Soccer Analytics

Predicting the outcome of soccer matches

Research Paper Business Analytics

December 2012

Nivard van Wijk

Supervisor: prof. dr. R.D. van der Mei

VU University Amsterdam
Faculty of Sciences
Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam
the Netherlands

VU UNIVERSITY AMSTERDAM

Soccer Analytics; Predicting the outcome of soccer matches

Nivard van Wijk

Research Paper Business Analytics

Supervised by: prof. dr. R.D. van der Mei

VU University Amsterdam
Faculty of Sciences
Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam
the Netherlands

December 2012

Preface

This research paper is written as part of the Master program Business Analytics at the VU University Amsterdam. The aim of the research paper is to combine all the knowledge built up during the years. Furthermore, it is a preparation for the final Master thesis.

The subject of this research paper is so called Sport Analytics, especially applied to soccer. After reading the book *Moneyball* by Michael Lewis, I became enthusiastic about the opportunities of analytics in sports. Therefore, I decided to write this research paper about this "new" application of analytical tools.

I would like to thank my supervisor prof. dr. R.D. van der Mei for his help during the research. Furthermore, I would like to thank Nick Groen and Michael McAssey Ph.D. for all the discussions we had.

Nivard van Wijk

Amsterdam, December 2012

Summary

Betting on sport matches has become one of the most popular forms of gambling. However, as far as we know, there has not been a lot of research done on the prediction of the outcome of sport matches. Throughout this paper we try to model the outcome of soccer matches in such a way that the results can be used to place a bet on the winner of the match. The main question is:

How can the outcome of soccer matches be predicted?

Based on data analysis of multiple seasons of the English Premier League we conclude that (1) there is a significant home ground advantage, (2) the number of goals scored by the home and away team can be described by a Poisson distribution, (3) there are differences between the teams in the ability of winning points, and (4) the number of goals scored by the home team is correlated with the number of goals scored by the away team.

Based on one or more of these conclusions, two types of models are formulated: (1) toto-models and (2) score-models. The first set consist of models that use only one of the conclusions to predict the outcome of a soccer match. These models are used as a benchmark for the more "complex" score-models. This set consist of three different implementations of the estimation of the parameters.

Based on four seasons of data, we conclude that one of the variants of the score-model (the so-called DE-model) has the highest performance percentage. This model takes into account the average number of goals, the attacking qualities of the team and the defensive skills of the opponent. Furthermore, we can conclude that using a time dependent method for estimating the parameters improves the performance of the model slightly. The biggest advantage of this time dependence is that if a team improves itself over time, the model will be able to represent this improvement in the parameter estimations.

As in only 54% of the matches the toto-outcome is predicted correctly, this raises the question whether it is possible to model the outcome of soccer matches. Soccer has a very low scoring rate which makes it difficult to diversify between a strong team (with an expected number of goals of 2) and a weak team (with an expected number of goals of 1). Because of the Poisson distributed number of goals, both teams will still have a reasonable probability of winning the match.

Contents

Chapter 1

Introduction

During the last decade, betting on sport matches has become more and more popular. Especially since the rise of the Internet, more people are betting on the outcome of sport matches. Because there is a lot of money involved in sport betting it would be very profitable if these outcomes could be predicted with a high certainty.

In Europe, soccer is one of the biggest sports and also one of the sports on which a lot of money is bet. In soccer two teams compete to each other by trying to score the most goals. Most gambling companies give the opportunity to bet on either the final score or just on the winner of the match. Using analytical skills might help predicting the correct winner.

1.1 Aim of this paper

In this paper we try to model the outcome of soccer matches in such a way that it is useful to for instance a gambler who wants to bet on these matches. The main question is:

How can the outcome of soccer matches be predicted?

To answer this question we first ask ourselves what characterizes the score of a soccer match. Next, we ask the question how these characteristics can be used to model the outcome of matches.

1.2 Previous research

In literature, there is a difference between the research aimed at predicting the winner (toto-models) and the research aimed at predicting the score (score-models). The first area is mostly researched by computer scientists, while the second area is researched by economists and econometricians. The two research areas are closely related; when the number of goals scored by both teams in a soccer match can be predicted, the winner can be derived from this result. Therefore, the second type of model, predicting the number of goals scored, seems to be more interesting. Goddard [5] showed that there is little difference in the predictive performance of both types of models, which makes the second type of model even more preferable.

1.2.1 Model the match scores

One of the earliest models that were used to predict the final score of a soccer match was described by Maher [7] in the early 1980s. In this model, the number of goals of both the home and away team are modeled as independent Poisson distributions. The parameters of these distribution are described as a product of an "attack parameter" and a "defense parameter." Each team has a total of four different parameters: an attack and defense for both home and away.

More recently, Rue and Salvesen [8] and Dixon and Coles [2] provided new models based on the model of Maher. Dixon and Coles assume the independence between the number of goals scored by the home and away team. However, they conclude that this does not imply that these goals are

independent of the team-profiles (the parameters for attack and defense). Therefore, they propose to use a joint conditional probability law, which also includes a correction factor for some of the possible scores. Rue and Salvesen use this same function, but extend it even further. They make the assumption that not all information of a match is based on the parameters of the teams playing and therefore form a mixture of laws. Only $100(1-\epsilon)\%$ of the information comes out of the law described by Dixon and Lee, while the other $100\epsilon\%$ is based on the average number of goals in the competition.

Karlis and Ntzoufras [6] used a Bivariate Poisson model in which the number of goals scored by the home and away teams are no longer independent. Because the probability of a draw was underestimated by their model, they proposed an inflated Bivariate Poisson model, in which the probability was corrected for this underestimation.

In all approaches mentioned, the parameters of the model are estimated based on the maximization of the likelihood function of the scores. However, because it is impossible to find these estimations analytically, a numerical method is used.

1.2.2 Model the score process

A completely different approach was described by Dixon and Robinson [3]. Instead of modeling the final score of a soccer match directly, they model the goal scoring process within the match as a Birth Process (i.e., at most one goal at a time interval). They claim that their approach improves the match outcome estimations of the models of Maher and Dixon and Coles. However, because we do not have the necessary data for these kind of models, we will not look further into the time processes within a match.

1.3 Outline

The remainder of this paper is outlined as follows. First, in chapter 2, a dataset filled with multiple seasons of match statistics is analyzed to find interesting facts that could be included in the models for soccer matches. With these observations in mind, we will present different models in chapter 3. Both toto- and score-models are considered and different implementations of these models are described. Based on the analyzed dataset, each model will be tested on its performance in predicting the winner and the correct score of all matches. These results and some analysis of the parameters are described in chapter 4. Finally, in chapter 5 an answer to the research question is formulated and discussed.

Chapter 2

Data analysis

Throughout this paper we use a dataset based on the English Premier League of the seasons 2007-2008 through 2010-2011 [1]. In this chapter we describe some of the characteristics of soccer matches.

Many different statistics are recorded during a soccer match which could all influence the outcome of the match. Which statistics could be used to predict an upcoming match depends on the dataset that is used. To create models that can be used within different soccer competitions and situations, we restrict ourselves to the number of goals scored by both teams and the number of points won in previous matches. Therefore we neglect other interesting statistics such as the number of corners received or the number of fouls committed. A second reason to neglect these statistics is that none of them has a high correlation with the number of goals scored during a match and there is no clear relationship with this number.

The final score of a match consists of a number of goals for both teams playing. Based on the 1520 matches in the dataset, we can conclude that the number of goals scored by the home team is correlated to the number of goals scored by the away team. Pearson's correlation coefficient is -0.0617, which is significantly different from 0 ($\alpha = 0.05$).

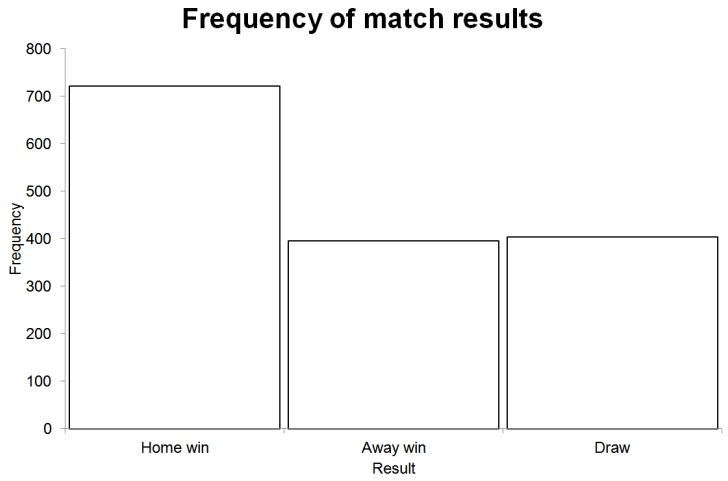


Figure 2.1: The results of 1520 matches in the Premier League from season 2007-2008 until season 2010-2011.

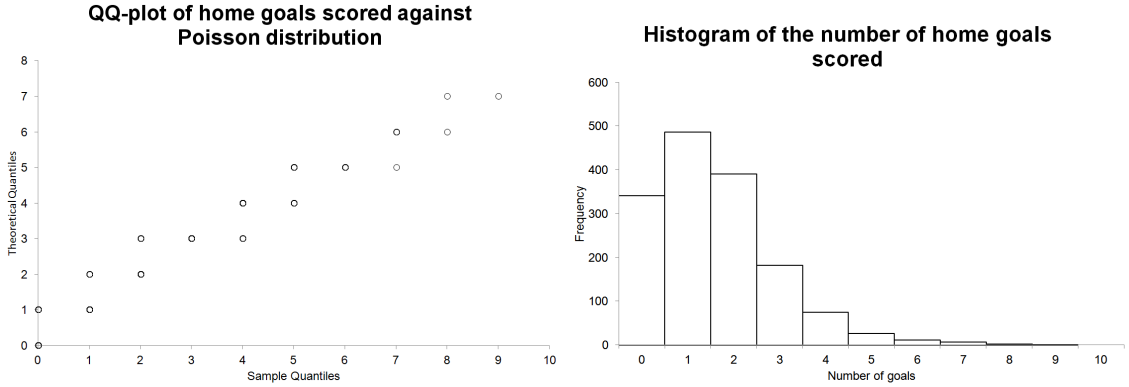
Figure 2.1 is a histogram showing the frequency of each possible result in the 1520 matches played. This figure clearly indicates a home advantage in soccer matches, because almost 50% of the matches is won by the home team. A draw and win for the away team are more or less equally likely to occur, namely in 25% of the matches. The possible existence of home advantage was also found by Clark and Norman [1].

[1] These results are downloaded from: http://football-data.co.uk/englandm.php

Using a Monte Carlo method as proposed by Dwass [4] with 10,000 simulation runs, we can test the hypothesis that there is a significant home advantage in soccer matches. This is done by testing whether or not the mean number of goals scored by the home team μ_{home} could be equal to the mean number of goals scored by the away team μ_{away}.

$$H_0 : \mu_{home} = \mu_{away}$$
$$H_1 : \mu_{home} \neq \mu_{away}$$

This two-sided test results in a test-statistic $\bar{X} - \bar{Y} = 0.4539$, while the 97.5% quantile of the emperical test-statistic distribution based on the simulation using μ_{home} is 0.0868. This means that we can conclude with a confidence of 95% that H_0 should be rejected and therefore the means are different. We conclude that there is indeed a home-advantage in the English Premier League.



(a) QQ plot

(b) Histogram

Figure 2.2: Distribution of the number of goals scored by the home team.

In figure 2.2 both a QQ-plot against the Poisson distribution and a frequency diagram of the number of goals scored by home teams are plotted. The frequency diagram shows that in most cases a total number of 0, 1 or 2 goals are scored, where 1 goal has the highest frequency. In some extreme cases 5 or more goals are scored (which is the case in 2% of the matches). Looking at the distribution of the number of goals scored by the home team, it becomes clear that it seems to follow a Poisson distribution with a mean equal to the average number of goals scored.

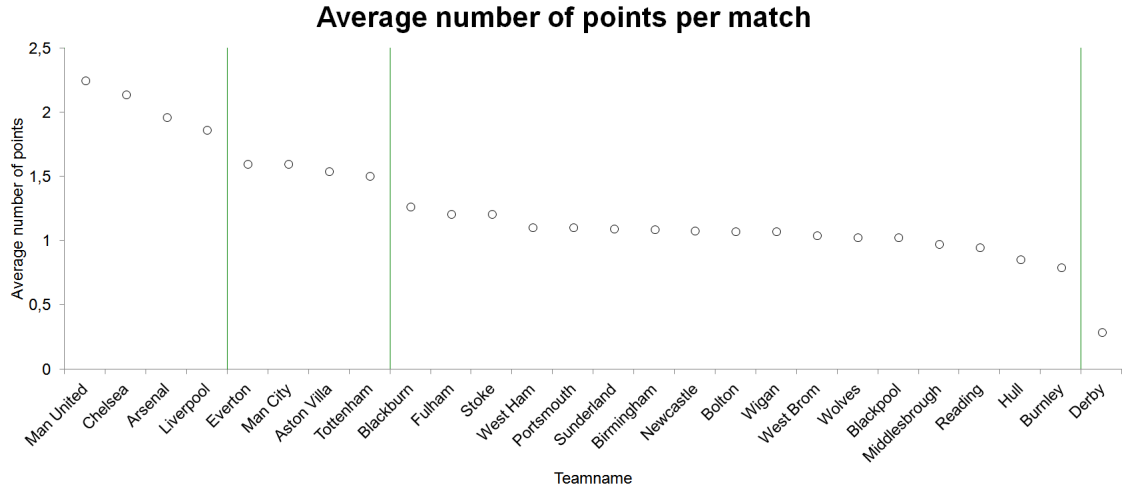


Figure 2.3: The average number of points won per match by each of the teams that played somewhere between 2007 and 2011 in the English Premier League.

When looking at the different teams in the competition, we assume that there are strong and

weak teams. Some teams are capable of winning more matches and therefore will collect more points per match. Figure 2.3 shows that indeed some teams are stronger than others. Especially *Manchester United*, *Arsenal*, *Chelsea* and *Liverpool* (which are the biggest four teams in the competition) seem to be able to win more matches than others. The figure also shows that there are a lot of teams which are more or less equally powered and win an average of 1 point per match (which equals a draw).

Based on the average number of points per match, we could divide all teams into four different categories. The categories are represented by the lines in figure 2.3 and divide the teams in groups of more or less equal strength.

We again use a Monte Carlo method to test the null hypothesis that each team has the same average number of points per match. However, we will not consider the averages of each team, but only the averages of *Manchester United* and *Derby*, which are respectively the team with the highest and lowest averages. If there are teams which averages are significantly different, it would be at least these teams.

$$H_0 : \mu_{ManUnited} = \mu_{Derby}$$
$$H_1 : \mu_{ManUnited} \neq \mu_{Derby}$$

For this test, we consider the test-statistic $Z = z_{Derby} - z_{ManUnited}$ where z_i denotes the average number of points won per match, which under the null hypothesis should be symmetrically distributed around 0. Based on 10,000 simulations of two samples drawn from the distribution found in figure 2.1, we approximate the distribution of the test-statistic. We find that the value of Z based on the observations is 1.9605, which is a lot bigger than the 97.5% quantile of the empirical distribution, for which we found a value of 0.4342. Therefore, we reject the null hypothesis with a 95% certainty and conclude that there is indeed a difference between the teams.

Based on the analysis throughout this chapter, we can draw the following conclusions:

- There is a significant home ground advantage in the English Premier League. Clarke and Norman [1] did also find indications towards this conclusion.
- The number of goals scored by the home and away team can be described by a Poisson distribution. This result is also found by Dixon and Coles [2] and Rue and Salvesen [8].
- In the English Premier League there are differences between the teams in the ability of winning points
- The number of goals scored by the home team is correlated with the number of goals scored by the away team.

Chapter 3

Match models

In this chapter we formulate different models for predicting the outcome of soccer matches. In particular we look at the match between teams A and B where A is playing at home. Different types of models are described, starting with the toto-models. These models do not take anything in account except for one of the characteristics found in chapter 2. Afterwards, we present a model that combines all conclusions from the previous chapter, using different assumptions and therefore different implementations.

3.1 Notation

Throughout this chapter, the following notation will be used for each team $i \in \{A, B\}$.

X number of goals scored by the home team

Y number of goals scored by the away team

T_i team profile of team i; a set of parameters indicating the team quality

a_i parameter for the attacking qualities of team i

d_i parameter for the defensive qualities of team i

3.2 Toto-models

Toto-models can be used as a benchmark for the performance of more complex models. Based on the information from the previous chapter we present two different toto-models. First, one using random probability and second, one groups of teams.

3.2.1 Random probability

When looking at the winner of a soccer match, there are three possible outcomes. Either the home team wins, the away team wins or the match finishes in a draw. Using no other information than this, one could only randomly predict one of those outcomes with probability $\frac{1}{3}$. However, from the data analysis in the previous chapter we know that these outcomes are not equally likely to occur. Based on figure 2.1 we can conclude that better estimates of the probability are $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{4}$ for respectively a home win, away win and draw. Choosing one possible outcome using these probabilities leads to a naive prediction for the match.

3.2.2 Team grouping

Based on figure 2.3 we can assume that all teams in the English Premier League can be divided into four different groups based on the average number of points won. With these groups, we can predict the outcome of a match based on simple rules. A team will always win against a team from a lower group and always lose against a team from a higher group. A match between two teams from the same group will always result in a draw.

3.3 Score models

Using all information from the previous chapter, we are able to define "smarter" models than the toto-models. In this section we will first present a general model and afterwards refine this model under different additional assumptions.

Based on the data analysis in chapter 2 we can assume that random variables X and Y conditioned on the team profiles are both Poisson distributed.

$$\begin{aligned} X|(T_A, T_B) &\sim Poisson(\theta_A) \\ Y|(T_A, T_B) &\sim Poisson(\theta_B) \end{aligned} \tag{3.1}$$

Although the previous chapter showed that the random variables are not independent (because there is a significant correlation between their realizations), we will nevertheless assume that they are independent. There are two reasons for this assumption:

- The previous chapter indicates that the correlation coefficient seems very low, which indicates that the assumption of independence probably has only a slight effect on the model performance.
- It is not clear how X and Y are dependent. This is a problem because there is no known joint distribution for two dependent not identically Poisson distributed random variables.

To predict the winner of the match, we want to be able to calculate the probability of each possible score (x, y), given the strengths and weaknesses of both teams. This distribution is denoted by $\pi(x, y|(T_A, T_B))$ and is defined as:

$$\begin{aligned} \pi(x, y|(T_A, T_B)) : &= \mathbb{P}(X = x, Y = y|(T_A, T_B)) \\ &= \mathbb{P}(X = x|(T_A, T_B))\mathbb{P}(Y = y|(T_A, T_B)) \end{aligned} \tag{3.2}$$

Together, equation (3.1) and (3.2) imply that we can see $\pi(x, y|(T_A, T_B))$ as a product of two independent Poisson distributions. However, it is still unknown how the team profiles influence the distribution function and to find this relationship, additional assumptions are needed.

3.3.1 Multi-Independent model

The first implementation of the general model is the Multi-Independent model (MI). This model assumes that the number of goals scored by either the home or away team is independent of the teams that are playing. That is, the number of goals scored by each team is assumed to be stochastically identical. Therefore, instead of heaving a different team profiles T_A and T_B, we can have one team profile describing both teams.

$$T_A = T_B = T = (a, d)$$

Because both teams are the same, so would be the effect of the team profile on the match. If the effect is always the same, we could as well assume that there is no effect at all. That makes that we could neglect the parameter a and d. Now only the fact whether the team is playing at home or not is influencing the expected number of goals scored.

$$\begin{aligned} \theta_A &= \lambda \\ \theta_B &= \mu \end{aligned}$$

To estimate the values of these parameters, we will use the Least-Square Estimator (LSE). In this function we use the fact that over time, each team plays multiple matches against other teams. The number of matches where team A played at home against team B is denoted by K_{AB}. For the specific match between team A and team B we need to consider all matches played in the competition, because all these matches influences the estimations of the team parameters. We denote the total number of teams in the competition by N.

$$LSE = \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K_{ij}} [(X_{ijk} - \theta_i)^2 + (Y_{ijk} - \theta_j)^2] \tag{3.3}$$

Using the LSE we can find estimations for λ and μ by differentiation with respect to each parameter and setting these equations equal to zero. In this way the estimations[1] will be such that the function in (3.3) is minimized.

$$\hat{\lambda} := \overline{X_{...}}$$
$$\hat{\mu} := \overline{Y_{...}}$$

Using these estimations in equation (3.2) results in:

$$\pi(x, y|(T_A, T_B)) = \mathbb{P}(X = x|\theta_A = \lambda)\mathbb{P}(Y = y|\theta_B = \mu) = \frac{\hat{\lambda}^x \hat{\mu}^y}{x!y!} \mathrm{e}^{-(\hat{\lambda}+\hat{\mu})}$$

This leaves a fairly simple model were each match has exactly the same prediction. Of course this is unlikely to be the best model, but we can use the results of this model in the same way as the results of the toto-models.

3.3.2 Single-Independent model

The Single-Independent model (SI) does not require the same strong assumption that the MI does. In this model, the parameters of each team can be different, and so each team has its own team profile. However, we still assume that the number of goals scored by the home team are independent of the away team and the other way around. Regardless of the opponent, each match will result in the same prediction for the number of goals scored. Therefore, the expected number of goals scored only depends on the attacking capability of the team and whether it is playing at its home ground. We will define a_A and a_B in such a way that their value makes sense. We first need to define λ and μ as the average number of goals scored by all home and away teams.

$$T_A = (a_A)$$
$$T_B = (a_B)$$
$$\theta_A = \lambda + a_A$$
$$\theta_B = \mu + a_B$$

In this way, a_A is the average number of goals scored thanks to the capability of team A, in addition to the average number of goals that is scored at home by all teams. For team B a same explanation holds, where it is an addition to the average number of goals scored away. We assume that $\sum_{i=1}^{N} a_i = 0$ and $K_{ij} = K_{ji}$ for each pair (i, j). That is, we assume that the sum over all attack parameters is equal to 0. Furthermore, we assume that the number of matches between teams i and j where team i played home is equal to the number of matches where team j played at home. In a competition where each team plays home once against each opponent, this assumption would make sense. With these assumptions, we get a unique solution to the set of equations that come out of the LSE:

$$\hat{\lambda} = \overline{X_{...}}$$
$$\hat{\mu} = \overline{Y_{...}}$$
$$\hat{a}_A = \frac{\overline{(X_{A..}} - \overline{X_{...})} + (\overline{Y_{.i.}} - \overline{Y_{...}})}{2}$$
$$\hat{a}_B = \frac{\overline{(X_{B..}} - \overline{X_{...})} + (\overline{Y_{.i.}} - \overline{Y_{...}})}{2}$$

[1] The dots define the index over which the average is taken, i.e.:

$$\overline{X_{...}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{N}\frac{1}{K_{ij}}\sum_{k=1}^{K_{ij}} X_{ijk}$$
$$\overline{X_{i..}} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{K_{ij}}\sum_{k=1}^{K_{ij}} X_{ijk}$$

With these estimations, equation (3.2) becomes:

$$\begin{aligned}\pi(x, y|(T_A, T_B)) &= \mathbb{P}(X = x|\theta_A = \lambda + a_A)\mathbb{P}(Y = y|\theta_B = \mu + a_B) \\ &= \frac{(\hat{\lambda} + \hat{a}_A)^x(\hat{\mu} + \hat{a}_B)^y}{x!y!}\mathrm{e}^{-(\hat{\lambda}+\hat{a}_A+\hat{\mu}+\hat{a}_B)}\end{aligned}$$

3.3.3 Dependent model

In the third and last model there is a dependence between the teams that are playing and the number of goals that will be scored, called the Dependent model (DE). However, there still is no dependence between the number of goals scored by the home and away team, as was assumed for the general model.

The dependence between the number of goals scored by one team and the opponent is reflected by the parameter for the defensive qualities of the opponent. This parameter either lowers the expected number of goals, when the opponent has a strong defense, or raises the expectation, when the defense is less than average. The team profile now consist of two parameters, one for the attack and one for defense:

$$\begin{aligned}T_A &= (a_A, d_A) \\ T_B &= (a_B, d_B) \\ \theta_A &= \lambda + a_A - d_B \\ \theta_B &= \mu + a_B - d_A\end{aligned}$$

Again we need some additional constraint on the factors in order to get a unique solution out of the set of partial derivative equations of the LSE. We again assume that $\sum_{i=1}^{N} a_i = 0$ and $K_{ij} = K_{ji}$ for each pair (i, j). For the DE model we need the additional assumption that $\sum_{i=1}^{N} d_i = 0$. With these constraints and the LSE, we can derive the following estimations for the parameters:

$$\begin{aligned}\hat{\lambda} &= \overline{X_{...}} \\ \hat{\mu} &= \overline{Y_{...}} \\ \hat{a}_A &= \frac{(\overline{X_{A..}} - \overline{X_{...}}) + (\overline{Y_{.A.}} - \overline{Y_{...}})}{2} \\ \hat{d}_A &= \frac{(\overline{X_{.A.}} - \overline{X_{...}}) + (\overline{Y_{A..}} - \overline{Y_{...}})}{2} \\ \hat{a}_B &= \frac{(\overline{X_{B..}} - \overline{X_{...}}) + (\overline{Y_{.B.}} - \overline{Y_{...}})}{2} \\ \hat{d}_B &= \frac{(\overline{X_{.B.}} - \overline{X_{...}}) + (\overline{Y_{B..}} - \overline{Y_{...}})}{2}\end{aligned}$$

When these values are put together with equation (3.2), we get the following distribution function:

$$\begin{aligned}\pi(x, y|(T_A, T_B)) &= \mathbb{P}(X = x|\theta_A = \lambda + a_A - d_B)\mathbb{P}(Y = y|\theta_B = \mu + a_B - d_A) \\ &= \frac{(\hat{\lambda} + \hat{a}_A - \hat{d}_B)^x(\hat{\mu} + \hat{a}_B - \hat{d}_A)^y}{x!y!}\mathrm{e}^{-(\hat{\lambda}+\hat{a}_A-\hat{d}_B+\hat{\mu}+\hat{a}_B-\hat{d}_A)}\end{aligned}$$

3.4 Pseudo Least-Square Estimator

One could argue that teams develop their skills over time in a way such that matches played for instance three years ago do not have much information about the attack and defense qualities now. The current estimators for the team-performance parameters do not account for this: a match played years ago has the same informative value as a match played yesterday.

However, using Pseudo Least-Square estimators, this problem can be fixed. Instead of using the Least-Square function defined in (3.3), we could use the following function:

$$PLSE = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K_{ij}}[(X_{ijk} - \theta_i)^2 + (Y_{ijk} - \theta_j)^2]\mathrm{e}^{-\beta(T-t_k)} \tag{3.4}$$

Here T stands for the current time period (i.e. in weeks calculated from the start of the dataset) and t_k represents the time period in which the k'th match between team i and j is played. The parameter $\beta \in [0, \infty)$ is the rate at which information from the past should be discounted. When β is set to 0, the PLSE will be equal to the LSE, because the information will not be discounted at all.

The PLSE has one drawback in the way that the optimal value of β can not be determined by taking the derivative with respect to this parameter. Therefore, the optimal value of this parameter has to be found iteratively.

Chapter 4

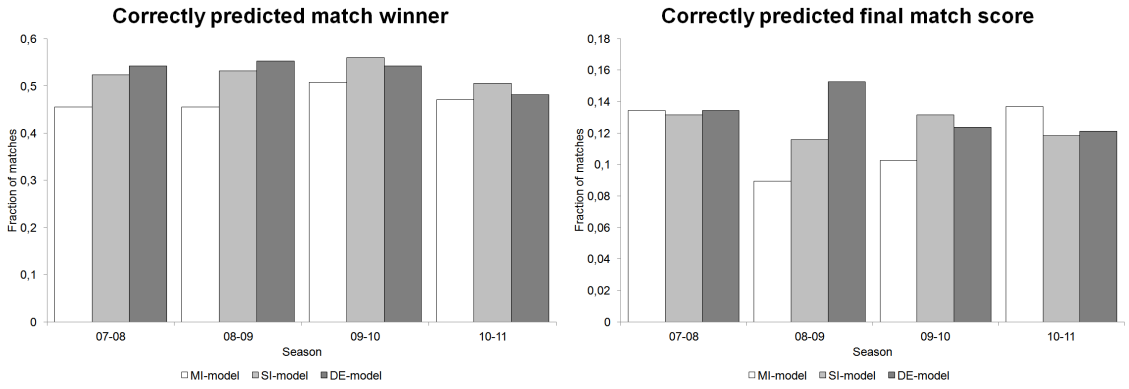
Results

With the data of all soccer matches played during the seasons 2007 until 2011 in the English Premier Leauge, we are able to test the models presented in chapter 3. The results of these tests will be presented in the following order: we will start by looking at the different score-models presented, using the LSE. Afterwards, we will look at the advantage of adding a time effect in the model parameters by using the PLSE. At the end of this chapter, a comparison between the toto- and score-models is made.

4.1 Score-model choice using LSE method

In chapter 3, three different score-models were presented, the MI-, SI- and DE-model. In the MI-model, the outcome of the match is independent of the teams that are playing. The SI-model is an extension to this model, were the parameter of the Poisson distribution is corrected for the team that plays. In the DE-model again an extension is added, were the expected number of goals to be scored by a team is also corrected for the opponent.

Based on the fact that the DE-model contains the most information, one would expect this model to outperform the other models. To see whether this is the case, we have used the models to predict both the score and the winner of each match in our dataset. For the estimations of the parameters, we have used the fact that all information about the matches on previous dates are known. For example, when predicting a match played on 20-10-2007, all matches played before this date could be used in the estimations. This means that more information is used (based on the LSE) when time progresses.



(a) Winner prediction

(b) Score prediction

Figure 4.1: Fraction of matches for which the predictions are correct in each of the four seasons.

In figure 4.1 the performance of each of the three models is stated. Figure 4.1a shows what fraction of matches in each of the four seasons has a correctly predicted winner. It becomes clear

that the MI-model performs not as good as the SI- and DE-model. These last two models have a more or less equal performance, where the DE-model outperforms the SI-model in the first two seasons and the SI-model outperforms the DE-model in the last two seasons. Both models will predict the winner correctly in over 50% of the matches during a season. We still prefer the DE-model over the SI-model, because this model does hold more information.

Figure 4.1b shows the performance of the models in predicting the correct score of a match. From this figure it can be concluded that there is no model that outperforms all others. Each model has at least one season were it performs better than the other two. Furthermore, the figure shows that in none of the four seasons more than 15% of the matches had a correctly predicted score (where both X and Y are predicted correctly). Therefore we will no longer look at the score prediction and focus on the winner prediction.

4.2 Optimizing PLSE method for DE model

Using the DE-model and the fraction of matches where the winner is predicted correctly, we will try to find a β value for which the PLSE performs best.

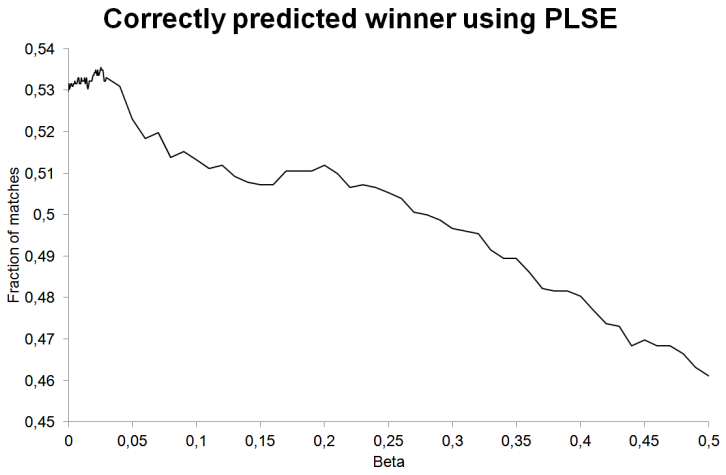


Figure 4.2: Fraction of matches with a correctly predicted match winner under the DE-model over all four seasons against the value of β

In figure 4.2, the performance of the model in all four seasons is plotted against the β-value used in the PLSE. A value of 0 equals the LSE method and the figure shows that this method has a good performance. Using a small β-value (i.e. $\beta = 0.02$) gives a slightly better performance. However, less than a fraction of 0,01 is added to the performance indicator. When the value of β grows, the performance of the model will decline.

The optimal value of β found under the DE-model is 0.0255. Using this β, the correct winner was predicted in 53.55% of the matches, which is a slightly better result than the 52.96% that was correctly predicted using the LSE method. An intuitive explanation why the optimal β-value is found around 0 is given in the next section.

4.3 Parameter analysis

One would assume that the parameter estimations do converge to a certain value over time. This is due to the fact that most teams stay either strong or weak during a four seasons time period. It seems unlikely that a certain team will lose most of its matches during the 07-08 season and wins most of its matches during the 10-11 season. However, when this happens, it should be possible to be reflected in the parameter estimations. Therefore, the value of the parameters are likely to converge to a certain value, but fluctuate around this value rather than really converge.

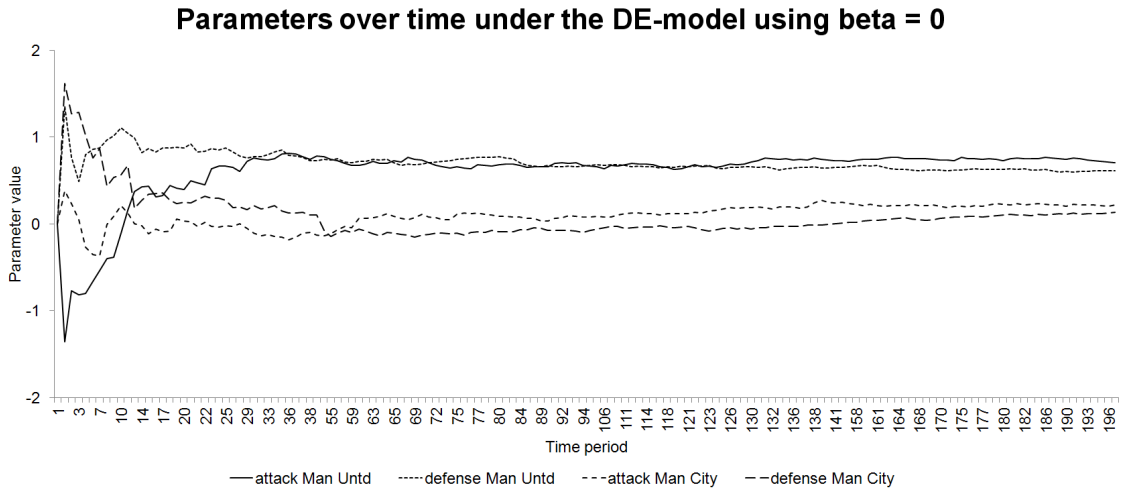


Figure 4.3: Parameter estimations of two teams in the English Premier League over time using the DE-model and the PLSE method with $\beta = 0$.

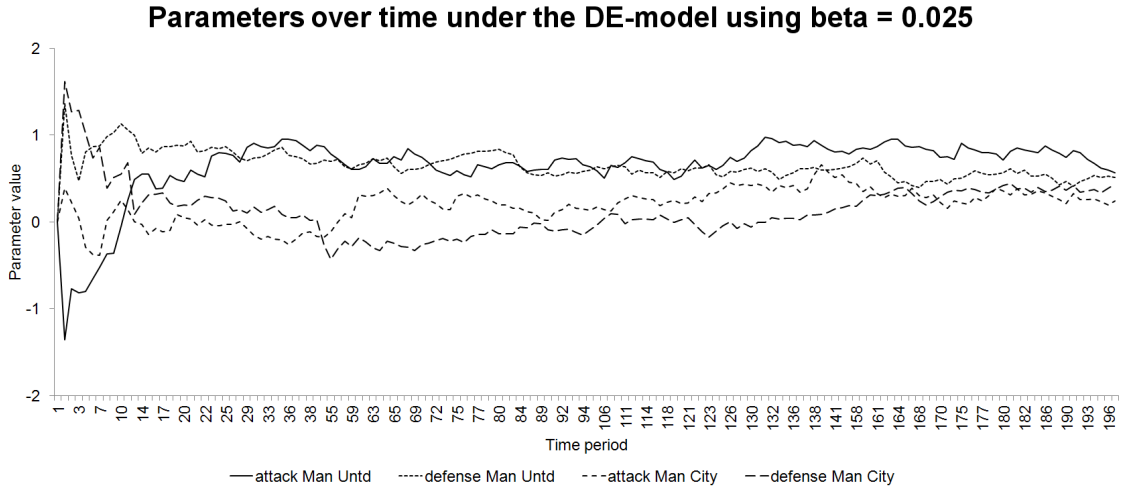


Figure 4.4: Parameter estimations of two teams in the English Premier League over time using the DE-model and the PLSE method with $\beta = 0.0255$.

Figures 4.3 and 4.4 show the values of both the attack and defense parameter of two teams in the English Premier League over time (where each period is a new week). The graphs show that for the estimations of *Man United* it takes more or less 25 weeks to converge to a certain value. Still over time, we see in figure 4.4 these estimations fluctuating within a certain region. This is due to the fact that the value of β is different from 0 and therefore the last matches have a (slightly) higher influence on the estimation values. In figure 4.3 there is less fluctuation.

The *Man City* squad has improved itself during the period drawn in figures 4.3 and 4.4. This team finished 9th in the 07-08 season and 3rd in the 10-11 season. This is a special case which shows why a value of β different from 0 is preferred. In figure 4.4 the estimation of the parameters also shows the teams improvement, because there is a slightly upwards trend in the lines belonging to *Man City*. However, figure 4.3 does not show this trend because "old" information is not discounted and therefore the parameters are not representing the current situation well enough.

4.4 Model comparison

To evaluate the performance of the score-models, we will benchmark them against the toto-models presented in chapter 3. We have already seen that using more information within the score-models results in a higher predictive performance of the model. Therefore we expect the toto-models to have less predictive power than the DE-model using PLSE.

Table 4.1: Comparison between the score and toto-models. The performance statistic equals the percentage of correct predicted match outcomes.

Model	Additional details	Performance
Random chance	$p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ 5000 simulations	36.77%
Team grouping		47.37%
MI model using LSE		47.24%
SI model using LSE		53.03%
DE model using LSE		52.96%
DE model using PLSE	$\beta = 0.0255$	53.55%

Table 4.1 shows the performance metric of each of the toto- and score-models and the parameter settings used to get these results. As expected, the DE-model using the PLSE method with $\beta = 0.0255$ has the best performance. Moreover, its performance is better than the performance of both toto-models. This indicates that using more information gives a better model performance.

Chapter 5

Conclusion

Throughout this paper we looked at different models which could be used to predict the outcome of soccer matches. The main question we tried to answer was: *How can the outcome of soccer matches be predicted?* The best performing model presented is this paper is the DE-model using a PLSE method. This model outperforms all other score-models as well as the toto-models, which we used as a benchmark. Furthermore, this model has a big advantage over other model from previous research The parameters of the DE-model are estimated directly from the dataset, while the other models require a numerical method to find the estimations.

The advantage of using a Pseudo Least-Squares estimator over using a normal Least-Squares estimator is that the first one takes time into account. Therefore, this method is able to make a better estimation of the parameters values, if the strength of a team chances.

5.1 Discussion

In order to generalize the performance of the DE-model, we used this model on a different dataset containing all international FIFA matches played between 2002 and may 2012. This dataset contains over 9000 matches between 208 different nations. The performance of the DE-model is more or less equal to the performance on the Premier League dataset; in both cases the correct winner is predicted in more than 54% of the matches.

Returning to the main question, we can conclude that the model described in this paper models the outcomes of soccer matches. However, the model slightly outperforms a "foolish" strategy of always predicting a win for the home team (MI-model). One could argue that it is not possible at all to find a good model for soccer matches. Because the score is always low in relation to for instance the score in a basketball match and can be described by a Poisson distribution, there is a high probability for different number of goals. For example, when the expected number of goals is equal to 1.5, there is a high chance that only 1 goal is scored. However, this probability is almost equal to the probability of 2 goals to be scored. Also the numbers 0, 3 and even 4 have a realistic chance. This makes that there is little difference between the best and the worst team in competition, which means that there is a reasonable probability that the worst team wins from the best team.

5.2 Further research

Future research should look at this hypothesis and try to find a better way to estimate the expected number of goals. Moreover, we have modeled the home and away score throughout this paper as two independent Poisson distributions. However, we know that these variables have significant correlation and therefore should not be independent. A solution to this problem, by means of a probability distribution, needs to be found.

Bibliography

[1] Clarke, S.R. and J.M. Norman: *Home ground advantage of individual clubs in english soccer.* The Statistician, 44(4):509–521, 1995.

[2] Dixon, M.J. and S.G. Coles: *Modelling association football scores and inefficiencies in the football betting market.* Appl. Statist., 46(2):265–280, 1997.

[3] Dixon, M.J. and M.E. Robinson: *A birth process model for association football matches.* The Statistician, 47(3):523–538, 1998.

[4] Dwass, M.: *Modified randomization tests for nonparametric hypotheses.* Anals of Mathematical Statistics, 28:181–187, 1957.

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

[6] Karlis, D. and I. Ntzoufras: *Analysis of sports data by using bivariate poisson models.* The Statistician, 52(3):381–393, 2003.

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

[8] Rue, H. and Ø. Salvesen: *Prediction and retrospective analysis of soccer matches in a league.* The Statistician, 49:399–418, 2000.