Elo Rating System for UEFA Women's Euro 2017

The Predictive Power of Elo Ratings for the Performance of Teams and Players in the 2017 UEFA Women's Championship

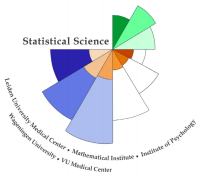
Chang Heng Chen (s1668919)

Thesis advisor: Dr. Joost N. Kok

Second advisor: Dr. Willem J. Heiser

MASTER THESIS

Defended on Month Day, Year



STATISTICAL SCIENCE

FOR THE LIFE AND BEHAVIOURAL SCIENCES

Acknowledgements

I would like to sincerely thank my first supervisor, Joost Kok, the founder of the Dutch Sport Data Center (SDC), for his amazingly administrative supervision and patience. He supported my work, guided me to the right direction, and helped me get results of better quality.

I also would like to say thank you to my second supervisor, Willem Heiser for his time, effort, and patience to instruct me throughout my time spent in writing this thesis project. He suggested certain statistical methods to be used for the analysis of this project and for the ideas on the simulation study of this thesis.

I had a wonderful time working on this project under their supervision. They both broadened my horizon of sports statistics.

Next, I want to thank Arie - Willem, another sport data scientist, for important discussions with regard to the algorithm of this thesis and the results of the statistical analysis.

Lastly, I would like to thank my Dutch family, team Sportja, for your support and knowledge about football tournament. You guys were always there for me whenever I needed to talk about the results of my statistical algorithm that I found.

Abstract

The Elo rating system has been used in various sports / games, such as chess, soccer, tennis and even video games, to calculate the relative playing strengths of players / teams. Originally, the Elo system was invented by a Hungarian physics professor, Arpad Elo, to improve chess rating system. Now many rating systems used in sports are based on the Elo rating system with modifications.

The objective of this thesis project is to examine the Elo rating system for soccer tournaments and how it can be applied to the 2017 UEFA Women's Championship (short for UEFA Women's Euro 2017). More specifically, two primary interests lie in this project.

The first interest lies in determining the strength of each team by assigning an Elo rating to the each competing team after tournament. In addition, it is interesting to see how home-field advantage helped the Netherlands (the host country) win the championship of UEFA Women's Euro 2017 by incorporating the home-field advantage in the Elo formula.

Secondly, strengths of the players of all teams are also of interest. In order to estimate the strengths of the players, each player is assigned a rating (Not an Elo rating) to represent how strong every player is. We can then compare the players among all teams.

In order to access the reliability of our ideas and methodology, a simulation study

will follow after the theoretical part of our research.

In Chapter 1 I will first describe the basic concepts of the Elo rating system. Then a short summary of the relevant literature papers will be presented. Finally I will discuss the source of the data, the arrangement of the tournament, and the process that will take to go through the algorithm / methodology.

In Chapter 2 the basic Elo formula and some modified Elo models are proposed, which allows us later on to determine the most appropriate model for estimating the strengths of every single competing country and the players of all teams. In the end of this chapter, I develop an ordered probit regression model for forecasting match results in UEFA Women's Euro 2017.

Chapter 3 suggests a simulation study for estimating the strengths of all the participant countries of the tournament and the strengths of football players of all teams. Chapter 4 presents the main conclusions drawn from the model computations and suggests some further research of this thesis project.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Without a doubt, a healthy nation is always a wealthy nation. The importance of sports has been addressed throughout the history. Nowadays, with the help of advanced media, the popularity of various sports has grown faster than ever before, and sports has become one of the biggest lucrative businesses in the world. Take football business in Europe for example. Clubs are spending large sums of euros on transfers. Professional athletes and coaches are making a huge profit each year. Although a single game outcome can be significantly influenced by a great deal of factors, such as the quality of players involved and circumstances of the game, to name a few, a single game outcome could reshape the business of a football club. With all these interests, if a system or process with some quantitative property that could fairly help us judge, determine, or predict which player / team is preferred or has a great amount of strength to win a match when comparing athletes / teams in pairs during a sports tournament, would that not be a great thing?

One of the conventional, scientific approaches is to use pairwise comparison, which was first introduced by Thurstone (1927) in his research paper, "A Law of Comparative Judgement". During the method pairwise comparisons, each entity is compared or matched head-to-head with each of the other entities. Such method can easily help

us make decisions that require comparing alternatives (or sports teams in a tournament) regarding a set of rules in order to rank from the most preferred entity (or the strongest player / team) to the least preferred entity (or the weakest player / team).

However, as David (1988) states in his book, "The Method of Paired Comparisons", that in the design of paired-comparison experiments, it becomes more difficult to answer questions of design when collecting information from all possible comparisons is not an option. For example, the soccer tournament data (UEFA Women's Euro 2017) used for this thesis project consists of only 31 matches of 16 national European teams, where the first 24 matches are group stage competitions (within each group, teams are matched in single round-robin arrangement), and the last 7 matches are knockout stage competitions (single-elimination tournament). This data set is therefore incomplete design that we only observe 30 different pairs (see Figure 1.1) instead of all possible 120 comparisons ($\binom{16}{2} = 120$), or even 240 comparisons if each team plays twice against all other teams (one home game and one away game). If we would like to compare the strengths of any two teams that did not play against each other during the tournament, utilizing the method of pair comparison is not feasible. Figure 1.1 shows all the matches that were played during UEFA Women's Euro 2017. The first column and the first row represents the abbreviations of all the 16 teams. All the matches are shown in the upper-right corner where letter A, B, C, and D are group stage competitions, and QF, SF, and F are knockout stage competitions for quarter-finals, semi-finals and final, respectively. For instance, there are 6 games played in group A by 4 countries (NL, DK, BE, and NO), and the final was played by NL and DK.

Moreover, a typical scoring system, three points for a win, is a standard scoring system used mainly in sports leagues, especially in association footballs. This system

	NL	DK	BE	NO	DE	SE	RU	IT	AT	FR	CH	IS	UK	ES	SCT	PT
NL		A / F	A	A		QF							SF			
DK			A	A	QF											
BE				A												
NO																
DE						B	B	B	SF							
SE							B	B								
RU								B								
IT																
AT										C	C	C		QF		
FR											C	C	QF			
CH												C				
IS																
UK														D	D	D
ES															D	D
SCT																D
PT																

Figure 1.1: The matches of UEFA Women's Euro 2017

was first proposed by Jimmy Hill in 1981 for the English Football League in which a team gets 3 points for winning a match, 1 point for a draw, and 0 for losing a game. Although the primary incentive of this system is to encourage more offensive play during a match since teams would not go for a draw if winning two extra points benefits the winning team more than having a draw, it is still far from clear whether this system has succeeded in promoting more exciting football games.

In this section, I will introduce an existing algorithm, known as the Elo rating system, that not only can avoid the problem caused by incomplete design of pairwise comparison, but also implements an alternative, universal scoring system (1 for a win, 1/2 for a draw, and 0 for a loss) across different tournaments that gives the property of zero-sum games. The Properties and advantages of the Elo rating system will be discussed in detail in Section 1.1.1.

1.1 Elo Rating Fomula

The Elo rating system is at present one of the best-known rating systems for calculating the relative strength of players in two - player (or - team) matches. Originally, the Elo system was invented by a Hungarian physics professor, Arpad Elo, to improve chess rating system (see Elo (1978)). Elo assumes that the performance of each player / team is symmetrically distributed with a mean of the player's / the team's true strength. Although a player's / team's true performance could be improved with a lot of sports practice (or worsened without practice), another assumption of the Elo system is the mean value of a players / team's performance (relative strength) only slowly changes throughout the time. Meanwhile, the Elo system rewards a weaker player / team more rating for winning a stronger counterpart than it assigns a stronger player / team for defeating a weaker opponent.

The Elo updating formula is stated as follows. When i plays against j, we have

$$r_i^{\text{new}} = r_i^{\text{old}} + K\left(S_{ij} - E_{ij}\right) \qquad \text{and} \qquad r_j^{\text{new}} = r_j^{\text{old}} + K\left(S_{ji} - E_{ji}\right) \tag{1.1}$$

In the left part of equation (1.1), r_i^{old} is the current (old) rating, and r_i^{new} is the updated (new) rating for i. K is a factor constant, S_{ij} represents the competition result (1 for a win, 0 for a loss, and 1/2 for a draw) for i, and E_{ij} is the expected score for i against j. In the right part, the variables r_j^{old}, r_j^{new}, and S_{ji}, the fixed parameter K, and the expected score E_{ji} are all based on j (against i). In the following, I will outline each of the parameters and variables in equation (1.1) in a more detailed fashion.

The K

The role that the factor K plays here is to regulate the deviance between old and new ratings. If K is set too large, the sensitivity of updating rating will be high, meaning that if a player performs slightly better or worse than expected, his / her rating can change drastically. However, if K is set too small, the Elo updating rating system will not be fast enough to catch up with the true ability of the players. One advantage of the Elo rating system is that it allows an experienced player (a player who has played several matches) could perform somewhat better or worse than his usual performance without causing considerable consequences on the players rating by setting a relatively small K-value than a relatively big K-value for newbies.

The K-factor is often different according to the importance or level of the competition. According to the fact sheet from FIFA (2017a), different K values are being applied in women football matches to indicate the importance of the match. The larger K is, the more important and more competitive the match is; and the smaller K is, the less important the competition is. Table 1.1 shows different K values for the importance of the match.

The Actual Score S

When the competition result, S, is described as 1 for a win, 0 for a loss, and 1/2 for a draw, then this can be mathematically expressed as:

$$S_{ij} = \begin{cases} 1 & if\ i\ wins\ over\ j, \\ 1/2 & if\ i\ and\ j\ draw\ a\ tie, \\ 0 & if\ i\ loses\ to\ j. \end{cases} \tag{1.2}$$

Table 1.1: Different K-values for the level of the competition

Match Importance	K-value
FIFA Women's World Cup match	60
Women's Olympic football tournament	60
FIFA Women's World Cup qualifier	45
Women's Olympic football qualifier	45
Women's Continental finals match	45
Women's Continental qualifier	30
Women's friendly match between two Top 10 teams	30
Women's friendly match	15

Equation (1.2) is based on the match result; however, in Section 1.4.1 I will introduce that the actual score can also be modified based on the scores of i and j.

$$S_{ij}^* = \frac{P_i + 1}{P_i + P_j + 2}, \tag{1.3}$$

where P_i and P_j are the scores of i and j. Both actual score equations indicate that the sum of the actual scores for two teams playing against each other is always one ($S_{ij} + S_{ji} = 1$ and $S_{ij}^* + S_{ji}^* = 1$). For $S_{ij} + S_{ji} = 1$, this is pretty straightforward. For $S_{ij}^* + S_{ji}^* = 1$, the proof is shown in Section 1.1.1.

The Expected Score E

In the Elo formula (see equation (1.1)), E_{ij} represents the expected score or expected winning probability that player i gets when playing against player j, and is usually assumed that it is a logistic function of the rating difference of two players / teams.

Before digging into the mathematical details of the E_{ij} in equation (1.1), I would

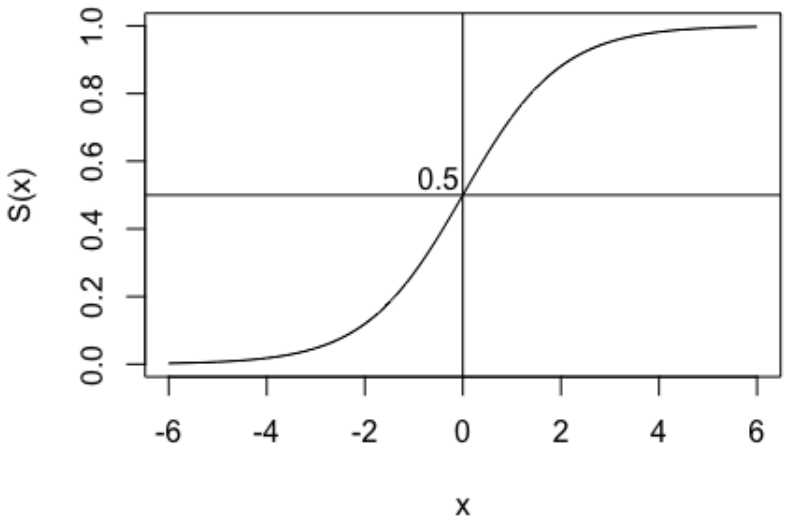


Figure 1.2: The sigmoid curve

like to introduce the standard logistic function $S(x) = 1/(1 + e^{-x})$. It is also named sigmoid function or sigmoid curve because of its s-shaped curve. Sigmoid function is a special case of the logistic function shown in Figure 1.2. Note that when x goes to ∞, $S(x)$ goes to 1; when x goes to $-\infty$, $S(x)$ goes to -1; and when x goes to 0, $S(x)$ goes to 0.5.

When player / team i plays against player / team j, and they have ratings of r_i and r_j respectively, the definition of the expected score E_{ij} (for player i against j) is stated as follows:

$$E_{ij} = \frac{1}{1 + 10^{-(r_i - r_j)/400}}. \tag{1.4}$$

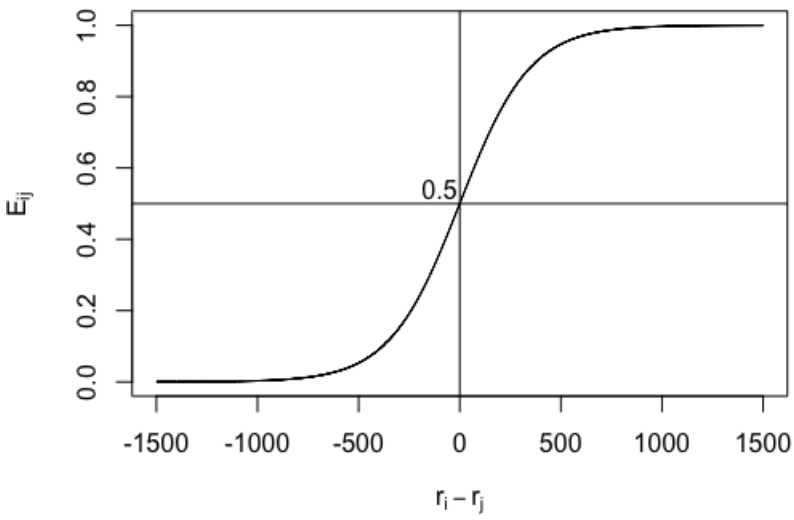


Figure 1.3: The expected score E_{ij}

Note that equation (1.4) is a base-ten version of a logistic function of the difference in ratings of two players; however, in Section 1.4.1 I will introduce that home-field advantage can also be implemented into the expected score function, stated as follows:

$$E_h = \frac{1}{1 + 10^{-(h + r_h - r_a)/400}}, \tag{1.5}$$

where h is the amount of home advantage boost. The constants 10 and 400 in the formula (1.4) come from the empirical distribution of the chess world. Together, they simply imply a change in the scale of the x-axis in Figure 1.2.

In Figure 1.3, it is easy to see that the expected score model (see (1.4)) is a logistic model in terms of the rating differences $r_i - r_j$, and that the constants 400 and 10 in (1.4) only lead to a rescaling of the x-axis.

When two players / teams (i and j) play against each other, if the rating difference of the two players is exceedingly high, then the expected score of the stronger player is 1 (a win), and if it is about zero, the expected score for both players is 1/2 (a draw), and if it is extremely low, then the expected score of the weaker player is 0 (a loss).

Clearly, we must have $E_{ij}+E_{ji} = 1$ because if i wins over j, j loses to i (complementary relationship). It follows that $E_{ji} = 1 - E_{ij}$, so that we have $E_{ji} = 1 - \frac{1}{1+10^{-(r_i-r_j)/400}} = \frac{10^{r_j/400}}{10^{r_i/400}+10^{r_j/400}}$. Therefore, we can express the odds of i winning over j as

$$\frac{E_{ij}}{E_{ji}} = \frac{\frac{10^{r_i/400}}{10^{r_i/400}10^{r_j/400}}}{\frac{10^{r_j/400}}{10^{r_i/400}+10^{r_j/400}}} = \frac{10^{r_i/400}}{10^{r_j/400}} = 10^{(r_i-r_j)/400}, \tag{1.6}$$

and the log-odds as

$$\log_{10} \frac{E_{ij}}{E_{ji}} = \frac{(r_i - r_j)}{400}. \tag{1.7}$$

So if the rating difference of two players is about 400 ($r_i - r_j = 400$), then the probability of the player with a higher rating winning over the other player with a lower rating is expected to be 10 times more likely than the probability of the weaker player beating the stronger player. For instance, if the Elo rating for i is 500 ($r_i = 500$) and for j is 100 ($r_j = 500$), and we plug 500 and 100 into the formula (1.4), we have $E_{ij} = 0.9090$ so that $E_{jj} = 1 - E_{ij} = E_{ij}/10 = 0.0909$.

1.1.1 The Important Properties of the Elo Rule

We are now ready to prove three important properties of the Elo rating system:

1. The amount of rating gained from one team is always equal to the amount lost from the other team, and

2. The sum of Elo ratings of all pairs of teams will always reach a constant value, in spite of how many games have been played, and how many time ratings have been updated.

3. The Elo rating system rewards weaker players more than stronger players for winning or drawing a tie, such that

   (a) a weaker player is rewarded more of the Elo rating for winning over a stronger player than the amount of Elo rating rewarded to the stronger player for beating the weaker player,

   (b) a weaker player is rewarded more of the Elo rating for drawing a tie with a strong player than the stronger player, and

   (c) a stronger player is rewarded less of the Elo rating for winning over a weaker player than the amount of Elo rating rewarded to the weaker player for beating the stronger player.

For property 1, the gain of i against j can be expressed as

$$r_i^{\text{new}} - r_i^{\text{old}} = K\left(S_{ij} - E_{ij}\right). \tag{1.8}$$

Using the fact that $S_{ij} + S_{ji} = 1$ and $E_{ij} + E_{ji} = 1$, we obtain from the right-hand part of (1.8)

$$K(S_{ij} - E_{ij}) = K((1 - S_{ji}) - (1 - E_{ji})) = -K(S_{ji} - E_{ji}), \tag{1.9}$$

which is equal to the negative gain of j against i.

For property 2, we find

$$r_i^{\text{new}} + r_j^{\text{new}} = r_i^{\text{old}} + r_j^{\text{old}} + K(S_{ij} - E_{ij}) + K(S_{ji} - E_{ji}), \tag{1.10}$$

and the last two items at the right side of (1.10) vanish because of (1.9) (therefore, $r_i^{\text{new}} + r_j^{\text{new}} = r_i^{\text{old}} + r_j^{\text{old}}$).

For property 3a, suppose that i is stronger than j according to their current (old) ratings before a match so that $r_i^{\text{old}} > r_j^{\text{old}}$, which implies

$$\begin{matrix} -(r_i^{\text{old}} - r_j^{\text{old}})/400 < 0 \\ -(r_j^{\text{old}} - r_i^{\text{old}})/400 > 0 \end{matrix} \Longrightarrow 10^{-(r_i^{\text{old}} - r_j^{\text{old}})/400} < 10^{-(r_j^{\text{old}} - r_i^{\text{old}})/400}. \tag{1.11}$$

Adding 1 to both sides of (1.11) and then dividing them with 1, we find

$$\frac{1}{1 + 10^{-(r_i^{\text{old}} - r_j^{\text{old}})/400}} > \frac{1}{1 + 10^{-(r_j^{\text{old}} - r_i^{\text{old}})/400}} \Longrightarrow E_{ij} > E_{ji}. \tag{1.12}$$

Now if j wins over i (the weaker wins), the updated rating for j becomes

$$r_j^{\text{new}} = r_j^{\text{old}} + K(1 - E_{ji}) \tag{1.13}$$

while if i wins (the stronger wins), the updated rating for i becomes

$$r_i^{\text{new}} = r_i^{\text{old}} + K(1 - E_{ij}). \tag{1.14}$$

From (1.12) (and a fixed, positivie K), we can conclude that $K(1-E_{ji}) > K(1-E_{ij})$. The proofs for 3b and 3c are skipped here since they are similar to the proof of 3a.

Finally, it is of interest to note that all these properties are valid regardless of the

specific model used for the relationship between E_{ij} and $r_i - r_j$. Any probability model as a function of the difference in ratings has these properties. In this thesis project, the expected score function E_{ij} is a simple logistic function (the Rasch model), yet some people use the Thurstone model (see Thurstone (1927)) for E_{ij} where it is the cumulative normal distribution. The difference between these two is (exceedingly) small.

1.2 Literature Study

A detailed Elo rating formula algorithm and history are introduced by Glickman (1995). The main focus is the Elo rating system applied in different chess associations. The author also talks about the property of each parameter in the Elo formula from the point of view of chess world.

Aside from using the Elo rating formula to estimate the strengths of players / teams, most people are also interested in forecasting match results. Goddard (2005) applied two approaches of predicting and modelling match outcomes in soccer associations. The first is bivariate Poisson regression, which is used to estimate predicting models based on goals scored and conceded. The second approach is ordered probit regression for estimating forecasting models based on match results. A 25-year longitudinal data set (25 seasons) on English league football match outcomes is used to estimate both types of models.

The popularity of sports betting has soared over the past decade. Nowadays, sports betting business alone is a multibillion-dollar industry in many developed, industrialized countries. Leitner et al. (2010) compared two methods to evaluate the strengths / abilities of sports participants and their corresponding probabilities of winning are

applied to compare their predictive performance. First the Elo ratings are augmented by a simulation approach producing probabilities of winning for the full tournament. Second, under a consensus model, tournament winning probabilities based on bookmakers odds are calculated. Both methods are applied to predict the results of UEFA EURO 2008. The conclusion of their case study was that the method based on bookmakers odds outperformed the method based on the Elo rating.

It seems apparent that prediction of competition results in association soccer is popular topic in sports statistics. Hvattum and Arntzen (2010) suggested two Elo based prediction methods and Six benchmark prediction methods to examine ratings assigned to teams in order to forecast game outcomes in association soccer. The two Elo based methods are to use an ordered logit regression model (Greene (1999)) with one single covariate, rating difference, to make match results predictions. The ratings are generated under the two Elo based methods, $\mathbf{ELO_b}$ and $\mathbf{ELO_g}$, which are the basic Elo formula and goal based Elo formulate respectively. The the basic Elo formula (from Leitner et al. (2010)) is the same as equation (1.1) with (1.2) and (1.4). The goal based Elo formulate is also the same except the K-factor is modified as $K = K_0(1 + \delta)^\lambda$, where δ is the absolute difference in the goals of two players / teams, and $K_0 > 0$ and $\lambda > 0$ are the fixed parameters (K_0 and λ can be obtained using mathematical optimization). **UNI** is the first of the six benchmark methods. It ignores any information and gives an equal predicted probability of 1/3 for each possible outcome (home win, draw, away win). The second is **FRQ** in which the predicted probability is calculated based on the observed frequency of each outcome (home win, draw, away win). The next two approaches, $\mathbf{GOD_b}$ and $\mathbf{GOD_g}$, are derived from ordered probit regressions to generate predictions for football match results. The covariates that are used are based on only the past match results. 50 results based on team performance covariates in model 4 of the paper, "Regression

models for forecasting goals and match results in association football", by Goddard (2005) are used for $\mathbf{GOD_b}$ while 100 goal based team performance covariates in models 3 of the same paper are used in $\mathbf{GOD_g}$. The last two methods, **AVG** and **MAX** are based on the odds from a set of bookmakers. The difference is that in **AVG** the average odds from the bookmakers are considered while the maximum odds for each home win, draw and away win are applied in **MAX**.

The science of sports rating and ranking plays an important role in sports world. It was not until Langville and Meyer (2012) that an overall comprehensive overview was given of the methodology and mathematical algorithms applied for sports rating and ranking in their compendium, "Who's #1?: The Science of Rating and Ranking". Chapter 5 gives a comprehensive history and overview of the mathematical background of Elo rating formula. Two main special properties, constant sums and the K-factor are presented in there. This chapter of the book also provides a modified Elo rating formula where home-field advantage boost, various K-factors based on the importance of the match, and game scores differences are incorporated. A case study is also presented in the end of the chapter, Elo in the NFL (the National Football League) where readers can easily understand how the authors processed this study via the basic Elo formula and the modified Elo scheme, and how they measure the prediction accuracy to compare the results.

The impact of certain parameters in the Elo rating formula has been evaluated in various studies. Sullivan and Cronin (2016) provided the examination of the Elo rating system and its application to the English Premier League to forecast the match result. In their study, they investigate four different methods of modifying the basic Elo scheme: incorporating home-field advantage, adapting variable K-factor at different level in a season, rewarding and penalizing the winning and losing streaks, and

incorporating game scores (rewarding a win proportionally to the margin of the win). These additional parameters are expected to increase in prediction accuracy rate over the basic Elo system.

The fundamental science in sports rating and sports betting is probability theory. While Elo ratings and the sports Model had been a neglected topic in applied probability, Aldous (2017) was the first to look at the Elo ratings from the point of view of probability theory. He proposes a broad spectrum of research questions suggested under a model that the probability of player i beats j is a set function of their difference in strength. This can be mathematically shown as $P(A > B) = W(x_A - x_B)$. For the win probability function W satisfies:

$$\begin{cases} 0 \leq W(x) \leq 1, \\ W(-x) + W(x) = 1, \text{ and} \\ \lim_{x\to\infty} W(x) = 1. \end{cases} \tag{1.15}$$

A common choice for W would be the logistic function. In addition, he also focuses on using Elo-type rating system to track changing strengths (in a long run) since it is realistic to think that the mean of the performance of a player could change over time.

1.3 Research Questions

The UEFA European Women's Championship (abbreviated: the UEFA Women's Euro) is the female soccer tournament for the women's national teams under the Union of European Football Associations (UEFA). The competition takes place every four years, and is commensurate with the European cup, except that it is only for

the female soccer teams. In 2017, the tournament was the 12th edition, extended to 16 qualified teams with 31 matches played, and was hosted by the country of the Netherlands in July and August.

At the end of the tournament, the Netherlands won the championship for the first time ever, after winning Denmark in the final.

The prime goal of this research thesis will center on applying the Elo rating system to the data set of the 2017 UEFA Women's Championship in order to estimate and monitor the strengths of both teams and players as the tournament goes along.

This thesis first investigates the team strengths by assigning ratings to teams according to their match performance in this tournament in order to understand how strong each team was in the tournament. Moreover, since the Netherlands was the host country, one of the main interests is to observe how significantly home-field advantage helped the Netherlands (the host country) win the championship.

Next, another research interest is to access the strengths of players of all teams. For the purpose of estimating the strengths of players, all players are assigned values of ratings based upon their performance in each event and the result of each match. A comparison of the strengths of all players can be then examined afterward.

After estimating the strength of each team and each player based on their performance in the 2017 UEFA Women's Championship, a simulation study will be carried out to access the reliability of the Elo rating estimation for the teams and the players.

In the next section, a detailed description and process will be explained on how to

use the Elo rating formula to generate the ratings of the 16 teams and how to use those ratings to estimate the strengths of players.

1.4 Methodology

The algorithm structure of this thesis project consists of 3 main parts: Elo ratings for the strengths of teams, ratings for the strengths of players of all teams, and simulation for the reliability of the estimations of team and player strengths.

1.4.1 Team Elo Ratings

To calculate the Elo ratings for all 16 teams, we need to fill in the Elo updating formula (see (1.1)) with: initial ratings r (variable), the K-factor (fixed parameter), the actual score S (see (1.2)), and the expected score E (see (1.4)).

The initial rating r can be set with different values since the Elo ratings represent the relative strength levels of teams in a competition. Here 2 scenarios of the initial rating r are considered: 1000 for all teams, and their FIFA / Coca-Cola world ratings before this tournament. The initial rating 1000 is set under the assumption that we do not have any information about the strength of all the teams before the tournament and therefore, all the teams are assumed equally strong before the tournament. The second scenario is to implement the information of all the participant teams prior to the tournament into our algorithm.

The formula for the outcome (the actual score) S (see (1.2)) is based on the game result (win, draw and loss). Langville and Meyer (2012) suggest in chapter 5 of their book that the outcome (the actual score) S defined in (1.2) can be modified based on

the goal score difference of two players or two teams (i and j).

$$S_{ij}^{*} = \frac{P_i + 1}{P_i + P_j + 2}, \tag{1.16}$$

where P_i and P_j are the number of goals scored by team i and team j. Note that S_{ij}^{*} still have the constant sum property ($S_{ij}^{*} + S_{ji}^{*} = \frac{P_i+1}{P_i+P_j+2} + \frac{P_j+1}{P_i+P_j+2} = 1$).

The formula for the expected score E ((1.4)) is a function of the difference in ratings of team i and team j. However, many authors (such as Sullivan and Cronin (2016)) recommend that the home-field advantage can be incorporated in the equation (1.4) by adding some additional points to the rating of the home team, and the expected score formula with home advantage parameter, h, can be stated as:

$$E_h = \frac{1}{1 + 10^{-(h+r_h-r_a)/400}}. \tag{1.17}$$

The amount of home advantage boost h is a constant where if it is a positive value the winning probability (expected score) of the home team E_h becomes higher than the expected score calculated without h. If a match is taken place on a neutral territory, h is set to 0, and then formula (1.17) is exactly same as (1.16). Note that the sum of the expected scores of home team and away team is still 1. Suppose that r_h and r_a in (1.17) are the ratings of home team and away team before a match. While the expected score for home team is expressed in (1.17), the expected score for away team is $E_a = \frac{1}{1+10^{-(r_a-h-r_h)/400}} = 1 - E_h$.

Now to implement the Elo ratings, it is important to choose the inherent parameters K and h to keep the updated ratings r^{new} as precise as possible. For instance, if K is set too small, the calculated, updated rating r^{new} will be primarily determined by the experience or old rating r and will not catch up the recent development of current

strength of a team fast enough. Conversely, if K is set too large, a lot of weight will be on the results of the most recent competitions (because of the term $(S-E)$ in equation (1.1)) which means that the results of the most recent competitions will be more dominant than the past experience (the old rating r), and the Elo updating rating system will be either highly overestimate and underestimate the current strengths of teams. The same principle for the home-field advantage boost h in equation (1.17). If h is set too high , the expected winning probability (the expected score) for the home team is estimated higher than what it should be. However, if it is set too low the expected winning probability for the home team is underestimated. In practice, the K-factor and the home-field advantage boost h can be estimated under mathematical optimization method (see Section 2.2).

To measure the accuracy of ratings, certain criterion for the model prediction accuracy needs to be well-defined in this research project. Fortunately the term $(S - E)$ in equation (1.1) can be used for prediction error. This term is the difference between the competition actual results and the expected score (or the expected results) calculated by the difference in ratings of two match teams. The smaller this term means better the prediction.

The prediction (squared) error then can be defined as follows:

$$\begin{aligned} PSE(i,j) &= \sum_{m=1}^{M} (S_{ijm} - E_{ijm})^2 + (S_{jim} - E_{jim})^2 \\ &= \sum_{m=1}^{M} (S_{ijm} - E_{ijm})^2 + [(1 - S_{ijm}) - (1 - E_{ijm})]^2 \\ &= \sum_{m=1}^{M} (S_{ijm} - E_{ijm})^2 + (E_{ijm} - S_{ijm})^2 \\ &= \sum_{m=1}^{M} 2(S_{ijm} - E_{ijm})^2, \end{aligned} \tag{1.18}$$

where i and j are two match teams from the 1st match to the Mth match, and S_{ijm} and E_{ijm} are the actual score and the expected score, respectively, for i against j at match m (also $S_{jim} = 1 - S_{ijm}$ and $E_{jim} = 1 - E_{ijm}$). PSE is also known as cost function or Brier score proposed by Brier (1950) in the context of forecasting the weather. So the basic idea is to get the optimal parameters K-factor and home-field advantage boost h by the lowest prediction squared error under the Elo updating formula (1.1) with 4 different settings:

- the initial rating 1000 for all teams and the actual score S based on the results of match (see (1.2))
- the initial rating 1000 for all teams and the actual score S^* based on the difference in scores of two match teams (see (1.16))
- taking the FIFA / Coca-Cola world ratings of teams (FIFA (2017b)) before this tournament as thier initial ratings and the actual score S based on the results of match (see (1.2))
- taking the FIFA / Coca-Cola world ratings of teams (FIFA (2017b)) before this tournament as thier initial ratings and the actual score S^* based on the

difference in scores of two match teams (see (1.16))

1.4.2 Player Ratings

Once we compare the PSEs calculated from the 4 different models descirbed in the end of Section 1.4.1, we can then find the optimal Elo rating formula to update the ratings of each team after every match.

A team's Elo rating is a certain number that can go up, decrease, or sometimes remain the same depending the result of each match. After two opposing teams finish a match, the winning team takes certain amount of ratings from the losing counterpart. The amount is based on the old ratings of the two teams (for the E), the game outcome (the S) and the K-factor.

So to assign ratings for players, it seems reasonable to share the amount that a team gains or loses after a match with all participant players. For example, when a team wins a match, the Elo rating of the team increases, this increasing amount can be seen as a cake as a reward that is going to be shared with the players of the winning team. The player who performed the best during the match will get the biggest portion of the cake (in this case it means rating points), and vice versa, the player who did not perform so well will get only a little portion.

To determine players with best and worst performance for a match, an event data set for UEFA Women's Europ 2017 will be used for that. The data set contains the ratings of all participating players, based on their performance, in each of the 31 games in UEFA Women's Euro 2017. The event data set was collected under the method, STARSS: A Spatio-Temporal Action Rating System for Soccer by Decroos

et al. (2017), in a project[1] between the Dutch Sport Data Center (SDC (2017)), the Royal Dutch Football Association (KNVB (2017)), and the sports data analytics center, Sportinnovator (2017). In Section 2.4 a more detailed description for the STARSS approach will be presented.

1.4.3 Simulation

After finding out the best Elo updating formula and using it to calculate the team Elo ratings and the player ratings for this tournament, we can then simulate the whole tournament multiple times based on the team Elo ratings and the ratings of players. A more detailed explanation on how the tournament was arranged and how to decide the simulation match results will be presented in the next two chapters (Chapter 2 and Chapter 3).

[1] https://mediator.zonmw.nl/mediator-25-september-2017/wetenschap-van-de-teamsport/?utm_medium=email&utm_campaign=mediator&utm_content=mediator25&utm_source=nieuwsbrief

Chapter 2

Elo Rating Models and Ordered Probit Regression

2.1 Data Source

The data of the 2017 UEFA (The Union of European Football Associations) Women's Championship comes from the official website, Women's football - UEFA.com (see UEFA (2017a)). This year the UEFA Women's Euro 2017 was the 12th edition of the UEFA Women's Championship. The tournament was held in the Netherlands, and it consisted of 31 matches of 16 national European teams, where the first 24 matches were group stage competitions, and the last 7 matches belonged to knockout stage.

2.1.1 Group Stage

For the group stage competition, 4 groups (Group A, B, C, and D) of 4 teams are formed where each group consists of one team randomly drawn from each one of the 4 seeding pots (see Table 2.1). Which seeding pot a team is in is based on their ranking after the qualifying group stage competition[1]. Because the Netherlands is the host country of the tournament, it is placed to pot 1 in the draw. After the final draw (the

[1]The UEFA Women's Euro 2017 qualifying is outside of the research topic. For information about the qualifying, see UEFA (2015).

Table 2.1: The four seeding pots for the final draw

Pot 1	Pot 2	Pot 3	Pot 4
Netherlands	Norway	Italy	Austria
Germany	Sweden	Iceland	Belgium
France	Spain	Scotland	Russia
England	Switzerland	Denmark	Portugal

arrangement of the group stage competition), each group was played in a way where each team played against each other once only (round-robin tournament). Therefore 6 games were held within each group, and in total of 24 matches during group stage competition. Note that the match result during the group stage competition can be a draw for both teams in a game. The group ranking is determined by points: 3 for a win, 1 for a draw, and 0 for a loss. For example, Netherlands won all 3 matches during the group stage competition, so it scored 9 points; Germany had 2 wins and 1 draw, so Germany scored 7. However, if any of the two teams in each group have the same points, certain tiebreaking criteria[2] will be applied to decide the rankings. Table 2.2 shows the result of the final draw for group stage competition and the competition result.

2.1.2 Knockout Stage

The group winners (Netherlands, Germany, Austria, and England) and runners-up (Denmark, Sweden, France, and Spain), teams who came second in each group, moved to the quarter-finals. 4 games were held during the quarter-finals in a way where the first place in group A was against the second place in B, and the second place in

[2]For more details for the tiebreaking criteria, please check the articles 19.01 and 19.02 of the regulations of the UEFA European Women's Championship 2017 (see UEFA (2017b)).

Table 2.2: The final draw and group stage competition results

Ranking	Group A	Group B	Group C	Group D
1	Netherlands	Germany	Austria	England
2	Denmark	Sweden	France	Spain
3	Belgium	Russia	Switzerland	Scotland
4	Norway	Italy	Iceland	Portugal

group A was against the first place in B. The same arrangement was also for group C and D (see Table 2.3). After the quarter-finals the winners then advanced to the semi-finals. 2 games were held during the semi-finals where the winner in the first quarter-final match, 1st place in group A v.s 2nd place in group B, was against the winner in the third quarter-final match, 1st place in group D v.s 2nd place in group C, and the winner in the second quarter-final match, 1st place in group B v.s 2nd place in group A, was against the winner in the fourth quarter-final match, 1st place in group C v.s 2nd place in group D (see Table 2.4). The winners of the semi-finals were then for the final match (see Table 2.5). Therefore, a total of 7 matches were held during knockout stage competition. Note that the match results of the quarter-finals, the semi-finals, and the final can not be a draw for both teams. If the final scores of both teams are equal, such as 0 - 0, 1 - 1, and so on, for example, the winner and loser of the game are determined by penalty shots. The higher successful penalty shots a team has, wins. Because of this rule for team draw, the simulation method for group stage competition would differ from the one for knockout stage.

For more information about the data set and the whole tournament arrangement, please check the regulations of UEFA Women's Euro 2017 (see UEFA (2017b)) and the match results published on the UEFA offical website (see UEFA (2017a)).

Table 2.3: UEFA Women's Euro 2017 quarter-finals

Quarter-finals	
Winner	Loser
Netherlands	Sweden
England	France
Denmark	Germany
Austria	Spain

Table 2.4: UEFA Women's Euro 2017 semi-finals

Semi-finals	
Winner	Loser
Netherlands	England
Denmark	Austria

Table 2.5: UEFA Women's Euro 2017 final

Final	
Winner	Loser
Netherlands	Denmark

2.2 Model Selection

As mentioned in Section 1.4.1, we have four candidate models coming from the Elo updating formula (see (1.1)) with two different initial ratings and two actual score functions to try to get the optimal parameters K-factor and home-field advantage boost h by minimizing PSE (prediction squared error, see (1.18)). To start the Elo updated rating algorithm, an initial value is necessary. Because Elo ratings act as the relative strength levels of teams in a competition, and during the updating process the sum of the ratings of all match teams do not increase or decrease, only their points are exchanged (the amount of one team gains is the amount of the other team loses), it should not matter whether the starting rating chosen to be a high or a low number. In this research experiment, two initial ratings will be carried out to find the optimal parameters.

The first initial rating is to be set for all teams is 1000. This is to assume that all teams are equally good before the tournament. Although this assumption is not ideal nor optimal, the ratings will be updated to better represent the true strengths of the teams after a few matches have been played.

The second one is to use the FIFA / Coca-Cola world ratings of teams before this tournament as the initial ratings to start the Elo rating computation. The ratings were posted on the FIFA official website [3] on 2017-06-23. Table 2.6 shows all the qualified teams for the 2017 UEFA Women's Championship and their FIFA world ratings preliminary to the competition.

Aside from setting the initial rating, we also have defined two actual competition

[3]For the FIFA Women's world ratings published 23 June 2017, please check FIFA (2017b).

score functions S and S^*. Presumably, using the competition outcome function S^* would give us a lower PSE than using S since S gives us only the information of a match result, and S^* provides us more, the information of a match result and the scores of two competing teams.

With each combination of initial rating and outcome score function set-ups, we want to find the parameter K in (1.1) and the parameter h in (1.17) by getting the lowest PSE. During the process of model selection, we will use the root-finding algorithm, Brent's method (see Brent (1971)), to find the optimal K and h.

Table 2.6: Qualified teams for UEFA Women's Euro 2017 and their FIFA ratings before the tournament (published on 2017-06-23)

Team	FIFA women's rating before the tournament	Ranking
Germany	2111	2
France	2076	3
England	2024	5
Sweden	1956	9
Norway	1924	11
Netherlands	1918	12
Spain	1885	13
Denmark	1872	15
Switzerland	1858	17
Italy	1841	18
Iceland	1829	19
Scotland	1788	21
Belgium	1756	22
Austria	1746	24
Russia	1738	25
Portugal	1590	38

Initial Rating 1000 & Actual Score S

With the actual score function S (1.2), the expected score function E_h (1.17), and initial rating 1000, the Elo updating formula (1.1) is applied on the results of the group stage competition from the July 16th 2017 to the July 27th 2017, a total of 24 games. The reason only the group stage competition data (24 matches) were applied on the rating formula instead of the whole tournament (31 matches) was because the match results during the group stage competition could be a draw for both teams in a game. However, the match results during knockout stage competition could not be a draw for both teams.

In the first step, when the home advantage boost h was set to 0 (leave the home-field advantage factor out of the expected score formula) and the K-factor was set to 45 (suggested by Table 1.1), the PSE was calculated to be 10.26.

Then different values of the K-factor from 80 to 120 in increments of 1 were thrown into the updating computation, followed by the plot of prediction squared error (PSE) against the K-factor. We can clearly see from Figure 2.1 that the optimal K is 100.1, obtained by Brent's method (see Brent (1971)), and the PSE is 10.16.

Using the same strategy and the same optimization algorithm (Brent's method, Brent (1971)), with the K we just found ($K = 100.1$), we see that when the home-field advantage H is -7403.25 (see Figure 2.2), it is optimal with PSE, 9.76. However, this number -7403.25 does not seem realistic and reliable. When we later inspect the plot of prediction squared error (PSE) against home-field advantage H with a broader range (see Figure 2.3), we observe that when H passes certain positive or negative levels, the prediction squared error (PSE) converges to about 10.72 or 9.76. Hence, here finding the home advantage boost H giving the lowest PSE is not feasible.

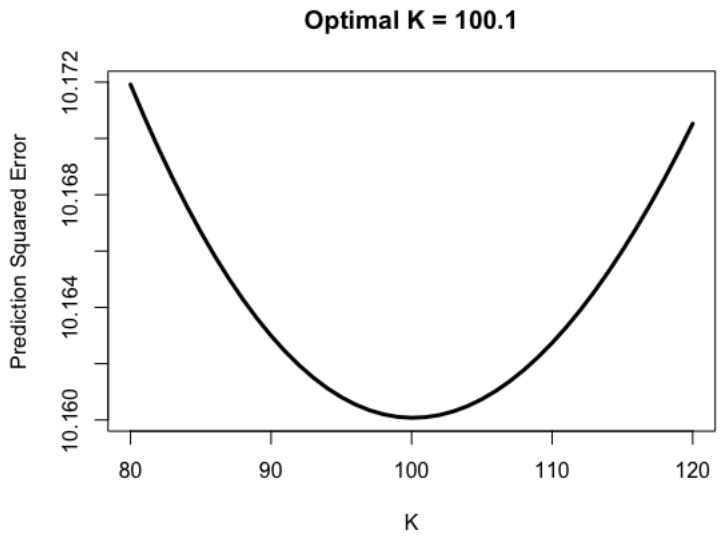


Figure 2.1: PSE v.s K given initial rating 1000 and S

In the next step, we tried to find the optimal K-factor and the home-field advantage H independently by throwing a number of possible combinations of different K and H values. Here we found that the best pair of K and H is (133.99, -3086.99) respectively (see Figure 2.4) with only slightly lower PSE, 9.72.

All the K values we found so far are way too big (even way bigger than all the possible options in Table 1.1). Also all the extreme negative H values we found do not really make sense meaning that the home team has a (way) smaller chance of winning because it is the host country than the away team. This does not add up at all.

Table 2.7 shows the prediction squared errors calculated under the optimal parameters K-factor and the home advantage H given initial Rating 1000 and the actual score function S .

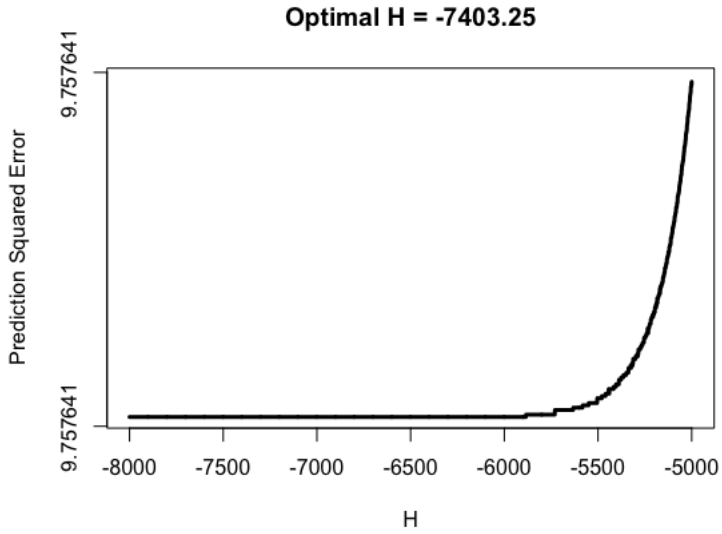


Figure 2.2: PSE v.s H given initial rating 1000, S, and $K = 100.1$

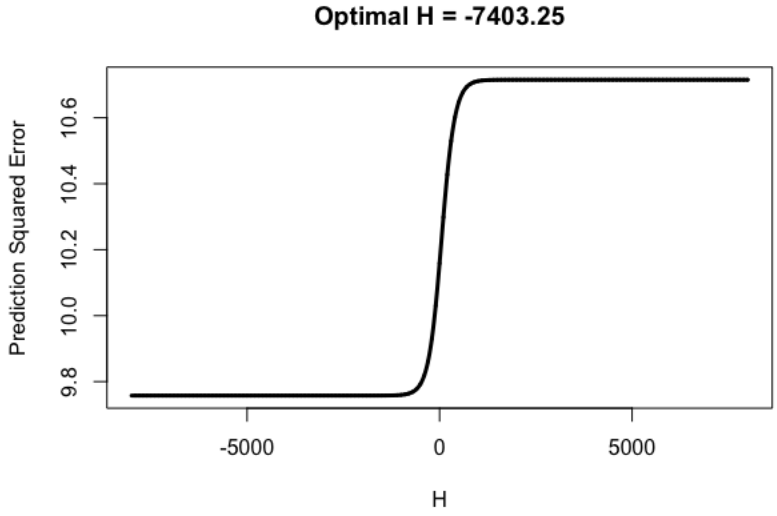


Figure 2.3: PSE v.s H given initial rating 1000, S, and $K = 100.1$

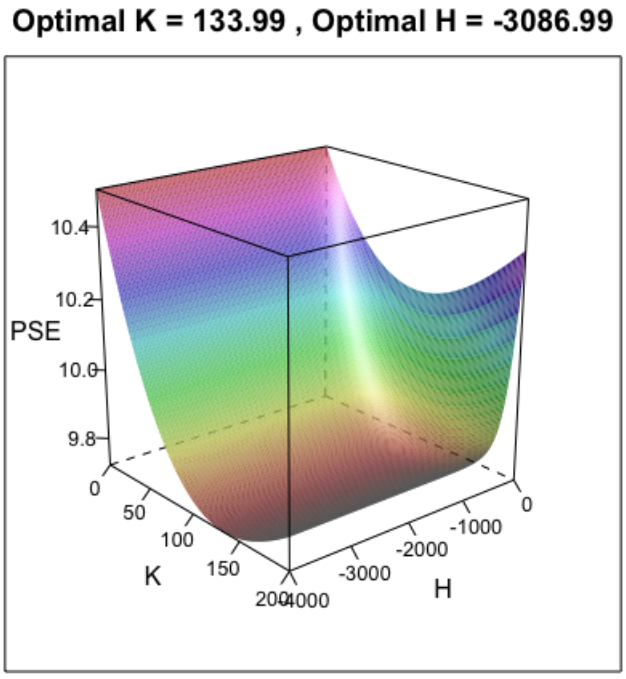


Figure 2.4: PSE v.s K and H given initial rating 1000 and S

Table 2.7: Optimal parameters K and H by using initial Rating 1000 & S

Initial rating 1000 & S	PSE
$K = 45$ (according to FIFA) & $H = 0$	10.26
Optimal $K = 100.1$ (given $H = 0$)	10.16
Optimal $H = -7403.25$ (given $K = 100.1$)	9.76
Optimal $(K, H) = (133.99, -3086.99)$ (grid search)	9.72

Initial Rating 1000 & Actual Score S^*

Now we replace the actual score function S based on only the results of matches with S^* based on the scores of two competing teams (see (1.16)), and go through the same process over again to find out the calculated PSE and the optimal parameters K and H.

We first set the parameter K as 45 suggested by Table 1.1 and H as 0, and then found the PSE was 1.55. Next, trying to find the optimal K given $H = 0$), we threw several values of K into the computation system and figured out which K provided us the smallest PSE. Here the optimal K was found to be 124, and the related PSE was 1.51. We then set K as 124 and tried to find the optimal H that would give us the lowest PSE by trying different values of H into the system. We found that the optimal H was -492.14 given a fixed K (=124) with PSE 1.42. Lastly, we tried to found the two parameters by grid search meaning that we threw all possible joint points of pairs of K and H in the system and found the best pair. The best pair

was found to give 1.42 of PSE with 117.6 and -571.5 for the K-factor and the home advantage boost H respectively (see Table 2.8).

The optimal values of K are still very large, and the optimal H are still very extremely small. However, the prediction squared error has decreased quite a lot (if we compare Table 2.7 and Table 2.8) since we utilized the actual score function S^* that provided us more information about the outcomes of the matches.

In the next two scenarios, the initial rating will no longer be 1000, but the teams' FIFA world ratings before the tournament (see Table 2.6) will be used as starting values of the Elo updating system.

Table 2.8: Optimal parameters K and H by using initial rating 1000 & S^*

Initial rating 1000 & S^*	PSE
$K = 45$ (according to FIFA) & $H = 0$	1.55
Optimal $K = 124$ (given $H = 0$)	1.51
Optimal $H = -492.14$ (given $K = 124$)	1.42
Optimal $(K, H) = (117.6, -571.5)$ (grid search)	1.42

FIFA World Ratings & Actual Score S

With the actual score function S (1.2), the expected score function E_h (1.17), and the teams' FIFA world ratings before the tournament, the Elo updating formula (1.1) is

once again applied on the results of the group stage competition from the July 16th 2017 to the July 27th 2017, a total of 24 games.

By setting the FIFA world ratings of teams as the initial ratings, we do not assume that all teams are equally strong at the start of the event anymore. We expect that with this bit of knowledge the prediction will improve, and the updating system will become more representative.

The main task here is still to discover the optimal parameters K and H by the optimization of slowest prediction sqaured error. Once again our first attempt was to set the parameter K as 45 (suggested by Table 1.1) and H as 0, and then the PSE was found to be 9.62. Following that we attempted to find the optimal K given $H = 0$. We used the same method of optimization by putting many values of K into the computation system and figured out which K provided us the smallest PSE. Here the optimal K was found to be 101.1, and the related PSE was 9.62. Using the optimal K we tried to find the optimal H that would give us the lowest PSE by trying different values of H into the system. We found that the optimal H was -7403.25 with PSE, 9.33. Our last attempt was to find the two parameters by grid search. We threw all possible joint points of pairs of K and H in the system and found the best pair. The best pair was found to give the PSE, 9.27 under the parameters $K = 154.12$ and $H = -2944.06$ (see Table 2.9).

FIFA World Ratings & Actual Score S^*

In this scenario, we expect to get the best result since here we are using the knowledge of strengths of all teams at the start of the tournament and the more informative actual score function S^*.

Table 2.9: Optimal parameters K and H by using FIFA world ratings & S

FIFA wolrd rating & S	PSE
$K = 45$ (according to FIFA) & $H = 0$	9.68
Optimal $K = 101.1$ (given $H = 0$)	9.62
Optimal $H = -7403.25$ (given $K = 101.1$)	9.33
Optimal $(K, H) = (114.12, -2944.06)$ (grid search)	9.27

Table 2.10 shows the prediction squared errors calculated with the optimal parameters K-factor and the home advantage H under the condition of using FIFA world ratings of teams as initial ratings and the actual score function S^*. Surprisingly, the result seems fairly good yet not best possible in terms of PSE (compared to Table 2.8). This does not reflect what we initially expected.

Based on results, the actual score function S^* does help improve the prediction. First of all, if we calculate the variances of the PSEs in each table, all the variance values are close to 0 meaning that the PSEs in each table are close to each other. So by comparing Table 2.7 with Table 2.8, we can clearly see that the PSE drops by 85.39% from 9.72 to 1.42 (if we pick the lowest PSEs in the tables). And the PSE drops by 77.13% between Table 2.9 and Table 2.10. However, using teams' FIFA world ratings before the tournament as initial ratings does not seem very helpful. We can see from Table 2.7 and Table 2.9 that under the same score function S although the PSE gets lower because of using FIFA world ratings, but it is not much lower at all. Seeing from Table 2.8 and Table 2.10, we find out that the PSE even increases a little bit

after replacing 1000 with teams' FIFA world ratings as initial ratings.

Therefore, the optimal model will be to use 1000 as our initial rating and the actual score function S^* based on the points of two match teams (see Table 2.8). Even though we find the best model for our data, all the parameters K found are way too extremely large and H are either way too extremely small. Unfortunately, we could not have identified any home-field advantage effect up to this point. Under such circumstances, we need to compromise by choosing $K = 45$ (suggested by FIFA in Table 1.1) and $H = 0$. Although this is not optimal, yet reasonable, and compared its related PSE, 1.55 with the lowest one, 1.42, the PSE is only higher by 9.15%.

Table 2.10: Optimal parameters K and H by using FIFA world ratings & S^*

FIFA wolrd rating & S^*	PSE
$K = 45$ (according to FIFA) & $H = 0$	2.71
Optimal $K = 358.01$ (given $H = 0$)	2.17
Optimal $H = -130.24$ (given $K = 358.01$)	2.12
Optimal $(K, H) = (358.89, -129.83)$ (grid search)	2.12

2.3 Team Strength Estimation

After deciding the K-factor ($K = 45$), the home advantage boost h ($h = 0$), the initial rating (1000), and the actual score function (S^*), we can then begin the Elo

rating algorithm for the strengths of teams (i and j) with the following formula:

$$r_i^{\text{new}} = r_i^{\text{old}} + 45\left(S_{ij}^* - E_{ij}\right) \qquad \text{and} \qquad r_j^{\text{new}} = r_j^{\text{old}} + 45\left(S_{ji}^* - E_{ji}\right) \tag{2.1}$$

where $S_{ij}^* = \frac{P_i+1}{P_i+P_j+2}$ ($S_{ji}^* = 1 - S_{ij}^*$) and $E_{ij} = \frac{1}{1+10^{-(r_i-r_j)/400}}$ ($E_{ji} = 1 - E_{ij}$). r_i^{new} represents the updated rating of team i after having a new match with team j. r_i^{old} is its old rating before the new match. Note that $r_i^{\text{old}} = r_j^{\text{old}} = 1000$ for each team before this tournament (the initial ratings). r_i and r_j are the current (old) ratings of team i and team j; P_i and P_j are their scores in a match.

Table 2.11 gives the results of the Elo ratings of all teams after UEFA Women's Euro 2017. Here we can see that after the Elo updating system Netherlands has the highest Elo rating of 1046.79 since Netherlands never lost one single match and won the title after beating Denmark in the final. Although England did not get into the final (England got eliminated by Netherlands in the semi-finals), its Elo rating is higher than Denmark's Elo rating. This is because during the tournament, England scored a total of 11 goals, and Denmark, 6. This makes a big difference for updating the Elo ratings since the actual score function S^* is based on the scores of two competing teams. Also, in terms of total number of wins, England has 5, which is one more than the total number of wins of Denmark.

Not only do these Elo ratings represent the true strengths of all teams of the tournament, but later on they are also used for estimating the strength of each player of the teams and for building an ordered probit regression model to predict the game results. To examine the reliability of these Elo ratings, a simulation study will be present in Chapter 3.

Table 2.11: Elo ratings of all teams of UEFA Women's Euro 2017

Team	Elo Rating
Netherlands	1046.79
England	1022.07
Austria	1017.49
Germany	1009.68
Denmark	1008.17
France	1000.82
Belgium	999.98
Switzerland	997.68
Sweden	996.53
Italy	995.82
Spain	993.31
Portugal	990.64
Scotland	987.87
Russia	982.44
Iceland	975.76
Norway	974.96

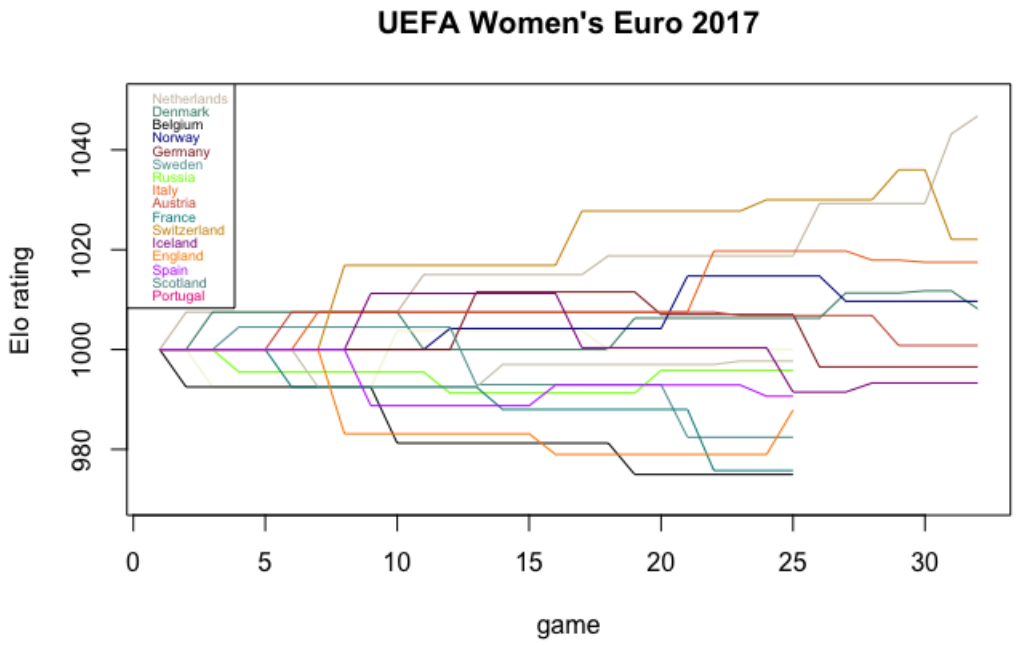


Figure 2.5: Team Elo ratings of UEFA Women's Euro 2017

Figure 2.5 gives a clear view of the trend of the team Elo ratings after every single match. Note the Elo ratings of half of the teams stopped being updated after game 24 because these teams were eliminated out of the tournament after the group stage competition. Another interesting point is that before the match of Netherlands against England in the semi-finals, both teams did not lose a single match (both teams possessed 4 wins each), and the Elo rating of England was still higher than the Netherlands Elo rating. However, with a lower Elo rating, Netherlands still managed to win over England in the semi-finals with a significantly great result (3 vs. 0).

2.4 Player Strength Estimation

Besides team-performance evaluation, another paramount core in sports analytics is to assess the performance and strength of players. In this section, we are going to present how the Elo ratings can be shifted from teams to estimate the competitive strength of all players competing in UEFA Women's Euro 2017. In order to do so, we need to utilize the updated team Elo ratings after each match and an event data set of UEFA Women's Euro 2017.

The data set consists of 856 data points containing the ratings of all participating players in every single match of UEFA Women's Euro 2017. While the ratings in the data set are completely irrelevant to the Elo rating system that is being discussed in this thesis project, they were generated under the methodology of STARSS: Spatio-Temporal Action Rating System for Soccer by Decroos et al. (2017), which assigns ratings to actions of players performed in a match by using past soccer match data.

The STARSS approach to rate the performance of players in a soccer match considers the spatio-temporal context where all player actions that deliver the offensive outcome

of a team are taken into account. The presented rating method contains three major steps:

1. dividing action stream into phrases,
2. assigning a rating to each phrase, and
3. sharing the phrase rating with individual athletes of actions that form the phrase.

STARSS splits each match of a stream of actions performed by players into several phrases based on the time on ball possession of one team (temporal aspect). Then it rates each phrase based on the most 100 similar phrases from historical data as regard to their spatial spots on the soccer field (spatial aspect). After that, the phrase rating is distributed over the players that perform the actions in the phrase with exponential decay. For more detailed information on STARSS, please check the paper, "STARSS: A Spatio-Temporal Action Rating System for Soccer" by Decroos et al. (2017).

Although the STARSS approach primarily focuses on the actions of players, it does not consider the result of a match. It seems more realistic to also incorporate the team result in a match when rating the performance and strengths of players. After all, good team work leads to team success and vice versa.

To estimate the strengths of players after a match, we first calculate the Elo ratings of two competing teams. Then we distribute the team Elo rating over the players according to each one's share. In order to calculate each player's share of their team Elo rating, we apply the following approach. After each match, the Elo rating of the winning team increases, and the share of every player who was in the match is calculated by normalizing their event ratings calculated with STARSS. Therefore, the

higher a player's event rating (from the data set) is, the more share of the Elo rating she will get if and only if her team wins a match. When a team loses, the team loses certain amount of their Elo rating, and the amount to lose is shared with its players in the match. The share of the players is computed by taking the multiplicative inverse of their ratings out of STARSS. By doing so, we can make sure that a well-performed player will not need to lose (or "pay") as much rating as her poorly-performed teammates when her team gets defeated.

However, there is a small problem. Around 1 % of the data set is 0[4], and if a player's event rating is 0, her share will be infinity ($1/0 = \infty$) when her team loses in a match. To avoide this problem, we had changed the 1 % of the data from rating 0.000000 to 0.000001 before the estimation algorithm started.

Table 2.12 presents the top 20 players of UEFA Women's Euro 2017 from our calculation result. It is clear that these top 20 players belong to 7 countries: Netherlands, England, Austria, Germany, Denmark, France, Belgium. Note that these 7 countries are also the top 7 in terms of their Elo ratings in Table 2.11. It also shows that 7 players from the champion Netherlands are in the best 20 list while Sherida Spitse from Netherlands is top-ranked player. Although England did not get into the final, 5 of England players are ranked among the best 20.

[4]Some players' event ratings from STARSS are rated zero.

Table 2.12: Top 20 players of UEFA Women's Euro 2017

Rank	Player	Team	Rating
1	Sherida Spitse	Netherlands	52.22
2	Vivianne Miedema	Netherlands	50.32
3	Jordan Nobbs	England	50.02
4	Lieke Martens	Netherlands	49.10
5	Shanice van de Sanden	Netherlands	47.94
6	Pernille Harder	Denmark	47.76
7	Desiree van Lunteren	Netherlands	47.51
8	Danielle van de Donk	Netherlands	47.46
9	Millie Bright	England	47.42
10	Jackie Groenen	Netherlands	47.40
11	Babett Peter	Germany	47.33
12	Jill Scott	England	47.04
13	Dzsenifer Marozsan	Germany	46.92
14	Nina Burger	Austria	46.70
15	Laura Feiersinger	Austria	46.40
16	Sarah Puntigam	Austria	46.36
17	Stephanie Houghton	England	46.08
18	Jodie Taylor	England	46.01
19	Eugenie Le Sommer	France	45.91
20	Elke Van Gorp	Belgium	45.90

2.5 Regression on UEFA Women's Euro 2017

2.5.1 Ordered Probit Regression

In this section, we are going to inspect the Elo ratings assigned to the 16 teams in regard to their past performance in the games that have been played, as a means to predict the results of matches in UEFA Women's Euro 2017. The Elo rating formula is applied to extract the Elo ratings of teams as covariates used to be predictive in ordered probit regression models, as was suggested by Goddard (2005), and Hvattum and Arntzen (2010).

The ordered probit regression models are used when the dependent variable is an ordered outcome of more than two possible levels. For example, an education evaluation system where people's opinions about it could be poor, fair, good, and excellent (4 levels). It also comes in a lot in opinion surveys where people are asked about a statement, and they need to answer by saying: strongly agree, agree, neutral, disagree, or strongly disagree (5 levels). In these cases, the dependent variable would be an ordered, categorical variable that the levels of it can be arranged from the highest to the lowest. Because the outcome variable would be a ranking, it is typically needed to provide a code or a value so that the statistical software can know which level is higher or lower than which one. However, the codes do not mean anything other than providing ranking order because the deviance between the highest to the second highest levels might not be the same as the difference between the second lowest and the lowest levels, and so on.

Here is the overview of the ordered probit regression model. Supposed we have a sample data set: $(\mathbf{x_1}, y_1)$, $(\mathbf{x_2}, y_2)$, ..., $(\mathbf{x_n}, y_n)$ where y_i has J possible outcomes, i.e $y_i = j$ for $j = 1, 2, ..., J$; $\mathbf{x_i}$ is a vector of independent variable where it can be

continuous level or non-continuous level. The equation for ordered probit models is stated as follows:

$$y^* = \mathbf{x}^{\mathrm{T}}\beta + \epsilon, \tag{2.2}$$

where y^* is a single latent variable which is unobservable (we only know when it crosses thresholds μ). Although we do not observe the latent variable itself, we can only observe the level of dependent varibale in response. Therefore $y_i = j$ if $\mu_{j-1} \leq y_i^* \leq \mu_j$ ($\mu_0 = -\infty$ and $\mu_J = \infty$). For example, while we do not observe exactly how people feel about a statement in a survey, we only observe one of the five categories: strongly agree, agree, neutral, disagree, and strongly disagree.

The probability of observation i will select a certain category j is then calculated as follows:

$$P(y_i = j) = P(\mu_{j-1} \leq y_i^* \leq \mu_j) = F(\mu_j - \mathbf{x}_i^{\mathrm{T}}\beta) - F(\mu_{j-1} - \mathbf{x}_i^{\mathrm{T}}\beta), \tag{2.3}$$

where F is the cumulative distribution function (CDF) of the standard normal distribution[5]. Unlike the linear regression, the parameters μ (cutoff points) and β (regression coefficients) can not be estimated under ordinary least square (OLS), yet is eatimated using maximum likelihood estimation (MLE). Fortunately, a lot of statistical software programs can help us obstain these parameters. For more detailed information on how to estimate the parameters, please check Winship and Mare (1984).

[5]If F is the CDF of the logistic distribution, i.e $F(z) = \frac{1}{1+e^{-z}}$, then (2.2) is the ordered logistic regression.

2.5.2 OPR on UEFA Women's Euro 2017

OPR with One Covariate, Elodiff

With a specific end goal to utilize the Elo ratings for forecasting match result, we will fit an ordered probit regression model. We are going to use the idea of the $\mathbf{ELO_b}$ model proposed by Hvattum and Arntzen (2010). We first calculate the Elo ratings of all teams prior to each match of the group stage competition (24 games). Then the first 24 matches are used to estimate the parameters of the ordered probit model with one single covariate, the difference in ratings of team 1 and team 2,

$$x = r_{\text{team 1}} - r_{\text{team 2}},$$

and with dependent variable, the match result of team 1,

$$y = \begin{cases} 1 & loss\ for\ team\ 1\ (win\ for\ team\ 2), \\ 2 & draw\ for\ team\ 1\ (draw\ for\ team\ 2), \\ 3 & win\ for\ team\ 1\ (loss\ for\ team\ 2). \end{cases}$$

With the parameters obtained by the first 24 games, we can then make the prediction of the first match of the knockout stage compeition (the 25th match) with the corresponding covariate x (see Elodiff in Table 2.13) from the two teams of the match. After prediction of the first knockout stage match (whether we predict it correctly or incorrectly), we then once again estimate the parameters of the ordered probit model with the first 25 matches and then try to predict the result of the second match of the knockout stage compeition (the 26th match) with the corresponding covariate x. We continue the process until the end of the tournament; therefore, we in total make the

prediction of the last 7 matches, the whole knockout stage. Note that the covariate x used (along with dependent variable) to estimate the parameters always comes from the most recently updated ratings, before a match, of two teams, and we always use as much data as possible to employ the method to predict the next match.

After running the ordered probit regression with one covariate and making match result predictions, the prediction result shows that our regression model gives a prediction rate of 57% where only 4 matches out of 7 were predicted correctly. The 3 matches that the regression model did not predict correctly are on 07-30 Germany v.s Denmark, 08-03 Denmark v.s Austria, and 08-03 Netherlands v.s England (see Table 2.13). In these 3 matches, the model forecast the winners on the teams with higher Elo ratings, yet it is the opposit of the reality. Next when we used the first 30 matches to build another order probit model with only one Elo rating difference covariate, and use the estimated parameters and the corresponding covariate to predict the winner of the final, the result showed that around 90% that Netherlands would win over Denmark.

OPR with Two Covariates, Elodiff and Home Advantage

Now we are building another ordered probit regression model with the same dependent variable, the match result of team 1, and with two covariates, the rating difference between team 1 and team 2,

$$x_1 = r_{\text{team 1}} - r_{\text{team 2}},$$

and home advantage with 3 possible outcomes:

- non of the teams is the host country,
- team 1 is the host country, and

- team 2 is the host country,

$$x_2 = \begin{cases} 0 & \textit{no home advantage exists}, \\ 1 & \textit{team 1 with home advantage}, \\ 2 & \textit{team 2 with home advantage}, \end{cases}$$

to predict the knockout stage competition. The process here is the same as the previous ordered probit regression model, only except that we are using two covariates instead of one like the previous model. The first 24 matches along with the Elo ratings of the teams prior to each match are used to estimate the parameters of regression. we then use the parameter estimations and the the corresponding covariates x_1 andx_2 of the 25th match to make the prediction of the first match of the knockout stage compeition. We continue the same process to predict the each following match until the end of the tournament.

After adding the home advantage covariate to the model, the prediction rate increases by about 15% from 57% to 71.4%, and 5 matches out of 7 were predicted correctly by the model. The 2 matches that the regression model did not predict correctly are on 07-30 Germany v.s Denmark and 08-03 Denmark v.s Austria (see Table 2.13). In these 2 matches, the model still forecast the winners on the teams with higher Elo ratings, yet the real winners were the teams of lower Elo ratings. Interestingly, the model predicts the match on 08-03 Netherlands v.s England correctly this time. Although Netherlands had a slightly lower Elo rating than England before the match, the model still predicted that Netherlands would win the match because of its home-field advantage. Next, when we used the first 30 matches to build another order probit model with the Elo rating difference and home advantage covariates, and use the estimated parameters and the corresponding covariates to predict the winner of

Table 2.13: Knockout stage competition with team Elo ratings prior to each match

Date	Team 1	Team 2	Home Advantage	Team 1 Result	Team 1 Elo	Team 2 Elo	Elodiff
07-29	NL	SE	team 1	win	1018.77	1007.02	11.75
07-30	DE	DK	no	loss	1014.73	1006.29	8.44
07-30	AT	ES	no	win	1019.74	991.48	28.26
07-30	UK	FR	no	win	1030.01	1006.82	23.19
08-03	DK	AT	no	win	1011.33	1017.91	-6.58
08-03	NL	UK	team 1	win	1029.26	1036.01	-6.75
08-06	NL	DK	team 1	win	1043.20	1011.76	31.44

the final, the result showed a nearly 100% chance that Netherlands would win over Denmark.

Chapter 3

Simulation Study

3.1 Simulation of Team Strength

In order to examine how reliable the team Elo ratings that we estimated in Section 2.3 are (see Table 2.11), and how well they represent the true strengths of each team for UEFA Women's Euro 2017, we need to be able to imitate the whole tournament over and over again under the relevant, real-tournament process and see how similar the results are to the reality.

The number of the computer simulations was set to 100. After each simulation, the Elo ratings of all teams were calculated using the Elo updating formula (1.1) with

- 1000 as initial rating for all teams,
- the K-factor set to 45,
- the actual score function S based on match result (1.2),
- the expected score function without home advantage (1.4).

We chose this model with the actual score function S based on match result (instead of S^* based on match scores), which is different from the optimal model discussed in

Section 2.2, was because of the scarce number of data points, simulating match scores would be out of the question. With all the 100 sets of team Elo ratings, a number of questions then can be answered. Out of the 100 simulations, what is the estimated probability that Netherlands would win the championship if the tournament was repeated 100 times? What are the probabilities that each country get to advance to quarter-finals, semi-finals, or even final? If we take the average of the 100 sets of team Elo ratings, what are the estimated strengths for each team? Are they any different from what we earlier estimated in Table 2.11?

The simulation process consists of two parts: the group stage (24 matches) and the knockout stage (7 matches). While the group stage is simulation as described in Section 2.1.1, if any of two teams within each group are tied on points (3 for a win, 1 for a draw, and 0 for a loss), the team with higher Elo rating that we estimated in Section 2.3 (see Table 2.11) is higher in rank. Note that here we do not apply any of the tiebreaking criteria to decide the rankings of teams with equal points since the tiebreaking rules are applied with goals scored in matches, penalty shoot-out system, disciplinary points, or teams' coefficient rankings after the qualifying group stage.

To determine the result of each simulated match, we first build an ordered probit regression model with the two covariates x_1 and x_2 described in Section 2.5.2 using the group stage matches and the Elo ratings of all teams (see Table 2.11) to estimate the coefficients β (see (2.2)) and the intercepts μ (see (2.3)). The reason we used the first 24 matches instead of the whole 31 to build the regression model was because only the first 24 games were involved with all the participating countries. Also, the results of the first 24 games consist of three possible outcomes: loss, draw, and win, whereas the last 7 matches only has two: loss and win. Second, we calculate the probabilities of loss, draw, and win for team 1 using the coefficients β, the intercepts μ, and the two

covariates of each match. Third, we draw a uniform random sample from the scale of 0 to 1 for each match, and each match outcome is determined by which interval (loss, draw, or win) the sample falls into. For example, if the probabilities of loss, draw, and win for team 1 on a certain match are 0.1, 0.3, and 0.6, and the random sample happens to be 0.5, then this match is a win for team 1 ($0.1 + 0.3 = 0.4 < 0.5$).

After simulating the group stage competition with results and counting the points for each team, we can then simulate the second part of the tournaments, the knockout stage competition. While the overall process of simulating the knockout stage is described in Section 2.1.2 and is similar to the group stage simulation, we also use the coefficients β, the intercepts μ, and the two covariates of each match to help us calculate the match outcomes probabilities for team 1 during quarter-finals, semi-finals, and final. After obtaining the probabilities of loss, draw, and win, we need to modify the probability distribution with three levels (loss, draw, and win) into two levels (loss and win) by merging half of the draw probability with loss and the other half with win since a draw is not allowed for the knockout stage.

After we simulated the tournament 100 times, calculated the Elo ratings, and averaged them out for each team, the probability of Netherlands winning the championship is 1. The order of the team Elo ratings generated by simulation (see Table 3.1) is nearly identical as our estimation (see Table 2.11). The only difference is that Sweden has about 0.71 higher in Elo rating than Italy in our estimation whereas in our simulation, Italy is about 0.71 higher than Sweden. Table 3.2, Table 3.3, and Table 3.4 also show the probabilities that each country gets into quarter-finals, semi-finals, and final. We can clearly see that the order of the probabilities of each country getting into the quarter-finals, semi-finals, and final is also similar to the order of the team Elo ratings (Table 2.11), which could be argued that the higher Elo rating a team possesses, the

Table 3.1: Simulation Elo ratings of all teams of UEFA Women's Euro 2017

Team	Elo Rating
Netherlands	1125.55
England	1060.55
Austria	1033.14
Germany	1029.47
Denmark	1008.33
France	1007.21
Belgium	991.87
Switzerland	988.58
Italy	983.93
Sweden	983.22
Spain	978.81
Portugal	975.73
Scotland	967.13
Russia	964.18
Iceland	951.83
Norway	950.48

higher probability that it moves up to the quarter-finals, semi-finals, and final (and even win the championship).

Table 3.1 gives the simulation results of the average Elo ratings for each country. Figure 3.1 and Figure 3.2 show the distributions and the variabilities of the simulation data grouped by each countries. Figure 3.1, box plot, displays the locations of the minimum, first quartile, median, third quartile, and maximum of the simulation data

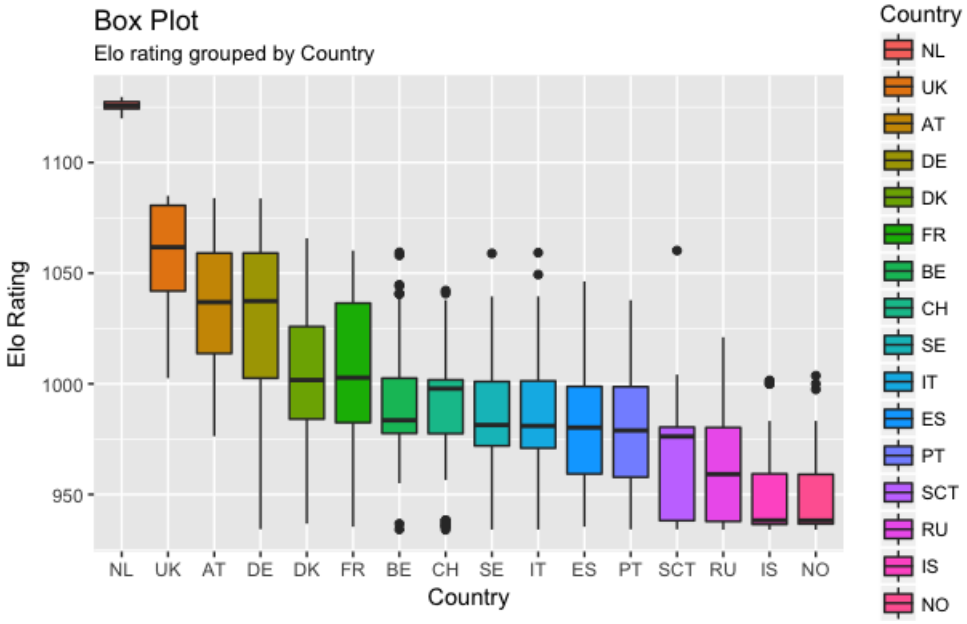


Figure 3.1: Box plot for the simulated Elo ratings of all countries

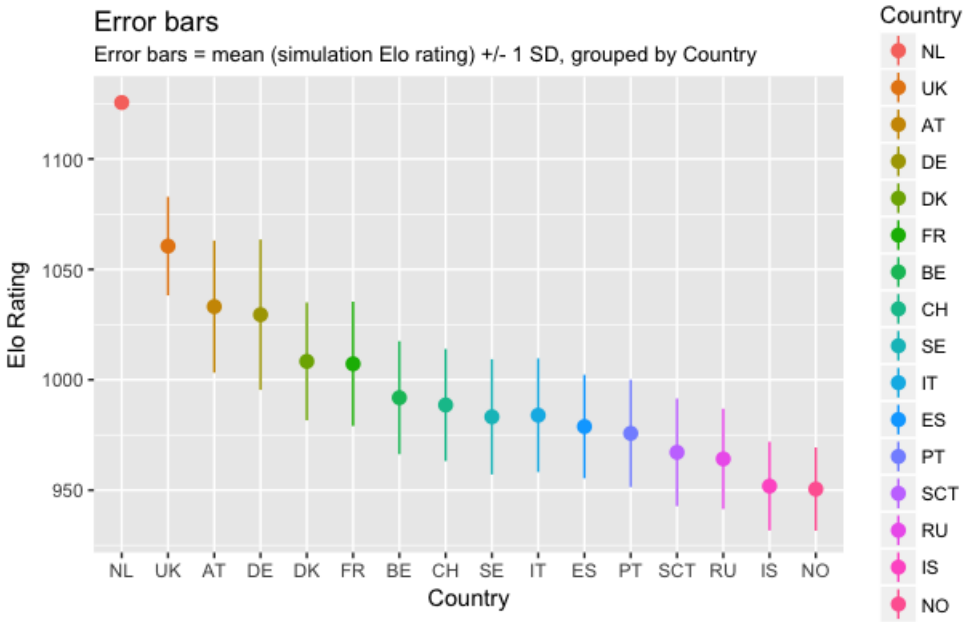


Figure 3.2: Error bars for the simulated Elo ratings of all countries

(for each team). In Figure 3.2, the length of each error bar is the mean simulation Elo rating (see Table 3.1) of a country, plus and minus one standard deviation (SD) of its mean.

It is clear that the mean simulation Elo rating of the Netherlands is about at least 50 higher than the ones of the rest of the countries. In addition, the standard deviation (SD) of the simulated Elo ratings for the Netherlands gives the highest precision of these measurements since the SD is the lowest. This implies that the simulation data for the Netherlands tend to be very close to the mean, which is a reliable estimate. We can also see that although the strongest and the weakest countries have slightly lower SD, the variabilities of all the countries, except for the Netherlands, are more or less the same. Note that in Figure 3.2, the measurement for the variability of the simulation team Elo rating data is standard deviation (SD), instead of standard error (SE). The reason why SD is presented here is that in this case, using SD is easier for us to visualize the variability of the data since SD is larger than SE, and it is just proportional to SE ($SE = SD/\sqrt{n}$, where n is the simulation size, and in this case it is 100). Lastly, neither the box-and-whisker plot nor the error bar of the Dutch team overlaps with the rest. This indicates that the difference between the mean simulation Elo rating for the Netherlands and the others may be significant, yet this conclusion can not be drawn until a statistical test is performed.

Table 3.2: Probabilities of each country getting into the final

Final	
Country	Probability
Netherlands	1.00
England	0.50
Austria	0.23
Germany	0.17
Denmark	0.05
Belgium	0.01
France	0.01
Italy	0.01
Spain	0.01
Switzerland	0.01

Table 3.3: Probabilities of each country getting into the semi-finals

Semi-finals	
Country	Probability
Netherlands	1.00
England	0.85
Austria	0.65
Germany	0.54
Denmark	0.33
France	0.16
Belgium	0.13
Switzerland	0.13
Sweden	0.07
Spain	0.06
Italy	0.05
Scotland	0.02
Portugal	0.01

Table 3.4: Probabilities of each country getting into the quarter-finals

Quarter-finals	
Country	Probability
Netherlands	1.00
England	1.00
Austria	0.94
Germany	0.94
Denmark	0.74
France	0.74
Belgium	0.58
Switzerland	0.47
Italy	0.39
Sweden	0.39
Spain	0.28
Portugal	0.18
Russia	0.15
Scotland	0.14
Norway	0.05
Iceland	0.01

3.2 Simulation of Player Strength

After we simulate the whole tournament 100 times and compute the updated team Elo ratings after a match, we can now simulate the performance and strengths of all players for each match. For a given simulated tournament, we now have the most current team Elo ratings after each match, and to simulate players' strengths for a given match, our approach proceeds in three simple steps.

First, we need to simulate players of two competing teams for each match by choosing each player with the probability calculated from the frequency of matches that she appeared in divided by the number of matches that her team played. For example, the midfielder Sherida Spitse from Netherlands played all the 6 matches during the tournament, so the probability of her getting chosen for each simulated match is 1, while her teammate, Dominique Janssen only appeared in 1 of the 6 matches so that her probability is around 17%.

Next, we assume that each player's strength follows a normal distribution with a mean of her rating (see Table 2.12) calculated in Section 2.4 and unit deviation. We randomly draw a number from the distribution of every chosen player playing in each match and then calculate the share of each one of them by normalizing their ratings if the team wins or taking the multiplicative inverse of the ratings if the team loses.

Lastly, for each simulated match in a simulated tournament, we distribute the team Elo rating over the players. After we get the 100 sets of the most updated player ratings from the 100 simulated tournaments, for each player, we average their results to get their usual average ratings for the simulated tournaments.

Table 3.5: Top 20 players of UEFA Women's Euro 2017 after 100 simulations

Rank	Player	Team	Rating
1	Sherida Spitse	Netherlands	52.52
2	Vivianne Miedema	Netherlands	52.07
3	Lieke Martens	Netherlands	51.98
4	Danielle van de Donk	Netherlands	51.94
5	Shanice van de Sanden	Netherlands	51.89
6	Jackie Groenen	Netherlands	51.80
7	Desiree van Lunteren	Netherlands	51.69
8	Kika van Es	Netherlands	51.32
9	Stefanie van der Gragt	Netherlands	51.28
10	Sari van Veenendaal	Netherlands	51.27
11	Anouk Dekker	Netherlands	51.21
12	Mandy van den Berg	Netherlands	50.29
13	Renate Jansen	Netherlands	50.11
14	Kelly Zeeman	Netherlands	48.08
15	Jill Roord	Netherlands	48.00
16	Lineth Beerensteyn	Netherlands	47.95
17	Jordan Nobbs	England	47.83
18	Millie Bright	England	47.46
19	Jodie Taylor	England	47.11
20	Francesca Kirby	England	47.07

Table 3.5 shows the results of the ratings of top 20 players from our simulation. It is obvious that Netherlands and England take all the top 20 spots. It is no surprise that 80% of the list are from Netherlands since our simulation of team Elo ratings (Table 3.1) exhibits a bigger gap between Netherlands and the rest of the competing countries than the estimation of their Elo ratings (Table 2.11). This shows that the Netherlands is even more dominant from the simulation results. While the first 16 players are from Netherlands, Sherida Spitse is still the top-ranked player.

Looking at Figure 3.3 and Figure 3.4, we can see that the best 4 athletes during this tournament are from the Netherlands (orange for the Dutch team), and their performance and ability during this tournament is very similar. This is judged from the similar spread and variability of their simulation rating distributions, and also the

overlap of their box-and-whisker plots and error bars. Note that in Figure 3.4, the error bar of Sherida Spitse does not seem to overlap with the other three athletes. This is due to the measurement of variability standard error (SE) which is proportionally smaller than standard deviation (SD). If SE is changed into SD in this case, the four error bars overlap. However, the conclusion that the difference between the ratings of the four best athletes is not statistically significant can not be drawn until a statistical test.

Figure 3.5 and Figure 3.6 are the box plot and the error bar plot for the 4 weakest athletes (in terms of their ratings). The first 3 (Guro Reiten, Ingrid Moe Wold, and Ingrid Hjelmseth) are from Norway (red), and the last one (Harpa Thorsteinsdottir) is from Iceland (blue). Their simulation rating results are also very similar, and the overlap of their box plots and error bars may be the indication that it is not significantly different in their ratings although this can not be confirmed until a valid statistical test is carried out.

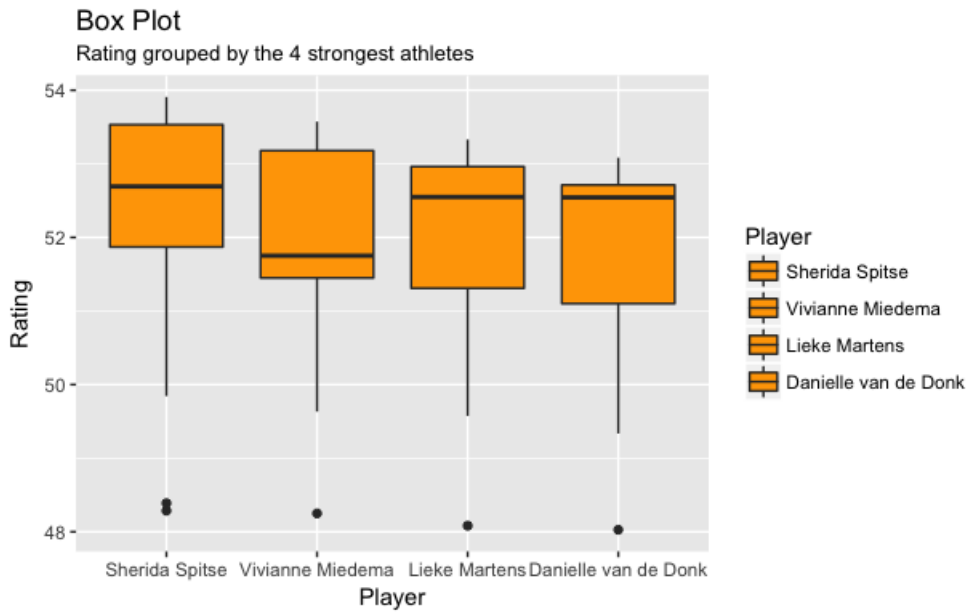


Figure 3.3: Box plot for the simulated ratings of the 4 strongest athletes

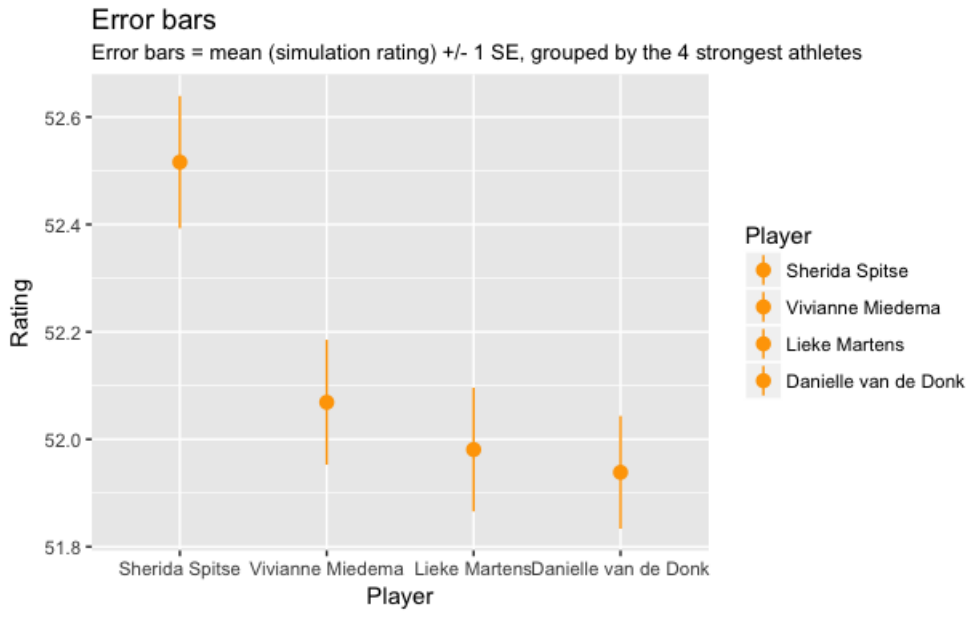


Figure 3.4: Error bars for the simulated ratings of the 4 strongest athletes

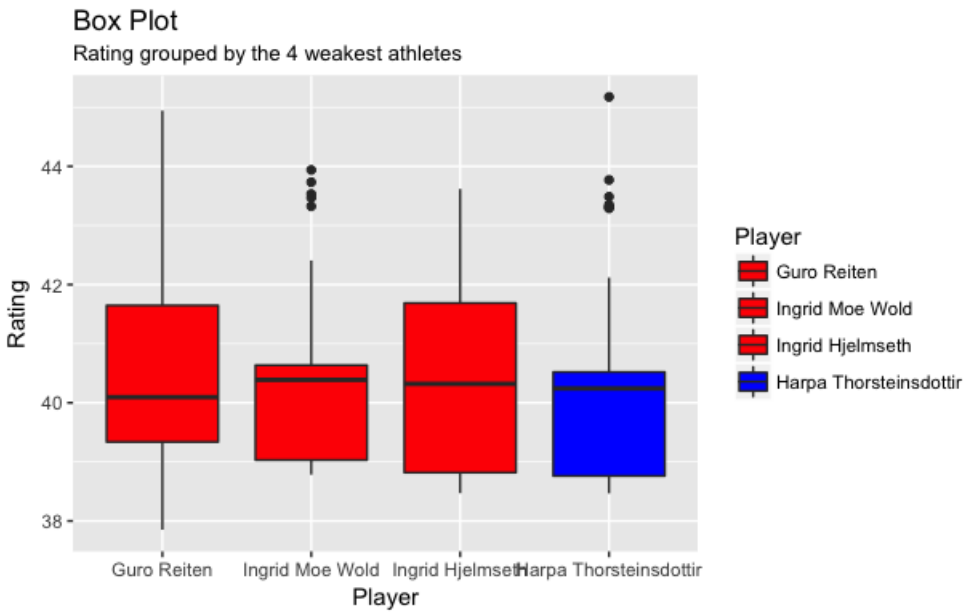


Figure 3.5: Box plot for the simulated ratings of the 4 weakest athletes

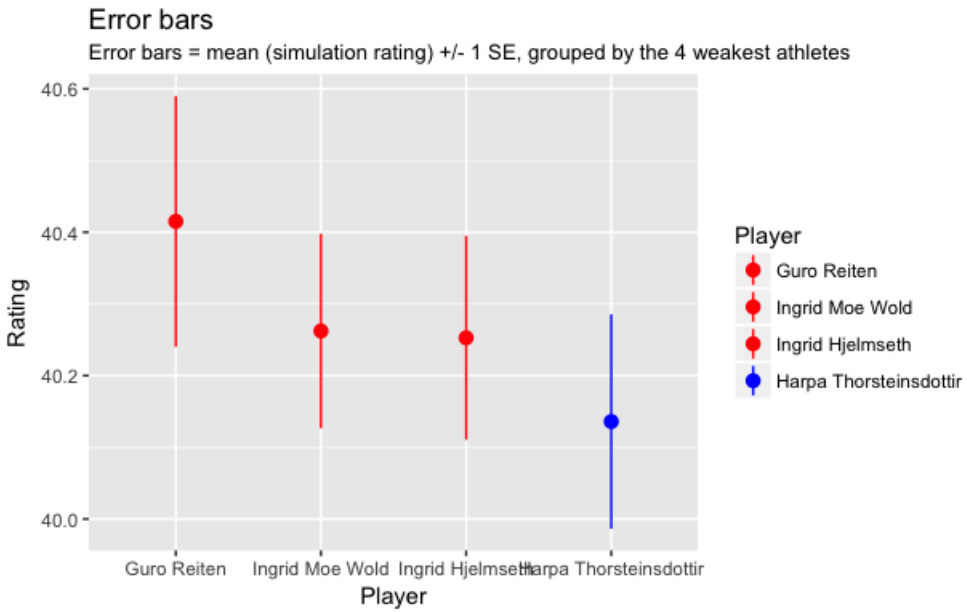


Figure 3.6: Error bars for the simulated ratings of the 4 weakest athletes

Chapter 4

Conclusion

This thesis project has shown how the Elo rating system can be applied to the 2017 UEFA Women's Championship to estimate the strengths and performance of the participating teams and players. The Elo updating formula was set with various parameters in order to find the best combination of parameters that produce the lowest prediction squared error. We found out that the optimal combination for the formula was to use 1000 as the initial ratings for all teams, 45 for the K-factor, 0 for the home advantage boost h, and the goals-based actual score function S^*. With the parameters found, we implemented the Elo updating formula to assign a rating to each team and player. The results showed that Netherlands possesses the highest Elo rating among all teams. The top 20 players are from the top 7 countries with the 7 highest Elo ratings, and Sherida Spitse from Netherlands is the best player.

During the search of the parameters, we were hoping to detect any home-field advantage for Netherlands to win the tournament, and due to the size of our data set (31 matches), the home advantage was not identified until the ordered probit regression approach. In the case of UEFA Women's Euro 2017, the home-advantage is a highly significant predictor of match results. This discovery can be interpreted as Netherlands had the advantage of being the host country to win the tournament, and our

prediction that Netherlands would win the championship is 1.

The team Elo ratings (see Table 2.11) that we estimated in Section 2.3 appear to be highly reliable according to the simulation study of team Elo ratings; however, the result of the other simulation study for player strength estimation does not seem to be much in line with Table 2.12. This can be due to the limited number of matches during the tournament. In addition, developing a reliable and objective measure to estimate the productivity of soccer players is extremely complex and challenging since only the team output can be directly observed. Also, soccer matches have the nature of low scoring, identifying individual successful movements and actions during a game is not difficult to define.

Further Research

This research project mainly focuses on applying Elo rating system on UEFA Women's Euro 2017, which consists of only 31 matches. The methodology presented in this project to estimate the strengths and performance of both teams and players is also suggested for other soccer tournaments or even other sports. One of the further steps for this project would be to use the algorithm and methodology on other soccer match tournament or soccer league with more matches. With more information from the data, it would be more efficient to train an optimal Elo rating model on all of the parameters.

Another further step that we are interested in is applying for information prior to the tournament. During our experimentation, although we tried to utilize the FIFA world ratings of all teams before the tournament as the initial ratings needed for the algorithm, the FIFA world ratings did not seem to improve the prediction after all (again, it may be because of the limited number of matches). So since we did

not have any extra knowledge about the strengths of the teams, we set equal initial ratings. This is obviously not ideal. A better initial rating set-up would have been applying the Elo rating formula on UEFA Women's Euro 2017 qualifying competition and using the most updated team Elo ratings at the end of the qualifying tournament as the initial ratings for our research.

In this research, we used the goals-based actual score function S^*, along with other parameters, to estimate the Elo ratings of the teams. We then used the ordered probit regression model based on the Elo ratings to forecast match results. It has shown that using the goals-based actual score function S^* gives better results than the results-based actual score function S. Yet during the tournament simulation process, we could also have chosen to apply bivariate Poisson regression model to estimate probability of the goals scored and conceded, and to predict the match results if more data have been allowed. This is suggested by Goddard (2005).

The biggest challenge in this research process was to estimate the strengths of player, the individual productivity in a team. Although we needed to rely on the event data set of UEFA Women's Euro 2017 generated with STARSS: A Spatio-Temporal Action Rating System for Soccer by Decroos et al. (2017) to calculate the share of each player in a match, a further study could also include other information of players on a match; for example, their positions during a match, number of matches played, minutes played, number of goals and fouls, number of goals, passes and tackles; or other personal factors such as the fitness levels of athletes, years of soccer playing, injury history, and so on.

Appendix

Here is the R code function for the Elo updated rating system. Note that this is the main code for the Elo rating computations of this thesis project. For people who are interested in the complete code and data files, please send an email to: jeff73511@msn.com.

```
## Elo1 rating function with S (1 for a win, 1/2 for a
draw, and 0 for a loss) for Team1 (1-S for Team2),
according to the match results. It returns prediction
error, mse (Brier score) and ratings of all teams after
each game ##

Elo1 <- function (data, K, initratings, H) {

  # data: Team1 vs Team2, Team1.Home,
          Team2.Home, Winner, Loser, Draw
  # K: the K-factor
  # initratings: initial ratings
  # H: home-field advantage boost
```

```
GroupA <- c("Netherlands", "Denmark",
            "Belgium", "Norway")
GroupB <- c("Germany", "Sweden",
            "Russia", "Italy")
GroupC <- c("Austria", "France",
            "Switzerland", "Iceland")
GroupD <- c("England", "Spain",
            "Scotland", "Portugal")
Teams <- c(GroupA, GroupB, GroupC, GroupD)

# Team1 expected score function
E1 <- function(r1, r2)
 10^(r1/400) / (10^(r1/400) + 10^(r2/400))

# Team2 expected score function
E2 <- function(r1, r2)
 1 - E1(r1, r2)

# Elo updating formula
RatingNew <- function (RatingOld, K, S, E)
 RatingOld + K*(S - E)

# Teams' ratings
ratings <- matrix(NA, nrow = dim(data)[1] + 1,
                  ncol = length(Teams))
colnames(ratings) <- Teams
rownames(ratings) <- paste("game",
```

```
                              0:dim(data)[1],
                              sep = "")

  # Team1 actual score function
  S1 <- rep(NA, dim(data)[1])
  S1[which(as.character(data$Winner) ==
             as.character(data$Team1))] <- 1
  S1[which(is.na(data$Winner) )] <- 1/2
  S1[which(as.character(data$Winner) !=
             as.character(data$Team1))] <- 0

  ratings[1, ] <- initratings

  for (i in 1:dim(data)[1]) {

    r1 <- if (data[i, ]$Team1.Home == 1)
      ratings[i, as.character(data[i, ]$Team1)] +
      H
      else
        ratings[i, as.character(data[i, ]$Team1)]

    r2 <- if (data[i, ]$Team2.Home == 1)
      ratings[i, as.character(data[i, ]$Team2)] +
      H
      else
        ratings[i, as.character(data[i, ]$Team2)]
```

```
    ratings[i + 1, as.character(data[i, ]$Team1)] <-
      RatingNew(ratings[i,
                        as.character(data[i, ]$Team1)],
                K,
                S = S1[i],
                E = E1(r1, r2)
      )

    ratings[i + 1, as.character(data[i, ]$Team2)] <-
      RatingNew(ratings[i,
                        as.character(data[i, ]$Team2)],
                K,
                S = 1 - S1[i],
                E = E2(r1, r2)
      )

    ratings[i + 1, ][is.na(ratings[i + 1, ])] <-
      ratings[i, ][is.na(ratings[i + 1, ])]

  }

  # MSE (Brier score)
  mse <- sum((S1 - sapply(1:dim(data)[1], function(i)
   E1(
      ratings[i, as.character(data[i, ]$Team1)],
      ratings[i, as.character(data[i, ]$Team2)]
     )
```

```
                )
  )^2,
  ((1 -S1) -
     sapply(1:dim(data)[1], function(i)
       E2(
         ratings[i, as.character(data[i, ]$Team1)],
         ratings[i, as.character(data[i, ]$Team2)]
       )
     )
  )^2
  )

  return(list(ratings = ratings,
              mse = mse))
}
```

References

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

Brent, R. P. (1971). An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, 14(4):422–425.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1):1–3.

David, H. (1988). *The method of paired comparisons*. Griffin's statistical monographs & courses. C. Griffin.

Decroos, T., Van Haaren, J., Dzyuba, V., and Davis, J. (2017). Starss: A spatio-temporal action rating system for soccer. In *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop*.

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

FIFA (2017a). *Fact Sheet, FIFA Women's World Ranking*. `https://www.fifa.com/mm/document/fifafacts/r&a-wwr/52/00/99/fs-590_06e_wwr-new.pdf` [Accessed: 2017-11-30].

FIFA (2017b). *The FIFA Women's World Ranking - 23 June 2017*. `http://www.fifa.com/fifa-world-ranking/ranking-table/women/rank=558/index.html` [Accessed: 2017-11-30].

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

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2):331–340.

Greene, W. H. (1999). *Econometric Analysis (4th Edition)*. Prentice Hall.

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

KNVB (2017). *The Dutch Sport Data Center*. `https://www.knvb.nl` [Accessed: 2017-12-5].

Langville, A. N. and Meyer, C. D. (2012). *Who's #1?: The Science of Rating and Ranking*, chapter 5. Princeton University Press.

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

SDC (2017). *The Dutch Sport Data Center*. `https://www.universiteitleiden.nl/en/science/sport-data-center` [Accessed: 2017-12-5].

Sportinnovator (2017). *Sportinnovator*. `https://www.sportinnovator.nl` [Accessed: 2017-12-5].

Sullivan, C. and Cronin, C. (2016). Improving elo rankings for sports experimenting on the english premier league. *Virginia Tech*. `http://courses.cs.vt.edu/cs5824/Fall15/project_reports/sullivan_cronin.pdf` [Accessed: 2017-10-30].

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273.

UEFA (2017a). *Match Results - UEFA Women's Euro 2017*. http://www.uefa.com/womenseuro/season=2017/statistics/round=2000623/matches/index.html#order=0asc [Accessed: 2017-11-30].

UEFA, Q. (2015). *Qualifying Results of UEFA European Women's 2017*. https://www.uefa.com/womenseuro/season=2017/standings/round=2000628/index.html#/ [Accessed: 2017-11-30].

UEFA, R. (2017b). *Regulations of the UEFA European Women's Championship 2017*. http://www.uefa.com/MultimediaFiles/Download/Regulations/uefaorg/Regulations/02/16/53/77/2165377_DOWNLOAD.pdf [Accessed: 2017-11-30].

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, pages 512–525.