Faculty of Sciences

Improving the FIFA ranking system using probabilistic modeling and prediction of the UEFA EURO 2016 tournament

Tom Van de Wiele

Master dissertation submitted to
obtain the degree of
Master of Statistical Data Analysis

Promotor: Prof. Dr. Christophe Ley

Department of Applied Mathematics, Computer Science and Statistics

Academic year 2015 - 2016

Tom Van de Wiele
June 7, 2016

Foreword

First of all I would like to thank Christophe Ley for the continued support throughout the analysis. It has been a pleasure to research a shared passion since the very start of this project. I would also like to thank Sofie Verrewaere for remaining a wonderful girlfriend despite the frequent football (association football or soccer, as you like) discussions with friends that resulted from talking about my thesis topic. People throughout my company, Eastman, have inspired and motivated me to continue with my studies while working full-time, thank you all. Finally, I would like to thank my parents to whom I own everything.

The club team data for the analysis was provided by James Curley, the author of the open source R package engsoccerdata. The match data for the European national football team matches was collected from http://eu-football.info/.

Table of Contents

Abstract

The goal of this research is to find an improved football ranking system with the FIFA (*Fédération Internationale de Football Association*) male ranking[1] system as the baseline reference. Predictive performance is used to assess the quality of the proposed ranking systems. Both structured settings where all teams play each other a fixed number of times and unstructured match settings are researched. Premier League data is analyzed in the structured setting and European national matches are analyzed in the unstructured setting. Three main classes of models are considered for all analyses: Bradley-Terry (BT) models that only consider the match outcomes (home win, draw or home loss), Poisson models that use the match results (e.g. 3-2 home win) and model both the number of home and away goals using a Poisson distribution, and Elo models which also use the match results. The model parameters for BT and Poisson models are estimated using maximum likelihood estimation whereas the team strength ratings are updated after each match in Elo models.

Six classes of statistical methods (2 BT, 2 Poisson, one blend of BT and Poisson, one Elo) are considered to predict football matches in the predictive part of the research. Prediction of the second half of Premier League matches over a 15 season period (2000-2014) revealed that the best Poisson models outperform the best BT and Elo type models with respect to the log loss criterion.

The same six classes of models are considered to predict the upcoming UEFA EURO 2016 tournament in France. The preferred model is again selected using the log loss criterion over 20 years (1986-2015) of European international matches. A simulation study using the preferred model is used in the final stage in order to come up with a detailed prediction of the expected performance of all participating teams. Simulating the tournament 100,000 times without reassessing the model after each round of matches revealed that France is the current favorite to win the tournament followed by Germany and Belgium. The ranking of the tournament participants of the simulation study aligns better with the predicted ranking by the bookmakers, which can be treated as the gold standard, than the most recent FIFA ranking. The simulation study rank has a mean absolute rank difference of 2.7 compared to the bookmaker rank while the mean absolute rank difference between the FIFA rank and the bookmaker rank is 4.5.

1 Introduction

The official FIFA announcement on November 5, 2015 revealed that Belgium was top of the male ranking for the first time in its history. Although nobody questions the great talent of the current Belgian generation it is still a remarkable observation given the fact that the best official tournament result of the team over the preceding 30 years was a quarter final exit in the 2014 FIFA world cup in Brazil. Belgium had not been able to qualify for a single major tournament (European or world cup) since exiting in the round of 16 in the 2002 FIFA world cup in Japan/South Korea.
The FIFA ranking and its major disadvantages will be discussed in the next section after which statistical alternative rankings will be discussed. At this point it is important to stress what is probably the main drawback of the FIFA ranking: it cannot be used to predict the match outcomes (home win, draw or home loss), leave alone the match results (e.g. 3-2 home win). The most important resources and an outline of the remainder of the text makes up the last part of this chapter.

1.1 FIFA ranking

The Coca-Cola World Ranking, which is the commercial name for the FIFA ranking, is based on the weighted average of ranking points per game a team has won over each of the last four years. The current ranking system has been in place since the 2006 World cup in Germany. A new official ranking is published monthly, usually on a Thursday. The average ranking points over the last 12 month period makes up half of the ranking points. The average ranking points

in the 12-24 months before the update counts for 25% leaving 15% for the 24-36 month period before the update and 10% for the 36-48 month period before the update. The arbitrary decay function is a first major criticism of the FIFA ranking: a similar match of eleven months ago has approximately twice the contribution as a match played twelve months ago.
Ranking points are calculated by the following formula:
$Result \cdot Importance \cdot Opposition\,Strength \cdot Opposition\,Confederation$
The details of all four factors will now be discussed and the major disadvantages will be pointed out.

Result *Value*: 3 for a win, 1 for a draw and zero for a loss. This system rewards winning teams generously and omits important indicators of the true team strength like goal difference, the home advantage or the duration spent leading. It is also important to point out that the average ranking points are calculated over a minimum of five matches. Consequently, if a team plays less than five matches over each of the twelve month periods, it will be treated as if it lost the other of the remaining five matches. This rule is particularly disadvantageous for countries in remote locations which find it hard to get to the five international official matches.

Importance *Value*: 1 for a friendly, 2.5 for a confederation or a world cup qualifier, 3 for a confederation tournament (e.g. UEFA EURO 2016 or the Africa Cup of Nations) or the confederations cup and 4 for world cup matches. The major criticism of this factor relates to the fact that hosts of some major tournaments do not take part in qualifying rounds and can consequently only collect points in friendly matches over a two-year period. Another way to abuse the ranking system is to obtain a higher average by playing less friendly matches since friendlies only have an importance value of 1.

Opposition strength *Value*: $max(0.5, \frac{200-Oppositionranking}{100})$ With the small correction that playing against the number one in the FIFA ranking receives a multiplier of 2 (compared to 1.99 in the formula). Opponents with rankings of 150 and lower are assigned with the same opponent strength multiplier. The major criticism of this part of the formula relates to the fact that the better teams in weaker regions such as Asia, Africa and Oceania find it very hard to collect points from the confederation and world cup qualifiers as well as confederation tournaments leading to a ranking inertia. Also, this approach fails to leverage the more detailed ranking points. The difference between a team ranked 30th (1000 points) and a team ranked 31st (also 1000 points) is considered the same as the 31st ranked team and the 32nd ranked team (900 points).

Opposition confederation *Value*: 1 for South America, 0.99 for Europe and 0.85 for other regions. This multiplier is calculated based on the performance of the teams in the previous three world cups. It is a vital part of the ranking since teams from different regions only play

each other occasionally in friendlies and every four years in the world cup tournament. An often heard critic from the smaller regions is similar to the critic from the previous paragraph: the opposition confederation factor leads to ranking inertia by benefiting European and South American teams.

1.2 Wang and Vandebroek - A model based ranking system for soccer teams

Wang and Vandebroek (2011)[2] proposed a statistical approach with a strength parameter for each football team to overcome some of the major drawbacks of the FIFA ranking. They use an adapted Bradley-Terry (BT) model for their purpose.

The details of their approach will be discussed in chapter 2 but the most important aspects of the approach are covered here. The main idea is that the outcome probabilities are a direct function of the team strengths. For now it is assumed that draws cannot occur in the interest of simplicity. This results in the simplest BT structure and arguably the simplest possible statistical ranking structure. The probability that team 1 beats team 2 is modeled as $\frac{\beta_1}{\beta_1+\beta_2}$ where team i has strength parameter β_i. The strengths are restricted to be strictly positive in order to translate to match outcome probabilities.

The team strengths β are estimated using maximum likelihood estimation given past match outcomes. A ranking of the teams follows naturally by sorting the teams by their estimated team strengths.

It is remarkable that the simple BT model is able to overcome some of the major drawbacks of the FIFA ranking. The most obvious improvement results from the fact that the BT ranking can be used to predict match outcomes. Another clear improvement is the observation that organising teams of major tournaments are no longer penalized for not playing regional qualifiers.

The BT approach of Wang and Vandebroek will be used as a starting point and main reference for all the proposed statistical methods in the remainder of this text.

1.3 Other football ranking approaches

There are numerous alternative ranking procedures so the discussion will be limited to the most popular alternatives in the literature that only consider previous match results.

Poisson models Poisson models were first suggested by Maher (1982)[3] to model football match results. He assumed the number of scored goals by both teams to be **independent** Poisson distributed variables. Let G_i be the goals scored by the home team i and G_j be the goals scored by the away team j. With those assumptions the density function can be written as

$$P(G_i = x, G_j = y) = \frac{\lambda^x}{x!} exp(-\lambda) \cdot \frac{\mu^y}{y!} exp(-\mu) \tag{1.1}$$

In this formula, the means of the scored goals for the home and the away team are λ and μ respectively. Maher assumed a log-linear model for the scoring rates: $log(\lambda) = c + H + o_i - d_j$ and $log(\mu) = c + o_j - d_i$ where H represents the home advantage and c is a common intercept. The parameters o_i, o_j, d_i and d_j stand for offensive and defensive capabilities of teams i and j respectively. The basic Poisson model as described above is also able to overcome the major drawbacks of the FIFA model. The basic Poisson model has been used frequently since Maher's original publication and several possible improvements have been suggested.
Karlis and Ntzoufras (2003)[4] studied diagonal inflated bivariate Poisson models as they noticed that the independence assumption caused underestimation of the draw probabilities in some cases.
Crowder et al. (2002)[5] extended the basic Poisson model by allowing the strength parameters to vary over time in order to account for the fact that the shape of a football team is subject to fluctuations.

Elo models Interestingly, the women's FIFA ranking[6] uses a ranking procedure which is based on the Elo rating system. The Elo system was originally invented as an improved chess rating system. The ratings are updated after each match using the formula

$$R_n = R_o + K \cdot (O_{act} - O_{exp}) \tag{1.2}$$

R_n represents the new rating in this formula, which is the update from the old rating R_o and K represents the match importance. The values for K agree closely with a multiple of the importance factor of the men's ranking procedure. O_{act} denotes the actual outcome and O_{exp} represents the expected match outcome. A team will receive a better rating if it performs better than expected and the change is proportional to the match importance. The women's ranking procedure overcomes all major disadvantages of the men's ranking procedure. It allows for prediction of match outcomes as a function of the difference in Elo ratings and incorporates the home advantage and goal difference into the outcome score O_{act}. A win for the home team by $1 - 0$ results in $O_{act} = 0.85$ and $O_{act} = 1 - 0.85 = 0.15$ for the losing team. The calculation details of O_{act} are discussed in chapter 2. Winning by $3 - 1$ results in an increased O_{act} (0.893) and $O_{act} = 1 - 0.893 = 0.103$ for the losing team. Playing few important matches no longer penalizes a team as this will result in little change of its ranking. The arbitrary opposition strength calculation is performed more granularly and teams of weaker regions no longer suffer from the regional drawbacks in the men's ranking.
A similar Elo ranking system for men's football is maintained on http://eloratings.net/[7]. The main difference is found in the calculation of the O_{act} value. It is equal to 1 for

a win, 0.5 for a draw and 0 for a loss whereas the female FIFA ranking ranges between 0.01 (loss by 0-6/6+) and 0.99 as a function of the goals scored and the goal difference. The match importance score K is also adjusted slightly and gives a slightly higher weight to friendlies compared to the FIFA importance score.

1.4 Resources

The statistical programming language R[8] was used for all the analyses in this research.
The data source of the English Premier League was supplied by James Curley through the engsoccerdata[9] package. Bookmaker odds and match statistics for the English Premier League were collected from http://www.football-data.co.uk/englandm.php/[10].
National team match results were scraped from http://eu-football.info/[11] using Hadley Wickham's R package rvest[12].
All the graphs in this text were generated using Hadley Wickham's R package ggplot2[13].

1.5 Outline

The details of the three considered model types (Bradley-Terry, Poisson and Elo) are discussed in chapter 2. All considered model hyperparameters will be discussed and a simulation study is performed to verify the consistency of the estimators. Next, the prediction performance of the three considered model types is analyzed in chapter 3. Bookmaker predictions of 15 years of English Premier League seasons and a model using match data variables will be used to benchmark the model performances. An analysis of the predictive performance of European male national team matches makes up the second part of the chapter and a single model will be selected as the preferred predictive model for European male national team matches. This selected model will then be used to simulate the UEFA EURO 2016 tournament. The last part of chapter 3 compares the results of the UEFA EURO 2016 simulation study to the FIFA ranking and bookmaker odds. A summary of the most important findings and possible future improvements is provided in chapter 4.

2 Models

The three studied model families (Bradley-Terry, Poisson and Elo) with the considered hyperparameters will be discussed in detail in the first part of the chapter. Next, the parameter estimation procedure will be described. A simulation study is performed in the last part of the chapter to make sure that the parameter estimates of all three researched model families are consistent.

2.1 Considered models

2.1.1 Wang and Vandebroek - A modified extended Bradley Terry model

Wang and Vandebroek proposed a statistical approach based on maximum likelihood estimation with a strength parameter for each national team plus two additional parameters to overcome some of the major drawbacks of the FIFA ranking. They used an adapted Bradley-Terry (BT) model for their purpose.

The standard BT model only allows two possible outcomes (home win or home loss). In football however, the probability of a tie needs to be modeled as well. The tie probability is modeled similarly to Davidson (1970)[14] in such a way that the modeled probability of a draw increases as the strengths of both teams are more similar. The adapted BT model also takes a home effect into account. From an analysis of 10 recent Italian Serie A seasons it was apparent that the home advantage is proportional to the strength of the home team. The analysis of a BT model with separate weights for home and away matches later on in this text reveals that the multiplicative home effect is a valid simplification. A total of M team strengths (β) needs to be estimated

when M teams are analyzed. From now on we denote the home team as team 1 and the away team as team 2. If we call P_{i1} the probability of a home win for team 1 in match i, P_{i2} the probability for a draw in match i and P_{i3} the probability of an away win for team 2 in match i, then the outcome probabilities are

$$P_{i1} = \frac{H\beta_{1i}}{H\beta_{1i} + K\sqrt{H\beta_{1i}\beta_{2i}} + \beta_{2i}}$$
$$P_{i2} = \frac{K\sqrt{H\beta_{1i}\beta_{2i}}}{H\beta_{1i} + K\sqrt{H\beta_{1i}\beta_{2i}} + \beta_{2i}}$$
$$P_{i3} = \frac{\beta_{2i}}{H\beta_{1i} + K\sqrt{H\beta_{1i}\beta_{2i}} + \beta_{2i}}$$

The home effect is represented by H and the tie effect is represented by K. A home effect $H > 1$ inflates the strength of the home team and increases its modeled probability to win the match. The tie effect is best understood by assuming similar strengths in the absence of a home effect. In that case P_{i1} is similar to P_{i3} and the relative probability of P_{i2} compared to a home win or loss is approximately equal to K. H is typically greater than one since playing at home gives the benefit of familiar surroundings, the support of the home crowd and the lack of traveling. Matches on neutral ground are modeled by dropping the home effect H. All strength parameters β, H and K cannot be negative.

The strength parameters β, H and K are estimated using maximum likelihood (ML) estimation of the previous match outcomes. Two important extensions to make the extended BT model more realistic are added. The first extension to multinomial ML estimation is to give more weight to more important matches. The FIFA weights seem reasonable for this purpose and will be employed whenever national team matches are analyzed. The relative importance of a national match is indicated by $w_{type,i}$ and can take the values 1, 2.5, 3 and 4. The second extension is the use of a continuous depreciation function which gives less weight to older matches with a maximum weight of 1. Specifically, the time weight for a match which is played x_i days back is calculated as $w_{time,i}(x_i) = exp(\frac{x_i}{Half\,period}\, log(\frac{1}{2}))$. This means that a match played $Half\,period$ days ago only contributes half as much as a match played today and a match played $3 \cdot Half\,period$ days ago contributes 12.5% of a match played today. Figure 2.1 shows a graphical comparison of a continuous time decay function versus the arbitrary FIFA decay function.

Let y_{ij} be 1 in the likelihood formula below if the result of game i is j and $y_{ij} = 0$ otherwise. Once again $j = 1$ denotes a home win, $j = 2$ a draw and $j = 3$ an away win. The likelihood becomes as follows if N matches are being considered

$$L = \prod_{i=1}^{N} \prod_{j=1}^{3} P_{ij}^{y_{ij} \cdot w_{type,i} \cdot w_{time,i}} \tag{2.1}$$

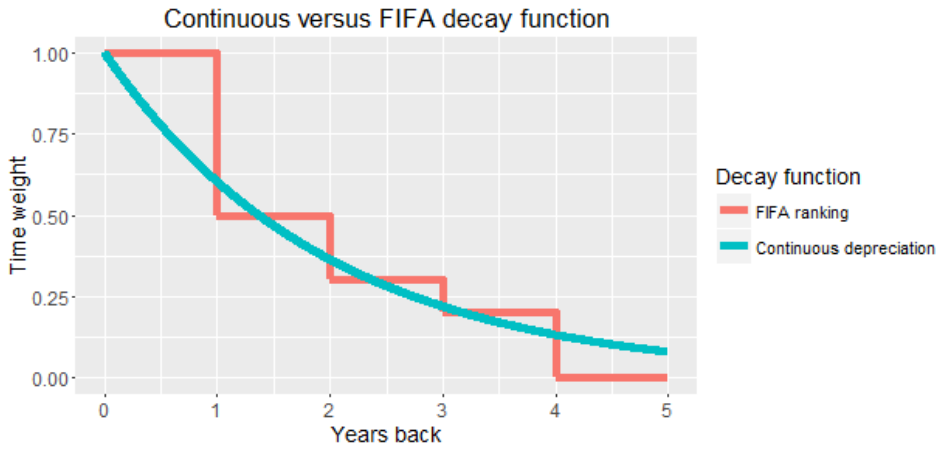

Figure 2.1: Comparison of the FIFA ranking decay function versus an exponential smoother. The blue continuous depreciation line uses a half period of 500 days.

The model parameters p (a strength parameter β for each team, H and K) are replaced with $exp(p)$ in the estimation procedure to make sure that the estimated parameters are positive.

2.1.2 Poisson model

Maher's (1982)[3] Poisson model is used as a starting point to model outcome probabilities. The model specification is however adapted in order to match the BT specification. A single strength parameter for each team will also be enforced rather than a defensive and attacking strength parameter for each team in Maher's model. Additionally, the extensions of the adapted BT model to give more weight to more recent matches and take the match importance into account are also incorporated.

Let G_i be the goals scored by the home team i and G_j be the goals scored by the away team j. Assuming an independent Poisson distribution for the goals scored by each team, the density function can be written according to formula 1.1. The assumption of a log-linear model for the scoring rates: $log(\lambda) = c + H + \beta_i - \beta_j$ and $log(\mu) = c + \beta_j - \beta_i$ is equivalent to $\lambda = c \cdot H \cdot \frac{\beta_i}{\beta_j}$ and $\mu = c \cdot \frac{\beta_j}{\beta_i}$ when the constraint that all parameters are positive is enforced. This constraint is satisfied by replacing the model parameters p (a strength parameter β for each team, c and H) by $exp(p)$ in the estimation procedure. Matches on neutral ground are modeled by dropping the home effect H.

Next, the probability of all three possible outcomes P_{ij} is computed. If $D = G_i - G_j$, then the probability of a win of team i over team j and the probability of a draw is computed as $P(D > 0)$ and $P(D = 0)$ respectively. The Skellam distribution, the discrete probability distribution of the difference of two Poisson distributed variables, is used to derive these probabilities given λ and μ.

The likelihood function can now be written in the exact same form as formula 2.1 from the BT discussion. The model parameters are estimated using maximum likelihood estimation. It

is important to notice that the Poisson model uses two observations for each match (the goals scored by each team) while using the same number of parameters (number of teams + 2). The BT approach only uses a single observation for each match.

2.1.3 Derived BT and Poisson models

The proposed BT and Poisson models both contain a single team strength. A derived model with two strength parameters for each team is analyzed as well. The third derived model modifies the BT model to give more weight to matches with greater goal differences.

Extended Bradley-Terry model (EBT)

The extended BT model contains a separate team strength parameter β for home and away team strengths. This way, the home team strength is not influenced by performances on the road and the away team strength is not influenced by results in home matches. A team is modeled as if it were a separate team playing at home or away from home. This contrasts with the BT model where the strength parameter at home is multiplied by a constant H for all teams. The tie effect is modeled similarly leading to the following outcome probabilities for the home team 1 playing the away team 2 in match i

$$P_{i1} = \frac{\beta h_{1i}}{\beta h_{1i} + K\sqrt{\beta h_{1i}\beta a_{2i}} + \beta a_{2i}}$$
$$P_{i2} = \frac{K\sqrt{\beta h_{1i}\beta a_{2i}}}{\beta h_{1i} + K\sqrt{\beta h_{1i}\beta a_{2i}} + \beta a_{2i}}$$
$$P_{i3} = \frac{\beta a_{2i}}{\beta h_{1i} + K\sqrt{\beta h_{1i}\beta a_{2i}} + \beta a_{2i}}$$

The home strengths are indicated by the βh parameters and the away strengths are written as βa. Estimation of the parameters is identical to the procedure used in the BT and Poisson sections since the likelihood function can be written as a function of the model parameters (a home strength βh and an away strength parameter βa for each team as well as a common tie effect K) using the match outcome P_{ij}.
The strength of a team on neutral ground is modeled by combining the home and away strengths. All three Pythagorean means were considered for this purpose (the arithmetic, geometric and harmonic means). Analysis of all available Premier League data (1892-2014) revealed that the geometric mean is best suited for the combination of team strengths. The absolute difference in ranks between the Pythagorean means of the modeled team strengths (two for each team) and the actual 2-point for a win final league ranking was used to identify the preferred approach. Team strengths were estimated using all season matches for each separate season. Table 2.1 summarizes the results and reveals that the geometric mean strongly outperforms the other considered Pythagorean means. The mean ranking difference between the geometric mean of

	Method		
	Arithmetic mean	Geometric mean	Harmonic mean
Absolute ranking differences sum	3466	1004	3512
Mean ranking difference	1.49	0.43	1.51

Table 2.1: Comparison of the three Pythagorean means to combine home and away team strengths. The actual 2-point ranking of 2333 teams over a total of 112 seasons is compared to the ranking of the combined team strengths. The geometric mean clearly outperforms the other two approaches.

the team strengths and the actual 2-point ranking at the end of the season was found to be 0.43. This is much less than the mean ranking difference between the actual 2-point ranking and the arithmetic and harmonic means of the modeled team strengths (1.49 and 1.51 respectively).

Extended Poisson model

Extending the Poisson model to include two team strengths was performed using Maher's (1982)[3] approach. He modeled the distribution for the goals scored by each team using separate attacking and defensive strengths of the teams. Maher also used a multiplicative home effect.

Let G_i be the goals scored by the home team i and G_j be the goals scored by the away team j. Assuming independent Poisson distributions for the goals scored by each team, the density function remains unchanged from formula 1.1. The expected value of the goal counts is modeled as $\lambda = c \cdot H \cdot \frac{\beta a_i}{\beta d_j}$ and $\mu = c \cdot \frac{\beta a_j}{\beta d_i}$. The βa parameters refer to the attacking strengths and the defensive strengths are represented by βd. Modeling matches on neutral ground is enforced by dropping the home effect H.

Combined Bradley-Terry - Poisson model (Combined BTP)

The basic BT model and the extended BT model do not use all of the available information. They only take the match outcome into account, omitting likely valuable information present in the goal difference. A team that wins by 8-0 and loses the return match by 1-0 is probably stronger than the other team. The combined Bradley-Terry - Poisson model tries to take this information into account by using the basic BT model where the matches are given an increasing weight when the goal difference grows. The likelihood function is calculated as

$$L = \prod_{i=1}^{N} \prod_{j=1}^{3} P_{ij}^{y_{ij} \cdot w_{goalDiffscaled,i} \cdot w_{type,i} \cdot w_{time,i}}$$

The goal difference weight is the only difference compared to the basic BT model. The considered formula for the goal weight as a function of the goal difference is defined as $w_{goalDiff,i} = log2(goalDiff, i + 1)$. This way, a goal difference of 1 receives a goal difference weight of 1 and every additional increment in goal difference results in a smaller increase of the goal dif-

ference weight. A goal difference of 7 goals receives a goal difference weight of 3. All weights where the home team won the match are scaled in the estimation procedure so that the average goal difference weight for matches where the home team came out as the winner is equal to 1. The same process is repeated for matches where the away team came out as the winner. Draws are given a weight of 1. Scaling the goal difference weights in the estimation procedure is necessary in order to obtain unbiased estimates of the home effect H and the tie effect K. Not rescaling the weights for the non-tied matches would result in an underestimate of the true tie effect K. The home effect $\hat{H}_{ml}$ maximum likelihood estimators would be overestimated if the true home effect $H > 1$ and the maximum likelihood estimators would be an underestimate of the true home effect for $H < 1$.

2.1.4 Elo model

The Elo model is essentially different compared to the BT and Poisson models. Team strengths are updated after each match whereas the team strengths in the BT and Poisson models are estimated using all matches in the considered time frame. Consequently, more recent matches are inherently more important in the Elo model. Elo ratings are updated after each match using formula 1.2. A team will receive a higher rating if it performs better than expected and the change is proportional to the match importance. The rating change of the other team is identical but it is applied in the opposite direction, leaving the mean Elo ranking constant.
The female FIFA[6] procedure to calculate O_{act} will be used in this research rather than the estimation used at http://eloratings.net/[7] since the women's FIFA procedure incorporates the goals scored and the goal difference.
Table 2.2 shows the calculation of the O_{act} values. Values range between 0.01 and 0.99.

	Goal difference						
	0	1	2	3	4	5	6/6+
Goals scored	**Actual result (percentage)**						
0	47	15	8	4	3	2	1
1	50	16	8.9	4.8	3.7	2.6	1.5
2	51	17	9.8	5.6	4.4	3.2	2
3	52	18	10.7	6.4	5.1	3.8	2.5
4	52.5	19	11.6	7.2	5.8	4.4	3
5/5+	53	20	12.5	8	6.5	5	3.5

*Table 2.2: Percentage scores of O_{act} for the **drawn or losing team**. The winner is awarded 1-$O_{act,loser}$, except for a draw (goal difference = 0) when the opponent receives the same number of points.*

The expected match outcome is calculated as

$$O_{exp} = \frac{1}{1 + 10^{-x/2}}$$
$$x = \frac{R_o - A_o + H}{200}$$

A_o is the opponent rating before the match in this formula. H represents the home effect and is set to 100 if the team for which the change in Elo is calculated plays at home, -100 if it is the visiting team and 0 if the match is played on neutral ground. Figure 2.2 illustrates the relation between O_{exp} and the numerator of x, the rating difference between the teams after taking the home effect into account. A rating difference of 100 after adjusting for the home effect results in $O_{exp} = 0.64$ and a rating difference of 300 after adjusting for the home effect translates to $O_{exp} = 0.85$. A couple of examples are written out to further clarify the Elo calculation. A team with rating R plays a team with rating $R - 200$ on away ground and wins by $1 - 3$. O_{exp} for the away team is equal to 0.64 and O_{exp} for the home team equals 0.36. O_{act} for the visiting team is equal to 0.911 since O_{act} for the home team equals 0.089. The visiting team will receive a rating update of $K \cdot (0.911 - 0.64)$ and the home team will be subject to an identical drop in ratings. The Elo ratings remain unchanged when a team with rating R plays a team with rating $R - 200$ at home and wins by $1 - 0$, no matter the importance of the match.

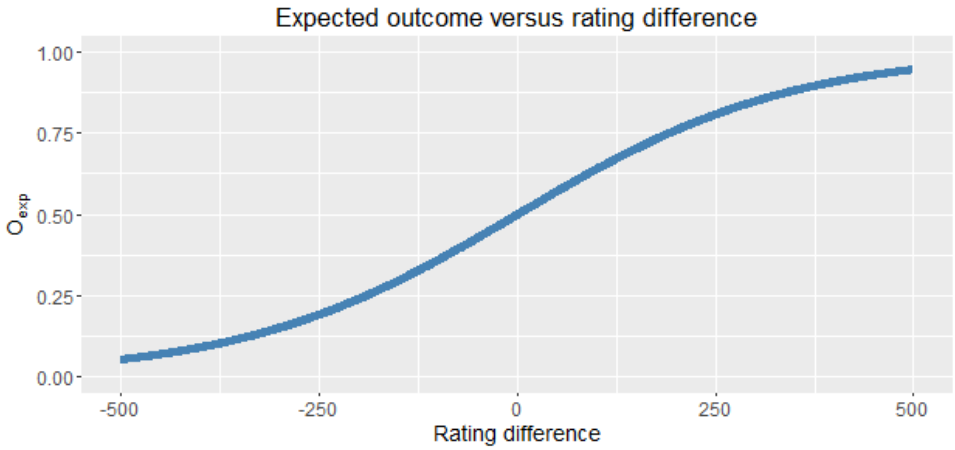

Figure 2.2: Relation between the Elo ranking difference of two teams after taking the home effect into account and the expected outcome O_{exp}.

The match importance factor K in the Elo ratings will be set to $K = c{\cdot}matchImportanceFIFA$, a multiple of the match importance scores from the male FIFA ranking. The FIFA match importance factor can take the values 1, 2.5, 3 and 4. The constant c will be referred to as the Elo update constant and will be varied to study different time depreciation effects.
Elo ratings are easy to interpret and update but it takes some time to converge to stable ratings. The convergence rate strongly depends on the multiplicative constant of the match importance factor K. The team ratings are initialized at zero but it should be noted that any constant could

have been chosen.
Translating Elo ratings to match outcome predictions is not a straightforward task in the presence of ties. The followed procedure is based on the reasoning of Hvattum and Arntzen (2010)[15]. They decided to use a proportional odds logistic regression model with a single covariate (difference in Elo ratings). Hvattum and Arntzen used part of their data to estimate the proportional odds logistic regression model parameters after making sure that the Elo ratings had converged. European national team matches from the period 1980-1985 are used to make sure that the Elo ratings are converged. The period 1986-1991 is then used to estimate the proportional odds logistic regression parameters. The model was calculated using the 1980-1991 data in order to avoid using the same match data used in the national team prediction calculations of chapter 3. The predictive performance of the considered models for national team matches is assessed using matches between 1992 and 2015. A value of 15 is used for the multiplicative constant of the match importance factor K since this is the constant used by the female FIFA ranking. The model resulted in a fit which was centered on a rating difference of 15. The model parameters were adjusted manually to enforce centering on an Elo rating difference of 0. The outcome probabilities versus the Elo rating difference for the manually adapted model are depicted in figure 2.3.

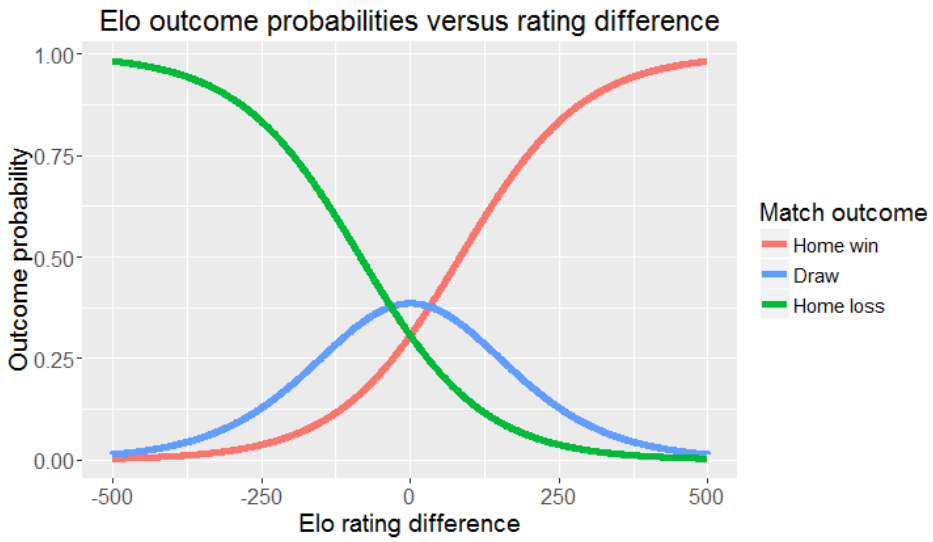

Figure 2.3: Outcome probabilities versus Elo rating difference after taking the home effect into account.

2.1.5 Hyperparameters

The purpose of the hyperparameters is similar for all considered methods. Each predictive model contains two hyperparameters: one that takes depreciation of matches into account and one that relaxes the predicted outcome probabilities.

Match depreciation

The match depreciation is modeled directly in the BT and Poisson models since it is a part of the likelihood function 2.1. A larger half period gives relatively more weight to older matches. The limiting case where the half period is infinity gives equal weights to all matches.
Match depreciation is modeled differently in the Elo model 1.2. The update value K is set to $K = constant \cdot matchImportance$. The constant in the previous formula is varied to study different match depreciation effects.

Probability relaxation

Predicted outcome probabilities for a given match depreciation effect are relaxed linearly using the hyperparameter k. This linear relaxation process is also known as Laplace or additive smoothing. It is commonly used in naive Bayes classification. The set of the three predicted outcome probabilities P_{ij} for each match i is transformed to $\frac{P_{ij}+k}{1+3\cdot k}$. The transformed probabilities equal the original probabilities for $k = 0$. Increasing k shrinks all three outcome probabilities for a given match towards $\frac{1}{3}$. All models use the same relaxation transformation. Figure 2.4 shows the relation between the model and relaxed probabilities for $k = 0.1$. The relaxed probability for a model probability of 1 is depicted in figure 2.5 for different values of k.

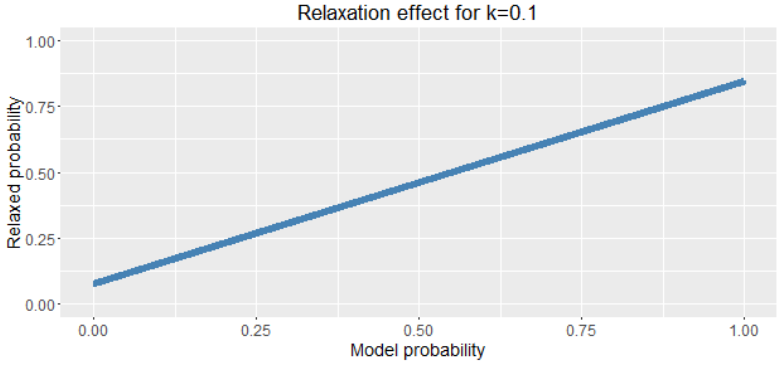

Figure 2.4: Linear relaxation of model probabilities between 0 and 1 for $k = 0.1$.

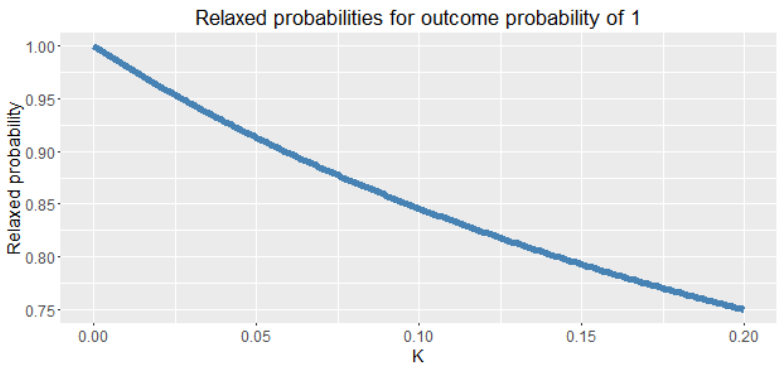

Figure 2.5: Relaxed probabilities for different values of k and a model probability of 1.

2.2 Parameter estimation

Parameters in the BT and Poisson models are estimated using maximum likelihood estimation. A custom function calculates the negative log likelihood given the model parameters. The negative of the log likelihood is calculated since the base R $optim$ function calculates the minimum of a function by default.
BFGS, a quasi-Newton method, is the preferred method to estimate the parameters due to its robust properties. On complex problems it may however not converge to the maximum likelihood estimators. The conjugate gradient method is used to estimate the model parameters when convergence does not occur using the BFGS method. All model parameters should be positive. This is enforced by the exp transformation of the model parameters in the likelihood calculation. All optimized parameters are initialized at 0 (1 in the estimation procedure after applying the exp transformation). The maximum likelihood estimators for given hyperparameters are used as initial values in subsequent estimations of the model parameters in order to speed up the convergence. The denominators of the density functions for the Poisson models (x! and y! with x and y the goals scored for the home and away team respectively) are omitted since they are independent of the model parameters. Estimation of the team strengths in Elo models are calculated by updating the Elo formula 1.2.
Table 2.3 summarizes the parameter estimation procedure for the different methods.

Method	**# Parameters**	**Estimation procedure**
BT	$M + 2$	Maximum likelihood
Poisson	$M + 2$	Maximum likelihood
EBT	$2M + 1$	Maximum likelihood
Extended Poisson	$2M + 1$	Maximum likelihood
Combined BTP	$M + 2$	Maximum likelihood
Elo	M	Elo update formula 1.2

Table 2.3: Summary of the parameter estimation of the studied models. M represents the number of teams.

2.3 Consistency study

A consistency study is performed for the base version of all three main model types: BT, Poisson and Elo. The difference in randomized model parameters is compared to the estimated model parameters for an increasing number of matches N. All matches are simulated using the randomized model parameters and all matches are assumed to have a home effect (so no matches on neutral ground). Ten sets of randomized model parameters are used to simulate the match outcomes/results. The same 10 randomized parameter sets are used for all considered values of N. A total of 54 teams is considered in the consistency study since this corresponds

to the most complex estimation setting (national team matches).
The randomized model parameters are drawn from a distribution that approximates the model parameters for national team matches in the period between 1992 and 2015. BT team strengths are drawn randomly from the distribution $exp(RandomUniform[-3, 3])$. All other model parameters (including all Poisson and Elo model parameters) are drawn from a uniform distribution. The tie effect K is drawn from the interval $[0.8, 1]$ and H is drawn randomly from $[1, 1.5]$. Team strengths in the Poisson model are drawn from $[\frac{1}{2}, 2]$, c is picked from the interval $[1, 1.5]$ and H is also selected randomly from $[1, 1.5]$. True Elo ratings are drawn randomly from the interval $[-300, 300]$.
All hyperparameters are set to their default values: the half period is fixed at infinity, giving each match the same weight for BT and Poisson models. K in the Elo update formula 1.2 is set to $2.5 \cdot 15 = 37.5$. This corresponds with simulating only confederation or world cup qualifiers using the female FIFA ranking base constant of 15.
Simulating random match outcomes using the BT model and random match results using the Poisson model given the true model parameters is straightforward. The Elo ratings can however only be used to simulate match outcomes using the proportional odds logistic regression model. This requires special attention since the update formula also takes the goals scored by both teams into account. Outcomes where the home team is simulated as a winner is coded as a 1-0 win. Draws are coded as 1-1 (O_{act} of 0.5) and home losses are coded as 0-1.
The estimated team strength parameters β in the BT and Poisson model are multiplied by a constant to make sure that the mean team strength is equal to the mean team strength of the randomized model parameters that were used to simulate the matches. This transformation enabled a meaningful interpretation of the mean absolute parameter error for the model parameter estimators. Scaling the estimated team strength parameters has no influence on the modeled match outcome calculations since the numerator and denominator of the relevant formulas are multiplied by the same constant.
Figure 2.6 shows the mean absolute model parameter error for the consistency study of the Bradley-Terry model. It can clearly be seen that the estimated model parameters converge to the randomized model parameters that were used to simulate the matches. The Poisson consistency study leads to the same conclusion as can be seen from figure 2.7. A different trend is observed in the Elo consistency study, which is presented in figure 2.8. The mean absolute Elo rating error drops to a constant rate for $N \geq 2560$. The residual rating variance can be explained by the inherent random variation of the match outcomes. It was verified that the parameter estimates are unbiased for $N \geq 2560$.

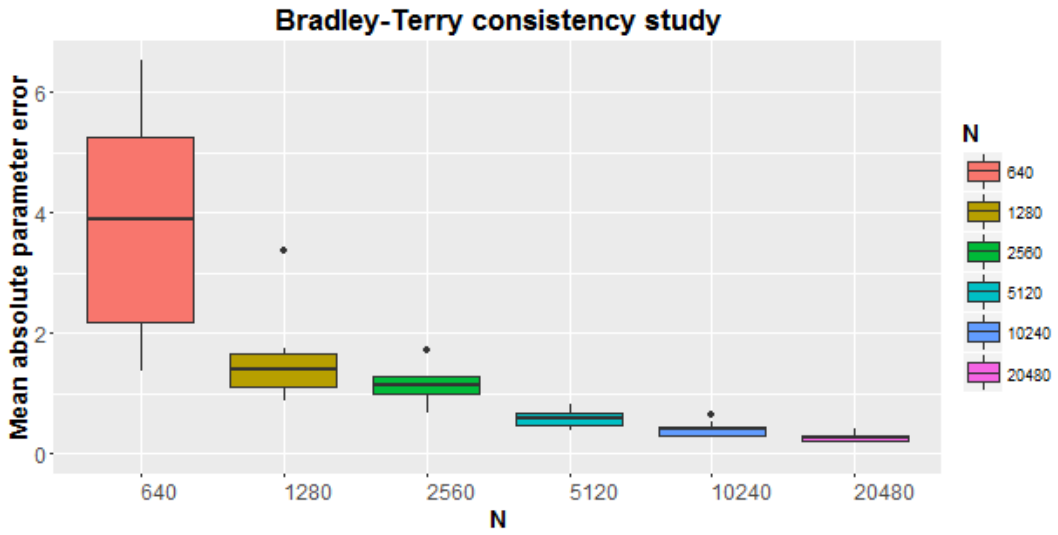

Figure 2.6: Bradley-Terry consistency study. N matches are simulated for 54 teams. This is repeated 10 times for each value of N. The box plot of the mean absolute parameter error indicates consistency of the parameter estimates.

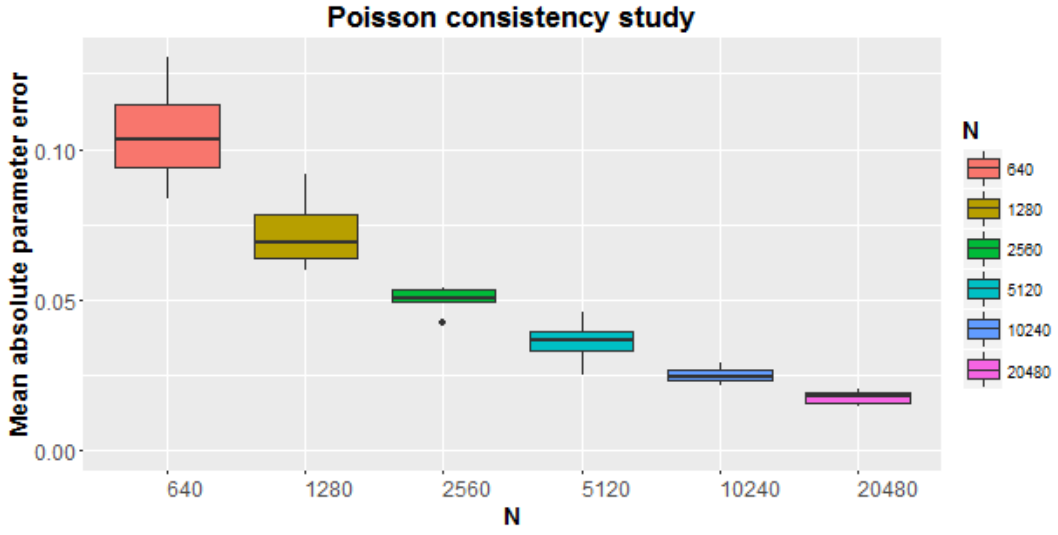

Figure 2.7: Poisson consistency study. N matches are simulated for 54 teams. This is repeated 10 times for each value of N. The box plot of the mean absolute parameter error indicates consistency of the parameter estimates.

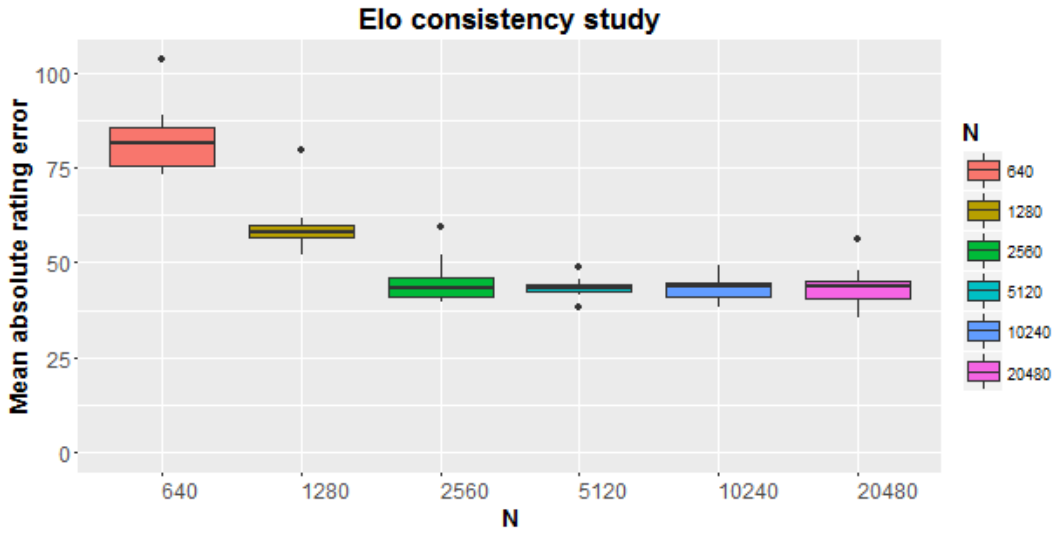

Figure 2.8: Elo consistency study. N matches are simulated for 54 teams. This is repeated 10 times for each value of N. The box plot of the mean absolute rating error indicates convergence of the Elo ratings.

3

Prediction of football matches

Six classes of statistical methods (2 Bradley-Terry, 2 Poisson, one blend of BT and Poisson and one Elo) are compared in this chapter with respect to their predictive performance. Only the actual outcome is used to compare the quality of the models. The first part of the chapter lists different measures of predictive performance in this setting and explains why the log loss is used as the preferred predictive performance metric. Both English Premier League matches and European national team matches are studied next using a full grid of the hyperparameters (match depreciation and probability relaxation). A simulation study using the preferred national team model is used in the final part of the chapter in order to come up with a detailed prediction of the expected performance of the participating teams in the UEFA EURO 2016 tournament.

3.1 Measures of predictive performance

The studied models are built to perform three way outcome prediction (home win, draw or home loss). Each of the three possible match outcomes is predicted with a certain probability but only the actual outcome is observed. The predicted probability of the outcome that was actually observed is used in all the measures of predictive performance. A specific case where the home team has a win probability of 0.5, a draw probability of 0.3 and a loss probability of 0.2 will be used to explain and compare four different measures of predictive performance. Outcome probabilities will be written as [prob home win, prob draw, prob home loss]. The ideal predictive performance metric is able to select the model which approximates the true outcome probabilities the best.

3.1.1 Majority rule

The majority rule is the simplest of the predictive performance measures and is used by Wang and Vandebroek (2011)[2] to compare their BT models to the FIFA ranking. The measure considers the highest of the three predicted outcome probabilities as the predicted outcome and counts the fraction of matches corresponding to the actual outcome. A higher majority rule value indicates a better predictive performance.
The majority rule for a model that predicts the three outcome probabilities as [0.8,0.1,0.1] would be 0.5 on average for the studied case of true outcome probabilities [0.5,0.3,0.2]. It would also be 0.5 on average had the actual probabilities been modeled. The majority rule can be used to determine whether a model identifies the favorite of a match correctly on average but fails to incorporate an appropriate penalty on the predicted probabilities of the observed outcomes.

3.1.2 Accuracy

The accuracy metric calculates the average predicted probabilities of the actual outcomes. The higher the accuracy, the better the predictive performance of the considered model. Assuming that the predicted probabilities are [0.8,0.1,0.1] for the studied case, average predicted probabilities of the actual outcomes are expected to be $0.5 \cdot 0.8 + 0.3 \cdot 0.1 + 0.2 \cdot 0.1 = 0.45$. The average accuracy would however be $0.5 \cdot 0.5 + 0.3 \cdot 0.3 + 0.2 \cdot 0.1 = 0.39 < 0.45$ on average, when the correct probabilities would have been predicted. The average accuracy metric improves the majority rule by taking the relative differences of the predicted probabilities into account but fails to be optimal for the true probabilities. Interestingly enough, the average accuracy is optimized and equal to the majority rule if the true favorite outcome is predicted with a probability of 1.

3.1.3 Bet return

The bet return assumes that the modeled probabilities are converted to decimal odds using $decimal\ odds = 1/probabilities$ and calculates the average bet return that would be paid out (based on the actual outcomes). It can be interpreted as three times the average fraction of the stakes that has to be reimbursed to the gambler, assuming no overround by the bookmaker. It is further assumed that the gamblers bet on outcomes proportional to their true probabilities, independently of the offered odds. If the true probabilities equal the modeled probabilities, the average bet return is always 3: $\frac{True\ prob\ 1}{True\ prob\ 1} + \frac{True\ prob\ 2}{True\ prob\ 2} + \frac{True\ prob\ 3}{True\ prob\ 3}$. Bet returns should be minimized to improve the predictive performance. Bet returns are higher than 3 for *most* other modeled sets of outcome probabilities. If the modeled probabilities are [0.8,0.1,0.1] for the studied case, the average bet return equals $0.5 \cdot \frac{5}{4} + 0.3 \cdot 10 + 0.2 \cdot 10 = 5.625 > 3$. The modeled outcome probabilities that match the minimum of the bet return do however not correspond with the true probabilities. The same issue was mentioned in the discussion of the *Accuracy* metric. If linear relaxation is used, the optimum is found with a relaxation parameter of 0.32 at

[0.418,0.316,0.265], leading to a bet return of 2.897. The lowest possible bet return decreases asymptotically to 1 for matches with an increasingly strong favorite.

3.1.4 Log loss

The average log loss criterion for N matches is defined as $log\ loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{3} y_{ij} \cdot log(P_{ij})$ where P_{ij} represents the modeled probabilities and y_{ij} is 1 if the result of game i is j and $y_{ij} = 0$ otherwise. The log loss criterion should be minimized to improve the predictive performance. Assuming all match outcome probabilities to be identical in the different matches and treating the outcome probabilities as the target estimators results in a minimized log loss criterion by setting P_{ij} equal to the true probabilities. The log loss for the studied case is expected to be 1.26 on average for the predicted probabilities [0.8,0.1,0.1] and 1.03 for the true probabilities. It was verified that no other set of probabilities resulted in a lower log loss value. The log loss criterion is typically used as the performance metric in data science competitions[16] when the goal is to predict probabilities for a set of possible outcomes.

3.1.5 Conclusion

The log loss criterion is the preferred metric as it is the only one of the four considered performance metrics having a unique optimum at the true probabilities. It should however be noted that optimizing the log loss criterion for matches with different true probabilities (as is typically the case) gives different weights to the matches. This is the case because the variation in log loss values heavily depends on the true probabilities. For a match with a heavy favorite where the true probabilities are [0.8,0.1,0.1], the optimum log loss lies at $-(0.8 \cdot log(0.8) + 2 \cdot 0.1 \cdot log(0.1)) = 0.64$ and getting all probabilities wrong by a little like in [0.7,0.15,0.15] leads to an inflated log loss value of 0.85, relatively far off from the optimum. If the true probabilities are however [$\frac{1}{3}$,$\frac{1}{3}$,$\frac{1}{3}$], the optimum log loss lies at 1.10, and getting the probabilities wrong by a considerable shift like in predicted probabilities [0.5,0.3,0.2] leads to a log loss value of 1.17, which is relatively close to the optimum. Matches with a strong favorite have more influence on the average log loss than matches without a pronounced favorite.

3.2 Premier League prediction

A total of 2850 Premier League matches are predicted using the six considered statistical methods. The analyzed period includes the seasons between 2000-2001 and the season 2014-2015. The results of the predictive performance comparison are presented and discussed after a short overview of the match data, the bookmaker data and the match data benchmark model.

3.2.1 Data

Premier League match results data

The data source of the English Premier League was supplied by James Curley through the engsoccerdata[9] package and contains results of all top 4 tier football leagues in England since 1888. The dataset contains the date of the match, the teams that played, the tier and division of the match as well as the result. No preprocessing was required.

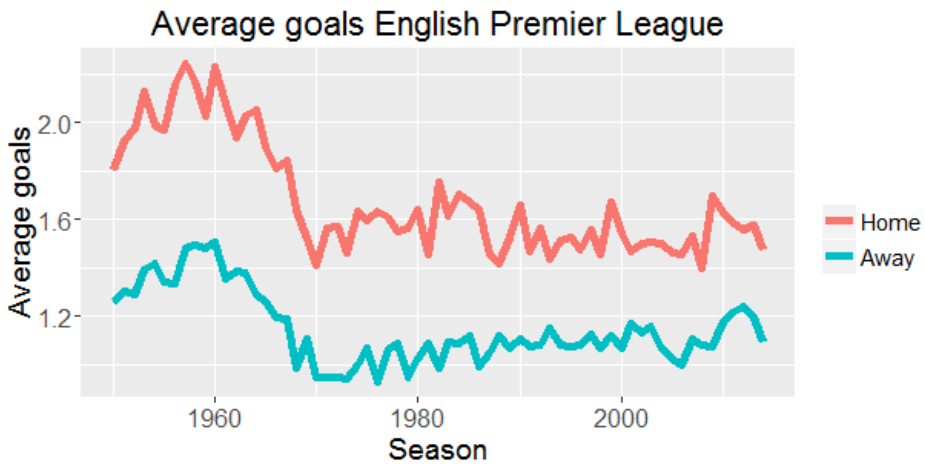

Figure 3.1: Average goals scored by teams in the English Premier League in the period 1960-2015. The number of goals seem more or less stable since 1970 with a marginal upward trend in the average away goals.

Figure 3.1 illustrates the evolution of the average number of goals scored by English Premier League teams in the period 1960-2015. Teams used to score significantly more goals before 1970. Naturally, this resulted in less draws compared to the period after 1970. It should be noted that the English Premier League was the first European football league that converted its ranking system from 2 to 3 points for a win. The English Premier League made the conversion in the 1981 season.

Univariate analysis of the home and away goals, scored during the modeled period (2000-2015) revealed that the Poisson distribution is a good approximation. This is illustrated in Figure 3.2 for the distribution of the home goals. However, testing for the goodness of fit of the Poisson

distribution for the home goals using a Chi-squared test based on 100,000 Monte-Carlo replicates ($p < 0.0001$) indicates a major deviation from the Poisson distribution. The observed goal counts are zero inflated with heavy tails for high numbers of goals scored. A possible explanation could be that the number of goals scored is a mixture of different Poisson distributions. Such a mixture distribution also contains zero inflated counts and heavy high end tails when it is approximated using a single Poisson distribution.

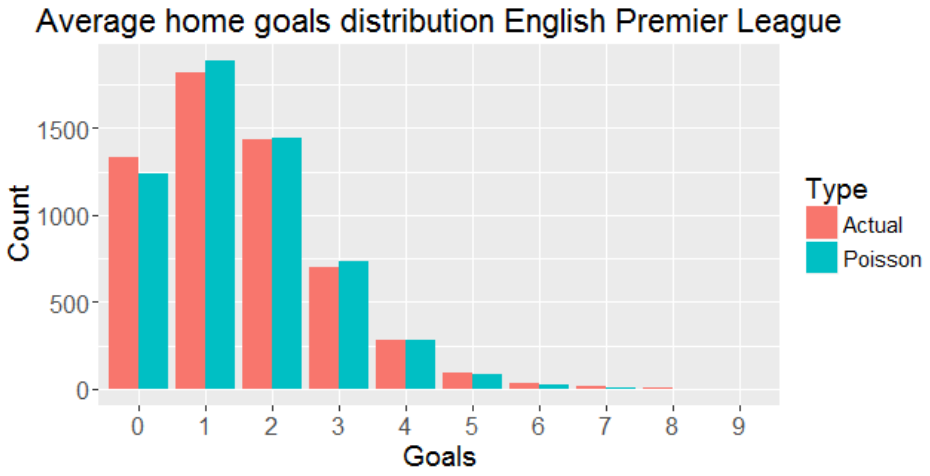

Figure 3.2: Comparison of the actual home goals count to the modeled counts using a Poisson model without overdispersion. The analyzed period includes the seasons between 2000-2001 and the season 2014-2015.

Testing for independence of the home and away goals using a Chi-squared test based on 100,000 Monte-Carlo replicates resulted in a p-value of 0.17. It cannot be concluded that there is enough evidence of a statistically significant relationship between the home and away goals at $\alpha = 0.05$. This is interesting since one would intuitively expect a significant negative correlation between the number of home and away goals. Testing the significance of the correlation $\rho = -0.05$ did confirm this intuitive thought with a p-value < 0.001 and a 95% confidence interval $[-0.072, -0.021]$.

Bookmaker data and match statistics

The website http://www.football-data.co.uk/englandm.php/[10] hosts a separate excel file with match data and odds of each English Premier League season since the season 2000-2001. Decimal odds are available for thirteen major betting companies. Nine bookmakers (Bet365, Bet & Win, Gamebookers, Interwetten, Ladbrokes, Sportingbet, Stan-James, VCBet and William Hill) contained odds for at least half of the matches, the other four were discarded. Decimal odds are translated to outcome probabilities assuming fixed over-round margins on all three outcomes. If the observed odds are [2,3,3.5] for a home win, a draw and a home loss respectively, the transformed probabilities become $[\frac{1}{2}, \frac{1}{3}, \frac{1}{3.5}]/(\frac{1}{2} + \frac{1}{3} + \frac{1}{3.5}) \approx$

$[0.447, 0.298, 0.255]$.
Several match statistics are available in the same datasets but the analysis is restricted to match statistics that are available in all matches. These statistics cover the total shots, shots on target, corners, fouls as well as the yellow and red cards for both the home and away teams.

3.2.2 Match data model

The match statistics from the previous paragraph are used in a proportional odds logistic regression model as a benchmark for the predictive model performance. The difference in total shots, shots on target, corners, fouls as well as yellow and red cards are used as the predictors. Predictive performance of the model is measured using the average log loss of 10-fold cross validation. The folds are kept constant throughout all analyses in order to obtain reproducible results. The match data model can obviously not be used to predict the match result but serves as a pure benchmark since the match data statistics are only available as soon as the match is finished. Analysis of the entire data set reveals that the difference in shots on target is the most important predictor for the match outcome with a relative importance of 1 for each shot on target. Teams with more shots on target have higher probabilities to win the match. Interestingly enough, having additional shots off target reduces the probability of winning with a relative importance of 0.14. Having additional corners compared to the opponent also reduces the probability of winning with a relative importance of 0.28 for each additional corner. Having corners and shots off target can both be seen as indicators of a team's failure to finishing off chances conditional on the number of shots on target. The difference in the number of fouls is the only variable that has no statistically significant impact on the match outcome at $\alpha = 0.05$. The difference in yellow cards has a relative importance of 0.45 for each additional card. Teams with less yellow cards tend to do better on average. The relative importance of red cards is the highest of all variables at 2.38 in favor of the team with less red cards. The modeling process was repeated using all the two-way interactions but this did not result in significant interactions at $\alpha = 0.05$.

3.2.3 Predictive performance comparison

The number of teams equals 20 for each of the seasons. Matches are predicted for one season at a time and the first prediction takes place when half of the matches of that season have been played. Accordingly, every season is analyzed separately from previous years. This avoids the issue of assigning strengths to promoted teams and takes team changes during the summer transfer season into account. Matches are predicted in blocks such that the chronological blocks are as large as possible, containing each team only once. Predicting in blocks ensures that all the available match information is used in the training phase.
Models are varied by considering a full grid of the hyperparameters. The half period parameter, the match depreciation hyperparameter for all models except the Elo model, is varied between

20 and 600 in steps of 20. Match depreciation in the Elo model 1.2 is performed by setting K to $K = constant \cdot matchImportance$ where match importance is fixed at 2.5 and the constant is varied between 1 and 50 in steps of 1. The linear relaxation values are varied between 0 and 0.2 in steps of 0.005 for all models. Models do not have to be retrained with different values for the linear relaxation since the predicted probabilities are a direct function of the model without relaxation. A total of $(5 \cdot 30 + 50) \cdot 41 = 8200$ models are analyzed and the models with the lowest log loss values of all six method classes are compared to the two benchmark models: the match data model and the average bookmaker performance. The average bookmaker outcome is considered to be the entitlement of the considered models since it only considers information prior to the start of the match. The linear correlation coefficient between the modeled actual outcome probabilities of the considered models and the benchmark models is used to measure the association. The average bookmaker predictive performance is calculated by transforming the odds of nine major bookmaker firms to bookmaker probabilities and calculating the log loss for that combined probability.
Two analyses are performed. The first analysis includes all 2850 matches. Next, the two last rounds are eliminated from the analysis since many teams have nothing to play for in the last rounds. This could lead to a drop in the team performance and it is worth looking at the major trends without the last two rounds. A total of 2534 matches are played by teams where neither of the two plays one of the last two matches of the season. An interactive shiny[17] application was written for the purpose of analyzing the predictive model performances in a dynamic way.

All data

A visual representation of the log loss values of all models except the Elo variations is shown in Figure 3.3. Elo models are analyzed separately since the time decay unit is different. The figure shows 5967 of the modeled $5 \cdot 30 \cdot 41 = 6150$ log loss values that are less than 1.05. Every model is shown in a different color and the average log loss is plotted against the half period for all considered relaxation constants. The shiny application allows the user to choose one of the two considered hyperparameters on the x-axis and also offers the option to subset the data based on the other hyperparameter. Each of the three other alternative predictive performance measures can be chosen for the y-axis instead of the log loss criterion. It is apparent from the plot that the Poisson model is performing best and that the EBT model performs the worst of the five considered model classes. The plot itself is however quite crowded and is best inspected interactively. The shiny application allows zooming in and conditional hiding of methods which allows the user to inspect the model performances in great detail. The log loss values seem to level off for higher half periods but all five model classes were found to have an optimum between 180 and 340 days after zooming in. A pronounced interaction is present between the relaxation constant and the half period. The log loss value is optimal for lower levels of either the relaxation constant or the half period if the other one is higher. This makes sense since both hyperparameters can be considered to have a smoothing effect on the predictions. The log loss

values of the two benchmark models outperform all considered models. This should not come as a surprise since the benchmark models are constructed using a lot more information. The best of the two benchmark models is the match data model (log loss of 0.942). Interestingly enough, an averaging model blend of the match data model and the average bookmaker model (log loss of 0.957) improves the predictive performance drastically to 0.922. Match data alone results in better predictions than the average bookmaker model but adding the bookmaker odds prior to the match further improves the outcome predictions. A possible interpretation could be that better teams have a higher shot conversion rate. Figure 3.4 plots the actual outcome probabilities of the best model, a Poisson model with a half period of 180 days and a relaxation constant of 0.02, against the average bookmaker outcome probabilities. The linear association is quite strong with a Pearson correlation coefficient of 0.941.

Figure 3.5 shows the results for the Elo models. The log loss is analyzed against the Elo update constant for various relaxation constants. The main findings are identical to the analysis above which excluded Elo models.

Table 3.1 summarizes the analysis by comparing the best performing models of each of the six considered classes to the two benchmark models. The best non-benchmark model is the Poisson model, followed by the extended Poisson, combined BTP, BT, Elo and EBT models respectively. Poisson models generally outperform BT models. This was to be expected since Poisson models use more information. Scaling the match importances by considering the win margin in the combined BTP model slightly improved the performance of the BT model. The best Elo model is ranked fifth and falls between the BT and EBT models. Increasing the number of team strengths from 1 to 2 resulted in worse average log loss ratings for the best Poisson model but the decline in predictive performance is a lot worse for the best BT model.

Using the log loss as the predictive performance metric is encouraged by considering the correlation between the average bookmaker model and the best performing model of the six considered model classes. The ranking of the best models of each class aligns perfectly with the correlation coefficient: a higher correlation with the average bookmaker model results in a lower log loss. A similar conclusion can be drawn for the relaxation of the best in class models. Higher relaxation constants for the best in class models result in higher log loss values, indicating worse predictive performance. The size of the relaxation can be seen as a measure of how much the models overfit the historical results. The strength of the association between the best in class models and the benchmarks is very different. Correlation with the bookmaker models is typically high (> 0.8) and indicates that the considered models pick up on similar historical match information as the bookmakers. The correlation with the match statistics model however is a lot lower (between 0.4 and 0.5). The correlation between the benchmark models is a little bit higher at 0.534. This does not say much about the quality of the considered model and average bookmaker predictions since different information is used to build the models.

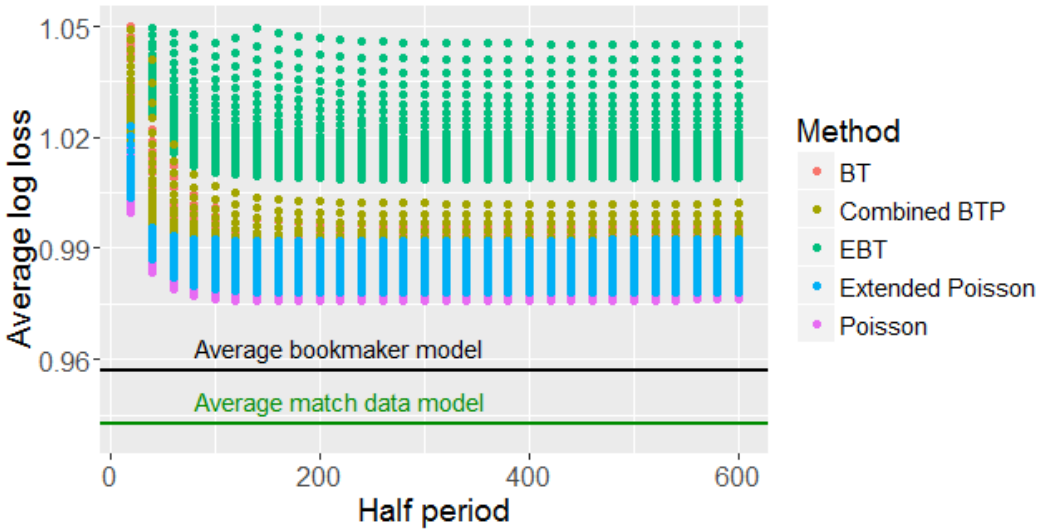

Figure 3.3: Model comparison graph using all 2850 second season half English Premier League matches in the period between the 2000-2001 and the 2014-2015 season. Elo models are excluded since they have a different time decay unit. Each model is fit using different half periods and relaxation parameters. The model with the lowest log loss value is a Poisson model with a half period of 180 days and a relaxation constant of 0.02.

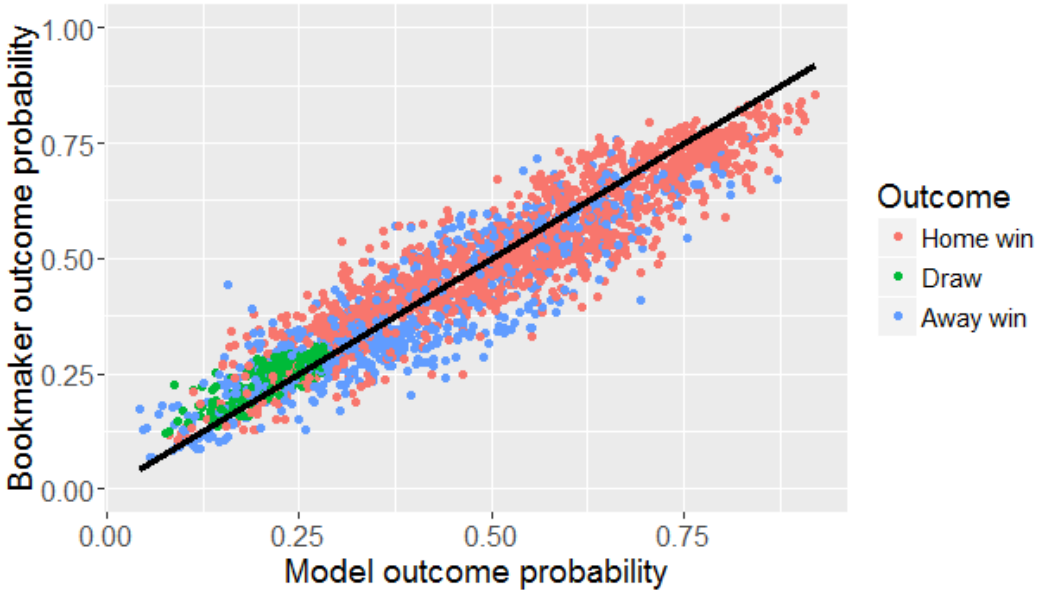

Figure 3.4: Comparison of the average bookmaker actual outcome probability and the actual outcome probability for the Poisson model with half period 180 days and relaxation constant 0.02. All 2850 outcome probabilities for the second season half English Premier League matches in the period between the 2000-2001 and the 2014-2015 season are shown. The Pearson linear correlation coefficient is 0.941 *and the black line* $y = x$ *represents coinciding predictions of the two models.*

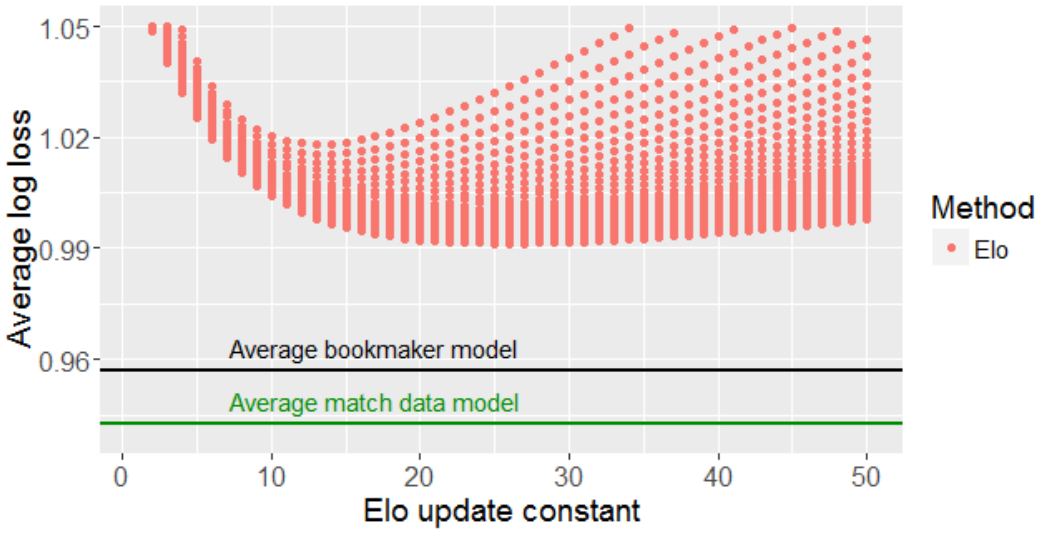

Figure 3.5: Elo model comparison graph using all 2850 second season half English Premier League matches in the period between the 2000-2001 and the 2014-2015 season. The Elo models are fit using a full grid of Elo update constants and relaxation parameters. The model with the lowest log loss value is the model with an Elo update constant of 26 and a relaxation constant of 0.16.

Model Class	**Lowest log loss**	**Model Class Rank**	**Time depreciation constant best model**	**Relaxation constant best model**	**Correlation match statistics model**	**Correlation average bookmakers model**
BT	0.986	6	HP = 340	0.075	0.459	0.913
Poisson	0.976	3	HP = 180	0.02	0.501	0.941
EBT	1.009	8	HP = 280	0.185	0.406	0.810
Extended Poisson	0.978	4	HP = 300	0.025	0.492	0.931
Combined BTP	0.985	5	HP = 220	0.100	0.449	0.918
Elo	0.991	7	Elo const = 26	0.160	0.495	0.878
Match statistics	0.942	1	/	/	1	0.534
Bookmakers avg	0.957	2	/	/	0.534	1

Table 3.1: Comparison table for the best performing models of each of the six considered classes with respect to the average log loss criterion and two benchmark models. The 2534 matches where none of the two teams played one of their last two season matches during the period between the 2000-2001 and the 2014-2015 season are considered. The best non-benchmark model is the Poisson model with a half period of 180 days and relaxation constant 0.02.

Model Class	Lowest log loss	Model Class Rank	Time depreciation constant best model	Relaxation constant best model	Correlation match statistics model	Correlation average bookmakers model
BT	0.985	6	HP = 580	0.075	0.457	0.916
Poisson	0.975	3	HP = 300	0.015	0.497	0.945
EBT	1.009	8	HP = 540	0.190	0.401	0.810
Extended Poisson	0.977	4	HP = 580	0.015	0.487	0.933
Combined BTP	0.984	5	HP = 360	0.095	0.446	0.921
Elo	0.991	7	Elo const = 26	0.155	0.496	0.880
Match statistics	0.945	1	/	/	1	0.535
Bookmakers avg	0.957	2	/	/	0.535	1

Table 3.2: Comparison table for the best performing models of each of the six considered classes with respect to the average log loss criterion and the two benchmark models. The table covers 2534 matches from the second season half of English Premier League matches in the period between the 2000-2001 and the 2014-2015 seasons. Matches where at least one of the two teams played one of their final two matches of the season are excluded from the calculations (316 excluded matches). The best non-benchmark model is the Poisson model with a half period of 300 days and a relaxation constant of 0.015.

Exclude last season matches

Excluding the last two season matches from the analysis did not result in major changes of the conclusions. The Poisson model is still outperforming the other considered models. Table 3.2 summarizes the analysis by comparing the best performing models of each of the six considered classes to the two benchmark models. The only apparent difference versus including all matches is observed in the half period on the best performing models for each class. The optimum is shifted to higher half period values for the best non-Elo models.

3.3 National team match prediction

A total of 4401 European national matches are predicted using the six considered statistical methods. The analyzed period includes all football matches between two European teams in the 1992-2015 period that were recognized as official by a national football association. The section kicks off with an overview of the data extraction and preprocessing steps followed by a short discussion of the extracted data and the results of the predictive performance comparison. The section concludes with the selection process of the preferred statistical model for national team match predictions. This model is used in the next section to simulate the results of the UEFA EURO 2016 tournament.

3.3.1 Data

Data extraction

National team match results were scraped from the impressive website http://eu-football.info/[11]. The platform contains a complete archive of all European national football team results (1872-2016) where at least one of the playing teams was European. The website also contains historical results of all European and domestic club competitions. Results are organized by national team and are extracted by parsing through all the national team pages using the R package rvest[12]. Variable time sleep intervals were added to the extraction logic in order to avoid overloading the server.

Data preprocessing

A total of 26929 match records were extracted by iterating over all national team pages. The first preprocessing step was to restrict the data to matches that are played amongst national teams that are currently still competing (e.g. drop matches of East Germany). Matches where one or both of the two teams were not the senior national team were dropped as well. Records played before 1980 were also discarded. Next, matches were paired by looking at records where the national team names and the date played were identical. This step is necessary since all national team matches are extracted twice, once from each national team page. All matches were paired perfectly so half of the records could be dropped. Finally, all matches before 1980 were dropped. A total of 6117 unique matches were retained at this point.
The match type is contained in the scraped data so match importance weights, according to the FIFA match importance factor, could be calculated easily. Determining if a match was played at home or on a neutral location was less straightforward. The approach considers a match a home match if the home team has played at that location during one of its qualifier matches or during one of the ordinary friendly matches. The remaining match locations of neutral matches were mapped to a country using the R package maps[18]. The matches where a match location was located in the country of the home team were assumed to be non-neutral matches.

The specific match page for all retained matches was visited in the second scraping iteration in order to extract additional match information. The detailed competition type as well as the goal scorer information (the name of the goal scorer and the match minute of the goal) were extracted at this point. The additional scraped information is not used in the analysis but is considered a nice addition to the interactive analysis of the predictive performance comparison.

Extracted data discussion

Figure 3.6 illustrates the evolution of the average number of goals scored by national teams between 1980 and 2015. Matches on neutral ground (356 out of 6117) were excluded from the analysis. The number of goals scored by home and away teams shows less variation after 1990. A slight downward trend for the number of home goals is apparent after 1990. The opposite observation can be made for the goals scored by the away team. It seems like the home effect becomes slightly less important. This was not investigated further since the trend does not appear to be overly strong.

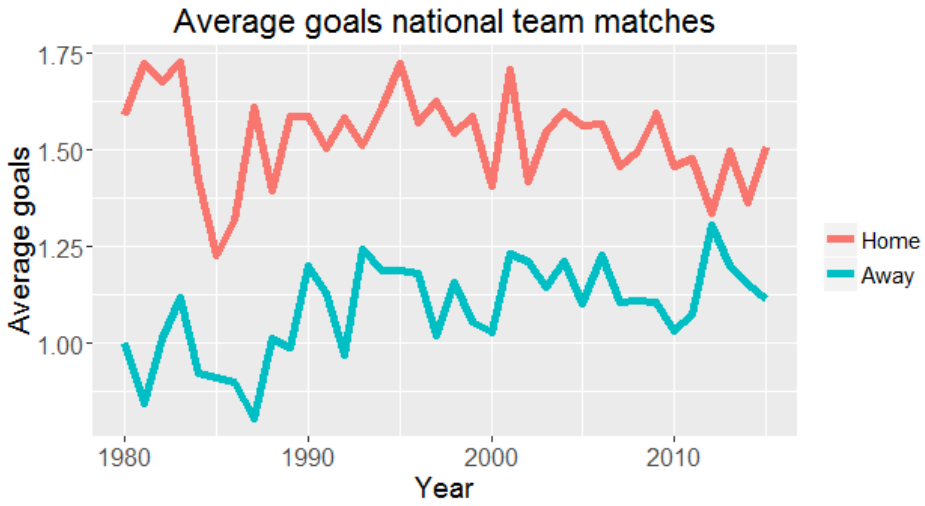

Figure 3.6: Average goals scored by national teams between 1980 and 2015. The number of goals scored by home and away teams shows less variation after 1990.

Univariate analysis of the home and away goals scored during the analyzed period (1992-2015) revealed that the Poisson distribution is an acceptable approximation. This is illustrated in Figure 3.7 for the home goals. However, testing for the goodness of fit of the Poisson distribution for the home goals using a Chi-squared test based on 100,000 Monte-Carlo replicates ($p < 0.0001$) indicates a major deviation from the Poisson distribution. The observed goal counts are zero inflated with heavy tails for high numbers of goals scored. The distribution is similar to the distribution of the home goals in English Premier League matches as shown in Figure 3.2 and is similar to the distribution of a mixture of Poisson variables.

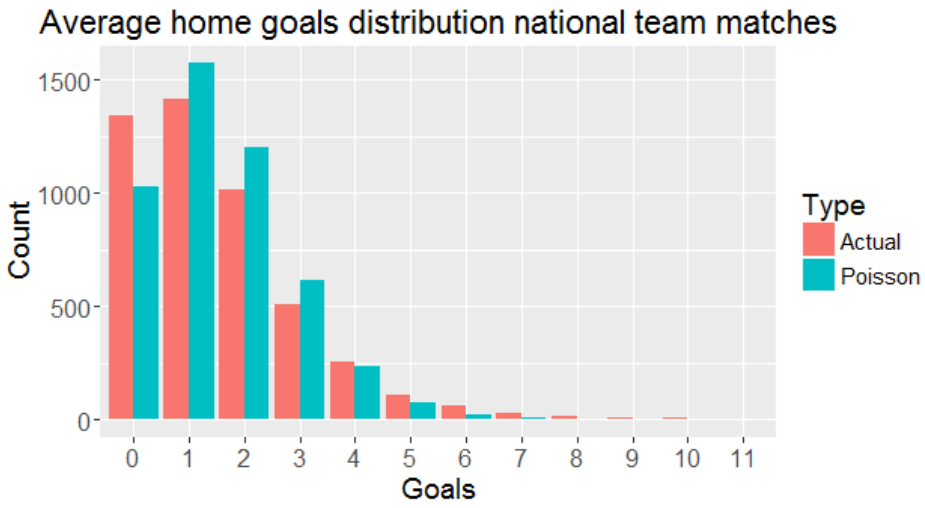

Figure 3.7: Comparison of the actual home goals count to the modeled counts using a Poisson model without overdispersion. The analyzed period covers all national team matches between 1992 and 2015.

Testing for independence of the home and away goals using a Chi-squared test based on 100,000 Monte-Carlo replicates resulted in a p-value of 0.02. This indicates that there is a statistically significant relationship between the home and away goals at $\alpha = 0.05$. Inspection of the residuals confirmed the intuitive thought of a significant negative correlation between the number of home and away goals. The correlation $\rho = -0.22$ was found to be highly significant ($p < 0.0001$) with a 95% confidence interval $[-0.248, -0.194]$.

3.3.2 Predictive performance comparison

Match data (1057 matches) from the period between 1980 and 1991 was used to construct the proportional odds logistic regression model that translates Elo strengths to match outcome probabilities as described in chapter 2. The remaining 5060 national team matches from the period 1992-2015 are used to train and evaluate the six considered model classes. Matches are predicted in blocks such that the chronological blocks are as large as possible while only containing each team once. The training period for all blocks is set to four years prior to the first predicted match. Consequently, the predictive performance is assessed by considering the 20-year period 1996-2015. The 1996-2015 period contains 4401 national team matches between two European teams and was predicted using a total of 446 blocks of matches. A training period of four years corresponds with the FIFA ranking which ignores matches that were played more than four years ago. Elo models are trained using the four year period 1992-1995 before predicting the first block of matches. The Elo strengths are updated after comparing the actual outcome to the expected outcomes of each block of matches and can be considered to have an ever increasing training period. A total of 54 national teams are considered in the analysis. Match importance

is taken into account using the FIFA weights.

Models are varied similarly as in the analysis of English Premier League matches by considering a full grid of the hyperparameters. The half period parameter, the match depreciation hyperparameter for all models except the Elo model, is varied between 20 and 600 in steps of 20. Match depreciation in the Elo model 1.2 is performed by setting K to $K = constant \cdot matchImportance$ where match importance is fixed at 2.5 and the constant is varied between 1 and 50 in steps of 1. The linear relaxation values are varied between 0 and 0.2 in steps of 0.005 for all models. Models do not have to be retrained with different values for the linear relaxation since the predicted probabilities are a direct function of the model without relaxation. A total of $(5 \cdot 30 + 50) \cdot 41 = 8200$ models are analyzed and the models with the lowest log loss values of all six method classes are discussed in further detail.
No extensive bookmaker or match data source was found for the national team matches in the considered timeframe. Consequently, no benchmark models can be constructed and the analysis is restricted to a relative comparison of the considered model classes.
Two analyses are performed and one additional filtering step is performed in both analyses. Matches are only used to assess the predictive performance of the models if there is considerable historical information available for both teams. National team matches where at least one of the teams has played less than 20 matches were excluded. This filtering reduced the number of matches to assess the predictive performance from 4401 to 4238. A similar filtering was not required when analyzing English Premier League data since new teams cannot be founded during the season resulting in a similar number of training matches for all teams. The first analysis includes all 4238 matches that contain sufficient historical information of both teams. The second analysis further restricts the considered matches to the ones where both teams do not belong to a national team region with a relatively limited number of inhabitants. The limit was set at 100,000 inhabitants such that as many countries as possible were excluded while still including all countries that will participate in the UEFA Euro 2016 championship. Excluding matches where one of the teams was Gibraltar, San Marino, Liechtenstein, Faroe Islands or Andorra reduced the number of considered matches to 3785.
A shiny application was developed for the purpose of analyzing the predictive performance comparison of national team matches interactively. It strongly resembles the shiny application to analyze the predictive performance when analyzing English Premier League data and contains some additional data selection and visualization features.

All data

Analysis of the predictive performance of the different considered model classes was performed in a similar way to the analysis of the English Premier League data and the discussion will be restricted to the main observations.
A clear interaction is present between the relaxation constant and the half period. The log

Model Class	Lowest log loss	Model Class Rank	Time depreciation constant best model	Relaxation constant best model
BT	0.876	4	600	0.05
Poisson	0.862	1	600	0
EBT	0.907	6	600	0.11
Extended Poisson	0.868	2	600	0
Combined BTP	0.875	3	600	0.06
Elo	0.884	5	Elo const = 20	0.07

Table 3.3: Comparison table for the best performing models of each of the six considered classes with respect to the average log loss criterion. A total of 4238 matches are used to compare the predictive performance of the models. The best performing model is the Poisson model with a half period of 600 days and a relaxation constant of 0.

loss value is optimal for lower levels of either the relaxation constant or the half period if the other one is higher. The interaction was also present in the analysis of English Premier League matches.

Table 3.3 summarizes the analysis by comparing the best performing models of each of the six considered classes. The best model is the Poisson model, followed by the extended Poisson, combined BTP, BT, Elo and EBT models respectively. It should be noted that this ranking is identical to the ranking when English Premier League data was analyzed and that the other elements from the discussion also apply to national team match models.

The finding that higher relaxation constants for the best in class models typically result in higher log loss values also applies to national team match models. The size of the relaxation constants can be seen as a measure of how much the models overfit the historical results.

A cause for concern is the observation that all the best in class models except the Elo model coincide at a half period of 600 days. This is the boundary of the hyperparameter space for the time depreciation constant. Inspecting the relation between the log loss and the half period at the boundary of the half period dimension did reveal that the log loss was decreasing marginally for increasing values of the half period. The second derivative of the log loss in function of the half period was found to be positive for all models. It seems safe to conclude that the order of the best in class models is likely to remain unchanged when larger half periods are also considered given the relatively high difference in log loss values with respect to the log loss changes at the boundary of the half period dimension. At this point it is important to remember that the main goal is to select a preferred model and that the relative ranking is of less interest. The best Poisson model strongly outperforms the class with the second best model (Extended Poisson). The observation that the relative ranking remains unchanged in the English Premier League analyses and the national team analysis with all the available data justifies the decision to restrict the analysis to the half period range 20-600 days for all but the Poisson model. The best Elo model was found for an Elo update constant of 20 which is not at the boundary of the parameter space.

Model Class	Lowest log loss	Model Class Rank	Time depreciation constant best model	Relaxation constant best model
BT	0.928	4	600	0.075
Poisson	0.917	1	600	0.005
EBT	0.955	6	600	0.145
Extended Poisson	0.923	2	600	0.015
Combined BTP	0.926	3	600	0.085
Elo	0.934	5	Elo const = 19	0.10

Table 3.4: Comparison table for the best performing models of each of the six considered classes with respect to the average log loss criterion. Matches where Gibraltar, San Marino, Liechtenstein, Faroe Islands or Andorra are competing are excluded from the analysis resulting in a total of 3785 matches to compare the predictive performance of the models. The best performing model is the Poisson model with a half period of 600 days and a relaxation constant of 0.005.

Excluding small nations

Excluding national teams with a population of less than 100,000 inhabitants from the analysis did not result in major changes of the conclusions. The Poisson model is still outperforming the other considered models and the half periods of the best in class models are still situated at the boundary of the parameter space. Table 3.4 summarizes the analysis by comparing the best performing models of each of the six considered classes. The average log loss value for the best in class models is higher than the analysis that did not exclude matches from teams with a small number of inhabitants. An increase of the log loss was expected since the optimal average log loss value is higher for matches that are more balanced. Relaxation constants for all best in class models are also higher than the relaxation constants for the best in class analysis that considered matches between all European teams.

3.3.3 Preferred model selection

The selection of the preferred model to predict national team matches during the UEFA EURO 2016 tournament is restricted to Poisson models with a single team strength parameter. Poisson models with a single team strength parameter outperform all other five considered model classes and the analyses so far and require the least amount of relaxation to minimize the log loss criterion.

The restriction to Poisson models was encouraged by considering the extensive calculation time to fit the training models. Every additional half period - model combinations requires 446 maximum likelihood estimations of the model parameters since there are 446 blocks of matches to predict. Finding the maximum likelihood estimators takes about 8 seconds on average on a Lenovo S540 Thinkpad with 8 GB RAM, 4 logical processors and a clock speed of 1.80 GHz. Calculating an additional half period - model combination consequently takes about an hour. The code was parallelized and distributed across multiple systems but still requires extensive

computation time.
Matches against teams with less than 100,000 inhabitants are not considered since these matches are not representative for the matches during the UEFA EURO 2016 tournament.
The search space of Poisson models needs to be extended to find the optimal half period value. Half period values between 620 and 1600 in steps of 20 days were used to fit the Poisson models. Figure 3.8 plots the average log loss values of Poisson models for different combinations of the half period and relaxation hyperparameters. It can be concluded that the best log loss value corresponds with a Poisson model with a half period value of 1320 days and a relaxation constant of 0.01. The optimal log loss value was improved from 0.917 (half period 600 days and relaxation constant 0.005) to 0.916 by extending the search space. The improvement is very small so it makes more sense to select a model which performs slightly worse on the studied period but has an easier interpretation. The Poisson model with relaxation constant 0 serves this purpose. It is also optimized for a half period of 1320 days and this model is selected as the preferred model to predict national team matches.

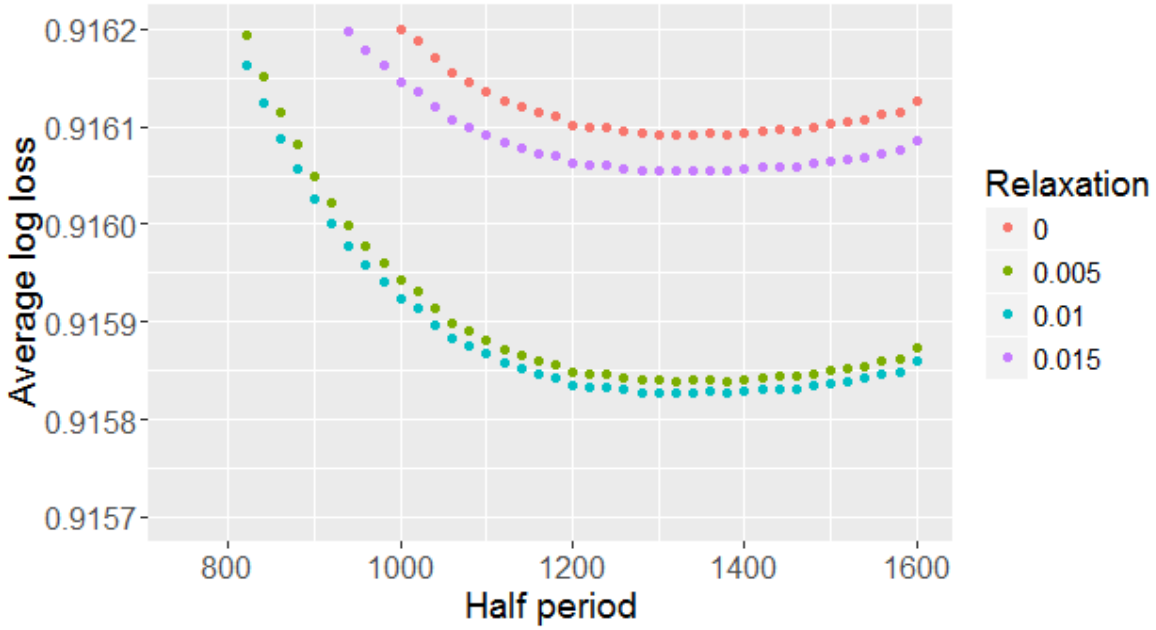

Figure 3.8: Comparison graph of the predictive performance comparison of various Poisson models for national team matches. Matches where Gibraltar, San Marino, Liechtenstein, Faroe Islands or Andorra are competing are excluded from the analysis resulting in a total of 3785 matches. The best Poisson model has a half period of 1320 days and a relaxation constant of 0.01.

3.4 UEFA EURO 2016 simulation study

The selection of a preferred model to predict national team match outcomes enables the prediction of outcome probabilities for any match played between two of the 54 modeled European teams. Furthermore, the selected model allows predicting the probability of actual match results since the selected model is a Poisson model that assumes independence of the goals scored by the home and away team, conditional on the modeled team strengths. This section uses the preferred model to predict matches from the UEFA EURO 2016 tournament repeatedly. Simulating the tournament many times gives valuable insight into the likely outcomes for the teams and the relative difficulty of the draw. The first part of this section explains the simulation procedure. The simulation results using the actual and the permuted draw are discussed next. The section concludes with a comparison between the ranking of the likely winners according to the simulation study, the bookmaker odds and the FIFA ranking.

3.4.1 Simulation details

The UEFA EURO 2016 tournament is played in France by 24 teams. Six groups of four teams were drawn on December 12, 2015 in Paris. The first two ranked teams progress to the knockout rounds as well as the four best third ranked teams of the six groups. Four knockout stages are played to select a winner.

The simulation logic starts with fitting the preferred national team match model parameters using the training period of 4 years prior to the start of the UEFA EURO 2016 championship (June 10, 2016). The friendly matches played in January and March of 2016 are included in the analysis but the friendly matches played in May and June were not considered due to the submission deadline of this text.

Team strengths are converted to the expected number of goals scored by both teams according to formula 1.1. The home effect is only applied to France in the calculation of the expected goal counts of both teams. Simulated goal counts for both teams are generated next using independent Poisson distributions.

Ideally, the Poisson model should be refit after each of the three group stage rounds and after the first three of the four knockout stages. Recalculating the model parameters using maximum likelihood estimation does however prove to be slow. Not refitting the model allows more simulations which reduces the variance of the estimated team outcomes at the cost of introducing biased estimates of the same team outcomes.

Articles 16-19 of the rules[19] for the UEFA EURO 2016 tournament were used to determine the ranking of the teams based on the group match results. The rules were also used to calculate the best four thirds of the six groups and pair the correct teams in the knockout rounds.

Knockout matches that end in a draw after ninety minutes are extended by two 15 minute periods. The simulation study handles this by simulating a new match with the expected goals

scored set to one third of the main part of the match. Shootouts are simulated by assigning a random winner.

3.4.2 Simulation results

Models are not refit throughout the tournament. It was judged that the reduction in variance of the team outcome estimates was more important than the introduction of a small bias. The preferred model to predict national team matches has a relatively high half period at 1320 days. Matches played four years ago are considered about half as influential as matches played today. Consequently, the estimates of the outcomes are not expected to be biased heavily by not refitting the model.
The UEFA EURO 2016 tournament was simulated 100,000 times and the final outcomes of the 24 teams will be analyzed using both the actual draw and different permutations of the draw.
The permuted draw was calculated according to the draw rules where the 24 teams are divided into four pots based on UEFA national coefficients. One team from each pot is assigned to each group (France is always assigned to Pot 1).

Actual draw

A bar chart of the win frequency of the participating teams is shown in Figure 3.9. France and Germany are the main favorites to win the tournament according to the simulation study, each winning in about 15% of the simulations. Belgium and Spain are expected to win the tournament with a probability close to 12% and are followed by England that is ascribed a win probability of about 10%. Austria, Russia and Ukraine are the main outsiders according to the simulation study.
Detailed outcomes for Belgium are depicted in Figure 3.10. The plot reveals that the most likely outcome for Belgium is a round of 16 exit even though Belgium is on average the favorite to win each knockout round.

Permuted draws

A bar chart of the win frequency of the participating teams is shown in Figure 3.11. France is still the main favorite but is not followed as closely by Germany assuming permuted draws. Belgium and Spain are still second tier favorites that are expected to win the tournament with a probability of about 12.5% and England is still ascribed a win probability of about 10%. Austria, Russia and Ukraine remain the main outsiders according to the simulation study.
Detailed outcomes for Belgium are depicted in Figure 3.12. Comparison with the actual draw learns that Belgium is much more likely to exit the tournament in the group phase assuming

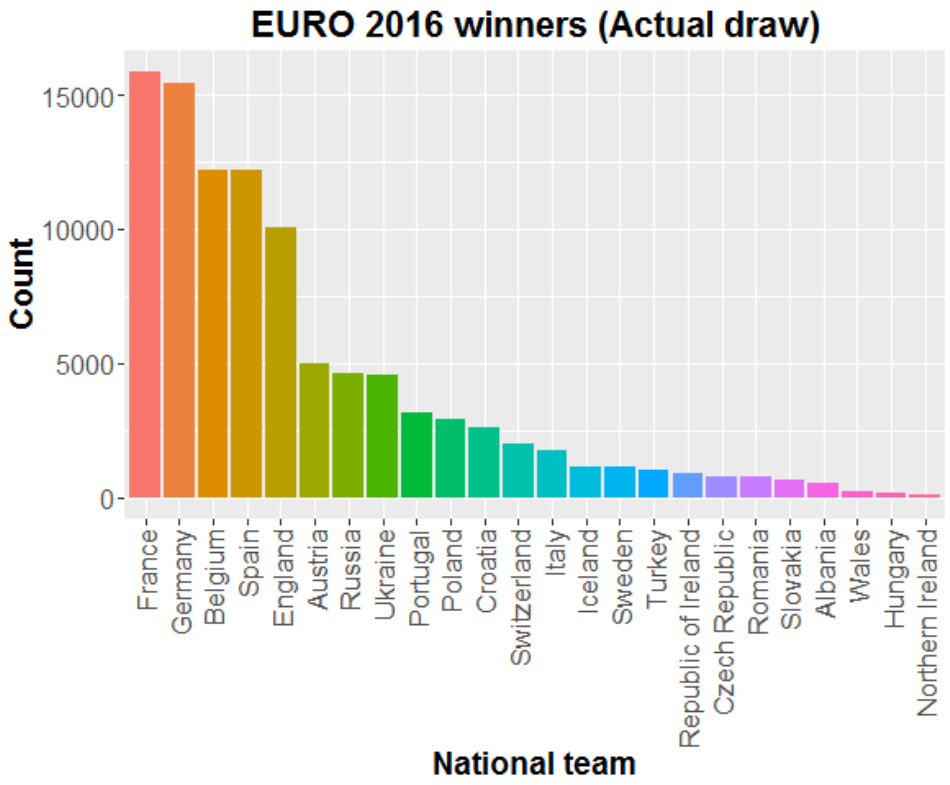

Figure 3.9: Simulated winners using the actual draw. France and Germany are the main favorites to win the tournament based on 100,000 simulations.

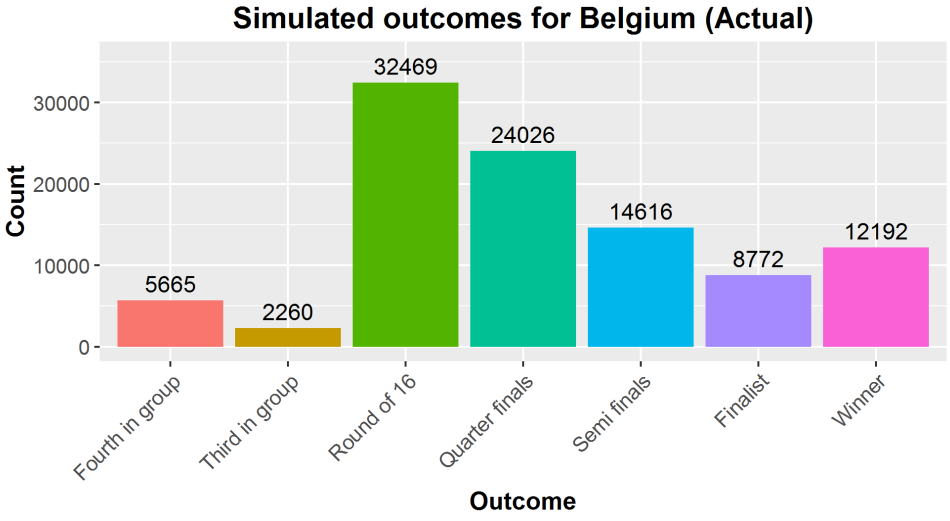

Figure 3.10: Simulated outcomes for Belgium using the actual draw. Belgium came out as the winner of the tournament in 12,192 of the 100,000 simulations.

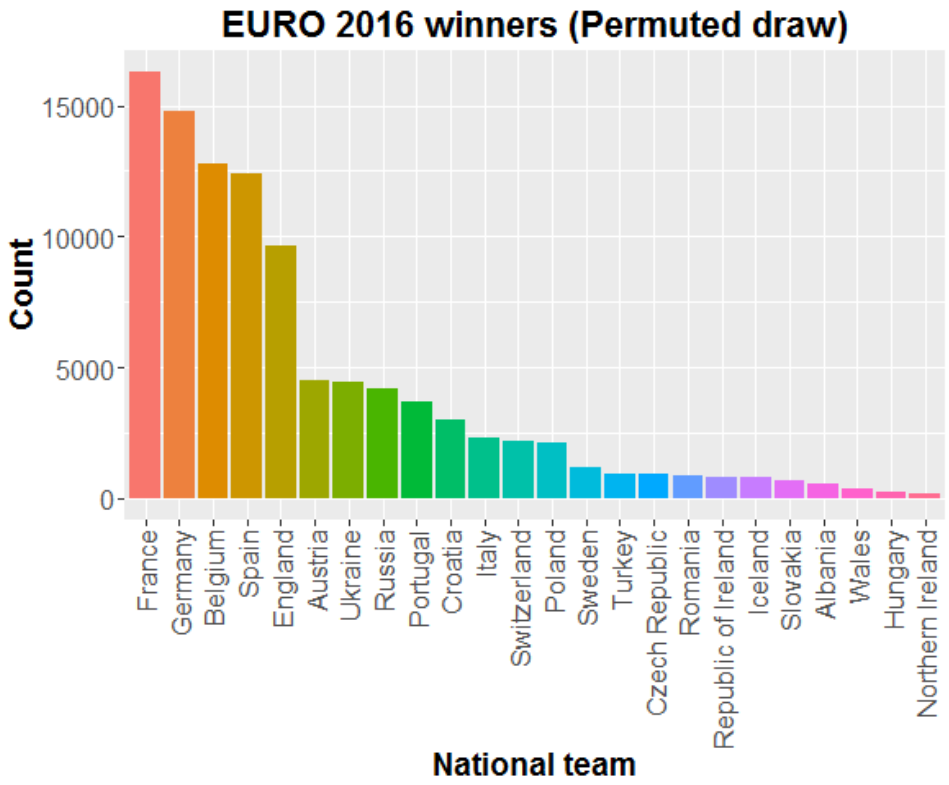

Figure 3.11: Simulated winners using permuted draws. France is still the main favorite to win the tournament followed by Germany based on 100,000 simulations of the tournament.

permuted draws than the case that considers the actual draw (10.4% versus 7.9 %). It can be concluded that Belgium was assigned to a relatively easy group. The exit rate of Belgium in the rounds of 16 and 8, conditional on making it through the group stages, is however a lot higher when the actual draw is considered opposed to the permuted draw (61.4% versus 59.2%). The elevated exit rate in the round of 16 of the actual draw can be explained by the group Belgium is assigned to. Belgium plays in Group E and the winners of Group E and F play the second ranked team of the other group. Winners of the other four groups play a third ranked team of another group. Lower ranked teams in the group are more likely to be weaker so this rule is a disadvantage for Belgium since it is the major favorite in its group. The elevated exit rate of Belgium in the quarter finals in the actual draw compared to the permuted draws can be explained by the likely opponent. If both Belgium and Germany win their group and progress to the quarter finals, they will play for a semi-final ticket. This is a likely scenario in which Germany would be a slight favorite according to the preferred national team match model. All mentioned differences between the actual and the permuted draws are highly significant ($p < 0.0001$) given the large sample size.

Figure 3.13 digs deeper into the difference in win proportion for all participating teams. P-value analysis of the proportion tests reveals that Poland shows the biggest relative difference in win

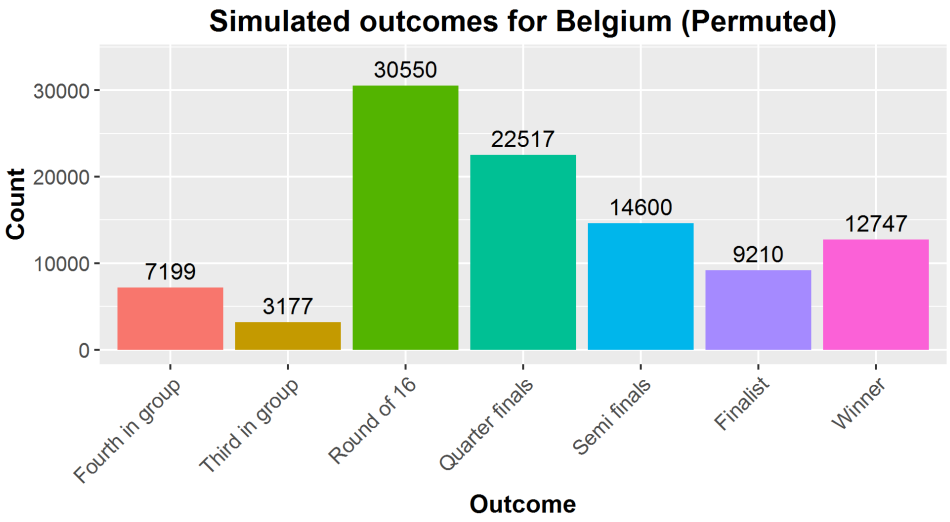

Figure 3.12: Simulated outcomes for Belgium using permuted draws. Belgium came out as the winner of the tournament in 12,747 of the 100,000 simulations.

proportion when comparing the actual draw to the permuted draws. Poland has a win probability of 2.1% according to the simulation study of the permuted draws and a win probability of 2.9% according to the actual draw. The difference can be largely explained by the weakness of Northern Ireland, which is also assigned to Group C. The probability to finish last in the group is estimated at 34.4% assuming permuted draws and 17.8% in the actual draw.

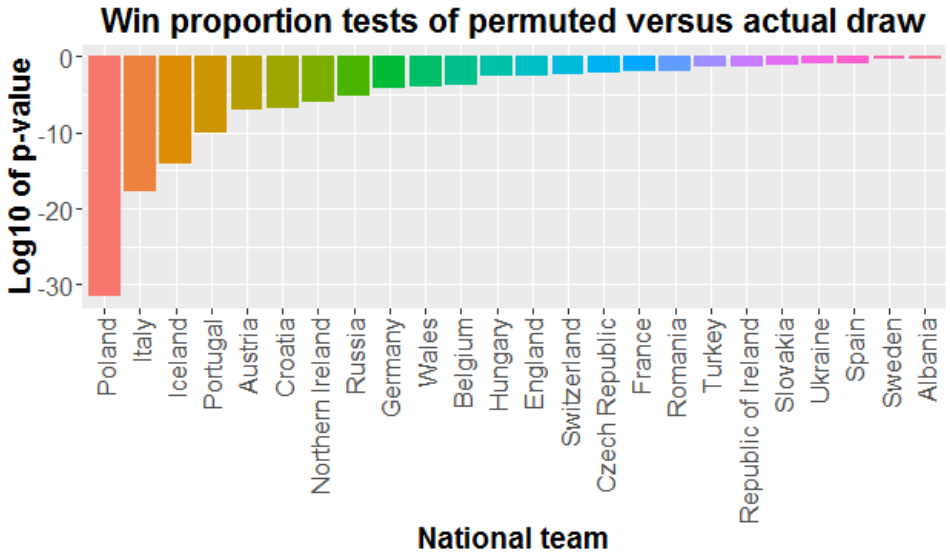

Figure 3.13: P-values of comparing the win proportion for all participating teams on a log scale. Significant p-values indicate either a favorable or disadvantageous draw.

3.4.3 Comparison of winner predictions

The ranking of the simulated winners using the actual draw is compared to the FIFA ranking and the ranking according to bookmaker odds. Bookmaker odds are assumed to be the gold standard and the quality of the ranking is judged by the similarity with the bookmaker ranking. Bookmaker odds were collected from the website of bwin https://www.bwin.be on May 15, 2016. The FIFA ranking of May 5, 2016 is used for the comparison. Tied ranks were averaged for both the bookmaker odds and the FIFA ranking.

Table 3.5 summarizes the ranks of the three considered ranking strategies. The ranks of the simulation study are strongly correlated with the bookmaker ranks ($\rho = 0.85$) with a 95% confidence interval $[0.684, 0.934]$. The correlation between the FIFA ranks and the bookmaker ranks was found to be lower ($\rho = 0.67$) with a 95% confidence interval $[0.372, 0.847]$. Testing for significant difference of the correlations yielded a p-value of 0.15. It cannot be concluded that the simulation study rank has a different correlation with the bookmaker rank than the correlation between the FIFA rank and the bookmaker rank at $\alpha = 0.05$.

The simulation study rank has a mean absolute rank difference of 2.7 compared to the bookmaker rank while the mean absolute rank difference between the FIFA rank and the bookmaker rank is 4.5.

Team	Bookmaker rank	Simulation study rank	FIFA rank
Germany	1	2	2
France	2	1	12
Spain	3	4	3
England	4	5	5
Belgium	5	3	1
Italy	6	11	9
Portugal	7	9	4
Croatia	8	10	14
Austria	9	6	6
Poland	10	13	17.5
Russia	12	8	17.5
Wales	12	22	15
Switzerland	12	12	8
Czech Republic	15	16	19
Iceland	15	19	22
Turkey	15	15	7
Sweden	17.5	14	23
Ukraine	17.5	7	13
Romania	20	17	11
Slovakia	20	20	21
Republic of Ireland	20	18	20
Hungary	22	23	10
Northern Ireland	23	24	16
Albania	24	21	24

Table 3.5: Ranking comparison of the bookmaker odds of the UEFA EURO 2016 tournament with the simulated winners using the preferred national team match model and the FIFA ranking of May 5, 2016. The simulation study rank has a mean absolute rank difference of 2.7 compared to the bookmaker rank while the mean absolute rank difference between the FIFA rank and the bookmaker rank is 4.5.

4

Conclusion

Three main classes of models are considered for all analyses: Bradley-Terry (BT) models that only consider the match outcomes (home win, draw or home loss), Poisson models that use the match results (e.g. 3-2 home win) and model both the number of home and away goals using a Poisson distribution, and Elo models which also use the match results. All studied statistical models are optimized with respect to the predictive performance on future matches. The FIFA ranking however is not optimized to generate accurate predictions. This major difference and the presence of other flaws in the FIFA ranking make the considered statistical models better indicators of actual team strengths.

The predictive comparison of the considered models revealed that Poisson models outperform BT models. Modified BT models that also take the goal difference into account were found to have a slightly better predictive performance than the BT models, but a clearly worse predictive performance than Poisson models. The BT and Poisson models were both fit using one and two team strength parameters. The best predictive models were found to contain a single team strength. The reduction in bias by using two team strengths can be concluded to be less important than the introduction of additional variance by using two strength parameters for each team.

Elo models are widely used to rank teams using a rating scheme and an interesting application is the female FIFA ranking. The simple nature of the update formula 1.1 makes it easy to understand and apply. Poisson models were found to generate better predictions than Elo models on average at the cost of a more complex model. Consequently, Poisson models can be concluded to be superior to Elo models if predictive performance is more important than simplicity.

The predictive performance comparison of English Premier League matches in the period 2000-2014 between the studied models and the odds of the bookmakers revealed that bookmakers generate slightly better predictions on average. This was expected since bookmakers can model many likely important factors such as injuries, the weather or suspensions directly. The outcome predictions of the studied models are however strongly correlated with the bookmaker predictions. The best Poisson model was found to result in a correlation of $\rho = 0.94$ between the Poisson actual outcome probability and the bookmaker actual outcome probability.
An extensive study of national team matches was used to select a preferred model to simulate the UEFA EURO 2016 tournament. The simulation study resulted in valuable insights of likely team outcomes and the relative luck of the draw compared to permuted draws. France is a slight favorite to win the tournament, closely followed by Germany. Belgium and Spain are the major outsiders. Analysis of the permuted draw revealed that Poland has benefited the most from the actual draw.

Future work

The research findings can be further developed in various ways. A first improvement could be to explore other modeling approaches. One way to achieve this would be to incorporate additional predictors such as the weather, suspensions and injuries. Previous match specific parameters such as the shots on target are also likely to improve the predictive performance. Several of the generated models can also be combined to improve the predictive performance.
Currently, linear relaxation is used to make the predicted outcome probabilities less extreme. However, it would likely be better to take the order of match outcomes into account. If the first team is a strong favorite according to the model, it makes more sense to inflate the probability of a draw by a greater margin than the inflation margin of the outcome where the other team wins.
Elo rating differences are transformed to outcome probabilities using a proportional odds logistic regression model. The transformation could also be performed by dropping the restriction of proportionality. Extending the data set to train the model is another relevant next step to consider.
The home effect constant could be analyzed in further detail as well. It would be interesting to research if the home effect is related to the required travel distance. Is the home effect as strong during a derby as during a match where the away team had to travel for 24 hours? Other fatigue related predictors are also likely to improve the outcome predictions.

References

[1] FIFA. Calculation of points fact sheet. Available at http://resources.fifa.com/mm/document/fifafacts/r%26a-wr/52/00/97/fs-590_10e_wrpoints_english.pdf. Accessed: 23/04/2016.

[2] C. Wang and M. Vandebroek. A model based ranking system for soccer teams. *Available at SSRN 2273471*, 2013.

[3] M. Maher. Modelling association football scores. *Statist. Neerland*, 36:109–118, 1982.

[4] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society. Series D*, 52:381–393, 2003.

[5] M. Crowder et al. Dynamic Modelling and Prediction of English Football League Matches for Betting. *Journal of the Royal Statistical Society. Series D*, 51:157–168, 2002.

[6] FIFA. Calculation of points fact sheet. Available at http://resources.fifa.com/mm/document/fifafacts/r%26a-wwr/52/00/99/fs-590_06e_wwr-new.pdf. Accessed: 23/04/2016.

[7] EloRatings.net. The world football elo rating system, 2016. Available at http://www.eloratings.net/system.html. Accessed: 23/04/2016.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[9] James Curley. *engsoccerdata: Soccer Data 1871-2015*, 2015. R package version 0.1.4.

[10] http://www.football data.co.uk/. English premier league data files. Available at http://http://www.football-data.co.uk/englandm.php/. Accessed: 05/05/2016.

[11] eu football.info. European national football teams 1872-2016 matches database. Available at http://eu-football.info/. Accessed: 24/04/2016.

[12] Hadley Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2015. R package version 0.3.1.

[13] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[14] R. Davidson. On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments. *Journal of the American Statistical Association*, 65:317–328, 1970.

[15] M. Hvattum and H. Arntzen. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26:460–470, 2010.

[16] kaggle.com. Multi class log loss. Available at https://www.kaggle.com/wiki/MultiClassLogLoss/. Accessed: 05/05/2016.

[17] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2016. R package version 0.13.2.

[18] Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka, and Alex Deckmyn. *maps: Draw Geographical Maps*, 2016. R package version 3.1.0.

[19] UEFA.com. Regulations of the uefa european football championship 2014-2016. Available at http://www.uefa.com/MultimediaFiles/Download/Regulations/uefaorg/Regulations/02/03/92/81/2039281_DOWNLOAD.pdf. Accessed: 15/05/2016.