KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

FACULTY OF ECONOMICS AND BUSINESS

Prediction model for the Jupiler Pro League

Steven Cassimon

r0260620

Thesis submitted to obtain
the degree of

MASTER OF ECONOMICS
Major Research in Economics

Promoter: Prof. Dr. Geert Dhaene
Assistant: Yutao Sun

Academic year 2013-2014

KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

FACULTY OF ECONOMICS AND BUSINESS

Prediction model for the Jupiler Pro league

Attack and defense strengths of football teams vary over time due to changes of players, managers or chairmen. After looking into the already existing models for the prediction of football matches, we implement a statistical model for the analysis and forecasting of football match results. These results are assumed to come from a double Poisson distribution (the product of 2 independent Poisson distributions) with means that change over time. Our treatment is based on least squares estimation. The out of sample performance of our methodology is verified in a betting strategy that is applied to the match outcomes of the season 2013/2014 of the Jupiler Pro League. We show that our statistical modelling framework can produce a positive mean return over the bookmaker's odds.

MASTER OF ECONOMICS
Major Research in Economics

Academic year 2013-2014

Acknowledgements

First, I would like to thank Prof. Dhaene for the opportunity to work on this thesis and Yutao Sun for his guidance and help throughout the whole thesis. I enjoyed working on the thesis as it was a unique opportunity to combine my interest in football with my interest in statistics and probability theory.

Next I would like to thank my parents for supporting me throughout the whole process. Also many thanks to my brother and nephew for rereading the thesis.

Last but not least I would like to thank my roommates and friends for listing to my plans, theories and problems along the way, although it might not have interested them that much.

I have learned a lot from this thesis: my English writing skills have improved, my knowledge about statistics has expanded and my programming skills have most certainly improved. It was a very instructive first contact with doing research and I will take this experience with me in further projects.

Table of Contents

General Introduction

Football or soccer is a sport where it is said that a game is never won before it is played. One of the main quotes in football used by the pundit is “anything is possible”. Is this truly so? Are we unable to predict or have a good shot at predicting the outcome of a football match? As can be seen from the number of bets, many people believe that they can actually gain a positive return on the betting market just by using their knowledge of the game. However, can we go further than our intuition to predict the outcome of a game? Is it possible to predict the outcome of a football match by making use of statistics? In this thesis we develop a model that is capable of predicting match outcomes of the Jupiler Pro league. Probably this thesis is one of many to come as the money going around in the betting industry is growing at a fast rate nowadays, which will increase the interest in the prediction of football matches.

In this thesis we study a history of football match results from the Jupiler Pro league in the last 9 years. The number of goals scored by a team may depend on the attack strength of a team, the defense strength of the opposing team and the home ground advantage if applicable. Match results are analyzed on the basis of a dynamic statistical modelling framework in which the attack and defense strengths of the teams are supposed to vary over time. We will explore two models and show that at least one is accurate enough to gain a positive mean return over the bookmaker’s odds.

The remainder of the thesis proceeds as follows:
In the first section we will discuss the literature about sports betting and the prediction of football matches. Here we will summarize some papers in the field of sport statistics. In the second section the data used in the thesis will be described. Also some descriptive statistics of the Belgian Jupiler Pro league will be presented. In the third section our statistical modelling framework is introduced and discussed in detail. The estimation procedure is based on least squares estimation. Finally in the fourth and last section we look at the results we obtain from our estimation and we will discuss whether these results are realistic and if our modelling framework can be used in practice. We end this section by making predictions for the season 2013/2014 of the Jupiler Pro League based on our model and setting up a betting strategy.

1 Literature review

In this first section we give an overview of the literature in the field of sport statistics. Before getting started, it may be convenient to look at some key features of a prediction model for football matches, as summarized by Dixon and Coles [1].

a) The model should take into account the different abilities of both teams in the match.

b) There should be allowance for the fact that teams playing at home generally have some advantage: the so called home ground advantage.

c) The most reasonable measure of a team's ability is likely to be based on a summary measure of its recent performance.

d) The nature of football is such that a team's ability is likely to be best summarized in separate measures of their ability to attack (to score goals) and their ability to defend (not to concede goals).

e) In summarizing a team's performance by recent results, account should be taken of the ability of the teams that they have played against.

1.1 Skill based model

In an early contribution to the literature on modelling goals scoring, Maher [2] used univariate Poisson distributions with means reflecting the attacking and defensive capabilities of the two teams.

Before this paper was published, the idea ruled that the number of goals was best approximated using a negative binomial distribution. Reep, Pollard and Benjamin [3] confirmed this using data from the English Football League First Division for four seasons, and then proceeded to apply the negative binomial distribution to other ball games like cricket, ice hockey, baseball and lawn tennis. In an earlier paper, Reep and Benjamin [4] even remarked that "chance does dominate the game". From an intuitive point of view this makes no sense. It is easily shown as was done by Hill [5] that people can expect the league table at the end of the season quite well in advance. Thus, over a whole season, skill rather than chance is likely to dominate the game, as argued by Maher[1] [2].

In contrast to Reep, Pollard and Benjamin [3], Maher [2] thought that the number of goals scored by a team in a football match is likely to follow a Poisson distribution. The mean of this variable will vary according to the quality of the team. An assumption Maher [2] made in his basic model is that there is no dependence between the number of goals scored by the teams in a match. Thus when we have an observed match score, the number of goals scored by the home team is Poisson-distributed with the mean depending on the attack

[1] Therefore throughout the rest of the thesis we will consider goal counts to be fitted by a Poisson distribution which is also what we find in most of today's literature.

strength of the home team and the defensive weakness of the away team. The number of goals scored by the away team is also Poisson-distributed with the mean depending on the attack strength of the away team and the defending weakness of the home team. Both scores are assumed to be independent.

Using this model he studies data of the English premier league season 1971/1972 which back then still consisted out of 22 teams. By making use of a hierarchy of models in which every model includes more parameters, Maher [2] tests these parameters and shows that although home ground advantage is a highly significant factor, it applies with equal effect to all teams. This implies that a single parameter can be used to describe the quality of a team's attack strength no matter if the team is playing at home or away. The same applies for the defensive weakness of a team.

Although this model gives a good start to think about how to model football scores by using a Poisson model, there are still some points of criticism. Maher uses a static model, whereas Dixon and Coles [1] and others argue in favor of a dynamic model. At the very end of his paper Maher also explores the existence of a small dependence between home and away scores. He found a considerable improvement in the model fit by trying a range of values for the dependence parameter, suggesting that the independence assumption in his original model is invalid. However he did not provide estimates of the dependence parameter, therefore this is another point which can be improved.

1.2 Making the model dynamic

There have been several authors who have extended the basic framework of Maher [2] discussed in the previous subsection. Just like Maher, Dixon and Coles [1] start with a double Poisson model assuming that there is no dependence between the scores. They study English league and cup football data from 1992 to 1996.

On the model specification, Dixon and Coles [1] make 2 important extensions to Maher's model. First they provide a modification of the independence assumption since they found that under this assumption low scoring draws were underestimated. They take care of this problem by implementing a correction factor in the model. This correction factor increases the probabilities of the scores which are underestimated and lowers the probabilities of the scores which are overestimated by the double Poisson model.

Another modification is that they make the model dynamic. In reality, a team's performance is likely to vary from one time period to another. In particular, they argue that a team's performance is likely to be more closely related to their performance in recent matches than in earlier matches. This behavior has to be implemented in the model. Instead of formalizing a stochastic development of the model parameters, Dixon and Coles [1] estimate these parameters for each time point t using data on the history of match scores up to time t. For example, when estimating the scores of the first week of the season 1995/1996 they use information up to the first week whereas when estimating scores of the second week of the season 1995/1996 they use information up to the second week of that season. Thus the estimates change over time since they are estimated using different data. The downside is that they can only forecast scores one period ahead. In addition they introduce a weight to the log-likelihood function which puts more weight on recent performances.

Using the proposed model, Dixon and Coles [1] found that the sequence of estimates of attack and defense strengths over time are non-uniform for the different teams which implies that team performances are in fact dynamic. They also show that the home ground advantage remains practically constant over time.

Rue and Salvesen [6] also extended the basic framework of Maher [2] by making it dynamic. They propose a Bayesian dynamic generalized linear model for predicting football match outcomes. Just as Maher [2] and Dixon and Coles [1] , they assume that a match result is an observation generated from a double Poisson distribution. As Dixon and Coles [1], they find that the number of draws are underestimated using a double Poisson model. To cope with this problem they use a correction factor similar to the one proposed by Dixon and Coles [1].

Just as Maher [2] and Dixon and Coles [1], Rue and Salvesen [6] consider the attack and defense strengths to be the two most important explanatory variables. In addition they assumed these strengths to be time-varying, like Dixon and Coles [1]. The time variation is modelled differently however. While Dixon and Coles [1] downweight the likelihood to mimic the time varying properties, Rue and Salvesen [6] propose a Brownian equation of motion to model the time variation of the attack and defense strengths. They argue that the main purpose is to predict matches in the near future, therefore only a recent behavior for the properties in time is needed. Thus they choose for an equation of motion in continuous time.

A novelty the authors introduce in their paper is the use of a psychological effect. Apart from the usual explanatory variables (attack strength, defense strength and home ground advantage) they include psychological effect as an extra explanatory variable. This extra feature tends to capture the effect in football that team A is likely to underestimate the strength of team B if A is a stronger team then team B. For example, when RSC Anderlecht is playing against Oud-Heverlee Leuven, the players of Anderlecht could become “lazy”. The authors argue that in this case the strengths of the team A and B are not so different since they are analyzing teams in the same league (Premier League) thus it is reasonable to expect that the psychological effect has a negative impact on the number of goals scored by team A. This is also applicable to our data. (The opposite effect, a positive effect on the number of goals scored by team A might arise when team A is so superior compared to team B that the latter develops an inferiority complex facing team A)[2].

Another modification Rue and Salvesen [6] make is truncating the numbers of goals at 5, thus a score of 9-6 is interpreted as 5-5 for example. They argue that when one team scores many goals, the independence assumption may be violated. Since this is highly demotivating for the other team and therefore has an impact on the goal scoring intensity (mean of the Poisson distribution).

A final modification they propose is a mixture model in which the previously described form is mixed with a similar form based on average scores. The interpretation of this mixture model is that not all information in a match result is informative on the goal scoring intensities.

---

[2] An example of this within the same league might be KFCO Beerschot-Wilrijk nowadays. But it is not applicable in our data.

Because of the equation of motion and the extra features in the model the estimation of the parameters becomes more complicated. In order to make inference for the properties of each team conditionally on the observed matches, one needs the conditional density of the properties. Also the skills of all teams need to be estimated simultaneously as they are dependent in a Bayesian model. To still be able to cope with this problem the authors made use of Markov chain Monte Carlo methods (MCMC).

In 2009 Owen [7] developed a model similar to the previously described model of Rue and Salvesen [6]. Like Rue and Salvesen, Owen [7] criticizes the static modelling framework of Maher [2]. Owen [7] also adopts a dynamic generalized linear model and also uses MCMC methods for estimation. The main difference however is that Owen [7] argues strongly for a model in the discrete time. The evolution component is specified as a random walk for both the attack and defense parameters.

In his paper he finds that the evolution of parameters over time, the role of attack and defense strengths and the impact of home and away scores is more effectively analyzed in discrete time. For this observation we will also formulate our model in discrete time.

Another attempt to develop a dynamic double Poisson model was undertaken by Crowder et al. [8]. They extend the model of Dixon and Coles [1] by allowing for a stochastic development of the attack and defense performances of the teams. Crowder et al. [8] propose a non-Gaussian state space model with time-varying attack and defense strengths. For the evolution of attack and defense strengths over time, they propose an $AR(1)$ process. For estimation they use approximate methods because they state that an exact analysis would be computationally too expensive.

1.3 A bivariate Poisson distribution

While some authors focused on making Maher's model dynamic, others like Karlis and Ntzoufras [9] concerned themselves with another problem of the basic model. As discussed previously, Maher [2] found that by allowing for a small dependence between the goals, the model fit improved. Later Dixon and Coles [1] and Rue and Salvesen [6] confirmed this and found that the number of draws are underestimated when using a double Poisson model. They tried to solve this by using a correction factor.

Another way to improve the model fit of the number of draws is by using a bivariate Poisson distribution, which allows for dependence between two random variables. Karlis and Ntzoufras [9] used a bivariate Poisson distribution for modelling sports data, thus more general than football data. Their paper can however be applied to the field of statistical football prediction. Karlis and Ntzoufras [9] argue that in team sports, such as football, it is reasonable to assume that the two outcome variables are dependent since the two teams interact during the game. As an extreme example, they take basketball where teams have to score sequentially, and thus the speed of one team leads to more opportunities for both teams to score. For the season 2000-2001, a correlation of 0.41 between scores was found for the National Basket Association [9]. For the result of a football match the measure of dependence is represented by the covariance between the number of home goals and the number of away goals. When this is zero the bivariate Poisson distribution is equivalent to a double Poisson distribution. This dependence parameter is able to reflect the game conditions.

Using the bivariate Poisson model, Karlis and Ntzoufras [9] study matches played in the Italian Serie A season. They find that the bivariate Poisson model gives a better fit than a double Poisson model. Another finding is that the dependence parameter remains constant over time, which simplifies the estimation procedure. Karlis and Ntzoufras [9] also investigate the effect of considering a double Poisson model while the bivariate Poisson model is a better representation. They find that the probability of a draw under the bivariate Poisson model is larger than the corresponding probability under the double Poisson model, even for small values of the dependence parameter. Also notable is the fact that the larger the dependence parameter, the larger the relative change of the number of draws is. Thus Karlis and Ntzoufras [9] argue that this may explain the empirical fact that the observed number of draws is usually larger than those predicted under a double Poisson model.

Karlis and Ntzoufras [9] use a bivariate Poisson distribution to model results from football matches and take care of the problem of under prediction of the number of draws. However, attack and defense strengths are kept static over time in their analysis.

Goddard [10] compares the forecasting performance of two types of models that have been used to model match outcomes in football. The first approach is to model the goals scored and conceded by each team and the second approach is to model the win-draw-lose match results directly. Before setting up the model, Goddard [10] suggests that a goals-based model should outperform a results-based model because the former draws on more extensive data than the latter. However he also argues that given that league points are awarded for results and not goals, wins of 2-0 and 7-2 are equal of worth. The crucial part is which team wins, the number of goals scored and conceded are incidental. Therefore goals data might contain more noise than results data.

Goddard [10], as Karlis and Ntzoufras, [10] uses a bivariate Poisson regression to estimate forecasting models for goals scored and conceded. To estimate forecasting models for match results he uses an ordered probit regression. The datasets he uses for both models are identical in all respects except from their emphasis on goals or results. He finds that the difference in forecasting performance between the two approaches is rather small and argues that this may explain why both kind of models are used. Goddard [10] himself suggests to use a hybrid specification of both approaches.

1.4 A dynamic bivariate Poisson model

In this subsection we will consider the model developed by Koopman and Lit [11]. Just as Goddard [10] and Karlis and Ntzoufras [9], they assume that the observed pairs of counts come from a bivariate Poisson distribution. In addition, they use discrete stochastic time-varying attack and defense strengths similar to the models considered by Owen [7] and Crowder et al. [8]. This approach is a novelty in the literature. In particular Koopman and Lit [11] propose a dynamic state space bivariate Poisson model for predicting football outcomes and study the matches of the English Premier League played in the period of 2003/2004 to 2011/2012.

In the bivariate Poisson model the mean of the home score and the mean of the away score (goal scoring intensities) depend on both the mean of the marginal Poisson distribution (the intensity coefficient) as well as the dependence parameter. The

dependence parameter is assumed to be constant over time, as also suggested by Karlis and Ntzoufras [9], and is the same for all matches played. The intensity coefficients vary with the pairs of teams that play against each other. They are also allowed to change slowly over time since the composition and performance of a team will also change over time. The intensity coefficient of the home team depends on a home ground advantage, which is the same for all teams and is constant over time, on their own attack strength and on the defense strength of the away team. The intensity coefficient of the away team only depends on its own attack strength and the defense strength of the home team. The states in the state space model framework can be interpreted as the attack and defense strengths that are normally not observable but are influential to the scores of teams in a match. The dynamic property of the state variable reflects the assumption that the attack and defense strength of a team evolves dynamically in time. Just as Crowder et. Al. [8] they let the attack and defense strengths follow an $AR(1)$ process. In their setting the bivariate Poisson distribution is the measurement equation of the state space model and the autoregressive process of the states can be seen as the transition equation.

The football match outcomes in the framework of Koopman and Lit [11], briefly discussed above, rely on the stochastic and dynamic properties of the attack and defense strengths of the teams. These dynamic properties of the states depend on the autoregressive coefficients and the variance of the innovation of the autoregressive process. The authors argue that they can restrict these autoregressive coefficients and disturbance variances to be the same amongst all teams. They argue that these restrictions are not strong since they expect the attack and defense strengths for all teams to be evolving slowly over time. Also to be noted is that the attack and defense time paths for all teams can still change very differently over time, since we have the innovation term in the autoregressive process. Furthermore the model depends on the dependence parameter of the bivariate Poisson model and on the home ground advantage. These are the parameters that need to be estimated via maximum likelihood. For the evaluation of the likelihood function the authors require simulation methods because the multivariate model is non-Gaussian and nonlinear and hence one cannot directly rely on linear estimation methods for dynamic models such as the Kalman filter.

In a first approach to this problem the authors set up a naïve Monte Carlo estimate with simulated paths from the state vector. This is possible since explicit expressions for the densities are known. However the authors argued this Monte Carlo estimate is not efficient (nor feasible) since the simulated paths of states were having no support from the data. Therefore the authors proposed a computationally more efficient method. In order to make computation efficient they construct an approximating linear Gaussian model. This approximation adopts the importance sampling, as developed by Shephard and Pitt [12] and Durbin and Koopman [13], in order to give better support to the numerical integration of the joint density, which is necessary for their log-likelihood function, of the state variable and the Poisson variable. This approximating linear Gaussian model allows us to compute the approximate likelihood function by means of the Kalman filter and to simulate random samples for the states by means of the simulation smoother. When using a linear model, it is possible to use Kalman filtering. This enables us to find estimates for the state vector conditional on the data. The simulation smoother [14] is based on simulating from the posterior distribution of the disturbances of the model which then allows us, as required, to form the simulations from the states.

In general the advantage of their approach is that the integrand behaves in a much more regulated fashion under this type of sampling, as compared to a naïve sampling, such that the evaluation of the log-likelihood becomes much easier and more accurate. The two types of sampling are possibly asymptotically equivalent whereas the naïve sampling is not always feasible. The computational cost of the naïve sampling for given parameter values is heavy. Because of the high dimensionality of the state vector , the integration, and hence the log-likelihood function, behaves like a random variable unless when the number of replications of such sampling is huge.

Using this computationally more efficient approach, Koopman and Lit [11] find that the persistence of the dynamic state is quite strong. This implies that the attack and defense strengths of the current period rely heavily on that of the previous period. The estimated disturbance variances for the signals are relatively small. This shows that the attack and defense strengths do not vary much over time. They also find a significant home ground advantage. This implies that indeed teams tend to score more goals when they play in their home stadium instead of another location. A discussion on what this home ground advantage exactly represents is given by Pollard [15]. He states the main hypothesized explanations for home advantage in football are given by: crowd effects, travel effects, familiarity, referee bias, territoriality, specific tactics, rule factors and psychological factors. They also find that the dependence parameter is significant and therefore contributes to the model. Thus the phenomenon that the ability or the effort of a team during a match is influenced by the other team or by the way the match progresses is found realistic. For example, if the home team leads with 1-0 and there is only 10 minutes left to play, the away team can become more determined and can take more risk in an effort to end the match in a draw. This possible change in the score due to a change in the behavior of one team or both teams is then captured by the dependence parameter. However when the authors compared the out-of-sample forecasting performance, they could not reject the hypothesis that a double-Poisson model was equally accurate as a bivariate Poisson.

2 Data

In this section we will discuss the data used when implementing our model. After having discussed the data we will represent some descriptive statistics about the Jupiler Pro League which will motivate us to set up the model.

2.1 Data description

We use a panel time series of 9 years of football match results from the Jupiler Pro League. In the first 4 seasons, the season 2005/2006 till the season 2008/2009, the system was straightforward. There were 18 football clubs active in each season. The teams playing in each season vary, since at the end of the competition the team that is ranked last in the Jupiler Pro League relegates and is replaced by the champion of the Belgacom League. The team that finishes on the 17^{th} place in the Jupiler Pro League has to play a small competition with the second, third and fourth of the Belgacom League. The winner of this small competition can play in the Jupiler Pro League for the next season. Thus each season, minimum one and maximum two teams vary. Also each team plays against another team twice, hence every team plays 34 matches in a season.

However, in the season 2009/2010 there was a change in the system. The Belgian football society reduced the number of teams in the league and introduced a play-off system. In the reshaped system there are 16 instead of 18 teams. Firstly a regular competition is played in which each team plays 30 matches. After this regular competition there are play-off matches. The fourteen highest ranked teams of the regular competition are divided into groups. The six highest ranked teams play for the title of the Jupiler Pro League in play-off I, in which each team again plays 10 matches. At the start of this competition the number of points gained in the regular competition is halved. The numbers 7 till 14 are divided into 2 groups of 4 teams play-off II, which we will not discuss since they do not appear in our data. For which we will not go into detail since we did not include it in our data. The two teams that are ranked last in the Jupiler Pro League compete against each other in 5 matches. The one with the most points at the end gets to play in the small competition with the teams from the Belgacom league, as described above. The other team immediately relegates. Thus again each season minimum one and maximum two teams vary. Throughout our analysis we will always use the regular competition data. Only at the end of chapter 4 we will use our model to say something about the play-off I matches currently being played.

For the season 2009/2010 we note that there were only 15 teams competing in that season since Excelsior Mouscron went bankrupt and all their matches were scratched. In time dimension we span a period from the season 2005/2006 to the season 2013/2014. And the total number of different teams in our dataset is 26 due to relegating and promoting teams. Most games are played in the evenings of Fridays, Saturdays or Sundays. The total number of matches played in our dataset is $4x306 + 1x210 + 4x240 = 2394$ matches. The data used in the thesis can be found on *http://www.football-data.co.uk*.

2.2 Some descriptive statistics

Let's start by looking at the number of goals scored in a single match in the Jupiler Pro League.

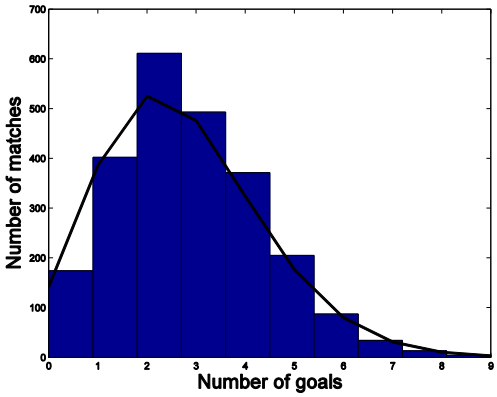

Figure 1: Histogram of the number of goals scored in a Jupiler Pro League match.

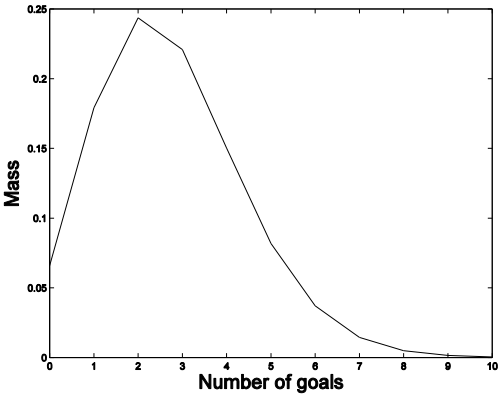

Figure 2: Continuous plot of the Poisson probability mass function.

Figure 1 shows the sum of the number of goals scored by both teams during a game in the Jupiler Pro League. The shape of the histogram confirms and illustrates that a good way to start modelling the goals would be by using a Poisson distribution. Figure 2 gives a continuous plot of the probability mass function of the Poisson distribution. The mean of the Poisson distribution is 2.7201, the mean of goals scored in a game. Using this distribution we can already calculate the probability of goals scored in a Jupiler Pro League game. Let's take the game SV Zulte Waregem vs Royal Standard de Liège played on the 4th of april 2014 as an example and compare the probabilities we calculated from our Poisson mass function with the transformed odds of the gambling site Bwin[3].

Total goals	0	1	2	3	4
Poisspdf(2.7201)	0.0659	0.1792	0.2473	0.2209	0.1502
Bwin	0.1176	0.2105	0.2779	0.2381	0.1600

Table 1: Comparing probabilities

In making bets the challenge is to find good bets, in which the calculated probabilities are higher than the corresponding probability determined by the bookmakers odds, so that there is a positive expected return. Here we cannot find such an opportunity therefore we would not be tempted to bet. Apart from the probability of a match without any goals, the probabilities from the estimated distribution presented in Figure 2 are close to the odds offered by Bwin. This suggests that this model, although basic, captures the distribution of goals and is a reasonable starting point for predicting football matches, as also argued by Whitacker [16].

[3] https://www.bwin.be/nl?trid=ex10151.

However caution is necessary as we cannot simply look at the average amount of goals of all matches to predict the number of goals of a random match, thus for a betting strategy this is simply not good enough. However we are more interested in being able to predict the number of goals for an individual team. Let's go a step further and take a look at the distribution of goals scored by individual teams. Let's take the game SV Zulte Waregem vs Royal Standard de Liège again as an example.

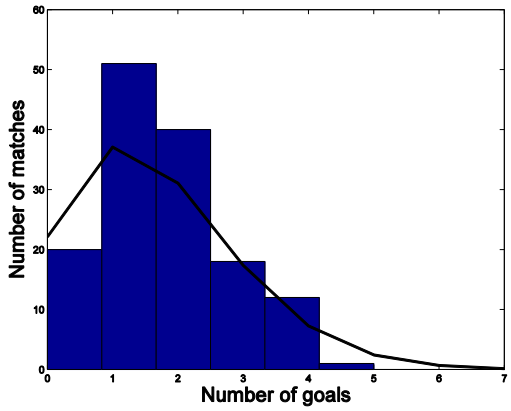

Figure 3: Number of home goals scored by SV Zulte Waregem in the Jupiler Pro League

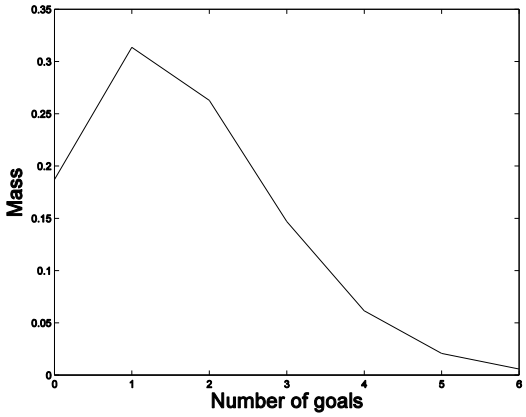

Figure 4: Continuous plot of the Poisson probability mass function.

SV Zulte Waregem goals	0	1	2	3 or more
Poisspdf(1.6761)	0.1871	0.3136	0.2628	0.2365
Bwin	0.37	0.2857	0.227	0.1143

Table 2: Comparing probabilities

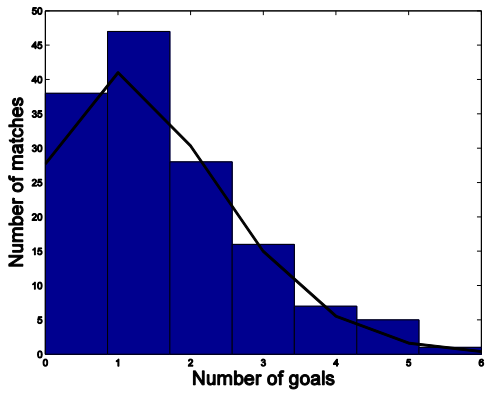

Figure 5: Number of away goals scored by Standard de Liège in the Jupiler Pro League

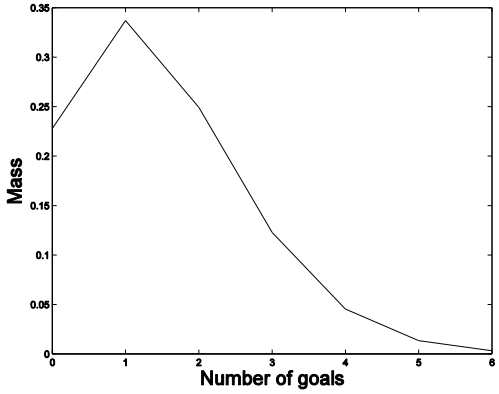

Figure 6: Continuous plot of the Poisson probability mass function

Royal Standard de Liège goals	0	1	2	3 or more
Poisspdf(1.1412)	0.2279	0.3370	0.2492	0.1229
Bwin	0.2564	0.3636	0.2174	0.2105

Table 3: Comparing probabilities

As can be seen we have more data for Royal Standard de Liège than for SV Zulte Waregem, this is due to the fact that Royal Standard de Liège played in the Jupiler Pro League for a longer period. For Figure 3 and 5 we can again suggest a Poisson distribution. Figure 4 and 6 again represent a continuous plot of the probability mass function of the Poisson distribution. In Figure 4, the mean of the Poisson distribution is 1.6761, the mean of the home goals scored by SV Zulte Waregem in a game. In Figure 6, the mean of the Poisson distribution is 2.7201, the mean of away goals scored by Royal Standard de Liège in a game.

Will these probability distributions enable us to predict football matches more accurately? For 1 and 2 goals, the probabilities are again rather close to the ones offered by Bwin. In other cases however, we have a wide spread. Our calculated probabilities here are even less accurate then in the previous case. This is because teams have different levels of attack and defense. We cannot simply look at the average amount of goals scored by a team and extrapolate that to all games. Every opponent is different so we will have to take the strength of the opponent into account. For example, SV Zulte Waregem is expected to score less goals against Royal Standard de Liège then against an average team considering the good defense of Royal Standard de Liège [4]. Also recent performances will be a much more important factor in predicting the scoring abilities of a team then performances of 7 years ago. This is another fact we will have to take into account when setting up our model.

Thus after going through these descriptive statistics about the Jupiler Pro League it seems clear that we will need a more extended model to be able to accurately predict football matches. A necessary extension will be to assign levels to attack and defense as previously stated.

[4] http://www.sport.be/nl/jupilerproleague/ranking.html?comp=2692#competition-ranking. Royal Standard de Liège conceded the least goals in the Jupiler pro league this year.

3 Methodology

In this section we will present our dynamic statistical modelling framework, justify the assumptions and elaborate our prediction method.

3.1 Statistical model

Let's denote (X_{it}, Y_{jt}) as the goals scored by home team i and visiting team j in week t. with X_{it} an Y_{jt} being independent Poisson random variables.

Each observed match outcome is assumed to be sampled from a double Poisson distribution with probability mass function.

$$P(X_{it} = x, Y_{jt} = y) = \frac{{\lambda_{it}}^{x}}{x!} e^{-\lambda_{it}} \frac{{\lambda_{jt}}^{y}}{y!} e^{-\lambda_{jt}}$$

Firstly we will consider a linearized version of the Poisson model. As Maher [2] we include attack and defense strengths, and we account for a home ground advantage. Let's denote this by model 1.

$$\mathrm{X_{it}} = \delta_x + \alpha_{it} - \beta_{jt} + \varepsilon_{ijt}^{x}, \qquad E(\varepsilon_{ijt}^{x} | \alpha_{it}\beta_{jt}) = 0 \qquad 5.1$$

$$\mathrm{Y_{jt}} = \delta_y + \alpha_{jt} - \beta_{it} + \varepsilon_{ijt}^{y}, \qquad E(\varepsilon_{ijt}^{y} | \alpha_{jt}\beta_{it}) = 0 \qquad 5.2$$

$\delta_x - \delta_y$ is the home ground advantage, we take the home ground advantage to be equal amongst all teams and constant over time as was shown by Dixon and Coles [1]. α_{it} and β_{it} denote the latent strength of team $i's$ attack and defence respectively. Since we include an intercept in the equations we may set the unconditional means of α_{it} and β_{it} to zero. $E[\alpha_{it}\,] = 0$ and $E[\beta_{it}\,] = 0$. It follows then that when we take the expected value with respect to the goals scored in a match.

$$\hat{\delta}_x = \hat{E}[X_{it}]$$

$$\hat{\delta}_y = \hat{E}[Y_{jt}]$$

We make the model dynamic by allowing the team properties (the attack and defense strengths) to change over time. Suppose that α_{it} and β_{it} evolve according to

$$\alpha_{it} = \phi_\alpha \alpha_{it-1} + \psi_\alpha \varepsilon_{ij,t-1}^{\alpha} \qquad 5.3$$

$$\beta_{it} = \phi_\beta \beta_{it-1} + \psi_\beta \varepsilon_{ij,t-1}^{\beta} \qquad 5.4$$

where

$$\varepsilon_{ij,t-1}^{\alpha} = h_{i,t-1}\varepsilon_{ij,t-1}^{x} + (1 - h_{i,t-1})\varepsilon_{ij,t-1}^{y} \qquad 5.5$$

$$\varepsilon_{ij,t-1}^{\beta} = -h_{i,t-1}\varepsilon_{ij,t-1}^{y} - (1 - h_{i,t-1})\varepsilon_{ij,t-1}^{x} \qquad 5.6$$

in which $h_{i,t-1} = 1$ if i played at home at $t-1$ and $h_{i,t-1} = 0$ otherwise. For model 1 we will have to estimate four coefficients. Let's collect all unknown coefficients of model 1 in the parameter vector $\theta_l = (\phi_\alpha, \phi_\beta, \psi_\alpha, \psi_\beta)$'.

Note that we assume ϕ_α and ϕ_β as well as all ψ_α and ψ_β to be equal for all teams. As Koopman and Lit [11], we expect the attack and defense strengths to be evolving slowly over time. Also this still leaves room for the attack and defense strengths of different teams to evolve differently over time since they also depend on game and team specific error terms. For example, when home team A scores more goals than expected, the error term of equation 5.1 will be positive which translates through equation 5.5 into a positive value of the error term of the dynamic process of the attack strength. And this will update the attack strength of team A positively by the amount $\psi_\alpha \varepsilon_{ij,t-1}^{\alpha}$. Given that $\psi_\alpha > 0$.

The transition of the attack and defense strengths of the teams is determined by pseudo weeks which means that the previous period $t-1$ actually is the previous match played. For example, in an ideal world, every team would play every week such that every week the attack and defense strengths would be updated. However in reality, it happens that a match is postponed to a later date due to unforeseen circumstances (weather conditions, hooliganism,..). When this happens the teams don't play that week, let's call this week 1. Suppose then that their next game is played a week after, week 2. We then base the attack and defense strengths of that team in week 2 on the last updated strengths which are found from the previous game 2 weeks ago, week 0. This also implies that when a new season starts we set the attack and defense strengths of the teams equal to their values at the end of the previous season. For the promoting teams that haven't played in the Jupiler Pro League before, we set the attack and defense strengths equal to zero.

Next we implement a simplified Poisson model. Let's denote this by model 2.

$$\mathrm{X_{it}} = \exp(\delta_x + \alpha_{it} - \beta_{jt}) + \varepsilon_{ijt}^{x}, \qquad E(\varepsilon_{ij,t}^{x} | \alpha_{it}\beta_{jt}) = 0 \qquad 5.5$$

$$\mathrm{Y_{jt}} = \exp(\delta_y + \alpha_{jt} - \beta_{it}) + \varepsilon_{ijt}^{y}, \qquad E(\varepsilon_{ij,t}^{y} | \alpha_{jt}\beta_{it}) = 0 \qquad 5.6$$

Again we assume the home ground advantage to be equal amongst all teams and constant over time. However, the home ground advantage is computed differently and must be interpreted in an alternative way. Since

$$\exp(\delta_x + \alpha_{it} - \beta_{jt}) + \varepsilon_{ijt}^{x} = \exp(\delta_x) * \exp(\alpha_{it} - \beta_{jt}) + \varepsilon_{ijt}^{x}$$

the home advantage $\exp(\delta_x) - \exp(\delta_y)$ becomes a measure of the difference in the multiplicative terms $\exp(\delta_x)$ and $\exp(\delta_y)$, whereas in model 1, δ_x and δ_y are additive such that the measure is of the difference of two additive terms.

For model 2 we will have to estimate 6 coefficients. We collect all unknown coefficients of model 2 in the parameter vector $\theta_{nl} = (\phi_\alpha, \phi_\beta, \psi_\alpha, \psi_{\beta,} \delta_x, \delta_y)$'.

3.2 Estimation procedure

For model 1 we take the following approach. At the beginning of the first season we set attack and defense strengths equal to their unconditional mean.

$\alpha_{i1} = \beta_{i1} = 0$ for all i

For the next seasons we set α_{is} and β_{is} for s being the beginning of next season to be equal to their value at the end of the previous season. α_{is} and β_{is} are already conditional on previous seasons then and thus no longer zero. Now, after each week $t \geq 1$,the residuals of model 1 can be computed from 5.1 and 5.2.

$$\hat{\varepsilon}_{ijt}^{x} = X_{it} - \hat{\delta}_x + \hat{\alpha}_{it} - \hat{\beta}_{jt}$$

$$\hat{\varepsilon}_{ijt}^{y} = Y_{it} - \hat{\delta}_y + \hat{\alpha}_{jt} - \hat{\beta}_{it}$$

When we have the model residuals $\hat{\varepsilon}_{ijt}^{x}$ and $\hat{\varepsilon}_{ijt}^{y}$ we can compute $\hat{\varepsilon}_{ij,t}^{\alpha}$ and $\hat{\varepsilon}_{ij,t}^{\beta}$ through equations 5.5 and 5.6. Using these residuals we obtain $\hat{\alpha}_{it+1}$ and $\hat{\beta}_{it+1}$ according to 5.3 and 5.4 for a given parameter vector $\hat{\theta}_l$. This implies that the model residuals ε_{ijt}^{x} and ε_{ijt}^{y} also depend on the parameter vector θ_l. Such that we can estimate the parameter vector using least squares.

$$\hat{\theta}_l = \underset{\theta_l \, \epsilon \, \Theta}{\operatorname{argmin}} \sum \left[\left(\varepsilon_{ijt}^{x} \right)^2 + \left(\varepsilon_{ijt}^{y} \right)^2 \right]$$

For model 2 we roughly follow the same approach, we now compute the residuals from 5.5 and 5.6.

$$\hat{\varepsilon}_{ijt}^{x} = X_{it} - \exp(\hat{\delta}_x + \hat{\alpha}_{it} - \hat{\beta}_{jt})$$

$$\hat{\varepsilon}_{ijt}^{y} = Y_{it} - \exp(\hat{\delta}_y + \hat{\alpha}_{jt} - \hat{\beta}_{it})$$

Again we apply least squares to estimate the parameter vector such that

$$\hat{\theta}_{nl} = \underset{\theta_{nl} \, \epsilon \, \Theta}{\operatorname{argmin}} \sum \left[\left(\varepsilon_{ijt}^{x} \right)^2 + \left(\varepsilon_{ijt}^{y} \right)^2 \right]$$

4 Results and discussion

In this section we will start by presenting our results. Next we will demonstrate how we can use these results to make predictions and we end by carrying out an out-of-sample forecast for the regular season 2013/2014 of the Jupiler Pro league.

4.1 Parameter estimates

For our time series panel of goals scored by teams in Belgian Jupiler Pro League during the 9 seasons from 2005/2006 till 2013/2014, the parameter estimates for model 1 are presented in Table 4 and the parameter estimates for model 2 are presented in table 5.

$\widehat{\theta}_l$		F-stat	p-value
$\widehat{\phi}_\alpha$	0.9980	226,8203	0.0000
$\widehat{\phi}_\beta$	0.9925	119,2214	0.0000
$\widehat{\psi}_\alpha$	0.0224	251,4463	0.0000
$\widehat{\psi}_\beta$	0.0270	142,3226	0.0000

Table 4: Parameter estimates model 1 with δ mean of goals over all seasons.

Table 4 presents the estimates of the coefficients of the latent dynamic processes for the attack and defense strengths. To check if our coefficients are significantly different from zero, we perform an F-test for each parameter. We use the following procedure: we first estimate the full model, then we set one parameter equal to zero and estimate the nested model. This gives us two SSR's which allows us to compute the F-ratio. We find that the autoregressive coefficients are significantly different from zero. The estimates are also close to one[5] which means that the attack and defense strengths are highly persistent. The impact of the residuals on the attack and defense strengths are also found to be significantly different from zero, this implies that the number of goals scored or conceded in the recent history plays a role in determining the attack and defense strengths in this period, confirming the literature. Although this effect is rather small as you can see from the estimated values. In model 1 we can simply calculate the home ground advantage by

$$\hat{\delta}_x - \hat{\delta}_y = \hat{E}[X_{it}] - \hat{E}[Y_{jt}] = 1.5789 - 1.1412 = 0.4378$$

[5] Using an F-test, we cannot reject the nullhypothesis that $\phi_\alpha = 1$ at the 5% level and thus we cannot rule out the case of a random walk for the attack strength. However we do reject the nullhypothesis of $\phi_\beta = 1$ at the 5% level.

$\widehat{\theta}_{nl}$		F-stat	p-value
$\widehat{\phi}_{\alpha}$	0.9979	227,6835	0.0000
$\widehat{\phi}_{\beta}$	0.9917	117,3395	0.0000
$\widehat{\psi}_{\alpha}$	0.0149	253,2063	0.0000
$\widehat{\psi}_{\beta}$	0.0187	141,6677	0.0000
$\widehat{\delta}_{x}$	0.4256	60,1575	0.0000
$\widehat{\delta}_{y}$	0.1114	4,5309	0.0333

Table 5: parameter estimates model 2

For model 2 we again used an F-test to check if our parameters are significantly different from zero. For inference on non-linear least squares estimation, we find that inference based on F-tests is more reliable than test or confidence intervals based on standard errors [17]. The F-test is calculated slightly different here since we have more parameters to be estimated. We again obtain autoregressive coefficients that are significantly different from zero and very persistent[6]. The impact of the residuals on the attack and defense strengths is again significantly different from zero but rather small. Important to note is that we are now estimating the home ground advantage by means of least squares. We find that the parameters which define the home ground advantage are significantly different from zero. This implies that the home ground advantage is

$$\exp(\hat{\delta}_x) - \exp(\hat{\delta}_y) = 1.5305 - 1.1178 = 0.4127$$

For the remainder of this thesis we will work with model 2.

4.2 Prediction

Now that we have found the estimates it becomes possible to make predictions about football matches. Using our estimated parameters we can find the estimates of the latent variables attack and defensive strengths for all i using equations 5.3 and 5.4. Since we use data up till the end of the regular competition of 2013-2014. Thus $t+1$ represents the first week of the play-off games. When we consider model 2 we obtain the following estimates for all 26 teams.

[6] Using an F-test, we cannot reject the null hypothesis that $\phi_{\alpha} = 1$ at the 5% level and thus we cannot rule out the case of a random walk for the attack strength. However we do reject the null hypothesis of $\phi_{\beta} = 1$ at the 5% level.

Team	$\widehat{\alpha}_{it+1}$	$\widehat{\beta}_{it+1}$
Royal Standard de Liège	0.2963	0.4013
Club Brugge KV	0.3470	0.2064
RSC Anderlecht	0.3894	0.2797
SV Zulte Waregem	0.1518	0.1506
KSC Lokeren	0.1748	0.2095
KRC Genk	0.2135	0.0308
KAA Gent	0.0668	0.1260
KV Kortrijk	-0.0241	0.0320
KV Oostende	-0.1135	-0.0302
R. Sporting du Pays de Charleroi	-0.1588	0.0215
Cercle Brugge KSV	-0.1703	-0.2285
K. Lierse SK	-0.1824	-0.1326
Yellow Red KV Mechelen	-0.0384	-0.0833
KVRS Waasland - SK Beveren	-0.2437	0.0885
Oud-Heverlee Leuven	-0.0687	-0.0878
RAEC Mons	-0.0367	-0.1711
K. Beerschot AC	-0.1331	-0.2607
KVC Westerlo	-0.1039	-0.1667
KAS Eupen	-0.1146	-0.0549
Sint-Truiden VV	-0.2291	-0.1744
KSV Roeselare	-0.1541	-0.1757
AFC Tubize	-0.0912	-0.2871
FCV Dender EH	-0.0551	-0.1214
RAA Louviéroise	-0.2163	-0.0426
Brussels FC	-0.2556	-0.1658
Exelsior Mouscron	-0.0675	-0.0382

Table 6: Latent variables

We find that RSC Anderlecht has the best attack which is intuitive since RSC Anderlecht has always been one of Belgium's top teams and therefore also in the last 8 years[7]. This year they didn't perform as good as previous years in the regular competition. However this was mostly due to their defense, they still scored the most goals in the regular competition which justifies the high value of the attack strength estimate. We also find that Royal Standard de Liège has the best defense. Last couple of years Royal Standard de Liège also was a certitude in the top league of the competition. This year they performed exceptionally, especially their defense (they only conceded 17 goals) which is also reflected by our findings.

We also obtain negative values for both the attack and defense strength estimates of the two teams ranked last in the regular competition: RAEC Mons and Oud-Heverlee Leuven[8]. However, we find that Cercle Brugge KSV and K. Lierse SK, who were ranked higher in the season 2013/2014, score even worse on both estimates. This is partly due to previous performances. These still have influence, since we obtained in the previous subsection that attack and defense strengths evolve slowly over time. While Cercle Brugge KSV and K. Lierse SK performed really bad in the season 2012/2013, they were ranked on the 16th and 14th place of the regular competition respectively, RAEC Mons and Oud-Heverlee Leuven had a good season and ended respectively on the 7th and 10th place. For example, we can compare the sequences of the attack and defense strength estimates of Cercle Brugge KSV and RAEC Mons during the season 2013/2014.

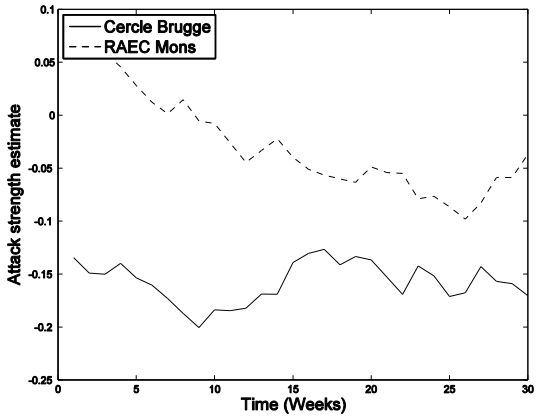

Figure 6: Attack strength estimate of KSV Cercle Brugge and RAEC Mons in the season

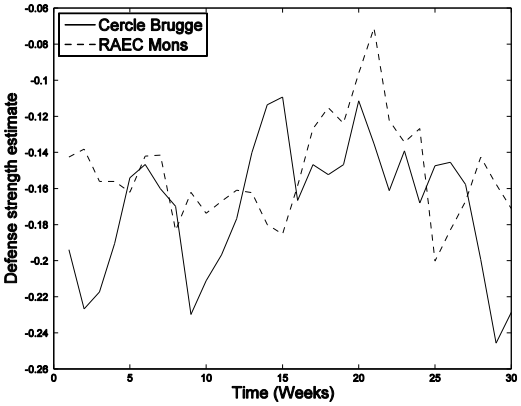

Figure 7: Defense strength estimate of KSV Cercle Brugge and RAEC Mons in the season

We see that RAEC Mons started off the season with higher values due to their good performance in the previous season, while KSV Cercle Brugge started off with really low values due to their bad performance previous season[9]. For the attack, the estimate of RAEC Mons reduced sharply, this might be due to the departure of their top goal scorer

[7] http://en.wikipedia.org/wiki/R.S.C._Anderlecht.

[8] http://www.sport.be/nl/jupilerproleague/ranking.html?comp=2692#competition-ranking.

[9] http://nl.wikipedia.org/wiki/Eerste_klasse_2012-13_(voetbal_Belgi%C3%AB).

Jérémey Perbet[10]. The attack strength estimate of KSV Cercle Brugge did not change much compared to last season, although at some parts of the season it was higher. However, they ended the season with some bad performances which is reflected by the end value of their attack strength estimate.

Apart from the higher starting value for RAEC Mons, the defense strength estimate of both teams was pretty similar throughout the whole season. The defense strength estimate of Cercle Brugge KSV was better than last year for most of the season. However, at the end of the season, as previously stated, they performed really bad and thus ended the season with a lower defense strength estimate than RAEC Mons. Also to be noted is that both attack and defense strengths estimates of RAEC Mons reduced as compared to the season 2012/2013, especially the attack strength estimate. Hence the model might suggest that if RAEC Mons would not have sold Jérémy Perbet they might still be playing in the Jupiler Pro league next year.

These latent variables allow us to find the expected number of home and away goals of a match played at $t+1$. When we take this expected number of goals to be the mean of the Poisson distribution, it allows us to compute the probabilities on the number of goals scored. Once computed we can compare these probabilities to the odds offered by the gambling sites to determine our strategy. Let's consider Club Brugge KV vs KSC Lokeren as an example. For both teams this was the first game they played after the regular competition so we can use the values of attack and defense strengths given in the table above. We get that for our Poisson distribution.

$$\hat{\lambda}_{it+1} = \exp\left(\hat{\delta}_x + \hat{\alpha}_{it+1} - \hat{\beta}_{jt+1}\right) = \exp(0.4256 + 0.3470 - 0.2095) = 1.7561$$

which is the estimated mean of the Poisson distribution for the goals scored by Club Brugge KV against KSC Lokeren.

$$\hat{\lambda}_{jt+1} = \exp(\hat{\delta}_y + \hat{\alpha}_{jt+1} - \hat{\beta}_{it+1}) = \exp(0.1114 + 0.1748 - 0.2064) = 1.0830$$

which is the estimated mean of the Poisson distribution for the goals scored by KSC Lokeren against KV Club Brugge. We can now calculate chances for the number of goals scored in a match or by a team using the same method as in subsection 2.2. Even more interesting is that we can define the probabilities of a win, a draw or a loss. We can find these probabilities by conducting the following calculations.

Probability of win for the home team is:

$$P\left(X_{it} > Y_{jt}\right) \qquad 4.1$$

Probability of a draw is :

$$P(X_{it} = Y_{jt}) \qquad 4.2$$

[10] 37 goals in 57 matches for RAEC Mons.

Probability of a loss for the home team is:

$$P(X_{it} < Y_{jt}) \qquad 4.3$$

	$P(Club\ Brugge\ KV\ wins)$	$P(draw)$	$P(KSC\ Lokeren\ wins)$
Club Brugge KV - KSC Lokeren	0.5314	0.2352	0.2334

Table 7: Probabilities

There exist several different betting strategies which can be followed. We consider a betting strategy similar to the one proposed by Koopman and Lit [11]. We calculate the expected values for each outcome. The expected value of a unity bet on an event A is given by

$$\begin{aligned} EV(A) &= P(A)\ x\ [odds(A) - 1] - P(Not\ A) \\ &= P(A)\ x\ odds(A) - 1 \end{aligned} \qquad 4.4$$

With the event A representing a win, a loss or a draw of the home team. A basic betting strategy could be to bet on all events for which the expected value is positive $E(A) > 0$. Using odds from Bwin[11] we obtain the following expected values.

	$EV(Club\ Brugge\ KV\ wins)$	$EV(draw)$	$EV(KSC\ Lokeren\ wins)$
Club Brugge KV - KSC Lokeren	-0.07005	-0.20032	-0.03988

Table 8: Expected values

Since all the expected values are negative we should not bet on this match given the odds from Bwin. When we apply the same reasoning for the second match of the first week of the play offs KRC Genk vs SV Zulte Waregem, we obtain the following expected values.

[11] http://www.betexplorer.com/soccer/belgium/jupiler-league/lokeren-club-brugge/ERrW67O5/.

	$EV(KRC\ Genk\ wins)$	$EV(draw)$	$EV(SV\ Zulte\ Waregem\ wins)$
KRC Genk-SV Zulte waregem	-0.0798	-0.2086	0.0078

Table 9: Expected values

This implies that according to our basic betting strategy we should bet on a victory of SV Zulte Waregem.

4.3 Prediction review

We end this section by testing the out-of-sample performance of our model for the betting on a win, a loss or a draw of the home team, for a selection of matches during the season 2013/2014. In our betting evaluation we carry out a one step ahead forecasting study. We forecast the probabilities of all possible outcomes (events) of a match by using a rolling window approach. At time t, we estimate all the parameters of the model and thus forecast the means of our double Poisson distribution using data up till time t. We then have the distributional properties of the next 8 games implied by our double Poisson model with the unknown means replaced by our forecasts. This enables us, as argued before, to forecast the probabilities of all possible outcomes of a match and hence the probabilities of a win, a loss or a draw for the home team by using equations 4.1, 4.2 and 4.3. Once these probabilities are calculated for all 8 next matches, we can visit the bookmaker's office. We apply this approach every week and thus we obtain probabilities of every outcome of a match for the entire season of 2013/2014. Using equation 4.4 we are now able as before to compute the expected value of every bet.

As argued before a basic betting strategy could be to bet on all events for which the expected value is positive (larger then 0). However, just like Koopman and Lit [11], we prefer a less risky betting strategy which is based on the following guidelines. First, we only bet on "quality" events which are defined as bets with EV's that exceed some benchmark. For example, when we take a benchmark of 0.1 we would not bet on the game SV Zulte Waregem vs KRC Genk, as it is not a quality bet, in contrast to a benchmark of 0. Second, we do not bet on longshots either which are defined as small probability events with such high odds that they pass as quality events. We take odds higher than 7 to be longshots.

When we follow the basic betting strategy defined above and bet an unity value on each quality event and do not bet on longshots, we can calculate the expected and actual profit for all our bets in the 2013/2014 regular season for a range of values for the benchmark. Since the games of the season 2013/2014 are already played we are able to compare the placed bet with the actual outcome and like that calculate the actual profit. The sample variance of the computed profit is obtained by the bootstrap method based on 10000

bootstrap samples, we have carried out a standard bootstrap method. The odds for betting are offered by bet365[12]

Also to be noted is that random betting ensures a considerable loss due to the bookmaker's edge. Let's take the game KRC Genk - SV Zulte Waregem from above as an example. The implied probabilities given by the bet365[13] odds have been $1/2$, $1/3.50$ and $1/3.20$ in the respective order of a win, a loss and a draw by the home team. The sum of these probabilities is 109.8%. Everything above 100% is the profit of the bookmaker. This means that the expected profit under random betting of a unity value is roughly -0.1. Thus in order to obtain a positive return our estimates will have to be sufficiently more accurate than the bookmakers.

When we calculate the actual percentage gain for different betting strategies and so for different benchmarks we find the following.

Minimum expected value (benchmark)	Actual percentage gain
-0.05	-0.0505
0	0.0446
0.05	0.0582
0.1	0.1180
0.2	0.3410

Table 10: Percentage gains

As expected we find that when the benchmark goes up, we get larger percentage gains. However since the number of bets also goes down in general, the variance on these bets will become larger.

[12] http://www.betexplorer.com/soccer/belgium/jupiler-league/results/?stage=SYdaMVJH.

[13] http://www.betexplorer.com/soccer/belgium/jupiler-league/matchdetails.php?matchid=j7l8Iq6K.

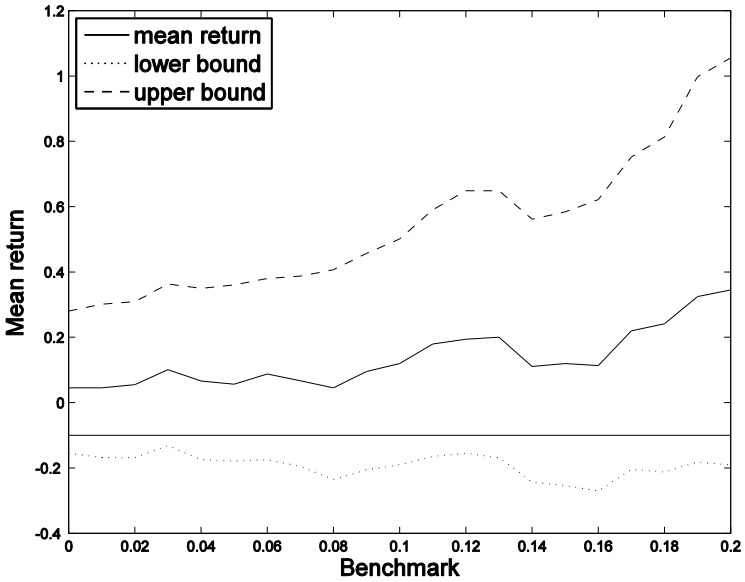

Figure 8: Mean return plotted against the benchmark; (return of -0.1 is the expected return under random betting, which is due to the bookmakers take of 10 % each match.) We have 90% bootstrap confidence intervals for the mean (this implies that we have generated the curve 10000 times). The mean return is generated only when there are at least 29 sample values.

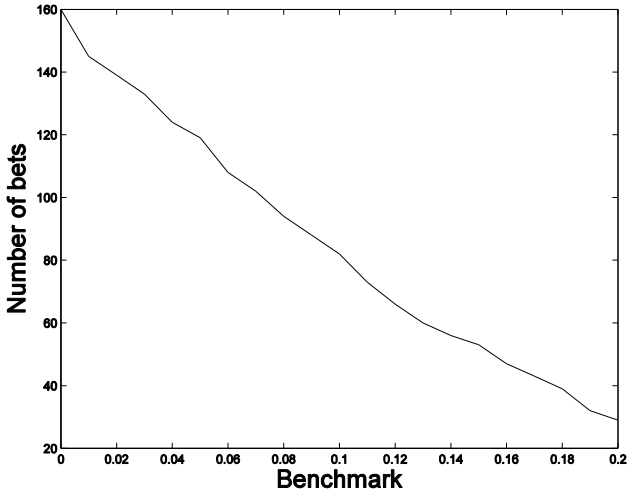

Figure 9: Number of bets

When we take a benchmark higher than 0.2 the events become too rare to further investigate. In Figure 8 we can see that the mean return is an increasing function of the benchmark. You also see the variance increasing with the benchmark, this is mainly due to the fact that the number of bets is decreasing as you can see on Figure 9. There is considerable variability in the plot which makes it difficult to draw definitive conclusions. However we do have a positive mean value for all benchmarks larger than 0. As Dixon and Coles [1] we thus claim that we have succeeded in deriving a model for estimating football match outcomes which is the basis of a betting strategy with positive returns.

A last remarkable note here is that following this strategy we find that we almost never bet on draws. This could mean that our model underestimates draws, as other authors found when using the double Poisson model. This leaves room for extensions on the

model. We could probably improve the fit by including a correction factor or using a bivariate Poisson distribution.

General Conclusion

We have presented a simplified Poisson model for the analysis and forecasting of football matches. Our model takes a match result as a pairwise observation that is assumed to come from a double Poisson distribution. The means of the Poisson distributions depend on the attack and defense strengths of the teams which are allowed to evolve over time. The means are also subject to a fixed coefficient for home ground advantage if applicable.

For our empirical study we used a dataset of match results from 9 seasons of the Belgian Jupiler Pro League ranging from 2005/2006 till 2012/2013. For our prediction review we used the same 9 seasons, only we now used the season 2013/2014 as out of sample evaluation period for the forecasting of football match results. Although we have presented promising results (prediction review) we believe further improvements can be made on our model. For example, we could use a correction factor like Dixon and Coles [1] and Rue and Salvesen [7] or use a bivariate Poisson distribution like Karlis and Ntzoufras [9], Goddard [10] and Koopman and Lit [11] to solve the problem of underestimation of the number of draws by taking into account the fact that goals scored by the home and away team are not independent and hence improve the model fit. Another interesting modification could be to include a psychological effect as extra explanatory variable, as Rue and Salvesen [6]. Or we could also include more information about matches. For example, the number of players injured before a game or the travelling distance of the visiting team.

For now we have used simple betting strategy: we bet on all outcomes for which the expected value of our bet exceeds a specified level (benchmark). For positive levels we have shown that this strategy yields positive mean returns. These mean returns can be improved even further when making use of more advanced betting strategies. For example, one can look for the highest odds provided at the various gambling sites. Also we have only bet on a possible win, draw or loss of the home team. A recent phenomenon on the betting market is the opportunity to bet on multiple games at once, one could investigate whether it is possible to obtain larger positive returns here. If one prefers riskier bets it is also worth investigating whether bets on the number of goals could be profitable.

Finally, before starting this thesis we asked the question if we can use statistics to predict football matches and make a positive return on the betting market. After setting up a statistical model and carrying out a prediction review we can now answer the question affirmative. It is possible to predict the outcomes of football matches and to set up a betting strategy. A basic betting strategy already leads to positive mean returns for all positive benchmarks. We have thus succeeded in deriving a model from which the forecasts are sufficiently accurate to gain a positive return over the bookmaker's odds.

Sources

[1] Dixon Mark.J. and Coles Stuart G. (1996) Modelling association football scores and inefficiencies in the football betting market Appl Statist 46, No2, pp265-280.

[2] Maher,M.J. (1982) Modelling association football scores, Statistica Neerlandica 36,nr.3.

[3] Reep C., Pollard R. and Benjamin B. (1971) Skill and Chance in Ball games, Journal of the Royal Statistical Society. Series A (General), Vol. 134, No 4 pp. 623-229.

[4] Reep C. and Benjamin B. (1968) Skill and chance in association football, Journal of the Royal Statistical Society. Series A (General), Vol. 131, No. 4 (1968), pp.581-585.

[5] Hill I. D. (1974) Miscellanea, Journal of the Royal Statistical Society. Series C(applied Statistics), Vol,23, No2 (1974), pp 203-208.

[6] Rue Havard and Salvesen Oyvind (2000), Prediction and retrospective analysis of soccer matches in a league, The statistician 49,pp399-418.

[7] Owen Allun (2011), Dynamic forecasting models of football match outcomes with estimation of the evolution variance parameter, IMA journal of Management Mathematics Advance Access published January 20, 2011.

[8] Crowder Martin, Dixon Mark, Ledford Anthony and Robinson Mike (2002), Journal of the Royal Statistical Society. Series D (The Statistician) Vol.51, No.2 pp 157-168.

[9] Karlis Dimitris and Ntzoufras Ioannis (2003), Analysis of sports data using bivariate poisson models, The statistician (2003) 52, part 3,pp381-393.

[10] Goddard John (2005) Regression models for forecasting goals and match results in association football, International journal of Forecasting 21 331-340

[11] Koopman Siem Jan and Lit Rutger (2012), A dynamic bivariate poisson model for analysing and forecasting match results in the English Premier league, working paper.

[12] Durbin, J. and S.J. Koopman (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models. Biometrika 84(3) 669-684.

[13] Shepard, N. and M.K. Pitt (1997) Likelihood analysis of non-Gaussian measurement time serie. Biometrika 84(3),653-667.

[14] Shepard Neil and De Jong Piet (1995), The simulation smoother for tiome series models, Biometrika, 82,2,pp339-350.

[15] Pollard Richard (2008), Home advantage in Football: A current review of an unsolved puzzle, The Open Sports Sciences Journal,1,12-14.

[16] Whitacker Gavin (2011), The bivariate Poisson Distribution and its Applications to Football, School of Mathematics and Statistics, Newcastle University.

[17] Smyth. Gordon K. (2002), Nonlinear regression, Encyclopedia of Environmetrics, John Wiley, Ltd, Chichester.

FACULTY OF BUSINESS AND ECONOMICS
Naamsetraat 69 bus 3500
3000 LEUVEN, België
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be