Studying some aspects of data from the professional Dutch soccer competition:

Is an artificial turf field advantageous for home performances.

By Dieuwe van Bergen en Henegouwen

December, 2014

Studying some aspects of data from the professional Dutch soccer competition:

Is an artificial turf field advantageous for home advantages?

Research Paper

Dieuwe van Bergen en Henegouwen
2513239
Dr. E.N. Belitser
Business analytics
VU University Amsterdam
Version 2
December 2014

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

Preface

This paper is written as part of the Master program Business Analytics at the VU University Amsterdam. The aim of the research paper is to use all the learned knowledge and write a paper about a self chosen topic. Furthermore, this paper is written as a preparation for the Master thesis.

The subject of this paper is about the influence of artificial turf field on the outcome of soccer matches. I chose this subject because of the growing amount of Dutch soccer clubs with an artificial turf field. Furthermore, I was curious whether a club with an artificial turf field have some advantage by playing on that type of field.

After following the course 'Statistical models' I asked Dr. E.N. Belitser to be my supervisor. I would like to thank Dr. E.N. Belitser for all the help and discussions we had during the research. Furthermore, I would like to thank Bertus Talsema from Ortec Sports to provide me their data.

Dieuwe van Bergen en Henegouwen

Amsterdam, December 2014

Summary

During the last years more and more professional Dutch soccer clubs impose an artificial turf field. In 2003 Heracles Almelo was the first club with an artificial turf field. Nowadays 50% of all the professional Dutch soccer clubs play on an artificial turf field because the maintenance costs for an artificial turf field are lower.

The aim of this paper is to figure out whether playing on an artificial turf field have additional home advantage. Therefore, the main question is:

Has playing soccer on an artificial turf field influence on the match outcome?

To answer the main question the following question has to be answered: is there a difference between passes on an artificial turf field and ordinary grass?

For this research all the matches of both professional Dutch soccer competitions during 2005 and 2014 are used. Based on the data analyses the data shows that there is no difference in the amount of earned points in the home and away matches at both types of surfaces. On the other hand the average amount of passes differs when the home playing team plays on artificial turf. A team that plays at home on an artificial turf field passes significantly more than a team that plays at home on ordinary grass. This information shows that it is easier passing on artificial turf field.

The goals difference of a single match is modeled as a Skellam distributed response variable. The factors type of surface, the division and the strength and shape difference of both teams and will been tested whether they have influence on a match outcome.

The Kruskal-Wallis method concludes that (1) there is a significant difference between the goal difference of a single match and the type of surface. (2) While there is a highly significant difference between the goal difference and strength or shape difference between the teams. It suggests that playing soccer on an artificial turf filed has influence on the match result.

By use of the Generalize Linear Model (GLM) the dataset is divided in two smaller datasets. The first dataset consists only positive goal differences and the second only the absolute values of the negative goal differences. Both of these data sets are not Poisson distributed, but because of there is no statistical model for the Skellam distribution the Poisson regression model is used to find some interaction between several parameters.

The GLM shows that (3) the strength difference of the teams has biggest influence on the goal difference of a single match. Thereby, the shape difference and types of surface has no influence on the outcome of the match when a team is significant stronger. But when two equal teams plays against each other it can be concluded that (4) the home playing team with an artificial turf field has advantage by playing a match on an artificial turf field.

The conclusion of this paper is that whether two teams with equal qualities plays against each other the home playing team with an artificial turf field has advantage by playing on an artificial turf field. But when two teams with different qualities plays against each other the type of surface has no influence on the outcome of the match.

Contents

Preface **3**

Summary **4**

Contents **5**

Introduction **6**

1. Data analysis **8**

1.1Match results 8

1.2Passes 9

2. Methods **11**

2.1 Variables 11

2.1.1 Response variable 11

2.1.2 Skellam distribution 13

2.1.3 Explanatory variables 16

2.2 Models **16**

2.2.1 Kruskal-Wallis test 17

2.2.2 Generalized Linear Models 17

3. Results **19**

3.1 Kruskal-Wallis test 19

3.2 Generalized Linear Models 19

4. Conclusion **22**

4.1 Further research 23

References **24**

Introduction

24 of May 2008, after regular and extra time the score was 1-1. John Terry the captain of Chelsea has to score the last penalty and Chelsea would win the Champions league 2007-2008. It all happens on the artificial turf field of the Luzhnike stadium in Moscow. At the moment that John Terry should become a Chelsea hero he slipped, Edwin van der Sar dived in the wrong direction, but the ball hits the post. Chelsea lost and Manchester United is the first team ever that won a Champions league title on an artificial turf field.

In the season 2003-2004 Heracles Almelo was the first professional Dutch soccer club with an artificial turf field. Nowadays 19 of the 38 professional Dutch soccer clubs have artificial turf. On the other hand in the five largest soccer competitions of Europe (German Bundesliga, Italian Serie A, English Premier league, Spanish Primera division and French Ligue 1) there are only two other clubs with an artificial field. While in countries like Switzerland, Austria Russia and north Europe more and more clubs have an artificial turf field. They have an artificial field to play in every weather condition. While in the Dutch competition most of the teams have an artificial field because of the lower maintenance costs.

During the last decade there are a lot of papers written about artificial turf fields. Within these years the quality of the artificial fields improved. One of the first was Winterbottom [1985][1] he compared how the ball rolls and bounced on the first-generation artificial fields and ordinary grass. Winterbottom [1985][1] figured out that it is more difficult to move on the first generation artificial turf field. With that in mind Vorstenbosch, Staal, Kolenburg and Meijer (2008)[3] and Steffen, Andersen and Bahr (2007)[4] investigated the risk of injuries on artificial turf field compared with ordinary grass. They revealed that there is no difference in the injury risk between ordinary grass and third generation artificial turf fields.

Since 2005 the FIFA and EUFA allowed artificial fields in European competitions. Commissioned by the FIFA ProZone [2006][2] wrote a technical study about the impact of an artificial turf field. They concluded that there is no significant difference between both types of surfaces, but they studied only two matches in the European league.

In the national soccer competitions all teams plays at least twice per year against every component. They play one time at home and the other time away. Because of this fact there is a lot of research about the influence of home advantage. Dowie [1982][5] was one of first that wrote about home advantage, he also did research about the possible reasons why home advantage can exists. After Dowie [1982][5] there were papers about the influence of crowed effects, travel effects and psychological factors of home advantage. Pollard [1986][6] did research to home advantage in professional team sports. Pollard [1986][6] found that the home advantage in soccer was greatest compared with other team sports. Nevill, Newell and Gale [1996][7] did research to the significant difference of the home advantage between clubs with small and big crowed. In 'Home ground advantage of individual clubs in England soccer' Clarck and Norman [1995][8] develop a model to calculate the amount of home advantage for each team. This model can also calculate the strength of each individual team.

In 1981 Queens Park Rangers was the first professional soccer club with an artificial turf field. In 1988 the English Football Association banned the artificial turf fields. The FA banned this type of surface because of possible competition distortion. The quality of the artificial field that time was not as high as it is right now. The bounce of the ball was difference and there was no possibility to slide. During that period Barnett and Hilditch [1993][9] did research and figured out that there was a significant difference between playing on the first generation artificial turf field and ordinary grass.

In this paper the connection between the match outcome and the type of surface will be made. The aim of this paper is to figure out whether playing on an artificial turf field has additional home advantage. Therefore, the main question is:

Has playing soccer on an artificial turf field influence on the match outcome?

Also determined is the pass difference between playing on artificial turf and ordinary grass (2).

The goal of this paper is to figure out whether it is still competition distortion by playing on artificial turf field. In earlier research is found that the quality of the third generation artificial field is even high as ordinary grass. Therefore, it is better to compare whether having an artificial turf field leads to competition distortion than Barnett and Hilditch [1993][9] did. It is interesting to see if there is a difference between passes on both types of surfaces. With that information the best way of playing on an artificial turf field will be found. It sounds logical that the amount of passes at an artificial turf field will increase because the ball rolls faster and is less fluctuated on artificial turf field. In chapter 'Data Analysis' this hypotheses will be tested and checked whether this assumptions are correct.

For this paper Ortec Sports provide the data. Ortec Sports develops computational software to analyze the performance in different sports. With this software Ortec Sports tries to increase the level of athletes. During different sports matches Ortec Sports measure the performance of all the players. Ortec Sports will deliver data of all the match result of botch Dutch professional soccer competitions. Also they deliver the pass statistics of all matches in the seasons 2012-2013 and 2013-2014.

This paper is divided in three chapters. First the dataset will been analyzed and the quality of the data will been tested. In this part the answer to the second research question will been found. In the chapter 'methods' the distribution of the variables will be researched. Thereby, the models that will be used are described. In chapter 'results' the results from the used models will be treated. These results become from the methods that are explained in 'methods'. Finally, in chapter 'conclusion' the research question will been answered and discussed.

1. Data analysis

The dataset for this paper consist of all the match result of both professional Dutch soccer competitions between 2005 and 2014. Between these years there are 5895 matches played in both professional Dutch soccer competitions. Within these years there was at least one club with an artificial turf field. The dataset consist all the amount of passes during each match of the seasons 2012-2013 and 2013-2014. This chapter gives a first impression about the data. First the match results are investigated for some interesting facts. In the second part of this chapter the statistics of the passes will be analyzed and tested.

In 2003 Heracles Almelo was the first professional Dutch soccer club with an artificial turf field. After Heracles Almelo more and more clubs decide to play on artificial turf. In the season 2005-2006 Cambuur Leeuwarden imposed artificial turf. But after the season 2007-2008 the club decides to change their surface back to ordinary grass. Cambuur Leeuwarden was not satisfied with the quality of the field. In the season 2013-2014 Cambuur Leeuwarden went back to artificial turf.

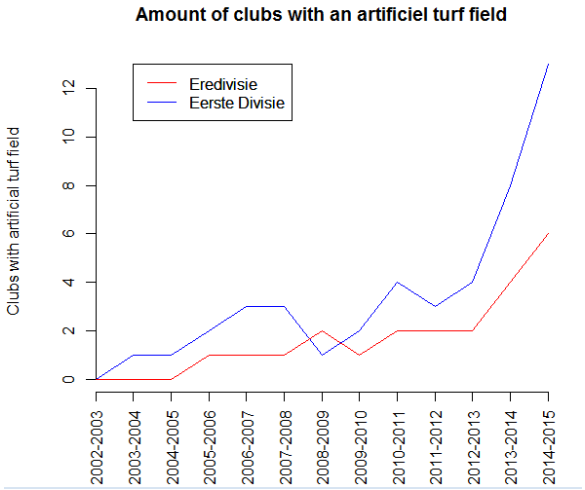


Figure 1.1: *The amount of clubs that plays on an artificial turf field in both professional Dutch soccer competitions during the seasons 2003 and 2015.*

1.1 Match results

To model the advantage of an artificial turf field there are several interesting statistical features. For example the average point achieved on both types of surfaces or the ratio of matches won compared in both types of surfaces. By considering the advantage of having artificial turf it is hard to model the difference in strength. In this part of the research the difference between the strength of both teams is not yet used. How this difference in strength is modeled is described in chapter 'methods'.

The ratio of home points achieved by teams with artificial turf is a good performance measure to indicate whether teams with artificial turf have advantage on the clubs with ordinary grass. Table 1.1 shows that there is now advantage for the home team with artificial turf. Despite the clubs with ordinary grass earned on average more points than the clubs on artificial turf. But the clubs in the first professional Dutch soccer division shows an advantage of 66% for clubs with an artificial turf field and 61% of clubs with a ordinary grass field. In the second division is the ratio of points for clubs with artificial turf only 54% while the ratio of points for teams with ordinary grass is 59%. This indicates that there is a difference between both divisions.

	Home	Away	Total
Artificial field	1225 (57,08%)	921 (42,92%)	2146
Ordinary grass	9746 (60,03%)	6489 (39,97%)	16235
Total points	10518	7431	18381

Table 1.1: *The amount of earned points on both types of surfaces in the both professional soccer competitions between 2005 and 2014.*

Another interesting performance measure is the ratio of matches won on both types of surfaces. Table 1.2 shows the amount of times a match ended in a victory, defeat or a draw at both types of surfaces. The amount of times a match ends in a draw is almost equal on both surfaces. In table 1.1 and 1.2 shows no additional home advantage for teams that play on artificial turf. For instance, the strength of the difference teams can be an explanation of these facts. The models that are described in chapter 'methods' uses a variable strength to model this problem.

	home				Away			
	Win	Draw	Lose	total	Win	Draw	Lose	total
Artificial field	345 (42,59%)	190 (23,46%)	275 (33,95%)	810	237 (24,79%)	210 (21,97%)	509 (53,24%)	956
Ordinary grass	2783 (47,61%)	1397 (23,9%)	1666 (28,5%)	5846	1704 (19,4%)	1377 (15,68%)	5701 (64,92%)	8782
Total points	3128	1587	1941	6656	1941	1587	6210	9738

Table 1.2: *The match outcomes of home and away matches at both types of surfaces in the both professional Dutch soccer competitions between 2005 and 2014*

1.2 Passes

The second research question of this paper is about the amount of passes on both types of surfaces. The dataset consist of the total amount of passes, the long passes, short passes and amount of turnovers during the seasons 2012-2013 and 2013-2014. An additional feature is the percentage of pass completion this feature is calculated by

$$percentage\ of\ passcompletion = 1 - \frac{turnovers}{total\ passes}.$$

The dataset is divided in three parts; the information about the home playing teams, the away playing team and the information of both together. These three parts are also divided in the matches that are played on ordinary grass and on artificial turf.

First the total statistics of both teams will be handled. The average of total passes at both types of surfaces is equal (841 vs. 841). There is a significant difference (p-value = 0.0011) between the amount of long passes between both types the surfaces. On average there are more long passes during the matches on artificial turf than on ordinary grass (122 vs. 111).

	Artificial turf field	Ordinary grass
Total passes	841	841
Short passes	719	730
Long passes	122	111
pass completions	77%	76%

Table 1.3: *The average amount of ball skills for each match during the seasons 2012-2013 and 2013-2014 in both Dutch soccer competitions.*

One of the questions was to figure out whether the ball rolls faster on an artificial turf field. This question is tested by the amount of total passes during a match on both types of surfaces. The amount of passes on artificial turf are significant (p-value = 0.004) higher than on ordinary grass. The percentage of pass completion on artificial turf is significant (p-value = 0.0003) higher in contrast with ordinary grass. Which sounds logical because of the ball will role faster on a flat and less fluctuated field.

	Artificial turf field	Ordinary grass
Total passes	484	437
Short passes	418	380
Long passes	65	57
pass completions	79%	75%

Table 1.4: *The average amount of ball skills of the home playing team for each match during the seasons 2012-2013 and 2013-2014 in the both Dutch soccer competitions.*

Conclusions about chapter 'data analysis'

- During the years the amount of clubs with an artificial turf fields increases.
- The amount of earned points is higher for the teams that played on ordinary grass compared with an artificial turf.
- There is a significant difference between the amount passes during the different matches on artificial turf.

2. Methods

This section of this paper introduces the methods that are used to analyze the data. Therefore, there are some statistical models used for the statistical analyses. A statistical model needs a response variable and some explanatory variables (sometimes called factors). In the first section of this chapter the response and explanatory variables will be introduced. The second part describes the different statistical models. These models describe how one or more variables are related to the response variable.

2.1 Variables

Given a random pair (X, Y), with Y is a set of observations from a one-dimensional vector and the X is a set of independent variables. These observations Y_i and independent variables X_{ij} is formulated by

$$Y_i = \beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \ldots + \beta_p(X_{ip}) + \varepsilon_i$$

With ε_i is the i^{th} normal distributed random error and p different factors.

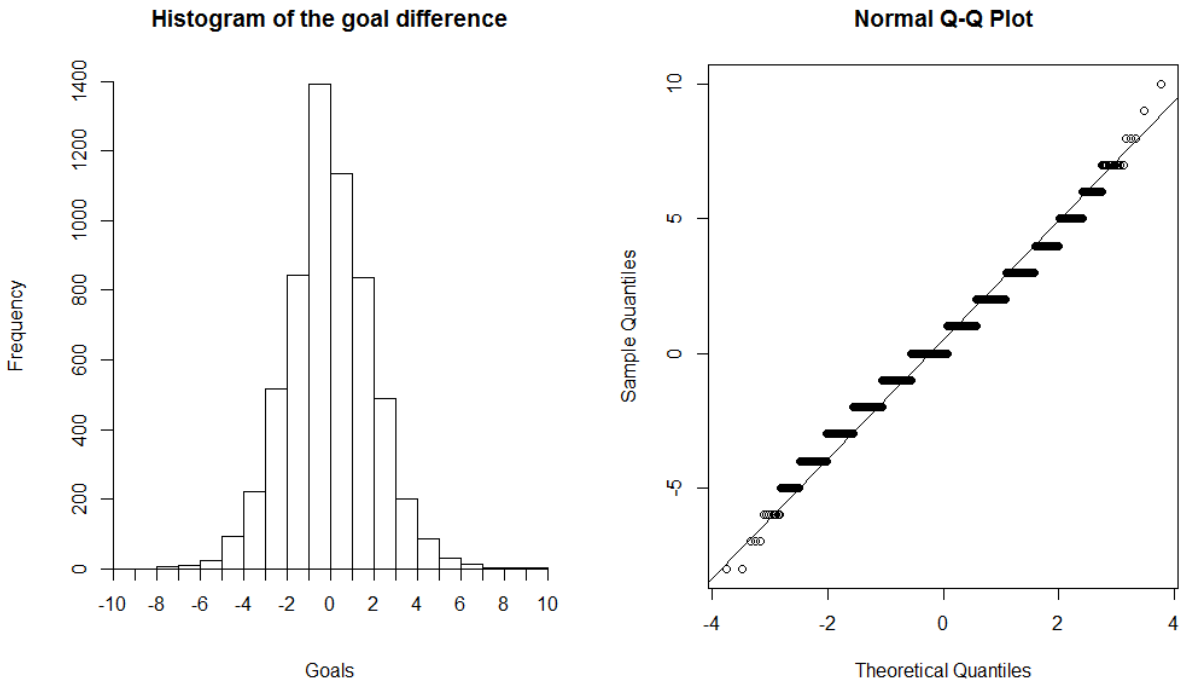
Which statistical model has to been chosen depend of the distribution the response variable come from. When the response variable comes from a normal distribution the linear regression model is most appropriate. But when the underlying distribution of the response variable comes from a exponential family (Exponential distribution, Poisson distribution, Gamma distribution or Binomial distribution) the Generalized linear regression model is most suitable.

2.1.1 Response variable

The response variable is defined as the goal difference of a single match in the professional Dutch soccer competition between 2005 and 201. This variable is defined by the difference between the number goals scored for both teams. Therefore, the goal difference can either be zero, a positive- or a negative value. When this variable is zero the match ended in a draw. But while the variable is positive the home playing teams won that game. Figure 2.1 shows this goal difference for all matches in the both Dutch soccer competitions between 2005 and 2014.

In most of the cases the goal difference between two clubs is zero. While in 31% of the cases a match ended with difference of only 1 goal. The highest goal difference in the data set is 10-0 of the match between PSV and Feyenoord in the season 2010-2011. Figure 2.1 shows a larger right tail, which indicates the home advantage that a lot of researchers[6, 7, 8] already proved.

This shape of the histogram and the straight line in the QQ-plot suggest that the difference between de scored goals comes from a normal distribution. However, the data does not fit the normal distribution because all the values are rounded integers in a small range. Which can be concluded by testing the data for normality (p-value = 2.2E-16) with the Shapiro-Wilk test.



Histogram *(2) QQ-plot*

Figure 2.1: *(1) a histogram and (2) QQ-plot of the goal difference of all the matches of the professional Dutch soccer competition during 2005 and 2014.*

The response variable is created by subtracting the scored goals of the home and the scored goals of the away playing team. There are many researchers like Heuer, Müller and Rubner [2010][13] that proved that the amount of goals scored by a home or away playing team come from a Poisson distribution.

Figure 2.2 shows a histogram and QQ-plots of the amount of goals scored by home and away playing teams of a single match. The both histograms shows that the probability that an away team does not score a goal is higher than in the case that the home playing team does not score. The estimated mean of the scored goals of the home team (λ_1= 1.77) is higher than the estimated mean of the scored goals of the away team (λ_2=1.30).

Both histograms suggest that these data come from a Poisson distribution. In the QQ-plot the difference between the real values and the estimated values are very small, which indicates that the data comes from a Poisson distribution. With a high certainty there can been concluded that the home team scored goals (p-value = 0.22) and the away teams scored goals (p-value = 0.35) come from a Poisson distribution.

This has been confirmed by testing the data with the Kolmogorov-Smirnov test method. This method test whether some sample belongs to some probability distribution function. The test calculates the difference between the empirical distribution function of the sample and the cumulative distribution function of the probability distribution. Whether this difference is to big the null hypotheses would be rejected.

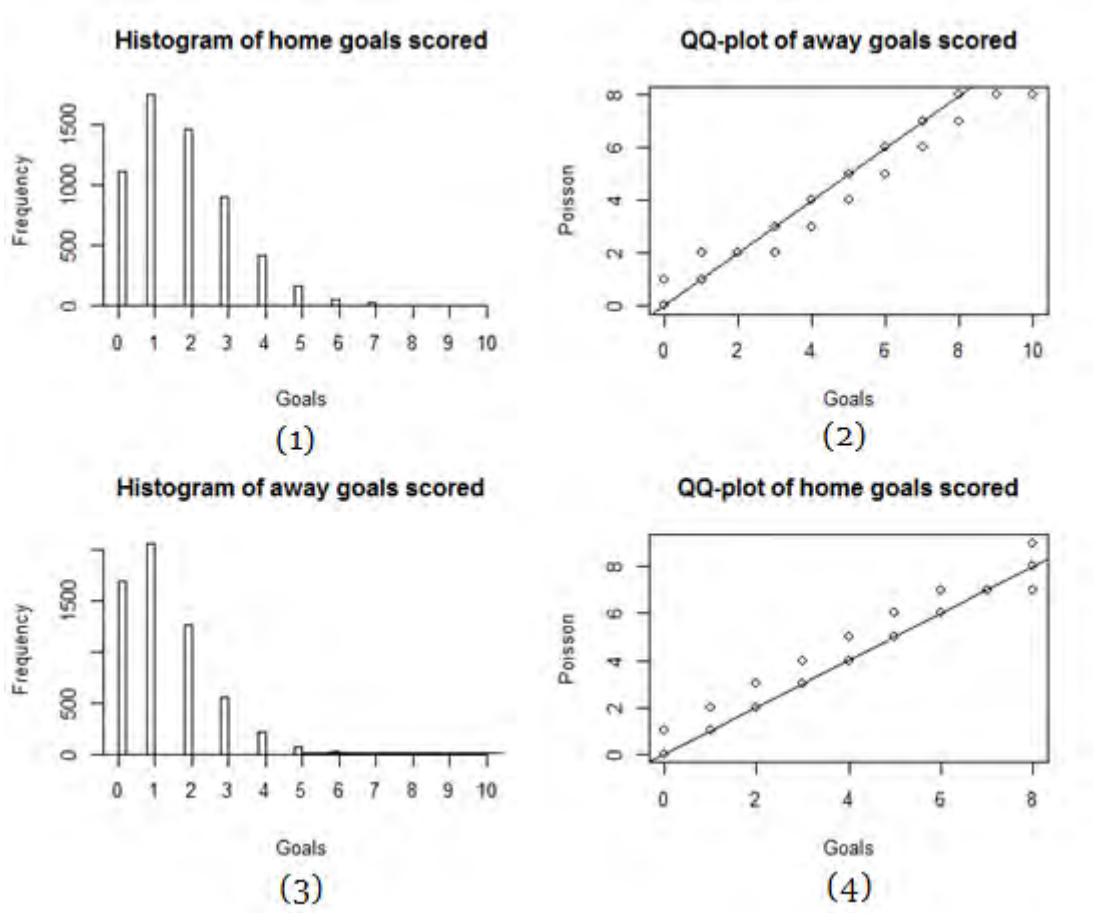


Figure 2.2: *the histogram (1) and QQ-plot (2) of the goals that a home team scored and the histogram (3) and QQ-plot (4) of the away team scored in both professional Dutch soccer competitions during the seasons 2005 and 2014.*

2.1.2 Skellam distribution

With the information that is found above there is another distribution that probably would fit the data. Because the response variable is defined as the difference between scored goals by the home and away playing team. These numbers of scoring goals are both Poisson distributed.

The Skellam distribution (or Poisson difference distribution) is a discrete probability distribution when a random variable X is de denoted by $Skellam(\lambda_1, \lambda_2)$, if and only if

$$X = U_1 - U_2,$$

With U_1 and U_2 are two independent random variables with $U_i \sim Poisson(\lambda_i)$. Because of the huge dataset the variables U_1 and U_2 are independent. The probability function of X is defined on a set of integer $\mathbb{Z}$ with $X = \{\dots, -2, -1, 0, 1, 2, \dots\}$. This probability distribution was created by Irvin (1937)[10] for the case of equal parameters and Skellam (1946)[11] with different parameters.

The probability mass function of the Skellam distribution X is given by

$$f(k;\lambda_1,\lambda_2) = e^{-\lambda_1-\lambda_2}\left(\frac{\lambda_1}{\lambda_2}\right)^{\frac{x}{2}} I_x\left(2\sqrt{\lambda_1\lambda_2}\right), \quad x = (\dots,-1,0,1,\dots)$$

For all $\lambda_1 > 0$ and $\lambda_2 > 0$ and $k = u_1 - u_2$ and where $I_y(x)$ is the modified Bessel function given by

$$I_y(x) = \left(\frac{x}{2}\right)^y \sum_{k=0}^{\infty} \frac{(x^2/4)^k}{k!\,(y+k)!}.$$

This probability mass function is derived out the difference of the probability mass functions of two different Poisson distributions. De expected value of a Skellam distribution is $\mathbb{E}(X) = \lambda_1 - \lambda_2$, while the variance is $Var(X) = \lambda_1+\lambda_2$. Often the Skellam distribution is used for describing the statistics of two counting distributions. This probability distribution is often used in sports to describe difference in match results of a single match.

To determine whether the data comes from a Skellam distribution it is not accurate enough to assume that both of the home and away scored goals become from a Poisson distribution. But the difference between het home and away scored goals also become from a Skellam distribution. Even though this is the definition of the Skellam distribution. The Kolmogorov-Smirnov test is used to test whether the data come from a Skellam distribution. First generate a sample of 5894 values from the Skellam distribution with estimated parameters. Figure 2.3 shows the goal difference of the real values, simulated values and the Skellam density function. This figure shows a small difference between real and simulated values. This indicates that the goal difference data set come from a Skellam distribution.

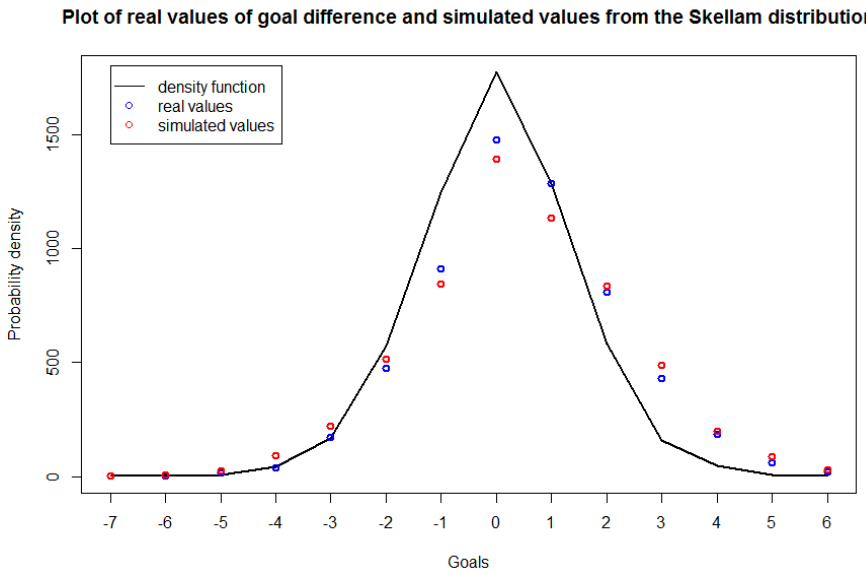


Figure 2.3: *plot of goal difference in the both professional Dutch soccer competition during the seasons 2005 and 2013, simulated vales from the Skellam distribution and the probability density function of the Skellam distribution with mean 0.*

The Kolmogorov-Smirnov test tests whether the goal difference sample comes from a Skellam distribution. This test has no build function for testing whether the data belongs to a Skellam distribution. However, the Kolmogorov-Smirnov test can test whether two sample belongs to the same distribution. This test does not reject (p-value = 0.18) the null hypotheses. Therefore, the data of the goal difference fits a Skellam distribution.

Another options to model the data is to use the Generalized Linear Models. For this model the distribution of the data must become from the exponential family. The Poisson distribution could fit this data because of the rounded values, but values of the Poisson distribution can only be positive. To test whether this data become from the Poisson distribution the data is dived in two new datasets. One of the new dataset consist only the positive goal differences and the other dataset consist of only the absolute values of the negative goal differences. The data from the matches that ended in a draw are added in both new datasets.

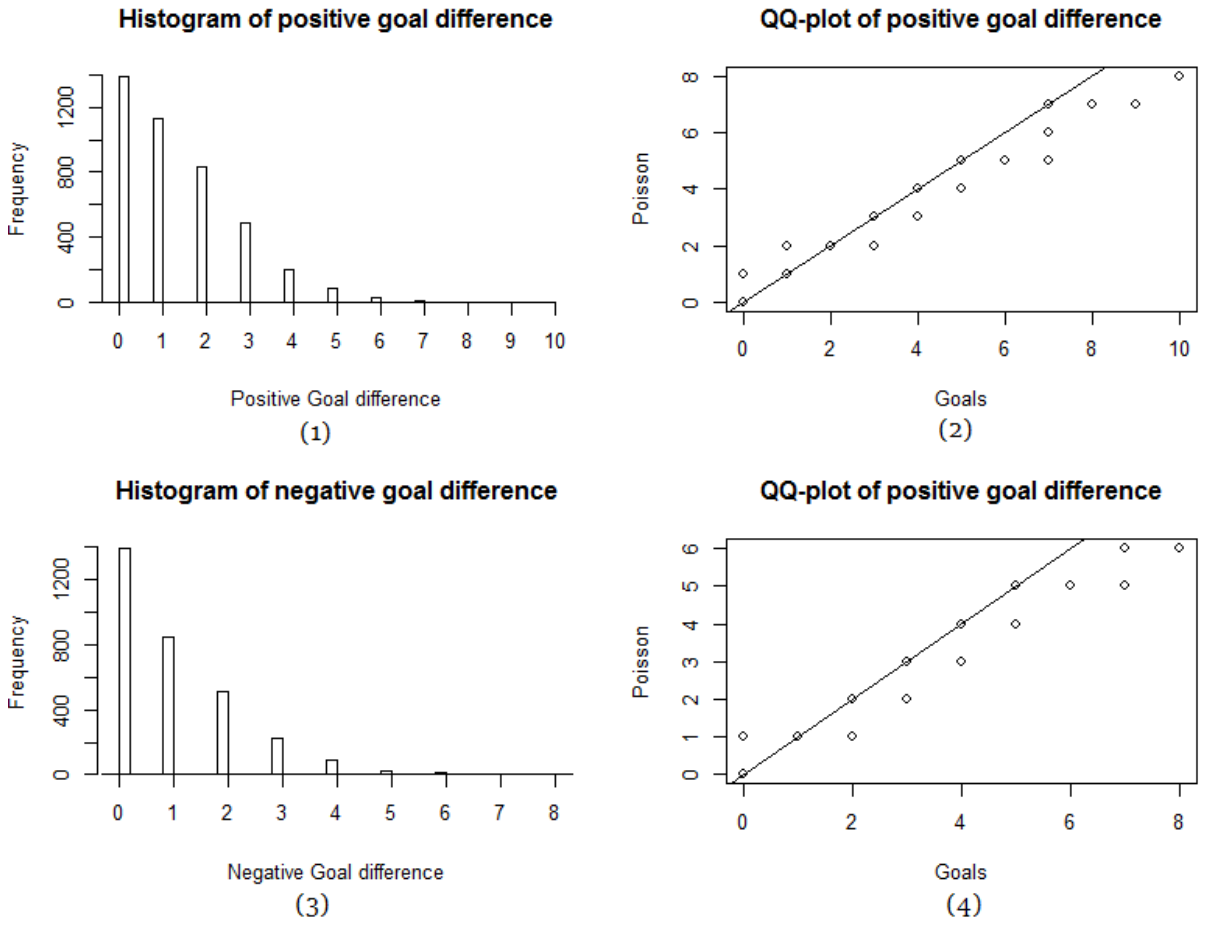


Figure 2.4: *the histogram (1) and QQ-plot (2) of the positive goal differences and the histogram (3) and QQ-plot (4) of the absolute values of the negative goal differences for all matches in both professional Dutch soccer competitions during 2005 and 2014.*

Figure 2.4 shows no indication whether both datasets fits a Poisson distribution. The histogram does not show that particular Poisson density function but the difference between the dotted QQ-vales and line through these points looks small. The Kolmogorov-Smirnov test is used to test whether these datasets fits a Poisson distribution. The positive (p-value = 1.18E-10) and the absolute values of the negative goal difference (p-value = 6.68E-8) does not fit a Poisson distribution. So the null hypotheses is been rejected.

2.1.3 Explanatory variables

The main research question is about the influence of the surface on a match result in soccer. It is logical to use the types of surfaces where the matches are played on as an explanatory variable (or factor). In only 13% of all matches it is played on an artificial turf field. The factor type of surface is determined by

$$type\ of\ surface = \begin{cases} 1\ if\ match\ is\ played\ on\ an\ artificiel\ turf\ field. \\ 0\ if\ match\ is\ played\ on\ natural\ grass. \end{cases}$$

It is interesting to figure out whether the additional home advantage differs between both divisions. This factor is determined by N_1 by playing in the first professional Dutch soccer division and N_2 by playing in the second division.
The type of surface and level of division are not the only factors that have influence on the outcome of a soccer match. The factor with the biggest influence is probably the strength of each team. The influence of the type of surface between a team with a high strength and a team with a low strength is less, than at the moment with equal teams. This strength is modeled on the basis of the final rankings of the competition. Every team gets a score between 3 and -3 based on the position of the final ranking of that particular year. The team that ended that particular year at place one get a score of 3 and the team that ended at the last place get a -3.

Whether a team wins a match does not only depend on the strength of a team or the type of surface they are playing on. The shape of the day can also be an decisive factor. This shape of a team is determined by the results of the last three matches they played. After a match each team gets a score and the sum of these three scores is the shape of that particular team. The score after each match is determined by

$$score\ for\ team\ i\ per\ match = \sum_{i=1}^{3} A_i$$

with

$$A_i = \begin{cases} 1\ if\ team\ i\ won\ the\ match \\ 0\ if\ match\ of\ team\ i\ ended\ in\ a\ draw \\ -1\ if\ team\ i\ lose\ the\ match \end{cases}$$

The shape of each team varies during the season, while the strength of a team is during the season fixed.

2.2 Models

This part describes the different models that are used to test the advantage of playing their home matches on an artificial turf field. These statistical models test whether some factors has influence on the response variable. The response variable comes from a uncommon distribution which has no statistical model. However, the Kruskal-Wallis Test, Mann-Whitney U-test and Friedman test tests without knowing which distribution fits the response variable. These methods are based by testing on ranks and can only test whether some factor has influence on the response variable.

More interesting are testing on more factors and test whether some factors has some interaction between each other. This type of test is not possible under the assumption that the response variable comes from the Skellam distribution. A statistical model that tests under the assumption that the response variable comes from the Skellam distribution is not yet developed. To test with multiple factors and interaction between the factors the Generalized Linear Model (GLM) can be used. This Model needs the assumption that the data comes from a distribution that belongs

of the exponential family. Already proved that the positive goal differences and absolute values of negative goal difference does not fit the Poisson distribution. Even the test will not work perfect, the tests and analyzes are interesting to do.

2.2.1 Kruskal-Wallis test

The Kruskal-Wallis test is a test based on ranks. Therefore, the underlying distribution of the response variable is not important. The Kruskal-Wallis test tests whether some population distributions are identical. When the test leads to no significant result, the tested factor has no influence on the response variable.

The Kruskal-Wallis test is based on ranks and the first step to evaluate the test is to rank all the data. The test statistic of Kruskal-Wallis One-Way Anova test is given by

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{m} n_i R_{i.}^2 - 3(n+1)$$

With n_i is the number of observations in group i and $n = \sum_{i=1}^{m} n_{i\,,}\, R_{ij}$ is the rank of observation j from group i with $R_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}$ as the average number of rank and m as the number of groups. The p-value of this test is given by the chi-squared distribution with $m-1$ degrees of freedom. The null hypotheses should been rejected when $K > \chi^2_{\alpha:m-1}$.

2.2.2 Generalized Linear Models

The generalized linear model (GLM) is a generalization of the ordinary linear regression model. The link function of the GLM is the only difference between the GLM and the ordinary linear regression model. The link function links the GLM Generalizes to a linear regression model. An assumption for working with the GLM is that the response variable comes from a distribution that belongs by the exponential family. Each distribution in the exponential family has another link function.

The structure of a GLM consists of three components. (1) The underlying probability distribution of the random variable. This distribution must fit a distribution from the exponential family. The probability mass function will been given in the canonical form

$$f_i(y) = f(y, \theta_i) = \exp\left[\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + C(y, \phi/A_i)\right]$$

In the case of the Poisson distribution the probability mass function in canonical form is given by

$$f_i(y, \theta_i) = \exp\left[\frac{y \log(\lambda) - b(\log(\lambda)}{\phi/A_i} + C(y, \phi/A_i)\right]$$

Take $\phi/A_i = 0$ and $b(\theta_i) = e^{\theta}$. then

$$f_i(y, \theta_i) = \exp[y log(\lambda) - \lambda - \log(y!)] = \frac{\lambda^y\, e^{-\lambda}}{y!}$$

This is the probability mass function of the Poisson distribution.

(2) A linear predictor (η) that is related to the expected value of the data through the link function. The η is a linear function with the following structure as linear model

$$\eta_i = \alpha + \beta_i X_{i1} + \cdots + \beta_k X_{ik}$$

(3) The link function $g(.)$ is a smooth and invertible function. This function provides the relationship between the linear predictor and the expectation of the response variable by

$$g(\mu_i) = \eta_i = \alpha + \beta_i X_{i1} + \cdots + \beta_k X_{ik}$$

The link function of the Poisson distribution is defined as followed

$$g(u) = (b'(u))^{-1} = (e^u)^{-1} = \log(u).$$

Which is correct because the $\mathbb{E}[Y] = b'(\theta) = e^{\theta} = e^{\log(\lambda)} = \lambda$.

Conclusions about chapter 'methods'

- The goal difference between the home team and away team determined as the response variable
- The Skellam distribution is a discrete probability distribution of the difference between two random variables U_1 and U_2 each having a Poisson distribution
- The Response variable fits the Skellam distribution.
- The positive and absolute values of the negative goal difference does not fit the Poisson distribution.
- Use the Kruskal-Wallis test for one way of factor analyzes.
- Use the Generalized linear model to test whether multiple factor has influence on the response variable. Without satisfy the model assumption of fitting an exponential family.

Results

In this part of the paper the result will been described. Even as the results of the one-way and multiple-way models is presented. The multiple-way models will test whether some factors has influence on the response variable.

3.1 Kruskal-Wallis test

The Kruskal-Wallis test can only compare two independent samples. With a p-value smaller than 0.05 concludes that there is a significant difference between two samples. There is been tested between the response variable and all the different factors.

Test	Kruskal-Wallis Chi-Squared (K)	degrees of freedom	p-value
Response ~ Surface	19,82	1	8,49E-06
Response ~ Strength	1206,56	12	2,00E-16
Response ~ Shape	265,16	12	2,00E-16
Response ~ Division	6,64	1	9,96E-03

Table 3.1: *Results of the Kruskal-Wallis test.*

The p-values of all the different variables are smaller than 0.05. All the tested variables are significant different than the response variable. The p-value of type of Surface is smaller than 0.05 and reject the Null hypotheses, which means that the average goal difference at artificial turf is not equal to the average of the goal difference on ordinary grass. The Kruskal-Wallis chi-squared of the factor strength is very high. The average goal difference for each quality class is different. That sounds very logical that a 'good' team has a higher average goal difference than a 'weak' team. The table shows a significant difference between the average goal differences in both Dutch professional soccer competitions, which does not make any sense.

3.2 Generalized Linear Models

The Generalized linear models (GLM) can test whether some variables has influence on the response variable. This can be done with an one-way or multiple-way test. In this part the dataset is divided in two parts, the part with only the positive goal differences and the part with the absolute values of the negative goal differences. Therefore, both of the datasets has positive values, which is needed by using Poisson regression. In this Poisson regression model the used link function is $g(u) = \log(u)$.

test	Positive goal differences p-value	negative goal differences p-value
Response ~ Surface	8,84E-03	7.48E-03
Response ~ Strength	2,00E-16	2.00E-16
Response ~ Shape	2.00E-16	1.06E-15
Response ~ Division	1.04E-05	2.73E-01

Table 3.2: *Results of the one-way Poisson Generalized Regression model.*

There is only a small difference between the p-values of the positive goal difference and the negative goal difference. Only at the factor division there is a difference between both. There is only one p-value bigger than 0.05, the response ~ division for the negative goal differences. This p-value means that the factor division has no influence on the response variable.

Because the P-values of the types of surfaces are both beyond 0.05 the type of surface has influence on the match results. This can also been concluded for the factors shape and strength.

The estimated regression function for the full model is

$$\log(\mu_i) = -0.074 + 0.126 * Surface - 0.0126 * Shape + 0.1366 * strenght - 0.044 * Division.$$

With this model the multiple-way Poisson generalized model is build.

Test	Factor	Positive goal differences p-value	negative goal differences p-value
Response ~ Surface + Strength	Surface	7,91E-01	9,13E-01
	Strength	2,00E-16	2,00E-16
Response ~ Surface + Shape	Surface	4,31E-02	1,94E-02
	Shape	2,00E-16	2,42E-15
Response ~ Surface + Division	Surface	2,79E-02	5,14E-03
	Division	3,38E-05	1,73E-02
Response ~ Surface + Strength + Shape	Surface	7,83E-01	9,18E-01
	Strength	2,00E-16	2,00E-16
	Shape	9,79E-02	7,59E-01
Response ~ Surface + Strength + Shape + Division	Surface	4,74E-01	9,23E-01
	Strength	2,00E-16	2,00E-16
	Shape	1,19E-01	7,60E-01
	Division	1,32E-04	9,63E-01

***Table 3.3**: Results of the multiple-way Poisson Generalized Regression model.*

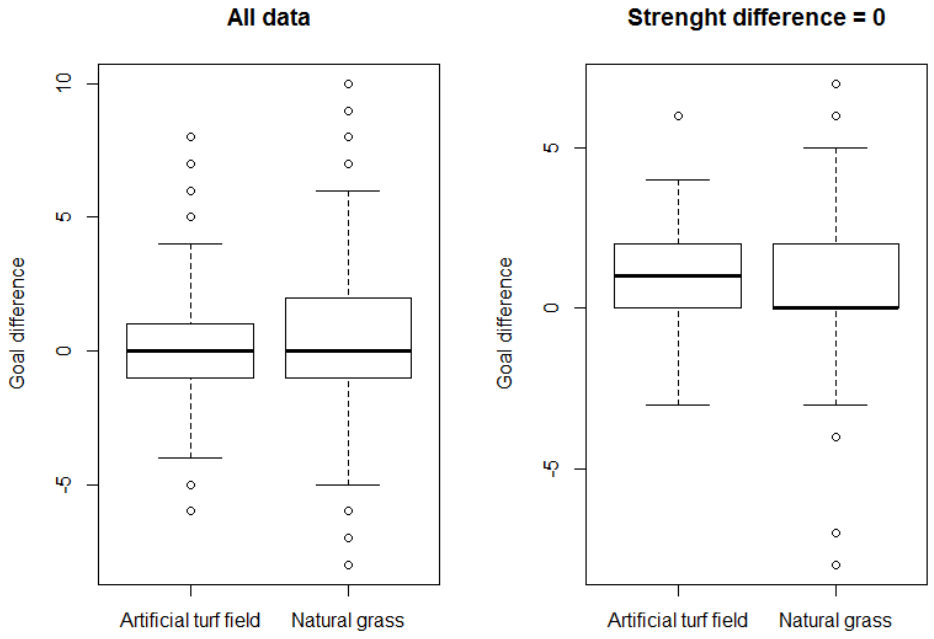
Table 3.3 gives the result of the multiple-way Poisson regression tests. Just like in table 3.2 there is a small difference between the p-values of the positive goal differences corresponding with p-values of the absolute values of the negative goal differences.

In table 3.2 is found that the type of surface has influence on the outcome of the match. But by adding the factor strength on the test, the influence of the type of surface (p-value = 0.791 or 0.913) disappeared. So the strength of the team has big influence (p-value = 2.00E-16) on outcome of the match. This indicates that there is a high probability that the strongest team wins the match. Testing the difference in shape and the type of surface, the factor type of surface has still a small influence (p-value = 0.043 or 0.0194) on the match outcome. The factor shape has even more influence (p-value = 2.00E-16 or 2.42E-15) on the match outcome.

Testing the variables type of surface, strength and shape shows again that the strength of the team is the most important factor. Type of surface (p-value = 0.783 or 0.918) and shape (p-value = 0.097 or 0.759) has no influence anymore in a model with factor strength. This fact shows again that the strength of the team is the most important factor to decide which team has highest probability to win a match.

The results of the full model shows that with the positive goal differences the factor division has influence (p-value = 0.0001) on the match results. While this factor has no influence (p-value = 0.963) at the data of the negative goal difference.

The factor strength is the most important factor to decide which team has the highest probability to win a match. But without this factor there was a significant difference between the match results of the matches played on both types of surface. But have a team playing at home on artificial turf an additional home advantage.



(2)

Figure 3.1: *box plots of the goal difference on different field types. (1) is for all the matches in the Dutch professional soccer competition during 2005 and 2014, (2) for all the matches in the Dutch professional soccer competition during 2005 and 2014 with a strength difference of zero.*

Figure 3.1 shows the boxplot of the goal difference on both types of surfaces with all data (1) and with only data from match with equal strength level (2). the box plot of all data (1) shows equal medians but a higher variance for goal difference on ordinary grass, which concludes that the average goal difference at both type of surface are almost equal. While the box plot with data of equal strength (2) shows that the median of the matches on an artificial turf field is much higher. With equal strength the average goal difference of artificial turf field is much higher (0.88) than the average of goal difference on ordinary grass (0.5).

Conclusions about chapter 'results'

- The Kruskal-Wallis test shows that there is a significant difference between the average goal difference between the type of surface, strength difference, shape difference and division.
- The one-way Poisson regression method shows that the type of surface, strength difference and shape difference has influence on the result of the match.
- The multiple-way Poisson regression method shows that the factor Strength has always the most influence on the match result.
- The average goal difference on artificial turf is higher than the mean goal difference at ordinary grass given the strength of both teams are equal.

4. Conclusion

The main goal of the research was to figure out whether a team that plays at home on artificial turf has additional advantage than a team that plays at home on ordinary grass. Therefore, first has to be determined whether playing on artificial turf has influence on the match outcome. The differences between passes at an artificial turf field an ordinary grass is been researched.

Every year more clubs has an artificial turf fields in the Dutch professional soccer competitions. In 2003 Heracles Almelo was the first team with an artificial turf field. In the season 2014-2015 50% of the Dutch professional soccer clubs has artificial turf. But has a club with artificial turf an additional home advantage. A team that played on artificial turf has earned fewer points than a team at ordinary grass, but it is also interesting to look at the strength of each team. Because on average the strength of the teams that played on an artificial turf field is lower than the strength of the teams at ordinary grass.

The difference between playing on different surfaces there are some interesting statistics, like the amount of passes for each team at both type of surface. On artificial turf the home team passes significant more than then a home teams which plays on ordinary grass. It can conclude that passing at artificial turf is easier than at ordinary grass. A result of previous statement is that artificial turf is always flat so the ball can roll faster.

The problem will be modeled as a statistical model. A statistical model can test whether some factors has influence on a response variable. In this paper the response variable is defined as the goal difference for a particular match. The response variable is either zero, positive or negative. The factors in the statistical model are: the type of surface $\in [1,0]$, strength difference between both teams $\in [-12,12]$, shape difference between both teams $\in [-12,12]$ and the division the match is played $\in [0,1]$.

The statistical model to use depends on the underlying distribution of the response variable. In this paper the response variable comes from a Skellam distribution (or Poisson difference distribution). The Skellam distribution is the difference between two independent Poisson distributions with different means $\lambda_1 \text{ and } \lambda_2$. Right now there is no statistical model developed that works under the condition that the response variable comes from a Skellam distribution. Therefore the Kruskal-Wallis test is used for one-way factor testing.

It is also interesting to test on more than one factor and find whether the factors have some interaction. For testing with a Generalized Linear Model (GLM) the response variable should come from a distribution of the exponential family. To solve this problem the dataset is divided in two smaller datasets. The first dataset consist of only the positive goal differences and the second dataset consist of only the absolute values of the negative goal differences. Both of these datasets does not come from a Poisson distribution. With that information one assumption of the generalized linear model is not satisfied. Still the tests are interesting to do and analyze the results.

With the Kruskal-Wallis test there is shown that there is a significant difference between the goal difference and all the four factors (surface, strength, shape and division). For the shape and strength is that very logic and shows a high significance. This significance tells that there is a difference between the average goal differences on both types of surfaces.

The link function that is used for the Generalized Linear model (GLM) is the $log(u)$ function. This link function suggested that the data comes from a Poisson distribution but that is not true. The one-way GLM gives the same answer as the Kruskal-Wallis test did. The only difference is that there is no significant influence of the division by the negative goal differences. In the case that the home team losses there is no difference in goal difference between the both divisions.

The multiple-way GLM gives more information about whether some factors have influence on the response variable. In the one-way GLM has found that there is a significant difference between the type of surface and goals difference. But the multiple-way GLM shows the strength of the team has the biggest influence on the goal difference of a single match. The shape and surface have no influence on the outcome of the match when a team is significant stronger. So when a there is a strong and a weak team the type of surface has no influence on the outcome of the match. But what will happens when the strength of the teams are the same. With two equal teams (strength difference = 0) the average goal difference of the teams that played on artificial turf field (mean = 0.88) is higher than the team that played on ordinary grass (mean = 0.50).

This information concludes that the match outcome depends most on the strength of each team. When a strong team plays against a weak team, the shape difference and type of surface have no influence on the outcome of the match. But when two equally good teams play against each other the team with artificial turf have more home advantage comparing by the team with ordinary grass.

4.1 Further research

The response variable in this paper is the goal difference between the home and away team during a single match. This variable comes from a Skellam distribution. For the Skellam distribution there is no statistical model to test whether some factor has influence on the response variable. In future research another response variable can be used. This variable should come from a well known distribution. Furthermore, it is possible to develop a statistical model with a Skellam distribution as underlying distribution. With this model the result should be recalculated and analyzed. With another response variable or statistical model the mistake that is made in this paper can be prevented.

References

1. Winterbottom W. (1985)., Artificial Grass Surfaces For Association Football, Sports Council, London
2. Di Salvio V., Collins A., McNeill B., Cardinale M. (2006)., Technical study with ProZone, http://ebookbrowsee.net/case-study-technical-analysis-351-pdf-d477644923
3. Vorstenbosch C.C.M., Staal J.B., Kolenburg L., Meijer K. (2008)., Voetbalblessures tijdens wedstrijden op kunstgras versus natuurgras, Sport&Geneeskunde, 2008:4, Utrecht
4. Steffen K., Andersen T.E., Bahr R. (2007), Risk of injury on artificial turf and ordinary grass in young female football players. Br J Sports.
5. Dowie J. (1982), Why Spain should win the world cup, New Scientists, 1982:94.
6. Pollard, R. (1986), Home advantage in soccer: a retrospective analysis, Journal of Sport sciences, 1986:4.
7. Nevill A.M., Newell S.M., Gale S. (1996), Factors associated with home advantage in English and Scottish soccer matches, Journal of Sports science, 1996:14.
8. Clarck S.R., Norman J. M. (1995), Home ground advantage of individual clubs in English soccer, The statistican, 1995:4.
9. Barnett V., Hilditch S. (1993), The effect of an artificial pitch surface on home performance in football (soccer).
10. Irwin W. (1937), The frequency distribution of the difference between two Poisson variates following the same Poisson distribution., Journal of Royal Statistical Society, 1937:100
11. Skellam J.G. (1946), The frequency distribution of the difference between tow Poisson variates belonging to different populations. Journal of the royal statistical society
12. Abdulhamid A., Azaid, Maha A., Omair (2010), On The Poisson Difference Distribution Inference and Applications., Bulletin of the Malaysian Mathematical Sciences Society, 2010:33
13. Heuer A., Müller C., Rubner O. (2010), Soccer: Is scoring goals a predictable poisonian process?, Europhysics, 2010:89