Can we predict the Eredivisie?

Predicting rankings in football using the Bradley-Terry model

H.C.N.J. Heijlema

Can we predict the Eredivisie?

Predicting rankings in football using the Bradley-Terry model

by

H.C.N.J. Heijlema

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Tuesday July 2, 2019 at 2:00 PM.

Student number:	4443926	
Project duration:	September 1, 2018 – June 27, 2019	
Thesis committee:	Dr. D. Kurowicka,	TU Delft, supervisor
	Dr. Ir. G. F. Nane,	TU Delft
	Drs. E. van Elderen,	TU Delft

This thesis is confidential and cannot be made public until July 2, 2019.

Preface

I'm a big football fan, who goes to the stadium at least once in every two weeks. Following football is one of my biggest hobbies. I can look at football statistics all day long. Using mathematics on football data therefore was a great opportunity, since I could combine my hobby with my study. This combination: mathematics and football, occurs more and more nowadays. The rise of mathematicians in football is happening as we speak. This made this research very relevant as well. I would like to thank my supervisor Dorotha Kurowicka for always being there to help me. Thank you for your patience.

H.C.N.J. Heijlema
Delft, June 2019

Contents

1
Introduction

Making predictions about football is a rapidly growing market. Bookies provide all different sorts of bets. The bets vary from the number of goals scored in a match to which team will win the league. More and more parties are getting involved in this market, using data and statistics to support their predictions.
Last year, the company Hypercube published a survey in which they talked about which team had the largest chance of relegating directly from the Dutch league at that moment. Their conclusion was that the team Roda JC had the largest chance of relegating directly, with a probability of 46.75%. At the end of the season however it was not Roda JC, but FC Twente that ended up at the bottom of the league. Hypercube assigned a probability of 29.36% to this event. Lots of people refer to this example to suggest that mathematics doesn't work for predicting football. This shows that general public has often difficulty interpreting probabilistic assessments. The probability of 46.75% means that if we would finish the league from that moment 100 times, on average Roda JC would relegate directly about 47 times. If one only looks at one season, where the probability of Roda JC to be relegated directly is estimated to be 46.75%, then this tells us only that there is about a fifty-fifty chance that this event will happen. We can use the method of hypothesis testing to decide if the probability 46.75% is estimated correctly. By setting the maximum probability of rejecting the hypothesis when in fact it is true, called the significance level p, (usually $p = 0.05$) we can formally state the following hypothesis:
Using a significance level of $p = 0.05$.

- H_0: The probability that Roda JC will relegate directly is 46.75%, so $P(Roda) = 0.4675$
- H_1: The probability that Roda JC will relegate directly is not 46.75%, so $P(Roda) \neq 0.4675$

The test tells us that the H_0 hypothesis would be rejected at our significance level in favor of H_1 if out of 100 times, Roda JC would have relegated less than 37 times or more than 57 times. Later we will go into more detail and show exactly how this test works. This test gives us a mathematical way to judge the quality of the probability estimate.

Despite the fact that a lot of people dismiss mathematics in football, because of their lack of knowledge, the last couple of years we can see an increasing trend in the usage of mathematics in football. Many teams now have a datascientist as part of their staff, to analyze the performances of their players, based on lots of gathered data. An extreme example of this development in the sport is the club FC Midtjylland. Matthew Benham the owner of a company called Smartodds, made millions using mathematical models to predict football results and he believed that these models could also have an impact on the transfer market and eventually on the pitch. Benham invested in FC Midtjylland and introduced his mathematical way of looking at football. They used analytics to see what was the best way to take a free kick, they analysed how every individual kicked the ball and how they could improve, moreover they used the database of Smartodds to find players with certain specific properties. The next season the team became the champions of their league. It is success stories likes the one from Midtjylland that persuade the whole football community to innovate and to start using analytics. There are, as always, some successful and some less successful stories about the use of mathematics in the sport.

Since I am a huge football supporter myself, the idea of combining football and mathematics really appealed to me. Ralph Allan Bradley and Milton E. Terry (1952) published a report about the method of paired comparisons [3]. Working with a list of arbitrary items, in this report they discuss a model that estimates the probabilities that the items are superior in a given comparison. These estimates allow to rate the items. The model has been used in the past to predict the outcome of tennis matches using the data of past results and the type of surface of the contest. Also in football the model has been experimented with. Gunther Schauberger, Andreas Groll and Gerhard Tutz [7] have used the Bradley-Terry model to predict football results in the German league.
In my research I will study the Bradley-Terry model to obtain rankings of teams using data from the Dutch league called the Eredivisie.

The original Bradley-Terry model has a few assumptions that we have to work with. The fundamental concept of the model is that each team has a so called " winning potential ". This winning potential is based on the past results and is independent of the winning potential of the opponent. A consequence of this independence is that we find a transitive ranking structure. This means that if team A is likely to win against team B, and team B is likely to win against team C, then team A is also likely to win against team C. The probability that team A beats team C should also be greater than the probability that team B beats team C, which also makes sense intuitively. The model assumes that if we compare two items, there is always one superior. If we translate this to football, it means that the model does not take the probability of a draw into account. If we compare team A with team B the model will only return a probability P that team A will win, giving a probability 1-P that team B will win. This means that there is a probability of 0 that the match between team A and team B will be a draw.

Since the model is very generic and can be applied in lots of different situations, we might need to adjust it to make it more suitable for football. Looking at the assumptions of the model, we observe that the model is very simplistic. The winning potentials obtained by the model are solely based on the wins and losses of all teams, regardless of the exact result of these matches. This means that it doesn't matter whether a match is won with a score of 1-0 or with a score of 10-0. These scores will have the same impact on calculating a winning potential of a team. Since the model doesn't deal with draws, we also have to exclude all the draws out of our data, meaning that those matches also do not contribute to the winning potentials of the teams. Another factor that we haven't yet discussed, but might impact the ranking as well is the advantage that teams playing at home have. In the model we deal with a constant winning potential for every team, regardless of playing a home game or an away game. Statistics show however that in general, teams that play at home have a bit of advantage over teams that play an away game. Even though, according to Richar Pollard [5], the home advantage differs over countries around the world.

The goal of this project is to make predictions about the ranking of the Dutch league using the Bradley-Terry model. In Chapter 2, we will study the original Bradley-Terry model, and look at some extensions of the model as well. If we know how the model works we can look at real data. We have complete rankings of the past 20 seasons of Eredivisie at our disposal, including all the matches played in these seasons. This data was provided by footballdata.co.uk [1]. In Chapter 3, we will look at this data in detail and we will discuss the best way to use it. After analyzing this data, we will start our simulation studies. In Chapter 4 the model is applied on different variations of simulated data to see how the model performs. This allows us to test the effect of excluding draws from our data, or the effect of the number of matches played, on recovering the winning potentials. In the simulation studies we should learn which extensions perform the best on certain types of data. In Chapter 5, this model is used on the real data to be able to make predictions about season 2018-2019 of the Eredivisie. The results will be discussed in Chapter 6. Here we will also talk about the limitations of the model and possible extensions, which might improve the model.

2
Theory

In this chapter we present the Bradley-Terry model of paired comparison introduced in [3]. The goal is to use this model to rank teams in a football league. We start with introducing notation and assumptions. Later on, some extensions of this model will be presented and discussed.

2.1. Original Bradley-Terry model

Suppose that there are m teams in the league. The model assigns so called 'winning potentials' to each of the m teams. Let us denote π_i as the winning potential of team i and let us denote $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ as the vector of all winning potentials. Each winning potential π_i is assumed to be positive and constant. This means that for every match, π_i is the same, independent of the opponent, the weather, the location of the game (home or away) or any other variable situations. Finally we constrain the winning potentials such that $\sum_i^m \pi_i = m$. We denote the probability of team i beating team j by p_{ij}. From the model assumptios it follows that:

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \tag{2.1}$$

The model only offers a probability for winning (hence also losing). This means that the model ignores draws.

Since the winning potentials are constant, there is a transitive relationship on the set of winning potentials, meaning that if $\pi_i > \pi_j$ and $\pi_j > \pi_k$ then $\pi_i > \pi_k$. Because of this transitive relationship we are able to rank the teams, by sorting their winning potentials. Let's do this for a small example.

Example 2.1.1. Example League:
This league consists of 3 teams. Team 1 that has a winning potential of $\pi_1 = \frac{27}{16}$, team 2 that has a winning potential of $\pi_2 = \frac{12}{16}$ and team 3 that has a winning potential of $\pi_3 = \frac{9}{16}$. Now we can simply rank the teams by sorting their winning potentials to get the following ranking:

1. Team 1
2. Team 2
3. Team 3

The probability of team 1 beating team 2 for example, would be $p_{12} = \frac{\pi_1}{\pi_1+\pi_2} = \frac{27/16}{27/16+12/16} \approx 0.69$

It is possible to simulate this 'mini-league' and see what happens. We let all the teams play each other five times. One simulation resulted in the following: Team 1 won 9 matches, Team 2 won 2 matches and Team 3 won 4 matches. We observe that Team 3 has had more victories than Team 2, while the winning potential of Team 2 is higher. Since the potentials of Team 2 and Team 3 are very close, these results are not surprising. If this simulated data was used to estimate the winning potentials and to find the ranking of this league, the obtained winning potentials would provide a different ranking than the ones we used to simulate the data with.

So by finding the winning potentials we can provide a ranking of the teams in the league, as seen in example 2.1.1. In section 2.1.1 the process of finding these winning potentials will be explained.

2.1.1. Estimation of Parameters

In order to obtain the winning potentials, the maximum likelihood method will be used. This means that we determine the probability of the real outcomes of the matches played in the league and then maximize this probability with respect to the winning potentials. To construct the likelihood we need to introduce some more notation. Let us denote w_{ij} as the number of victories team i has over team j and let us denote n_{ij} as the number of matches team i and team j played against each other. We denote $W_i = \sum_j w_{ij}$ as the total number of wins of team i. Since in this model there are only wins and losses, the relationship $w_{ij} + w_{ji} = n_{ij} = n_{ji}$ holds. For this model, an $m \times m$ matrix w containing the wins is enough to construct the likelihood. This matrix simply looks like:

$$
w = \begin{array}{c|cccc}
 & \text{Team 1} & \text{Team 2} & \dots & \text{Team } m \\
\hline
\text{Team 1} & 0 & w_{12} & \dots & w_{1m} \\
\text{Team 2} & w_{21} & 0 & \dots & w_{2m} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\text{Team } m & w_{m1} & w_{m2} & \dots & 0
\end{array}
$$

Let's first think of the likelihood for one match, played between team i and team j. The probability of the observed outcome of this match is

$$\left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{w_{ij}} \left(\frac{\pi_j}{\pi_i + \pi_j}\right)^{w_{ji}} = \frac{\pi_i^{w_{ij}} \pi_j^{w_{ji}}}{(\pi_i + \pi_j)} \tag{2.2}$$

This probability is equal to the probability of team i winning over team j or the probability of team j winning over team i depending on the result of the match. If team i wins we get $w_{ij} = 1$ and $w_{ji} = 0$. The expression above then becomes $\frac{\pi_i}{\pi_i+\pi_j}$, which indeed is the probability of team i winning as discussed earlier. Alternatively, if $w_{ij} = 0$ and $w_{ji} = 1$ the probability becomes $\frac{\pi_j}{\pi_i+\pi_j}$, which indeed is the probability of team j winning over team i. If team i plays team j more than once, say n_{ij} times, we simply obtain the probability:

$$\frac{\pi_i^{w_{ij}} \pi_j^{w_{ji}}}{(\pi_i + \pi_j)^{n_{ij}}} \tag{2.3}$$

Because p_{ij} is not influenced by p_{ik} for all i, j, k, the probability for all matches team i plays can now be found by taking the product of expression (2.3) over all teams m. We obtain:

$$\prod_{j=1}^{m} \frac{\pi_i^{w_{ij}} \pi_j^{w_{ji}}}{(\pi_i + \pi_j)^{n_{ij}}} \tag{2.4}$$

$$= \pi_i^{w_{i1}+w_{i2}+\dots+w_{im}} * \pi_1^{w_{1i}} * \pi_2^{w_{2i}} * \dots * \pi_t^{w_{ti}} * (\pi_i + \pi_1)^{-n_{i1}} * (\pi_i + \pi_2)^{-n_{i2}} * \dots * (\pi_i + \pi_m)^{-n_{im}} \tag{2.5}$$

$$= \pi_i^{W_i} * \pi_1^{w_{1i}} * \pi_2^{w_{2i}} * \dots * \pi_t^{w_{mi}} * (\pi_i + \pi_1)^{-n_{i1}} * (\pi_i + \pi_2)^{-n_{i2}} * \dots * (\pi_i + \pi_m)^{-n_{im}} \tag{2.6}$$

Now that we found the probability of the results of one team, as shown in (2.6) we can take the product over all teams to obtain the probability of all results. This suggests the likelihood will be $\prod_{i=1}^m \prod_{j=1}^m \frac{\pi_i^{w_{ij}} \pi_j^{w_{ji}}}{(\pi_i+\pi_j)^{n_{ij}}}$. However in this case we would count every match twice, because we would have a factor $\frac{\pi_i^{w_{ij}} \pi_j^{w_{ji}}}{(\pi_i+\pi_j)^{n_{ij}}}$ as well as a factor $\frac{\pi_i^{w_{ji}} \pi_j^{w_{ij}}}{(\pi_j+\pi_i)^{n_{ji}}}$, hence counting the probability outcome of the matches between team i and team j twice. Therefore we should take the square root of the product of expression (2.6). The resulting likelihood is:

$$L(\boldsymbol{\pi}) = \left(\prod_i \pi_i^{W_i}\right) \prod_i \prod_j (\pi_i + \pi_j)^{-\frac{1}{2} n_{ij}} \tag{2.7}$$

The log-likelihood, which is often easier to work with, then is:

$$\ell(\boldsymbol{\pi}) = \sum_i W_i * \ln(\pi_i) + \sum_i \sum_j -\frac{1}{2} n_{ij} * \ln(\pi_i + \pi_j) \tag{2.8}$$

Now that we've established the log-likelihood we should simply maximize it in order to find our maximum likelihood estimates.
In order to maximize the log-likelihood we need to find the partial derivatives and set these equal to 0.

The partial derivative for π_i is:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\pi})}{\partial \pi_i} &= \frac{W_i}{\pi_i} - \frac{1}{2}\sum_j \frac{n_{ij}}{\pi_i + \pi_j} - \frac{1}{2}\sum_j \frac{n_{ji}}{\pi_j + \pi_i} \\ &= \frac{W_i}{\pi_i} - \sum_j \frac{n_{ij}}{\pi_i + \pi_j}\end{aligned} \tag{2.9}$$

Setting the partial derivatives equal to zero we find the following system of equations:

$$\pi_i = W_i * \Big[\sum_j \frac{n_{ij}}{\pi_i + \pi_j}\Big]^{-1}, \; i = 1, \ldots, m \tag{2.10}$$

This means that we have m nonlinear equations, that each contain m unknown variables.
If the league only consisted of 2 teams, this system of equations could easily be analytically solved. In the next example this is shown.

Example 2.1.2. Let's assume that we have the following data of a league, where each team played 10 matches:

$w =$

	Team 1	Team 2
Team 1	0	7
Team 2	3	0

This data sets up the following log-likelihood:

$$\begin{aligned}\ell(\pi) &= 7 * \ln(\pi_1) + 3 * \ln(\pi_2) - 5 * \ln(\pi_1 + \pi_2) - 5 * \ln(\pi_2 + \pi_1) \\ &= 7 * \ln(\pi_1) + 3 * \ln(\pi_2) - 10 * \ln(\pi_1 + \pi_2)\end{aligned} \tag{2.11}$$

By taking partial derivatives and solving the system of equations we find:

$$\begin{cases} 7\pi_2 = 3\pi_1 \\ \pi_1 + \pi_2 = 2 \end{cases} \tag{2.12}$$

Solving this system we find $\boldsymbol{\pi} = \begin{bmatrix} \frac{7}{5} \\ \frac{3}{5} \end{bmatrix}$

If there are more than two teams, the system will become way more difficult, with multiple nonlinear equations. In general solving such a system of equations would be done numerically. We choose to use an iterative method to approach the solutions of the winning potentials.

This iterative method is called an MM algorithm, where the MM stand for Minorizing and Maximizing. To approach the maximum of a function, the first step in this algorithm is choosing a starting point. The function is then minorized to obtain a minorizing function, that contains the starting point. The minorizing function should be chosen such that it is easy to maximize and such that it is tangent to the original function at the starting point. The maximum of the minorizing function replaces the starting point in the next iteration, making sure that the updated minorizing function is now tangent to the original function in the new point. This updated minorizing function is again maximized and updated with the newly found maximum. Basically every iteration consists of updating the minorizing function and maximizing this function to obtain a new maximum, which is closer to the real maximum.

To give a visual impression of how an MM algorithm works we approached the maximum of a function $g(x) = -x^2 + \ln(x) + 10$ with an MM algorithm. This can be seen in Figure 2.1. The minorizing function for this MM algorithm is defined as $Q_k(x) = -x^2 + \ln(x^{(k)}) - \frac{x^{(k)}}{x} + 11$. The plots in black are iterations of the MM algorithm, in which we maximize and update our minorizing function. Function g is plotted in red. The starting point of the algorithm is $x^{(1)} = 3$. You can see that the starting point $x^{(1)}$ is quite far from the actual maximum x^*. After updating the starting point twice, we are already a lot closer to x^* with the obtained $x^{(3)}$. The idea is that $x^{(k)} \longrightarrow x^*, \; if \, k \longrightarrow \infty$ Note that g is tangent to Q at the current iterate, so at $Q_k(x^{(k)})$.

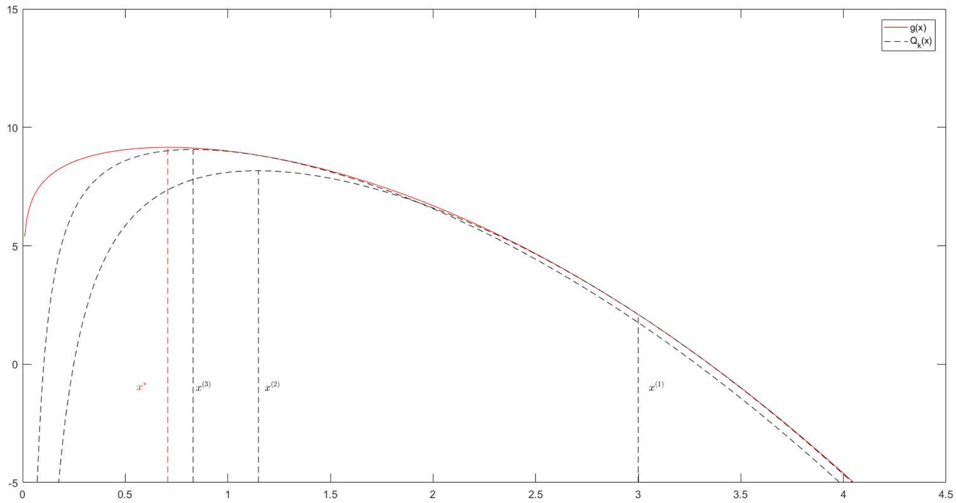


Figure 2.1: Plot of MM algorithm used to approach the maximum of function g

To design an iterative algorithm to find the maximum likelihood estimator of $\boldsymbol{\pi}$, first the minorizing function should be found. To construct this minorizing function we will use the following inequality:

$$-\ln(x) \geq 1 - \ln(y) - (x/y) \text{ with equality if and only if x = y.} \tag{2.13}$$

We can prove that this equality holds:

Proof. To proof that (2.13) holds it is enough to show that $\ln(\frac{x}{y}) \leq \frac{x}{y} - 1$. Since the log function is strictly concave and differentiable it is bounded by its first order Taylor expansion. Take $z = \frac{x}{y}$ and expand it in the neighbourhood of $a = 1$.
We then get $\ln(\frac{x}{y}) = \ln(z) \leq \ln(1) + 1 * (z-1) = z - 1 = \frac{x}{y} - 1$. □

We apply inequality (2.13) on $\ell(\boldsymbol{\pi})$ from (2.8), where $x = (\pi_i + \pi_j)$, $y = (\pi_i^{(k)} + \pi_j^{(k)})$, to find a function

$$Q_k(\boldsymbol{\pi}) = \sum_i W_i * \ln(\pi_i) + \sum_i \sum_j \frac{1}{2} n_{ij} \Big[-\ln(\pi_i^{(k)} + \pi_j^{(k)}) - \frac{\pi_i + \pi_j}{\pi_i^{(k)} + \pi_j^{(k)}} + 1 \Big] \tag{2.14}$$

Now because of inequality (2.13) we find that

$$Q_k(\boldsymbol{\pi}) \leq \ell(\boldsymbol{\pi}) \text{ with equality if } \boldsymbol{\pi} = \boldsymbol{\pi^{(k)}}, \tag{2.15}$$

The function $Q_k(\boldsymbol{\pi})$ which satisfies condition (2.15) minorizes $\ell(\boldsymbol{\pi})$ at the point $\boldsymbol{\pi}^{(k)}$. It is easy to see that, for any $Q_k(\boldsymbol{\pi})$, which satisfies the minorizing condition (2.15)

$$Q_k(\boldsymbol{\pi}) \geq Q_k(\boldsymbol{\pi}^{(k)}) \text{ implies } \ell(\boldsymbol{\pi}) \geq \ell(\boldsymbol{\pi}^{(k)}) \tag{2.16}$$

$Q_k(\boldsymbol{\pi})$ is easy to maximize, because it separates the components of the parameter vector $\boldsymbol{\pi}$. Therefore maximization of $Q_k(\boldsymbol{\pi})$ is equivalent to maximization for each component π_i separately.
To find the next iteration $\pi_i^{(k+1)}$ we differentiate $Q_k(\boldsymbol{\pi})$ with respect to π_i and set this equal to 0. Since w_{ij} is constant and $-\ln(\pi_i^{(k)} + \pi_j^{(k)})$ is constant we get:

$$\begin{aligned}
\frac{d}{d\pi_i} Q_k(\boldsymbol{\pi}) &= \frac{W_i}{\pi_i} - \sum_j (\frac{1}{2} n_{ij} + \frac{1}{2} n_{ji}) * \frac{1}{\pi_i^{(k)} + \pi_j^{(k)}} = 0 \\
&\Rightarrow \frac{W_i}{\pi_i} - \sum_j n_{ij} * \frac{1}{\pi_i^{(k)} + \pi_j^{(k)}} = 0 \\
&\Rightarrow W_i = \sum_j \frac{n_{ij}}{\pi_i^{(k)} + \pi_j^{(k)}} * \pi_i \\
&\Rightarrow \pi_i = W_i * \Big[\sum_j \frac{n_{ij}}{\pi_i^{(k)} + \pi_j^{(k)}} \Big]^{-1}
\end{aligned}$$

Hence the updated winning potential will be:

$$\pi_i^{(k+1)} = W_i * \Big[\sum_j \frac{n_{ij}}{\pi_i^{(k)} + \pi_j^{(k)}} \Big]^{-1} \tag{2.17}$$

Because it doesn't matter where the starting point is located in order to converge, we choose the starting point to be 'uniform', meaning that $\boldsymbol{\pi}^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
The proof of convergence of the algorithm 2.17 follows from Liapounov's theorem. We say that an MM algorithm converges if $\boldsymbol{\pi}^* = \lim_k \boldsymbol{\pi}^{(k)}$.

Theorem 1. Liapounov's theorem:
Suppose M: $\Omega \to \Omega$ is continuous and $\ell : \Omega \to \mathbf{R}$ is differentiable and for all $\boldsymbol{\pi} \in \Omega$ we have $\ell\Big[\boldsymbol{M}(\boldsymbol{\pi})\Big] \geq \ell(\boldsymbol{\pi})$, with equality only if $\boldsymbol{\pi}$ is a stationary point of $\ell(\cdot)$ (i.e., the gradient is $\mathbf{0}$ at $\boldsymbol{\pi}$). Then, for abitrary $\boldsymbol{\pi}^{(1)} \in \Omega$, any limit point of the sequence $\{\boldsymbol{\pi}^{(k+1)} = M(\boldsymbol{\pi}^{(k)})\}_{k\geq 1}$ is a stationary point of $\ell(\boldsymbol{\pi})$.

In our case the function M updates the winning potentials from $\boldsymbol{\pi}^{(k)}$ to $\boldsymbol{\pi}^{(k+1)}$, it is a continuous function. If $M(\boldsymbol{\pi}^{(\boldsymbol{k})}) = \boldsymbol{\pi}^{(k)}$, then we know that $\boldsymbol{\pi}^{(k+1)} = \boldsymbol{\pi}^{(k)}$, so it is a stationary point of $Q_k(\boldsymbol{\pi})$, because then the gradient would be $\mathbf{0}$ at $\boldsymbol{\pi}^{(k)}$. But then $\pi^{(k)}$ is also a stationary point of $\ell(\boldsymbol{\pi})$, because Q_k is tangent to the log-likelihood at the current iterate.
Finally we need two conditions:

- A sufficient condition for upper compactness (ℓ is defined to be upper compact if for any constant c, the set $\{\boldsymbol{\pi} \in \Omega : \ell(\boldsymbol{\pi}) > c\}$ is a compact subset of Ω)
- A sufficient condition for strict concavity of the log-likelihood.

If the log-likelihood is upper compact, this means that there is at least one limit point of the sequence $\{\boldsymbol{\pi}^{(k+1)} = M(\boldsymbol{\pi}^{(k)})\}_{k\geq 1}$. Because the sequence $\ell(\boldsymbol{\pi}^{(k)})$ is strictly increasing until the maximum is found, there is a K such that for every $k \geq K$ the inequality $\ell(\boldsymbol{\pi}^{(k)}) \geq c$ holds. The sequence then has a limit point, since it is a sequence on a compact set.

The strict concavity on the other hand says that there is at most one stationary point of $\ell(\boldsymbol{\pi})$ (the maximizer).

Suppose that $\ell(\boldsymbol{\pi})$ is upper compact and strictly concave. Liapounov's theorem says that for any starting point $\boldsymbol{\pi}^{(1)} \in \Omega$ any limit point of the sequence $\{\boldsymbol{\pi}^{(k+1)} = M(\boldsymbol{\pi}^{(k)})\}_{k\geq 1}$ is a stationary point of $\ell(\boldsymbol{\pi})$. Now because of the upper compactness, there is at least one limit point, so this is a stationary point. Because of the strict concavity of $\ell(\boldsymbol{\pi})$ there is at most one stationary point. So we can conclude that there is only one limit point, which is a stationary point. Because of the strict concavity this must then be the maximum.

To make sure that the log-likelihood satisfies the two conditions the data must satisfy a data assumption, which we describe here:

Data Assumption: In every possible partition of the teams into two nonempty subsets, some team in the second set beats a team in the first set at least once. This can be represented as a directed graph with the m teams as the nodes. There is a directed edge (i, j), from team i to team j whenever team i has beaten team j at least once. Now we assume this graph to be strongly connected, meaning that there is a directed path from i to j for all i and j.

The proof that the log-likelihood is upper compact and strictly concave under these two conditions can be found in Hunter (2004) [4]. However by looking at some examples in which the assumptions above are not satisfied, we can try to understand what goes wrong.

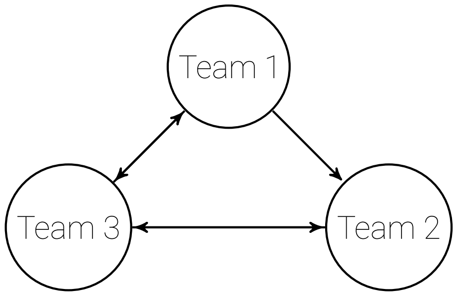
Let's first take a look at an example that does satisfy the data assumption.

Example 2.1.3. In this example, the teams played 20 matches against each other.

$w =$

	Team 1	Team 2	Team 3
Team 1	0	20	13
Team 2	0	0	8
Team 3	7	12	0

It is easy to check whether this data satisfies the data assumption by creating a graph based on this data. The following graph then shows us the relation between the teams:



It is clear that this graph is strongly connected, hence the data satisfies the data assumption.

The log-likelihood that follows from the data is:

$$33\ln(\pi_1) + 8\ln(\pi_2) + 19\ln(\pi_3) - 20\ln(\pi_1 + \pi_2) - 20\ln(\pi_1 + \pi_3) - 20\ln(\pi_2 + \pi_3) \tag{2.18}$$

We have the constraint $\sum_{i=1}^{3} \pi_i = 3$ and all the $\pi_i > 0$ for all i. In Figure 2.2 a visualisation of the parameter space is presented for a league that consists of 3 teams.

The parameter space is an open set, because all the winning potentials are positive. This means that if the log-likelihood has a maximum 'on the edge' of the parameter space, we would have a problem regarding upper compactness. We could then take a constant c, smaller than our maximum and find that the set $\{\boldsymbol{\pi} \in \Omega : \ell(\boldsymbol{\pi}) > c\}$ is open, meaning that this set is not compact, hence the log-likelihood is not upper compact.
This also tells us that if the log-likelihood is upper compact, there is no maximum 'on' the edge.

If we go near the edge of the parameter space, for some i, π_i will go to 0, meaning that $\ln(\pi_i)$ will go to negative infinity. This means that the log-likelihood (2.18) diverges to negative infinity, so we have no maximum on the edge.

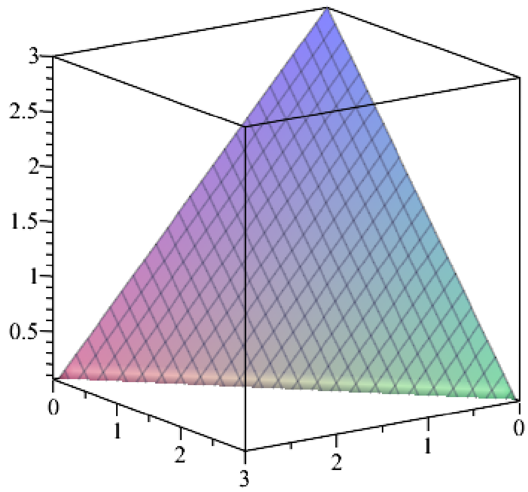


Figure 2.2: Visualisation of the parameter space for a league consisting of 3 teams

Since the data assumption is satisfied, we know that the algorithm will converge, so we can apply it on the data. We start with a uniform vector as our starting point.

$$\boldsymbol{\pi}^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{2.19}$$

Now that $\boldsymbol{\pi}^{(1)}$ is known, next iterations can be found:

$$\pi_1^{(2)} = 33 * \left[\frac{20}{2} + \frac{20}{2}\right]^{-1} = \frac{33}{20}$$
$$\pi_2^{(2)} = 8 * \left[\frac{20}{2} + \frac{20}{2}\right]^{-1} = \frac{8}{20}$$
$$\pi_3^{(2)} = 19 * \left[\frac{20}{2} + \frac{20}{2}\right]^{-1} = \frac{19}{20}$$

The following iteration gives us:

$$\pi_1^{(3)} = 33 * \left[\frac{20}{41/20} + \frac{20}{52/20}\right]^{-1} = \frac{5863}{3100}$$
$$\pi_2^{(3)} = 8 * \left[\frac{20}{41/20} + \frac{20}{27/20}\right]^{-1} = \frac{1107}{3400}$$
$$\pi_3^{(3)} = 19 * \left[\frac{20}{52/20} + \frac{20}{27/20}\right]^{-1} = \frac{6669}{7900}$$

Finally after 12 iterations, our algorithm converged and we find the estimates:

$\pi_1 \approx 2.0839187$

$\pi_2 \approx 0.2637063$

$\pi_3 \approx 0.6523750$

The algorithm is instructed to terminate if the result of $\frac{\pi_i^{(k+1)}}{\pi_i^{(k)}} - 1$ is smaller than ϵ. For this thesis $\epsilon = 0.001$ will be used for every MM algorithm. In the example above this threshold is crossed after 12 iterations.

After seeing an example in which the data assumption was satisfied, let's look at an example where this is not the case. In the next example one team wins all its matches, hence the data assumption is not satisfied.

Example 2.1.4. In this example, the following data is provided:

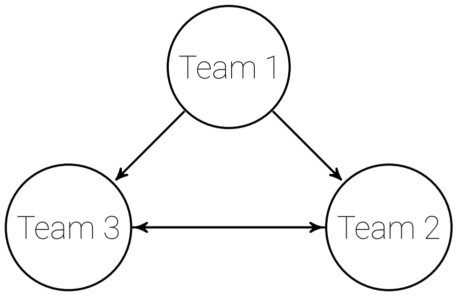
$w =$

	Team 1	Team 2	Team 3
Team 1	0	20	20
Team 2	0	0	8
Team 3	0	12	0

The log-likelihood of this example is:

$$\ell(\boldsymbol{\pi}) = 40\ln(\pi_1) + 8\ln(\pi_2) + 12\ln(\pi_3) - 20\ln(\pi_1+\pi_2) - 20\ln(\pi_1+\pi_3) - 20\ln(\pi_2+\pi_3) \tag{2.20}$$

If we represent the data as a graph, we find:



Since there is no path from Team 2 to Team 1, this graph is not strongly connected, therefore the data assumption is not satisfied.

We want to show that if one team wins all its matches, the winning potential tends to go to the edge of the parameter space.

When one team wins all its matches, the algorithm gives the following iterative values for the winning potentials:

$$\begin{aligned}
\pi_1^{(k+1)} &= W_1\left[\frac{n_{12}}{\pi_1^{(k)}+\pi_2^{(k)}} + \frac{n_{13}}{\pi_1^{(k)}+\pi_3^{(k)}}\right]^{-1}\\
\pi_2^{(k+1)} &= W_2\left[\frac{n_{21}}{\pi_2^{(k)}+\pi_1^{(k)}} + \frac{n_{23}}{\pi_2^{(k)}+\pi_3^{(k)}}\right]^{-1}\\
\pi_3^{(k+1)} &= W_3\left[\frac{n_{31}}{\pi_3^{(k)}+\pi_1^{(k)}} + \frac{n_{32}}{\pi_3^{(k)}+\pi_2^{(k)}}\right]^{-1}
\end{aligned}$$

Because Team 1 won all its matches, note that $W_1 = n_{12} + n_{13}$. This means that we find

$$\pi_1^{(k+1)} = \frac{n_{12}+n_{13}}{\frac{n_{12}}{\pi_1^{(k)}+\pi_2^{(k)}} + \frac{n_{13}}{\pi_1^{(k)}+\pi_3^{(k)}}} \tag{2.21}$$

Now we claim that $\pi_1^{(k)} \longrightarrow m$ if $k \longrightarrow \infty$. Since $\sum \pi_i = m$, this would imply that $\pi_2^{(k)} \longrightarrow 0$ and $\pi_3^{(k)} \longrightarrow 0$ if $k \longrightarrow \infty$. In order to proof this claim we just need to proof equation (2.22) holds.

$$\pi_1^{(k)} < \pi_1^{(k+1)} \quad \forall k > 1 \tag{2.22}$$

We proof this by showing that:

$$\frac{\pi_1^{(k+1)}}{\pi_1^{(k)}} > 1 \tag{2.23}$$

Using (2.21) we find:

$$\frac{\pi_1^{(k+1)}}{\pi_1^{(k)}} = \frac{n_{12}+n_{13}}{\frac{\pi_1^{(k)}}{\pi_1^{(k)}+\pi_2^{(k)}}n_{12} + \frac{\pi_1^{(k)}}{\pi_1^{(k)}+\pi_3^{(k)}}}n_{13}$$

But $\frac{\pi_1^{(k)}}{\pi_1^{(k)}+\pi_2^{(k)}} \leq 1$ and $\frac{\pi_1^{(k)}}{\pi_1^{(k)}+\pi_3^{(k)}} \leq 1$
So:

$$\frac{n_{12}+n_{13}}{\frac{\pi_1^{(k)}}{\pi_1^{(k)}+\pi_2^{(k)}}n_{12} + \frac{\pi_1^{(k)}}{\pi_1^{(k)}+\pi_3^{(k)}}}n_{13} \geq \frac{n_{12}+n_{13}}{n_{12}+n_{13}} = 1$$

Which proofs our claim.

Now since $\boldsymbol{\pi} \to \begin{bmatrix} m \\ 0 \\ 0 \end{bmatrix}$, the maximum will be located on the edge of Figure 2.2, which is not in our parameter space. Hence we have no maximum in our parameter space. If the maximum is located on the edge, this also means that the log-likelihood is not upper compact.

For the final example of this section, we look at a scenario where a team lost all its matches.

Example 2.1.5. In this example, the following data is provided:

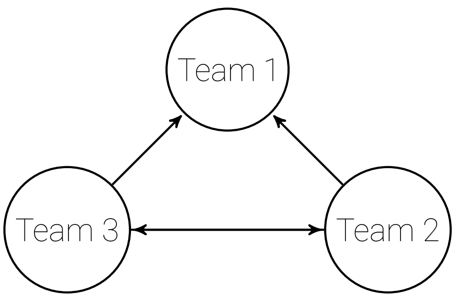
$w =$

	Team 1	Team 2	Team 3
Team 1	0	0	0
Team 2	20	0	8
Team 3	20	12	0

It is clear that Team 1 lost all its matches in this example. The log-likelihood will be:

$$\ell(\boldsymbol{\pi}) = 28 * \ln(\pi_2) + 32 * \ln(\pi_3) - 20 * \ln(\pi_1+\pi_2) - 20 * \ln(\pi_1+\pi_3) - 20 * \ln(\pi_2+\pi_3) \tag{2.24}$$

Translating the data into a graph gives us:



This graph is not strongly connected because there is no path from Team 1 to Team 2, therefore our data does not satisfy the data assumption. We still operate on the parameter space shown in Figure 2.2. Since Team 1 lost all its matches, we expect that the maximum would be on the edge where $\pi_1 = 0$. This edge lies outside the parameter space.
To find the maximum one of the equations we should solve is:

$$\frac{\mathrm{d}}{\mathrm{d}\pi_1}\ell(\boldsymbol{\pi}) = -\frac{5}{\pi_1+\pi_2} - \frac{5}{\pi_1+\pi_3} = 0 \tag{2.25}$$

However since $\pi_i > 0 \; \forall i$, it's trivial that there are no solutions to this equation in our parameter space. We can conclude from this example that if a team loses all its matches, there is no maximum in the parameter space.

2.1.2. Alternative representation

To extend the model, it is easier to work with a different representation of the probability p_{ij}. This different representation can be compared to the representation of a different model, called Thurstone's model.
In Thurstone's model every team i has a corresponding sensation S_i, which is a random variable that has a normal distribution. In this model the probability of winning depends on these sensations. The higher the expectation of sensation, the better the team. When teams i and j are compared in this model, the probability for i beating j is of the form:

$$p_{ij} = P(S_i > S_j) = P(S_i - S_j > 0) = \frac{1}{\sqrt{2\pi}} \int_{-(S_i - S_j)}^{\infty} e^{-\frac{1}{2}y^2} \mathrm{d}y \tag{2.26}$$

Because both S_i and S_j are normally distributed, the difference between them will also be normally distributed, hence we get a probability as in (2.26).
In the Bradley-Terry model, we define these 'sensations' as $V_i = \ln(\pi_i)$ and $V_j = \ln(\pi_j)$.
Our p_{ij} can actually be rewritten as an integral, just as in Thurstone's model, giving us:

$$p_{ij} = \frac{1}{4} \int_{-(V_i - V_j)}^{+\infty} \operatorname{sech}^2(y/2) \mathrm{d}y \tag{2.27}$$

Where $sech(x) = \frac{1}{cosh(x)} = \frac{2}{e^x + e^{-x}}$ is the hyperbolic secant.
We can show that this is exactly the same as $\frac{\pi_i}{\pi_i + \pi_j}$:

$$\begin{aligned}
p_{ij} &= \frac{1}{4} \int_{-(V_i - V_j)}^{+\infty} \operatorname{sech}^2(y/2) \mathrm{d}y \\
&= \frac{1}{4} \Big[2 \tanh(\frac{y}{2}) \Big]_{-(V_i - V_j)}^{+\infty} \\
&= \frac{1}{2} \Big[\frac{\sinh(\frac{y}{2})}{\cosh(\frac{y}{2})} \Big]_{-(V_i - V_j)}^{+\infty} \\
&= \frac{1}{2} \Big[\frac{\frac{1}{2}(-e^{-y/2} + e^{y/2})}{\frac{1}{2}(e^{-y/2} + e^{y/2}}) \Big]_{-(V_i - V_j)}^{+\infty} \\
&= \frac{1}{2} \Big[\frac{1 - e^{-y}}{1 + e^{-y}} \Big]_{-(V_i - V_j)}^{+\infty} \\
&= \frac{1}{2} \Big(1 - (\frac{1 - \frac{\pi_i}{\pi_j}}{1 + \frac{\pi_i}{\pi_j}}) \Big) \\
&= \frac{1}{2} \Big(1 - \frac{\frac{\pi_j - \pi_i}{\pi_j}}{\frac{\pi_j + \pi_i}{\pi_j}} \Big) \\
&= \frac{1}{2} \Big(\frac{\pi_i + \pi_j}{\pi_i + \pi_j} - \frac{\pi_j - \pi_i}{\pi_i + \pi_j} \Big) \\
&= \frac{1}{2} \frac{2\pi_i}{\pi_i + \pi_j} \\
&= \frac{\pi_i}{\pi_i + \pi_j}
\end{aligned}$$

Now that we've established a different representation of p_{ij} it becomes easier to extend and adjust this probability. In the next sections we will introduce different extensions on the original model.

2.2. Model Including draws

Unfortunately the original Bradley-Terry model only provides a probability for winning and losing, so any draws in data are ignored and can also not be simulated. Because of this, we base the winning potentials of the teams only on a percentage of all the data available, namely the wins and losses. To make sure the draws also contribute to the winning potentials we need to adjust the model. In this section we will introduce an extension, based on the model of Rao and Kupper [6] using the alternative form of p_{ij}.

First we need to introduce some new notation. Let us denote $p_{i=j}$ to be the probability that the match between team i and team j ends in a draw. The rest of the notation will be the same as in the original model. As you can see in equation 2.27, the probability p_{ij} is based on the difference between V_i and V_j. We want to do the same thing for $p_{i=j}$. When team i and team j are almost equally 'good', we think a draw would occur. Therefore we introduce the parameter η. We say that if the difference between V_i and V_j, so $|V_i - V_j|$ is smaller than η the match ends in a draw. Implementing this in the alternative form, we get:

$$p_{i=j} = \frac{1}{4}\int_{-(V_i-V_j)-\eta}^{-(V_i-V_j)+\eta} \operatorname{sech}^2(y/2)\mathrm{d}y \tag{2.28}$$

Rewriting this and using $\theta = e^{\eta}$ we find:

$$\begin{aligned}
p_{i=j} &= \frac{1}{4}\int_{-(V_i-V_j)-\eta}^{-(V_i-V_j)+\eta} \operatorname{sech}^2(y/2)\mathrm{d}y \\
&= \frac{1}{4} * 2 * \Big[\tanh(y/2)\Big]_{-(V_i-V_j)-\eta}^{-(V_i-V_j)+\eta} \\
&= \frac{1}{2}\Big[\frac{e^y-1}{e^y+1}\Big]_{-(V_i-V_j)-\eta}^{-(V_i-V_j)+\eta} \\
&= \frac{1}{2}\Big[(\frac{e^{-V_i} * e^{V_j} * e^{\eta} - 1}{e^{-V_i} * e^{V_j} * e^{\eta} + 1}) - (\frac{e^{-V_i} * e^{V_j} * e^{-\eta} - 1}{e^{-V_i} * e^{V_j} * e^{-\eta} + 1})\Big] \\
&= \frac{1}{2}\Big[\frac{\frac{\theta\pi_j}{\pi_i} - 1}{\frac{\theta\pi_j}{\pi_i} + 1} - \frac{\frac{\pi_j}{\theta\pi_i} - 1}{\frac{\pi_j}{\theta\pi_i} + 1}\Big] \\
&= \frac{1}{2}\Big[\frac{\theta\pi_j - \pi_i}{\theta\pi_j + \pi_i} - \frac{\pi_j - \pi_i\theta}{\pi_j + \pi_i\theta}\Big] \\
&= \frac{1}{2}\frac{2 * \theta^2\pi_i\pi_j - 2\pi_i\pi_j}{\theta^2\pi_i\pi_j + \theta\pi_i^2 + \theta\pi_j^2 + \pi_i\pi_j} \\
&= \frac{(\theta^2-1)\pi_i\pi_j}{(\pi_i+\theta\pi_j)(\pi_j+\theta\pi_i)}
\end{aligned}$$

The probability p_{ij} now will be $p_{ij} = \frac{1}{4}\int_{-(V_i-V_j)+\eta}^{+\infty} \operatorname{sech}^2(y/2)\mathrm{d}y$. Rewriting this probability in the same way as shown before, we find:

$$p_{ij} = \frac{\pi_i}{\pi_i + \theta\pi_j} \tag{2.29}$$

So to include draws, we've added one extra parameter, θ which is actually e^{η}. Since η is positive, we know that θ is always bigger than 1. In the next section we discuss how we can estimate θ, and the other parameters.

2.2.1. Estimation of parameters

To find the winning potentials $\pi_1, \ldots, \pi_m$ and the parameter θ we again use the method of maximum likelihood, combined with an MM algorithm.
We first need some extra notation. The $m \times m$ matrix w still stores the wins, where w_{ij} represents the number of wins of team i over team j. Let us denote t_{ij} as the number of draws between team i and team j. We store this data in an $m \times m$ matrix t. Note that t is a symmetric matrix, meaning that $t_{ij} = t_{ji}$. Finally let us denote $s_{ij} = w_{ij} + t_{ij}$ to be the number of times team i beat or tied team j and let us denote T to be the total number of ties among all matches.

For the likelihood, we need to remember that $t_{ij} = t_{ji}$. Keeping this in mind we find the likelihood:

$$L(\boldsymbol{\pi}, \theta) = \prod_i \prod_j \Big[\Big(\frac{\pi_i}{\pi_i + \theta\pi_j}\Big)^{w_{ij}} \Big(\frac{(\theta^2-1)\pi_i\pi_j}{(\pi_i+\theta\pi_j)(\pi_j+\theta\pi_i)}\Big)^{\frac{1}{2}t_{ij}}\Big] \tag{2.30}$$

The according log-likelihood of this model then is:

$$\ell(\boldsymbol{\pi},\theta) = \frac{1}{2}\sum_i\sum_j \Big[2w_{ij}\ln\big(\frac{\pi_i}{\pi_i+\theta\pi_j}\big) + t_{ij}\ln\big(\frac{(\theta^2-1)\pi_i\pi_j}{(\theta\pi_i+\pi_j)(\pi_i+\theta\pi_j)}\big)\Big] \tag{2.31}$$

We can rewrite this log-likelihood to be:

$$\ell(\boldsymbol{\pi},\theta) = \frac{1}{2}\sum_i\sum_j \Big[2w_{ij}\ln(\pi_i) - 2w_{ij}\ln(\pi_i+\theta\pi_j) + t_{ij}\ln(\theta^2-1) + t_{ij}\ln(\pi_i) + t_{ij}\ln(\pi_j) - t_{ij}\ln(\theta\pi_i+\pi_j) - t_{ij}\ln(\pi_i+\theta\pi_j)\Big]$$

$$= \sum_i\sum_j \Big[w_{ij}\ln(\pi_i) - w_{ij}\ln(\pi_i+\theta\pi_j)\Big] + \frac{1}{2}\sum_i\sum_j \Big[t_{ij}\ln(\theta^2-1) + t_{ij}\ln(\pi_i) + t_{ij}\ln(\pi_j) - t_{ij}\ln(\theta\pi_i+\pi_j) - t_{ij}\ln(\pi_i+\theta\pi_j)\Big]$$

Now using $\sum_i\sum_j t_{ij}\ln(\theta\pi_i+\pi_j) = \sum_i\sum_j t_{ij}\ln(\pi_i+\theta\pi_j)$
and $\sum_i\sum_j t_{ij}\ln(\pi_i) = \sum_i\sum_j t_{ij}\ln(\pi_j)$, we find that

$$\ell(\boldsymbol{\pi},\theta) = \sum_i\sum_j \Big[w_{ij}\ln(\pi_i) - w_{ij}\ln(\pi_i+\theta\pi_j)] + \frac{1}{2}\Big[t_{ij}\ln(\theta^2-1) + 2t_{ij}\ln(\pi_i) - 2t_{ij}\ln(\theta\pi_i+\pi_j)\Big] \tag{2.32}$$

$$= \sum_i\sum_j \Big[(w_{ij}+t_{ij})(\ln(\pi_i) - \ln(\pi_i+\theta\pi_j)) + \frac{1}{2}t_{ij}\ln(\theta^2-1)\Big] \tag{2.33}$$

We now need to minorize this, so we can find a minorizing function for our MM algorithm.
Applying inequality (2.13), with $x = \pi_i + \theta\pi_j$ and $y = \pi_i^{(k)} + \theta^{(k)}\pi_j^{(k)}$ gives us

$$\ell(\boldsymbol{\pi},\theta) \geq \sum_i\sum_j \Big[(w_{ij}+t_{ij})\big(\ln(\pi_i) + 1 - \ln(\pi_i^{(k)} + \theta^{(k)}\pi_j^{(k)}) - \frac{\pi_i+\theta\pi_j}{\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}\big) + \frac{1}{2}t_{ij}\ln(\theta^2-1)\Big] \tag{2.34}$$

By leaving out the constants we now find our minorizing function

$$Q_k(\boldsymbol{\pi},\theta) = \sum_i\sum_j \Big[(w_{ij}+t_{ij})\big(\ln(\pi_i) - \frac{\pi_i+\theta\pi_j}{\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}\big) + \frac{1}{2}t_{ij}\ln(\theta^2-1)\Big] \tag{2.35}$$

which minorizes the log-likelihood up to a constant. Because of the factor $\theta\pi_j$ in the minorizing function, the parameters are not completely separated.
The solution is to alternately maximize $Q_k(\boldsymbol{\pi},\theta^{(k)})$ as a function of π and $Q_k(\boldsymbol{\pi}^{(k+1)},\theta)$ as a function of θ. This way we obtain a cyclic MM algorithm. Maximizing $Q_k(\boldsymbol{\pi},\theta^{(k)})$ with respect to $\boldsymbol{\pi}$ gives us

$$\pi_i^{(k+1)} = \Big[\sum_j s_{ij}\Big]\Big[\sum_j\big(\frac{s_{ij}}{\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}} + \frac{\theta^{(k)}s_{ji}}{\theta^{(k)}\pi_i^{(k)}+\pi_j^{(k)}}\big)\Big]^{-1} \tag{2.36}$$

Maximizing $Q_k(\boldsymbol{\pi}^{(k+1)},\theta)$ with respect to θ gives us

$$\theta^{(k+1)} = \frac{1}{2C_k} + \sqrt{1+\frac{1}{4C_k^2}} \;\vee\; \theta^{(k+1)} = \frac{1}{2C_k} - \sqrt{1+\frac{1}{4C_k^2}} \tag{2.37}$$

where

$$C_k = \frac{1}{2T}\sum_i\sum_j \frac{\pi_j^{(k+1)}s_{ij}}{\pi_i^{(k+1)}+\theta^{(k)}\pi_j^{(k+1)}} \tag{2.38}$$

Notice that

$$\frac{1}{2C_k} = \sqrt{\frac{1}{4C_k^2}} \leq \sqrt{1+\frac{1}{4C_k^2}}$$

Which means that

$$\frac{1}{2C_k} - \sqrt{1+\frac{1}{4C_k^2}} < 0$$

However θ needs to be greater than 1, so we have only one solution that suffices:

$$\theta^{(k+1)} = \frac{1}{2C_k} + \sqrt{1 + \frac{1}{4C_k^2}} \tag{2.39}$$

Just as in the original model the data needs to satisfy certain constraints, in order to let the algorithm converge to the maximum. In this case the data should satisfy the data assumption as formulated in the original model and there should be at least one draw.

Data assumption model including draws:
In every possible partition of the teams into two nonempty subsets, some team in the second set beats some team in the first set at least once. You can look at this as a directed graph with the m teams as the nodes. There is a directed edge (i, j) from team i to team j whenever team i has beaten team j at least once. Now we assume this graph to be strongly connected, meaning that there is a directed path from i to j $\forall i, j$. There also needs to be at least one draw in the data.

The proof that the algorithm will converge under these conditions can be found in Hunter (2004) [4].

Now that the MM algorithm is constructed, let's take a look at an example in which the data assumption is satisfied.

Example 2.2.1. In this example, we've changed one win from example 2.2.1 into a draw. This gives us the following data:

$w =$

	Team 1	Team 2	Team 3
Team 1	0	19	13
Team 2	0	0	8
Team 3	7	12	0

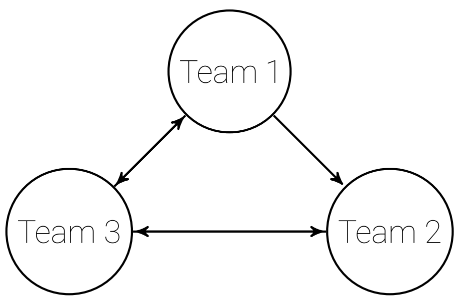
$t =$

	Team 1	Team 2	Team 3
Team 1	0	1	0
Team 2	1	0	0
Team 3	0	0	0

This data causes the log-likelihood to be:

$$\ell(\boldsymbol{\pi},\theta) = \frac{1}{2}\Big[38\ln(\frac{\pi_1}{\pi_1+\theta\pi_2}) + \ln(\frac{(\theta^2-1)\pi_1\pi_2}{(\theta\pi_1+\pi_2)(\pi_1+\theta\pi_2)} + 26\ln(\frac{\pi_1}{\pi_1+\theta\pi_3})$$
$$+\ln(\frac{(\theta^2-1)\pi_2\pi_1}{(\theta\pi_2+\pi_1)(\pi_2+\theta\pi_1)} + 16\ln(\frac{\pi_2}{\pi_2+\theta\pi_3}) + 14\ln(\frac{\pi_3}{\pi_3+\theta\pi_1}) + 24\ln(\frac{\pi_3}{\pi_3+\theta\pi_2})\Big]$$

The wins represented in a graph give:



Since this graph is strongly connected and there is at least 1 draw (there is exactly 1), we know that the algorithm will converge to a unique maximum.

The new MM algorithm terminated after 11 iterations. This means that both $\boldsymbol{\pi}$ as well as θ were updated 11 times. This example resulted in the following winning potentials:

$$\boldsymbol{\pi} = \begin{bmatrix} 2.0092723 \\ 0.3025041 \\ 0.6882237 \end{bmatrix} \tag{2.40}$$

θ converged to 1.0494253.

Because we only changed example 2.1.3 a little bit we can compare the obtained winning potentials very well. Team 3 for example has exactly the same number of wins and losses as in example 2.1.3 and has no draws. The winning potential π_3 however is slightly higher in this example. Even though the results of team 3 didn't change it still has a different winning potential. This has to do with the fact that the winning potentials are relative to the winning potentials of the other teams. If other teams perform less good and there winning potential goes down, this automatically means that the winning potential of team 3 goes up, even though they performed exactly the same.

To show that there need to be draws, for the MM-algorithm to converge we look at the data of example 2.1.3 and we add no draws.

Example 2.2.2. The data for this example then is:

$w =$

	Team 1	Team 2	Team 3
Team 1	0	20	13
Team 2	0	0	8
Team 3	7	12	0

$t =$

	Team 1	Team 2	Team 3
Team 1	0	0	0
Team 2	0	0	0
Team 3	0	0	0

The log-likelihood will then be:

$$\ell(\boldsymbol{\pi}) = 20\ln(\frac{\pi_1}{\pi_1+\theta\pi_2}) + 13\ln(\frac{\pi_1}{\pi_1+\theta\pi_3}) + 8\ln(\frac{\pi_2}{\pi_2+\theta\pi_3}) + 7\ln(\frac{\pi_3}{\pi_3+\theta\pi_1}) + 12\ln(\frac{\pi_3}{\pi_3+\theta\pi_2})\Big] \tag{2.41}$$

One of the equations we need to solve is:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\boldsymbol{\pi}) = -20\frac{\pi_2}{\pi_1+\theta\pi_2} - 16\frac{\pi_3}{\pi_1+\theta\pi_3} - 8\frac{\pi_3}{\pi_2+\theta\pi_3} - 7\frac{\pi_1}{\pi_3+\theta\pi_1} - 12\frac{\pi_2}{\pi_3+\theta\pi_2} = 0 \tag{2.42}$$

Since $\pi_i > 0$ for all i and $\theta > 1$, there are no solutions for this equation. This means that there is no solution in the parameter space. So because there are no draws, the MM algorithm doesn't work.

2.3. Model Including Home Advantage

In this section, the original model will be extended to include a home advantage. Results from the past have shown that in general the team that plays at home has a slight advantage over the opponent. Therefore the home team needs to have an upgrade. The extension introduced here is based on a model of Agresti (1990) [2].
We need to introduce some new notation, because in this model there will be a difference in playing at home or away. Let us denote p_{ij}^{i} to be the probability of team i beating team j while team i plays at home. Let us denote p_{ij}^{j} to be the probability of team i beating team j while team j plays at home.
The general advantage that the home team gets will be provided by a parameter $\gamma > 0$. The winning potential of the team playing at home will be multiplied by γ, such that if team i plays at home, $p_{ij}^{i} = \frac{\gamma\pi_i}{\gamma\pi_i+\pi_j}$. The alternative form then is: $p_{ij}^{i} = \frac{1}{4}\int_{-(V_i-V_j)}^{+\infty}\mathrm{sech}^2(y/2)\mathrm{d}y$, where $V_i = \ln(\gamma\pi_i)$ and $V_j = \ln(\pi_j)$.
In this model, like in the original model, there are no draws.

2.3.1. Estimation of Parameters

To estimate the winning potentials and γ, we use the method of maximum likelihood combined with an MM algorithm. To be able to estimate γ, we need to know whether a team won at home or won away. We therefore store the data in two different matrices. Matrix a contains the home wins and matrix b contains the home losses. The matrix w that only contains the total wins is not necessary anymore. The relationship $w_{ij} = a_{ij} + b_{ji}$ holds as well as $W_i = \sum_j a_{ij} + \sum_j b_{ji}$. Let us denote H to be the total number of home wins. With the data stored in the 3 $m \times m$ matrices, we can determine the likelihood.

The constructed likelihood is:

$$\prod_i \prod_j \Big(\frac{\gamma\pi_i}{\gamma\pi_i + \pi_j}\Big)^{a_{ij}} \Big(\frac{\pi_j}{\gamma\pi_i + \pi_j}\Big)^{b_{ij}} \tag{2.43}$$

The according log-likelihood will then be:

$$\ell(\boldsymbol{\pi}, \boldsymbol{\gamma}) = \sum_i \sum_j \Big[a_{ij} \ln \frac{\gamma\pi_i}{\gamma\pi_i + \pi_j} + b_{ij} \ln \frac{\pi_j}{\gamma\pi_i + \pi_j}\Big] \tag{2.44}$$

We can rewrite log-likelihood (2.44) to obtain

$$\begin{aligned}
\ell(\boldsymbol{\pi}, \gamma) &= \sum_i \sum_j \Big[a_{ij} \ln(\gamma) + a_{ij} \ln(\pi_i) - a_{ij} \ln(\gamma\pi_i + \pi_j) + b_{ij} \ln(\pi_i) - b_{ij} \ln(\gamma\pi_i + \pi_j)\Big] \\
&= \sum_i \sum_j \Big[(a_{ij} + b_{ij}) \ln(\pi_i) + a_{ij} \ln(\gamma) - (a_{ij} + b_{ij}) \ln(\gamma\pi_i + \pi_j)\Big] \\
&= \sum_i \sum_j (a_{ij} + b_{ij}) \ln(\pi_i) + \sum_i \sum_j a_{ij} \ln(\gamma) - \sum_i \sum_j (a_{ij} + b_{ij}) \ln(\gamma\pi_i + \pi_j)
\end{aligned}$$

Using $\sum_i \sum_j (a_{ij} + b_{ij}) = \sum_i W_i$ (the total number of wins) and $H := \sum_i \sum_j a_{ij}$ (the total number of home wins), we then find:

$$\ell(\boldsymbol{\pi}, \gamma) = H \ln(\gamma) + \sum_i W_i \ln(\pi_i) - \sum_i \sum_j (a_{ij} + bij) \ln(\gamma\pi_i + \pi_j)$$

To construct a minorizing function, we again apply inequality (2.13), where $x = \gamma\pi_i + \pi_j$ and $y = \gamma^{(k)}\pi_i^{(k)} + \pi_j^{(k)}$. This way we obtain the minorizing function:

$$Q_k(\boldsymbol{\pi}, \gamma) = H \ln \gamma + \sum_i W_i \ln \pi_i - \sum_i \sum_j \Big[\frac{(a_{ij} + b_{ij})(\gamma\pi_i + \pi_j)}{\gamma^{(k)}\pi_i^{(k)} + \pi_j^{(k)}}\Big] \tag{2.45}$$

Q_k minorizes the log-likelihood up to a constant.
Because we have a factor $\gamma\pi_i$ in the function, the parameters are not separated. Therefore we construct a cyclic MM algorithm that maximizes $Q_k(\boldsymbol{\pi}, \gamma^{(k)})$ as a function of $\boldsymbol{\pi}$ and $Q_k(\boldsymbol{\pi}^{(k+1)}, \gamma)$ as a function of γ. Maximizing $Q_k(\boldsymbol{\pi}, \gamma^{(k)}$ with respect to $\boldsymbol{\pi}$ gives us

$$\pi_i^{(k+1)} = W_i \Big[\sum_j \frac{(a_{ij} + b_{ij})\gamma^{(k)}}{\gamma^{(k)}\pi_i^{(k)} + \pi_j^{(k)}} + \sum_j \frac{a_{ji} + b_{ji}}{\gamma^{(k)}\pi_j^{(k)} + \pi_i^{(k)}}\Big]^{-1} \tag{2.46}$$

Maximizing $Q_k(\boldsymbol{\pi}^{(k+1)}, \gamma)$ with respect to γ gives us

$$\gamma^{(k+1)} = H \Big[\sum_i \sum_j \frac{(a_{ij} + b_{ij})\pi_i^{(k+1)}}{\gamma^{(k)}\pi_i^{(k+1)} + \pi_j^{(k+1)}}\Big]^{-1} \tag{2.47}$$

The data also needs to satisfy certain conditions in order for this new MM algorithm to converge to the maximum. We've stated these conditions in the following data assumption.

Data assumption model including home advantage:
In every possible partition of the teams into two nonempty subsets A and B, some team in A beats some team in B as home team, and some team in A beats some team in B as visiting team.

In other words, both matrix a and b need to satisfy the data assumption of the original model. Hence we could also represent matrices a and b as directed graphs. In practice it's easy to see if the assumption is not satisfied by inspecting the matrices. If there is a row or column solely consisting of zero's in matrices a or b, then the assumption is not satisfied.

The proof that the MM algorithm will converge under this assumption can be found in Hunter(2004) [4].

To show an example we use the data of Example 2.1.3, but now each win is specified to be at home or away.

Example 2.3.1. The following data is given:

$a =$

	Team 1	Team 2	Team 3
Team 1	0	15	10
Team 2	0	0	5
Team 3	4	7	0

$b =$

	Team 1	Team 2	Team 3
Team 1	0	0	3
Team 2	5	0	5
Team 3	3	3	0

For this example, the log-likelihood is:

$$\begin{aligned}\ell(\boldsymbol{\pi},\gamma) = 15\ln(\frac{\gamma\pi_1}{\gamma\pi_1+\pi_2}) + 10\ln(\frac{\gamma\pi_1}{\gamma\pi_1+\pi_3}) + 3\ln(\frac{\pi_3}{\gamma\pi_1+\pi_3}) + 5\ln(\frac{\pi_1}{\gamma\pi_2+\pi_1})\\ +5\ln(\frac{\gamma\pi_2}{\gamma\pi_2+\pi_3}) + 4\ln(\frac{\gamma\pi_3}{\gamma\pi_3+\pi_1}) + 7\ln(\frac{\gamma\pi_3}{\gamma\pi_3+\pi_2}) + 5\ln(\frac{\pi_2}{\gamma\pi_3+\pi_2})\end{aligned}$$

It turns out that the data in this example satisfies the data assumption for this extension. Hence we can use the MM algorithm to determine the maximum likelihood parameters. The algorithm terminates after 13 iterations, returning the winning potentials:

$$\boldsymbol{\pi} = \begin{bmatrix} 2.0015922 \\ 0.2805631 \\ 0.7178447 \end{bmatrix} \tag{2.48}$$

γ converged to 1.7979249, which means that in this example there is quite a significant home advantage.
We also see that, even though the teams had the same amount of wins against each other as in Example 2.1.3, the winning potentials differ a bit. We can conclude that winning at home or away, contribute differently to the winning potentials due to γ.

2.4. Model including Draws and Home Advantage

Since both the idea of involving draws as well as the idea of a home adventage seem very plausible to use when predicting football, ideally we want to combine those two extensions to create a new model by adding both parameters θ (= $^{\eta}$) and γ to the model. In order to do this, we need to use the ideas of both extensions and combine them.
Let's start with the alternative form $p_{ij} = \frac{1}{4}\int_{-(V_i-V_j)-\eta}^{-(V_i-V_j)+\eta} \text{sech}^2(y/2)\mathrm{d}y$. First of all, we use that if team i plays at home, $V_i = \ln(\gamma\pi_i)$ and $V_j = \ln(\pi_j)$ while if team j plays at home, $V_i = \ln(\pi_i)$ and $V_j = \ln(\gamma\pi_j)$. This is basically what we've done for our extension for home advantage. Now suppose team i plays at home. Using the idea of the extension for draws we say that there would be a draw if $|V_i - V_j| < \eta$. Hence the probability for a draw while team i plays it home, which we will denote as $p^i_{i=j}$ will be

$$\begin{aligned} p^i_{i=j} &= \frac{1}{4}\int_{-(V_i-V_j)-\eta}^{-(V_i-V_j)+\eta} \text{sech}^2(y/2)\mathrm{d}y \\ &= \frac{1}{4}\int_{-(\ln(\gamma\pi_i)-\ln(\pi_j))-\eta}^{-(\ln(\gamma\pi_i)-\ln(\pi_j))+\eta} \text{sech}^2(y/2)\mathrm{d}y \end{aligned}$$

Rewriting this, just as in section 2.2 we find:

$$\begin{aligned} p^i_{i=j} &= \frac{1}{4}\int_{-(V_i-V_j)-\eta}^{-(V_i-V_j)+\eta} \operatorname{sech}^2(y/2)\mathrm{d}y \\ &= \frac{(\theta^2-1)\gamma\pi_i\pi_j}{(\gamma\pi_i+\theta\pi_j)(\pi_j+\theta\gamma\pi_i)} \end{aligned}$$

For the probability of team i beating team j, while i plays at home we find:

$$p^i_{ij} = \frac{1}{4}\int_{-(V_i-V_j)+\eta}^{+\infty} \operatorname{sech}^2(y/2)\mathrm{d}y = \frac{1}{4}\Big[2\tanh(\frac{y}{2})\Big]_{-(V_i-V_j)+\eta}^{+\infty} \tag{2.49}$$

Rewriting this gives us:

$$p^i_{ij} = \frac{\gamma\pi_i}{\gamma\pi_i+\theta\pi_j} \tag{2.50}$$

The probability that team j beats team i, while team i plays at home, denoted as p^i_{ji} will be:

$$p^i_{ji} = \frac{\pi_j}{\pi_j+\theta\gamma\pi_i} \tag{2.51}$$

In this model, besides the winning potentials, both θ and γ need to be estimated.

2.4.1. Estimation of Parameters

In this model we require 3 different $m\times m$ matrices to store the data in. Matrix a which contains the home wins, b which contains the home losses and matrix v that contains the home draws. Let us denote T as the total number of draws of all teams and H as the total number of home wins. Also let us denote T_i as the total number of draws of team i, W_i as the total number of wins of team i. To estimate the parameters, just as in the previous models we use the method of maximum likelihood in combination with an MM algorithm.

The likelihood for this model will be:

$$L(\boldsymbol{\pi},\gamma,\theta) = \prod_i\prod_j\Big[\Big(\frac{\gamma\pi_i}{\gamma\pi_i+\theta\pi_j}\Big)^{a_{ij}}\Big(\frac{\pi_j}{\pi_j+\theta\gamma\pi_i}\Big)^{b_{ij}}\Big(\frac{(\theta^2-1)\gamma\pi_i\pi_j}{(\gamma\pi_i+\theta\pi_j)(\pi_j+\theta\gamma\pi_i)}\Big)^{v_{ij}}\Big] \tag{2.52}$$

The matching log-likelihood then will be:

$$\ell(\boldsymbol{\pi},\gamma,\theta) = \sum_i\sum_j\Big[a_{ij}\ln(\frac{\gamma\pi_i}{\gamma\pi_i+\theta\pi_j}) + b_{ij}\ln(\frac{\pi_j}{\pi_j+\theta\gamma\pi_i}) + v_{ij}\ln(\frac{(\theta^2-1)\gamma\pi_i\pi_j}{(\gamma\pi_i+\theta\pi_j)(\pi_j+\theta\gamma\pi_i)})\Big] \tag{2.53}$$

We can rewrite this log-likelihood until we get:

$$\begin{aligned} \ell(\boldsymbol{\pi},\gamma,\theta) &= \sum_i\sum_j\Big[a_{ij}\ln(\frac{\gamma\pi_i}{\gamma\pi_i+\theta\pi_j}) + b_{ij}\ln(\frac{\pi_j}{\pi_j+\theta\gamma\pi_i}) + v_{ij}\ln(\frac{(\theta^2-1)\gamma\pi_i\pi_j}{(\gamma\pi_i+\theta\pi_j)(\pi_j+\theta\gamma\pi_i)})\Big] \\ &= \sum_i\sum_j\Big[a_{ij}\ln(\gamma\pi_i) - a_{ij}\ln(\gamma\pi_i+\theta\pi_j) + b_{ij}\ln(\pi_j) - b_{ij}\ln(\pi_j+\theta\gamma\pi_i) \\ &\quad + v_{ij}\ln(\theta^2-1) + v_{ij}\ln(\gamma\pi_i) + v_{ij}\ln(\pi_j) - v_{ij}\ln(\gamma\pi_i+\theta\pi_j) - v_{ij}\ln(\pi_j+\theta\gamma\pi_i)\Big] \\ &= \sum_i\sum_j\Big[(a_{ij}+v_{ij})\ln(\gamma\pi_i) - (a_{ij}+v_{ij})\ln(\gamma\pi_i+\theta\pi_j) + (b_{ij}+v_{ij})\ln(\pi_j) - (b_{ij}+v_{ij})\ln(\pi_j+\theta\gamma\pi_i) + v_{ij}\ln(\theta^2-1)\Big] \end{aligned}$$

Now that we've found the log-likelihood we want to construct a minorizing function. To do this, we apply inequality (2.13) twice. The first time we choose $x=\gamma\pi_i+\theta\pi_j$ and $y=\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}$. The second time we choose $x=\pi_j+\theta\gamma\pi_i$ and $y=\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}$. This gives us:

$$\begin{aligned} \ell(\boldsymbol{\pi},\gamma,\theta) \geq \sum_i\sum_j\Big[&(a_{ij}+v_{ij})\ln(\gamma\pi_i) + (a_{ij}+v_{ij})(1-\ln(\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}) - \frac{\gamma\pi_i+\theta\pi_j}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}) + (b_{ij}+v_{ij})\ln(\pi_j) \\ &+(b_{ij}+v_{ij})(1-\ln(\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}) - \frac{\pi_j+\theta\gamma\pi_i}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}) + v_{ij}\ln(\theta^2-1)\Big] \end{aligned}$$

By removing the constant terms, we find the desired function Q_k that minorizes $\ell(\boldsymbol{\pi},\gamma,\theta)$ up to a constant.

$$Q_k(\boldsymbol{\pi},\gamma,\theta)=\sum_i\sum_j\Big[(a_{ij}+v_{ij})\ln(\gamma\pi_i)-(a_{ij}+v_{ij})\frac{\gamma\pi_i+\theta\pi_j}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}$$
$$+(b_{ij}+v_{ij})\ln(\pi_j)-(b_{ij}+v_{ij})\frac{\pi_j+\theta\gamma\pi_i}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}+v_{ij}\ln(\theta^2-1)\Big]$$

Because there are factors that include multiple variables like $\gamma\pi_i$, $\theta\pi_j$ and $\theta\gamma\pi_i$, we need to use a cyclic algorithm that maximizes $Q_k(\boldsymbol{\pi},\gamma^{(k)},\theta^{(k)})$ as function of $\boldsymbol{\pi}$, $Q_k(\boldsymbol{\pi}^{(k)},\gamma,\theta^{(k)})$ as a function of γ and $Q_k(\boldsymbol{\pi}^{(k)},\gamma^{(k)},\theta)$ as a function of θ. By finding these maxima, updating Q_k and then finding new maxima we construct the new MM algorithm.

To maximize Q_k as a function of $\boldsymbol{\pi}$ we need to solve:

$$\frac{\mathrm{d}}{\mathrm{d}\pi_i}Q_k(\boldsymbol{\pi},\gamma^{(k)},\theta^{(k)})=0$$

for every i. We find that $\frac{\mathrm{d}}{\mathrm{d}\pi_i}Q_k(\boldsymbol{\pi},\gamma^{(k)},\theta^{(k)})=$

$$\sum_j\Big[\frac{(a_{ij}+v_{ij})}{\pi_i}-\frac{(a_{ij}+v_{ij})\gamma^{(k)}}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}-\frac{(a_{ji}+v_{ji})\theta^{(k)}}{\gamma^{(k)}\pi_j^{(k)}+\theta^{(k)}\pi_i^{(k)}}+\frac{(b_{ji}+v_{ji})}{\pi_i}-\frac{(b_{ij}+v_{ij})\theta^{(k)}\gamma^{(k)}}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}-\frac{(b_{ji}+v_{ji})}{\pi_i^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_j^{(k)}}\Big]$$

In this equation there's a factor $\frac{(a_{ij}+v_{ij})}{\pi_i}$ and a factor $\frac{(b_{ji}+v_{ji})}{\pi_i}$. Looking at $\sum_j a_{ij}+b_{ji}+v_{ij}+v_{ji}$, we can say that $\sum_j a_{ij}+b_{ji}=W_i$, which is the total number of wins of team i, while $sum_j=v_{ij}+v_{ji}=T_i$. So $\sum_j a_{ij}+b_{ji}+v_{ij}+v_{ji}=W_i+T_i$. Using this we find:

$$\frac{\mathrm{d}}{\mathrm{d}\pi_i}Q_k(\boldsymbol{\pi},\gamma^{(k)},\theta^{(k)})=\frac{W_i+T_i}{\pi_i}-\sum_j\Big[\frac{(a_{ij}+v_{ij})\gamma^{(k)}}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}+\frac{(a_{ji}+v_{ji})\theta^{(k)}}{\gamma^{(k)}\pi_j^{(k)}+\theta^{(k)}\pi_i^{(k)}}+\frac{(b_{ij}+v_{ij})\theta^{(k)}\gamma^{(k)}}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}+\frac{(b_{ji}+v_{ji})}{\pi_i^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_j^{(k)}}\Big]=0$$

Hence the next iteration for π_i in the MM algorithm will be:

$$\pi_i^{(k+1)}=(W_i+T_i)\Big[\sum_j\Big(\frac{(a_{ij}+v_{ij})\gamma^{(k)}}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}+\frac{(a_{ji}+v_{ji})\theta^{(k)}}{\gamma^{(k)}\pi_j^{(k)}+\theta^{(k)}\pi_i^{(k)}}+\frac{(b_{ij}+v_{ij})\theta^{(k)}\gamma^{(k)}}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}+\frac{(b_{ji}+v_{ji})}{\pi_i^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_j^{(k)}}\Big)\Big]^{-1} \tag{2.54}$$

To maximize γ we solve:

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}Q_k(\boldsymbol{\pi}^{(k)},\gamma,\theta^{(k)})=0$$

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}Q_k(\boldsymbol{\pi}^{(k)},\gamma,\theta^{(k)})=\sum_i\sum_j\Big[\frac{(a_{ij}+v_{ij})}{\gamma}-\frac{(a_{ij}+v_{ij})\pi_i^{(k)}}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}-\frac{(b_{ij}+v_{ij})\theta^{(k)}\pi_i^{(k)}}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}\Big]$$

Now $\sum_i\sum_j(a_{ij}+v_{ij})$ is the sum over all the home wins and the home draws. The sum over all the home draws is actually equal to the total number of draws, because the total number of home draws equals the total number of draws, since in every draw one team plays at home. Therefore $\sum_i\sum_j(a_{ij}+v_{ij})=H+T$. Rewriting then shows:

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}Q_k(\boldsymbol{\pi}^{(k)},\gamma,\theta^{(k)})=\frac{H+T}{\gamma}-\sum_i\sum_j\Big[\frac{(a_{ij}+v_{ij})\pi_i^{(k)}}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}+\frac{(b_{ij}+v_{ij})\theta^{(k)}\pi_i^{(k)}}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}\Big]=0$$

By solving the equation above, the next iteration of γ then would be:

$$\gamma^{(k+1)}=(H+T)\Big[\sum_i\sum_j\Big(\frac{(a_{ij}+v_{ij})\pi_i^{(k)}}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}+\frac{(b_{ij}+v_{ij})\theta^{(k)}\pi_i^{(k)}}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}\Big)\Big]^{-1} \tag{2.55}$$

Finally to maximize θ we solve:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}Q_k(\boldsymbol{\pi}^{(k)},\gamma^{(k)},\theta)=0$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta}Q_k(\boldsymbol{\pi}^{(k)},\gamma^{(k)},\theta)=\sum_i\sum_j\Big[-\frac{(a_{ij}+v_{ij})\pi_j^{(k)}}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}-\frac{(b_{ij}+v_{ij})\gamma^{(k)}\pi_i^{(k)}}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}+\frac{v_{ij}2\theta}{\theta^2-1}\Big]$$

From earlier it is known that $\sum_i\sum_j v_{ij}=T$. Using this and using notation $S_k:=\sum_i\sum_j\Big[\frac{(a_{ij}+v_{ij})\pi_j^{(k)}}{\gamma^{(k)}\pi_i^{(k)}+\theta^{(k)}\pi_j^{(k)}}+\frac{(b_{ij}+v_{ij})\gamma^{(k)}\pi_i^{(k)}}{\pi_j^{(k)}+\theta^{(k)}\gamma^{(k)}\pi_i^{(k)}}\Big]$ we find that we need to solve:

$$\frac{2T\theta}{\theta^2-1}-S_k=0$$

This gives us:

$$S_k\theta^2-2T\theta-S_k=0$$

Solving this equation, we find two potential maxima for θ:

$$\theta_1=\frac{2T+\sqrt{4T^2+4S_k^2}}{2S_k} or \theta_2=\frac{2T-\sqrt{4T^2+4S_k^2}}{2S_k}$$

But since we want θ to be larger than 1 and $2T-\sqrt{4T^2+4S_k^2}<0$, the solution $\theta_2=\frac{2T-\sqrt{4T^2+4S_k^2}}{2S_k}$ does not suffice. Therefore the next iteration for θ in the MM algorithm will be:

$$\theta^{(k+1)}=\frac{2T+\sqrt{4T^2+4S_k^2}}{2S_k} \tag{2.56}$$

Again, the data needs to satisfy certain conditions in order for this MM algorithm to converge to the maximum. These conditions are exactly the same as the conditions for the previous models, but combined.

Data assumption model including draws and home advantage In every possible partition of the teams into two nonempty subsets A and B, some team in A beats some team in B as home team, and some team in A beats some team in B as visiting team.

In other words, both matrix a and b need to satisfy the data assumption of the original model. Hence we could also represent matrices a and b as directed graphs. In practice it's easy to see if the assumption is not satisfied by inspecting the matrices. If there is a row or column solely consisting of zero's in matrices a or b, then the assumption is not satisfied.
There also needs to be at least one draw in the data.

To show an example, we changed one win into a draw from the example used in the home advantage model.

Example 2.4.1. The data then looks like this:

$a=$

	Team 1	Team 2	Team 3
Team 1	0	14	10
Team 2	0	0	5
Team 3	4	7	0

$b=$

	Team 1	Team 2	Team 3
Team 1	0	0	3
Team 2	5	0	5
Team 3	3	3	0

$v=$

	Team 1	Team 2	Team 3
Team 1	0	1	0
Team 2	0	0	0
Team 3	0	0	0

The log-likelihood following from this data then is:

$$\ell(\boldsymbol{\pi},\gamma,\theta) = 14\ln(\frac{\gamma\pi_1}{\gamma\pi_1+\theta\pi_2}) + \frac{\theta^2-1\gamma\pi_1\pi_2}{(\gamma\pi_1+\theta\pi_2)(\pi_2+\theta\gamma\pi_1} + 10\ln(\frac{\gamma\pi_1}{\gamma\pi_1+\theta\pi_2}) + 3\ln(\frac{\pi_3}{\pi_3+\theta\gamma\pi_1}) + 5\ln(\frac{\pi_1}{\pi_1+\theta\gamma\pi_2})$$
$$+5\ln(\frac{\gamma\pi_2}{\gamma\pi_2+\theta\pi_3} + 5\ln(\frac{\pi_3}{\pi_3+\theta\gamma\pi_2}) + 4\ln(\frac{\gamma\pi_3}{\gamma\pi_3+\theta\pi_1}) + 3\ln(\frac{\pi_1}{\pi_1+\theta\gamma\pi_3}) + 7\ln(\frac{\gamma\pi_3}{\gamma\pi_3+\theta\pi_2}) + 3\ln(\frac{\pi_2}{\pi_2+\theta\gamma\pi_3})$$

The data suffices the data assumption for this model, so the MM algorithm converges to a unique maximum. Our algorithm terminated after 9 iterations to return the winning potentials:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.9213365 \\ 0.3273345 \\ 0.7513291 \end{bmatrix} \tag{2.57}$$

The returned θ, the threshold for a draw was 1.0518555, while the returned γ, the home advantage factor was 1.6282503.

In this chapter we've introduced the original Bradley-Terry model. It is shown that this model uses winning potentials to assign probabilities to events. Using these winning potentials, the teams in a league can be ranked. The winning potentials were obtained using the maximum likelihood method in combination with an MM algorithm. Already existing extensions of the model to include draws and to include home advantage have been introduced. We combined these two extensions to create a new model that could include draws as well as home advantage.

In the next chapter we will look closely at the data of the Eredivisie, which is the league we want to do predictions about. We need to determine which data has the most relevance to base new winning potentials upon and whether this data satisfies the data assumptions introduced in this chapter.

3
Data Analysis

Now that has been explained how the Bradley-Terry model and its extensions work, it is time to look at the available data we have to apply the model on. In this chapter we will analyze this data thoroughly. The data we have at our disposal is found on the website http://www.football-data.co.uk/netherlandsm.php. They offer the results of all the matches played in the Eredivisie, starting from season 1993/1994 up until now.
However there is no point in using data from season 1993/1994 to predict the winning potential of a team right now. It's obvious that data from over 20 years ago is irrelevant for predicting the performances of teams right now. Teams have completely changed in over time and using data from such a long time ago could actually make our predictions worse. We have to decide which data is actually still relevant to use for the coming season. It makes sense to use the most recent seasons, because the teams still look quite like how they were then and therefore give us the best idea of how the teams will perform in the next season. This actually implies that the 'best' data to use would only be the data from last season, because the teams have the fewest changes compared to that season. On the other hand however, the more data we use, the more accurate we can make the estimates of the parameters. This is the first dilemma: on the one hand we don't want to use data that's relatively old, but on the other hand we want to use as much data as possible.
To decide which data to use we have chosen some criteria:

- There is no point in using data from 2009/2010 or earlier because the teams have just changed too much.
- The performances of the teams should be relatively stable over the chosen seasons.

The reason we say that the performances of the teams should be stable is because the model bases a winning potential on the data we choose. Let's say a team is very unstable and it had the following rankings:

- Season 1: 1
- Season 2: 17
- Season 3: 5
- Season 4: 12
- Season 5: 16
- Season 6: 3

Our model might rank this team 7th for the next season. But looking at the rankings above, the team could just as well be ranked 2nd or 15th in the next season. So there's not really a point in ranking this team with the BT model, because the team is just too unstable. In Figure 3.1 we see the rankings of all the seasons we're considering.

Ranking	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18
1.	Ajax	Ajax	Ajax	Ajax	PSV	PSV	Feyenoord	PSV
2.	FC Twente	Feyenoord	PSV	Feyenoord	Ajax	Ajax	Ajax	Ajax
3.	PSV	PSV	Feyenoord	FC Twente	AZ	Feyenoord	PSV	AZ
4.	AZ	AZ	Vitesse	PSV	Vitesse	AZ	FC Utrecht	Feyenoord
5.	ADO Den Haag	sc Heerenveen	FC Utrecht	FC Groningen	sc Heerenveen	Heracles Almelo	Vitesse	Vitesse
6.	FC Groningen	Vitesse	FC Twente	AZ	Feyenoord	FC Utrecht	AZ	FC Utrecht
7.	Roda JC	RKC Waalwijk	FC Groningen	sc Heerenveen	PEC Zwolle	FC Groningen	FC Twente	ADO Den Haag
8.	Heracles Almelo	FC Twente	sc Heerenveen	Vitesse	FC Groningen	PEC Zwolle	FC Groningen	sc Heerenveen
9.	FC Utrecht	N.E.C.	ADO Den Haag	ADO Den Haag	Willem II	Vitesse	sc Heerenveen	PEC Zwolle
10.	Feyenoord	Roda JC	AZ	FC Utrecht	FC Twente	N.E.C.	Heracles Almelo	Heracles Almelo
11.	N.E.C.	FC Utrecht	PEC Zwolle	PEC Zwolle	FC Utrecht	ADO Den Haag	ADO Den Haag	Excelsior
12.	sc Heerenveen	Heracles Almelo	Heracles Almelo	SC Cambuur	SC Cambuur	sc Heerenveen	Excelsior	FC Groningen
13.	NAC Breda	NAC Breda	NAC Breda	Go Ahead Eagles	ADO Den Haag	FC Twente	Willem II	Willem II
14.	De Graafschap	FC Groningen	RKC Waalwijk	Heracles Almelo	Heracles Almelo	Roda JC	PEC Zwolle	NAC Breda
15.	Vitesse	ADO Den Haag	N.E.C.	NAC Breda	Excelsior	Excelsior	Sparta R'dam	VVV-Venlo
16.	Excelsior	VVV-Venlo	Roda JC	RKC Waalwijk	NAC Breda	Willem II	Roda JC	Roda JC
17.	VVV-Venlo	De Graafschap	VVV-Venlo	N.E.C.	Go Ahead Eagles	De Graafschap	N.E.C.	Sparta R'dam
18.	Willem II	Excelsior	Willem II	Roda JC	FC Dordrecht	SC Cambuur	Go Ahead Eagles	FC Twente

Figure 3.1: The complete rankings of the past 8 seasons

There are 10 teams that occur in every of these past 8 seasons. For every team out of these 10 teams we've calculated their average ranking, which can be seen in Table 3.1.

Team	Average rank
ADO Den Haag	10
Ajax	1.5
AZ	5
FC Groningen	8.375
FC Utrecht	7.75
Feyenoord	3.875
Heracles	10.625
PSV	2.25
SC Heerenveen	8.25
Vitesse	7

Table 3.1: Average rank over the past 8 seasons

We then look at the ranking of a team in each of these 8 seasons and calculate the absolute difference in rank between the rank of that season and the average rank. This way we see how far off a team is ranked, compared to its average rank. By summing these absolute differences for each season over all the 10 teams we have an indication of how much every season deviates from the average ranking. We will call this sum the 'deviation' from the average ranking. The deviation gives us an indication of how useful the data is, because if there's a huge deviation from the average the data is likely to only worsen our predictions. So the larger the deviation of a season is, the less useful the data of that season is. In Table 3.2 you can see the calculated deviations.

As a result of the obtained deviations we've decided to use the data of the past 6 seasons. We want to use the most 'stable' data as possible and the seasons 2010/2011 and 2011/2012 deviate the most from the average ranking. This fact, combined with the fact that they also are the two 'oldest' seasons gives us the reason to remove them from the data.

So the relevant data consists of seasons 2012/13 up until 2017/18. Let's look at these six seasons in more detail, since we will use them to do predictions about the ranking of season 2018/2019.

In the 2018/2019 season, the league consists of the teams found in Table 3.3.

One of the most important things we need to deal with, is the fact that not all teams have played in the Eredivisie for the past six seasons. In Table 3.3, for every team the number of seasons they played in the Eredivisie out of the six seasons we chose is shown between brackets.

Season	Deviation
2010/2011	31.375
2011/2012	23.625
2012/2013	16.375
2013/2014	17.375
2014/2015	22.125
2015/2016	19.125
2016/2017	13.625
2017/2018	15.125

Table 3.2: The total deviation from the average ranking per season

Teams (Number of seasons in Eredivisie out of selected 6)	
PSV (6)	SC Heerenveen (6)
Ajax (6)	ADO Den Haag (6)
Feyenoord (6)	Willem II (5)
FC Utrecht (6)	FC Emmen (0)
Vitesse (6)	Excelsior (4)
Heracles Almelo (6)	FC Groningen (6)
AZ (6)	PEC Zwolle (6)
VVV-Venlo (2)	NAC Breda (4)
Fortuna Sittard (0)	De Graafschap (1)

Table 3.3: All teams of the Eredivisie season 2018/19

We observe that there is quite some variation in the number of seasons we have about each team. This means that we would base the winning potential for teams like PSV on six seasons, while the winning potential of De Graafschap can be based only on one season. Even worse is that FC Emmen and Fortuna Sittard haven't played Eredivisie in the past 6 seasons. This means that we have no data at all of these two teams. Something else we need to deal with is that if a team is not in the Eredivisie for a season, this means that all other teams haven't played against this team in that season. In principle, a team plays 34 matches per season. However we will not use data of matches against a team that's not in season 2018/2019 of the Eredivisie. This means that if one of our 2018/2019 teams has not been in the league for one or more of the past 6 seasons, we have less data for all teams. The total number of matches per team in our data can be seen in Table 3.4.

PSV: 152	Ajax: 152	Feyenoord: 152
FC Utrecht: 152	Vitesse: 152	Heracles: 152
AZ: 152	VVV Venlo: 54	Fortuna Sittard: 0
Heerenveen: 152	ADO Den Haag: 152	Willem II: 130
FC Emmen: 0	Excelsior: 104	FC Groningen: 152
PEC Zwolle: 152	NAC Breda: 102	De Graafschap: 26

Table 3.4: The total number of matches per team in the chosen data

Not only do not all teams play the same amount of matches, but besides FC Emmen and Fortuna Sittard there are also other teams that don't play against each other in our data. In Figure 3.1 you can see that we have no data of De Graafschap and VVV-Venlo playing in the same season, or De Graafschap and NAC Breda. We want to find out whether it's still possible to rank all teams (beside FC Emmen and Fortuna Sittard) based on our data.

In order for the original Bradley Terry model to obtain a clear ranking structure, our data needs to satisfy the Data Assumption from this model, which can be found in Chapter 2. We can check whether the data satisfies this assumption, by representing the data as a directed graph as shown before. If the graph is strongly connected, we know that the data satisfies the Data Assumption and we can apply the BT model, meaning that we can find winning potentials for the teams.

In Figure 3.2 the data is represented as a directed graph as described before. It turns out that this graph is strongly connected. The graph has connectivity $k = 3$, which means that at least 3 teams need to be removed, for the graph to be disconnected. Because the graph is strongly connected, the selected data (excluding FC Emmen and Fortuna Sittard) satisfies the data assumption for the original Bradley Terry model, so we can find winning potentials using that model.

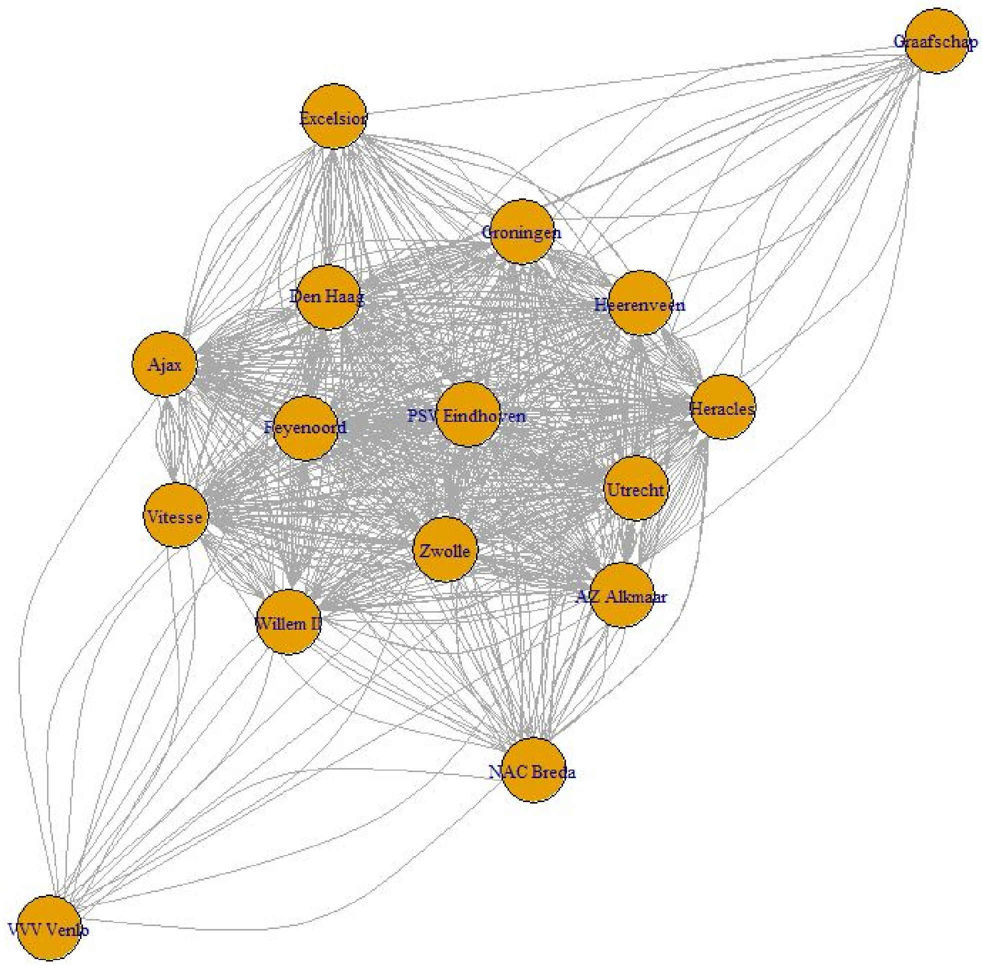


Figure 3.2: The representation of the data as a directed graph

3.1. Home advantage and Draws

As just shown, the basic Bradley-Terry model is applicable on the selected data. In Chapter 2 however, we've also introduced some extensions of this model. To check if these extensions are needed we've examined the chosen data, that contains 1836 matches.
Out of these 1836 matches, 455 matches ended in a draw. This is almost a quarter. If we would use the original BT model, which ignores draws, this means that we would ignore almost 25% of the data.

Out of the remaining 1381 matches, which all ended in a win, we found that 836 of those were Home wins. That's over 60% of all wins. By using a hypothesis test, we could show that this data indeed gives reason to include a home advantage in the model. In Table 3.5 we see how these wins were divided over the different teams. We observe that almost every team seems to have a home advantage. It is however not evident that

Team	Home wins	Away wins
Ajax	77	61
AZ	50	44
ADO Den Haag	40	24
Excelsior	17	16
Feyenoord	73	50
FC Groningen	44	25
sc Heerenveen	45	30
Heracles Almelo	41	26
NAC Breda	23	10
PSV	82	61
FC Utrecht	55	33
Vitesse	48	42
VVV Venlo	6	7
Willem II	27	16
PEC Zwolle	44	26
De Graafschap	3	2

Table 3.5: Home and Away wins per team

every home advantage is the same. Excelsior and NAC Breda for example both have 33 wins, but NAC had 23 of them at home while Excelsior only had 17. And VVV Venlo even had more wins away than wins at home.

To make sure that the extensions can be used, the extra Data Assumptions must be checked. For the extensions including draws, there must be at least one draw. There are 455 draws in the data, so this assumption is satisfied. For the extension including home advantage, every team must win and lose, both home and away at least once. We can again check this by representing the data as a graph. This time we should make two graphs, one where the edges are represented by the home wins and one where the edges are represented by the away wins. Both these graphs should be strongly connected in order for the model to work.
In Figure 3.3 we plotted the graph where the edges represent the home wins. The graph is strongly connected, with connectivity $k = 2$.
In Figure 3.4 the graph where the edges represent away wins is plotted. This graph is also strongly connected, the according connectivity is $k = 1$.

Since all the assumptions are satisfied, we can apply all the extensions we introduced on this data.

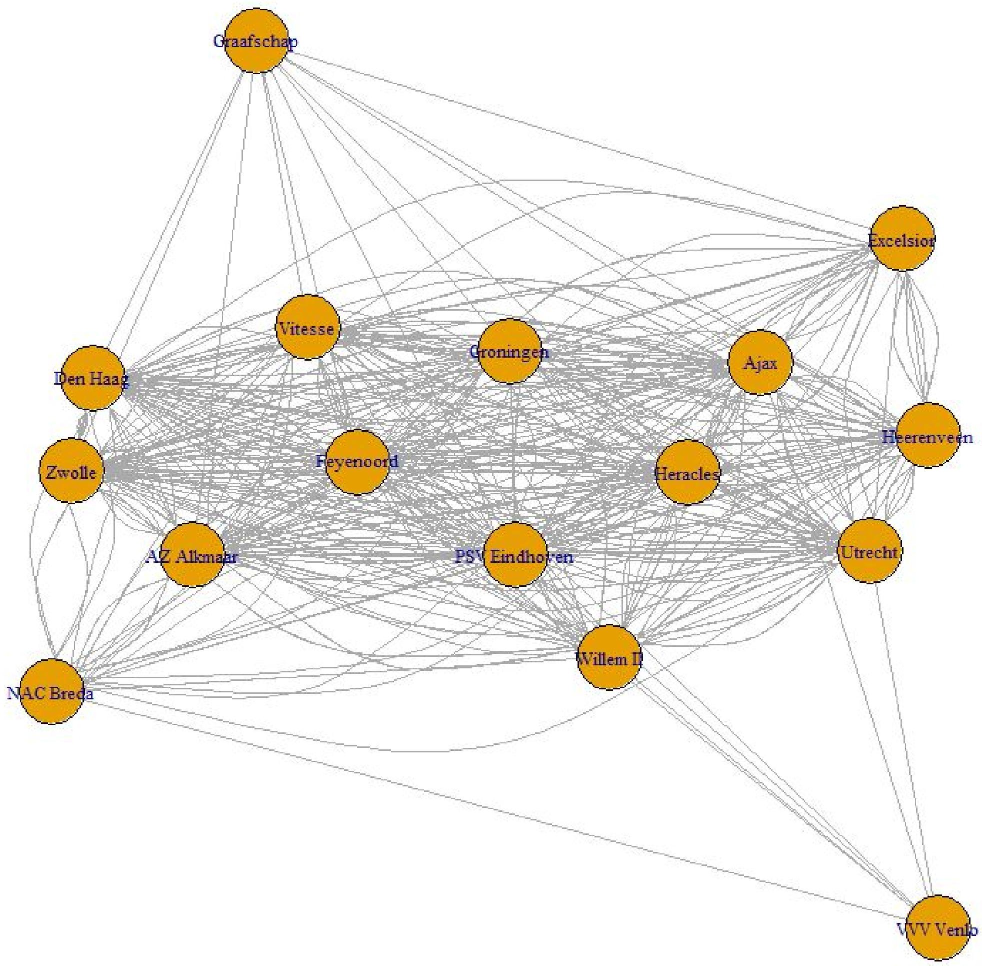


Figure 3.3: The representation of the data as a directed graph

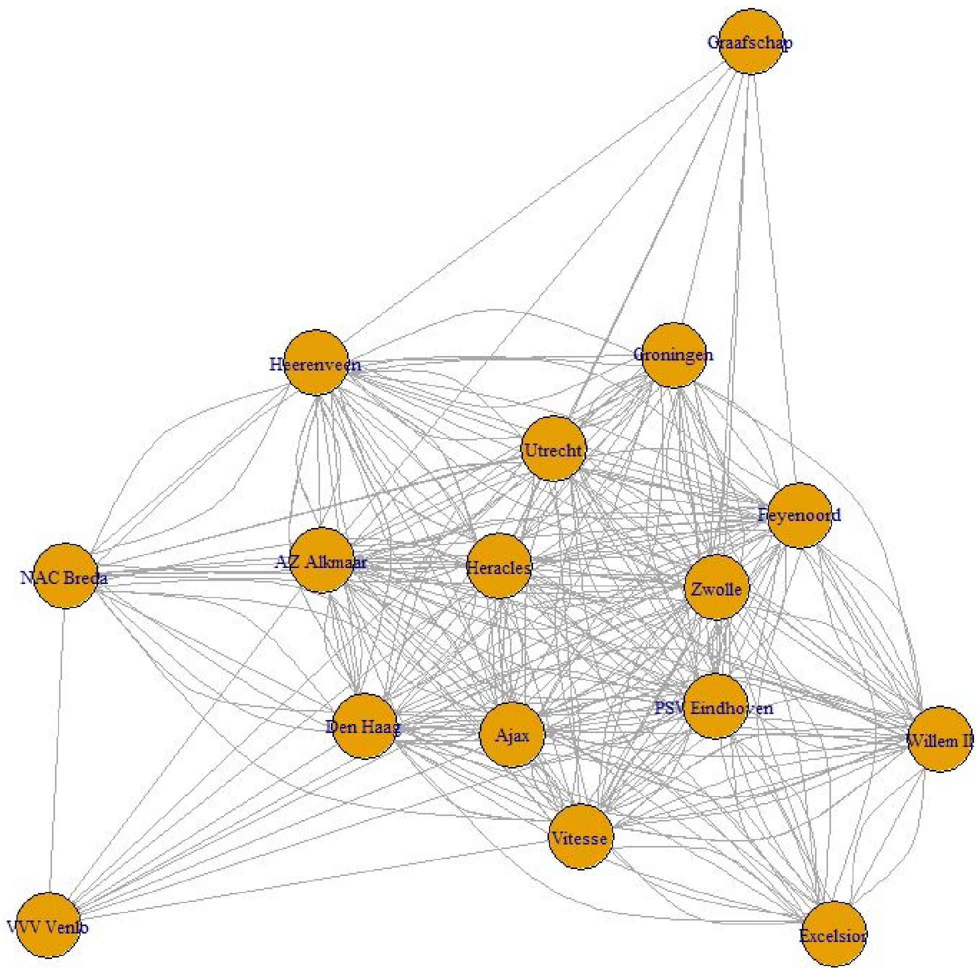


Figure 3.4: The representation of the data as a directed graph

3.2. Reliability of each Team

As discussed earlier in this chapter, not every team has the same amount of matches in the selected data. We've also discussed 'stability' of teams. Using 'deviation' we decided which data we wanted to use for predictions. But that deviation only said something about the stability of the 10 teams that played all 6 seasons combined. In this section we zoom in on each team to check whether the selected data would help us with good predictions for this team.

In Figure 3.5, the points of each team per season are shown. If a line is quite horizontal, this means that a team has had approximately the same amount of points in each season, thus being very stable. On the other hand, if a line goes up and down very steeply, this means that there are big differences in points gained per season.

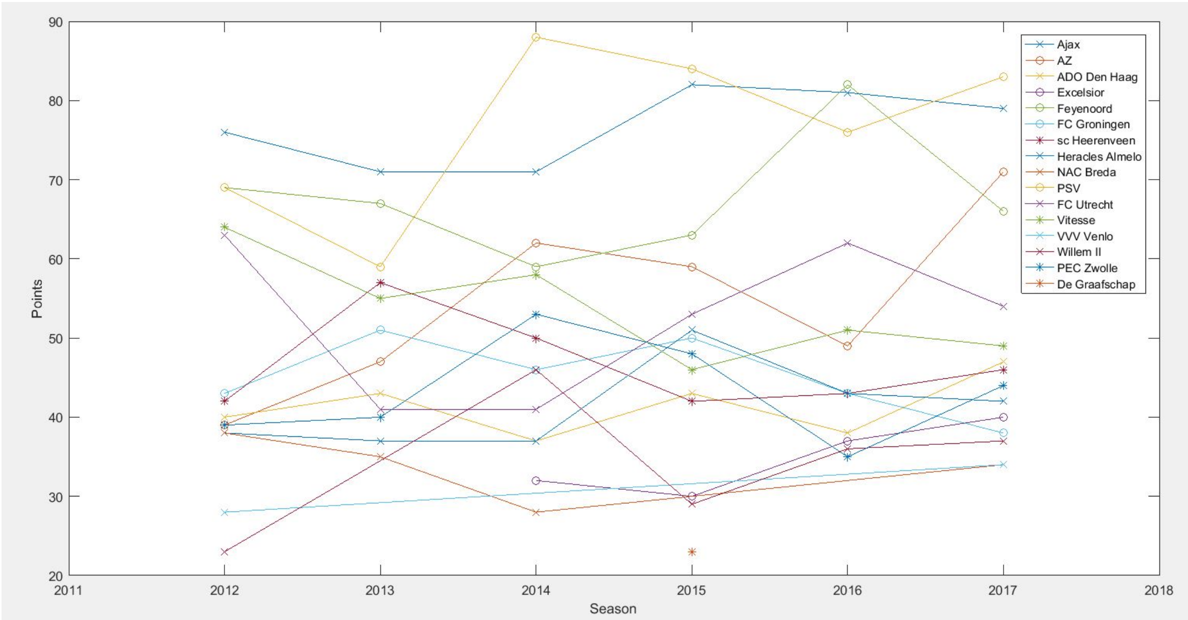


Figure 3.5: The points per season per team, the year specifies the start year of the season

The reason we also look at points and not only at rank is simple. Looking closely at Figure 3.5 shows that many of the teams get between 34 and 50 points per season. In season 2012/2013, or season 2012 in the figure, the team ranked 7th had 43 points, while the team ranked 15th had 37 points. In this season a difference of 6 points (which is really not a lot) makes a difference of 8 places on the ranking. In the season 2015/2016, there was one team with 43 points and it was ranked 11th. This shows that even though a team could have the same amount of points in different seasons, if we would only look at rank, we could judge this team as 'unstable', while this is clearly not the case. In order to prevent misinterpretation of ranking like this, we've calculated a new deviation, this time looking at the points. We did this by calculating the average points collected per team per season over the past six seasons and comparing this with the points collected in each season. We must also take into account the number of seasons we sum over. De Graafschap for example will have a deviation of 0, since it only played one season. We chose to calculate the deviation as some sort of mean squared error, meaning that we square the difference in points with the average and we average those differences. The obtained deviations can be found in Table 3.6.

Table 3.6 shows us that AZ has a deviation in points of 112.58. This deviation is very big, compared to teams as Ajax, FC Groningen and especially ADO Den Haag. A high deviation means that the differences in points over the past 6 seasons were high as well, while a low deviation means that the points were approximately the same over those seasons. This is also clearly visible in Figure 3.5, where it shows that ADO Den Haag is a very stable, almost horizontal line. Based on Table 3.6 and Figure 3.5 it is way easier to intuitively predict the number of points ADO Den Haag will get next season, than the number of points AZ will get.

This intuition should not be ignored, because if a team has been very stable over the past few years, it makes

Team (Number of seasons)	Deviation in Points
Ajax (6)	19.56
AZ (6)	112.58
ADO Den Haag (6)	11.56
Excelsior (4)	15.69
Feyenoord (6)	51.22
FC Groningen (6)	19.81
sc Heerenveen (6)	29.22
Heracles Almelo (6)	24.22
NAC Breda (4)	13.19
PSV ((6)	98.92
FC Utrecht (6)	77.89
Vitesse (6)	35.81
VVV Venlo (2)	9
Willem II (5)	60.56
PEC Zwolle (6)	35.81
De Graafschap (1)	NA

Table 3.6: Deviation in points

sense that their performance of the coming year would be comparable, while if a team's performance changes a lot over the years, it's more difficult to predict their performance for the next year.
Another team that has a very low deviation is VVV Venlo, however this is based on only 2 seasons, so we should ask ourselves how much that tells us. It is important to keep this table with deviations in the back of our head when doing predictions, because it can give us an indication of how trustworthy the predictions will be.

In some cases there are clear causes that can explain the certain change in performance of a team. These causes are more difficult to model, but may have great impacts.
Let's take a look at FC Twente. This team is not in the league anymore, since it just relegated from the Eredivisie. FC Twente was however present in all previous 6 seasons. Just to show the clear change in the performance of this team, we've plotted the number of points FC Twente got in the past 14 seasons in Figure 3.6.

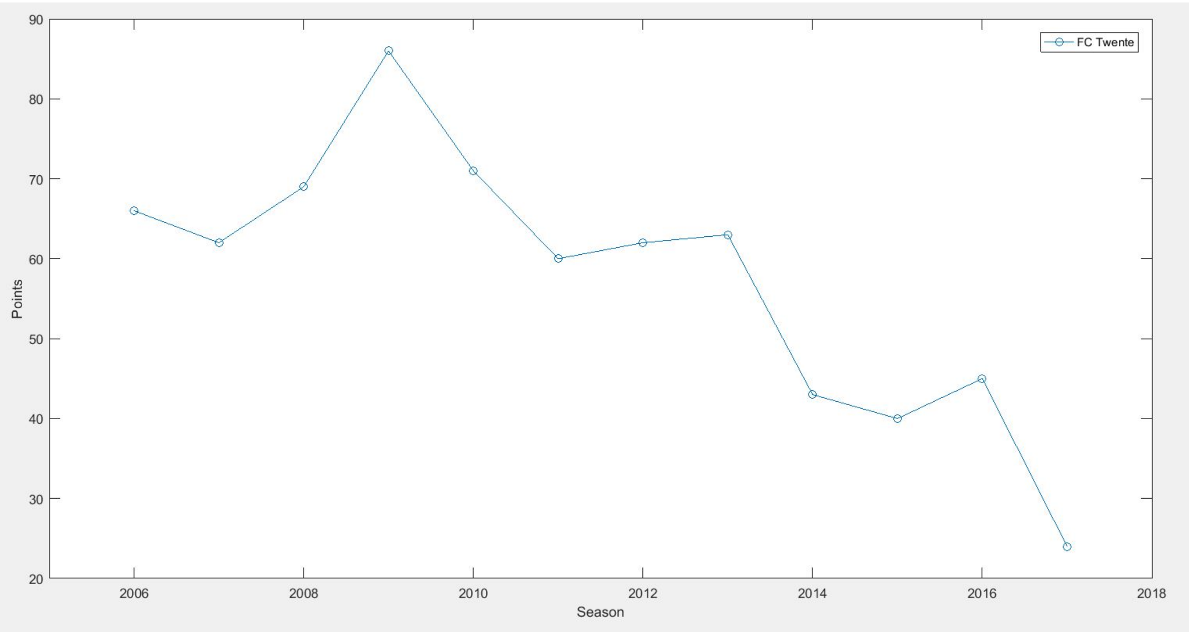


Figure 3.6: The points of FC Twente per season, 2012 is season 2012/2013, 2013 is 2013/2014 etc.

If we look closely at Figure 3.6, we see that FC Twente performed quite well until season 2013/2014. They even performed very well in 2009/2010 by gathering 86 points and becoming the champions of that season. Unfortunately from season 2014/2015 on their number of points decreased drastically. Where in 2013/2014 they still ended on the 3rd place with 63 points, the season after they only ended 10th with a meager 43 points. This became the new standard until their all time low in 2017/2018 where they only gathered 24 points to finish at the bottom of the table.
By just looking at the data of the results until 2013/2014 no model would have predicted this sudden decrease in points. After the championship of FC Twente, two of their board members, Van der Laan en Munsterman, who were also succesful businessmen got a taste for the succes. To make Twente structurally one of the best clubs of the Netherlands they took irresponsible financial risks. This eventually led the club to get in financial problems. Because of these problems, the Royal Dutch Football Association even deducts a total of 6 points of FC Twente in the season 2014/2015. With help of Dutch scout Ted van Leeuwen the club managed to survive the coming years, but when he left in October 2016, the club wasn't able to get a good enough selection for the next season which caused them to relegate in season 2017/2018 [9] [8].

What the story of FC Twente shows, is that there are a lot of different factors that determine the performance of a team. If the budget of a team changes relatively to other teams, the performance is very likely to change as well. Unfortunately, data about budget for players is not available from many clubs. It is clear however that in general, the larger the players budget, the better the performance of a team.

In this chapter we've decided which data we want to select to base predictions upon. Subsequently we checked whether this data satisfied the data assumptions that are necessary to apply the model and its extensions. Furthermore we looked at the stability of the different teams to get an idea how 'certain' a prediction can be. In the next chapter we'd like to get to know more about this uncertainty using simulations. Also the different models will be compared to see which one best fits the data.

4
Simulation Study

In this chapter the different models will be tested on simulated data. The main thing that will be tested is how well the model recovers winning potentials. This can be tested easily by simulating data with established winning potentials. We want to see how different factors impact this recovery. If we know how well a winning potential is recovered, we know how well the selected data represents the 'true' winning potential of a team. Furthermore different methods and extensions of the model will be compared, to see how they contribute to recovering the winning potentials. The chapter is divided in different sections. In each section we focus on another factor that might impact the recovery of the winning potentials.

4.1. Varying the number of matches

In this section we test the original Bradley-Terry model, so there will be no home advantage or draws. The first test will be varying in the number of matches. To test this we choose to work with three teams. The distance in the winning potentials between the teams will be 0.2. The winning potentials for all scenarios will thus be:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.2 \\ 1 \\ 0.8 \end{bmatrix} \tag{4.1}$$

The expectation is that the higher the number of matches will be, the better the recovery of the winning potentials. Each scenario will be simulated 1000 times, so that we get 1000 recoveries of each winning potential. The 1000 recoveries of winning potentials can be looked at as realizations of a random variable. Using the recoveries we can estimate the probability density function of this random variable.

4.1.1. Scenario 1

In the first scenario simulated, we choose $n_{ij} = 4$, for all i, j which means that every team plays 4 matches against each other.
Based on the 1000 simulations, we estimated the density functions of the winning potentials. We did this by fitting kernel density plots on the data, which are basically a smoothed Histograms. The estimated functions for this scenario can be found in Figure 4.1.
From these simulations, we found that the intervals for each winning potential such that 95% of the recoveries were in these intervals were as follows:

- $0.2667066 < \pi_1 < 2.329047$
- $0.1353166 < \pi_2 < 2.304991$
- $0.0014970 < \pi_3 < 1.799062$

These intervals can also be seen in Figure 4.1, where the 2.5 % quantile and the 97.5% quantile are clearly indicated. There are a few things that we need to address. First of all, because there were only 3 teams and they played so little matches, it occurred 20 times that the data assumption wasn't satisfied. In these 20 cases, one team lost every match, which caused the algorithm to recover a winning potential of '0' for this team.

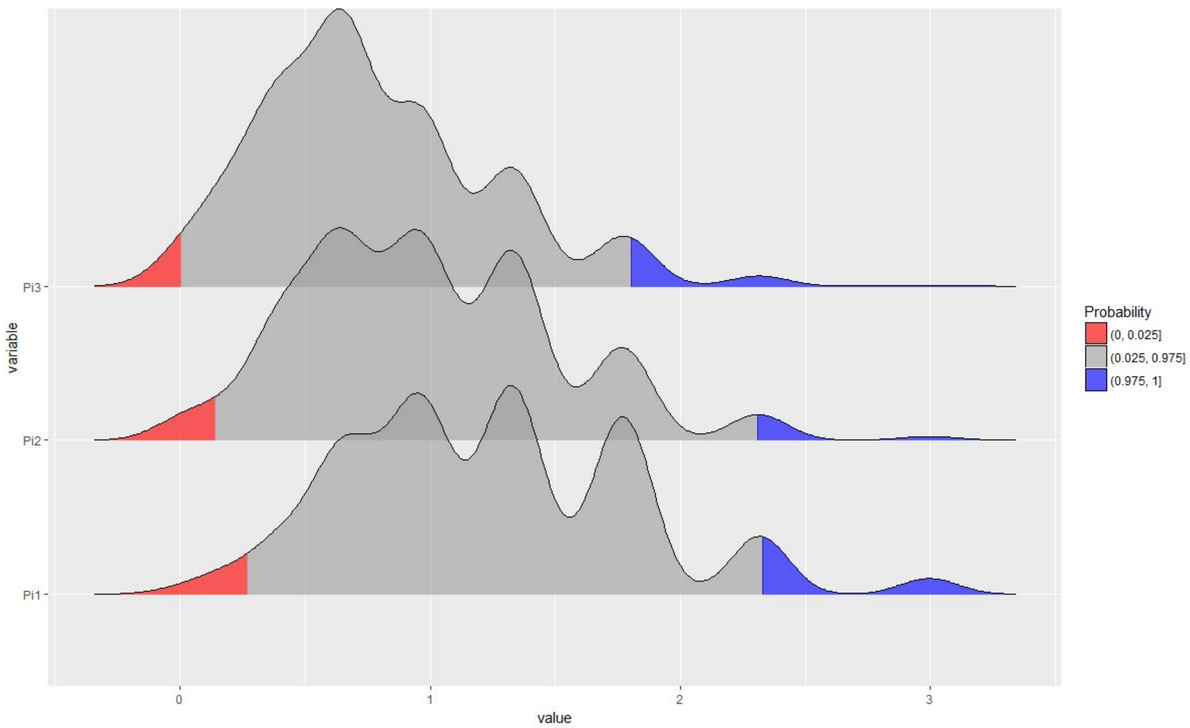


Figure 4.1: Estimates of density functions for Scenario 1.

This explains why the tails of the densities are in negative range of the axis.
Furthermore we see that the 95% intervals are very big, so the recoveries were very spread out. What we had hoped to get for the density estimation was a normal distribution, with the means equal to the winning potentials we used to simulate with. The winning potential that comes closest to such a distribution is π_3, although it's far from perfect.
What also stands out is that there is not even a clear difference in the distributions of the different teams. That has to do with the fact that the differences between the original winning potentials are not very big, but hopefully also with the number of matches played between the teams.

The size of the 95 % interval can be used as a measure of 'uncertainty'. By uncertainty we mean that, if this interval is very big, a prediction for the winning potential will be less accurate, thus more uncertain.

4.1.2. Scenario 2

Where in the Scenario 1, $n_{ij} = 4$ was used for all i, j, in this scenario we simulate with $n_{ij} = 12$. The results show that 95% of the retrieved winning potentials were in the following intervals:

- $0.6913259 < \pi_1 < 1.799062$
- $0.5111798 < \pi_2 < 1.625078$
- $0.3479295 < \pi_3 < 1.354870$

These intervals are already a lot smaller than the intervals we saw in Scenario 1, just as we expected.

4.1.3. Scenario 3

For the third and final scenario of this section, we've simulated the league with $n_{ij} = 40$. The 95% intervals then are:

- $0.8910294 < \pi_1 < 1.524884$
- $0.7183569 < \pi_2 < 1.313815$

- $0.5509996 < \pi_3 < 1.067110$

These intervals are significantly smaller than the intervals in the two previous scenarios, again showing that the number of matches plays a big role in the recovery of the winning potentials. The difference in this recovery is clearly demonstrated in Figure 4.2, where we compare the estimated densities of Scenario 2 and 3.

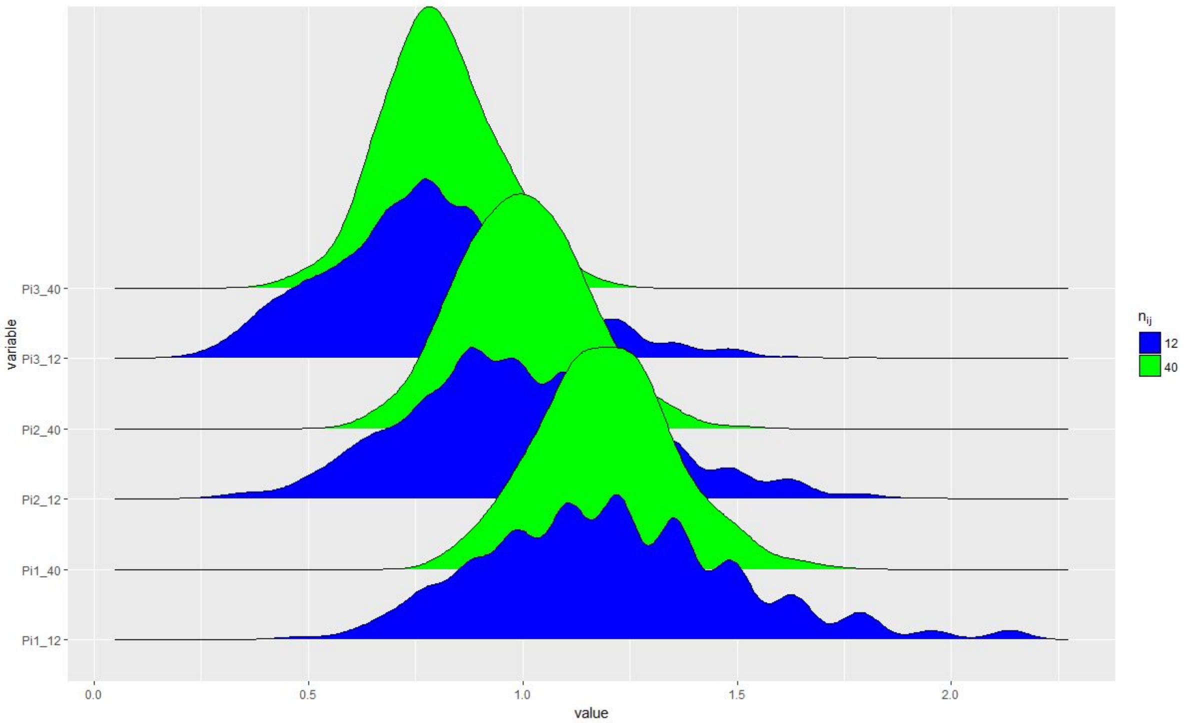


Figure 4.2: Comparison of estimated densities of the winning potentials, between Scenario 2 and Scenario 3.

What also stands out is the shape of the distribution. In general we see that the distribution of Scenario 3, where $n_{ij} = 40$ looks quite normally distributed, especially compared to the distributions of Scenario 2. Something else that catches the eye is that for both scenarios the recovery of π_3 seems to be the most accurate one. This can be explained by looking at probabilities p_{12} and p_{23}. Namely, $p_{12} = \frac{1.2}{1.2+1.0} \approx 0.5454$, where as $p_{23} \approx 0.5555$. This tells us that the probability of Team 2 beating Team 3 is greater than the probability of Team 1 beating Team 2, which suggests that Team 1 loses more often to Team 2, than Team 2 loses to Team 3. Hence the chance that a recovered π_2 is greater than a recovered π_1 is bigger than the chance that a recovered π_3 is greater than a recovered π_2. This explains why the recovery of Team 3 is generally better than the recovery of Team 1 and Team 2.

Based on the simulations in this section, it is clear what the effect of the number of matches n_{ij} is on the recovery of winning potentials. A higher number of matches contributes to a higher accuracy of the recoveries. We can conclude that an increase of the number of matches played between each team, contributes to a decrease in the uncertainty we earlier mentioned.

Now that we know the effect of n_{ij}, we go on to test another factor. In the scenarios we simulated in this section, 3 teams were used, while in reality we want to do predictions about 16 teams. What is also interesting about the number of teams is the number of matches that correspond with those teams. If there are 16 teams, and $n_{ij} = 4$ for all i, j, each team plays 60 matches, while in the scenario with 3 teams, each team only plays 8. In the next section we want to test whether this affects the recovery in a positive way, or whether the impact of having more teams influences the recovery in a negative way.

.

4.2. Varying the number of teams

To test the effect of the number of teams m, we choose to use the same n_{ij} for all i, j in each scenario. This means that we can also use Scenario 2 from section 4.1 to compare results with. The distance between the winning potentials used to simulate with, will also be kept the same for all scenarios. Just as in section 4.1 we choose this distance to be 0.2. To choose the starting winning potentials, while keeping this distance the same we use: $\pi_m + \pi_m + d + \pi_m + 2d + \cdots + \pi_m + (m-1)d = m$, with π_m being the lowest winning potential. Solving this we then find the other winning potentials by adding either $d, 2d, 3d, \ldots, (m-1)d$. Every scenario will again be simulated 1000 times, in order to make statements about it.

4.2.1. Scenario 1

As first scenario we pick Scenario 2 from the section 4.1. In this scenario we used $m = 3$ with the winning potentials:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.2 \\ 1 \\ 0.8 \end{bmatrix} \tag{4.2}$$

In this scenario every team plays a total of 24 matches. The estimated densities of the winning potentials are shown in blue in Figure 4.2.

4.2.2. Scenario 2

In this scenario we pick $m = 6$. Solving $6\pi_6 + 15d = 6$ we find $\pi_6 = 0.5$. This means that we simulate with the following winning potentials:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.5 \\ 1.3 \\ 1.1 \\ 0.9 \\ 0.7 \\ 0.5 \end{bmatrix} \tag{4.3}$$

Note that in this scenario every team plays 60 matches. To be able to compare the results with the results of Scenario 1, we plotted both the recoveries of the winning potentials of this scenario as well as the recoveries of scenario 1 in one Figure 4.3. Fitting boxplots on the data helps in comparing the recoveries of the scenarios.

What Figure 4.3 illustrates is that the differences in spread are not that great. From inspection at least this isn't visible. To check if there is a significant difference, we made an overview of the interquartile ranges or "H-spread" of the different winning potentials as a measure of uncertainty. These values can be found in Table 4.1. The interquartile range is the difference between the upper quartile and the lower quartile, or the length of the box as shown in Figure 4.3.

Winning Potential	Scenario	Original Value	Median	Interquartile range
π_1	2	1.5	1.482456	0.4388485
π_2	2	1.3	1.280515	0.3630921
π_1	1	1.2	1.199986	0.3661554
π_3	2	1.1	1.063565	0.2856789
π_2	1	1.0	0.983476	0.3984525
π_4	2	0.9	0.893923	0.2921043
π_3	1	0.8	0.785256	0.2527701
π_5	2	0.7	0.682874	0.2151648
π_6	2	0.5	0.481330	0.1633035

Table 4.1: Details about the recoveries displayed in Figure 4.3

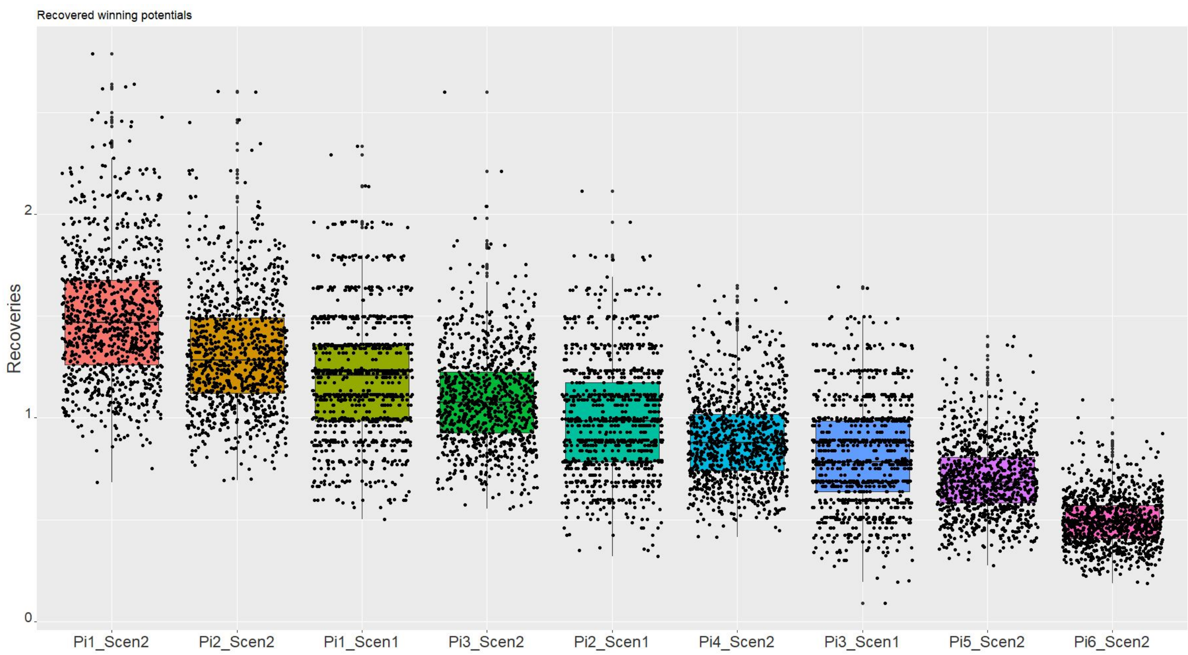


Figure 4.3: Comparison of recovered winning potentials, between Scenario 1 and Scenario 2.

From the information in 4.1, we can conclude a few things. First of all, as already discussed before, we observe a trend that says that the lower the winning potential, the lower the spread. Again this can be explained by observing that $p_{12} < p_{23} < p_{34} \cdots < p_{56}$.
Only one value of the interquartile ranges, namely the value of Pi2_Scen1 (Pi2_Scen1 represents the recoveries of π_2 from Scenario 1), actually supports the claim that says that the spread of Scenario 1 is greater than the spread of Scenario 2. The other two values don't really stand out compared to the values of Scenario 2. Therefore we shouldn't jump to a conclusion, based on these two scenarios alone.
Something that does stand out however, is which values are recovered. For the winning potentials of Scenario 1, it is clear that there is only a small number of values recovered for each winning potential. Notice that the recoveries of those winning potentials lay on lines. This can however easily be explained, because since there are only 3 teams in this scenario, there are only a relatively small number of outcomes of a league possible. Hence the winning potentials will also only have a relatively small number of recoveries.

4.2.3. Scenario 3

For the final scenario in this section, we simulate with $m = 9$. This gives us the winning potentials:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.8 \\ 1.6 \\ 1.4 \\ 1.2 \\ 1.0 \\ 0.8 \\ 0.6 \\ 0.4 \\ 0.2 \end{bmatrix} \tag{4.4}$$

The obtained interquartile ranges for the recoveries of the winning potentials are respectively:

$$\begin{bmatrix} 0.45792364 \\ 0.40826634 \\ 0.34736552 \\ 0.33298261 \\ 0.26477971 \\ 0.21298628 \\ 0.17119502 \\ 0.12171170 \\ 0.07746351 \end{bmatrix} \tag{4.5}$$

Again, comparing this to Scenario 1, this data doesn't support the claim that more teams, means lower spread in recovery.

4.2.4. Scenario 4

So far we couldn't figure out what exactly the influence is of the number of teams on the recovery of the winning potentials. This also has to do with the trend that the more teams we use, the higher the spread gets for the high winning potentials and the lower the spread gets for the low winning potentials. This makes it difficult to compare the different scenarios in this section. For this reason, in this scenario we compare simulations from two leagues, that both consist of teams with only the same winning potential. League 1 consists of the 4 teams with winning potentials:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \end{bmatrix} \tag{4.6}$$

While league 2 consists of 8 teams with winning potentials:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \\ 1.0 \end{bmatrix} \tag{4.7}$$

The retrieved winning potentials can be found in Figure 4.4.
While we were not able to find clear confirmation that the number of teams had an impact on the recovery before, Figure 4.4 shows that there might be some influence. The first four boxplots of the Figure show the recovery of the winning potentials from the league that consisted of 4 teams, while the last 8 represent the recovery of the winning potentials from the league with 8 teams. What the boxplots already show is confirmed by Table 4.2, which displays the interquartile ranges as a measure of spread. Finally using the simulations of this scenario, we may conclude that there is a positive influence of the number of teams on the recovery of the winning potentials. In other words: an increase of the number of teams might increase the accuracy of recoveries.

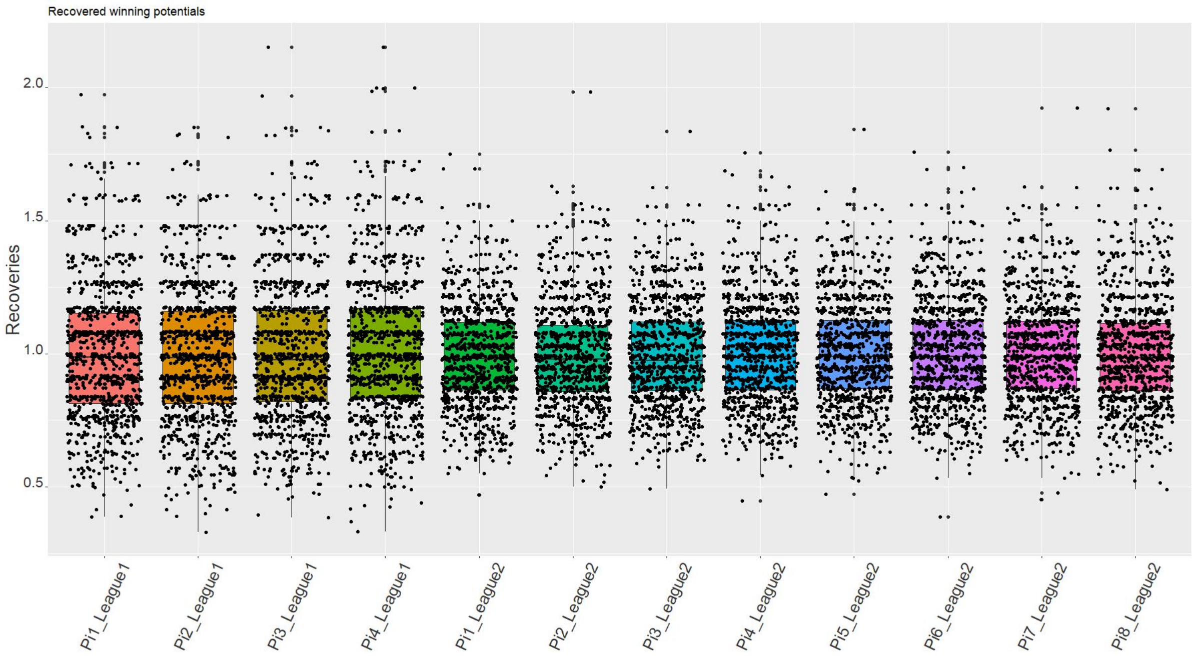


Figure 4.4: The recovered winning potentials of league 1 and league 2

Name in Figure 4.4	Interquartile range
Pi1_League1	0.3403439
Pi2_League1	0.3410978
Pi3_League1	0.3451674
Pi4_League1	0.3530046
Pi1_League2	0.2450154
Pi2_League2	0.2652737
Pi3_League2	0.2539361
Pi4_League2	0.2614527
Pi5_League2	0.2576141
Pi6_League2	0.2592142
Pi7_League2	0.2552210
Pi8_League2	0.2582007

Table 4.2: Details about the recoveries displayed in Figure 4.4

4.3. Simulating with Draws

In this section we compare several ways to deal with draws in the data. Using the extension introduced in Section 2.2, we simulated data. This means that draws are also included in that data. To see whether it's really necessary to use the extension in order to recover the original winning potentials in a good way we apply different models and compare the results. The first model we will use is the original model. This model can be applied on this data by simply ignoring the draws. After all, if the model in this way still recovers the winning potentials very well, there is no point in using the extension. We will refer to this model as 'Method 1'. For 'Method 2' we will again apply the original model, but this time we will not ignore the draws, but count them as half a win for both teams. Finally the extension for draws will be applied. We refer to t his as 'Method 3'.

The expectation of course is that both methods of the original model will recover the winning potentials significantly worse than the extension including draws. For all scenarios in this section $n_{ij} = 12$, for all i and j, and $m = 6$ will be used. For these simulations we again choose the constant distance of 0.2 between the winning potentials which means that the vector of winning potentials we will simulate with is:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.5 \\ 1.3 \\ 1.1 \\ 0.9 \\ 0.7 \\ 0.5 \end{bmatrix} \tag{4.8}$$

The parameter θ will be defined differently for the different scenarios.

4.3.1. Scenario 1

In this scenario we want to test the different models on data that contains many draws. Hence we choose $\theta = 5.0$ for the simulations in this scenario. As always we simulate the scenario 1000 times. In every scenario a total of 180 matches were played. The average number of draws with this θ was 114.6. In Table 4.3, for each winning potential we've calculated the average retrieved value for each method.

Original Winning Potential	1.5	1.3	1.1	0.9	0.7	0.5
Method 1	1.8877	1.4797	1.0869	0.7660	0.4938	0.2857
Method 2	1.2805	1.1839	1.0701	0.9516	0.8302	0.6835
Method 3	1.4980	1.3105	1.0984	0.8917	0.7005	0.5008

Table 4.3: The average retrieved winning potentials for each method in Scenario 1.

The 95% interval for the recoveries of θ is: $3.96351 < \theta < 6.715706$. As expected, Table 4.3 shows that the recoveries of Method 3 are way better than the recoveries of the other methods in this scenario. The average recoveries using Method 3 are very close to the original value. It makes sense that this method is the best one, independent of the number of draws, because when there are no draws, this method will just be the same as Method 1.

4.3.2. Scenario 2

In this scenario we want to check whether the performances of Method 1 and Method 2 become better when there are fewer draws. For this scenario we've therefore chosen $\theta = 1.1$. The expectation is that Methods 1 and 2 will perform better in this scenario than in Scenario 1, however we still expect Method 3 perform better than the other methods.

The average number of draws in the 1000 simulations of this scenario is 8. In Table 4.4, the average recoveries per method are shown. The 95% interval for θ is $1.037457 < \theta < 1.18521$. It is clear from 4.4 that the recoveries

Original Winning Potential	1.5	1.3	1.1	0.9	0.7	0.5
Method 1	1.5372	1.3221	1.1105	0.8777	0.6771	0.4751
Method 2	1.5106	1.3076	1.1088	0.8861	0.6926	0.4941
Method 3	1.5116	1.3078	1.1091	0.8856	0.6923	0.4933

Table 4.4: The average retrieved winning potentials for each method in Scenario 2.

of Method 1 and Method 2 are significantly better than their recoveries in Scenario 1. What surprises is that Method 2 even seems to perform better on this data than Method 3. Counting each draw as half a win might be a method worth considering, looking at these results. In Figure 4.5 the recoveries of both Method 2 and Method 3 are plotted. This Figure shows that also the spread for both methods is very similar.

In Figure 4.6 we plotted the densitiy estimates of the recoveries of Method 1 both from Scenario 1 and Scenario 2. This plot clearly shows the difference in performance on the different data sets, both in spread as well as in getting the right mean.

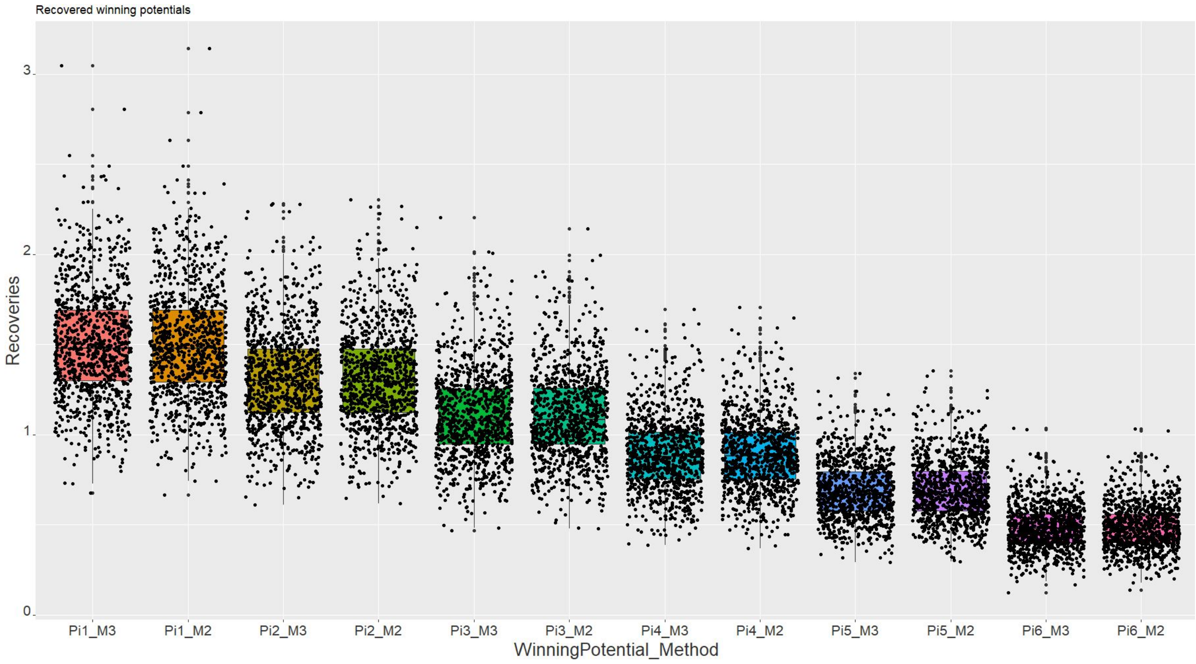


Figure 4.5: The recoveries of winning potentials, of Method 2 and Method 3 for Scenario 2.

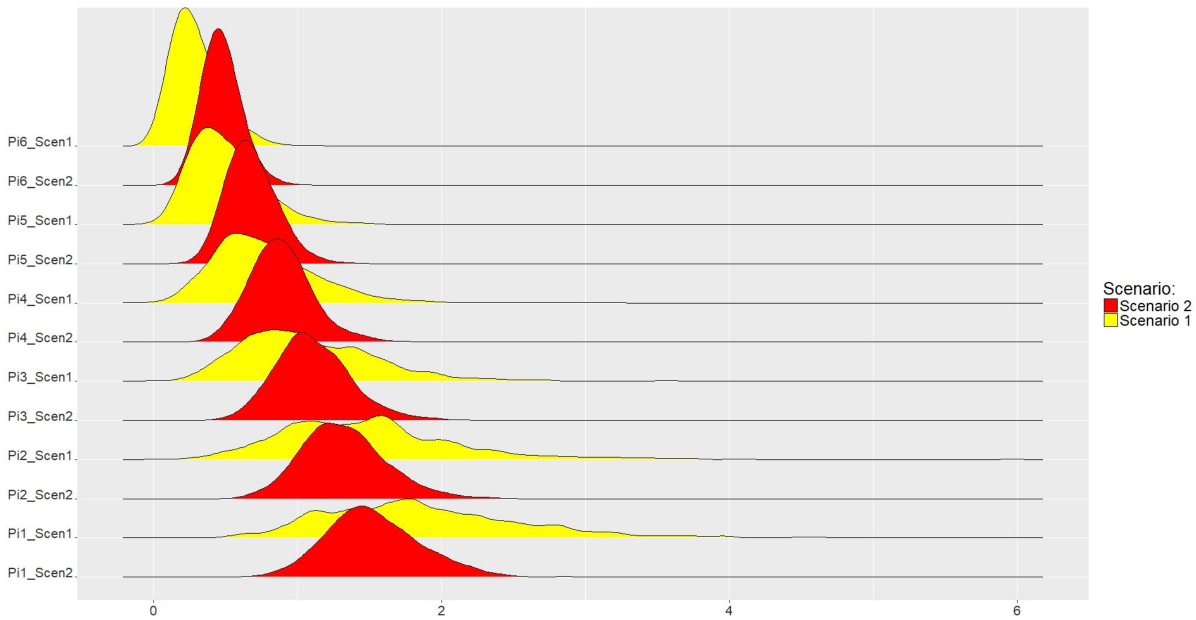


Figure 4.6: The density estimates of Method 1, for Scenario 1 and Scenario 2

4.3.3. Scenario 3

So far we've discovered that if there are lot of draws, like in Scenario 1, Method 3 is clearly the best method. In Scenario 2 however we've discovered that if there are only a few draws, Method 2 seems to perform just as well or even better than Method 3. To decide which Method is the best to use on our real data, we should simulate data that looks like our real data. As mentioned in Section 3.1 about 25% of the matches of our real data, ended in a draw. For the 180 matches we simulate here, this means that about 45 of them should end in a draw. By simulating this data and applying both Method 2 and Method 3, hopefully we can decide which method recovers the winning potentials best.

Using $\theta = 1.7$ we had an average of 43.576 draws per simulation. This is almost 25%, so we chose to apply

Methods 2 and 3 on this data. In Table 4.5, the average recoveries are displayed. The table again shows that

Original Winning Potential	1.5	1.3	1.1	0.9	0.7	0.5
Method 2	1.4842	1.2812	1.0937	0.9093	0.7172	0.5141
Method 3	1.5220	1.2980	1.0955	0.8986	0.6974	0.4883

Table 4.5: The average retrieved winning potentials for Method 2 and Method 3 in Scenario 3.

both methods do a good job of recovering the winning potentials. Method 3 seems to perform slightly better in this scenario, but not significantly. We can conclude that both options are worth considering. However the performance of Method 2 seems to be correlated with the number of draws in the data, while Method 3 performs well independently of the number of draws. Another 'plus' for Method 3 is that this method also estimates the θ. This method therefore tells us something about the expected number of draws for a league, contrary to Method 2.

4.4. Simulating with Home Advantage

Just as we've tested in the previous section whether the extension to include draws is better than the original model, in this section we want to test whether the extension that includes Home Advantage performs better than the original model. This means that we only compare 2 methods in this section. We will refer to the original model as 'Method 1' and to the home advantage extension as 'Method 2'. The setup for the scenarios in this section is approximately the same as in Section 4.3. For the simulations we use $m = 6$ with the respective winning potentials being:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.5 \\ 1.3 \\ 1.1 \\ 0.9 \\ 0.7 \\ 0.5 \end{bmatrix} \tag{4.9}$$

Furthermore we use $n_{ij} = 12$ again. In the simulations of this section however we simulate the data with the extension for Home Advantage from Section 2.3. This means that in the data there will be no draws, but only wins and losses. In the different scenarios the home advantage parameter γ is being tweaked to get different sorts of data.

4.4.1. Scenario 1

In the first scenario we would like to compare the two models, while choosing γ to be large, namely $\gamma = 5.0$. This means that there will be a significant home advantage in the data. The question is whether this influences the performance of the original model or whether this model performs equally well. The scenario, as always, is simulated 1000 times. From the simulations we find that out of the 180 wins, on average 147.23 were at home in this data.

The average recoveries of both methods can be found in Table 4.6. Table 4.6 shows that for this scenario

Original Winning Potential	1.5	1.3	1.1	0.9	0.7	0.5
Method 1	1.2840	1.1796	1.0774	0.9594	0.8251	0.6743
Method 2	1.5091	1.3037	1.1077	0.9044	0.6882	0.4866

Table 4.6: The average retrieved winning potentials for both methods in Scenario 1.

Method 2 clearly outperforms Method 1 on this simulated data. Since Method 1 ignores home advantage this was to be expected.

To show that indeed, the model used in Method 2 is preferred on this data, we use something called the Aikake Information Criterion, or AIC for short. The AIC value which we use to compare different models on a given dataset. The value can be calculated by $2k - 2\log(L)$, where k represents the number of parameters and L the likelihood. The model that has the lowest AIC value is preferred. In this Scenario the model used in Method 1 uses $k = 6$ parameters, while Method 2 uses $k = 7$. For every simulation we calculated the likelihood for both methods and thus we could calculate the AIC value of both models for every simulation. It turns out that for this scenario in every one of the 1000 simulations the AIC value for Method 2 is lower, which means that for this data the model that includes home advantage is preferred.

In this scenario however we have chosen a γ such that more than 80% of the wins took place at home. This is not very common in real data. Therefore we should also compare the methods on more realistic data.

4.4.2. Scenario 2

In this scenario we want to test both methods on data that is comparable with our real data. As mentioned in Section 3.1, in our data about 60% of the wins were booked at home. In this scenario we want to simulate a similar result of home wins.

Using $\gamma = 1.6$ we simulated data in which the average number of home wins was 109.49 out of 180. This is about 60%, so this data is suitable to apply the models on.

We expected the original model to perform better on this data, than on the data of Scenario 4.4.1. In Figure 4.7 we've visualized the recoveries of both methods. Table 4.7 displays the average values.

Original Winning Potential	1.5	1.3	1.1	0.9	0.7	0.5
Method 1	1.4793	1.2763	1.1197	0.9089	0.7072	0.5084
Method 2	1.5093	1.2908	1.1236	0.9002	0.6898	0.4861

Table 4.7: The average retrieved winning potentials for both methods in Scenario 2.

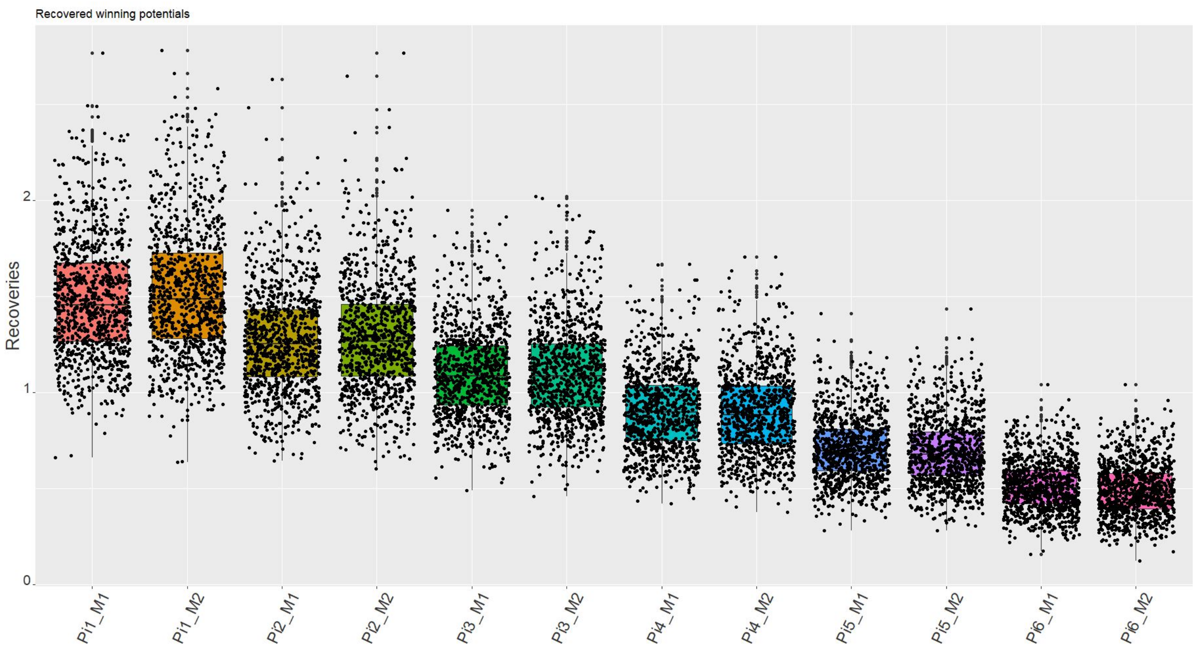


Figure 4.7: The recoveries of Methods 1 and 2, for Scenario 2.

The original model definitely performs better in this scenario. For the simulated data of this scenario, the recovered values with Method 1 actually don't differ that much from the values recovered with Method 2. This suggests that for a data set that contains only wins, and has no big home advantage, the original model could be used just as well as the extension for home advantage. By again checking the AIC values for both methods on this data we can determine which model is preferred. According to the AIC values, out of the

1000 simulations, 942 times the model that includes home advantage had a lower AIC value, which means that only 58 times, Method 1 was preferred.

4.5. Simulating with Home Advantage and Draws

In this section we test different aspects of the extension that includes both home advantage and draws. This extension is introduced in Section 2.4. Since this extension of the original model is merely a combination of the extension that includes draws and the extension that includes home advantage, we assume that all findings we've done in this chapter also apply to this model. In this section we not only want to focus on the recovery of the winning potentials, but also on some properties that the winning potentials have. One thing we want to look at, is the correlation between the winning potentials. Something else we want to check is the ranking that follows from the obtained winning potentials. In all scenario's we will use $m = 6$ and $n_{ij} = 12$. Based on Section 4.3 and Section 4.4 we will use $\theta = 1.7$ and $\gamma = 1.6$, since we found that these parameters cause realistic data. The winning potentials will vary per scenario.

4.5.1. Scenario 1

For the first scenario, we've chosen the 'basic' winning potentials we've often used in this chapter to simulate with:

$$\boldsymbol{\pi} = \begin{bmatrix} 1.5 \\ 1.3 \\ 1.1 \\ 0.9 \\ 0.7 \\ 0.5 \end{bmatrix} \tag{4.10}$$

Figure 4.8 shows us the density estimates of the recovered winning potentials.

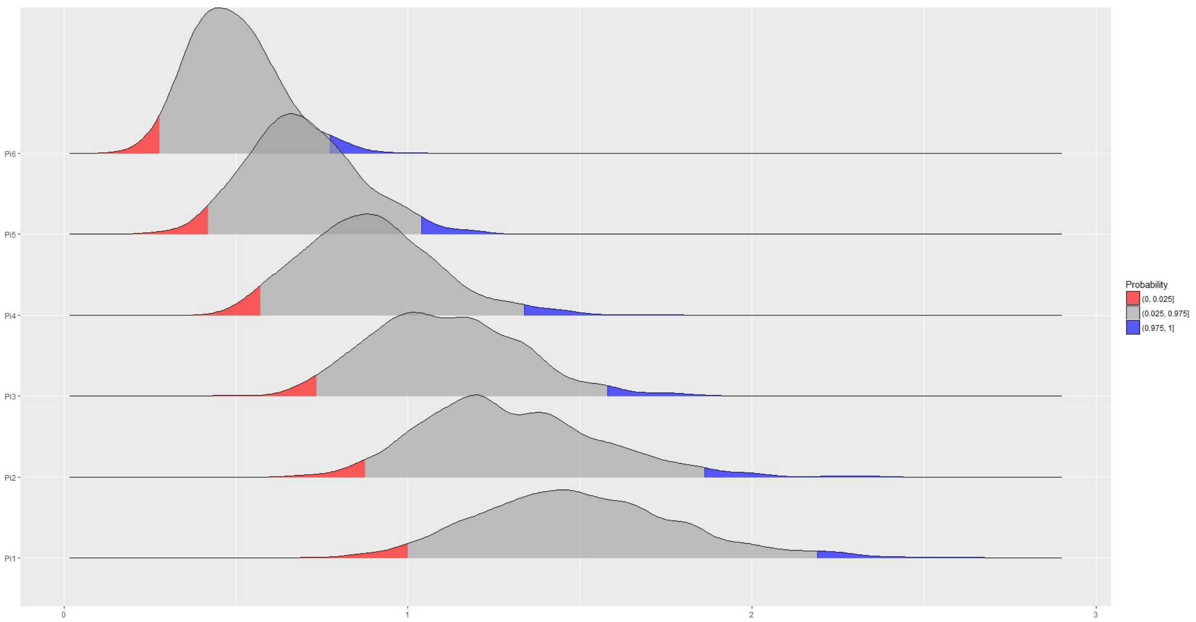


Figure 4.8: The density estimates of the recovered values for the winning potentials in Scenario 1.

As you can see in this figure, the densities often overlap each other. On places where two densities overlap, both teams have recovered winning potentials. In Figure 4.8 for example if we look at the recoveries of π_1 and π_2, since they overlap there is a chance that π_2 had a higher recovery than π_1 in a certain simulation. This means that even though they π_1's original winning potential was higher, based on that particular simulation π_2 would be ranked higher than π_1. If there is an overlap between the densities of two winning potentials, there is a chance that there rank switches for a certain simulation. Looking at Figure 4.8 there even is a chance that π_1 and π_5 switch rank in one or more simulations, since their densities have overlap.

To determine the probability of π_1 and π_5 switching rank is more complicated however. This has to do with the fact that the winning potentials are not independent of each other. There is a negative correlation between the winning potentials. This has to do with the constraint introduced in Section 2.1, that says that $\sum_i^m \pi_i = m$. In the correlation matrix (4.11) the correlations between the winning potentials, based on the simulated data, are displayed.

$$\begin{array}{c} \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \\ \pi_6 \end{array} \begin{array}{c} \begin{array}{cccccc} \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 & \pi_6 \end{array} \\ \begin{bmatrix} & & & & & \\ -0.3432 & & & & & \\ -0.3352 & -0.2609 & & & & \\ -0.3166 & -0.2171 & -0.1628 & & & \\ -0.2906 & -0.2136 & -0.0571 & -0.0594 & & \\ -0.2011 & -0.1753 & -0.1008 & -0.0001 & 0.0229 & \end{bmatrix} \end{array} \tag{4.11}$$

For every simulation we could rank the teams, due to their obtained winning potential. These rankings could help us with doing predictions about the rank of a team for next season.
Let's refer to the winning potentials we used to simulate the data with as the 'original winning potentials'. If we would rank these winning potentials, $\pi_1 = 1.5$ would have rank 1, while $\pi_6 = 0.5$ would have rank 6. In Figure 4.9, we've displayed the number of rankings the each winning potential got, out of the 1000 simulations. π_1 that has original rank 1, has been ranked number 1, 603 times for example.

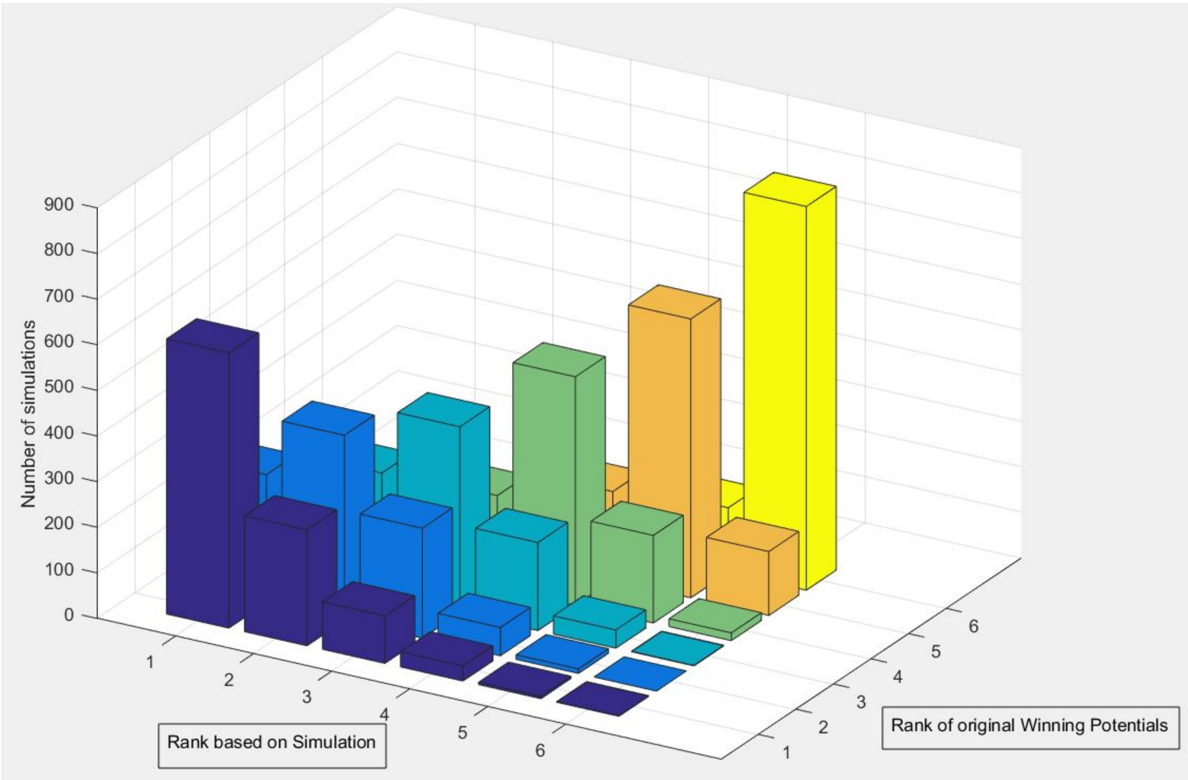


Figure 4.9: The number of times each rank was obtained for each winning potential.

What stands out is that the rank of π_1, and π_6 are recovered as their original rank relatively often, while π_3 has the lowest number of recoveries as rank 3. If the team's original rank is recovered by a simulation we call this a 'true recovery'. There seems to be a correlation between the true recoveries of a winning potential and the overlap of its density, which can be found in Figure 4.8. We see that π_6 has the smallest overlap with other winning potentials and that it has almost no overlap with π_1, π_2 and π_3. Consequently we find that π_6 had zero recoveries with rank 1, 2 or 3 and 841 with rank 6. π_3 on the other hand only had 408 true recoveries. The fact that the teams with the most overlay have the most variation in recovered ranking might be important for our predictions.

4.5.2. Scenario 2

In this scenario we experiment with different winning potentials. Opposed to the previous sections in which mostly a constant of 0.2 was used between the winning potentials, here we want to try different distances between the winning potentials of the teams.

In practice it often happens that there are a few top teams, which are significantly better than the other teams. Therefore in this scenario we want two 'top teams' that have very small probabilities of losing to the other teams. We're also curious to see if the correlations are similar to those found in Scenario 1. In order to simulate a scenario as described we've chosen the following winning potentials to simulate with:

$$\boldsymbol{\pi} = \begin{bmatrix} 2.5 \\ 2.3 \\ 0.34 \\ 0.31 \\ 0.28 \\ 0.27 \end{bmatrix} \tag{4.12}$$

We would expect the two top teams to be ranked either first or second, almost all the time based on the recovered winning potentials. The expectation is also that the densities of the other 4 winning potentials will significantly overlap. Based on the previous scenario this would also mean that the recovered rankings will vary strongly as well.

The obtained density estimates can be found in Figure 4.10.

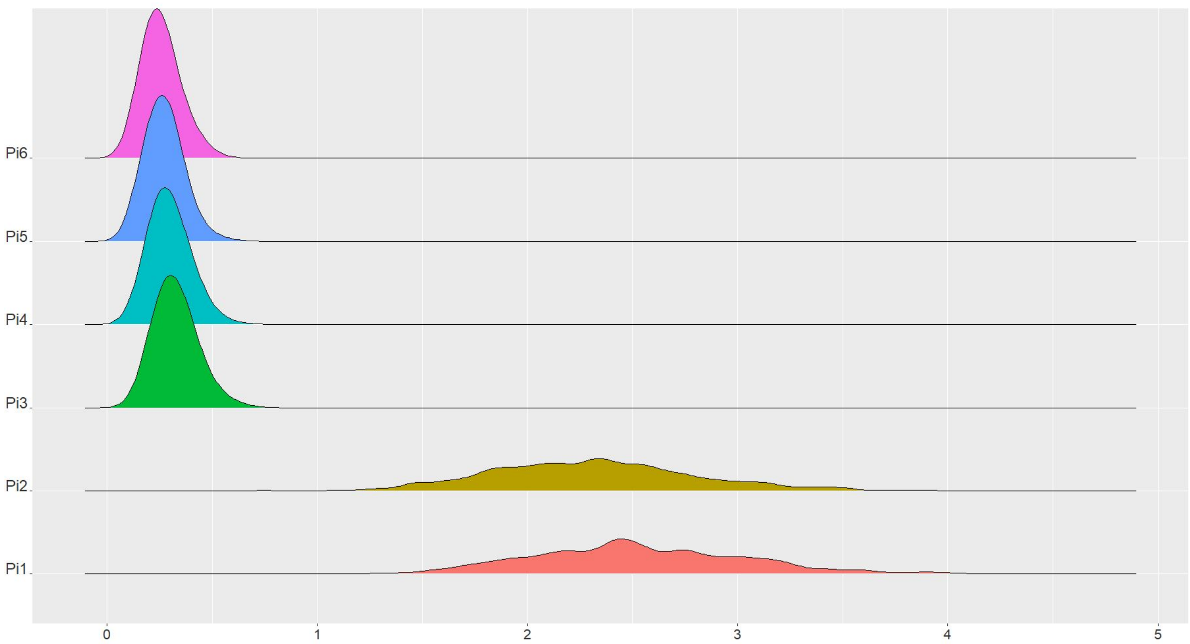


Figure 4.10: The density estimates of the recovered winning potentials of Scenario 1

Figure 4.10 shows what we were already expecting, a lot of overlap between the two top teams, as well as between the lower 4 teams. There seems to be no overlap between a top team and a lower team however. This also means that no top team is ranked 3rd or lower. Even though the difference in winning potentials between the two top teams and the four other teams is significant, the algorithm converged every single time. This means that the data assumption has been satisfied for every simulation.
What stands out as well is the 'range' of the winning potentials. The range of the recoveries of the two top

teams is way bigger than the range of the smaller teams.

The correlations between the winning potentials in this scenario differ greatly from the correlations of the previous scenario. They can be found in the correlation matrix (4.13).

$$
\begin{array}{c}
\\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \\ \pi_6
\end{array}
\begin{array}{c}
\begin{array}{cccccc} \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 & \pi_6 \end{array} \\
\begin{bmatrix}
 & & & & & \\
-0.8506 & & & & & \\
-0.2726 & -0.1369 & & & & \\
-0.2872 & -0.1056 & 0.3417 & & & \\
-0.2829 & -0.1155 & 0.4057 & 0.3758 & & \\
-0.2647 & -0.1124 & 0.3568 & 0.3705 & 0.3887 &
\end{bmatrix}
\end{array}
\tag{4.13}
$$

We see that π_1 and π_2 are strongly negatively correlated. This means that if π_1 gets bigger, π_2 gets smaller. All the lower teams surprisingly are positively correlated. Since each one of them has a negative correlation with the top teams, this means that if a top team has a smaller recovery, all the lower teams get a higher one in general.

The recovered rankings in this scenario can be found in Figure 4.11.

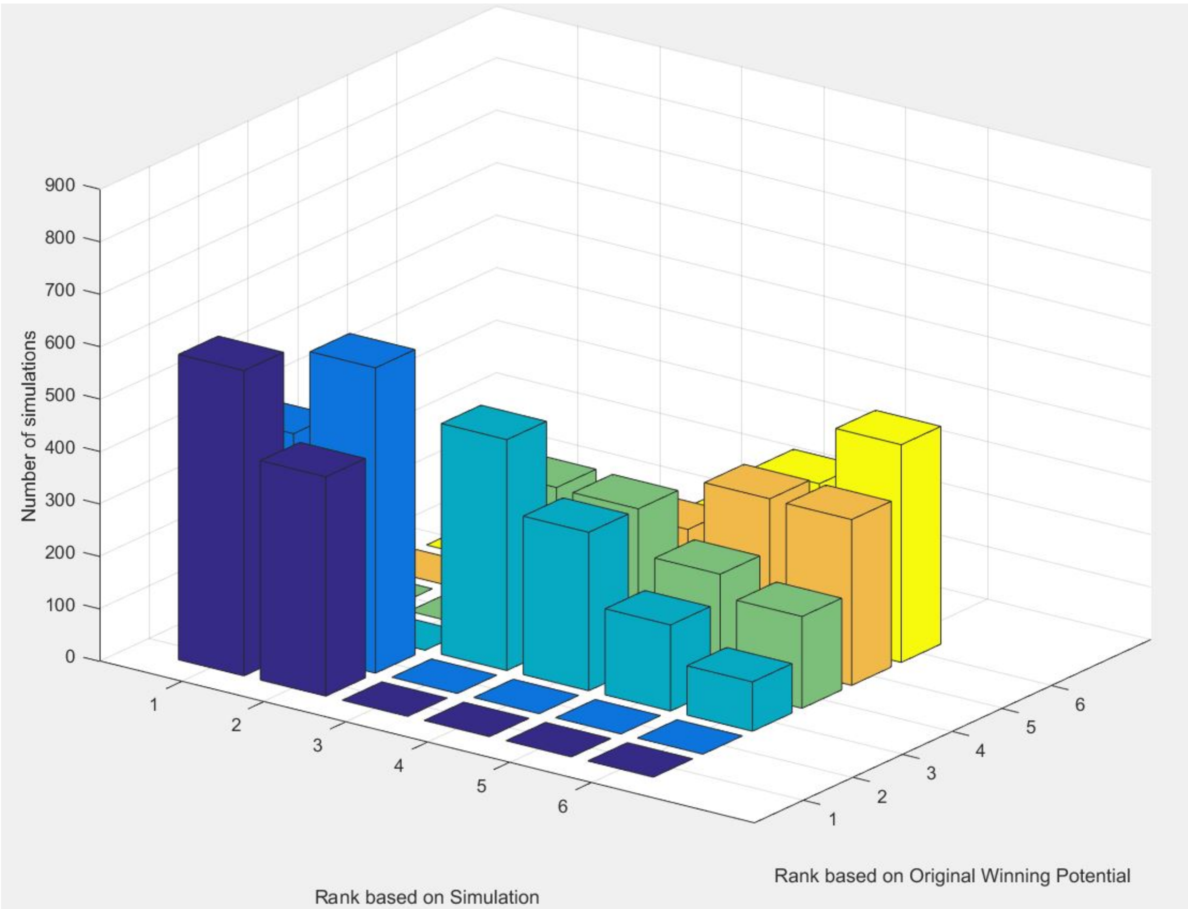


Figure 4.11: The number of times each rank was obtained for each winning potential.

Just as in the previous scenario we see that the winning potentials that have the most overlap in their density, in this case π_4 and π_5 have the lowest true recoveries.

In this chapter we've simulated different sorts of scenarios. We have varied in the number of matches, as well as the number of teams, to see the impact of those factors. We have also compared the performances of the extensions with the performances of the original model on different data sets. The simulations taught us that, even though the original model performs well under some circumstances, the model with extensions is the better option to use for our real data. Finally we have seen how the overlap of the densities correlates with the rankings we retrieve.

In the next chapter we will apply the model that includes home advantage and draws on our real data, in order to do predictions for the next season.

5 Predictions

Now that we know how our model behaves on different scenarios of simulated data, it is now time to apply the model on real data. In this chapter we will compare the extended model that includes home advantage and draws, which was introduced in Section 2.4, with the model that only includes draws, which was introduced in Section 2.2. The preferred model, based on the AIC value on this data will then be used to recover the winning potentials of the teams. The goal of this chapter is to make predictions about the ranking of the teams, that play in the Eredivisie in season 2018/2019. Since there was no data available about two teams: FC Emmen and Fortuna Sittard, this means that we will try to rank 16 teams.

We should mention that the predictions we will do are solely based on our selected data. This data consists of all the matches that the 16 teams played against each other in the past 6 seasons. In Chapter 3 we have explained why we select those 6 seasons as our data set.

The first step we take, is to apply the models on our data set and compare their AIC values. The preferred model will then recover winning potentials for every team. These winning potentials give us indications of the quality of each team compared to the other teams, based on the past 6 seasons. We then want to use these winning potentials to predict the ranking of next season.
In order to quantify the uncertainty of these rankings, we do simulations with the obtained winning potentials. Finally, some statements can be made about next season.

5.1. The Winning Potentials

After applying both models on the data, we could calculate the AIC values. The model that includes home advantage and draws had an AIC value of 2024.13. The model that only includes draws had an AIC value of 2052.631. This means that our preferred model is the one that includes home advantage and draws. This model will be used to recover the winning potentials and it will also be used to do simulations. From now on if we refer to 'the model', we mean the extension for home advantage and draws.

In this section we discuss the winning potentials for the teams, based on our real data. After applying the model on our data, we recovered a winning potential for every team. A visualization of these winning potentials can be found in Figure 5.1. An overview with the exact winning potentials is given in Table 5.1. Since these winning potentials are the most likely, the most likely ranking for the teams of next season follows from them. The recovered values for γ and θ are 1.3931712 and 1.7979117 respectively.

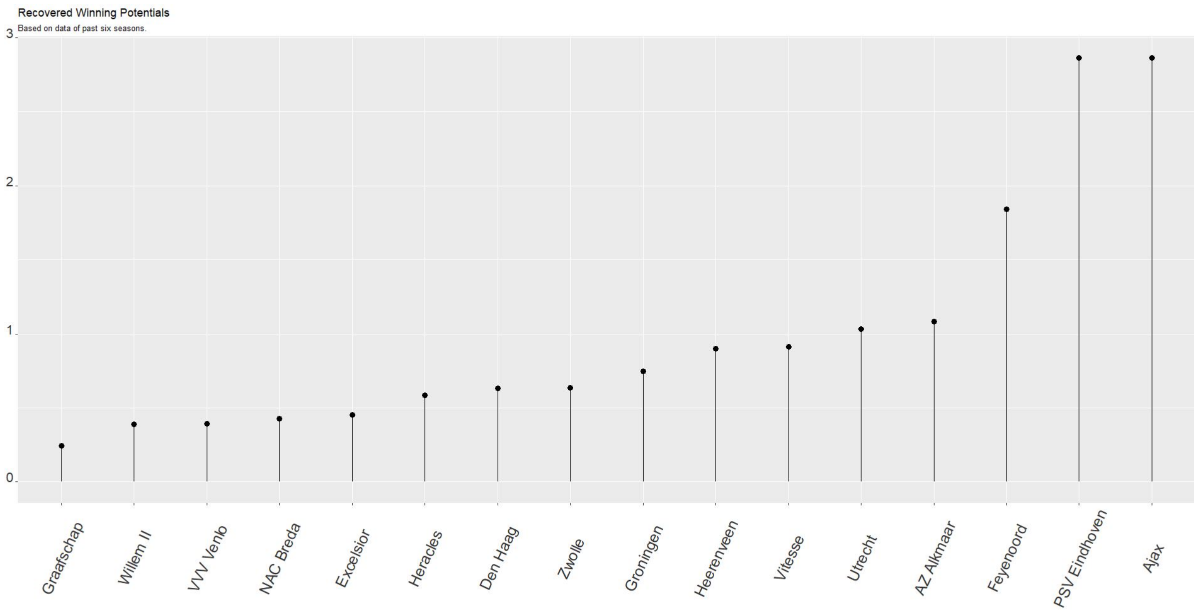


Figure 5.1: The recovered winning potentials for each team, based on the real data that we chose.

Team	Recovered Winning Potential	Most likely ranking	Number of matches in data set
Ajax	2.8649121	1	152
PSV	2.8638299	2	152
Feyenoord	1.8414906	3	152
AZ Alkmaar	1.0848692	4	152
FC Utrecht	1.0300534	5	152
Vitesse	0.9129733	6	152
SC Heerenveen	0.8985829	7	152
FC Groningen	0.7446745	8	152
PEC Zwolle	0.6364008	9	152
ADO Den Haag	0.6329242	10	152
Heracles Almelo	0.5835322	11	152
Excelsior	0.4541185	12	104
NAC Breda	0.4266287	13	102
VVV Venlo	0.3935700	14	54
Willem II	0.3894789	15	130
De Graafschap	0.2419608	16	26

Table 5.1: The recovered Winning Potentials for each team.

In Figure 5.1, there are some significant differences between the teams visible. Based on these differences, the teams can be divided into different groups, by looking at their winning potentials.

Group 1: Top Teams
In this group we would put Ajax and PSV. These two teams clearly have a significantly higher winning potential than the rest of the teams. Their winning potentials are almost equal as well. By just looking at the information we have already, these two teams are the favourites to win the title.
Scenario 4.5.2, from Chapter 4 actually comes into mind when looking at these two teams. In this particular scenario we also had two teams with a significantly higher winning potential than all the other teams. From simulating that scenario we found that in every simulation they were ranked first and second in no particular order.

Group 2: Feyenoord
The reason why Feyenoord is a group on its own is because it has a much smaller winning potential than the two top teams, but a far greater winning potential than all the other teams. From Figure 5.1, it seems that Feyenoord is our number one candidate for the number 3 spot on the ranking.

Group 3: Subtop
In this group we put AZ, FC Utrecht, Vitesse and Heerenveen. These four teams have very similar winning potentials. By looking at Figure 5.1, they are the favourites for ranks 4-7. However, the differences between the winning potentials of these teams, and those of lower teams are not as big as the differences between these teams and Feyenoord or the two top teams. This means that we shouldn't be surprised if a team that is ranked lower (based on the winning potentials we found here), will switch places with one of these teams.

Group 4: Middle-Ranking
To this group the teams FC Groningen, PEC Zwolle, ADO Den Haag and Heracles Almelo belong. There is, however small, a clear difference between these 4 teams and the lower teams from Figure 5.1. These teams are the favourites for ranks 8-11. The chances of being ranked in a higher or lower group however are very big as well, since the differences with the winning potentials of those groups are not big at all.

Group 5: Relegation zone
The teams in this group are Excelsior, NAC Breda, VVV Venlo, Willem II and De Graafschap. We call this group 'Relegation zone', because we think that these teams have the biggest chance of being ranked in one of the spots for relegation. The three teams that will be ranked 16th, 15th and 14th all have a chance of relegating. Out of all the teams in this group, De Graafschap has the significantly lowest winning potential. Intuitively, De Graafschap therefore has a small chance of switching to one of the other groups. The four other teams' winning potentials are not that far from the winning potentials of the teams in group 4.

5.2. Simulating the next season

In this section, the next year of the Eredivisie will be simulated, using the obtained parameters from Section 5.1. To simulate this year, in the simulations every team plays one home match and one away match against each other team, just as in real life. We do 10000 simulations of this season. In Figure 5.2, the density estimates, based on the simulations are displayed. In general, the teams with the higher winning potentials have more uncertainty in recovering them. From Chapter 4, we know that the uncertainty of the rankings are correlated with the overlap of the estimated densities of the winning potentials and not necessarily with the uncertainty of the winning potentials. Table 5.2 provides the 95% intervals for all recovered parameters.

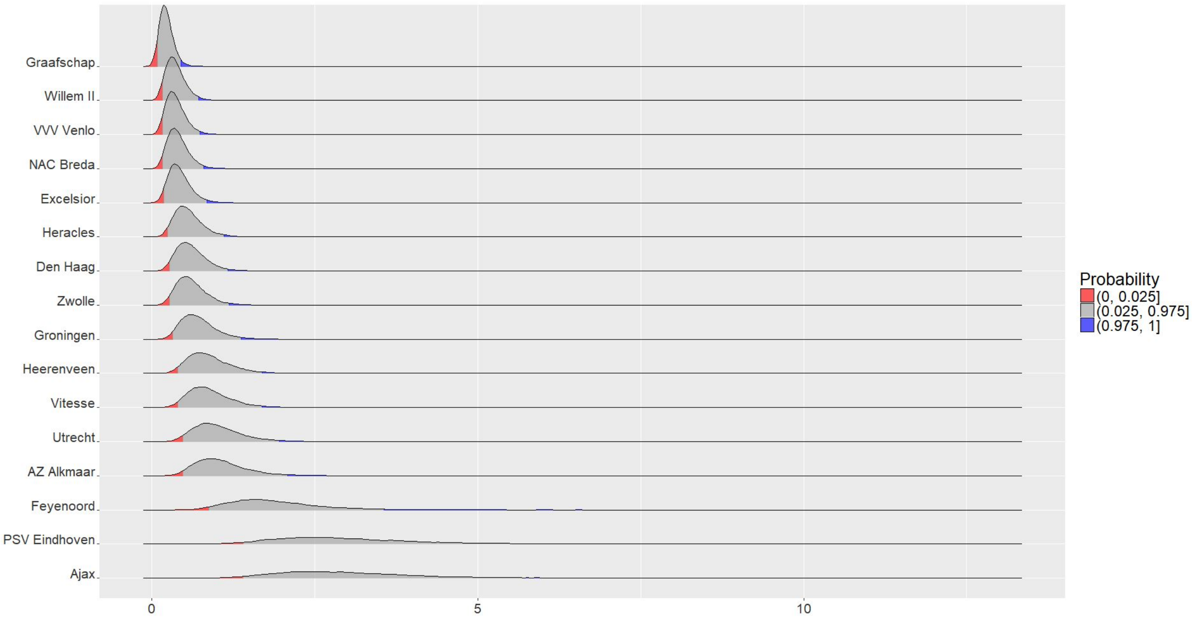


Figure 5.2: The densities of the recovered winning potentials for all 16 teams.

2.5% Quantile	Team	97.5% Quantile
1.40134148	Ajax	5.6752638
1.39779010	PSV	5.5722262
0.85694749	Feyenoord	3.5514376
0.47371577	AZ	2.0626182
0.45219771	FC Utrecht	1.9399500
0.39342326	Vitesse	1.6736144
0.38679169	SC Heerenveen	1.6757400
0.31496588	FC Groningen	1.3710513
0.25611264	PEC Zwolle	1.1799286
0.25712756	ADO Den Haag	1.1420656
0.23518850	Heracles Almelo	1.0921018
0.17448594	Excelsior	0.8223090
0.16235852	NAC Breda	0.7768850
0.14562917	VVV Venlo	0.7199392
0.14520683	Willem II	0.7029000
0.08058283	De Graafschap	0.4403679

Table 5.2: The 95% interval of the recoveries of the winning potentials.

The 95% intervals for the recoveries of the two other parameters were:

- $1.10362076 < \gamma < 1.8876908$
- $1.61123286 < \theta < 2.1848351$

To get a more detailed analysis of the results we will look at each group, introduced in Section 5.1, separately.

5.2.1. Group 1

In this section we analyze the results of the two top teams, Ajax and PSV. In the simulations we used winning potentials of 2.8649121 and 2.8638299 respectively for these two teams. The estimated densities of the winning potentials of these teams are shown in Figure 5.3.

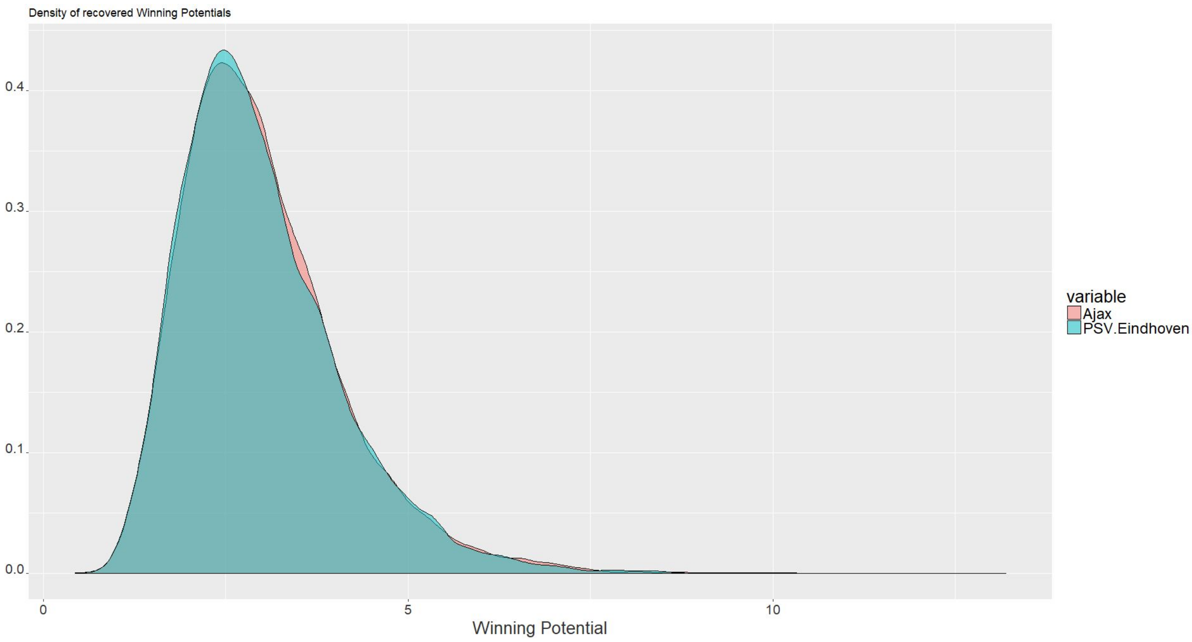


Figure 5.3: The densities of the recovered winning potentials for Ajax and PSV.

The densities show that the recoveries of both teams are very similar. They are almost the same, which could be expected since their winning potentials are also almost the same. The densities almost compltely overlap each other, which suggests that their rank switches often. Figure 5.2 also shows that, the only other team that has significant overlap with Ajax and PSV is Feyenoord. There is not a lot of overlap with the other teams, so Ajax and PSV have not been ranked at position 4 or lower very often. We can see this in Figure 5.4, which shows the number of times the two teams were ranked at each position. In Table 5.3 we provided some additional information about the rankings of this group.

Team	% of being ranked in Group 1 (Positions 1-2)	% of being ranked in Group 2 (Position 3)	Most frequent Rank (%)
Ajax	82.35%	13.70%	1 (45.75%)
PSV	81.84%	14.30%	1 (44.68%)

Table 5.3: Additional information about the recovered rankings of this group.

We learn from Figure 5.4 and Table 5.3, that even though we have seen that these two teams have the widest density estimates, the rankings do not vary a lot. Ajax and PSV were ranked in the top 3 in 96.05% and 96.14% of the simulations respectively.

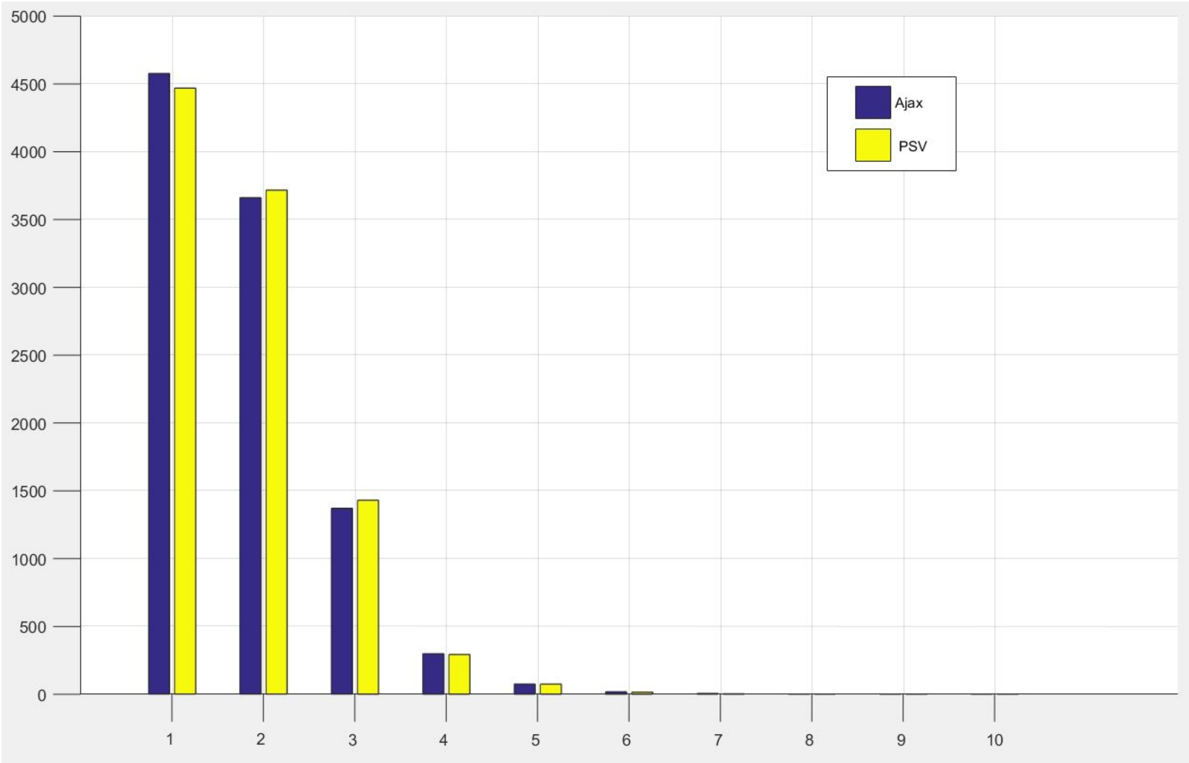


Figure 5.4: The number of times, Ajax and PSV were ranked at the respective positions. The x-axis represents the ranks, the y-axis the number of times the team was ranked at this position.

5.2.2. Group 2

This group only contains Feyenoord. As mentioned before, Feyenoord is a special team, based on its winning potential. If we look at Figure 5.5 as mentioned before, we observe that Feyenoord is the only other team that has serious overlap with the top two teams. The 95% interval of the recoveries of the winning potentials of Feyenoord is relatively big, just as the intervals of Ajax and PSV.

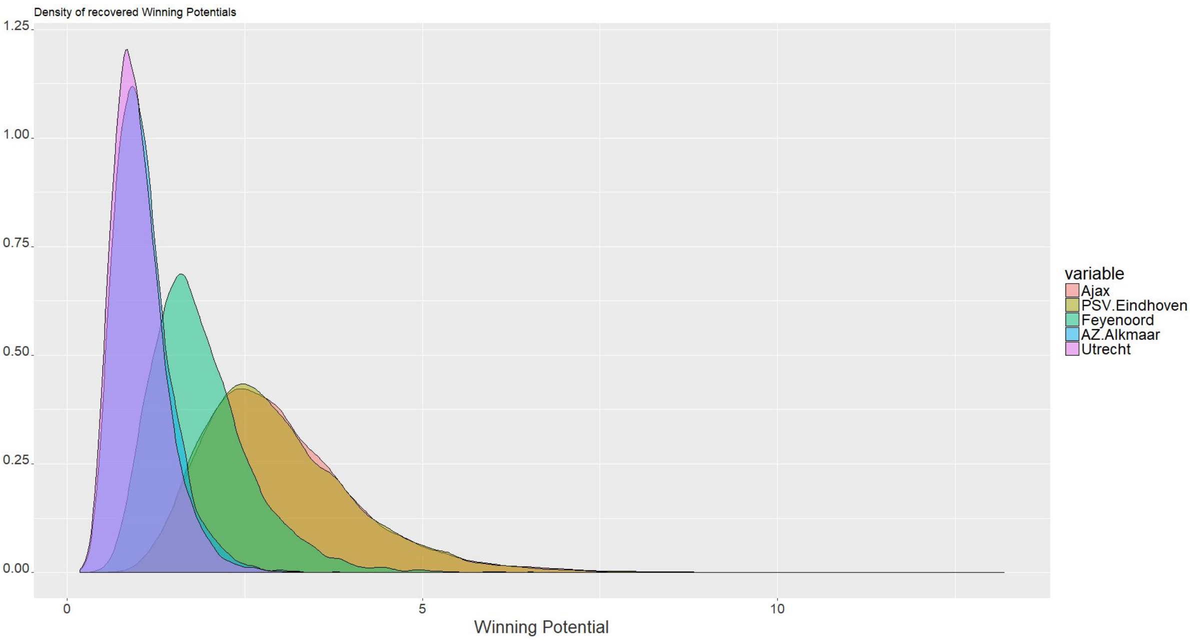


Figure 5.5: The density of Feyenoord and the closest teams.

Figure 5.5 shows how often Feyenoord was ranked at each position.

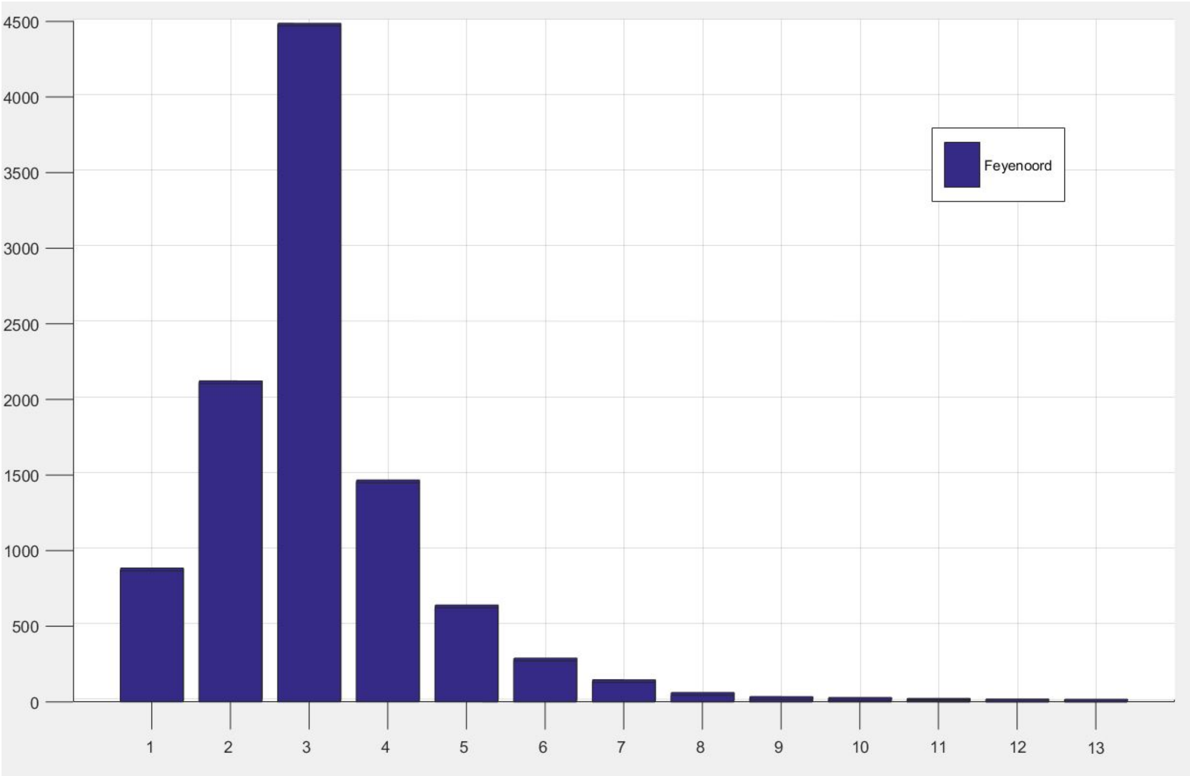


Figure 5.6: The number of times Feyenoord was ranked at the respective positions. The x-axis represents the ranks, the y-axis the number of times the team was ranked at this position.

We see in Figure 5.6, that Feyenoord is clearly ranked the most as number 3. This is due to the fact that the originally recovered winning potential is significantly lower than the two top teams, but significantly higher than all the other teams. In Table 5.4 some additional information about the rankings of Feyenoord is stated.

Team	% Group 1	% Group 2	% Group 3	Most frequent Rank (%)
Feyenoord	29.72%	44.70%	24.74%	3 (44.70%)

Table 5.4: Additional information about the recovered rankings of this group.

5.2.3. Group 3

In this group we find the so called 'subtop' of the Eredivisie. We see, compared to Group 1 and 2 that the densities of this group are already significantly less wide. What also stands out from Figure 5.2, compared to the earlier groups, is that the teams in this group not only have a lot of overlap with each other, but also relatively much with Group 4. In Figure 5.7 we plotted the densities of the teams in this group.

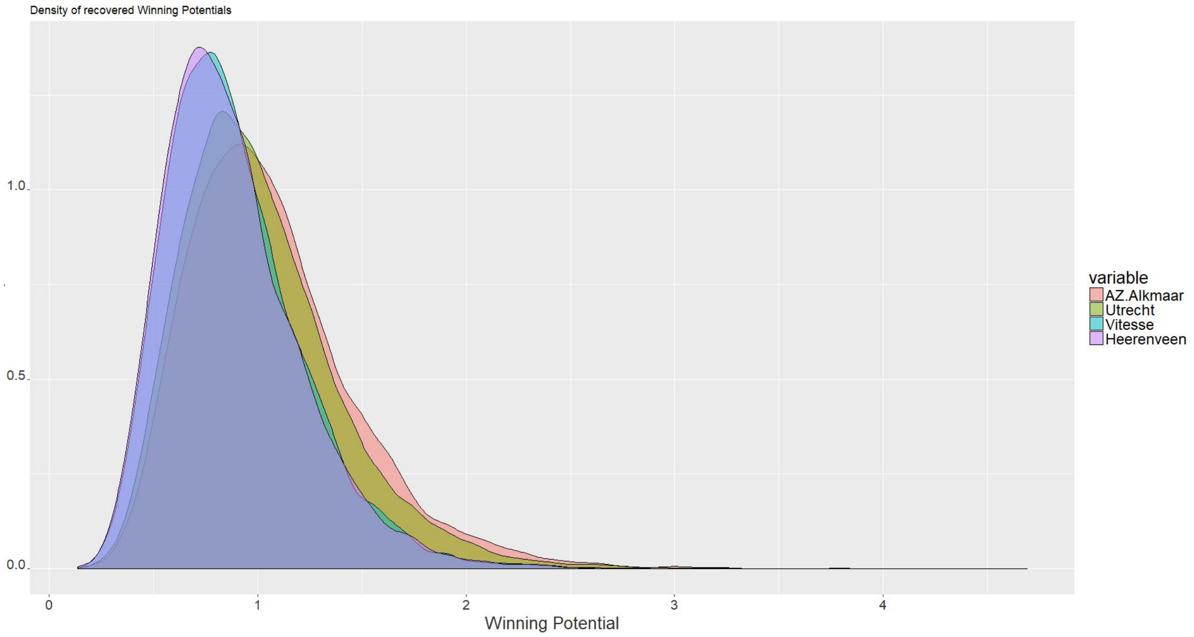


Figure 5.7: The density of the teams in Group 3.

There is a lot of overlap between the teams and we can see this back in Figure 5.8.
Compared to Group 1 and Group 2, the teams are ranked more often at different position. The highest frequency for a rank in this group was for the number of times AZ was ranked as number 4. This happened 2337 times, so compared to the Groups 1 and 2 there is a clear difference. In Table 5.5 we present some additional information about the rankings of this group. Table 5.5 tells us that for every team, the probability of being

Team	% Group 1	% Group 2	% Group 3	% Group 4	% Group 5	Most frequent Rank (%)
AZ	2.62%	9.32%	69.24%	17.48%	1.34%	4 (23.37%)
FC Utrecht	1.80%	7.48%	68.53%	20.03%	2.16%	4 (19.80%)
Vitesse	0.80%	4.10%	60.08%	30.50%	4.52%	5 (16.29%)
SC Heerenveen	0.63%	3.89%	58.30%	32.73%	4.45%	5 (15.68%)

Table 5.5: Additional information about the recovered rankings of this group.

ranked in a higher group (Group 1 or 2) is smaller than the probability of being ranked in a lower group (Group 4 or 5). This again tells us, as we already observed in Figure 5.2, that this group is closer to Group 4, than to the higher groups.

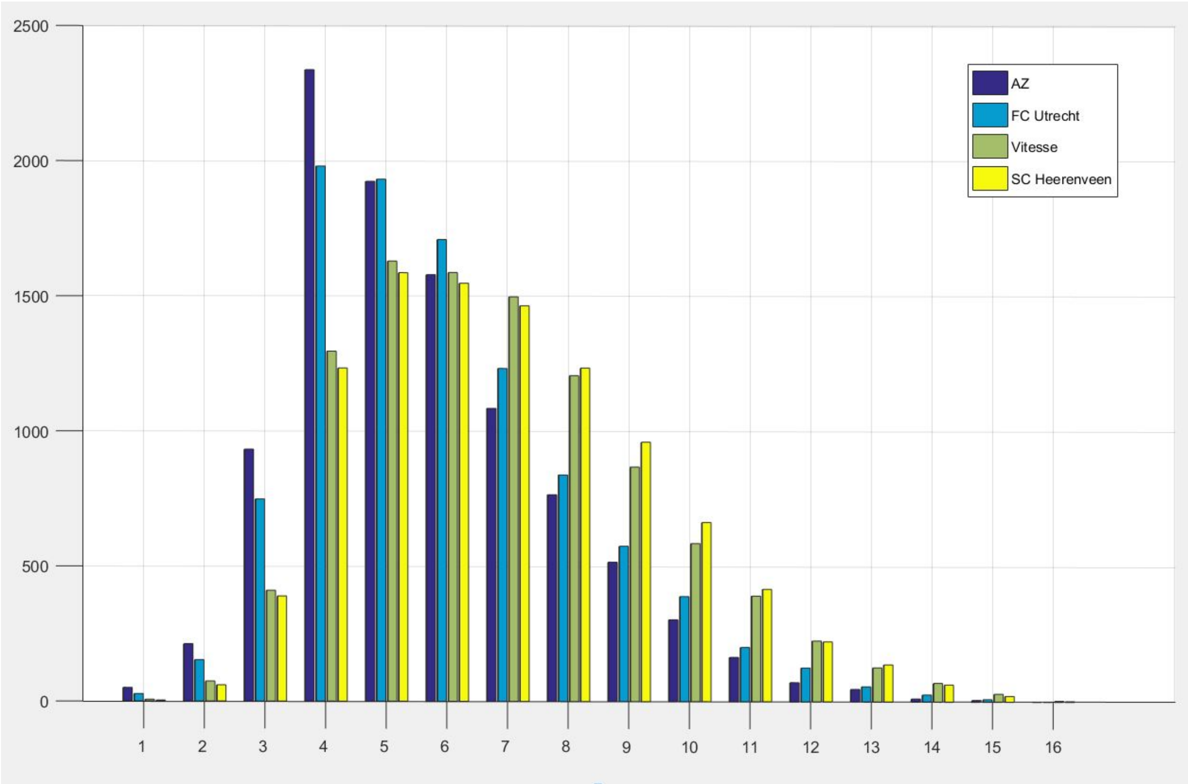


Figure 5.8: The number of times, the teams from this group were ranked at each position. The x-axis represent the rank, the y-axis the number of times the team was ranked at this position.

5.2.4. Group 4

Group 4 contains the middle-ranked teams. Based on the winning potentials from Figure 5.1, these teams take up ranks 8-11. From Figure 5.2, this groups seems to be right in between Group 3 and Group 5. It looks like there is about equally much overlap with those groups. This would mean that the probability of being ranked at the positions of Group 3 is approximately equal to the probability of being ranked at the positions of Group 5. The density functions of the teams can be found in Figure 5.9.

The density functions tell us that FC Groningen generally performs a bit better than the other teams, while Heracles generally performs the worst out of the four teams. ADO Den Haag and PEC Zwolle have almost got the same density estimate, which can be explained by their winning potentials from Table 5.1, which are 0.6329242 and 0.6364008 respectively. The expectation therefore is that they will be ranked the same way. The recovered rankings are shown in Figure 5.10.

Table 5.6 contains some additional information about the recoverd rankings of this group.

Team	% Group 1	% Group 2	% Group 3	% Group 4	% Group 5	Most frequent Rank (%)
FC Groningen	0.09%	1.31%	38.76%	47.24%	12.60%	8 (14.56%)
PEC Zwolle	0.06%	0.37%	22.90%	51.92%	24.75%	10 (14.24%)
ADO Den Haag	0.05%	0.54%	22.26%	51.59%	25.56%	10 (13.86%)
Heracles Almelo	0.04%	0.20%	16.52%	50.41%	32.83%	11 (14.50%)

Table 5.6: Additional information about the recovered rankings of this group.

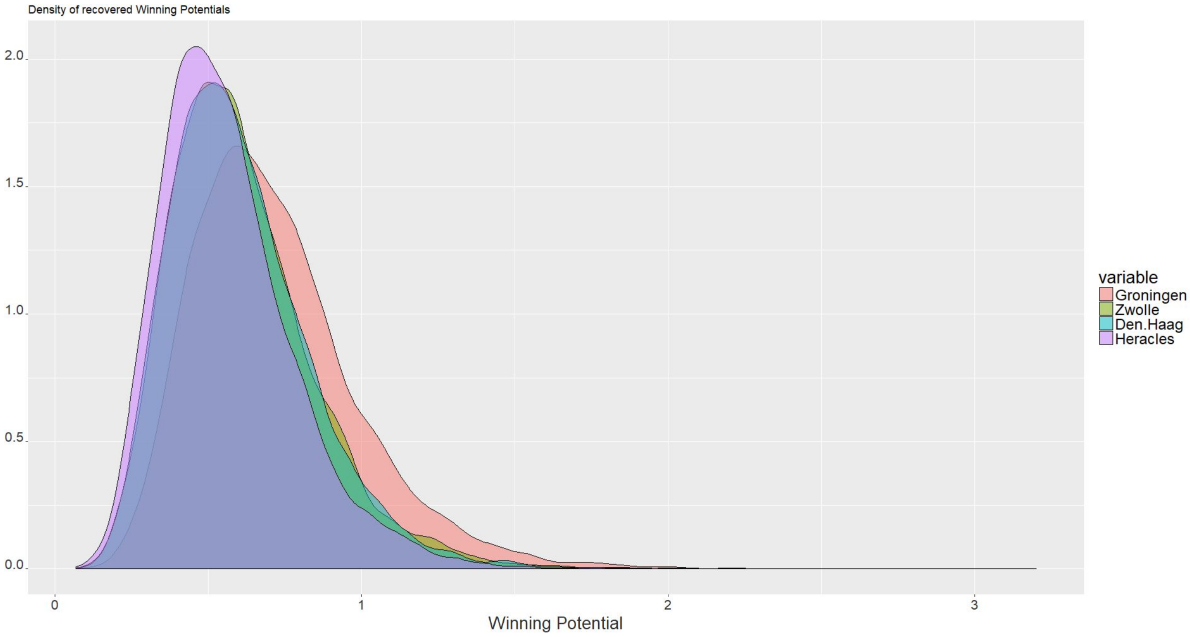


Figure 5.9: The density functions of the teams from Group 4.

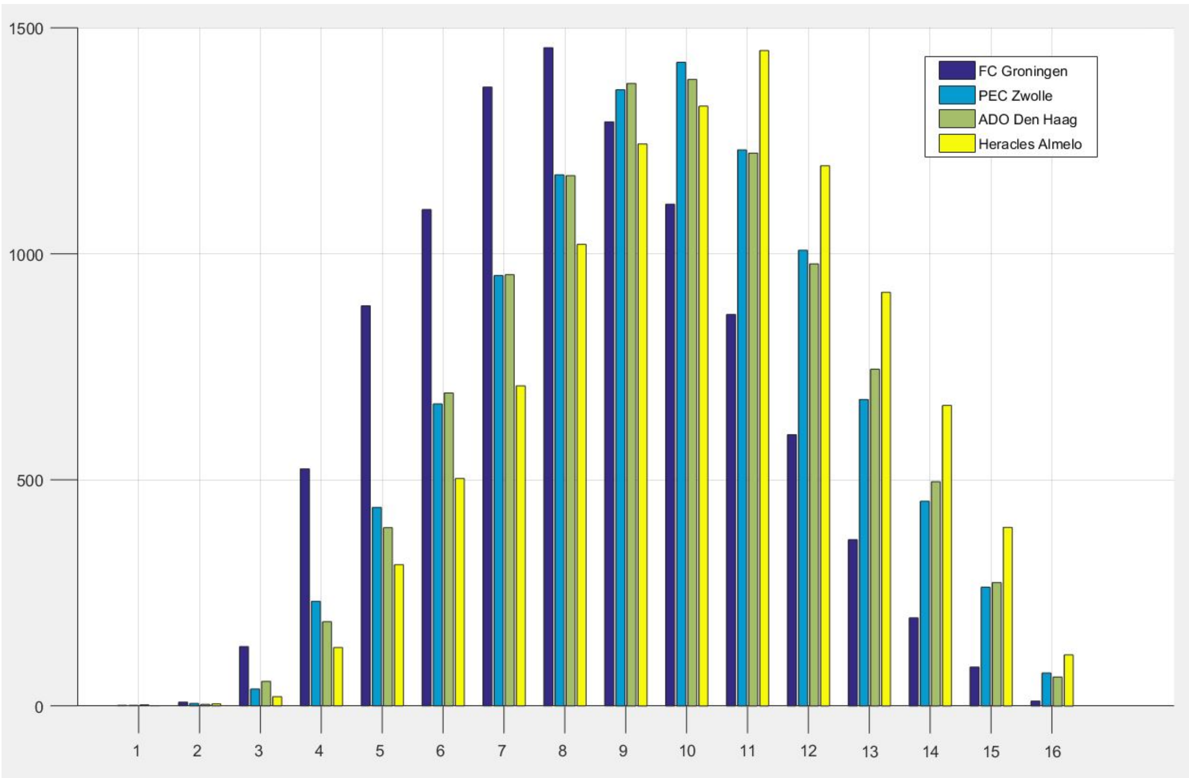


Figure 5.10: The number of times, the teams from this group were ranked at each position. The x-axis represent the rank, the y-axis the number of times the team was ranked at this position.

5.2.5. Group 5

The final group, which is also the largest group, consists of the teams that are in serious danger of relegating. Based on Figure 5.1, these teams have positions 12-16. If a team gets ranked at position 14-16 it is in danger of relegating. These positions we call the relegation zone. Figure 5.11 shows the densities of the the recovered winning potentials for the teams in this group.

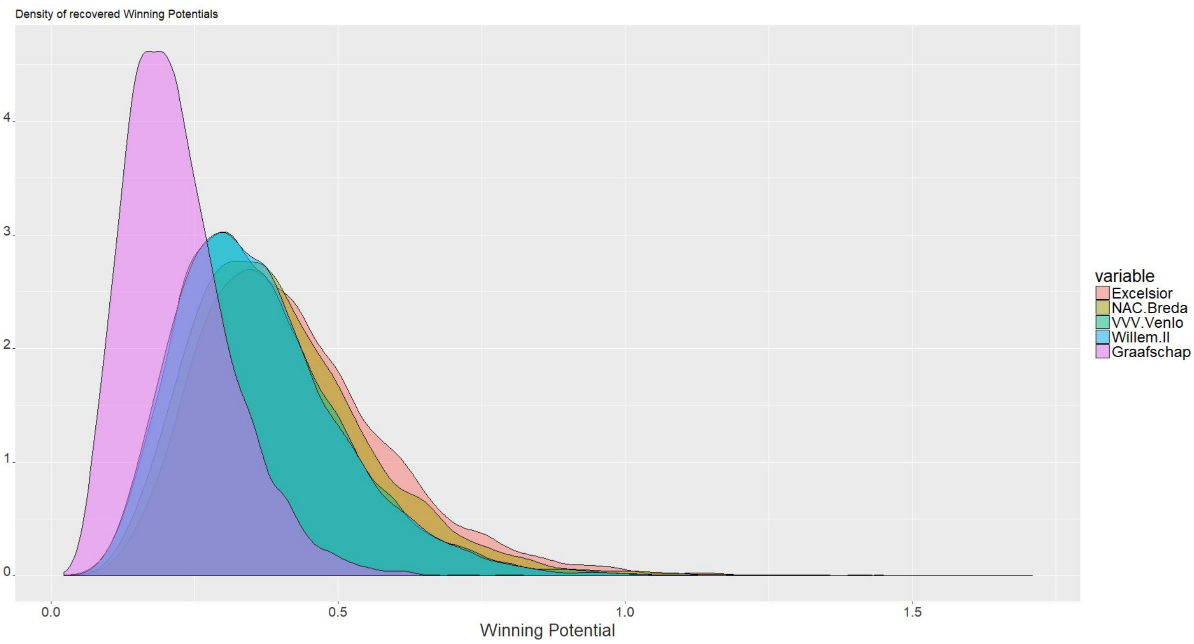


Figure 5.11: The density functions of the teams from Group 5.

If we look at these densities, we see that De Graafschap performs significantly worse than the other teams in this group. The densities of the other teams are more equal to each other. The rankings based on the simulations for the teams in this group can be found in Figure 5.12.

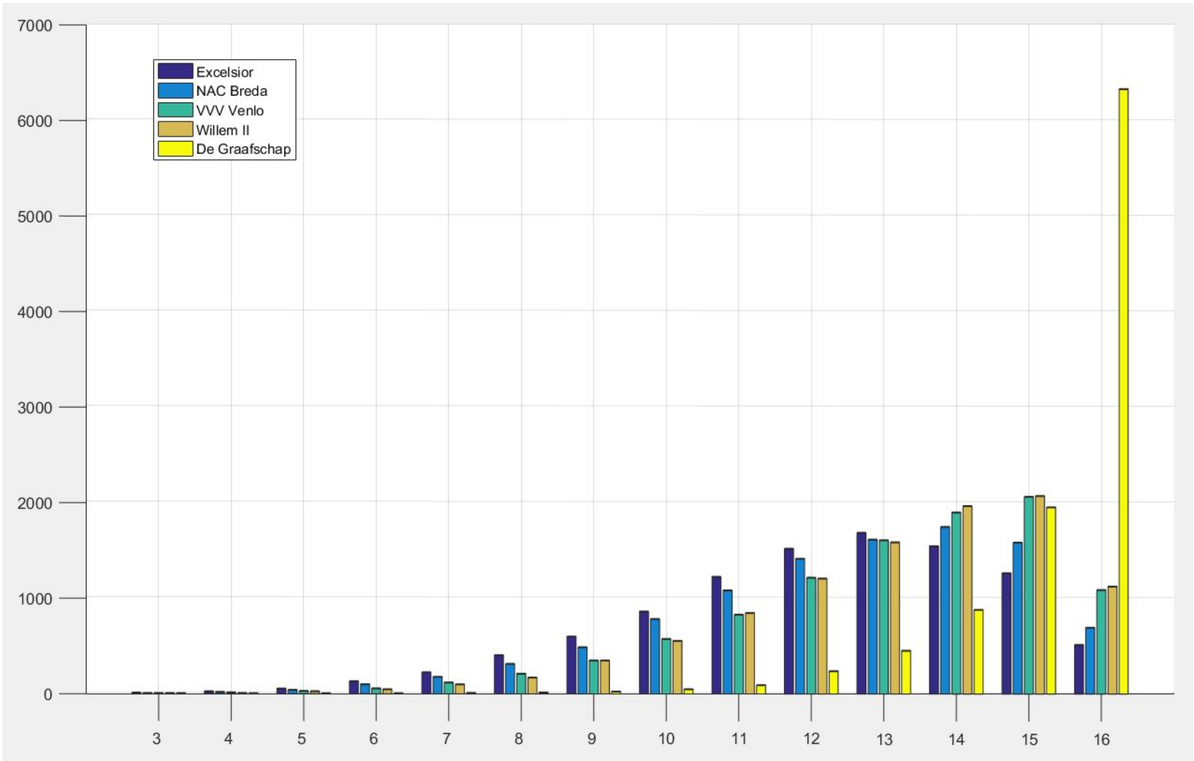


Figure 5.12: The number of times, the teams from this group were ranked at each position. The x-axis represent the rank, the y-axis the number of times the team was ranked at this position.

Figure 5.11 shows that, as we already saw in the density estimates, De Graafschap has significantly the biggest probability of ending at the bottom of the league, based on the simulations. Table 5.7 displays some additional information about the rankings.

Team	% Group 3	% Group 4	% Group 5	% Relegation Zone	Most frequent Rank (%)
Excelsior	4.14%	30.71%	65.09%	33.12%	13 (16.82%)
NAC Breda	3.15%	26.49%	70.35%	40.14%	14 (17.44%)
VVV Venlo	1.97%	19.45%	78.57%	50.41%	15 (20.60%)
Willem II	1.59%	19.02%	79.38%	51.52%	15 (20.68%)
De Graafschap	0.04%	1.59%	98.37%	91.55%	16 (63.28%)

Table 5.7: Additional information about the recovered rankings of this group.

5.2.6. Correlation Matrix

The recovered winning potentials are correlated, as we mentioned earlier already. The matrix from Figure 5.13 displays the correlations between the recovered winning potentials of the teams, based on the 10000 simulations from this Section.

	Ajax	PSV Eindhoven	Feyenoord	AZ Alkmaar	Utrecht	Vitesse	Heerenveen	Groningen	Zwolle	Den Haag	Heracles	Excelsior	NAC Breda	VVV Venlo	Willem II	Graafschap
Ajax	1.00	-0.35	-0.24	-0.19	-0.20	-0.17	-0.18	-0.16	-0.19	-0.18	-0.18	-0.16	-0.16	-0.16	-0.17	-0.15
PSV Eindhoven	-0.35	1.00	-0.24	-0.19	-0.18	-0.17	-0.18	-0.18	-0.17	-0.16	-0.17	-0.16	-0.17	-0.15	-0.14	-0.15
Feyenoord	-0.24	-0.24	1.00	-0.08	-0.06	-0.06	-0.05	-0.05	-0.06	-0.07	-0.03	-0.05	-0.04	-0.05	-0.03	-0.03
AZ Alkmaar	-0.19	-0.19	-0.08	1.00	0.00	0.00	0.02	0.02	0.04	0.03	0.04	0.03	0.03	0.05	0.04	0.04
Utrecht	-0.20	-0.18	-0.06	0.00	1.00	0.00	0.01	0.03	0.05	0.02	0.03	0.06	0.06	0.06	0.04	0.07
Vitesse	-0.17	-0.17	-0.06	0.00	0.00	1.00	0.02	0.02	0.05	0.06	0.04	0.07	0.06	0.04	0.07	0.05
Heerenveen	-0.18	-0.18	-0.05	0.02	0.01	0.02	1.00	0.04	0.05	0.06	0.05	0.06	0.06	0.04	0.07	0.06
Groningen	-0.16	-0.18	-0.05	0.02	0.03	0.02	0.04	1.00	0.07	0.08	0.07	0.07	0.08	0.08	0.07	0.09
Zwolle	-0.19	-0.17	-0.06	0.04	0.05	0.05	0.05	0.07	1.00	0.09	0.08	0.10	0.09	0.10	0.10	0.10
Den Haag	-0.18	-0.16	-0.07	0.03	0.02	0.06	0.06	0.08	0.09	1.00	0.07	0.09	0.09	0.09	0.09	0.09
Heracles	-0.18	-0.17	-0.03	0.04	0.03	0.04	0.05	0.07	0.08	0.07	1.00	0.11	0.10	0.10	0.08	0.11
Excelsior	-0.16	-0.16	-0.05	0.03	0.06	0.07	0.06	0.07	0.10	0.09	0.11	1.00	0.11	0.12	0.11	0.12
NAC Breda	-0.16	-0.17	-0.04	0.03	0.06	0.06	0.06	0.08	0.09	0.09	0.10	0.11	1.00	0.11	0.11	0.14
VVV Venlo	-0.16	-0.15	-0.05	0.05	0.06	0.04	0.04	0.08	0.10	0.09	0.10	0.12	0.11	1.00	0.10	0.13
Willem II	-0.17	-0.14	-0.03	0.04	0.04	0.07	0.07	0.07	0.10	0.09	0.08	0.11	0.11	0.10	1.00	0.12
Graafschap	-0.15	-0.15	-0.03	0.04	0.07	0.05	0.06	0.09	0.10	0.09	0.11	0.12	0.14	0.13	0.12	1.00

Figure 5.13: Correlation Matrix

By the set up we used for these simulations, we ignored the fact that in our real data not every team played the same number of matches. As discussed in Chapter 4, the number of matches a team plays has a big impact on the uncertainty of the recovery of its winning potential. Table 5.1 shows the number of matches in our real data per team. It shows that the recovered winning potential of De Graafschap is based on 26 matches, while the recovered winning potential of for example Vitesse, is based on 152 matches. Since we know that the number of matches has an impact, we know that the recovered Winning Potential of De Graafschap is much more 'uncertain' than the winning potential of Vitesse. It only seems reasonable to include this uncertainty in our results. In the next section we will discuss a way to do so.

5.3. Simulating, including the uncertainty

In this section we will simulate the next season again, but with a different set up. The goal of the method used in this section is to include the uncertainty, caused by the different amounts of matches on which the winning potentials are based. This means that in this simulation, the recoveries of the winning potentials of teams that have played less matches, should be relatively more uncertain compared to teams that have played more matches.

The idea is to first run 1000 simulations that cause the uncertainty. We will refer to these simulations as the 'uncertainty simulations'. In these simulations, teams play different amounts of matches, depending on the number of matches on which their winning potential of the real data is based. The matrix in Figure 5.14 contains the number of matches between the teams in each of these simulations. Each coordinate $[i, j]$ represents the number of matches team i plays at home against team j. This number is based on the minimum of the numbers of seasons the teams have played in our data. The winning potentials and the parameters γ and θ used in these simulations are the same as in the previous simulations and can be found in Table 5.1. The total number of matches each team plays, caused by this matrix is shown in Table 5.8. The number of matches played in the uncertainty simulations represent the number of matches in our real data quite well.

	Ajax	PSV Eindhoven	Feyenoord	AZ Alkmaar	Utrecht	Vitesse	Heerenveen	Groningen	Zwolle	Den Haag	Heracles	Excelsior	NAC Breda	VVV Venlo	Willem II	Graafschap
Ajax	0	6	6	6	6	6	6	6	6	6	6	4	4	2	5	1
PSV Eindhoven	6	0	6	6	6	6	6	6	6	6	6	4	4	2	5	1
Feyenoord	6	6	0	6	6	6	6	6	6	6	6	4	4	2	5	1
AZ Alkmaar	6	6	6	0	6	6	6	6	6	6	6	4	4	2	5	1
Utrecht	6	6	6	6	0	6	6	6	6	6	6	4	4	2	5	1
Vitesse	6	6	6	6	6	0	6	6	6	6	6	4	4	2	5	1
Heerenveen	6	6	6	6	6	6	0	6	6	6	6	4	4	2	5	1
Groningen	6	6	6	6	6	6	6	0	6	6	6	4	4	2	5	1
Zwolle	6	6	6	6	6	6	6	6	0	6	6	4	4	2	5	1
Den Haag	6	6	6	6	6	6	6	6	6	0	6	4	4	2	5	1
Heracles	6	6	6	6	6	6	6	6	6	6	0	4	4	2	5	1
Excelsior	4	4	4	4	4	4	4	4	4	4	4	0	4	2	4	1
NAC Breda	4	4	4	4	4	4	4	4	4	4	4	4	0	2	4	1
VVV Venlo	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	1
Willem II	5	5	5	5	5	5	5	5	5	5	5	4	4	2	0	1
Graafschap	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Figure 5.14: Matrix that represents the number of matches played in each simulation. Coordinate $[i, j]$ represents the number of matches between i and j where j played at home.

Team	Number of Matches in uncertainty simulations
Ajax	152
PSV	152
Feyenoord	152
AZ Alkmaar	152
FC Utrecht	152
Vitesse	152
SC Heerenveen	152
FC Groningen	152
PEC Zwolle	152
ADO Den Haag	152
Heracles Almelo	152
Excelsior	110
NAC Breda	110
VVV Venlo	58
Willem II	132
De Graafschap	30

Table 5.8: The number of matches each team plays in one simulation of the uncertainty simulations.

The next step is to simulate the next year of the Eredivisie, using the uncertainty simulations. In every uncertainty simulation we recover winning potentials, as well as a value for γ and a value for θ. This vector of recovered parameters will then be used as the values we will simulate with, when simulating next year. With each recovered parameter vector we will simulate the next year 10 times. This means that, just as in the previous simulation method, we do 10000 simulations in total. The difference is that this time we use a method to include the uncertainty caused by the different number of matches per team in our real data.

We found that using this method, the recoveries in general were more uncertain. This is not necessarily a bad thing, as this may be more realistic. The relative uncertainty which we wanted to achieve with this method also occurred. We quantified this relative uncertainty by comparing the 95% intervals of both methods. The relative uncertainty then can be calculated as a factor by dividing the length of the 95% interval of these simulations by the length of the 95% interval of the previous simulations. These values can be found in Table 5.9. The 95% interval for the recovered values of θ and γ are as follows:

- $1.0645862 < \gamma < 1.9551607$
- $1.5968056 < \theta < 2.2560513$

Results about the rankings will be discussed per group very briefly.

2.5% Quantile	Team	97.5% Quantile	Relative uncertainty factor
1.2984578	Ajax	6.0122475	1.102919
1.3036931	PSV	6.0305460	1.132333
0.7846610	Feyenoord	3.7359640	1.095310
0.4432185	AZ Alkmaar	2.1329203	1.063440
0.4092847	FC Utrecht	2.0575802	1.107910
0.3615584	Vitesse	1.7946256	1.119417
0.3509869	SC Heerenveen	1.7347495	1.073559
0.2780897	FC Groningen	1.4560280	1.115382
0.2321371	PEC Zwolle	1.2452122	1.096620
0.2316161	ADO Den Haag	1.1918874	1.085128
0.2144738	Heracles Almelo	1.1273397	1.065296
0.1552840	Excelsior	0.9019069	1.152510
0.1371790	NAC Breda	0.8189569	1.109436
0.1182060	VVV Venlo	0.8087335	1.202360
0.1248081	Willem II	0.7477193	1.116942
0.0472987	De Graafschap	0.5668559	1.444077

Table 5.9: The 95% intervals for the recovered winning potentials and the relative uncertainty factor.

5.3.1. Group 1

This group consists of the two top teams, Ajax and PSV. Their recovered rankings can be found in Table 5.10.

Team	#Rank1	#Rank2	% Group1	% Group2	% Group3	% Group4	% Group5
Ajax	4489	3565	80.54%	14.17%	5.22%	0.07%	0.00%
PSV	4442	3554	79.96%	14.61%	5.41%	0.02%	0.00%

Table 5.10: Additional information about the recovered rankings of this group.

5.3.2. Group 2

This is the group that only consists of Feyenoord. The recovered rankings can be found in Table 5.11.

Team	#Rank1	#Rank2	#Rank3	% Group1	% Group2	% Group3	% Group4	% Group 5
Feyenoord	915	2134	3953	30.49%	39.53%	28.47%	1.46%	0.05%

Table 5.11: Additional information about the recovered rankings of Feyenoord.

5.3.3. Group 3

In this group we find AZ, FC Utrecht, Vitesse and SC Heerenveen. Their recovered rankings can be found in Table 5.12.

Team	#Rank4	#Rank5	#Rank6	#Rank 7	% Group2	% Group3	% Group4	% Group5
AZ	2109	1847	1536	1105	9.95%	65.97%	18.65%	2.17%
FC Utrecht	1771	1877	1542	1198	8.67%	63.88%	21.61%	3.09%
Vitesse	1275	1484	1522	1353	4.98%	56.34%	31.95%	5.56%
SC Heerenveen	1199	1473	1506	1342	4.44%	55.20%	32.55%	6.61%

Table 5.12: Additional information about the recovered rankings of this group.

5.3.4. Group 4

This group contains the middle-ranked teams, which have ranks 8-11 according to the winning potentials based on the real data. The recovered rankings for the teams in this group are shown in Table 5.13.

Team	#Rank8	#Rank9	#Rank10	#Rank11	% Group2	% Group3	% Group4	% Group 5
FC Groningen	1261	1173	1101	884	1.77%	38.41%	44.19%	15.25%
PEC Zwolle	1123	1311	1288	1150	0.71%	23.40%	28.72%	27.09%
ADO Den Haag	1121	1227	1309	1177	0.52%	23.35%	48.34%	27.70%
Heracles Almelo	1002	1182	1308	1320	0.40%	17.28%	48.12%	34.16%

Table 5.13: Additional information about the recovered rankings of this group.

5.3.5. Group 5

The final group contains the 5 worst teams according to our model. Their recovered rankings are stored in Table 5.14.

Team	#Rank12	#Rank13	#Rank14	#Rank15	#Rank16	% Group3	% Group4	% Group5
Excelsior	1320	1488	1524	1316	599	6.14%	31.24%	62.47%
NAC Breda	1300	1522	1735	1545	877	4.06%	26.11%	69.79%
VVV Venlo	1080	1357	1714	2068	1261	3.55%	21.57%	74.80%
Willem II	1193	1548	1757	2089	1154	2.67%	19.90%	77.41%
De Graafschap	363	595	994	1737	5696	0.65%	5.50%	93.85%

Table 5.14: Additional information about the recovered rankings of this group.

5.4. Making Predictions

In this section we will make some predictions, based on the simulations we have done. Probabilities for teams ending at a certain rank, can be based on frequencies in those simulations. Besides the most likely ranking according to our model, which we have already given in Figure 5.1, there are a few things interesting to do predictions about.

- Who becomes the champion
- Who ends at the second place (This team plays Champions League in the next season).
- Who ends at third place (This team plays Europa League in the next season).
- Which teams end at places 4-8 (These teams have a chance of battling each other in a small tournament, to qualify for the Europa League).
- Which teams end at places 13-16 (These teams have a chance of relegating).

Who becomes the champion?
Based on the simulations of both methods there are only two clear candidates to become champion.

1. Ajax
The first team we have to name here is Ajax. In both methods of simulating, Ajax was the team that was ranked number one the most. this Based on the first simulations, where we didn't correct for the uncertainty, the probability that Ajax becomes champion next season is 0.4575.
Based on the second simulations, this probability is 0.4489.

2. PSV
The other team we should name is PSV. Ajax and PSV had very similar results in both simulations. Their winning potentials based on the real data were almost exactly the same. Based on the first simulations, the probability that PSV becomes champion next season is 0.4468. If we look at the second simulations, this

probability is 0.4442. Ajax and PSV almost have the same probability of becoming champion, based on our simulations, so there is no clear favourite between those two teams.

Looking at the results of both simulations, Ajax and PSV are the clear favourites to become champion. There is only one outsider that has a significant chance of beating them in the race for the title. This outsider is Feyenoord. Looking at the first simulations, they have a probability of 0.0867 of becoming the champions. If we would base the probability on the second set of simulations, the probability would be a bit higher: 0.0915.

Who ends at second place?
To answer this question, we again look at three teams: Ajax, PSV and Feyenoord.

1. PSV
Since PSV and Ajax, were clearly the two top teams, they are the favourites for the number 1 as well as the number 2 spot of next season. Also for the number 2 spot, their probabilities are very close. Based on the first simulations, PSV has a probability of 0.3716 to be ranked 2nd next season. The probability with regards to the second set of simulations is 0.3554.

2. Ajax
Just as when we answered the question who becomes champion, there is no clear favourite between Ajax and PSV for this spot. The probability that Ajax ends 2nd based on the first simulations is 0.3660, while this probability based on the second simulations is 0.3565.

3. Feyenoord
The outsider for this position is again Feyenoord. This time it's less of an outsider however and we could even call it 'one of the favourites', where this choice of words was not really applicable in predicting the champion. Based on the first simulations, the probability that Feyenoord will be ranked 2nd is 0.2105. The second simulations imply a probability of 0.2134.

Who ends at third place?
1. Feyenoord
As was already visible in the density plots, Feyenoord is the big favourite to end at the 3rd place. The first set of simulations suggests a probability of 0.4470 for this to happen, while the second set of simulations suggests a probability of 0.3953.

2/3. Ajax/PSV
The two top teams are the next two favourites for this position. For both teams the probability that they become 3rd next year is about 0.14, based on both sets of simulations.

Which teams end at places 4-8?
Since we discuss multiple positions here the probabilities grow.

1. AZ
The first team we name is AZ. This team has, according to the simulations we have done, the biggest probability of ending up at one of these spot on the ranking. The first simulations suggest a probability of 0.6924 for this to happen, while based on the second simulations, this probability is 0.6597.

2. FC Utrecht
Another team that has a significant probability of claiming one of these spots is FC Utrecht. They did so 6853 times in the first 10000 simulations. Based on the second 10000 simulations, the probability for ending at one of these places is 0.6388.

3. Vitesse
The team with the 3rd largest probability according to our simulations, of ending at one of these spots is

Vitesse. In the first 10000 simulations, Vitesse took one of these spots 6008 times. In the simulations where the uncertainty was included this happened 5634 times.

Which teams end at places 13-16?
We now arrive at the places on which you do not want to end. The teams that finish the league on one of these places are in danger of relegating from the Eredivisie. There is one team that is a clear favourite for one of these spots, but there are more contenders.

1. De Graafschap
Looking at the winning potentials, based on the real data, De Graafschap was already significantly the weakest team. We saw this back in our simulations. Based on the first simulations, De Graafschap had a probability of 0.9155 of ending up at one of the bottom 3 spots. When we included uncertainty, this still happened 8427 times out of 10000 simulations. Based on our model and our simulations, the prospect for De Graafschap is somber.

2. Willem II
The team with the second highest probability, of ending in the relegation zone, according to the simulations is Willem II. In the first simulations this happened 5152 times out of 10000, while the probability based on the second set of simulations is exactly 0.5. This is a significant probability, although relatively to De Graafschap, Willem II has a better chance of avoiding the relegation zone.

3. VVV Venlo
The third team we should mention here is VVV Venlo. Based on the first simulations, the probability for VVV Venlo of ending in the relegation zone is 0.5041. Based on the second 10000 simulations this probability is almost the same, since they ended in the relegation zone 5043 times.

6

Conclusion and Discussion

In this chapter, the main findings will be summarized and discussed. After the summary, the predictions made in Chapter 5 will be checked. In the discussion, the results are discussed, as well as the limitations of the model. Finally some recommendations for further work are given.

6.1. Summary

The goal of this thesis was to predict the ranking of the Eredivisie 2018-2019, using the Bradley-Terry model. In Chapter 2, the model was introduced. The one thing that made this, relatively simple model, difficult to work with was to determine the maximum likelihood estimates. In order to do this, a difficult system of equations needed to be solved, which consisted of m equations with m variables. This caused us to introduce an MM-algorithm which we used to solve this system and find the maximum likelihood estimates. Some limitations of the model were clear from the beginning. The most important one was that the model was not able to deal with draws in the data, so it would simply ignore them. Because of this we introduced an extension of the model in order to deal with the draws. For the new model, a new MM-algorithm needed to be introduced as well. Another extension which we introduced was an extension that included a parameter that measured the home advantage. For this model, also a new MM-algorithm needed to be introduced. Each algorithm had its own data assumptions. The data should satisfy these assumptions, otherwise the algorithm would not work. A final model was created by combining the extension for draws and for home advantage. The model contained 18 parameters, which were all obtained using, again, a new MM algorithm.

Chapter 3, was about examining the real data. This data consisted of all the matches played, in the past 20 seasons of Eredivisie. Reasoning was done to eventually choose to only use the past 6 seasons to get results and base predictions upon. The data set, that thus consisted of the past 6 seasons was then analyzed. It turned out that the data satisfied the data assumptions of all the MM-algorithms, which meant that the data was usable. A more detailed analysis of the stability of each team was done as well.

In Chapter 4, different scenarios were simulated to get certain types of data. On this data the model would be applied, to give insights on what the effects were of factors like number of matches and number of teams. Also the basic model was compared with some of the extensions, to make sure that the extended model was truly the better option.

In Chapter 5, the model that included draws and home advantage was applied on the real data. A winning potential for each team was obtained, which would be used for the final simulations. Two methods of simulating were used. In the first method, the obtained winning potentials were used. In the second method, some uncertainty was added to these winning potentials. With both methods, 10000 simulations were performed. From every simulation a ranking was obtained, which made it possible to add probabilities to these rankings, based on the number of times a team was ended at that rank in the simulations. Using these probabilities, some predictions about the ranking were made.

6.2. Checking the predictions

The Eredivisie season 2018-2019 for which we did predictions has ended. This means that the predictions that were done, using the extended Bradley-Terry model, can be checked. Table 6.1 shows the true rank of this season.

Rank	Team	Points
1.	Ajax	86
2.	PSV	83
3.	Feyenoord	65
4.	AZ	58
5.	Vitesse	53
6.	FC Utrecht	53
7.	Heracles Almelo	48
8.	FC Groningen	45
9.	ADO Den Haag	45
10.	Willem II	44
11.	SC Heerenveen	41
12.	VVV Venlo	41
13.	PEC Zwolle	39
14.	FC Emmen	38
15.	Fortuna Sittard	34
16.	Excelsior	33
17.	De Graafschap	29
18.	NAC Breda	23

Table 6.1: Real ranking of Season 2018-2019

According to our simulations, Ajax and PSV had the biggest probability of becoming champion. As shown in Table 6.1, Ajax performed slightly better in a very tight race for the title. PSV thus took second place, which was also the team that had the highest probability of doing so, according to our simulations.
Feyenoord ended at rank 3. This made sense, based on the simulations, because Feyenoord had the highest frequency of doing so in the simulations as well.
AZ, Vitesse and FC Utrecht we mentioned as favourites for the places 4-8. They succeeded in doing so. The probability that Heracles would end up at one of these places according to the simulations was about 0.165.
The bottom 3 teams were Excelsior, De Graafschap and NAC Breda. The probability that De Graafschap would end up at these bottom 3 was big according to the simulations: for the model including uncertainty this was still over 0.84. Following from the simulations the probabilities that Excelsior and NAC Breda would end at these spots were around 0.33 and 0.40 respectively.

6.3. Discussion

6.3.1. Results

The results obtained in this thesis may not look very impressive, but this was also not expected. Predicting football is very difficult and this is also the reason why so many people are working on this every day. Consider the fact that we started with the original Bradley-Terry model, which is a model of pairwise-comparison. This model is not very suitable for predicting football to say the least. Ultimately we succeeded in creating a model that can be applied on every football league. By applying this model on a data set of some football league, a person that has no knowledge of this league at all, gets a good indication of the relative strengths of all the teams in this league. There are however lots of factors that are not included in this model. Think about transfers of players, injuries of players, investers of clubs, etc. These factors probably give a person who is closely involved in a league and who keeps track of all the news an edge over people that blindly use rely on this model. There are also factors that we did not take into account, but that we could have. In Chapter 5, to get probabilities for rankings of teams we chose to do two different simulations. In the second one we tried to correct the winning potentials we simulated with for the uncertainty on which we based them. We think that this made the uncertainty recovered from the simulations more realistic. We knew from Chapter 4 that the impact of number of matches was quite significant, so we wanted to include correct for this effect. Something that we did not correct for unfortunately was the stability, which we talked about in Chapter 3. In this chapter we discovered that some teams perform very differently over the years, while others almost got the same number of points every year. We called this effect stability and we saw this as an influence on uncertainty. However we did not have time to come up with a good way to include this effect in our simulations. This is a shame, because this effect could have caused a difference in for example the recoveries of Ajax and PSV, while now they are almost the same. We have tried different ways of including the effect, but ultimately chose to let it out of this thesis, because we don't know what factor stability plays in the uncertainty of the performance of a team. Because of this it was difficult to decide how big a role we would let stability play in the simulations.

6.3.2. Limitations of the model

The model that includes home advantage and draws, already performs way better than the original model on our data. There are however some things which could make the model even better. One limitation is that every match has the same impact on the winning potential. What might be better is to 'weigh' each match relative to the season in which it was played. The matches from the most recent season should then be 'weighed' heavier than matches from older seasons. By correlating the weight with the number of years the season has ended, you should also be able to take more seasons into account, without being afraid that those seasons impact your winning potentials too much.

In Chapter 3, an overview is given of the number of wins at home and away. In this overview there are clear differences in the percentage of home wins, between the teams. In the model we used there is only one parameter that includes a general home advantage, while this overview suggests that a separate home advantage for each team might be more accurate. The number of parameters however will almost be doubled.

Bibliography

[1] http://www.football-data.co.uk/netherlandsm.php. Website, 2018.

[2] Alan Agresti. Models for matched pairs. *Symmetry Models: Categorical Data Analysis. New York: John Wiley & Sons*, pages 409–454, 1990.

[3] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[4] David R Hunter et al. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1): 384–406, 2004.

[5] Richard Pollard. Worldwide regional variations in home advantage in association football. *Journal of sports sciences*, 24(3):231–240, 2006.

[6] PV Rao and Lawrence L Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

[7] Gunther Schauberger, Andreas Groll, and Gerhard Tutz. Analysis of the importance of on-field covariates in the german bundesliga. *Journal of Applied Statistics*, 45(9):1561–1578, 2018. URL `https://EconPapers.repec.org/RePEc:taf:japsta:v:45:y:2018:i:9:p:1561-1578`.

[8] Unknown. Waar ging het allemaal mis met fc twente? *NOS*, 2015.

[9] Stefan Vermeulen. Hoe munsterman en van der laan fc twente ten gronde richtten. *Nieuwe Revu*, 2018.