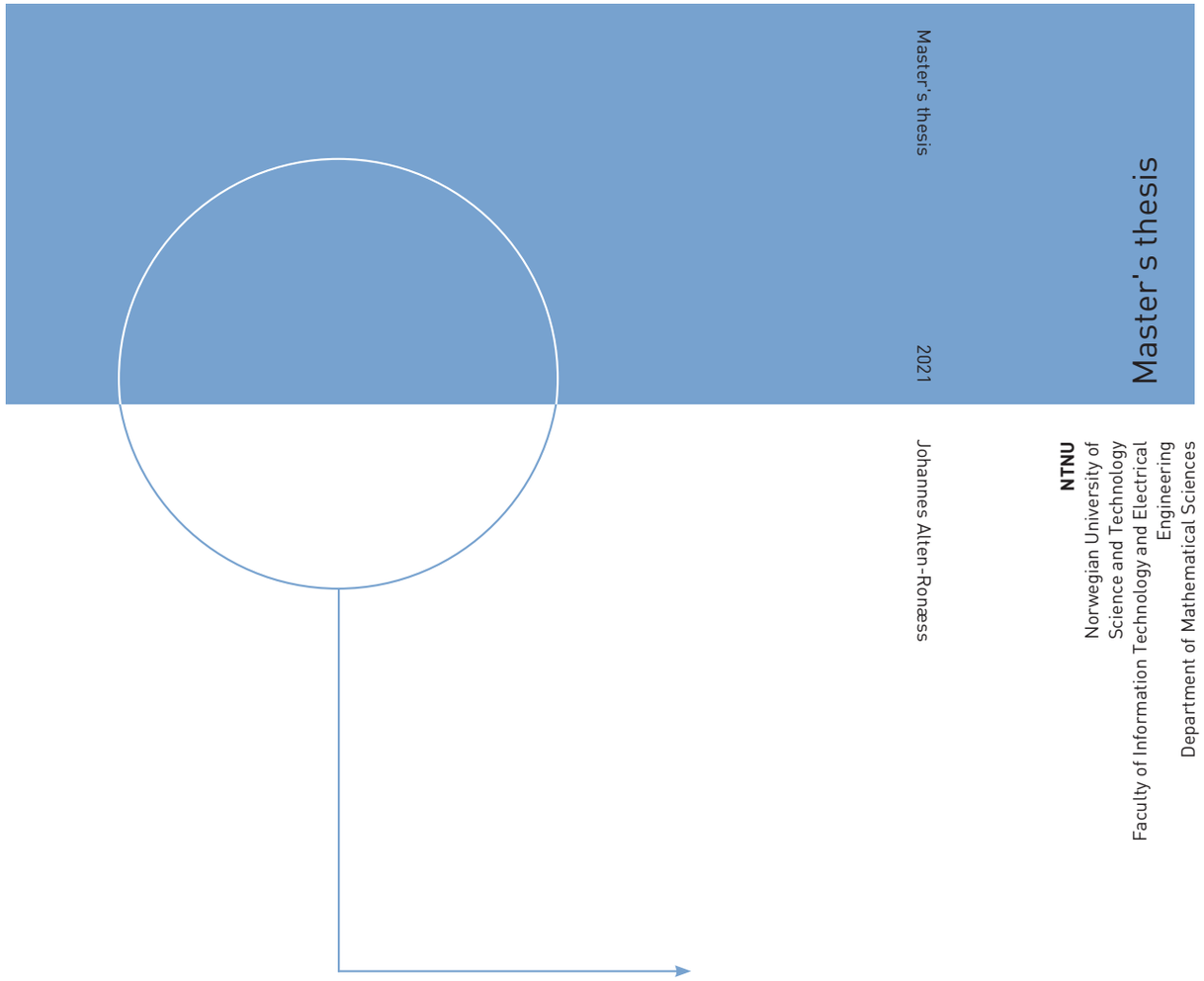


Master's thesis

2021

Johannes Alten-Ronæss

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Master's thesis

Johannes Alten-Ronæss

Predictions of football results in the Premier League and its use on the betting market

June 2021

NTNU
Norwegian University of Science and Technology

NTNU
Norwegian University of Science and Technology

NTNU

Norwegian University of
Science and Technology

Predictions of football results in the Premier League and its use on the betting market

Johannes Alten-Ronæss

Industrial Mathematics
Submission date: June 2021
Supervisor: Jarle Tufto

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

In this Master's thesis I introduce different parametric statistical models, similar to Generalized Linear mixed models, and I use the estimates from the models to predict the results in football matches in the Premier League. I compare different models from the literature, and investigate how the models perform on the betting-market the last 12 seasons of the Premier League (2009/2010-2020/2021). The models assume that the results in football matches can be predicted by the difference in strength for the teams, and what team plays at home. The models can be divided into two main categories. The first category estimates the number of goals scored by each team in the match, and uses this to make inferences about the probability for a home win, draw and away win. A Poisson distribution will be assumed for the number of goals scored. The second category estimates the probabilities for home victory; draw; and away victory directly, and a generalized Bradley-Terry model will be assumed for the probabilities. The strength of the teams can be constant, or follow Brownian motions.

There have been no supporters present at any matches the last year, due to the ongoing pandemic, and I find that the home field advantage has disappeared. Teams used to score 30% more goals playing at home than they did playing away, but now they do not score more goals at home than away. This is due to a combination of home teams scoring fewer, and away teams scoring more goals, than they used to do. The total amount of goals scored is the same as before the pandemic.

The results from the betting-market is that it is possible to earn money on the betting-market over a long time-period, but not much (with the models discussed in this project), and one does not earn money in every season. The best model earned 3.67% in total, over the defined time period of 11 years.

I find that the Poisson distribution is a good fit to model the number of goals scored by teams in the Premier League, but a small under-dispersion is detected. I propose the use of the generalized Poisson distribution to model the variance in the data correctly.

Contents

1 Introduction

1.1 Modelling sport results

Modelling sport results have interested many statisticians for decades, and football results are no exception. One of many cliches in football is "The ball is round", this is usually interpreted as anything can happen and no one can predict the result of a football match in advance. This is probably true, and a statistician would say that the result is a stochastic variable, but just because you can not deterministically predict the results of a football match in advance, does not mean that you can not predict the probability for the different outcomes. The two main ways to model football results are; (1) predicting the number of goals scored by each team, and from there infer about the result. (2) Directly predict the probabilities for the possible results. The possible results being home victory, draw and away victory. The aim for any model in sport results can be exploratory, but often the aim is to beat the bookie and try to make money on the betting market.

In 1982 Maher introduced a Poisson model for the number of goals scored by each team in different matches in England [Maher, 1982]. His proposed model assumed that the number of goals scored by a team is Poisson distributed, and the mean is dependent on how good the attacking team is at attacking and how good the defending team is at defence. Dixon and Coles made a more complex model in 1997 where they introduced the use of a weighted bivariat-Poisson for each match[Dixon and Coles, 1997]. The Poisson model was weighted to decrease the chances of low scoring draws and increase the chances of the results 1-0 and 0-1. They also introduced a time dependent model, where they weighted recent results more heavily in the likelihood. Rue and Salvesen introduces a new model. This model is also time dependent, but has more in common with Maher than Dixon and Coles [Rue and Salvesen, 2000]. They look at seasons in the two top divisions of English football, using almost the same model as Maher [1982]. However, they let the attacking and defending strength vary with time as a Brownian motion. They also introduce some betting strategies, meaning how best to allocate your money for maximum expected return with as little variance as possible. In Langseth [2013] he mainly focuses on the betting aspect, and he looks at different betting strategies. However, he does propose a new model. This model is more data intensive than the models previously mentioned. He incorporates the number of shots, shots on target and the number of goals scored in each match by each team, into the model. The idea behind the model is that to know how good a team is, you have to look beyond just the result of the matches. And see what their offensive and defensive output is, to be able to make better predictions for the future. All these articles have used some sort of goal counting model, often a variation of the Poisson distribution. Cattelan et al. [2013] models the result of a season of Serie A (The top division of Italy), using a generalized Bradley-Terry model. This is an example of a model that models the results directly and not by first finding the distribution for the goals scored by each team. In this project I will compare the results from different models on 12 different seasons of the Premier League, that is the top division in England. I will use both goal counting models and models that directly estimates the probabilities. I will try to find which model fits the observations best and which model is best suited to earn money on the betting market.

1.2 Premier League

Premier League (PL) is the top division in English football. It consist of 20 teams where every team plays against every other team two times each season, one home and one away match. A season normally lasts from August to May. There is a total of 38 rounds of 10 matches each, giving a total of 380 matches through a season. A team gets 3 points for a victory; 1 point for a draw; 0 points for a loss. A team gets the same number of points whether they win at home or away, but it is often assumed that a team is more likely to win at home. This is often called the home field advantage. At the end of the season the three teams with fewest points are relegated to the second tier league called the Championship. In the event of two teams getting the same number of points the goal difference, that is the number of goals scored minus the number of goals conceded, is used as a tie breaker. Three teams are promoted from Championship each season so the total number of teams in PL remains 20. The total data-set used in this project has the results, and

scores, for all matches in the seasons 2009/2010-2020/2021, a total of 12 seasons. PL was in 2019 the sports league with the fourth highest revenue in the world, only beaten by the three American sport leagues NFL (American football), MLB (Baseball) and NBA (Basketball). The possible outcomes for a match are always victory to the home team (H), a draw (D) or victory to the visiting team (A). The probabilities the models estimate will be compared to the probabilities used by BET365 in their pre-match odds. BET365 is one of the largest online betting sites for bets on football results. It was founded in 2000, and has its headquarters in England.

1.2.1 The home field advantage

Football is a game full of cliches. Among these are; "the football is round"; "we need to look past the results"; "one game at a time"; "like playing with an extra man". The last is often used to describe the effect the supporters have on the home team. It can also be used as an accusation that the referee is biased towards the opponent, but that is not the usual meaning. It is a seldom questioned fact that there exists a home field advantage in football, that teams win more at home than away. The reasons for this can be many; they can sleep in their own bed the night before the match; less travel time; they are accustomed to the field, even in the Premier League the fields differ in both size and quality; and the arenas are designed so that the dressing rooms are much more pleasant for the home teams. In at least one arena the away team wardrobe is put directly beneath the home teams supporter stand, with minimal sound isolation. Despite these factors being important, the one factor that always gets the most attention is the supporters. A lot of the tickets are reserved for the home team supporters, the home team is therefore always much better represented among the supporters. Usually the home field advantage is used as a synonym for the advantage of having your supporters root for you, and maybe as important, root against the other team. Because of the ongoing Covid-19 pandemic, the stands have been empty for over a year, but football have still been played down at the pitch. This gives a good opportunity to see if there was a home field advantage before Covid-19, and to see if the home field advantage disappeared (or at least changed) when the supporters disappeared from the stands.

1.3 The model

The aim of the project is to find a model that precisely estimates the probabilities for the outcomes of future football matches in the Premier League bases on previous performance. The results will be on the form H (home victory), D (draw) or A (away victory). The exact score of the match is not of interest in the prediction, only the result, but the score can be used to infer about the result. Two different ideas will be utilized in the models. The first is to model the results directly with a multinomial distribution. The second is to model the number of goals scored by each team. The most used distribution for count data is the Poisson distribution, where the Poisson distribution is the resulting distribution from a Poisson prosess in the time interval, $t \in [0, 90]$. It has been the distribution of choice for many of those who wants to model football results. The Poisson prosess has some limitations. One is that it has a constant intensity function, meaning that if the number of goals is modelled by the Poisson prosess, there is the same probability for goal in any equal size time interval, Δt. This problem has an easy solution. Since we only are interested in the number of goals scored in the match by each team, and not at what time it was scored, we can use a inhomogeneous Poisson prosess, with time-dependent intensity $\lambda_{ij}(t)$. Goals scored by team i against team j is then Poisson distributed with mean $\Lambda_{ij} = \int_0^{90} \lambda_{ij}(t)dt$. Another possible problem with modeling the goals scored by each teams as two independent Poisson prosesses, is that we neglect the effect an early goal might have on what is remaining of the match. Both for the team that scores, and for the team that concedes the goal. In the Poisson process the probability of an occurrence in a time segment is independent of what happens in any other non-overlapping time segment. A violation of this in the data could result in over- or under-dispersion, and will be discussed later. The Poisson distribution has only one parameter, and the mean is equal to the variance. In empirical data there are often more variance than would be expected with the Poisson distribution, this is referred to as over-dispersion. An alternative model that has often been proposed is to use the negative binomial distribution instead. This second model will give another parameter,

θ, that is meant to model the variance. Models using the negative binomial distribution only accommodates over-dispersion, not under-dispersion[Greenhough et al., 2002]. In the case of under-dispersion, meaning that there is less variance than is expected, the generalized Poisson distribution can be used[Harris et al., 2012]. If the model for goals scored is correct, we would expect it to give better results than the multinomial model, because there is more information in the score than in just the result.

I will introduce the basic models here, and in section 3 I will describe it in more detail. As earlier mentioned, λ_{ij} will be the expected number of goals scored by team i against team j. The Poisson model in its most basic form will have log expectation, $\log \lambda_{ij} = a_i - d_j + c + h \cdot I(i)$. a_i is the attacking strength of team i and d_j is the defending strength of team j. It is reasonable to assume that the number of goals a team scores is dependent on their own attacking capacity as well as their opponents defensive capacity. However, it can be difficult to estimate which one was the most important in a specific game. That is: If team i beats team j 4-0 it can be difficult to say if this was due to team i being good offensively or if team j was poor in defence. These parameters are not identifiable if the data consists of a single game. However, with data consisting of several matches it will be possible to estimate the parameters. If one team often scores more goals than others do against the same opponents, then that team is likely to get a higher attacking strength. And if one team often let in more goals than other teams does against the same opponents that team will get a low defensive value. Even if it is impossible to say for sure if one team is good offensively or the other team is weak defensively after a single match, it is easy to spot a trend when the teams play against other teams as well. When the number of matches increase and the teams play against many teams, the parameters will become identifiable, and easier to estimate precisely. h is the home field advantage. It is expected that $h > 0$. In the model described above the attacking and defending strengths for the teams are constant over time, this is not necessarily a reasonable assumptions, as we know that for example managerial changes and player injuries, affect the teams greatly. An alternative to this is to model the attacking and defending capacity as a time dependent stochastic variable, for example one following a binormal Brownian Motion. This will be discussed further in section 3.

1.4 Money Management

Using the estimated probabilities offered by the models on the betting market can work for testing how well the models performs, and it can be used later in "the real world", if the model makes money. The betting sites offer odds. The odds are how much money you get back for each unit wagered. So if the odds for home victory is 2.5, and you bet 1 on home victory, you get 2.5 back if the home team wins, and 0 back if the home team does not win. In this project the odds for home victory, draw, and away victory will be denoted $\omega_h, \omega_d, \omega_a$. The odds offered by the bookie is in relation to the the probabilities for the different outcomes they have predicted. If they gave "fair odds", meaning that on average they paid out as much as they took in, and assuming that their probabilities are true, the relationship would be $p_j = 1/\omega_j, j \in \{h, d, a\}$. However, as they are in business the betting companies only pay out 97.3% on average. If the probabilities used by the betting sites are true, then no one will beat them in the long run, as they pay out less than they take in. When betting, one assumes that the probabilities offered by your model is better than the one used by the betting company. And you assume that the probabilities estimated by your model, are the true probabilities for the different outcomes. I will denote these probability estimates, π_h, π_d, π_a. One should place a bet when there is a positive expected return. The return from betting on outcome j in match i can be denoted Δ_i^j. If one bets c_i^j on match i, the estimated expected profit is $\hat{E}(\Delta_i^j) = (\pi_i^j \omega_i^j - 1)c_i^j$. One should place the bet if the estimated expectation is higher than zero, with $c_i^j \geq 0$. How to calculate the amount c_i^j to place on bet i will be discussed in section 2.7.

2 Theory

2.1 Latent variables

Latent, or hidden, variables in a statistical model are quantities that are not directly observed, they are inferred through the directly measured variables. Examples of models including latent variables are LMM (Linear mixed models), and GLMM (Generalized linear mixed models). I will briefly go through GLMs (Generalized linear models) before adding the latent effects to make it a GLMM.

2.1.1 GLM & GLMM

All GLMs have some things in common. They have a response function h such that $E(Y) = \mu = h(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, and the distribution of Y need to be a part of the univariat exponential family. Here $\mathbf{x}_i$ are the covariates for observation i, $\boldsymbol{\beta}$ is the parameter vector of interest, and y_i is the response. The total data is $\mathbf{y}$ and the matrix X, where row i in X is $\mathbf{x}_i^{\mathrm{T}}$. We also assume that the observations are independent of each other. The data points can be both numerical and categorical. Most of the analysis will be about estimating $\boldsymbol{\beta}$. The probability density function(pdf), or probability mass function(pmf), for an univariat exponential family distributed variable can be written as

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi}w + c(y, \phi, w)\right).$$

Where θ is the natural parameter; we require that $b(\theta)$ be twice differentiable; ϕ is a dispersion parameter; and w is a number that can be interpreted as a weight. It can be shown that $\mathrm{E}(Y) = b'(\theta)$ and $\mathrm{Var}(Y) = \phi b''(\theta)/w$. It is worth noting that the normal distribution is a part of the the univariat exponential family with $b(\theta) = \theta^2/2$ and $\phi = \sigma^2$. However, the GLM with normal distributed response variable, is just linear regression(LM), on the form $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$. If one uses the identity link-function ϵ has a diagonal variance matrix because we have assumed independence between each observation. This is a special case and one of the few where we can find an analytical solution for the maximum likelihood estimator (MLE) of our parameter of interest, $\boldsymbol{\beta}$. Usually we need some numerical optimization to find the best estimate of $\boldsymbol{\beta}$, which we will call $\hat{\boldsymbol{\beta}}$. $\hat{\boldsymbol{\beta}}$ will be found as the solution, with respect to $\boldsymbol{\beta}$, of $\mathbf{s}(\boldsymbol{\beta}) = 0$. The score function, $\mathbf{s}$, is the gradient of the log-likelihood to the full model. As the observations are independent the likelihood is just the product of the densities for each observation. The MLE of $\boldsymbol{\beta}$ can be computed using some numerical optimization, e.g. the Fisher-scoring algorithm. The $\hat{\boldsymbol{\beta}}$'s will be asymptotically normal in distribution, $\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, F^{-1}(\boldsymbol{\beta}))$, where F^{-1} is the inverse of the expected Hessian for the log-likelihood (the fisher information matrix).

The idea of random effects are easiest illustrated with LMM and can then be generalized from there. Random effects are often included for longitudinal data or clustered data [Fahrmeir et al., 2013a]. For example when the data is collected by different people; at different locations; or any situation when one gets a cluster structure, it can be relevant to include it as a random effect. The random effects are often assumed to be normal distributed. If a linear mixed model includes a normal distributed random intercept, it will look like this,

$$y_i = x_i^{\mathrm{T}}\boldsymbol{\beta} + \gamma_{\alpha(i)} + \epsilon_i, \gamma_{\alpha(i)} \sim N(0, \tau^2), \epsilon_i \sim N(0, \sigma^2).$$

Here $\alpha(i)$ is a function taking i as an argument and returning what cluster i comes from. The index i is a specific observation and $\alpha(i)$ returns the cluster i is in. This could be that i is a specific student and $\alpha(i)$ returns what school i attends. In this example y_i could be the test score on some national test. It is worth noting that there is a difference between the marginal model, and the conditional model. The marginal model has the distribution $y_i \sim N(x_i^{\mathrm{T}}\boldsymbol{\beta}, \tau^2+\sigma^2)$ and the conditional will have the form $y_i|\gamma_{\alpha(i)} \sim N(x_i^{\mathrm{T}}\boldsymbol{\beta}+\gamma_{\alpha(i)}, \sigma^2)$. This means that the variance is higher for new observations with unknown origin, than for new observations with known origin. It is standard to assume that $\mathrm{cov}(\epsilon_i, \gamma_{\alpha(i)}) = 0$. If one also includes a random slope for one or more of the covariats, the covariance matrix for the random effects need not be diagonal. For example there can be covariances between the random slopes and random intercept, making the covariance matrix

blockdiagonal. However, as only random intercepts will be considered in this project, I will not go into more detail here.

For a GLMM with normal distributed random intercept we have $\mu = h(\eta_i) = h(x_i^{\mathrm{T}}\boldsymbol{\beta}+\gamma_{\alpha(i)})$, with $\gamma_{\alpha(i)}$ as defined above. The joint likelihood for the observed data and the random effects is then the distribution for all the observations dependent on the random effect, multiplied by the distributions for the random effects.

$$L_{\text{joint}}(\boldsymbol{\beta}, \tau^2) = f(\mathbf{y}, \boldsymbol{\gamma}; \boldsymbol{\beta}, \tau^2) = \prod_{i=1}^{n} f\Big(y_i|\mathbf{x}_i, \gamma_{\alpha(i)}\Big) \prod_{\alpha=1}^{m} f(\gamma_\alpha),$$

where m is the number of clusters. This is the joint likelihood. However, when finding the best estimation for the parameters the marginal likelihood is often used. When optimizing the joint likelihood one finds the set of parameters $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ that maximize the likelihood. However since we are interested in $\boldsymbol{\beta}$ and not $\boldsymbol{\gamma}$, because we have not observed $\boldsymbol{\gamma}$, it is better to find the $\boldsymbol{\beta}$ that maximize the likelihood over all $\boldsymbol{\gamma}$. This means that we optimize the marginal likelihood, the one with the random effects integrated out. The marginal likelihood can be written as

$$L_{\text{marginal}}(\boldsymbol{\beta}, \tau^2) = f(\mathbf{y}; \boldsymbol{\beta}, \tau^2) = \int_{\mathbb{R}^m} \prod_{i=1}^{n} f\Big(y_i|\mathbf{x}_i, \gamma_{\alpha(i)}\Big) \prod_{\alpha=1}^{m} f(\gamma_\alpha) d\boldsymbol{\gamma}.$$

This is a general approach to the simplest GLMM, the one with just a random intercept. The estimated values for the parameters can be found by numeric optimization such as the r-package **TMB** which will be discussed further down.

2.2 Numerical approximation

This section will focus on the numeric approximation techniques used in the project. The techniques are general, but will be used almost exclusively on log-likelihood functions. The techniques discussed in this part of the paper are the ones used in the r-package **TMB**. It is based on the article Kristensen et al. [2016], which discusses how and why one would use TMB.

2.2.1 Laplace approximation

The Laplace approximation is used to approximate the value of an integral. If we assume $h(x)$ is the function to be integrated, the function is written as $h(x) = e^{\log(h(x))}$. The second order Taylor expansion of $\log(h(x))$ around the maximizer $\hat{x} = \max(\log(h(x))$ is used to approximate the integral of $h(x)$.

To use the Laplace approximation of the marginal log-likelihood function there are some assumptions that need to be made

- The negative joint log likelihood is on the form $f(\boldsymbol{\theta}, \mathbf{u})$
- $\boldsymbol{\theta} \in \mathbb{R}^n$ is the parameter vector of interest
- $\mathbf{u} \in \mathbb{R}^d$ are the latent variables
- $\hat{\mathbf{u}}(\theta) = \text{argmin}_{\mathbf{u}} f(\boldsymbol{\theta}, \mathbf{u})$ is in the interior of the domain
- $\nabla_u^2 f(\boldsymbol{\theta}, \hat{\mathbf{u}})$ is positive definite

The joint likelihood function can be written as

$$L(\boldsymbol{\theta}, \mathbf{u}) = e^{-f(\boldsymbol{\theta}, \mathbf{u})}.$$

Since we are interested in the marginal likelihood, that is the one with the random effects integrated out. We write

$$L^*(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} e^{-f(\boldsymbol{\theta}, \mathbf{u})} \mathrm{d}\mathbf{u} \tag{1}$$

The function $f(\boldsymbol{\theta}, \mathbf{u})$ has a second order Taylor expansion with respect to $\mathbf{u}$ around $\hat{\mathbf{u}}(\boldsymbol{\theta})$ on the form

$$f(\boldsymbol{\theta}, \mathbf{u}) = f(\boldsymbol{\theta}, \hat{\mathbf{u}}) + (\mathbf{u} - \hat{\mathbf{u}})^{\mathrm{T}} \nabla_{\mathbf{u}} f(\boldsymbol{\theta}, \hat{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \hat{\mathbf{u}}) \nabla^2_{\mathbf{u}} f(\boldsymbol{\theta}, \hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}})^{\mathrm{T}} + O(\|\mathbf{u} - \hat{\mathbf{u}}\|^3).$$

Since $\hat{\mathbf{u}} = \text{argmin}_{\mathbf{u}} f(\boldsymbol{\theta}, \mathbf{u})$ and $\hat{\mathbf{u}}$ is not on the boundary, we know the gradient with respect to $\mathbf{u}$ has to be zero in this point. This gives

$$f(\boldsymbol{\theta}, \mathbf{u}) = f(\boldsymbol{\theta}, \hat{\mathbf{u}}) + \frac{1}{2}(\mathbf{u} - \hat{\mathbf{u}}) \nabla^2_{\mathbf{u}} f(\boldsymbol{\theta}, \hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}})^{\mathrm{T}} + O(\|\mathbf{u} - \hat{\mathbf{u}}\|^3).$$

Inserted into (1) and omitting the higher order terms this becomes

$$L^*(\boldsymbol{\theta}, \hat{\mathbf{u}}) \approx e^{-f(\boldsymbol{\theta}, \hat{\mathbf{u}})} \int_{\mathbb{R}^d} e^{-\frac{1}{2}(\mathbf{u} - \hat{\mathbf{u}}) \nabla^2_{\mathbf{u}} f(\boldsymbol{\theta}, \hat{\mathbf{u}})(\mathbf{u} - \hat{\mathbf{u}})^{\mathrm{T}}} \mathrm{d}\mathbf{u}.$$

The integrand is proportional to a multivariate Gaussian probability density function (pdf) with inverse variance matrix $\Sigma^{-1} = \nabla^2_{\mathbf{u}} f(\boldsymbol{\theta}, \hat{\mathbf{u}}(\boldsymbol{\theta}))$. Since $\nabla^2_{\mathbf{u}} f(\boldsymbol{\theta}, \hat{\mathbf{u}})$ is assumed to be positive definite this is not a problem. Since the integral of a Gaussian pdf is 1 we have

$$\int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi}^n \sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})} d\mathbf{x} = 1.$$

The whole integral becomes

$$L^*\big(\boldsymbol{\theta}, \hat{\mathbf{u}}(\boldsymbol{\theta})\big) \approx (2\pi)^{\frac{d}{2}} \cdot \det\big(\nabla^2_{\mathbf{u}} f(\boldsymbol{\theta}, \hat{\mathbf{u}})\big)^{\frac{1}{2}} e^{-f(\boldsymbol{\theta}, \hat{\mathbf{u}})}.$$

(2)

This function can then be optimized with respect to $\boldsymbol{\theta}$ by some optimization algorithm. It is worth noting that if the integral for $\mathbf{u}$ does not go over the whole of $\mathbb{R}^d$ this technique can still be used with the assumption that the function goes fast down to zero. When calculating the uncertainty of the estimate, the $\delta-$method can be used,

$$\mathrm{Var}\big(\phi(\hat{\boldsymbol{\theta}})\big) = -\phi'_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\big(\nabla^2 \mathrm{log} L^*(\hat{\boldsymbol{\theta}})\big)^{-1} \phi'_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})^{\mathrm{T}}.$$

For the variance of the estimate $\hat{\boldsymbol{\theta}}$, we have $\phi(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $\mathrm{Var}(\theta) = (\nabla^2 \mathrm{log} L^*(\hat{\boldsymbol{\theta}}))^{-1}$, as it should. In the description above it was assumed that $\hat{\mathbf{u}}(\boldsymbol{\theta})$ as a minimizer of the function was readily available. But in reality we need to estimate this vector by using an inner optimization technique. **TMB** does this automatically by using automatic differentiation which is the theme in the next subsection.

This subsection has introduced the Laplace approximation in general, but for the interest of this project it will only be used for finding the maximum likelihood estimation for the marginal likelihood. The Laplace approximation works well when the likelihood with respect to $\mathbf{u}$ (not necessarily the data) is approximately Gaussian, and when one can use the second order Taylor expansion. However, this is not always the case and there exists alternative strategies for integrating out the random effects. In Bolker et al. [2009], the authors look at the use of GLMM's for ecological and evolutionary data. Among other things they compare the different methods for integrating out random effects. Three of the techniques they consider, which are the ones used in most software is: Penalized quasilikelihood (PQL), Laplace method, and Gauss-Hermite quadrature (GHQ). In this project only the Laplace approximation is used, but the pros and cons of these other two strategies, will be briefly discussed here.

PQL uses a quasilikelihood instead of an estimation of the actual likelihood, and some statisticians do not like inference, such as AIC, to be made with quasilikelihood. The method of PQL alternates between two methods (i) estimating the fixed parameters by fitting a GLM with a covarinace matrix based on an LMM, (ii) estimating the covarince matrix by fitting an LMM that has unequal variances calculated from

the previous GLM. PQL is the most used teqnice, it is for example used in the **lme4**-library in R, and it is highly flexible. The disadvantages of the method is that it uses a quasilikelihood and that it gives biased estimates when the variance of the random effects are high.

The Laplace approximation has already been discussed at length, so all I will say here, is that it is more accurate than PQL, but it is at the same time slower and less flexible.

GHQ is a more classical integration technique. It finds optimal subdivisions to evaluate the integrand. It is even more accurate than Laplace, but again slower. It can not compute more than 2-3 different clusters modeled as random effects. GHQ is used in the **glmmML**-library in R.

A forth alternative is to use Markov Chain Monte Carlo (MCMC) and consider it from a Bayesian point of view. One can then have an arbitrary number of random effects. The disadvantage of modeling it this way is that it is often quite slow, compared to the other methods. Alternatively one can use MCMC from a non-Bayesian point of view by using "data cloning" and non-informative priors[Lele et al., 2010]. This can be useful for frequentists who wants to use MCMC.

2.2.2 Automatic differentiation

When given an algorithm that computes a given function, automatic differentiation (AD) finds derivatives to the function. There are two different approaches to AD. The first is source transformation that uses a preprocessor to create a derivative code that is run alongside the code to be differentiated. This is the fastest and it also uses less memory [Kristensen et al., 2016]. It is however more difficult to implement. The second is operator overload. The **TMB** package that is used in this project, has implemented AD with operator overloading.

AD works by breaking the function down into elementary operations. It can be easier to imagine that the function is broken down into nodes. A function has some input nodes, that would be the variables. The values the function returns, would be the output nodes. For a scalar function there would only be one output node. When AD "breaks a function down into elementary operators" it means that it introduces internal nodes that only has the function of performing some intermediate calculation. For example $y = (a + b)^2$ could be broken into an internal node that calculated $a + b$, and then the value in the internal node would be squared and sent to the output node.

AD has two modes for calculating derivatives, called "forward" and "backward". The mode "forward" calculates the values for all internal nodes and output nodes, based on the values of the variables (input nodes). The total function is broken down to elementary operations. In the mode "backward" the chain rule is used to calculate the derivative of higher order nodes with respect to lower order nodes. The output nodes has the highest order and the input nodes has the lowest orders. The idea is that it is easy to calculate derivatives for elementary operations.

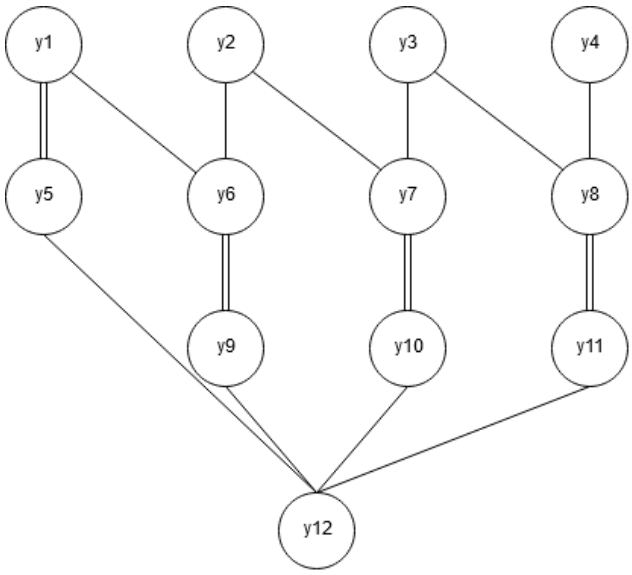


Figure 1: An illustration of $f(x)$. Double line indicates that the value of the node is squared to get to the next node. Single line means that the parent node is added or subtracted to the child node. Children have higher node numbers than parents.

This project will only focus on AD used on scalar functions, as AD will only be used to find the derivative of the negative log-likelihood. When a function is scalar it is enough to use each of the two modes "forward" and "backward", once.

An example of how this is done: A function $f(x) = x_1^2 + \sum_{i=2}^{4}(x_i - x_{i-1})^2$. This function is illustrated in figure 1 where the nodes $y_1 - y_4$ are the variables, node y_{12} is the output and the inner nodes is a breakdown of the function into elementary operations. One line between two nodes means that it is linearly included, and two lines means it is taken to the power 2, this is just to illustrate how the function $f(x)$ is broken down into elementary operations. What we are interested in is the derivative of the output with respect to each of the variables. That is $\frac{\partial y_{12}}{\partial y_i}$ for $i \in (1, 4)$. In the "forward" mode of this calculation the values for each node is calculated based on the four input nodes. In the "backward" mode it calculates the derivatives with the chain rule. It calculates all derivatives recursively given the partials of higher nodes. So for example, $\frac{\partial y_{12}}{\partial y_1} = \frac{\partial y_{12}}{\partial y_5}\frac{\partial y_5}{\partial y_1} + \frac{\partial y_{12}}{\partial y_9}\frac{\partial y_9}{\partial y_6}\frac{\partial y_6}{\partial y_1}$. The partials of the the output node, $\frac{\partial y_{12}}{\partial y_5}$, $\frac{\partial y_{12}}{\partial y_9}$, are calculated earlier in the "backward" mode since 5 and 9 are larger than 1.

For multiplication, the partial derivative will be dependent on the value of the node. Therefore the value from the "forward" mode will be needed in calculating the value in the "backward" mode.

For scalar functions such as this and the ones that will be discussed in this project, the evaluation of the the derivative is inexpensive using no more than 4 times the float-point operations used to evaluate f. This is referred to at the "Cheap gradient principle".

There exists alternative ways to estimate the derivative of a function. The two examples I will give here is Numerical differentiation and Symbolical differentiation. Numerical differentiation estimates the derivative by looking at the definition of the derivative,

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \lim_{h \to 0} \frac{f(\mathbf{x} + \mathbf{e}_i \cdot h) - f(\mathbf{x})}{h},$$

and picking a small value for $h > 0$, and evaluating the expression. $\mathbf{e}_i$ is the i-th unit vector. Numerical differentiation gives less precise estimates than automatic differentiation, and for evaluating a gradient in n dimensions it is in order $O(n)$. h can also be difficult too choose correctly[Baydin et al., 2018].

Symbolical differentiation, as used in the mathematical program Maple, uses the analytical rules of differentiation to compute the derivatives. One problem with this method is that it can use exponentially more

space on intermediate calculation than on the expression to be differentiated. And they are computationally inefficient[Baydin et al., 2018]. When one is concerned with the accurate numerical evaluation of the derivatives, and their symbolic form is of less importance, it is possible to save much memory and time by using Automatic Differentiation instead of Symbolic differentiation.

2.3 TMB

Template model builder (TMB) is an r-package using files written in c++. Given a c++ file that computes negative log-likelihoods, one can use TMB to create an object to be optimized using the function MakeADFun(). In this function the user includes the data and the parameters to be used in the likelihood. If the user specifies that any parameters should be treated as random effects, TMB will integrate them out using the Laplace approximation[Kristensen et al., 2016]. Automatic differentiation is used to differentiate the negative log-likelihood function. The object created by MakeADFun() can be optimized by some optimization function in R, for example optim(). The maximum likelihood estimates (MLE) can be found by sdreport(), a function from the TMB-library. The estimates for the fixed parameters will also have a covariance matrix, while the estimates for random effects will only include the standard deviation of each random effect. TMB is both fast and gives accurate estimations. There are other efficient r-packages as well, but few others give the user such freedom in the way of designing the function to be optimized.

2.4 Skellam distribution

If $X \sim \text{Pois}(\lambda)$ is a Poisson distributed stochastic variable with expectation λ, its probability density function takes the form

$$f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}.$$

Its cumulative density function does not have a nice analytical expression. Suppose that we also have another Poisson distributed variable, $Y \sim \text{Pois}(\theta)$, and we assume X and Y to be independent. The sum of the variables will also be Poisson distributed $Z = X+Y \sim \text{Pois}(\lambda+\theta)$. However, sometimes we are interested in the difference between two Poisson distributed variables and not the sum. So what is the distribution for $Z = X - Y$?

To find the distribution of Z we will find the joint probability density function of X and Y. We will introduce the change of variables $Z = X - Y$ and $Y = Y$. After finding the joint density function for (z, y) we will sum out y and what is remaining will be $f(z)$. $f(z)$ will be the probability density function for the Skellam distribution.

$$f(x,y) = \frac{\lambda^x e^{-\lambda}}{x!}\frac{\theta^y e^{-\theta}}{y!}$$

Introducing $Z = X-Y$ and $Y = Y$ means that we get $X = Z+Y$. Since this is a one-to-one transformation of discrete variables we get

$$f(z,y) = \frac{\lambda^{z+y} e^{-\lambda}}{(z+y)!}\frac{\theta^y e^{-\theta}}{y!}.$$

Using the law of total probability the probability mass function of z is

$$\begin{aligned} \text{f(z)} &= \lambda^z e^{-\lambda-\theta}\sum_{y=0}^{\infty}\frac{(\lambda\theta)^y}{(z+y)!y!} \\ &= \lambda^z e^{-\lambda-\theta}(\lambda\theta)^{-z/2}I_z(2\sqrt{\lambda\theta}) \\ &= \text{e}^{-\lambda}e^{-\theta}(\tfrac{\lambda}{\theta})^{z/2}I_z(2\sqrt{\lambda\theta}), \end{aligned}$$

where $I_z(\cdot)$ is the modified Bessel function of the first kind. The Skellam probability mass function and cumulative mass function is included in r with the **skellam**-library. The gamma distribution is used to calculate the modified Bessel function. This relationship was first found by Skellam in 1946[Skellam, 1946]. In the case of football results the value of Z is not really of interest as much as the sign of Z. If

X is the number of goals scored by the home team, and Y the number of goals scored by the away team; $Z = X - Y > 0$ gives a home victory; $Z = X - Y = 0$ gives a draw; $Z = X - Y < 0$ gives an away victory.

2.5 Overdispersion

For count data, the Poisson distribution is often used. As the Poisson distribution is part of the exponential family it is often used as a GLM with log link-function. This means that the response, Y_i, follows a Poisson distribution with log-expectation $\log(\lambda_i) = x_i^{\mathrm{T}}\beta$, where $\mathbf{x}_i$ is covaraits for observation i. Maximum likelihood is used to estimate the parameters β. A Poisson distributed variable has the same expectation and variance. Often, when fitting a Poisson GLM to real data, the conditional variance is higher than the conditional expectation. Conditional means that the data $\mathbf{x}_i$ is known. When the conditional variance is higher than the conditional expectation, it is said that it is over-dispersed. The dispersion should be approximately equal to the degrees of freedom in the model, when it is significantly higher we say it is over-dispersed. There are some main reasons for over-dispersion. The first I will consider is that some key factors is neglected from the observed data, $\mathbf{x}_i$. When key factors are neglected the model will not give good estimates, and $\boldsymbol{\beta}$ will be biased. The only solution when this is the case, is to include all the necessary information in $\mathbf{x}_i$[Berk and MacDonald, 2008].

The second reason for over-dispersion is positive dependence in the data [Berk and MacDonald, 2008]. In Berk and MacDonald [2008] they use criminology as an example: They counted the number of reported cases of misconduct of different inmates. The data used in estimating the expectation was among others; age, and membership in some prison-gang. The positive dependency they suspected was: If an inmate had done something wrong and been reported, the prison guards would in the future report them for lesser offenses than for inmates with a clean sheet. Such dependencies are not covered by the Poisson model. The strategy often used for such cases, is to assume a noise term to the log-expectation

$$\log \mu_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$$

$$\mu_i = \lambda_i \cdot e^{\epsilon_i} = \lambda_i \cdot u_i$$

with $e^{\epsilon_i} = u_i \sim \Gamma(\theta, \frac{1}{\theta})$. So u_i has expectation 1. Then $Y_i|\mathbf{x}_i, u_i \sim$Pois$(\mu_i)$. However as we want the distribution of Y_i to depend only on $\mathbf{x}_i$, we need to integrate u_i out.

$$f(y_i|\mathbf{x}_i) = \int_0^\infty f(y_i|\mathbf{x}_i, u_i) \cdot f(u_i) du_i$$

$$= \int_0^\infty \frac{(\lambda_i u_i)^{y_i} e^{-\lambda_i u_i}}{y_i!} \cdot \frac{\theta^\theta}{\Gamma(\theta)} u_i^{\theta-1} e^{-u_i\theta} du_i$$

The last integrand is proportional to a Gamma distribution with $\alpha = \theta + y_i$ and $\beta = \frac{1}{\lambda_i + \theta}$. This gives

$$f(y_i|\mathbf{x}_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(\theta)\Gamma(y_i + 1)} r_i^{y_i}(1 - r_i)^\theta,$$

$$r_i = \frac{\lambda_i}{\lambda_i + \theta}.$$

This can be recognised as a negative binomial distribution. The negative binomial distribution is often used for count data when overdispersion is present. It is however important to notice that it only helps if it truly is more variance in the data, than is expected by the Poisson distribution. The variance is accommodated by the extra parameter θ. The estimates for λ_i will be almost identical for both models. Therefore, if one is only interested in the estimate, and not in the variance, the models are interchangeable. We also see that when $\theta \to \infty$ the negative binomial distribution simplifies too the Poisson distribution. The negative binomial distribution can only fit models when the Poisson distribution fits, or an over-dispersion is present. However, sometimes we have under-dispersion in the data, meaning that we have less variance than would be expected for the Poisson distribution. In that case a Generalized Poisson distribution is a better alternative.

2.6 Generalized Poisson

The Poisson distribution has only one parameter, that parameter is both the mean and the variance. Often the variance for count data is higher than the mean to the data. This is called over-dispersion. The variance can also be smaller than the mean, this is called under-dispersion. Unlike the more used negative binomial distribution, generalized Poisson distributions can fit all cases; That is Ordinary Poisson, under- and over-dispersion. Under-dispersion is often caused be some dependency within the data. The Poisson distribution assumes that the occurrence of an event does not alter the probability of another event happening in the next time segment. However, in the real world this is not always the case. Over-dispersion and under-dispersion is seen in the data when there is correlation between observations. For positive correlation between non-overlapping time segments we get over-dispersion, and for negative correlation we get under-dispersion. In Harris et al. [2012] they formulate the probability mass function for a Generalized Poisson distributed stochastic variable. The parameterization they use, is the same that will be used in this project. If $X \sim GP(\lambda, \delta)$, its probability mass function takes the form

$$f(x; \lambda, \delta) = \frac{\lambda(\lambda + \delta x)^{x-1} e^{-\lambda - \delta x}}{x!}, x = 0, 1, 2.... \tag{3}$$

With $\max(-1, -\lambda/4) < \delta < 1$, and $\lambda > 0$. For $\delta = 0$ this becomes the ordinary Poisson distribution. We have the first moments as $\mathrm{E}(X) = \frac{\lambda}{1-\delta}$ and $\mathrm{Var}(X) = \frac{\lambda}{(1-\delta)^3} = \frac{\mathrm{E}(X)}{(1-\delta)^2} = \frac{1}{\phi}\mathrm{E}(X)$. With $\phi = \frac{1}{(1-\delta)^2}$. For $\delta > 0$ we have over-dispersion and for $\delta < 0$ we have under-dispersion. The parameters in this model can be estimated with **glmmTMB** and it will return ϕ as its dispersion parameter. λ will be modelled by a log link, $\lambda_i = \exp(\mathbf{x}_i^T \beta)$, where x_i is the data and β is the parameters.

Often, when studying Generalized Poisson distributed variables, the most important result is not the value of X, but which is the larger of X and another GP distributed variable, Y. An example of this is in football, where the number of goals scored is not as important as scoring more goals than the opponent. For ordinary Poisson distributed variables we could use the Skellam distribution for the difference between two variables. For GP distributed variables it is more difficult. In Consul [1986] they produce the probability density function for the difference, but the probability mass function(pmf) is not on closed form.

2.7 Money Management

The aim of this project is not only to create a model that fits football results well, but also to make a model that can beat the bookie and reliably earn money on the betting market. To do this one needs a good model, but also a good betting strategy. There are played 10 games each round of the Premier League, and it is assumed that one places all the bets for the 10 matches before the round begins. This section will focus on how to best allocate the money on each of these 10 possible games. This can be viewed as a short-term portfolio with known payouts. We need to use some notations. C is the total amount to be wagered in one round. A good bet is a bet with positive expectation for each unit wagered. The payout if you win is equal to the odds, ω_i. So the expected return is $p_i\omega_i - 1$ for each unit. Δ_i will be the profit on bet i. Negative values for Δ_i means that you lost money on this bet. n will be the number of bets to allocate money on, in each round. n will normally be around 10, but can theoretically be any number from 0 to 20. Both the expectation of the profit from each wager, and the variance of each wager, is important. $\mathrm{E}(\Delta_i) = c_i(\omega_i - 1)p_i$ and $\mathrm{Var}(\Delta_i) = (c_i\omega_i)^2 p_i(1 - p_i)$, with c_i being the amount wagered on bet i. We need $\sum_{i=1}^{n} c_i \leq C$.

2.7.1 Fixed bet

This is the simplest betting strategy used in this project. Every bet with a positive expected profit will have the same amount of C allocated to it, independent of how much the expected profit is, and how likely it is to win. This means that $c_i = C/n$, for all bets with positive expectation. This model is not very sophisticated. However, as it only places money on bets with positive expectation, it is expected to earn money, given the assumption that the estimated probabilities are the true probabilities.

2.7.2 Fixed return

When using fixed return one places money on each bet such that the return from a winning bet will always be the same, $c_i \propto 1/\omega_i$. For bets with higher odds less money will be allocated. This model uses the assumption that bets with high odds have low probabilities. This is often the case, but if a bet with high probability and high odds appeared, this model would place a small bet on it. Ideally the model should include both the probability and the odds.

2.7.3 Kelly criterion

Kelly [2007] proposed a betting strategy that is theoretically the optimal strategy, given the opportunity to play infinitely many times. He assumed that a player starts with C money. Given a probability, p_i, and profit in the case of victory per unit wagered, $b_i = \omega_i - 1$, he wanted to find the optimal fraction of his total wealth, f^*, to place on the bet. He optimized the expectation of the logarithm of his amount of money after the bet. This is equivalent to optimizing the expected geometric growth rate.

$$\mathrm{E} = p_i \log(C + Cfb_i) + (1-p)\log(C - Cf),$$

where fC is the money he bets, and $C(1-f)$ is the money he does not bet. Finding the derivative with respect to f and setting equal to zero, finds the maximum value.

$$\frac{\partial E}{\partial f} = \frac{p_i b_i}{1 + f b_i} - \frac{1-p}{1-f} = 0$$

This is solved for

$$f^* = \frac{p_i(b_i+1)-1}{b_i} = \frac{p_i\omega_i - 1}{\omega_i - 1}.$$

This can be viewed as (expected net winning)/(net winning if you win). An alternative way of looking at it is $f^* = p_i - (1-p_i)/b_i$, the probability of winning minus the probability of losing divided by the profit if you win. Meaning that even for fantastic odds, you should never bet a larger portion of your wealth than the probability of victory.

The most common criticisms of the Kelly Criterion in betting, is that players often overestimates their probabilities, p_i. This model assumes that the probabilities are the true probabilities, and that is not always the case when the probabilities are the outputs from statistical models. It has therefore been proposed to use fractional Kelly criterion. In this case players bet a fraction of the fraction found by the Kelly Criterion, often half, but it can be any fraction.

This betting strategy is meant for cases where you start with C money and consider a single bet at a time. However, often in sports-betting the matches are played simultaneous and then you will have to place bets on more than one match at a time. Chris Withrow considers this problem of finding the optimum betting fraction on each bet when placing bets simultaneous [Whitrow, 2007]. If one has n bets to play on, for small n he finds that the optimal fractions are identical to the fractions found by looking at them individually. And of course $\sum_{i=1}^{n} f_i \leq 1$. If the sum of the fractions come close to 1, this no longer holds. One would then need to optimize it simultaneously, as he does with Markov Chain Monte Carlo. An alternative to this is the one used by Langseth [2013], where he at most places $C_0 << C$ on all the bets. This is to ensure that the model does not loose all its money on a single round. The fraction to place on bet i is then $f_i C_0$ if $\sum_{j=1}^{n} f_j \leq 1$, else it is $\frac{f_i C_0}{\sum_{j=1}^{n} f_j}$. This betting strategy assumes independence of the possible outcomes and can not be used to bet on different outcomes in the same match, even if there is expected profit for two outcomes.

2.7.4 Rue and Salvesen betting strategy

Rue and Salvesen propose a betting strategy based on optimizing the expected profit minus the variance of the profit [Rue and Salvesen, 2000]. They limit the number of bets to place on a single match to one. If two outcomes in the same match has a positive expected profit, the one with the highest positive expected profit

will be chosen. If the total profit is $\mathbf{\Delta} = \sum_{i=1}^{n} \Delta_i$, then $\mathrm{E}(\mathbf{\Delta}) - \mathrm{Var}(\mathbf{\Delta})$ is the expression to be optimized with respect to $\boldsymbol{c}$, under the constraint that $\sum_{i=1}^{n} c_i \leq C$. Since the bets are placed on different matches, the Δ_i's are independent of each other. And we can write $E(\mathbf{\Delta}) - \mathrm{Var}(\mathbf{\Delta}) = E(\sum_{i=1}^{n} \Delta_i) - \mathrm{Var}(\sum_{i=1}^{n} \Delta_i) = \sum_{i=1}^{n} E(\Delta_i) - \sum_{i=1}^{n} \mathrm{Var}(\Delta_i)$, with the expectation and variance as defined above. Finding the optimal value with the given constraints, is then easy. We see that without the constrains $c_i' = \frac{p_i\omega_i - 1}{2p_i(1-p_i)\omega_i^2}$ is the optimal amount to place on bet i. If we introduce a scalar k and say that $c_i = kc_i'$, we get $\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} kc_i' \leq C \implies k = \frac{C}{\sum_{i=1}^{n} c_i'}$. So the amount to place on bet i is $c_i = c_i' \frac{C}{\sum_{j=1}^{n} c_j'}$.

This strategy differs from Kelly Criterion in the way that it assumes that you have C money to spend on each round independent of how earlier rounds went. While the amount of money you have to spend at any given round using the Kelly Criterion, is based on the amount of money you have earned in the rounds up to this point. Kelly criterion is better suited for situations where you bet a large portion of your money. While this strategy proposed by Rue and Salvesen assumes that even if you always loose there is always more money for you to spend.

2.8 Paired comparison

Bradley and Terry introduced a comparison model in 1952[Bradley and Terry, 1952]. The model assumes that we have a population of n individuals, and all the individuals in that population compete. In Each competition there is one winner and one looser. The Bradley-Terry model then predicts the probability of a victory for each participant. When i and j competes, let $i > j$ denote the event that i beats j. Then

$$P(i > j) = \frac{p_i}{p_i + p_j},$$

where $p_i > 0$ is a score assigned to individual i. This comparison model is often used for sports results or for predicting the result of fights between animals of the same species. An often used parameterization is $p_i = e^{\beta_i}$, where β_i is a linear function for individual i. When using this parameterization we have that $\mathrm{logit}(P(i > j)) = \beta_i - \beta_j$. When β_i is independent of the opponent j, we have that; $P(i > j) > 1/2$ if and only if $\beta_i > \beta_j$. If $P(i > j) > P(j > i)$ we say that i dominates j. A model is called transitive if i dominating j and j dominating k, implies that i dominates k. We see that if all the β's are constant over time this model is transitive, it will also be transitive at any given point in time, if the β's change over time.

In some settings there is also a a possibility for a draw, as well as victory to one of the two individuals. This could for example be the case in many sports.

Hankin [2020] uses a generelized Bradley-Terry model to model draws in chess. The way he does it is to introduce a third party that can also win the match, this third party represents the draw. An alternative way to do it, the one used in Cattelan et al. [2013], is to introduce two θ's. This extends the Bradley-Terry model to a ordinal multinomial GLM with logit link-function. Assume we have $i = 1, ..., m$, matches and each match has one home team and one away team. We will use the notation 1 for home victory, 2 for a draw, and 3 for away victory. The functions h_i and a_i will return what team plays at home and what teams play away, in match i. The result of the match Y_i will follow a multinomial distribution with an cumulative distribution function as shown below

$$P(Y_i \leq y) = \frac{\exp(\theta_y + \eta + \beta_{h_i} - \beta_{a_i})}{1 + \exp(\theta_y + \eta + \beta_{h_i} - \beta_{a_i})}, y \in \{1, 2, 3\}.$$

$-\infty < \theta_1 < \theta_2 < \theta_3 = \infty$, we put the restriction of symmetry $-\theta_1 = \theta_2$ so that the teams will have the same chances of victory on a neutral field. η will model the home field advantage if there is any. Each β can be a linear combination of covaraits, it is also possible to include random effects in each β_i. We see that for the special case of $\theta_1 = \theta_2 = 0$, we have a normal Bernoulli trial Bradley-Terry model.

Bradley-Terry models can be very useful. For example in sport leagues, where every team meets each other twice. The result alone for these two matches is not enough to give a good approximation for the true probabilities of the outcomes. However, by assuming that the probabilities for the different outcomes are dependent on the relative strengths of the teams, we can also infer about the matches with few or even no

observations. It is important to note that an ill stated model can give bad results. If one assume transativity when that is not the case in the data, the results can be wrong. For example in Tufto et al. [1998] the authors found that for deers fighting over food, there was a correlation between antler size and chances of victory. However, in disputes between a mother and her own child(daughter) the mother often lost. This meant that a mother could dominate another deer, this other deer could dominate the child and the child could dominate the mother. This would give us a circle of domination and would then be in-transitive. In Tufto et al. [1998] they introduced an extra predictor that was present when the fight was between mother and daughter, thus making a in-transative model to match the in-transative nature of the data.

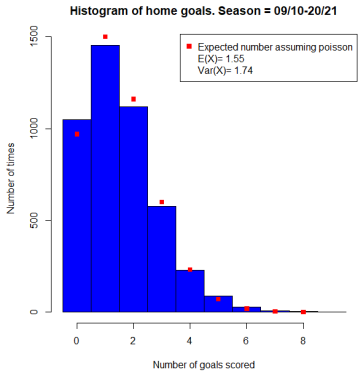
3 Model

This section will describe all the models used in the project. All the models are either counting models or multinomial models. The parameterization of these models can be both constant over the whole season, or following Brownian motions. We will first take a look at the data used in the project.

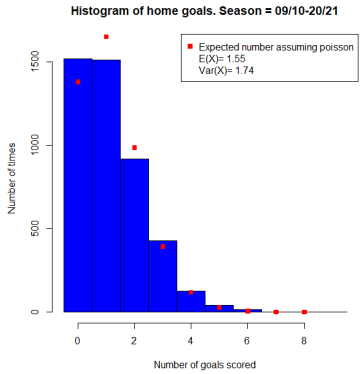
3.1 The data

The data set used in this project consist of 12 seasons of the Premier League from 2009/2010-2020/2021. Each entry into the data set contains information about what teams are playing (home and away), how many goals they scored, how many shots on target, and how many shots in total. It also has the pre-match betting odds from most of the big betting sites.

Histograms for home and away goals for all twelve seasons are given in figure 2 a and figure 2 b. If we assume that each score is the product of a Poisson process, the sum of goals should also follow a Poisson process. As can be seen by the plots it looks like the Poisson distribution fits fairly well, even when no covariat other than home/away has been used. However, it looks like there is an influx of zeros, and fewer games where the teams score 1 goal than the Poisson distribution would predict. The variance is also higher than the mean, this is contradictory to the theoretical values for the Poisson distribution. It does look like the higher variance is mainly caused by the increase in the number of zeros.



(a) Home goals



(b) Away goals

Figure 2: The distribution of home goals and away goals for the seasons 09/10-20/21 in the Premier League. We see that it is close to being Poisson distributed.

One would expect that a strong attacking team is strong in defense as well, it is therefore no surprise that the correlation of home and away goals scored, is negative. It is not very high however; -0.096, for all the seasons. This does not necessarily mean that a bivariat Poisson model is suited. We would expect this if the teams that are good at attacking is also good at defending.

In figure 3 we see how many percent of matches end in home win, draw, and away win. We see that the home team wins much more often than the away team. It is therefore reasonable to include a distinction between home and away team in any model. We also notice that for the current season, we see for the first time, more away matches have been won than home matches. This will be investigated further.

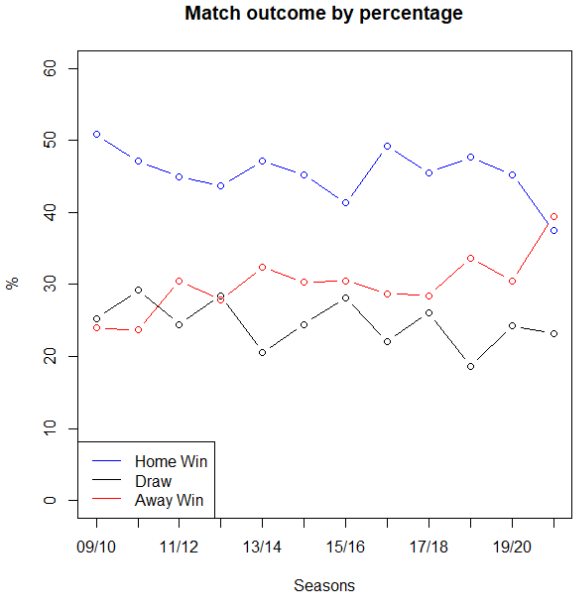


Figure 3: The percentage of matches that ended in home win; draw; away win, in the seasons 09/10-20/21. More games are won by the away team for the first time this last season.

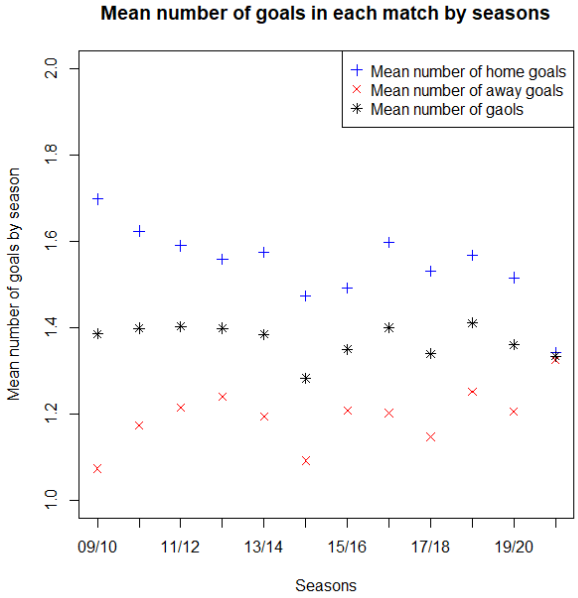


Figure 4: The mean number of goals scored in matches. We see that the mean number of goals scored by home team is the same as for away team for the first time, this last season.

In figure 4 the mean number of goals per match is plotted for all the seasons. We see that there is a clear trend that the teams score more goals at home than in away matches. However, at the current season there seems to be very little difference between the two. Any model used for predicting the outcomes of the current season, should probably not use a home field advantage estimated only on previous seasons.

3.2 Poisson Models

First I will introduce a constant Poisson model where each teams offensive and defensive strengths are modelled as a normal distributed random intercept. This model assumes that the teams strengths are constant over the whole season, and that there exist a hierarchy-structure among the teams. In other words if team i dominates team j and team j dominates team k, then team i dominates team k in this model. One can assume a dependence between the strength of the offensive and defensive ability for each team, the random intercepts will then come from a binormal distribution. If we let X_{ij} be the number of goals team i scores against team j, we have

$$X_{ij} \sim \text{Pois}(\lambda_{ij})$$

$$\log \lambda_{ij} | a_i, d_j = c + h \cdot h(i) + a_i - d_j$$

$$(a_i, d_i)^T \sim \text{N}(\mathbf{0}, \Sigma). \tag{4}$$

$h(i)$ is an indicator function returning 1 if team i plays at home, and zero otherwise. Restraints can be put on the variance matrix Σ, such as it being diagonal, i.e. no dependence between attacking and defending strengths. This is the simplest model uses in the project, and will be referred to as Model 1. It is similar to

the model used by Maher [1982], with the exception that I have included the strengths as random effects, while he included them as strengths relative to some randomly chosen reference team.

The Poisson model is often the model of choice for count data because of its simplicity, as it has only one parameter. This parameter is both the mean and the variance for data generated by a Poisson distribution. In the real world data is often more noisy than the data generated directly from this theoretical distribution. We therefore often observe over-dispersion, meaning that there are more variance in the data than is expected with the model. It has been suggested using the Negative binomial distribution instead of the Poisson when over-dispersion is present in the data. This has been used in for example Greenhough et al. [2002], and they conclude that the Poisson is a poor choice of distribution because its tail is not heavy enough. This means that the probability for high scoring games is too low with the Poisson compared with the empirical results. It is worth mentioning that their model is an even simpler model than the Model 1 described above. It assumes that all goals scored at the home field has one distribution, and all goals scored at the away field has one distribution, independent of the teams playing. With such a model there is no surprise that over-dispersion is present. However, as discussed in Berk and MacDonald [2008] there is a difference between over-dispersion due to an ill fitted model, and over-dispersion due to actual unexplained variance in the data. Berk and MacDonald [2008] looks at the the number of times inmates get a warning. They say that negative binomial is preferable in situations when an instance of what you are counting, increases the chance for another instance. In the case of football this would mean that the probability of a goal in the next time segment is higher if there has already been a goal. Conversely if the probability was lower we would expect under-dispersion. However, if the over-dispersion is due to important factors being neglected, the Negative Binomial will do no better than the Poisson. And it is difficult to advocate that what team is attacking and what team is defending is not important factors in the number of goals scored. Only if a model that included all necessary information is over-dispersed, will the negative binomial distribution be used in this project.

3.3 Time dependent models

The previous model was time independent. An assumption for such a model is that the teams have the same relative strength the whole season. We know that teams can have bad streaks, for example the defending champion in the Premier League has lost 6 home games in a row this season. One can wonder: has the team become worse than they were at the beginning of the season; or has the worsen results been due to bad luck? There are two main ways used to model time dependence in these models. The first, used in for example Dixon and Coles [1997], weights the results differently, and recent results are weighted more heavily in the likelihood, than older results. This model will give different scoring abilities for different times, but it will not return an estimate for the teams performance over time. The second alternative is to let the teams attacking and defensive ability be the result of a time series. This could for example be a binormal Brownian motion, as used in Rue and Salvesen [2000]. When finding the MLE for such a model one would find the estimated values for the time-series model. This second model would then also return the MLE for the attacking and defending strengths for the teams as a function of time.

A natural time dependent extension to model 1 would be to let the attacking and defending abilities to the different teams, follow a Brownian Motion. This would mean having the random effects from model 1 change from round to round.

$$(a_i^0, d_i^0)^{\mathrm{T}} \sim \mathrm{N}(0, \Sigma)$$

$$(a_i^t, d_i^t | a_i^{t'}, d_i^{t'}) \sim \mathrm{N}((a_i^{t'}, d_i^{t'}), \frac{(t-t')}{\tau}\Sigma), t > t'$$

where τ is a parameter describing how fast the teams performance changes. a_i^t is linear addition to the log-expectation for goals scored by team i at time t. And d_i^t is the linear addition to the log-expectation to the number of goals team i will concede at time t. It is important to use a random intercept for the different strengths at $t = 0$. A Brownian motion is a non stationary process, meaning that if $\{a_t\}$ is a solution, then so is $\{a_t\} + c$, where c is constant. And since we model the log-expected goals as the difference between the attacking teams attacking strength and the defending teams defensive strength, the solution would not

be identifiable without having the strengths grounded at some time, and time $t = 0$, is the obvious choice. τ and Σ can be parameters to be optimized, or they can be estimated from the data and kept fixed in the optimization of the likelihood. They were kept fixed by Rue and Salvesen in their article[Rue and Salvesen, 2000]. They estimated it from the data and found τ to be 100 and $\Sigma = \sigma^2 I = 1/37I$ to be diagonal with the same entry $\sigma^2 = 1/37$ on the diagonal. This would mean that after 100 days each team has as big a variance in its performance, as the different teams have at the beginning of the season. The reason it was not optimized in the likelihood together with the other parameters, was that there was too little information about τ in the dataset . I will try to optimize τ in the likelihood, by using all 12 season of the Premier League in the same model. The hope is that there was a lack of data that made it hard to estimate τ. Rue and Salvesen estimated other parameters from the distribution of historical data as well. Among these are a psychological effect where strong teams underestimates their opponents and therefore the chance of the weaker team scoring is increased, and the chance of the stronger team scoring is decreased. The parameter modeling this effect was also estimated from the distribution of historical data, and I will not include it in my model. I will try to estimate the parameters as the MLE by using TMB.

3.4 Generalized Poisson Model

Both the time independent and the time dependent models described above, assumes that the number of goals scored follows a Poisson distribution. The models can also assume a Generalized Poisson distribution. The linear additions to $\log \lambda$ will be the same, the only difference will be the extra parameter describing the variance in the data as described in section 2.6.

3.5 Data intensive model

The reason for using Poisson models and other goal counting models, as opposed to a more direct multinomial models, is to incorporate more information from the data into the model. In Langseth [2013] he takes this one step further by including, not only the goals scored, but also the number of shots and shots on target. The full model he uses has a Poisson distribution for the number of chances created. Each team has a different mean, λ_i. λ_i can change for home and away matches. This means that the number of chances team i gets against team j follows a Poisson distribution $C_{ij} \sim \text{Pois}(\lambda_i)$. It is worth noting that the data set does not include the number of chances created by each team, so this would have to be summed out of the model before calculating the MLE for the parameters. The number of shots fired, S, given the number of chances, $S_{ij}|C_{ij}$, follows a Binomial distribution with $n = C_{ij}$, and a probability $p = \Phi(\beta_j)$ that is dependent on the defending capacity of the opposing team. The shots can become shoots on target. The number of shots on target, T, given the number of shots, $T_{ij}|S_{ij}$, also follows a binomial distribution with $n = S_{ij}$, and probability of success is dependent on the attacking capacity of the team, $p = \Phi(\alpha_i)$. Lastly the number of goals, X, given the number of shots on target, $X_{ij}|T_{ij}$, also follows a binomial distribution with $n = T_{ij}$ and probability of success $p = \Phi(\gamma_j)$ dependent on the opposing teams goalkeeper. Φ is the cumulative density function for the standard normal, and it ensures that the probabilities are kept between 0 and 1.

A problem with this model is that it assumes that the goals is a subset of the shots on target, this is of course almost true, but because own-goals does not count as shots on target, there can be goals without shots on target. This is of course not a big problem and trying to correct it by modeling own goals as well might do more damage than good. However in a few matches in the data set the number of goals is larger than the number of shots on target, this is a problem as it breaks the model. I have therefore decided to use $T_{ij} = \max(T_{ij}, G_{ij})$. Of the 760 observation each season 3 games had this property, so I do not think it will affect the end result.

The model can be given more mathematically like this

$$f(x, t, s, c|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = f(x_{ij}|t_{ij}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) f(t_{ij}|s_{ij}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) f(s_{ij}|c_{ij}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) f(c_{ij}|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda})$$

$$f(x_{ij}, t_{ij}, s_{ij}, c_{ij}|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \text{Binom}(x; t_{ij}, \Phi(\gamma_j^t)) \, \text{Binom}(t_{ij}; s_{ij}, \Phi(\alpha_i^t)) \, \text{Binom}(s_{ij}; c_{ij}, \beta_j^t) \, \text{Pois}(c_{ij}; \lambda_i)$$

As we do not have c_{ij} in the data set, we need to sum it out.

We know that if $Y \sim \text{Pois}(y;\lambda)$ and $X|Y \sim \text{Binom}(x;y,p)$ then $X \sim \text{Pois}(x;p\lambda)$. A proof of this can be found in the Appendix.

Therefore

$$f(x_{ij},t_{ij},s_{ij}|\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda}) = \text{Binom}(x;t_{ij},\Phi(\gamma_j^t))\,\text{Binom}(t_{ij};s_{ij},\Phi(\alpha_i^t))\,\text{Pois}(s_{ij};\Phi(\beta_j^t)\lambda_i).$$

The likelihood to be estimated is then

$$L = \prod_{i=1}^{20}\prod_{j\neq i} f(x_{ij},t_{ij},s_{ij}|\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda}) \cdot f(\boldsymbol{\alpha})f(\boldsymbol{\gamma})f(\boldsymbol{\beta})f(\boldsymbol{\lambda})$$

The model uses shots and shots on target as well as goals. However, for predictions we only care about the number of goals. We therefore need to find the distribution of X alone. If we first think of S_{ij} as known then X_{ij} is the result of two sequential binomial distributions. It makes intuitive sense that two binomial distributions in a row results in a binomial distribution with probability equal the product of the two probabilities, and the number of trials is equal too the number of trials for the first binomial distribution. I did not find a proof of this online, but I have proven it myself. The proof can be found in the Appendix. Since $X_{ij}|S_{ij} \sim \text{Binom}(s_{ij},\Phi(\gamma_j^t)\Phi(\alpha_i^t))$ and $S_{ij} \sim \text{Pois}\,\Phi(\beta_j^t)\lambda_i)$, we get $X_{ij} \sim \text{Pois}(\Phi(\gamma_j^t)\Phi(\alpha_i^t)\Phi(\beta_j^t)\lambda_i)$. In other words X_{ij} is Poisson distributed dependent on the parameters.

3.6 Multinomial model

An alternative to counting the number of goals scored by each team, and using the joint distribution of goals to calculate probabilities for the outcomes, is to calculate the probabilities directly using a multinomial distribution. This method is used less often as it uses less of the information available, and it is therefore assumed to give poorer results. However, a goal-counting model will only outperform a multinomial model if one uses the right distribution for the number of goals scored. When using a multinomial model the teams of the Premier League can be viewed as part of a dominance structure. The dominance structure can be made to be trancative or intrancative, but the model should be justifiable based on common understanding of football. Normally a dominance structure only has a winner and a looser in each dispute. In football however we also have the possibility of a draw. This will be dealt with as discussed in section 2.8.

A multinomial GLM usually takes one of two forms depending on the structure of the data. It can either be categorical or ordinal [Fahrmeir et al., 2013a]. For football matches one usually choose ordinal. Each team will have their own strength, the strength can either be relative to some fixed team, or they can be included as latent variables assumed to come from the same normal distribution. Only the latter of these will be presented in this project. The probabilities for the three possible outcomes of each match will be decided by the difference in strengths of the teams and the parameters, θ, η, that gives us the probability for the results in general. This can be written as

$$\begin{aligned}
\text{P}(H) &= \text{F}(\beta_h - \beta_a - \theta + \eta) \\
\text{P}(D) &= \text{F}(\beta_h - \beta_a + \theta + \eta) - \text{F}(\beta_h - \beta_a - \theta + \eta) \\
\text{P}(A) &= 1 - \text{F}(\beta_h - \beta_a + \theta + \eta)
\end{aligned}$$

Here $\text{F}(\cdot)$ is the cumulative probability function for a logistic distributed variable. β_h, β_a is the strength of the home and away team. $\theta = 0$ would give the special case of binomial GLMM with logit link function. Each team will have their own strength β and that strength can be modelled to change over time, this will be done as a Brownian motion. The strengths β will be modeled as a random intercept, with a common normal distribution. The whole model will be a multinomial ordinal generalized linear mixed model. As we model the probabilities for the results directly, and not the number of goals first, in this model there is no way to find separate values for the attacking and defensive strength of the teams simultaneously.

4 Results

4.1 Home Field

The home field advantage is something all football supporters agree exists. The question is how large it is, and also what causes it. Most supporters would say that the home field advantage is caused by them, the supporters. Is this true, or is it just the supporters trying to make themselves more important than they truly are? After all there is many advantages playing at home, besides having more supporters on the stands. I will try to answer these questions here. Using Model 1, that is the Poisson GLMM with teams as random effects, I will estimate the parameters c and h. If $\exp(c)$ is the expected number of goals scored by an away team then $(\exp(h) - 1) \cdot 100\%$ is the percent increase in number of goals scored by the home team. Calculating this number for the seasons 2009/2010-2020/2021 we get the figure seen in figure 5. Based on this plot it looks like the home field advantage has been about 0.3, meaning a team scores about 30% more goals at home compared to away. However, in the current season(2020/2021) the home field advantage has disappeared. This came as a surprise. The fact that it decreased was to be expected, but its disappearance was surprising. It gives validity to the supporters claim of importance. Not only financially, but during the match. Not many players can say that the teams scores 30% more with them on the field, so the cliche about having a twelfth man in the stands, might have some merit.

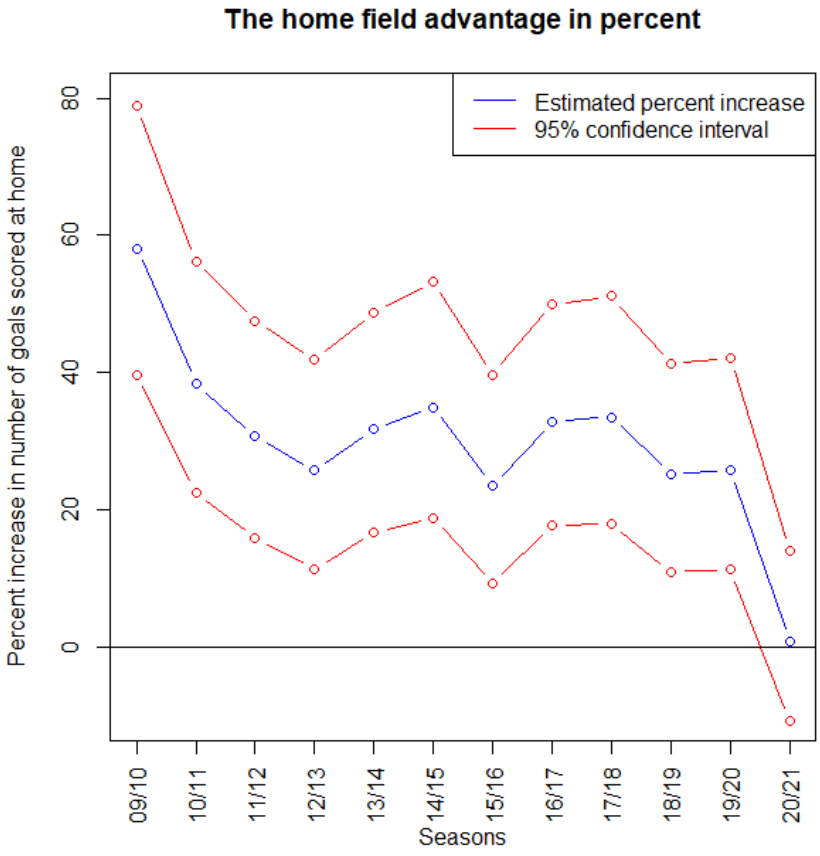


Figure 5: The home field advantage shown as the increase in expected goals scored, with a 95% confidence interval. The increase is reduced to almost zero during the pandemic.

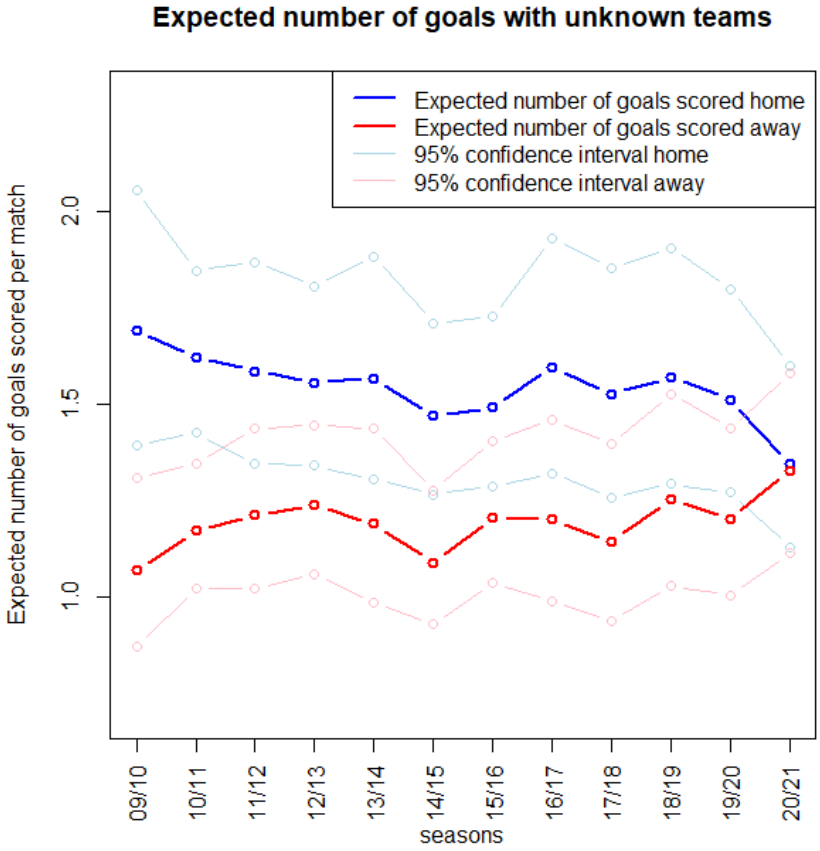


Figure 6: The expected number of goals scored by away and home team, with teams unknown. We see that the number of goals scored by home and away teams, is the same without the supporters.

In figure 6 the expected number of goals scored by the home team, and the expected number of goals scored by the away team in one match, for all the different seasons, is plotted. First it is important to say that since $\log(\lambda)$ has a normal distributed random intercept, we can say that $\log(\lambda)$ is normal distributed when the teams are unknown. Since $\log(\lambda)$ is normal distributed that makes λ log-normal distributed and the expectation is then $\exp(\mu + \frac{\sigma_a^2+\sigma_d^2}{2})$, where $\mu = c$ for the away team, and $\mu = c + h$ for the home team. We see that the expected number of goals for the home team has decreased, but at the same time the expected number of goals scored by the away team has increased. This could indicate that the presence of the supporters does not only make the home team play better, but also make the away team play worse. Alternatively it could be that the supporters cause both the home team defence and the home team attack to play better than they would without the supporters.

The effect of the Home field advantage can also be modelled using a multinomial ordinal generalized linear mixed model. This model was described in section 3.6. After the random effects are integrated out, we get the best estimates for the probabilities of home win, draw and away win. In figure 8 we see the ratio of the probabilities for home win and away win. One would assume this ratio to be at least one. We see that it goes down to one in the current season(20/21). In figure 7 the probabilities for the different results are plotted for all the seasons. The probabilities for home victory is always above the one for away victory, until the current season, where the probability for away victory is actually estimated to be a little higher, this difference is not significant.

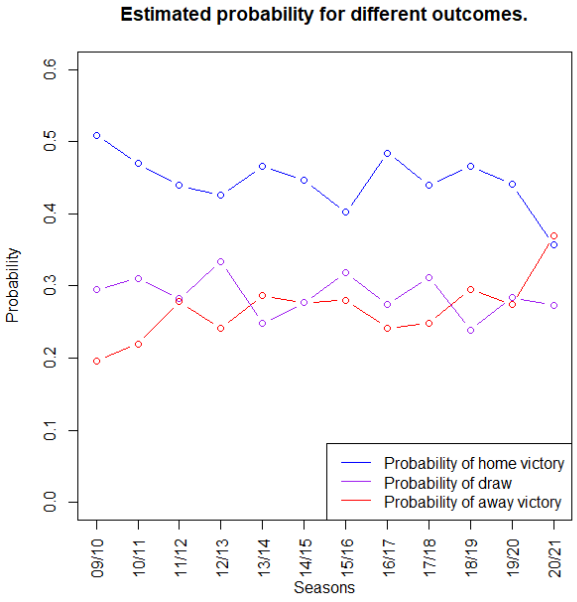


Figure 7: The estimated probability for the different outcomes using a Multinomial GLMM.

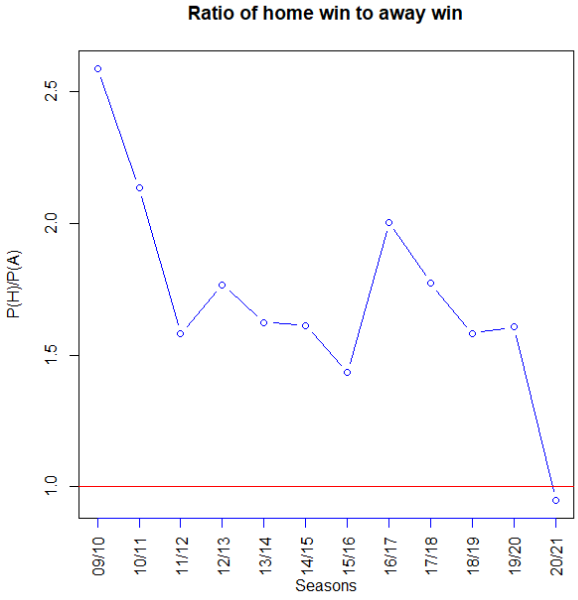


Figure 8: Probability of home victory divided by the probability of away victory. The ratio is estimated using a Multinomial GLMM.

In figure 8 we see that in earlier seasons when teams with the same strength met, the probability of victory for the home team was between 1.5 and 2 times as high as it was for away victory. However, during the pandemic where the stands have been empty, the probabilities for home victory and away victory are equal.

4.2 Model 1

Model 1 as described in section 3.2 is based on a the assumption that the number of goals scored by each team follows a GLMM Poisson distribution, with the teams as random effects. In this section I will see if the assumptions for a Poisson model is contradicted. I will also test the model with different betting strategies as described in the theory. The model assumes that each team has an offensive and a defensive strength, and the number of goals scored by a team is dependent on its own attacking strength, the opponents defensive strength and whether or not they play at home.

In figure 9 and 10 the attacking and defending strengths for all the teams in the 19/20 season of the Premier League are plotted. The left figure illustrates the attacking strengths for each team. It is estimated that a team with attacking strength 1, will expect to score 1.43 goals at home field and 1.13 as a visiting team, against a team with defending strength 1. The strength plotted in figure 9 gives the factor to multiply these expected values with. The reason it is multiplied is that the log expectation is linear, making the expectation a product of the exponentials. We see that Man City is the best attacking team in the league, and it is expected to score 19% more goals than Liverpool (the second best), against the same opponent. In figure 10 the defending capacity is plotted. High values means that the team is expected to concede more goals than those with lower values. We see a clear negative correlation between the goal scoring capacity and the defending capacity, this is likely the cause of the negative correlation between goals scored and goals conceded in the data. And this is thus not a contradiction of the assumptions in the model. When including a correlation in the attacking and defending strength, that means having Σ from equation (4) not being diagonal. The correlation was estimated to 1. This means that we only included one strength parameter for each team. The correlation was found to be significantly different from 0, but it performed worse on

the predictions of future results. Most teams will have a positive correlation for attacking and defending strength, but some teams score many goals and concede many, and some score few and concede few. I chose to assume Σ as diagonal, because assuming that all the teams have the same correlation between attacking strength and defending strength, performed worse on the prediction of future results, even if it did fit the data better. The reader is now informed that the rest of the results have not used any correlation between the attacking strength and defending strengths in the model.

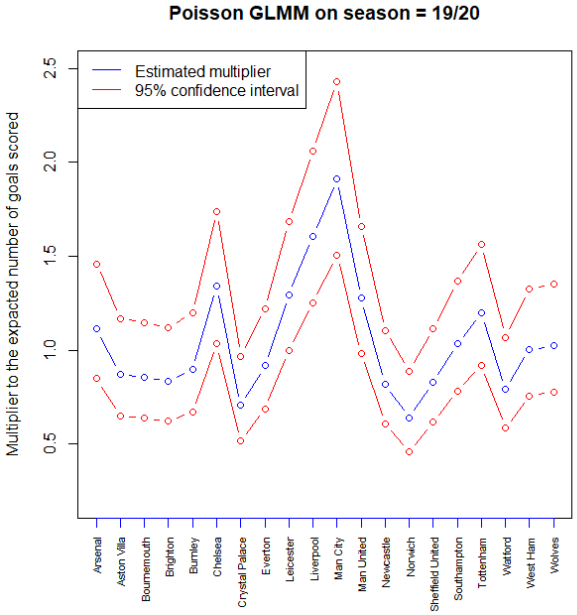


Figure 9: The product for expected number of goals scored for each team in the PL in the 19/20 season. The blue line is the estimate and the red lines are 95% confidence interval.

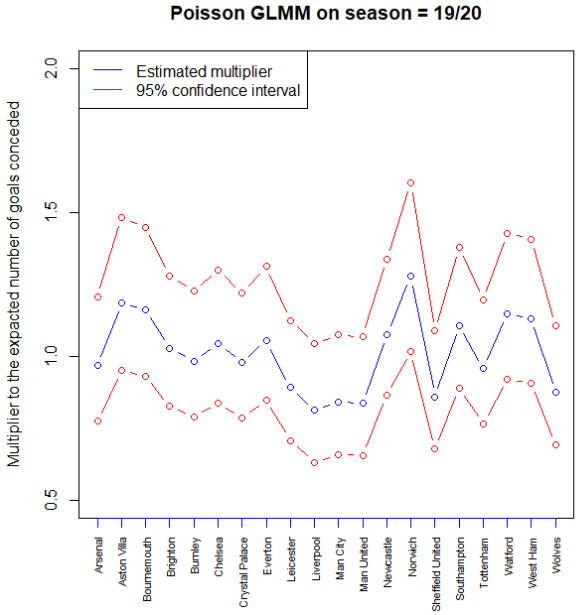


Figure 10: The product for expected number of goals conceded for each team in the PL in the 19/20 season. The blue line is the estimate and the red lines are 95% confidence interval.

The Poisson distribution has one parameter to represent both the expectation and the variance. If the variance dependent on the random effects is lower than the expectation dependent on the random effects, we say that the model is under-dispersed. And if the variance dependent on the random effects is higher, we call it over-dispersed. If over- or under-dispersion is present, we might want to use the generalized Poisson distribution. This will give the same expectation as the Poisson distribution, but the variance will be closer to the empirical variance. The reason variance is important in this model is that we do not really care about the number of goals scored by each team, only which team scores more goals. If over- or under-dispersion is present, the probabilities for the outcomes of the games will be affected.

If the data is under-dispersed, the true variance in the data is lower than the variance used by the model. This will give large deviations away from the mean higher probability than they truly have. The result of this is that the model will give a higher probability for victory to the assumed weakest team, than is the true probability. Conversely if the data is over-dispersed the estimated probability of victory to the assumed best team will be higher than it truly is.

A search for over-dispersion in a normal GLM is done by dividing the residual deviance by the degrees of freedom. If it is significantly higher than one, it is over-dispersed and lower than one it is under-dispersed. However, when random effects are included, the procedure is not as easy. I have used code from Daniel Lüdecke, and found this model to be under-dispersed. But not necessarily significantly so. The code only gives p-values for over-dispersion. If the model truly is under-dispersed it means that a model using generalized Poisson should give better estimates for the probabilities.

This model was used in this project to place bets on the betting market. How it did compared to the other models and across the seasons, will be discussed in section 4.7 In this part I will show some specific examples from different seasons. In figures 11 and 12 we see the results from testing the model on the last half of the seasons 19/20 and 20/21. The betting strategy used, is the one introduced by Rue and Salvesen [2000], the one that maximizes the expectation of the profit minus the variance of the profit. The model does make a profit in both seasons, but not by much in the 19/20 season. The grey lines in the plots are different realizations of the profit if the results had been randomly drawn with the probabilities estimated. If the probabilities given by the model was the true probabilities, the resulting profit(the red line in the plot) should resemble the grey lines. We see that for the 20/21 season the red line performs about as well as the grey lines, while in 19/20 the red line performs worse than all but 3 of the 20 different realizations.

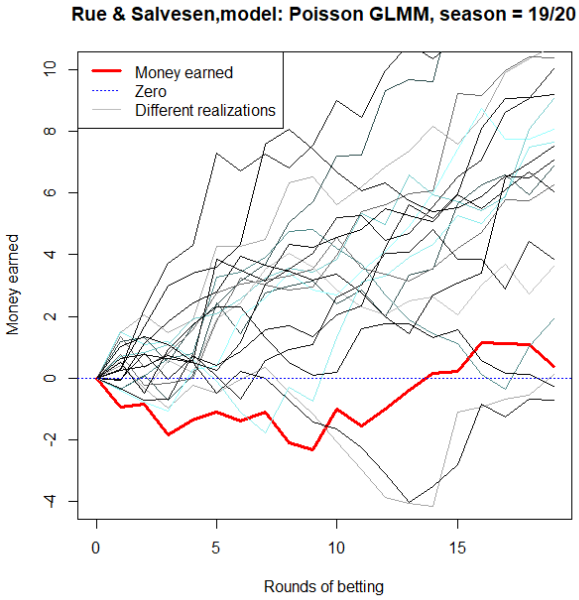


Figure 11: The red line is the amount of money that had been earned at different times. 1 unit of money was wagered at each round.

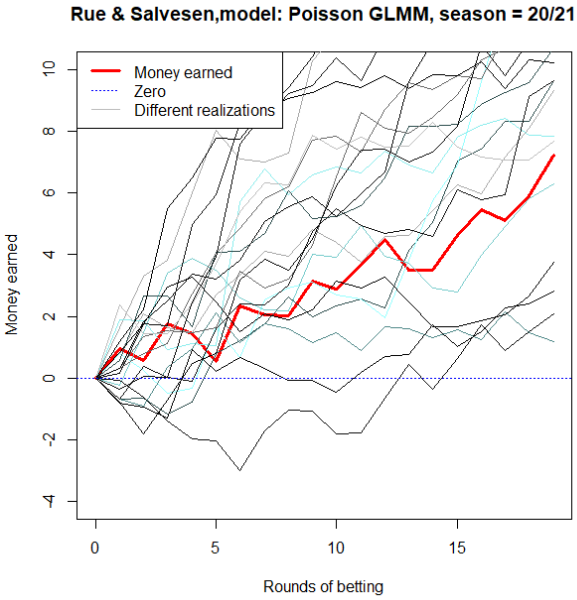


Figure 12: The red line is the amount of money that had been earned at different times. 1 unit of money was wagered at each round.

Based on the the plots in figure 11 and 12 we can see how much money the model earned for the two seasons, but not which bets the model made, and neither how many of the bets was won. In figure 13 and 14 the bets that were found to have positive expectation are plotted with the odds along the y-axis, and the estimated probability along the x-axis. We see that many more bets are lost than won for both seasons. This is to be expected as most of the bets have probabilities estimated to be low. Again it is important to stress that the goal is not to win as many bets as possible and betting on the most probable result, but rather to bet on the ones with high expectations. Had the model found the exact same probabilities as the betting companies, we would have placed no bets. This is because the betting site chooses the odds to be lower than the inverse of the probabilities. We notice that all the bets are close to the diagonal line, going from the top left to the bottom right. I have chosen to plot the odds-probability relationship on a log-scale. The reason for this is that then there exists a linear relationship between them. We know that $\omega_i \propto 1/p_i$ from there follows $\log(\omega_i) \propto -\log(p_i)$. All the bets with positive expected profit will therefore be in the right upper triangle of the plot. We see that my estimates for the probabilities are further away from the odds for high odds. This is to be expected, as small changes to a low probability will have huge effects on the odds. On the whole it looks like all the bets are close to the imaginary line going along the diagonal. This means that

the model is never completely in disagreement with the betting company about the likelihood of a result. If we compare the results of the bets with odds higher than 10, we see that in the 19/20 season the model was either unlucky or it overestimated the probability of unlikely results. It lost 23 of 24 bets where the odds were above 10. However, in 20/21 the model lost 16 out of 19 bets with expectation above 10. And this is one of the reasons the model did so well in the current season, compared to last years. The three shocking results in question are: Chelsea lost 2-5 at home to West Brom with 15.0 in odds. The model estimated an away victory to have probability 0.13, giving an expected gain of $p\omega - 1 = 0.13 \cdot 15 - 1 = 0.95$. Man. United lost 1-2 at home to Sheffield United with 11 in odds. The model estimated the probability of this outcome to be 0.16, giving an expected gain of $p\omega - 1 = 0.16 \cdot 11 - 1 = 0.76$ And lastly Man. City lost at home to newly promoted Leeds United with 10 in odds. The model estimated this to have a probability of 0.19, giving an expected gain of $p\omega - 1 = 0.19 \cdot 10 - 1 = 0.9$. We see that for all these bets the estimated probabilities are much higher than the ones used by the betting sites. This is what we would expect if the model was under-dispersed.

When a negative-binomial model was fitted to the data the θ tended to infinity, as the negative binomial model has no way to model under-dispersion, only over-dispersion. This is a clear sign that there is no over-dispersed.

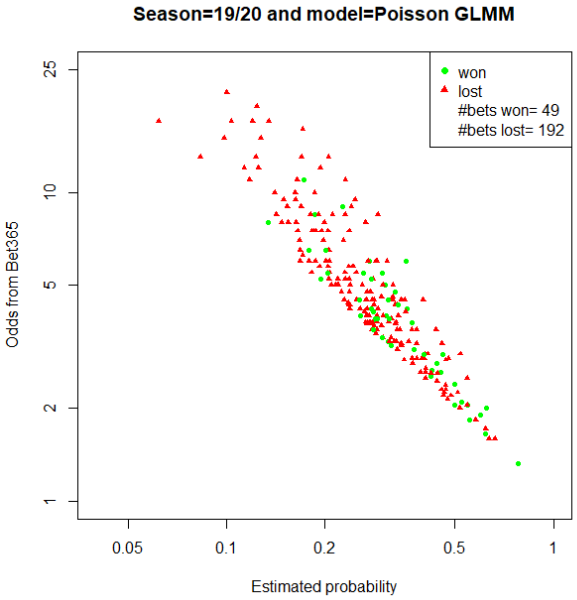


Figure 13: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

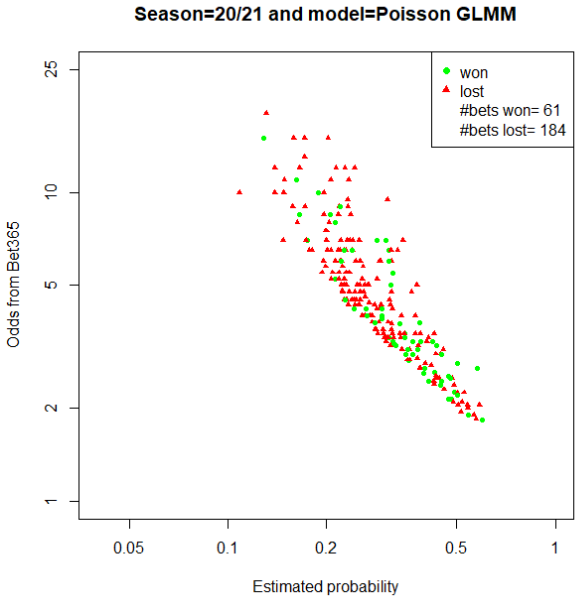


Figure 14: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

4.3 Time dependent model

Using the model described in Section 3.3 for the season 18/19 in the Premier League, I get $\hat{\tau} = 7683$ and $\hat{\sigma}_a^2 = 1/16$ and $\hat{\sigma}_d^2 = 1/33$. If the estimated value for τ is true, one would need to wait 21 years for the variance within a teams strength to equal the variance in the distribution in the strengths across the teams. Because of this very high value of τ we get almost the exact same attacking and defending strengths for this more advanced time model, as we got with just random intercept from Section 3.2. This has been found before by Rue and Salvesen [2000] and their solution was to set $\tau = 100$ and $\sigma = 1/37$, they found these

values by looking at distributions in the data in some way, it was not clear from the article [Rue and Salvesen, 2000]. This value is far lower than the estimated value based on a single season. There is a possibility that within just a single season there is not enough variation in the teams performance, and that any variation in results can be explained by the stochastic nature of the model. Therefore the value for the variance of the Brownian motion is found to be so low. If this is the case a model using all 12 seasons at the same time, should do better. By using all 12 seasons of the Premier League at the same time, we get $\hat{\tau} = 3024$. This is much lower than the value for just one season, but still much higher than 100. This much higher value might be caused by a lack of information in the data about this parameter. The risk of picking a too low value for τ is that the variation in the strength modeled by the Brownian motion, might just be a variation in result caused by the stochastic nature of the game. In other words we might over-fit. A more pragmatic way to compare them will be to see what model performs best at the betting market, and to choose that model. A comparison of the models will be made in section 4.7 and 4.8.

4.4 Generalized Poisson Model

The Generalized Poisson model was introduced in Section 3.4. The only change from the model used in section 4.2 is that this will use the Generalized Poisson model, and not the ordinary Poisson model. The reason to use a Generalized Poisson model is that under-dispersion was present in the Poisson model. To test if Generalized Poisson model fits better than the Poisson model one can perform a deviance test, as these are nested models. As the larger model has one parameter more to estimate than the smaller model, we get that twice the difference in negative log-likelihood of the models is χ^2 distributed with one degree of freedom, under the null-hypothesis that the two models are equally good [Fahrmeir et al., 2013b]. The null-hypothesis is that $\delta = 0$, and the alternative hypothesis is that $\delta \neq 0$.

By using all 12 seasons in the data-set, and estimating different constant random intercepts for all the teams each season, we get a p-value of 0.0312 and $\hat{\delta} = -0.0168$. There seems to be some under-dispersion when all the teams have a different random intercept each season. However, this under-dispersion might be the result of over-fitting. The value found for δ gives a 3.3% reduction in variance compared to a ordinary Poisson distribution. Next we will look at the results of the time-dependent model using Brownian motion, for a ordinary Poisson model and a Generalized Poisson model.

We fit a Generalized Poisson model, using all 12 seasons in the Premier League, where the attacking and defending strengths follow a Brownian motion, this is a similar model too my simplified version of the one used in Rue and Salvesen [2000]. Except in this case we use the generalized Poisson distribution, We also fit the same model using the Poisson distribution. The estimate $\hat{\delta} = -0.0144$, with standard deviation 0.00764, was obtained from the Generalized Poisson model. The standard deviation is found from the Hessian matrix to the score-function in TMB. The estimates are approximately normal with mean equal to the estimate, and variance equal to the standard deviation squared. Assuming this normal distribution, we get that the p-value is 0.0588, and is just to high for $\alpha = 0.05$. Using the deviance test described above we get $d = 3.537$, this gives a p-value of 0.0600, which again is just above the most used critical value for significance. We see that there is no indication of over-dispersion, which has been assumed in the past. And that any under-dispersion that is present, is not significant using the Brownian motion models. However, from a more pragmatic point of view, what matters the most, is how well the model does on the betting market. Sadly the Generalized Poisson model does not perform better than its simpler counterpart on the betting market. Possible reasons for this will be discussed later.

4.5 Multinomial

The multinomial model described in section 3.6 will be discussed here. Some of the results have already been discussed in section 4.1, and how it did on the betting market will be discussed in section 4.7. By using all twelve seasons of the Premier League at once, and having the strength of the teams modeled by a Brownian Motion, predictions for the future should be as good as possible. In figure 15 and 16 we see the results from the betting market on the second half of the seasons 18/19 and 19/20. The grey lines are the results of the betting strategy with the results simulated from the estimated probabilities. The grey lines are almost

all above zero, this is not surprising as all the bets have a positive expectation given that the probabilities are true. The fact that the actual results, the red line, does so poorly compared with the grey lines, could indicate that the model fits poorly.

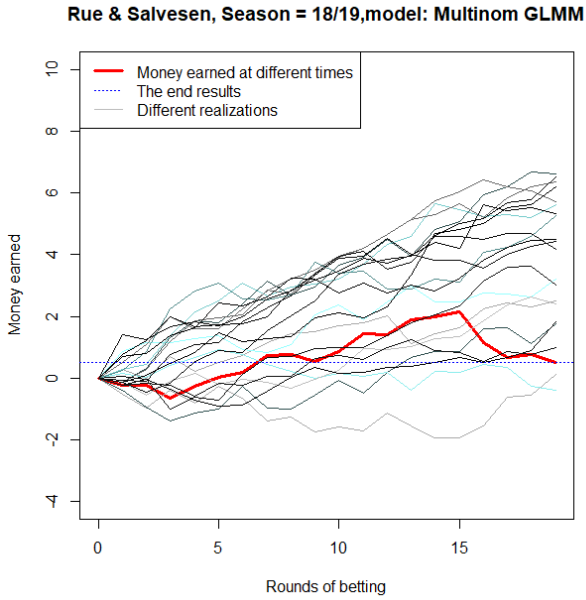


Figure 15: The red line is the amount of money that had been earned at different times. 1 unit of money was wagered at each round.

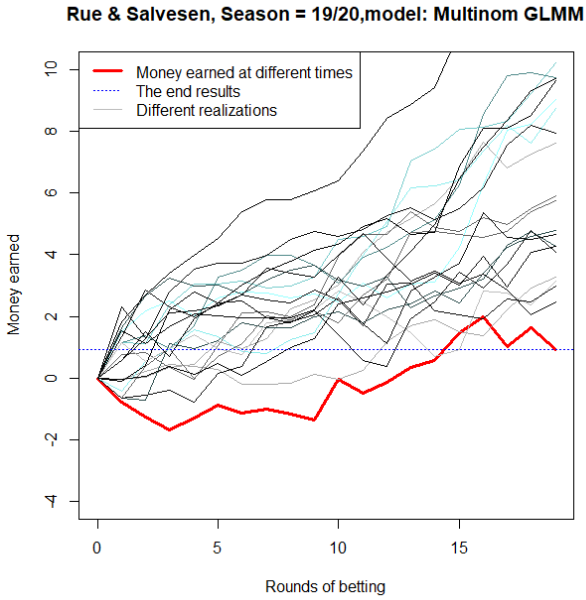


Figure 16: The red line is the amount of money that had been earned at different times. 1 unit of money was wagered at each round.

This can also be seen in figures 17 and 18. These two plots are meant to illustrate a pattern that is present in all of the seasons. We see that the probabilities estimated by the model is often much higher than the probability the odds are based on. For example one match in season 19/20 with 17 in odds, the model predicts it is a 20% chance of victory. This match is colored red meaning that the bet was lost, and the estimated probability was likely too high. It is difficult to say if the probability truly was too high, since even if the probability truly was 20%, we would still expect to loose the bet. We see that the model often estimates too high probabilities for unlikely events. However, for more likely events, that is games with less than 5 in odds, it does better than the Poisson model. The high probability estimated for unlikely results, will be punished when using the Kelly Criterion as described in section 2.7.3.

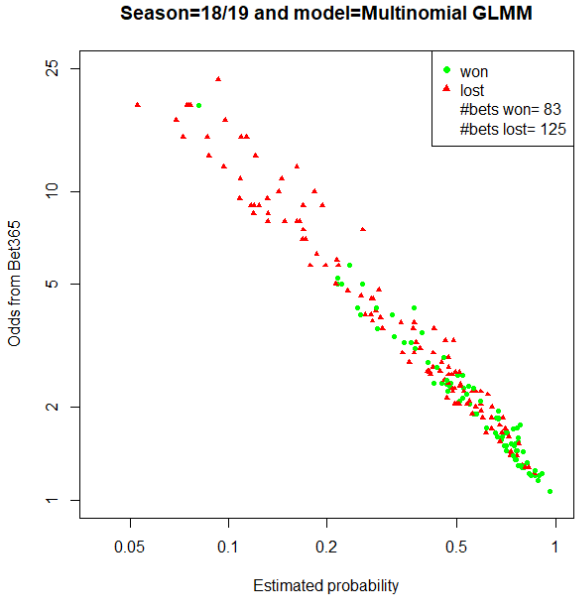


Figure 17: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

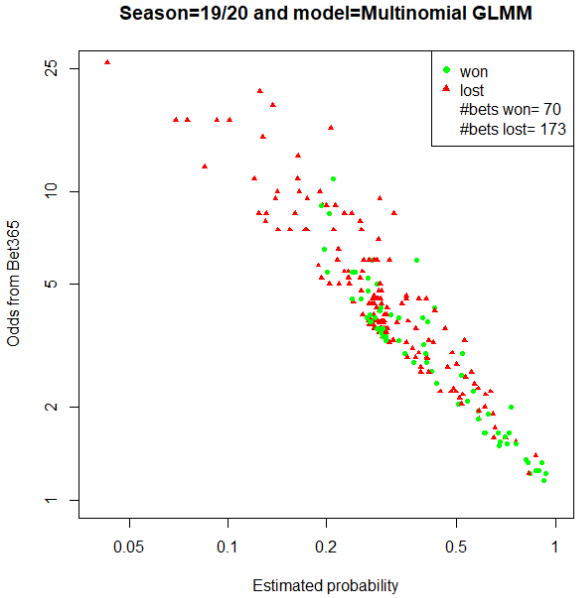


Figure 18: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

4.6 Data Intensive

The model discussed here is the one described in section 3.5. It is called data intensive because it uses more of the data available than any other model. It includes the number of shots, and the number of shots on target, as well as goals, in the model. It is a common understanding that more data included in the model will give better results. However, that is only the case if the data is included in a sensible way. The relationship proposed in Langseth [2013] was that a chance could become a goal for the team in a sequential way (chance→ shot→ shot on target→ goal). The number of chances was modeled with a Poisson distribution where the expected number of chances was specific to each team. This relationship seems strange, as it indicates that the expected number of chances is the same independently of the opponents strength. It was proposed that a binomial relationship existed between chances and shots, and that this relationship was only dependent on the strength of the defence of the defending team. The probability that the attacking team managed to end their chance with a shot, was only dependent on the defending teams strength. In the same way the relationship between shots and shots on target was also assumed to be binomial, but this time the probability was assumed to only depend on the attacking team. Again it seems unreasonable that a team will have the exact same chance of getting their shot on goal against all the different opponents. Some are good at defending and force the opponent to shoot from far away, for example. Lastly the binomial relationship between shots on target and goals can seem as a good approximation. The model does include both the attacking and defending teams in its approximation of number of goals scored, but only sequentially.

I will use the 19/20 season to illustrate some of the results the model gives. The estimated parameters are illustrated in figure 19.

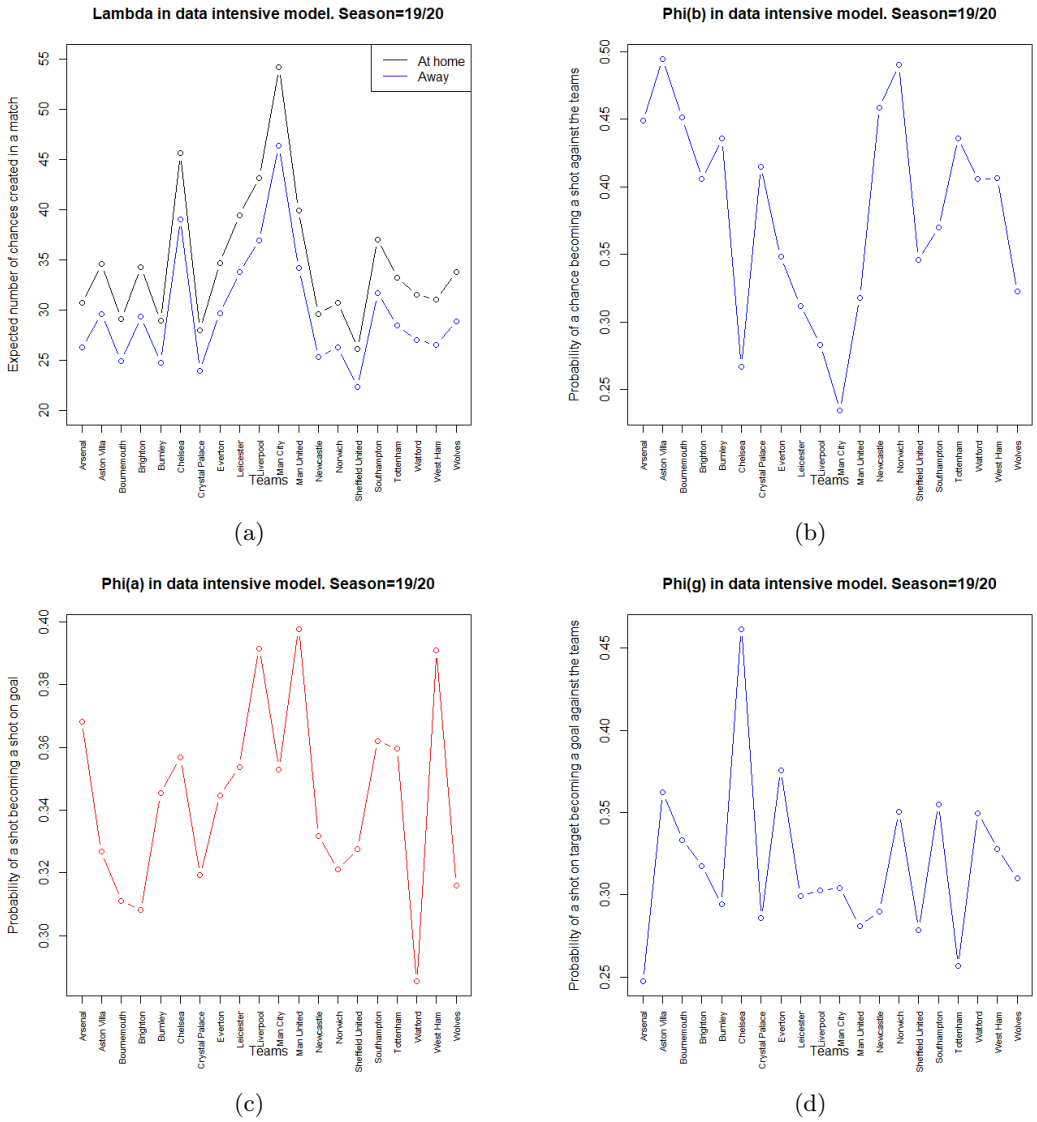


Figure 19: The estimated values using the Data intensive model on the 2019/2020 season of the Premier League.

We see from figure 19 (a) that there is a large difference in the number of expected chances created by each team, teams create more chances at home than away, which is as expected. From figure 19 (b) we see that some teams are much better at defending than others. Manchester City seems to be the team that has the lowest probability of letting a chance become a shot against them. And the probabilities vary from less than 25% to about 50%. In figure 19 (c) the probability of a shot becoming a shot on target for the attacking team, is plotted. Here the probabilities are very similar to each other. I do not think this is because the different teams in the Premier League are about equally good at attacking, but rather because using shots and shots on target as the only way to measure the attacking strength, is an inefficient way to model the attacking strength. Lastly how good the goalkeeper is, is modeled in figure 19 (d). We see that Chelsea is by far the worst, and that Arsenal and Tottenham are both rather good. If this is because Chelsea has a weak keeper compared to Arsenal and Tottenham, or if it is due to a playstyle where Arsenal and Tottenham more often force the opponents to shoot from farther away, is unclear from the results. This is one of the

possible weaknesses of the model.

The results in figures 19 (c) and (d) is not necessarily significant. We have used 20 parameters for each of these probabilities, and the probabilities are very similar for all the teams. The significanse of the parameters can be tested with a deviance test, where the simpler model uses the same value for α and the same value for γ, for all the teams. The deviance test will have 38 degrees of freedom. By using the same value for all the teams for α and γ we get a deviance of $d = 65.6$, and with 38 degrees of freedom that gives a p-value of 0.004. Therefore the model is significant. Testing the larger model against the one with the same attacking strength for all teams, also gives the result that the larger model is to be preferred, with a p-value of 0.01. We now know that the larger model is significant, but how well it fits the data compared to the other models, will be shown in section 4.7 and 4.8.

4.7 How it performed on the betting market

This subsection is dedicated to the results each of the models had on the betting market. The results for the following betting strategies: Fixed bet, Fixed return, and Rue and Salvesesens betting strategy; will be given in percentage gained or lost, as was done in Rue and Salvesen [2000] and Langseth [2013]. The models bets the same total amount(of money) in each round, this amount is called C. If $\mathbf{\Delta}_i$ is the profit from round i, the total profit after n rounds is $\sum_{i=1}^{n} \mathbf{\Delta}_i$. The percentage of gain or loss can be calculated as $\frac{\sum_{i=1}^{n} \mathbf{\Delta}_i}{nC} \cdot 100\%$. This calculation of profit assumes that when you place money on a bet you can never use the same money on another bet, even if you win the bet. If one developed a model that always won money every round, this way to calculate return of investment, would give an underestimate to the actual return of investment. Betting different sums based on how well the model has done previously is the idea used in the Kelly Criterion. The Kelly Criterion is included, not as an actual betting strategy, but as a test to see how well the models fit the results. Therefore the Kelly Criterion will be used on each individual match, and not on each round simultaneously as the other models are. The return of investment for the Kelly Criterion is therefore $\frac{K-C}{C} \cdot 100\%$, where C is the money you started with, and K is the money you ended up with. I have used a fractional Kelly Criterion where I only bet 20% of the fraction found by the actual Kelly Criterion. The results of the constant model can be viewed in table 1. The results are calculated by betting on the second half of the season. The estimated probabilities are based on all the previous information.

Season	Model	Fixed Bet	Fixed Return	Rue & Salvesen	Kelly Criterion
20/21	Poisson GLMM	11.19	1.36	38.05	214.95
20/21	Generalized Poisson	17.33	6.42	42.05	279.47
20/21	Multinomial GLMM	14.11	8.95	20.52	90.73
20/21	Data Intensive	-0.14	-14.98	-1.78	-2.44
19/20	Poisson GLMM	-16.14	-18.73	1.93	-20.57
19/20	Generalized Poisson	-17.70	-19.07	2.74	-22.14
19/20	Multinomial GLMM	-3.44	-3.31	4.94	-7.72
19/20	Data Intensive	-14.31	-13.25	-11.64	-23.92
18/19	Poisson GLMM	5.80	8.93	15.34	21.73
18/19	Generalized Poisson	2.90	4.40	7.06	10.00
18/19	Multinomial GLMM	-0.31	0.60	2.69	0.23
18/19	Data Intensive	-7.02	-4.70	6.56	21.63
17/18	Poisson GLMM	10.68	9.85	-3.49	-30.15
17/18	Generalized Poisson	8.48	5.02	-6.15	-36.02
17/18	Multinomial GLMM	-6.34	-6.66	-22.67	-68.15
17/18	Data Intensive	-0.45	-8.87	-0.43	-3.01
16/17	Poisson GLMM	-8.32	-2.83	-1.83	-35.35
16/17	Generalized Poisson	-10.79	-2.07	0.08	33.77
16/17	Multinomial GLMM	-2.53	7.53	17.55	21.55
16/17	Data Intensive	-16.52	-10.39	-2.13	-22.14
15/16	Poisson GLMM	-8.83	-7.07	-16.23	-47.04
15/16	Generalized Poisson	-8.10	-5.49	-15.94	-46.25
15/16	Multinomial GLMM	-1.45	-2.38	-4.26	-25.43
15/16	Data Intensive	-3.48	-2.53	-9.67	-36.61
14/15	Poisson GLMM	-16.60	-19.87	-7.67	-28.15
14/15	Generalized Poisson	-18.28	-21.12	-11.02	-27.32
14/15	Multinomial GLMM	-7.64	-12.14	-23.11	-50.63
14/15	Data Intensive	-4.76	-3.36	-12.30	-36.91
13/14	Poisson GLMM	-5.43	-3.76	-12.54	-33.35
13/14	Generalized Poisson	-8.50	-7.54	-11.44	-31.98
13/14	Multinomial GLMM	-10.38	-5.68	-2.02	-11.22
13/14	Data Intensive	5.28	12.56	8.24	12.58
12/13	Poisson GLMM	-18.79	-18.39	-15.08	-52.65
12/13	Generalized Poisson	-17.94	-18.10	-15.83	-50.12
12/13	Multinomial GLMM	-13.17	-9.02	-19.63	-51.46
12/13	Data Intensive	-26.62	-26.40	-28.19	-60.56
11/12	Poisson GLMM	24.30	23.28	32.36	143.67
11/12	Generalized Poisson	25.21	23.63	32.68	145.50
11/12	Multinomial GLMM	5.23	-0.31	19.53	36.84
11/12	Data Intensive	19.35	15.36	31.71	97.16
10/11	Poisson GLMM	-14.15	-9.16	-9.77	-52.31
10/11	Generalized Poisson	-15.45	-11.74	-9.91	-48.07
10/11	Multinomial GLMM	-9.16	-5.34	-5.06	-23.42
10/11	Data Intensive	-18.02	-13.44	-14.37	-35.46
09/10	Poisson GLMM	-18.69	-3.91	-8.03	-38.03
09/10	Generalized Poisson	-10.40	-0.06	-2.91	-32.64
09/10	Multinomial GLMM	-5.32	-0.67	-14.09	-46.31
09/10	Data Intensive	-15.11	0.16	–5.80	-34.34
total	Poisson GLMM	-4.58(-3.30)	-3.36(-3.31)	1.09(1.92)	3.56(7.34)
total	Generalized Poisson	-4.44(-3.89)	-3.81(-4.15)	0.95(1.30)	8.89(12.66)
total	Multinomial GLMM	-3.38(-3.21)	-2.37(-2.52)	-2.14(-1.06)	-11.25(-8.06)
total	Data Intensive	-6.82(-6.06)	-5.82(-6.36)	-3.32(-3.09)	-10.34(-8.15)

Table 1: The results with the different models used on the second half of the season. The total profit is included at the bottom. The value in the brackets are the result using the seasons 10/11-20/21. It is included to make comparison to the time-dependent models easier.

We see from the results in table 1 that the same models can earn a lot of money some seasons, and lose a lot in others. Over the 12 year period only the Poisson GLMM and the generalized Poisson GLMM managed to earn money in total, and that was only with the betting strategy of Rue and Salvesen. All the other models lost money in total. From the table it seems that the betting strategy proposed by Rue and Salvesen performs better than the simpler betting strategies. We also see that The Kelly criterion does well for the Poisson models thanks to the 11/12 season and the 20/21 season. We also see a clear trend that there is easy to earn money in some seasons and much harder in others.

We see that the data-intensive model performs slightly worse compared to the other models. This is especially noticeable in the last two seasons.

In figures 20, 21 and 22 the odds and estimated probabilities are plotted together for the last season in the data set, 20/21. We see that the data intensive model has lost a lot of bets in the 3-4 odds range. We also notice that the generalized Poisson model has estimated the probabilities of unlikely outcomes much higher than the other models, and much higher than the betting sites, especially for bets with high odds. In this season many unlikely results has occurred, and therefore the model has done very well.

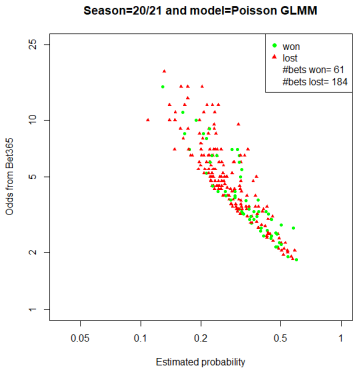


Figure 20: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

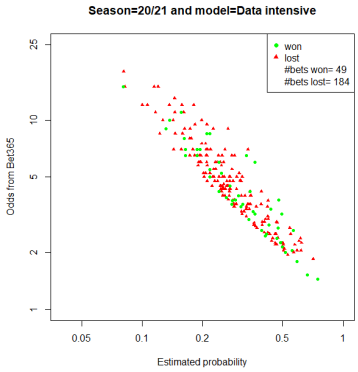


Figure 21: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

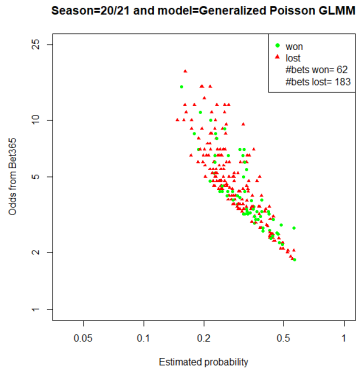


Figure 22: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

In table 2 and 3 the results from the models using Brownian motion, is shown. In table 3 the second half of the season is used, just like in table 1, while in table 2 bets have been placed in every round from 3 to 38. There are a total of three models compared in these tables. The first two are similar, they are based on the Poisson model with attacking and defending strength that varies with time as a Brownian Motion. One of the models estimate τ in the likelihood, and the other used $\tau = 100$. The estimate of τ in the likelihood is larger than 100. Therefore the model with constant τ will have strengths that vary more with time, than the model with τ estimated in the likelihood. The third model is Multinomial where the one strength parameter vary with time as a Brownian motion.

Season	Model	Fixed Bet	Fixed Return	Rue &Salvesen	Kelly Criterion
20/21	Const $\tau = 100$	7.19	0.22	13.42	89.38
20/21	τ form MLE	5.55	-4.70	-0.76	-25.82
20/21	Multinomial	7.98	-0.14	-4.64	-54.01
19/20	Const $\tau = 100$	14.21	6.44	21.62	55.07
19/20	τ from MLE	16.91	2.63	9.52	45.71
19/20	Multinomal	6.22	-5.39	5.20	66.83
18/19	Const $\tau = 100$	-4.21	2.84	12.85	39.26
18/19	τ from MLE	4.00	7.64	8.17	27.18
18/19	Multinomial	-6.19	-1.03	-7.70	-50.37
17/18	Const $\tau = 100$	-3.44	-2.86	-6.39	-11.75
17/18	τ from MLE	-8.53	-8.90	-6.97	-19.19
17/18	Multinomial	-1.64	-1.55	-19.95	-66.05
16/17	Const $\tau = 100$	-11.21	-7.98	-12.09	-62.43
16/17	τ from MLE	-18.43	-14.33	-12.56	-50.80
16/17	Multinomial	-21.73	-17.71	-12.30	-33.25
15/16	Const $\tau = 100$	8.11	0.30	3.70	23.83
15/16	τ from MLE	2.73	-1.36	-0.97	-14.33
15/16	Multinomial	1.60	-5.33	-16.09	-59.39
14/15	Const $\tau = 100$	-6.15	-7.67	-13.52	-49.33
14/15	τ from MLE	-3.78	-5.77	-6.41	-33.39
14/15	Multinomial	-5.22	-6.52	-5.15	-37.32
13/14	Const $\tau = 100$	-6.70	-0.63	-3.40	-17.48
13/14	τ from MLE	-4.61	-1.64	-4.33	-18.56
13/14	Multinomial	-6.67	-1.23	-5.73	-43.88
12/13	Const $\tau = 100$	-6.23	-4.78	-13.83	-55.26
12/13	τ from MLE	-11.79	-5.91	-7.67	-41.19
12/13	Multinomial	2.50	4.03	-0.69	-8.81
11/12	Const $\tau = 100$	4.46	4.82	5.25	34.37
11/12	τ from MLE	14.59	8.56	9.32	31.03
11/12	Multinomial	4.55	-1.24	8.21	24.82
10/11	Const $\tau = 100$	0.49	-6.12	-1.51	-35.33
10/11	τ from MLE	5.99	0.67	2.92	-20.99
10/11	Multinomial	6.80	2.27	0.23	-24.44
total	Const $\tau = 100$	-0.32	-1.54	0.55	0.93
total	τ from MLE	0.15	-2.28	-0.72	-10.00
total	Multinomial	-1.07	-3.08	-5.31	-25.99

Table 2: The results, given as percentage gained or lost, after betting on round 3-38, using the time-dependent models. All the data prior to the matches are used in the estimations of the probabilities.

Season	Model	Fixed Bet	Fixed Return	Rue &Salvesen	Kelly Criterion
20/21	Const $\tau = 100$	6.38	-5.20	21.75	72.94
20/21	τ form MLE	4.34	-6.21	0.76	-2.94
20/21	Multinomial	6.38	-5.20	21.75	72.94
19/20	Const $\tau = 100$	9.01	7.61	9.53	5.92
19/20	τ from MLE	4.58	3.98	14.44	33.57
19/20	Multinomal	0.73	3.07	18.97	74.24
18/19	Const $\tau = 100$	4.06	8.68	19.69	57.40
18/19	τ from MLE	3.00	6.72	12.54	35.13
18/19	Multinomial	1.22	5.46	-0.10	-4.25
17/18	Const $\tau = 100$	1.88	-0.45	-1.50	-3.67
17/18	τ from MLE	-2.23	-6.73	-2.83	4.92
17/18	Multinomial	-0.68	-7.12	-21.86	-43.27
16/17	Const $\tau = 100$	-2.19	-4.81	-7.01	-29.35
16/17	τ from MLE	-12.43	-13.16	-7.34	-14.82
16/17	Multinomial	-16.96	-18.03	-3.25	6.93
15/16	Const $\tau = 100$	9.63	3.94	-2.97	-10.72
15/16	τ from MLE	-0.33	-2.96	-8.02	-29.56
15/16	Multinomial	-4.35	-9.45	-13.61	-38.68
14/15	Const $\tau = 100$	-7.75	-7.02	-7.14	-14.13
14/15	τ from MLE	-3.70	-3.63	2.29	3.39
14/15	Multinomial	-12.62	-10.42	-4.40	-23.53
13/14	Const $\tau = 100$	-5.67	3.48	5.11	8.70
13/14	τ from MLE	-3.43	0.49	-5.32	-4.67
13/14	Multinomial	-11.26	-3.86	-9.92	-43.01
12/13	Const $\tau = 100$	-2.51	-6.51	-19.67	-38.11
12/13	τ from MLE	-9.69	-8.31	-14.55	-33.62
12/13	Multinomial	1.39	3.86	-6.23	-11.03
11/12	Const $\tau = 100$	29.26	24.60	34.83	123.40
11/12	τ from MLE	23.49	20.85	34.77	93.23
11/12	Multinomial	8.39	4.46	23.34	48.29
10/11	Const $\tau = 100$	-15.75	-13.09	-12.20	-35.31
10/11	τ from MLE	0.26	2.57	1.33	-22.73
10/11	Multinomial	2.57	-0.25	-13.32	-22.83
total	Const $\tau = 100$	2.40	1.03	3.67	12.46
total	τ from MLE	0.35	-0.64	2.94	6.91
total	Multinomial	-2.29	-3.41	-0.79	1.44

Table 3: The results, given as percentage gained or lost, after betting on round 20-38, using the time-dependent models. All the data prior to the matches are used in the estimations of the probabilities.

The results for the time dependent models are decent, at least for the second half of the season. We see that the models have problems in the first half of the season. Both the time dependent Poisson models managed to break even, or even have a small profit over the eleven year period. The multinomial model on the other hand did poorly. The odds given by the betting site is given in such a way that if their probabilities are correct, you will loose about 2.7% on average. Any model that looses less than that over a long time frame, could be said to work. The multinomial model does better than this, but not by much and it looses money in total over the whole period. That does not mean that the multinomial model is a useless model. It can be used to rank the different teams in the Premier League and look at how their strengths changed over time. However, on the betting market it does not give an edge against the betting sites. The model using $\tau = 100$ is the model that does best, with a profit over the eleven year period for all the betting strategies,

and a profit in 6 of the 11 years using the betting strategy proposed by Rue and Salvesen [2000]. But this combination earns a return of investment of 3.67%, which is less than just having the money in the bank over the same period, and with a lot more risk.

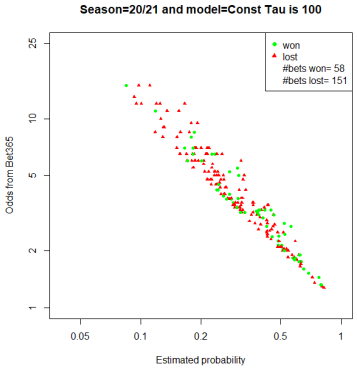


Figure 23: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectation. Red triangle means the bet was lost, green circle means the bet was won.

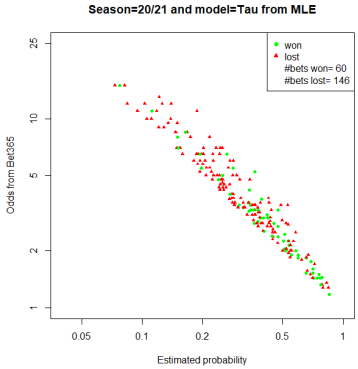


Figure 24: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

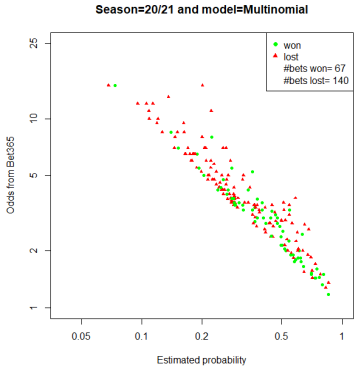


Figure 25: The odds from the betting site and the estimated probabilities plotted in the same figure. All the bets have positive expectations. Red triangle means the bet was lost, green circle means the bet was won.

The odds and probability of the bets with positive expectation from the season 20/21 is also plotted for the time dependent models in figures 23-25. If we compare it to the results in figures 20-22 we see that the time dependent models gives probabilities closer to the probabilities used by the betting sites, and that the constant models and the time dependent models have almost the same number of bets won across the season. However, the constant models have won more bets with higher odds this season, and won more money than the time dependent models.

4.8 Comparison

In subsection 4.7 we saw how the models did on the betting market. We saw some indications that the time dependent models did better than the constant models. Another way of testing which model has the best predictions for the outcome of the match, is the use of likelihood. When betting we estimate the probabilities of the outcome for the next 10 matches. By doing this for the last 19 rounds of the season, we get a total of 190 matches with estimated probabilities, per season. These probabilities are predictions and the actual results of the matches were not used when making the prediction. If $\mathbf{x}_i = (1, 0, 0)$ is the result of match i, if the home team wins; $x_i = (0, 0, 1)$ if the away team wins; $x_i = (0, 1, 0)$ for a draw. And $\mathbf{p}_i = (p_i^h, p_i^d, p_i^a)$ are the estimated probabilities for the different outcomes of match i, we can compare the models by a looking at the likelihood, L. $L = \prod_{i=1}^{190} f(\mathbf{x}_i|\mathbf{p}_i)$, where $f(\mathbf{x}_i|\mathbf{p}_i)$ is the probability mass function for a multinomial distributed random variable, evaluated at $\mathbf{x}_i$ given the probabilities $\mathbf{p}_i$. This is a way to compare the estimated probabilities from all the different models, and it is closely related to likelihood cross-validation. As we look at the predicted probabilities there are no risk of over-fitting, and the likelihoods can be compared directly. The total negative log-likelihoods for the seasons 10/11-20/21 using all the different models discussed in this project, can be viewed in table 4 and 5.

Model	Poison GLMM	Generelized Poisson	Data Intensive	Multinomial GLMM
nll	2044.04	2043.35	2047.40	2060.44

Table 4: The total negative log likelihood for the constant models. The likelihood is based on predictions for the last half of the last 11 seasons of the Premier League. Lower values indicate a better fit to the observations.

Model	Poisson Brownian motion $\tau = 100$	Poisson Brownian motion τ MLE	Multinomial Brownian Motion
nll	2015.45	2018.70	2048.99

Table 5: The total negative log likelihood for the time dependent models. The likelihood is based on predictions for the last half of the last 11 seasons of the Premier League. Lower values indicate a better fit to the observations.

Based on the negative log-likelihood we see that the time dependent models fit better to the data than the constant models. And that the Poisson model fits better than the multinomial ones. The Generalized Poisson model has the best fit of the constant models, and the ordinary Poisson model fits better than the data intensive model. The multinomial model does not fit the data well. We see that the models that give predictions that fit the data well, are the same models that made a profit on the betting market.

4.9 A discussion on the performance of the teams

The model discussed here is my simplified version of the one introduced by Rue and Salvesen [2000], and it has been explained in section 3.3. They used MCMC to estimate the parameters, while I used TMB which should be faster. They did not estimate τ, that is the "memory" of the model, as a parameter, but instead estimated it based on the distribution of the data. I estimated it as a parameter using a much larger data-set. Their estimate was 100 while mine was 3024.

Using this model I could see the relative strength of the different teams change over the 12 seasons in the data-set. Only 7 of the total 38 teams have been in the Premier League for all 12 seasons. The estimated strength of offence and defence for each team can be seen in Figures 26-32. We see that the teams strength vary over time, but not by much. There is seldom a large change within a season. This might explain why τ was estimated to be so large when only one season was used at a time. The plots are made with points, the days are marked along the x-axis and the multiplying factor is marked along the y-axis. When there is large white parts between the points it is due to a large time skip between matches. This is due to the summer break between the seasons, or if the team has been relegated to the lower league and then promoted back up. There is one large break due to Covid-19. There was no matches played between 08/03/20-17/06/20 in the 19/20 season. In the figures we can see that before a team is relegated they often have poor offensive and defensive outputs. An example of this is Norwich in figure 29 (f), they performed worse and worse and in the last seasons they were relegated shortly after promotion. This is not always the trend, in figure 29 (e) we see Newcastle's performance, and they were just slightly below average when they were relegated.

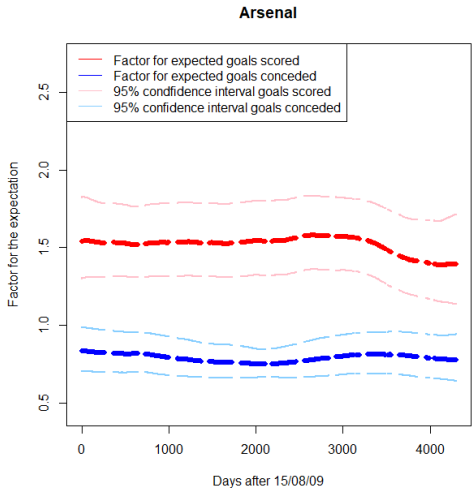
A team that has had a difficult decade behind them, is Manchester United. They started well by winning the league in 2010/11, and they won again in 2012/13. The last game in the 2012/13 season, a 5-5 game against Everton, was also the last game for their manager of 27 years, Sir Alex Ferguson. It had been predicted that his retirement would be problematic for Manchester United, and so it was. The last day he was manager corresponds with day 1373 in figure 29 (c). We see that from this day on Manchester United went from being the team with the best offence, to lagging far behind Manchester City and Liverpool. Based on the plots it looks like Manchester United are currently moving upward. This is also reflected in their placements the last seasons. They came second in the league this season behind city rival Manchester City. That is their best placement, tied with the 17/18 season, since Sir Alex Ferguson retired. Their current manager, the Norwegian and former Manchester United player Ole Gunnar Solskjær, took over the club from Jose Morinho 19/12/18(day 3413 in the plot). We see that the upward momentum has continued under

the Norwegian, and that they are almost as good now as they were under Sir Alex Ferguson, the difference being that the competition is much harder, as shown in the plots of the other teams.

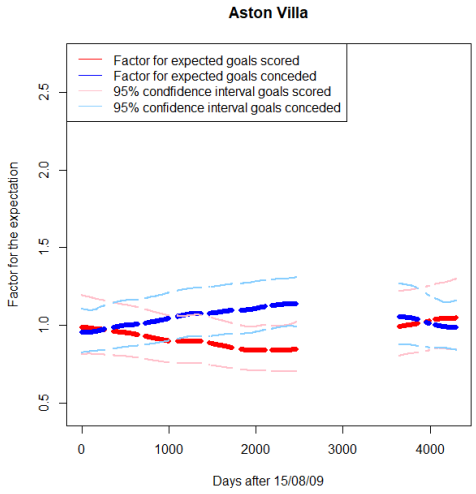
Arsenal is another fan favorite in Norway, they were at their best in the early 2000's, so their best period is not covered in the data set. Their manager of 22 years, Arsene Wenger, retired after much pressure form the fans in 2018. We can not say anything definitive based on these plots, but it can be of interest to see if Arsenal has performed better without Wenger, as many fans believed would be the case. His last day as manager was the last day of the 17/18 season, 14/05/18(day 3194 in figure 26 (a)). We recognise this in the plot as the exact point in time when both their attacking and defensive output, changed for the worse. Again there can be other factors than just the change in management, but Arsenal has performed worse since Wenger left.

The best teams the last seasons has been Manchester City and Liverpool, as can also be seen in figure 29 (a) and (b). Their performance have been incredible, and not just offensively. If we look at Manchester City for instance, not only do they have the effect of doubling the expected number of goals, but they also have had half the number of expected goals conceded. Liverpool had a local maximum in their runner up season of 13/14, but have become better since. Man City on the other hand has become better and better. Man City has won the league five times in this time period(11/12,13/14,17/18,18/19,20/21) while Liverpool has won once in the 19/20 season.

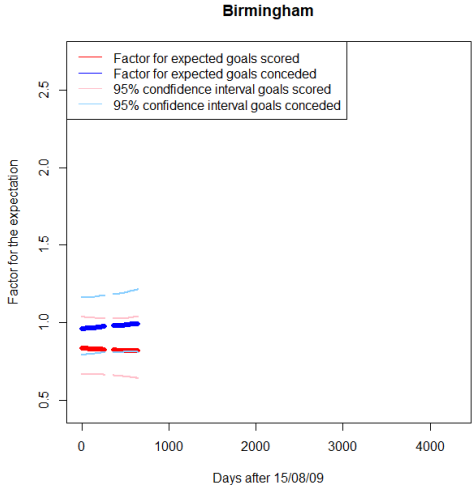
A possible weakness of the model can be seen in figure 27 (c), where the strengths of Chelsea is plotted. Based on the plot it looks like Chelsea has been one of the most consistent teams in the league. This is by no means the case. In the span of three seasons from 14/15-16/17 Chelsea won the league twice, with a 10th place in the middle season. We see this as the very tiny decrease in attacking strength. This consistency in the model, that was not present in the results, might indicate that Chelsea was extremely unlucky in the season of 15/16, and was about as good at the game as in the season before and after. This seems unlikely in my opinion, it seems more plausible that this is due to the difficulty of modeling correctly. τ is found to be very high meaning that the room for variation in strengths is small, so it can be difficult to model changes that goes up and down as fast as a single season. This might also be because we model the expected number of goals scored by the difference in the attacking teams attacking strength, and the defending teams defensive strength, and it can be difficult to say for sure if the number of goals scored was due to good offence or poor defence.



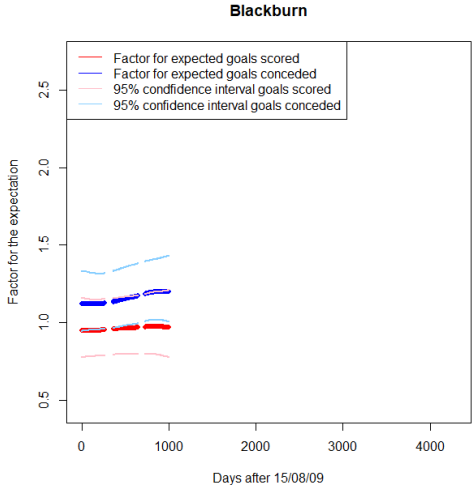
(a)



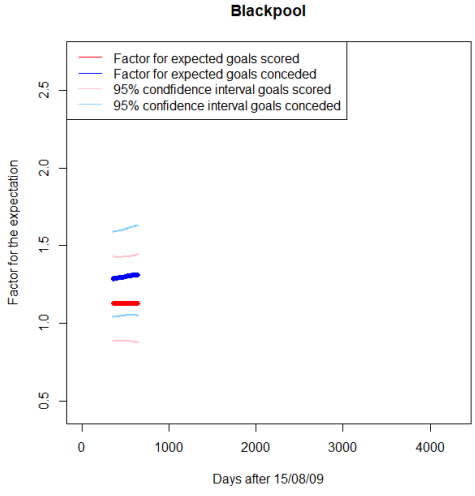
(b)



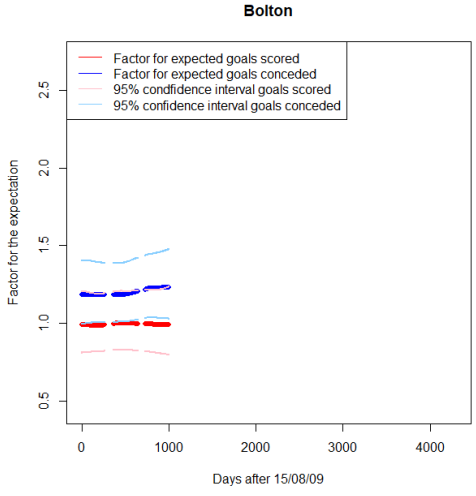
(c)



(d)

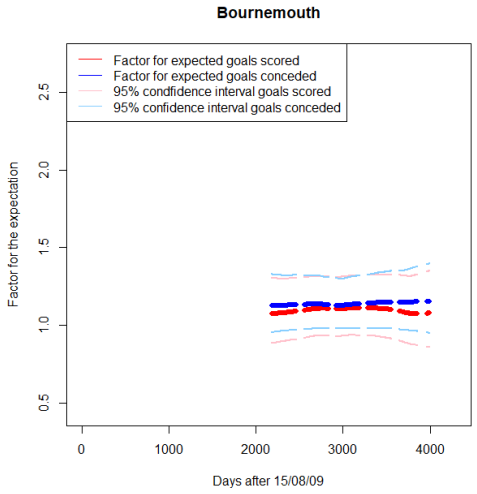


(e)

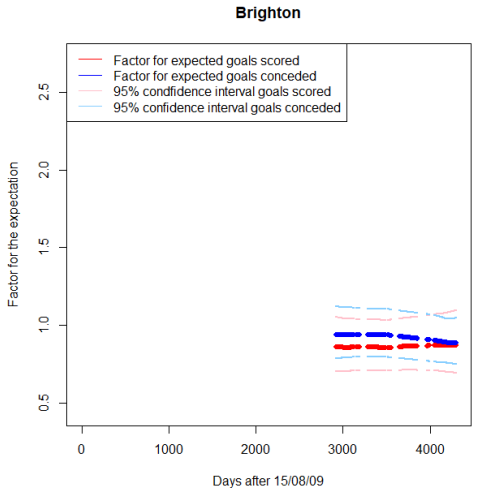


(f)

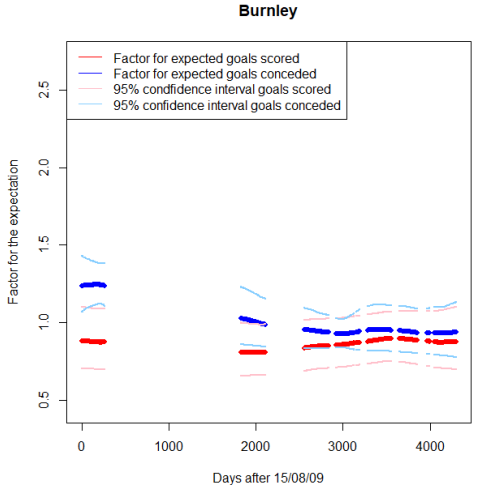
Figure 26



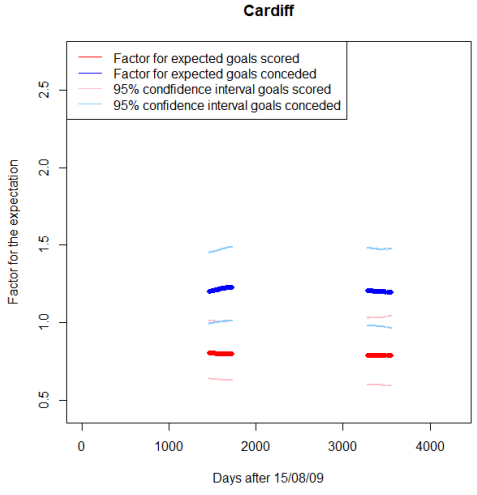
(a)



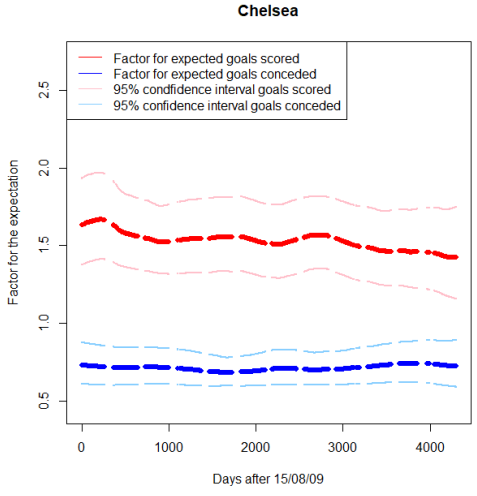
(b)



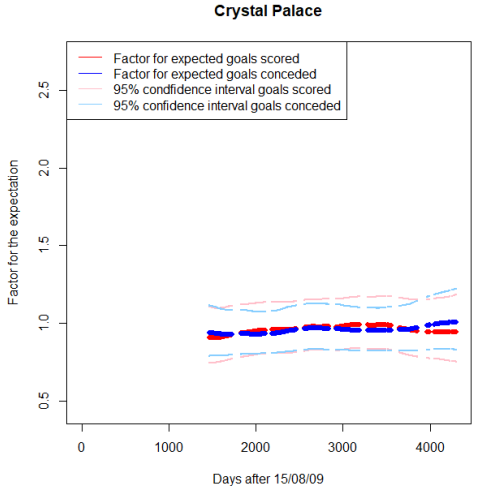
(c)



(d)

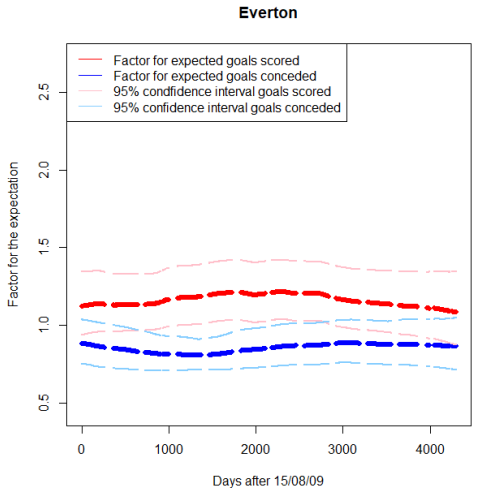


(e)

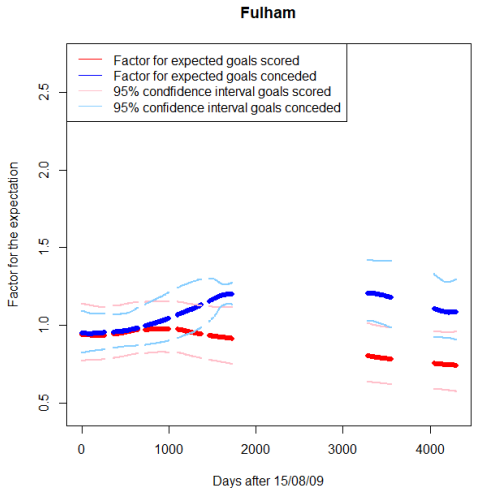


(f)

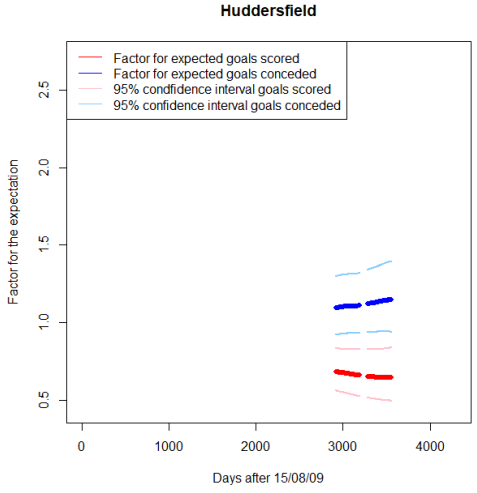
Figure 27



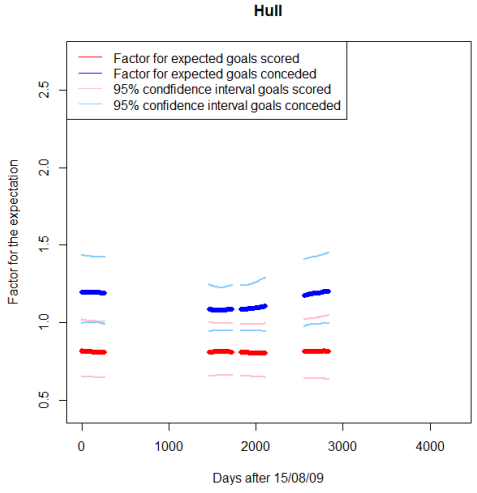
(a)



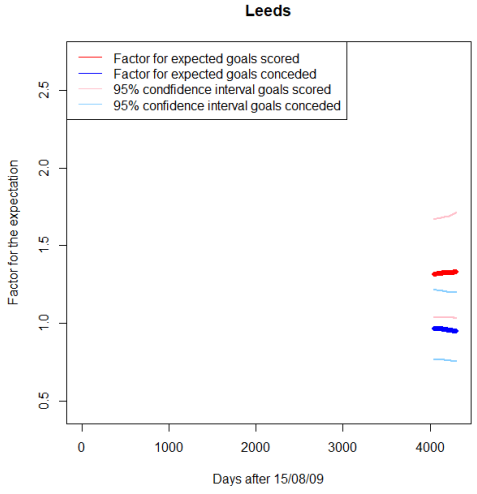
(b)



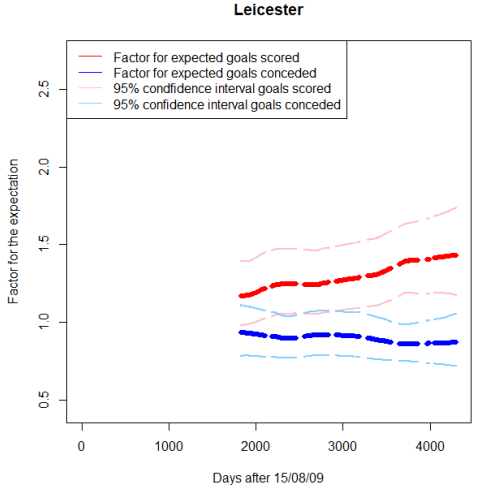
(c)



(d)

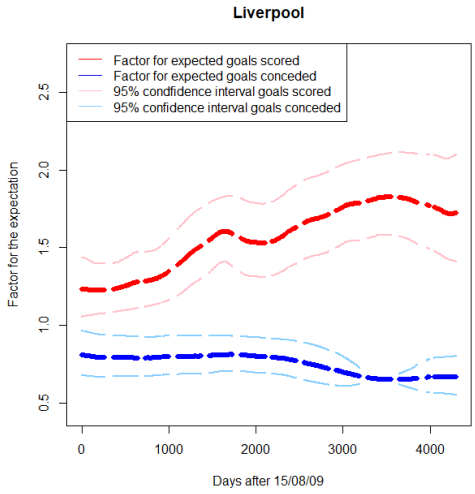


(e)

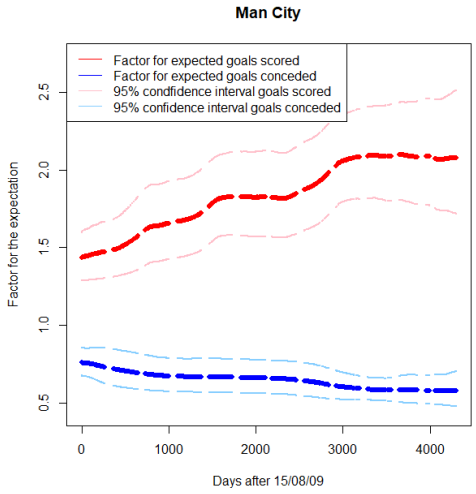


(f)

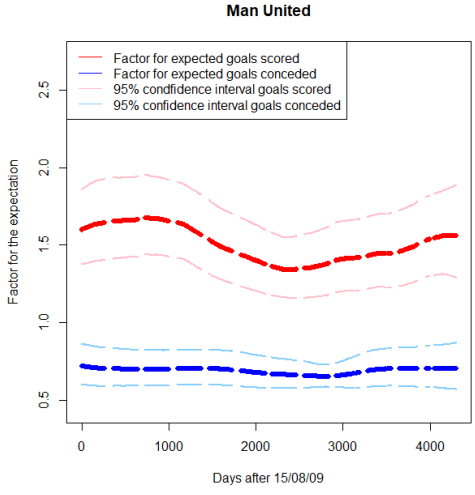
Figure 28



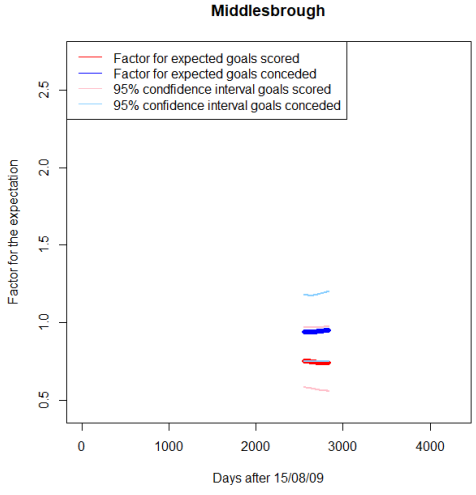
(a)



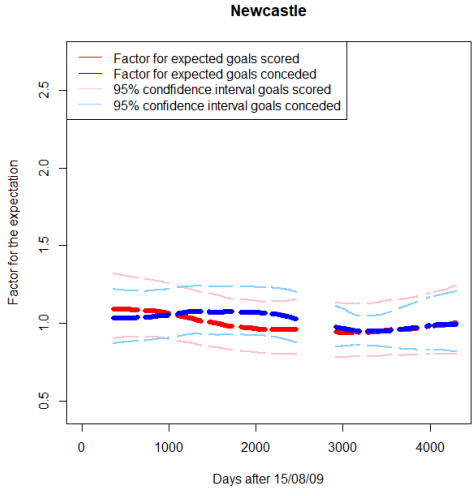
(b)



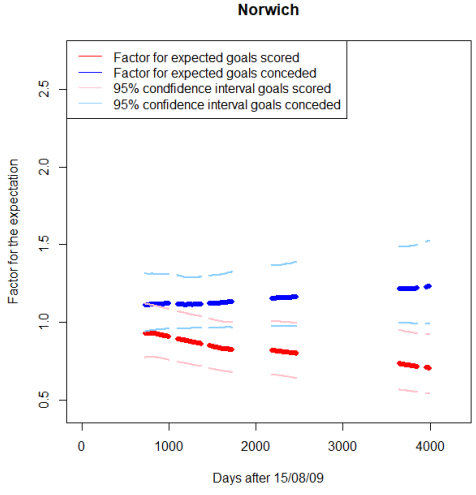
(c)



(d)

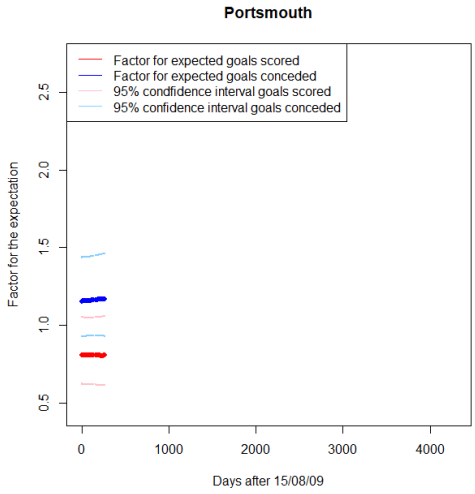


(e)

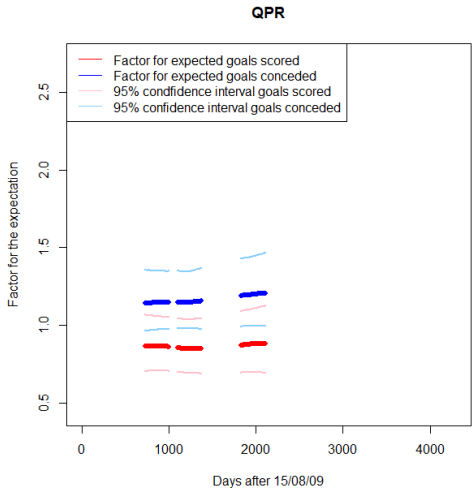


(f)

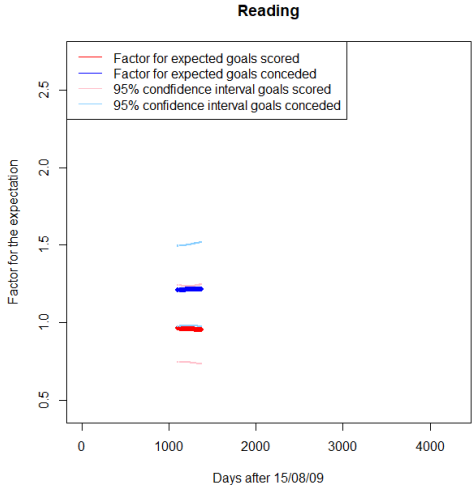
Figure 29



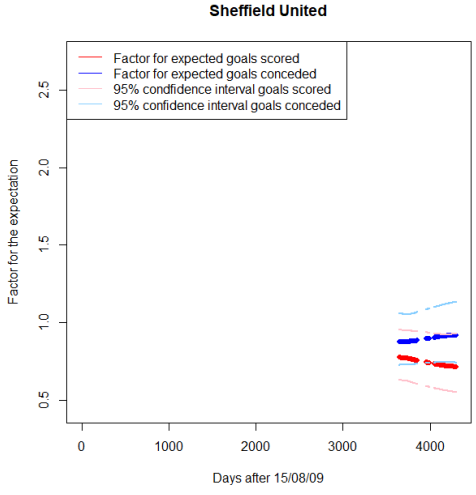
(a)



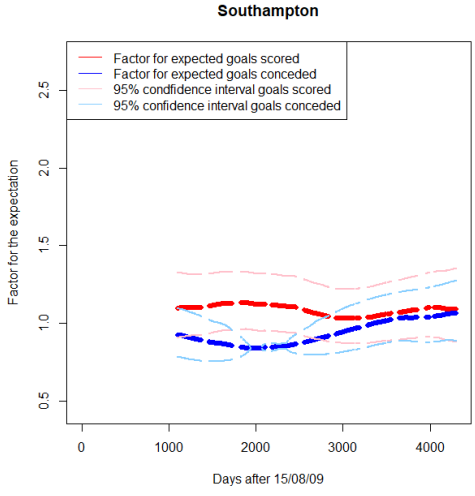
(b)



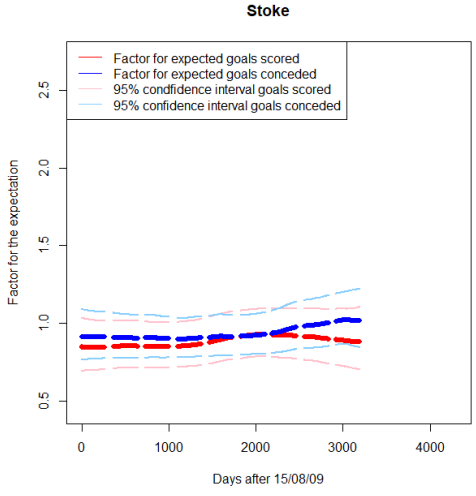
(c)



(d)

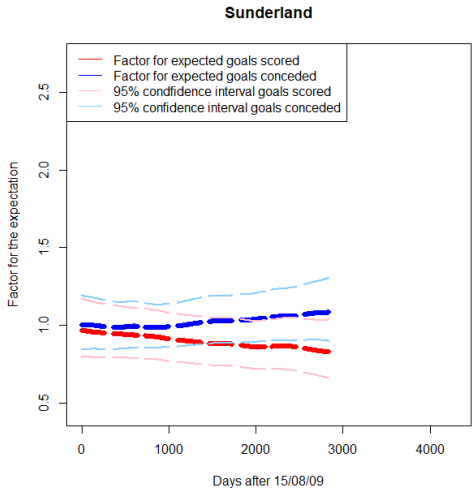


(e)

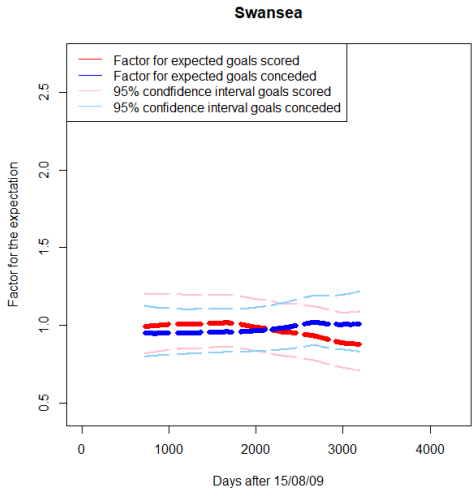


(f)

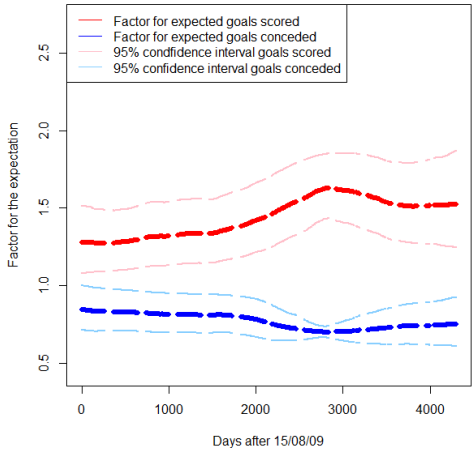
Figure 30



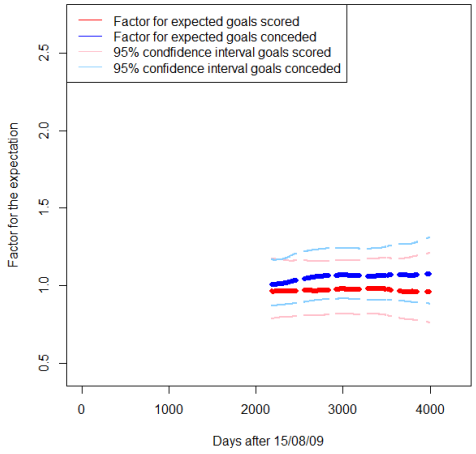
(a)



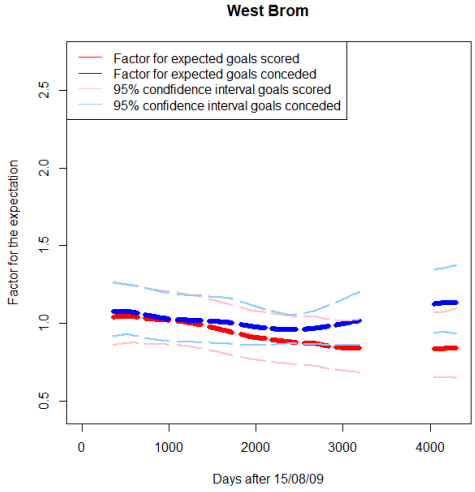
(b)



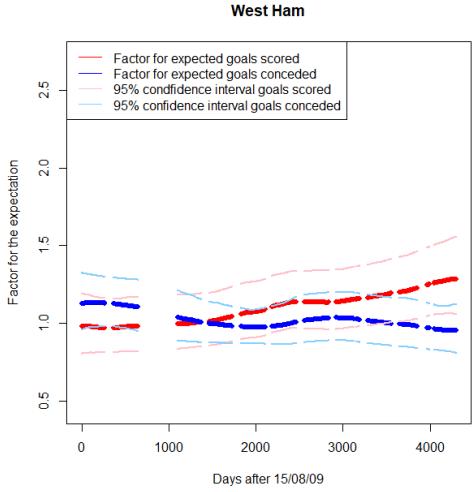
(c)



(d)

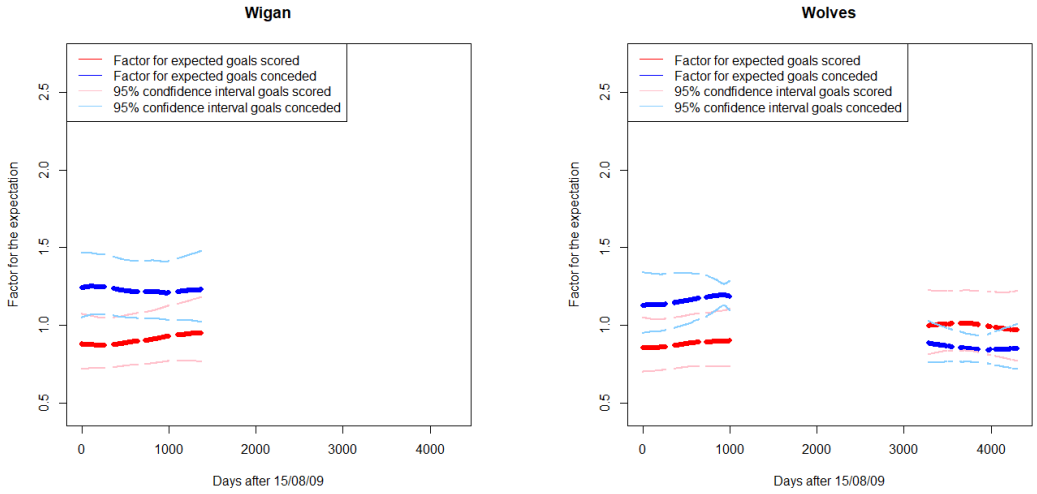


(e)



(f)

Figure 31



(a) (b)

Figure 32: A plot of the attacking and defending strength for all the teams, in all the 12 seasons of the Premier League. The light colors are 95% confidence intervals. A strong team will have high values for the attacking strength (red line in the plot) and low values for the inverse of the defending strength (blue lines in the plot).

5 Discussion

5.1 Home field advantage

Due to the ongoing pandemic, caused by SARS-Cov-2, there has been a lockdown in most countries. The English Premier League has still been played, but without spectators. The past year will likely be the best opportunity to test the effect supporters have on the outcome of the match, it is unlikely that the supporter stands will be empty again during matches. It does seem that the home field advantage was present and significant before the pandemic, with teams scoring about 30% more goals at home compared to what they do away, everything else being constant. According to the estimated models the home field advantage has disappeared during the pandemic. The only real change in football during the pandemic has been the empty stands. The disappearance of the home field advantage indicates that the supporters were the reason for the home field advantage. It does seem like the home field advantage was due to a combination of the home teams scoring more, and the away team scoring less, than they would have done without the supporters. I think that the home field advantage will come back next season when the supporters return, and if that happens it will be a clear indication that the supporters are the main driving force for the home field advantage. However, this can of course not be tested at the time of writing.

5.2 Best model

The models proposed in this project can easily be divided into categories. An obvious categorical split is time-dependent and non-time dependent models. We see in general that the time dependent models performs better than the constant models. But that the estimates for the strengths does not vary much, especially for the model where τ is estimated in the likelihood. The time dependent models use a larger data set for their estimation than the constant models, and we see that especially in the later seasons the time-dependent models earn more money than the constant ones.

Another possible division of the models is into direct and indirect models. The multinomial model tries to estimated the probabilities directly using a generalized Bradley-Terry model. While all the other models estimates something other than the probabilities, and then infer the probability distribution of the results from there. The different Poisson and generalized Poisson models try to model the number of goals scored by both teams, while the data intensive model, models the number of chances to each team and assumes a known relationship between the number of chances and the number of goals dependent on both teams. Of the constant models the Poisson model was the one that did best on the betting market in total, it gave a small profit over a twelve year period. We do not need to be impressed with the results, from a financial point of view. The generalized Poisson model did better on the likelihood test of section 4.8, than did the normal Poisson model. This could indicate that the Generalized Poisson model is a better fit than the ordinary Poisson model, even though it did not earn more on the betting market, as we saw in table 1. The fact that it did not do better at the betting market, even though it was a better fit to the probabilities, could indicate that the betting sites have given too high odds for unlikely outcomes and too low odds for likely outcomes. This could be the result of the fact that many people are betting on what they think will be the outcome of the match, without considering the odds.

It has earlier been discussed that the Poisson distribution is a poor fit for the score in football matches, because over-dispersion has been suspected. This is for example the case in Greenhough et al. [2002]. They found that the chances of high scoring games was not sufficiently covered by the Poisson distribution, and they proposed the use of the negative binomial distribution instead. They only used one parameter for the home team and one for the away team, and the estimate of the number of goals scored was independent of what teams was playing. However, in this project I found the number of goals score in football matches to be under-dispersed, and proposed the use of the Generalized Poisson model, to model the score in football matches. Under-dispersion means that there is less variance than with a normal Poisson model, and one can think of the results one gets with the Generalized Poisson model as more precise than the ones with the Poisson model. But sadly the model does not do better at the betting market. Its predictions does however fit the data better, indicating that the data is under-dispersed. This has not been found in football

results before, as I know of. Possible reasons for under-dispersion could be a negative covariance between non-overlapping time segments in the match. This would mean that if a team scores early it is less likely to score later, and if it does not score early it is more likely to score later. This fits well with the common understanding of football where a team will try harder to attack and score goals if it has no goals, than if they are currently leading and just run out the clock. The variance was found to be 3.3% smaller than for normal Poisson model, this is not a large reduction. And the Poisson distribution is a good approximation for the distribution of number of goals scored by teams in the Premier League.

5.3 Making money

This project uses more seasons than they do in any of the articles mentioned in this paper. Some of the articles only look at one or two season, and this is often the same seasons that most models earn money. For example we saw in section 4.7 that almost any model would have earned money in the 11/12 season of the Premier League. The exception to this rule is Rue and Salvesen. They looked at 5 seasons in both the Premier League and Championship. Their model consistently earned money in all five seasons [Rue and Salvesen, 2000]. However, the data set they used was from the middle to late 90's, and it is reasonable to assume that the betting sites are now better at estimating probabilities as well. When earning money at the betting market one needs an edge over the bookie, and in an ever increasing arms race the models of yesterday will not be good enough today. Over the 12 seasons only a few of the models managed to earn money in total, but just 3.67% in eleven years. The predictions made with the time dependent Poisson models was also the ones that best fitted the observations. This is indicating that the reason it earned more money, was that it was better at predicting matches.

Of the betting strategies used in this project, the one introduced by Rue and Salvesen performed best. But the Kelly criterion performed very well as well. It could be of interest to look at ways to use it on many bets simultaneously, then it could be used as a betting strategy where one needs to place 10 bets at the same time. One alternative could be to say that each bet can maximum place 10% of the total wealth, in that case one would never place more than 100% of the money, and probably much less.

5.4 Further works

We saw from the results that the models using a Brownian motion, worked better than the models using constant parameters, for the strengths of the teams. The Browinan motion, used to model the change in the strength of each team over time, is built on some assumptions. For instance it assumes that the variance of the change for each team is the same, and only dependent on the time difference between the observations. It also assumes that the expected change is zero, independent of possible trends. One alternative to an extension to the model, is to include trends in the time model for the strengths of the teams. This could for example be done as an AR model. Another alternative could be to let the variance in the change be dependent on other factors such as manager change, new players, or injuries. One could also assume that the change can have an expectation different from zero in some matches. For example if a key attacking player became injured in the last game the expected change in attacking strength could be slightly negative. Some of these changes are possibly difficult to model, and one would need a larger data-set that had all this information.

One factor that is known to be of importance, that have not been included in this project, is how frequent the matches have been played. The number of days between each match for the different teams have been included in the Browian motion, but only to calculate the variance. The players get tiered during the season and the teams have to be rotated when there are many games in a short time period. If seven of the eleven players are changed from one game to the next, the variance in the performance could be higher that the model would assume. A model that would try to include this, would need a data set that included all the possible tournaments for the teams in the Premier League. Some of the teams in the Premier League play over 60 matches in a season, and only 38 of the matches is actually in the Premier League. The rest is from domestic cups and the European cups (Champions League and Europa League).

Not every match is as important for both teams, this is especially noticeable in the last rounds of the seasons. At the end of the season some teams might play to remain in the league while others are secure in the middle of the table, with nothing to play for. For example the champions of this season, Manchester City, lost 3-2 to Brighton in the first match after securing the title. After winning the league Manchester City did not need to win the match, this resulted in them resting some of their best players. This is just an example, but it could be of interest to include the importance of the match for both teams, in the model in some way.

6 Conclusion

In this project different models have been utilized to predict the results of football matches, and the accuracy of the predictions have been tested on the betting market. The models were based on the assumption that the results of football matches could be predicted by looking at the difference in strength for the two teams, and what team plays at home. The strengths of the teams and the home field advantage, could be estimated based on previous results. The home field advantage was significant before the pandemic, but estimated to zero during the pandemic. Of the different models discussed, I found that the Poisson model made better predictions for future football matches than the generalized Bradly-Terry model did. This is because there is more information about the strengths of the teams in the score, than in just the result. By using the Poisson GLMM model one could also model both the attacking and defending strengths for each team separately, instead of just one total strength. The attacking and defending strength for each team was usually highly correlated, but not for every team. Some teams score many goals and concede many goals, these observations fit well with the common understanding of football teams. The teams strengths stayed almost constant, also when the strength could change with time, as a Brownian motion. However, letting the strength be time dependent made better predictions, and the models using time-dependent strengths, earned more money on the betting market. But even the time dependent models did not earn enough money over the time period for it to be worth the risk. In total the best model earned 3.67% over an 11 year period. This means that if you bet NOK100 every round for the last 19 rounds each season for 11 years, your total earnings would be NOK767. This is less than one would earn with interest in most banks. The model earns a small profit in total, but there is a large variance in the profit for each season. For this reason anyone that wants to test different models and different betting strategies on football results, should make predictions for many seasons. This is because a model can earn a lot of money in some seasons, but still loose money in aggregate over a longer time period.

Despite many critics to the Poisson model, I found it to fit the data well. It has been taught that the Poisson distribution would underestimate the variance in the data, and that the negative-binomial distribution would be a better fit. In this project I found the opposite to be the case. I found a small under-dispersion, and proposed the use of the Generalized Poisson distribution instead. The Generalized Poisson distribution did better on predicting the probabilities for different outcomes in the football matches, than the ordinary Poisson model did. This indicates that the reason for the under-dispersion was not over-fitting. However, the Generalized Poisson model did not outperform the ordinary Poisson model on the betting market in aggregate.

References

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018. URL http://jmlr.org/papers/v18/17-468.html.

Richard Berk and John M MacDonald. Overdispersion and poisson regression. *Journal of Quantitative Criminology*, 24(3):269–284, 2008.

Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology Evolution*, 24(3):127–135, 2009. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2008.10.008. URL https://www.sciencedirect.com/science/article/pii/S0169534709000196.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

Manuela Cattelan, Cristiano Varin, and David Firth. Dynamic bradley–terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):135–150, 2013. doi: https://doi.org/10.1111/j.1467-9876.2012.01046.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2012.01046.x.

P. C. Consul. On the differences of two generalized poisson variates. *Communications in Statistics - Simulation and Computation*, 15(3):761–767, 1986. doi: 10.1080/03610918608812538. URL https://doi.org/10.1080/03610918608812538.

Mattan S. Ben-Shachar Indrajeet Patil Philip Waggoner Brenton M. Wiernik. Daniel Lüdecke, Dominique Makowski. Check overdispersion of gl(m)m's. URL https://easystats.github.io/performance/reference/check_overdispersion.html.

Mark J. Dixon and Stuart G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280, 1997. doi: https://doi.org/10.1111/1467-9876.00065. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00065.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regrssion*. Springer, 2013a.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regrssion*. Springer, 2013b.

J Greenhough, P.C Birch, S.C Chapman, and G Rowlands. Football goal distributions and extremal statistics. *Physica A: Statistical Mechanics and its Applications*, 316(1):615–624, 2002. ISSN 0378-4371. doi: https://doi.org/10.1016/S0378-4371(02)01030-0. URL https://www.sciencedirect.com/science/article/pii/S0378437102010300.

Robin K.S. Hankin. A generalization of the bradley–terry model for draws in chess with an application to collusion. *Journal of Economic Behavior Organization*, 180:325–333, 2020. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2020.10.015. URL https://www.sciencedirect.com/science/article/pii/S0167268120303838.

Tammy Harris, Zhao Yang, and James W. Hardin. Modeling underdispersed count data with generalized poisson regression. *The Stata Journal*, 12(4):736–747, 2012. doi: 10.1177/1536867X1201200412. URL https://doi.org/10.1177/1536867X1201200412.

J. L. Kelly. A new interpretation of information rate. *he bell system technical journal,*, pages 917–926, 2007. doi: https://doi.org/10.1111/j.1467-9876.2007.00594.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2007.00594.x.

K. Kristensen, A. Nielsen C. W. Berg, H. Skaug, and B. M. Bell. Tmb: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(5), 2016.

Helge Langseth. Beating the bookie: A look at statistical models for prediction of football matches. *Frontiers in Artificial Intelligence and Applications*, 257:165–174, 01 2013. doi: 10.3233/978-1-61499-330-8-165.

Subhash R. Lele, Khurram Nadeem, and Byron Schmuland. Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105 (492):1617–1625, 2010. doi: 10.1198/jasa.2010.tm09757. URL https://doi.org/10.1198/jasa.2010.tm09757.

M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982. doi: https://doi.org/10.1111/j.1467-9574.1982.tb00782.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1982.tb00782.x.

Havard Rue and Oyvind Salvesen. Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):399–418, 2000. doi: https://doi.org/10.1111/1467-9884.00243. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00243.

J. G. Skellam. The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society*, 109(3):296–296, 1946. ISSN 09528385. URL http://www.jstor.org/stable/2981372.

Jarle Tufto, Erling Johan Solberg, and Thor-Harald Ringsby. Statistical models of transitive and intransitive dominance structures. *Animal Behaviour*, 55(6):1489–1498, 1998. ISSN 0003-3472. doi: https://doi.org/10.1006/anbe.1998.0755. URL https://www.sciencedirect.com/science/article/pii/S0003347298907552.

Chris Whitrow. Algorithms for optimal allocation of bets on many simultaneous events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(5):607–623, 2007. doi: https://doi.org/10.1111/j.1467-9876.2007.00594.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2007.00594.x.

A Appendix

A.1 Proof

A.1.1

If $X \sim \text{Pois}(\lambda)$ and $Y|X \sim \text{binom}(x, p)$. Then $X \sim \text{Pois}(p\lambda)$.

$$f(y) = \sum_{x=y}^{\infty} f(x, y) = \frac{e^{-\lambda}}{y!}(\frac{p}{1-p})^y \sum_{x=y}^{\infty} \frac{((1-p)\lambda)^x}{(x-y)!}$$

if we introduce $k = x - y$

$$f(y) = \frac{e^{-\lambda}}{y!}(p\lambda)^y \sum_{x=y}^{\infty} \frac{((1-p)\lambda)^k}{k!}.$$

We recognise the sum as the Taylor-expansion of $e^{(1-p)\lambda}$. This gives

$$f(y) = \frac{(p\lambda)^y e^{-p\lambda}}{y!}$$

This is the density of a Poisson distributed variable with $p\lambda$ as mean. I.e $Y \sim \text{Pois}(p\lambda)$.

A.1.2 Proof binomial in chain

I will show that if $X \sim \text{binom}(n, p)$ and $Y|X \sim \text{binom}(x, q)$, then $Y \sim \text{binom}(n, pq)$. Without loss of generality I will show that this is the case for two variables, but the chain can also have length m.

$$f(x, y) = f(x)f(y|x) = \binom{n}{x} p^x (1-p)^{n-x} \binom{x}{y} q^y (1-q)^{x-y}$$

$$f(y) = \sum_{x=y}^{n} f(x, y)$$

$$f(y) = \frac{n!}{(n-y)!y!}(pq)^y (1-pq)^{n-y} \sum_{x=y}^{n} \frac{(n-y)!}{(n-x)!(x-y)!} p^x (1-p)^{n-x} (1-q)^{x-y} (1-pq)^{y-n}$$

We need the sum to become 1 to show that $Y \sim \text{binom}(n, pq)$. If we introduce $z = x - y$ and $m = n - y$ we get

$$f(y) = \binom{n}{y} (pq)^y (1-pq)^{n-y} \sum_{z=0}^{m} \binom{m}{z} (\frac{p(1-q)}{1-p})^z (\frac{1-p}{1-pq})^m$$

The inside of the sum is an alternative formulation for the Binomial distribution. If $z \sim \text{binom}(n, p)$ then $f(z) = \omega^z \frac{1}{(1+\omega)^n}$, with $\omega = p/(1-p)$ equal to the odds. And since the sum goes over all possible values for z it is equal to 1. We therefor get

$$f(y) = \binom{n}{y} (pq)^y (1-pq)^{n-y},$$

and $Y \sim \text{binom}(n, pq)$.