's national teams.

Future work

Further study of the topic of ranking football teams can go into two opposite directions. First of all, we may suggest a better performing alternative. From a practical point of view, general merits in ratings and accuracy one of the two Elo ratings system seems to be adequate to this purpose. Another possibility would be to optimize the current FIFA ranking procedure. This can be done, for example, by tuning the parameters. We also note that the teams in FIFA ranking are barely penalized by losing games, even if top ranked team loses to the last team in the table. We made some experiments with deducting points for a lost game. However, initial trials were not successful as indicated by the increase in the error measurements.

Just another direction would be to exploit possible inefficiencies in the FIFA ranking. We have seen that the points are not awarded in the most accurate manner by that method. In Chapter 5 we sketch the strategies of how a team can capitalize on that by making an appropriate choice of the opponent to play against. This is just a draft of a possibly deep research topic. A number of interesting questions can be stated about scheduling of games by the teams. Future plan is to develop those ideas in greater detail.

Appendix A

Ranking tables

The tables below present the top 40 teams rated by the methods that we considered in the thesis. Ratings are derived from the whole dataset from years 2006-2012 (all the official international games between 15/07/2006 and 02/05/2012). For each method we provide one ranking that achieved the best performance on the test set with respect to the binomial deviance error measure. We comment briefly on the tables under consecutive headings.

Elo ratings

Table (A.1) presents ranking of the teams derived by the method that FIFA uses for rating teams in women football. Prior ratings were set to ranking points from 12 July 2006 FIFA ranking release. On that day Brazil was ranked first with 1630 points and Spain seventh with 1309 points. Despite the dominant role of Spain in football in recent years, this country does not overtake Brazil in the ranking. This shows the importance in the choice of prior ratings. Both intuition and recent successes of the Spanish team seems to agree that it should be ranked first by whatever rating method we choose.

If we decide to use an equal prior for each team, Spain is ranked first. However, due to little number of games played by some teams this is a rough choice and the teams tend to gain rank that seems to be not reliable in a relative comparison.

Table (A.2) presents the version of the Elo model maintained on the website `Eloratings.net`. Table (A.3) presents the ranking table with ratings derived from the Elo++ model, the winner of the first **Kaggle.com** competition on chess player ratings.

Least squares ratings

The least squares model encourages teams to score many goals. The teams with a large positive goal difference will have a high rank, especially when the goals are scored against strong opposition. Table (A.4) presents rankings derived from estimation of the model with correction for the advantage of the home team. The model indicates that on average the host has an advantage of scoring 0.55 goals more per game. The R^2 coefficient of determination is equal to 0.4819. It means that the method explains around 48% of the variance in the difference of goals scored between the teams.

Network-based ratings

The social network ratings are presented in the table (A.5). As we have seen in Chapter 3, teams around the world create local communities associated with six world football confederations. This rating method particularly suffers because of that reason. The ranking table seems to be reasonable when we look at the teams that belong to same confederation. On a global scale, the method struggles to provide reasonable ranking table.

Markovian ratings

Team rankings derived by the analysis of an appropriate Markov chain are presented in Table (A.6). Head-to-head transition probabilities are computed with the use of goals scored between the teams.

The Markov method for ratings rewards teams that play games against many different opponents like Japan or Mexico. This is particularly rewarding when playing against strong (the highest rated) teams. This is the case for Poland or South Africa. Both teams played top ranked teams in recent years mainly because they were hosts of two major football tournaments: World Cup 2010 in South Africa and European Championship 2012 in Poland, co-hosted with Ukraine.

The property of getting a high rank because of many connections in the network is especially appealing for rating the web pages, as in the PageRank algorithm. It is also intuitive when applied in rating football teams around the world. The information about the schedule is explored and used in determining the ratings points. Usually it happens that at tournaments or in friendly games the competing teams are of comparable strength.

Pos	**Team**	**Points**
1	Brazil	1151
2	Spain	1140
3	Netherlands	1105
4	England	1073
5	Germany	1040
6	Italy	1021
7	Argentina	995
8	France	951
9	Uruguay	947
10	Portugal	946
11	Chile	873
12	Czech Republic	872
13	Ukraine	848
14	Croatia	845
15	Sweden	842
16	Paraguay	823
17	Colombia	817
18	Mexico	811
19	Switzerland	810
20	Denmark	807
21	Ecuador	792
22	Russia	770
23	Cote d'Ivoire	767
24	Romania	748
25	Venezuela	746
26	Greece	745
27	Peru	738
28	Turkey	738
29	USA	733
30	Republic of Ireland	729
31	Cameroon	716
32	Nigeria	712
33	Serbia	706
34	Norway	698
35	Scotland	693
36	Poland	690
37	Ghana	682
38	Egypt	682
39	Bosnia-Herzegovina	681
40	Australia	666
	⋮	

Table A.1: Elo Women's World Rankings.

Pos	Team	Points
1	Spain	2096
2	Netherlands	2051
3	Brazil	2041
4	Germany	2040
5	Uruguay	1991
6	England	1918
7	Argentina	1907
8	Chile	1874
9	Portugal	1872
10	Sweden	1849
11	Italy	1845
12	Croatia	1844
13	Mexico	1837
14	France	1831
15	Cote d'Ivoire	1825
16	Paraguay	1823
17	Australia	1815
18	Russia	1809
19	Denmark	1782
20	Japan	1781
21	Korea Republic	1773
22	Ecuador	1771
23	CzechRepublic	1768
24	Republic of Ireland	1767
25	Colombia	1766
26	Ukraine	1765
27	Greece	1756
28	Norway	1739
29	USA	1729
30	Switzerland	1724
31	Egypt	1724
32	Iran	1719
33	Peru	1712
34	Turkey	1702
35	Serbia	1701
36	Romania	1695
37	Ghana	1691
38	Poland	1687
39	Venezuela	1682
40	Scotland	1670
	⋮	

Table A.2: Eloratings.net.

Pos	Team	Points
1	Spain	3.459
2	Brazil	3.209
3	Netherlands	2.957
4	Germany	2.803
5	Argentina	2.748
6	Uruguay	2.648
7	England	2.543
8	Chile	2.47
9	Korea Republic	2.268
10	Sweden	2.264
11	Cote d'Ivoire	2.23
12	France	2.203
13	Portugal	2.191
14	Paraguay	2.182
15	Australia	2.18
16	Mexico	2.058
17	Croatia	2.02
18	Japan	2.013
19	Colombia	2.004
20	Russia	1.919
21	Ecuador	1.887
22	Italy	1.879
23	Norway	1.811
24	Denmark	1.712
25	Czech Republic	1.699
26	Iran	1.698
27	Ukraine	1.681
28	Peru	1.645
29	Turkey	1.644
30	Venezuela	1.64
31	USA	1.632
32	Greece	1.597
33	Republic of Ireland	1.587
34	Egypt	1.56
35	Ghana	1.535
36	Serbia	1.431
37	Belgium	1.421
38	Iraq	1.419
39	Scotland	1.417
40	Bosnia-Herzegovina	1.413
	⋮	

Table A.3: Elo ++ ratings.

Pos	**Team**	**Points**
1	Brazil	3.51135
2	Spain	3.37631
3	Germany	3.2522
4	Netherlands	3.17168
5	Argentina	2.97543
6	England	2.95238
7	Uruguay	2.93974
8	Portugal	2.65248
9	Mexico	2.6263
10	Croatia	2.53871
11	Cote d'Ivoire	2.53798
12	France	2.39341
13	Sweden	2.37642
14	Chile	2.36588
15	Paraguay	2.34089
16	Italy	2.33266
17	Colombia	2.31331
18	Russia	2.2726
19	Denmark	2.26636
20	Japan	2.23836
21	Turkey	2.23312
22	Serbia	2.22425
23	Norway	2.21527
24	Australia	2.18006
25	Czech Republic	2.16545
26	Ukraine	2.12451
27	Romania	2.12112
28	Ecuador	2.10413
29	USA	2.06983
30	Korea Republic	2.04603
31	Ghana	1.99469
32	Iran	1.97755
33	Greece	1.97435
34	Cameroon	1.95714
35	Egypt	1.95712
36	Switzerland	1.91311
37	Republic of Ireland	1.88587
38	Poland	1.86552
39	Nigeria	1.83253
40	CostaRica	1.78298
	⋮	

Table A.4: Least squares ratings with home advantage.

Pos	Team	Points
1	Spain	102.57
2	Brazil	95.96
3	Iran	78.53
4	Netherlands	77.47
5	Germany	71.34
6	Japan	70.38
7	Argentina	65.8
8	Korea Republic	59.7
9	Saudi Arabia	57.71
10	Egypt	56.27
11	Mexico	56.13
12	Australia	55.49
13	Cote d'Ivoire	51.18
14	Croatia	49.83
15	England	47.84
16	USA	44.74
17	Chile	44.25
18	Uruguay	43.18
19	France	42.28
20	Portugal	36.53
21	Italy	34.95
22	China PR	33.85
23	Oman	33.44
24	Ghana	32.38
25	Russia	30.23
26	Uganda	29.39
27	Nigeria	28.48
28	Honduras	28.46
29	Sweden	28.09
30	Iraq	24.2
31	Denmark	24.15
32	Ukraine	23.84
33	Colombia	23.71
34	Morocco	22.98
35	Greece	22.64
36	Cameroon	22.58
37	Paraguay	22.04
38	Turkey	21.61
39	Norway	20.6
40	Romania	20.58
	⋮	

Table A.5: Social network ratings.

Pos	Team	Points
1	Spain	2.17388
2	Netherlands	2.07598
3	Brazil	2.04056
4	Germany	1.98419
5	Japan	1.62999
6	Argentina	1.53803
7	Mexico	1.53737
8	England	1.46263
9	France	1.38628
10	Portugal	1.35488
11	Uruguay	1.35431
12	Cote d'Ivoire	1.26864
13	South Africa	1.26257
14	Poland	1.24404
15	Egypt	1.23762
16	Sweden	1.22136
17	Italy	1.21027
18	Korea Republic	1.20789
19	Croatia	1.17142
20	Chile	1.16909
21	USA	1.16573
22	Ghana	1.13716
23	Turkey	1.13516
24	Switzerland	1.09205
25	Australia	1.07479
26	Iran	1.07313
27	Paraguay	1.07246
28	Serbia	1.0626
29	Norway	1.06066
30	Romania	1.04996
31	Saudi Arabia	1.02979
32	Nigeria	1.0233
33	Denmark	1.01714
34	Russia	1.00504
35	Czech Republic	0.97774
36	Republic of Ireland	0.97415
37	Ukraine	0.96746
38	ChinaPR	0.95753
39	Greece	0.94907
40	Oman	0.86172
	⋮	

Table A.6: Markovian ratings on goals scored (probabilities multiplied by 100).

Bibliography

[1] S. Brin, L. Page, R. Motwami, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-0120, Computer Science, Stanford University, 1999.

[2] T. Callaghan, P.J. Mucha, and M.A. Porter. Random walker ranking for NCAA division I-A football. *American Mathematical Monthly*, 114:761–777, 2007.

[3] Chessmetrics.net. Chessmetrics Rating System, 2012. `http://www.chessmetrics.com`, Last access date: 15 March 2012.

[4] R.R. Davidson. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. Technical report, FSU Statistics Report Ml69 ONR Technical Report No. 37, 1969.

[5] T. Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15, 2000.

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

[7] EloRatings.net. The World Football Elo Rating System, 2012. `http://www.eloratings.net/system.html`, Last access date: 3 March 2012.

[8] FIFA.com. FIFA Women's World Ranking Methodology, 2012. `http://www.fifa.com/worldranking/procedureandschedule/womenprocedure/index.html`, Last access date: 11 February 2012.

[9] FIFA.com. FIFA/Coca-Cola World Ranking Procedure, 2012. `http://www.fifa.com/worldranking/procedureandschedule/menprocedure/index.html`, Last access date: 29 January 2012.

[10] M.E. Glickman. A comprehensive guide to chess ratings. *American Chess Journal*, 3:59–102, 1995.

[11] M.E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, 48:377–394, 1999.

[12] R. Herbrich, T. Minka, and T. Graepel. Trueskill(tm): A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, 20:569–576, 1999.

[13] Kaggle.com. Chess ratings – Elo versus the Rest of the World, 2010. `http://www.kaggle.com/c/chess/details/`, Last access date: 25 January 2012.

[14] Kaggle.com. Deloitte/Fide Chess Rating Challenge, 2011. http://www.kaggle.com/c/ChessRatings2/details/, Last access date: 25 January 2012.

[15] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

[16] J.P. Kenner. The Perron-Frobenius Theorem and the ranking of football teams. *SIAM review*, 35(1):80–93, 1993.

[17] M.J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[18] K. Massey. Statistical models applied to the rating of sports teams. Master's thesis, Bluefield College, 1997.

[19] R.B. Mattingly and A.J. Murphy. A Markov method for ranking College Football conferences, 2010. http://www.mathaware.org/mam/2010/essays/, Last access date: 26 March 2012.

[20] P. McCullagh and J.A. Nelder. Generalized linear models. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1989.

[21] I. McHale and S. Davies. Statistical analysis of the effectiveness of the FIFA World Rankings. In J. Albert and R.H. Koning, editors, *Statistical Thinking in Sports*, pages 77–90. Chapman & Hall/CRC, Boca Raton, Florida, 2007.

[22] M.J. Moroney. Facts from figures. Penguin, 1956.

[23] J. Park and M.E.J. Newman. A network-based ranking system for US College Football. *Journal of Statistical Mechanics*, 2005(10):P10014–P10014, 2005.

[24] R. Pollard. Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, 1(1):12–14, 2008.

[25] R. Pollard, C.D. da Silva, and C.M. Nisio. Home advantage in football in Brazil: differences between teams and the effects of distance traveled. *The Brazilian Journal of Soccer Science*, 1(1):3–10, 2008.

[26] P.V. Rao and L.L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

[27] A. Seckin and R. Pollard. Home advantage in Turkish professional soccer. *Perceptual and Motor Skills*, 107(1):51–54, 2008.

[28] Y. Sismanis. How I won the 'Chess Ratings: Elo vs the rest of the world' Competition, 2011. http://blog.kaggle.com/2011/02/08/how-i-did-it-yannis-sismanis-on-winning-the-elo-chess-
Kaggle.com blog, Last access date 25 January 2012.

[29] J.C. Spall. Introduction to stochastic search and optimization: Estimation, simulation, and control. Wiley, 2003.

[30] R.T. Stefani. Football and basketball predictions using least squares. *IEEE Transactions on Systems, Man and Cybernetics*, 7(2):117–121, 1977.

[31] H. Stern. Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences*, 23(1):103–117, 1992.

[32] The Free Encyclopedia Wikipedia. Fifa world rankings. `http://en.wikipedia.org/wiki/FIFA_World_Rankings`, Last access date: 10 June 2012.