



NTNU – Trondheim
Norwegian University of
Science and Technology

A Study on Soccer Prediction using Goals and Shots on Target

Snorre Gebhardt Stenerud

Master of Science in Physics and Mathematics

Submission date: June 2015

Supervisor: Håvard Rue, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

A Study on Soccer Prediction using Goals and Shots on Target

Snorre Stenerud

June 28, 2015

Abstract

In this thesis I have developed a model for result prediction in soccer. The model is based on chances created being modeled as a Poisson process while goals scored is seen as a result of first creating chances and then converting them, here modeled as a Bernoulli trial. Compared to existing models, this one takes advantage of a number of data that previously have not been considered. As each team is described by four parameters, teams are able to be distinguished further allowing for better prediction of chances created and goals scored for each team in a given matchup.

Six different models are developed gradually with the goal of improving the model fit to data and its predictive ability. In the final model the parameters can change over time so as to explain how a team can go through periods of good or bad form. Parameters are assumed to be correlated to each other - reflecting how a good offensive team also often has a good defense. Red cards are included to explain why certain surprising results took place. And lastly, the model uses shots on target to predict goals, as this is shown to have a stronger connection than between shots and goals.

A betting strategy is implemented where the size of the bet decreases with increasing odds, while only placing bets when the expected return is above a certain value. The model struggles with consistency, but is still able to make a small profit over a run of five seasons, so it should be an excellent candidate for further development.

Sammendrag

I denne oppgaven har jeg utviklet en prediksjonsmodell for fotball. Modellen baserer seg på at målsjanser kan antas å følge en Poisson fordeling, og at hver målsjanse har en viss sannsynlighet for å gå inn, altså et Bernoulli-forsøk. Sammenlignet med eksisterende modeller har jeg valgt å inkludere mer interessante data som tidligere ikke har blitt benyttet. Ettersom hvert lag beskrives av fire parametere blir lagene beskrevet svært forskjellige i modellen, og dette tillater bedre prediksjon av målsjanser og mål for hvert lag i en gitt kamp.

Seks forskjellige modeller er gradvis utviklet med mål om forbedret tilpasning til data og økt treffsikkerhet i prediksjon av fremtidig kampers resultat. I den ferdig utviklede modellen lar jeg parametrene kunne endres over tid for å reflektere hvordan lag kan ha svingninger i form. Parametrene antas korrelerte - på samme måte som et godt offensivt lag også ofte har et godt forsvar. Røde kort er inkludert som en forklaringsvariabel ettersom det kan hjelpe modellen å forstå kamper der f.eks. gode lag har slitt mot dårlige lag. Og modellen benytter skudd på mål for å forutse mål, ettersom det har vist seg at skudd på mål har en sterkere tilknytning til mål enn kun skudd.

En bettingstrategi er implementert hvor innsatsen minker ved økende odds, mens man bare setter penger på de kampene der den forventede gevinsten er over en viss verdi. Modellen sliter med å konsekvent slå bookmakerene, men klarer å produsere en liten profitt gjennom de fem sesongene den ble testet på, så den virker som en utmerket kandidat for videre utvikling.

Acknowledgment

I would like to express my greatest gratitude to my advisor and supervisor Håvard Rue for his invaluable help and support throughout this project.

Contents

1	Introduction	11
2	Literature Review	13
2.1	Other studies of interest	18
3	Presentation of data	21
4	Quality Assessment of the models	27
4.1	Likelihood	27
4.2	Deviance Information Criterion (DIC)	27
4.3	WAIC - Watanabe Akaike Information Criterion	28
4.4	Second Half Pseudo-Likelihood	28
5	Designing a model for prediction	31
5.1	Model 1: Chances Poisson distributed and constant p of conversion	32
5.1.1	Priors	32
5.1.2	Performance with data	33
5.2	Model 2: Unique probability of conversion for each team	34
5.2.1	Priors	34
5.2.2	Performance with data	34
5.2.3	Validity of Model 2 over Model 1	34
5.3	Model 3: Probability of conversion depends on the opposition	36
5.3.1	Priors	36
5.3.2	Performance with data	36
5.3.3	Validity of Model 3 over Model 2	36
5.3.4	Correlation between parameters	40
5.3.5	Parameters change over time	42
5.4	Model 4: Team-specific strengths change over time	43
5.4.1	Priors	43
5.4.2	Performance with data	43
5.4.3	Validity of model 4	44
5.5	Model 5: Correlation between parameters	46
5.5.1	Performance with data	46
5.6	Model 6: Effect of red cards included	47
5.6.1	Performance with data	47

6	The prediction model	51
6.1	Chance model	51
6.2	Goal model	51
6.3	Parameter Properties	52
6.4	Chances - Shot or SOT?	52
7	Prediction model vs betting companies	53
7.1	The betting model	54
7.2	Results from betting	54
8	Results and discussion	59
8.1	Results	59
8.2	How to improve the model	62
8.2.1	Initial value for teams	62
8.2.2	Home Field Advantage for goal converting or goal preventing	62
8.2.3	Different correlations between parameters	62
8.2.4	Including non-quantitative information	62
8.2.5	Other risk strategies for betting	63
9	Conclusion	65
	Appendix A INLA	69
	Appendix B R Code	71

Nomenclature

α_0	Goal model offset
α_A	Goal scoring strength of team A
β_A	Goal preventing strength of team A
$\hat{\alpha}_0$	Chance model offset
$\hat{\alpha}_A$	Chance creating strength for team A
$\hat{\beta}_A$	Chance preventing strength for team A
$\hat{\delta}$	Home field advantage for chances
$\hat{\lambda}_{i,j}$	The estimated mean Chances Created by home team i
$\hat{\mu}_{i,j}$	The estimated mean Chances created by away team j
$\hat{\theta}_\alpha$	Impact of a red card on $\hat{\alpha}$
$\hat{\theta}_\beta$	Impact of a red card on $\hat{\beta}$
$\hat{X}_{i,j}$	Chances Created by team i on team j
$\hat{Y}_{i,j}$	Chances Created by team j on team i
ω	Betting cutoff - minimum margin to place a bet
ρ	Parameter Correlation
ρ_t	Time correlation
τ_t	Precision of ρ_t
θ_α	Impact of a red card on α
θ_β	Impact of a red card on β
i, j	Generic team names where i always is the home team and j always is the away team
p	Probability of conversion for home team
q	Probability of conversion for away team
r	Red cards received
$X_{i,j}$	Goals Scored by team i on team j
$Y_{i,j}$	Goals Scored by team j on team i

As for general terms used throughout the paper, a **chance** refers to either a shot or a shot on target (**SOT**). A teams **chance creating strength** is its ability to produce chances in a game, while its **chance preventing strength** is its ability to keep the opposing team from producing chances. A teams **goal converting strength** is its ability to turn chances into goals, while its **goal preventing strength** is its ability to keep the opposing team from converting chances into goals.

Chapter 1

Introduction

Association football (from now on referred to as soccer) is regarded as the biggest and most popular sport in the world. In the English Premier League (EPL), the highest level of soccer in England, 20 teams from England and Wales play a total of 380 games over a season. All teams play each other twice so that each team gets one game at home and one away, essentially all the possible permutations of teams not including repetition. A game gives three points to the winning team and zero to the losing team, or, in the case of a draw, one point to each team.

When all the matches have been played the points are counted and the teams are ranked based on the amount of points gathered over the season. The one with the most points is the years winner of the EPL. The top three get a direct qualification to the UEFA Champions League, and the team placing 4th has an opportunity to qualify through competing with other lower ranked teams from around Europe. Likewise, the team placing 5th gets a direct qualification to compete in the UEFA Europa League, whereas the teams placing 6th and 7th get a chance to qualify through competing with other European lower ranked teams. Twenty-five percent of the domestic broadcasting revenue is divided on a merit basis, meaning the higher ranked teams get more of the money. In the 2013-2014 season this amounted to about 25 million pounds for league winners Manchester City and about one million pounds for last placing Cardiff City, in addition to the 50 percent of the total revenue being spread equally and 25 percent spread based on matches broadcast in the UK. [1]

Existing models for forecasting typically attempt to describe a team using two explanatory variables (hereby referred to as the strengths of a team) - one for the offensive strength and one for the defensive, and as a team faces off against another team their offensive power is put to test against the opponents defensive power and vice versa. This difference in strength, along with a few constants like the home field advantage and the effect of underestimating the opponent, give us the expected goals scored for each team in a match. While some have achieved impressive results, very few fans of soccer would agree that a team composed of 11 different players with unique abilities, playing in their own formation and their own style with a tactic planned out for that specific match, could all be described by two parameters. A goalkeeper could be good at stopping long range efforts, but that wont help if the opponent excels at playing their way into the penalty area. A forward might be a clinical finisher, but if the defense is able to completely isolate him he won't get a single shot in. With this in mind, I will create a more complex model that can handle teams excelling in different aspects of attacking and defending, while taking advantage of the availability of interesting data like shots fired and red cards.

The approach used in this paper draws a lot of inspiration from the ones designed by Koopman, Lit (2015)[10], Rue, Salvesen (2000)[18] etc, which again are extensions of the model first created by Maher (1982)[12]. The biggest changes are that I am increasing the number of team-describing variables from two to four, provided that they make a significant impact. I have a full 14 seasons of interesting data that I want to exploit as well as possible, including looking at shots, shots on target and red cards. I want to model a teams offensive capabilities through their ability to create chances and later through their ability to convert chances rather than modeling goals scored directly. I'm taking advantage of recent advancements in statistical inference in the form of Integrated Nested Laplace Approximations (INLA), which allow for much faster approximations than Markov Chain Monte Carlo methods (MCMC).

The remainder of this paper will be structured as follows: Chapter 2 discusses the existing literature and the results achieved by others in the field of forecasting match results in soccer. Chapter 3 gives a description of the various types of data recorded during a soccer match, looking at their applications and relevance. Chapter 4 describes methods for comparing results produced by different models. Chapter 5 starts the process of developing an efficient model for predicting goals, beginning at a simple model where the probability of a chance becoming a goal is considered constant across all teams and matches. It then goes on to include time dependency, correlation between parameters and using red cards. Chapter 6 fully describes the chosen model, while discussing the advantages of using "shots on target" instead of shots to do prediction. Chapter 7 describes a basic betting strategy, and evaluates how well the model fares against the betting companies. Chapter 8 includes a rundown of the results collected throughout the paper and a discussion regarding further improvements to the model. Finally, Chapter 9 concludes the paper with an evaluation of the project as a whole along with a quick summary.

Chapter 2

Literature Review

Moroney (1956) - Facts From Figures [13] is the first record of statistical models being used to model results in professional soccer. Moroney shows that the Poisson distribution is unsuitable for describing goals scored, and that the Negative Binomial distribution is a much better fit. This is done by looking at the match result as a variable not dependent on which teams are playing, and the implication is that in the model every team is of the same strength. **Reep, Pollard, Benjamin (1971) - Skill and Chance in Ball Games**[16] verify these results and come to the conclusion that "chance dominates the game". **Hill(1974) - Association Football and Statistical Inference**[9] believes it is obvious simply from watching a game of soccer that both chance and skill have impact on the result, but that in the long run skill will be dominant. Hill attempts to prove this by comparing the predictions of the final tables by experts to the actual final tables for 4 divisions of the 1971-1972 Football League. He shows that there is at least a positive correlation between the two, and claims thereby to have debunked the theory that soccer is dominated by chance.

Maher (1982) - Modelling association football scores [12] shows that by giving each team an attacking strength α and a defensive strength β he is able to model the goals scored as a Poisson response variable with mean equal to the relative strength between the teams. He points out that assuming all teams to be equal when they are not would give a Poisson distribution with variable mean which indeed would look a lot like the Negative Binomial Distribution. Maher finds that the advantage of playing at home ground δ is significant and constant across all teams. If goals scored by home team i , x , against team j is generated by a Poisson-distributed random variable $X_{i,j}$ (and likewise for $Y_{i,j}$ for the away team, $X_{i,j}$ and $Y_{i,j}$ being independent), then $Pr[X_{i,j} = x, Y_{i,j} = y] = \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^y e^{-\mu}}{y!}$, where $\lambda = \alpha_i \beta_j \delta$ and $\mu = \alpha_j \beta_i$. Maher notes that the model underestimates scores where a team has one or two goals, while overestimates scores where a team has zero or more than three goals. A better distribution would be one that is slightly "narrower" than the Poisson distribution. To address this a bivariate Poisson model is tested and the correlation constant is estimated to be around 0.2.

Dixon, Robinson (1998) - A birth Process Model for Association Football Matches [3] looks at the rate of goals scored over the course of a match. This is done by modeling goals scored by both teams as interactive birth processes. An increasing number of goals are scored throughout the

90 minutes. This could be because the scoring rates gradually increase, or because the scoring rates are dependent on the current score.

The final score is modeled the same way as done by Maher(1982)[12] where the number of goals scored by the teams in a game are dependent Poisson variables determined by the strength of the attack and defense of the two sides. The actual goal scoring is modeled as a two dimensional birth process where home and away scores are different species. H_k and A_k for home and away goal processes at match k are modeled with $\lambda_k(t)$ and $\mu_k(t)$ respectively that are allowed to vary in time, t. This is simplified to $\lambda_k(t) = \lambda_{xy}\lambda_k$, where λ_{xy} holds the current score (x-y), $\lambda_k = \alpha_{i(k)}\beta_{i(k)}\delta$ and $\mu_k = \alpha_{j(k)}\beta_{i(k)}$. α and β are strengths in offense and defense, i and j refer to the home team and away team and δ is the home advantage factor. The intensities for minutes 45 and 90 are handled separately as added time is pushed onto these themselves.

Dixon and Robinson conclude that scoring rates generally increase for both teams throughout the match, they depend on the current score, and they generally increase when a goal is scored.

Lee 1997 - Modeling Scores in the Premier League: Is Manchester United Really the Best? [11] employs what is essentially the exact same model as used by Maher (1982)[12], but instead of merely looking at the goodness of fit, he goes further to simulate the season 1000 times to see what team actually deserved to win the 95/96 season in the Premier League. A problem with the approach is that the same estimated strengths of attack and defense are used for every simulated season, which doesn't consider the fact that the observed results were only one possible outcome.

Dixon, Coles (1997) - Modelling Association Football Scores and Inefficiencies in the Football Betting Market [4] build on the model proposed by Maher(1982)[12], but make certain improvements. Instead of continuing with the standard bivariate Poisson model, they make a direct modification to the joint probability distribution -

$$Pr[X_{i,j} = x, Y_{i,j} = y] = \tau_{\lambda,\mu}(x, y) \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^y e^{-\mu}}{y!} \quad (2.1)$$

$$\text{where the new term } \tau_{\lambda,\mu}(x, y) = \begin{cases} 1 - \lambda\mu\rho & \text{if } x = y = 0, \\ 1 + \lambda\rho & \text{if } x = 0, y = 1, \\ 1 + \mu\rho & \text{if } x = 1, y = 0, \\ 1 - \rho & \text{if } x = y = 1, \\ 1 & \text{otherwise.} \end{cases}$$

ρ here is a dependence parameter with $\max(-1/\lambda, -1/\mu) \leq \rho \leq \min(1/\rho\mu, 1)$ and $\rho = 0$ would mean total independence.

An even bigger improvement is the move from a static model with all team strengths being constant to a dynamic model that allows the parameters to vary in time. They let the importance of each game be decided by the weighting function $\phi(t) = e^{-\xi t}$ such that old results are given exponentially less weight than recent ones. $\xi = 0$ represents a static model, and when used to maximize chance of predicting correct results (home, draw, win) $\xi = 0.0065$ is estimated to be the optimal. The model is tested on its ability to consistently beat the bookmakers with a simple betting strategy - put money on

all bets that have an expected return above a certain level, and it is deemed to be adequate. Among possible improvements they mention the need for a Bayesian structure to incorporate additional covariate information and a potentially profitable betting strategy based on exact match results instead of just match outcomes.

Rue, Salvesen (2000) Prediction and Retrospective Analysis of Soccer Matches in a League [18] use a Bayesian model for calculating the time-variation of all strengths simultaneously. While the basis of the model resembles the ones used by both Maher (1982)[12] and Dixon, Coles(1997)[4], they make several changes to the underlying assumptions. A superior team will tend to underestimate an inferior team, and conversely the inferior team will be more prepared against a superior team. This is entered into the model by measuring the overall difference in strength between the two teams $\Delta_{i,j} = (\alpha_i + \beta_i - \alpha_j - \beta_j)/2$. The amount of goals scored by i against j, $x_{i,j}$ is connected to $\alpha_i - \beta_j - \theta_0 \Delta_{i,j}$, where θ_0 is a small constant $\theta_0 > 0$ giving the magnitude of the psychological effect. $\theta_0 = 0$ would mean that the psychological effect has no impact on the match.

Rue and Salvesen build on the modifications to the joint conditional law made by Dixon, Coles (1997) in two ways. They make all goals scored by a team beyond five be counted as five, as they consider such scores to be demotivating to an extent where the underlying assumptions of goals scored being independent no longer hold up. A win of 7-5 will therefore be handled as a 5-5 draw. π_g^* is the resulting truncated law. They also infer that the match results are not as informative on the strengths of teams as was assumed by Dixon and Coles, and they introduce a parameter ϵ with the interpretation that only $(1 - \epsilon) \times 100\%$ of the information in a match result is informative. The goal model then changes to

$$\begin{aligned} \pi_g(x_{i,j}, y_{i,j} | \lambda_{i,j}, \mu_{i,j}) &= (1 - \epsilon) \pi_g^*(x_{i,j}, y_{i,j} | \lambda_{i,j}, \mu_{i,j}) \\ &+ \epsilon \pi_g^*(x_{i,j}, y_{i,j} | \exp(c^{(x)}), \exp(c^{(y)})), \end{aligned} \quad (2.2)$$

where $c^{(x)}$ and $c^{(y)}$ are the league averages for home and away goals respectively.

To allow the parameters to vary in time, they use Brownian motion for $\alpha_A^{t''}$, the attacking strength of team A at time t'' , and tie it to α_A for $t' (\leq t'')$.

$$\alpha_A^{t''} \stackrel{d}{=} \alpha_A^{t'} + (B_{\alpha,A}(t''/K) - B_{\alpha,A}(t'/K)) \frac{\sigma_{\alpha,A}}{\sqrt{1 - \theta_0(1 - \theta_0/2)}}, \quad (2.3)$$

where $B_{\cdot}(t), t \geq 0$ is standard Brownian motion starting at level zero and K is the inverse loss of memory rate. The strength of the teams have to follow equation 2.2 for each game, and 2.3 describes the time development. Markov Chain Monte Carlo methods are used to handle how the strengths of the teams are updated after a game. 1684 matches of Premier League to decide the values of constants $c^{(x)}, c^{(y)}, K, \theta_0$ and ϵ . As a betting strategy they choose to maximize the expected profit minus the variance of the profit. This is attempted on both single bets and combo-bets with multiple games, but single bets is concluded to be easier and more reliable.

Rue, Salvesen bring up several possible improvements to their model. Most notably; The use of more interesting data than simply the final match-result and that the home field advantage should play a part. The goal-model could be improved upon by perhaps using the birth-process approach

of Dixon and Robinson (1998)[3], and the time-model should be updated to include the local trend (first derivative) in its predictions.

Timmaraju, Palnitkar, Knahha - Game on! Predicting English Premier League Match Outcomes [19] take a machine learning approach where the match result is used in combination with corners and shots on target. They test out two ways of using these parameters:

- KKP (k-past Performances) simply uses the average values for the team of the k last matches played by that team. For instance, the goal related parameter is the sum of goals scored the last k games divided by k. This is collected in a vector $P_A = [g_{avg}; c_{avg}; st_{avg}]$ and the ordered difference $P_A - P_B$ is what's taken as the feature (measurable property of observed phenomenon).
- TGKPP (Temporal Gradient k-Past Performances) uses the same k last matches, but applies what they call a temporal differencing operator on the data. $g_{dA} = (g_2 - g_1, g_3 - g_2, \dots, g_k - g_{k-1})$, and $P_{diff} = [mean(g_{dA}) - mean(g_{dB}); mean(c_{dA}) - mean(c_{dB}); mean(st_{dA}) - mean(st_{dB})]$. The feature is the 6-element vector $[P_A - P_B; P_{diff}]$. The reasons behind these choices are not really made clear.

From here they run the features through various standard machine learning algorithms with different values of k. The best results appear to be using the Radial Basis Function kernel on k=7 with the TGKPP, where they achieve an 66.67% prediction accuracy (ignoring the 2-class prediction where they omit all games ending in draws). This looks impressive but is on a tiny sample, they have only tested on 51 different games. There is also a clear weakness that the feature doesn't contain any information about the strength of the teams that were played - a decent team having played bad teams will have a better feature than a good team having played good teams.

Owen - Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter (2011) [15] uses an approach very similar to the one described by Rue, Salvesen (2000)[18], but instead uses it on the Scottish Premier League (SPL) and also makes some alterations to the model. Firstly they discard the changes to the probability distribution for different scores as they did not find any evidence of them applying to the SPL.

Whereas Rue and Salvesen have the prior strengths of teams spread uniformly on (-0.2, 0.2) based on the rankings of the previous year, Owen decides them by applying the model to the previous years data. This brings up an issue where the teams promoted from the 1st division won't have any prior strength. This is solved by giving them the same strength as the team they were replacing. This is not a flawless tactic, as the whole point of having separate strengths for attack and defense is that two teams of equal skill (i.e. expected to end up at the same rank at the end of the season) can have their strength built on a strong attack or a strong defense or a slightly weaker combination of the two. So while the overall strength of the team they are replacing might be similar, directly inheriting the strength composition of an entirely different team is not necessarily the best idea. A different approach might be to keep the overall strength (i.e. $\alpha_A + \beta_A$ the same) but base the relationship of the strengths on the results from the particular teams last season. It is unclear how much better this estimation will be, it may be sufficient to use the simple approach as long as the prior selected reflects the uncertainty

Instead of using continuous time, Owen concludes that using discrete time offers more or less the same predictive probabilities. The reasoning being that the time between two matches for a team is typically restricted to 3-4 days, 7 days or 14 days, and that this simplification speeds up the simulation. Owen lets the evolution variance σ^2 be kept as a parameter in the model, effectively allowing it to better adapt to a team suddenly showing signs of a rapid change of skill. This is especially useful for correcting badly estimated prior strengths, such as for promoted teams.

Koopman, Lit (2015) - A dynamic bivariate Poisson model for analyzing and forecasting match results in the English Premier League [10] is the latest in a long line of models built on the one first proposed by Maher (1982)[12]. The result (X, Y) of a football match between teams i (at home) and j (away) in week t is assumed to be generated from the bivariate Poisson distribution with probability density function

$$P_{BP}(X, Y; \lambda_{i,j}, \mu_{i,j}, \gamma) = \exp(-\lambda_{i,j} - \mu_{i,j} - \gamma) \frac{\lambda_{i,j}^X \mu_{i,j}^Y}{X! Y!} \sum_{k=0}^{\min(X,Y)} \binom{X}{k} \binom{Y}{k} k! \left(\frac{\gamma}{\lambda_{i,j} \mu_{i,j}} \right)^k, \quad (2.4)$$

with $\lambda_{i,j}$ and $\mu_{i,j}$ being the intensities for X and Y respectively and p a coefficient for the dependency between X and Y , $Cov(X, Y) = \gamma$.

The correlation coefficient between X and Y is thereby $\rho = \frac{\gamma}{\sqrt{(\lambda_{i,j} + \gamma)(\mu_{i,j} + \gamma)}}$.

The goal intensities for home team i and away team j in week t are $\lambda_{i,j,t} = \exp(\delta + \alpha_{i,t} - \beta_{j,t})$ and $\mu_{i,j,t} = \exp(\alpha_{j,t} - \beta_{i,t})$, where the δ is the home advantage parameter that can be unique for every team or equal.

To allow the strengths to change over time, the strength parameters α and β are described as auto-regressive processes $\alpha_{i,t} = \kappa_{\alpha,i} + \phi_{\alpha,i} \alpha_{i,t-1} + \eta_{\alpha,i,t}$ and $\beta_{i,t} = \kappa_{\beta,i} + \phi_{\beta,i} \beta_{i,t-1} + \eta_{\beta,i,t}$. κ are team specific unknown constants, ϕ are auto-regressive coefficients and η are normally distributed and independent error terms. They implement the same modification to the joint probability distribution as Dixon, Coles (1997), and they allow the random shocks η to vary in scale so that large changes in strength over winter and summer breaks are accepted. To reduce the number of parameters, the home field advantage of a team is taken from a set of two values, one for the typical top 5 teams (Arsenal, Manchester City, Manchester United, Liverpool and Chelsea), as they are expected to have a larger home advantage, and one for the rest.

For practicality the model is presented in general state space form. The strengths of the teams are stored in a $2J \times 1$ matrix $z_t = (\alpha_{1t}, \dots, \alpha_{Jt}, \beta_{1t}, \dots, \beta_{Jt})$ where J is the number of teams (20), holding the strengths of each team at time t . In this form, $z_t = \kappa + \Phi z_{t-1} + \eta_t$, with $\eta_t \sim NID(0, H)$ and $NID(c, d)$ means a normal distribution with mean c and variance d . κ , Φ and H are matrices defined as:

$$\kappa = (\kappa_{\alpha,1}, \dots, \kappa_{\alpha,J}, \kappa_{\beta,1}, \dots, \kappa_{\beta,J})$$

$$\Phi = \text{diag}(\phi_{\alpha,1}, \dots, \phi_{\alpha,J}, \phi_{\beta,1}, \dots, \phi_{\beta,J})$$

$$H = \text{diag}(\sigma_{\alpha,1}^2, \dots, \sigma_{\alpha,J}^2, \sigma_{\beta,1}^2, \dots, \sigma_{\beta,J}^2)$$

The remaining unknown parameters are placed in the parameter vector $\psi = (\phi', h', \delta, p)'$ where ϕ' and h' are vectors containing the diagonal elements of Φ and H respectively.

To estimate the parameters they maximize the likelihood equation. If g_t is a vector containing the match results from one week ($J/2$ games), then the observation density of g_t given z_t is $p(g_t|z_t; \psi) = \sum_{k=1}^{J/2} p_{BBP}(\lambda_{i,j,t}, \mu_{i,j,t}, p)$. The signal vector is expressed as $\mathbb{E}(g_t|z_t; \psi) = \exp(a_t\delta + W_t z_t)$, where a_t consists of elements of value 1 when the corresponding score in y_t is for a home team and vice versa, and W_t is a matrix that selects the appropriate α, β values from z_t .

For $g = (g'_1, \dots, g'_n)$ and $z = (z'_1, \dots, z'_n)$, the joint density becomes $p(g, z; \psi) = p(g|z; \psi)p(z; \psi)$ where $p(z; \psi) = p(z_1; \psi) \prod_{t=2}^n p(z_t|z_1, \dots, z_{t-1}; \psi)$, which leads to the likelihood function

$$l(\psi) = \int p(g|z; \psi)p(z; \psi)dz \quad (2.5)$$

This has no analytic solutions and as numerical integration is unfeasible Monte Carlo simulation methods are used to evaluate for different values of ψ .

For testing the out-of-sample performance of the model, meaning how well it predicts future, unseen games, they adopt a conservative betting strategy with slight modifications from the one used by Rue, Salvessen(2000)[18]. They maximize expected profit, but only accept bets where the expected value (EV) is above a threshold ω for some $\omega > 0$. Also they consider any bet with odds higher than 7 a "long shot", and even for $EV > \omega$ they only bet 0.3 units on these instead of the 1.0 placed on normal bets. For $\omega = 0.4$ this means playing 50 bets over two seasons and an expected profit of 50%.

Of possible improvements they mention the use of more information from games, testing out other dynamic models such as random walks long memory processes, using Bayesian Markov Chain Monte Carlo methods for making predictions that account for parameter uncertainty, and lastly exploiting the lack of efficiency in the betting market - i.e. using the highest odds on the market instead of the average.

2.1 Other studies of interest

Oberstone(2009) - Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success [14] seeks to uncover what separates a good team from a bad one in the BPL - basically, what data is worth looking at. Data is collected from the 2007-2008 season of BPL, and a regression model is devised using the amount of point gathered over the season, Y , as the dependent variable. Starting out with 17 independent variables, this is narrowed down to 6 statistically significant pitch interactions; (1) goals to shots ratio, (2) % goals scored outside of the box, (3) short to long pass ratio, (4) total crosses, (5) average goals conceded and (6) yellow cards. These values are all taken over the whole season, and only variables (5) and (6) have negative impacts on Y . The model is used to retrodict the same season and the results are very good with $R^2 = 0.990$ and $p < 0.0001$, though I would have liked to see the model be tested on some other season than the one the data was drawn from. By design the model will do a good job at 'predicting' itself, but doing well on an other season would mean that the model has uncovered some underlying success factors of being a good team.

Oberstone also runs an ANOVA to see if any of the pitch interactions are significant for a team to be a part of the top 4, the bottom 4 or the middle 12. This highlighted a number of variables that significantly have an impact on what tier a team places in. Average shots fired per game and basically everything to do with passing (number of passes, short to long ratio, pass completion) are all higher in the higher tiers. Crossing seems to be at an even level throughout the league and not have any real impact. Defensively the good teams make more tackles while receiving fewer cards.

Goddard (2005) - Regression models for forecasting goals and match results in association football [8] focuses on deciding whether (1) result modeling (H,D,A) or (2) goal modeling is the best. One could assume that the goal-model should be better because it's built on more extensive data, but on the other hand league points are awarded for winning (and drawing) games and not for scoring goals, meaning the goal data might include a lot of noise that isn't really relevant. One also avoids relying on models like the Poisson distribution to describe something that not everyone agrees is Poisson distribution.

To decide which method is the best, the comparison has to be done on the basis of how to best predict match result (H,D,A) as only the goal-focused method can predict the amount of goals. To measure the forecasting performance, Goddard adopts the pseudo-likelihood introduced by Rue, Salvesen(2000) - the geometric mean of the estimated probabilities of the actual results. So if the results of matches M_1 and M_2 were H and D respectively and a model had probabilities for these two results as $P_1(H) = 0.3$ and $P_2(D) = 0.25$ then pseudo-likelihood is $\sqrt[3]{0.3 * 0.25}$. In the study, the best method seems to be a combination of the two - using goal data to decide the predictors and using those predictor to forecast the match result directly. This method has the best results on the majority of the seasons tested, although it doesn't always give the best results, and Goddard concludes that the forecasting ability of the different approaches are rather similar.

Fong, Rue, Wakefield (2009) - Bayesian inference for generalized linear mixed models [5] describe how Integrated Nested Laplace Approximations (INLA) has made Bayesian inference for Generalized Linear Mixed Models (GLMM) feasible. Markov Chain Monte Carlo methods have long been the gold standard for simulation, as they are easy to implement and one can achieve an arbitrary accuracy by running the algorithm long enough. However, this comes at a price of severe computational cost. INLA resolves this problem by approximating the posterior and then evaluating it using Laplace approximations. The output is the posterior marginal distributions for each parameter.

As they time the computations, the differences in runtimes are enormous. For an example problem with temporal smoothing, INLA used 45 seconds on a single core while MCMC required 15 hours to achieve similar accuracy. For an example in B-spline nonparametric regression, INLA took 5 seconds to run while MCMC required 40 hours to reach the same accuracy. So INLA has some very attractive properties, especially if one doesn't require a specific accuracy and if the user lacks access to a super computer.

Chapter 3

Presentation of data

While others have been able to forecast match results with decent success, they have, with few exceptions, only used the final score as the independent variable. Over the years, the amount of data available has become much greater - both in terms of size and variety. For the largest soccer leagues, anyone can download massive data sets for no charge, and this means there is room for models taking advantage of these new data. The data used in this paper are openly available at www.football-data.co.uk. [6] I am interested in a wide variety of data, including the match result, shots by each team, shots on target and the bookmaker odds for each of these games. All of this is available from the season 2000-2001 until present day, meaning 14 completed seasons are available for optimizing my model.

The actual data is a matrix where each row represents a match, and the columns group the following statistics:

- Date
- (H) Home Team
- (A) Away Team
- (FTHG) Full Time Home Team Goals
- (FTR) Full Time Result (H, D or A)
- (HTHG) Half Time Home Team Goals
- (HTR) Half Time Results (H, D or A)
- (HS) Home Team Shots
- (HSOT) Home Team Shots on Target
- (HC) Home Team Corners
- (HF) Home Team Fouls
- (HY) Home Team Yellow Cards
- (HR) Home Team Red Cards

and similarly, FTAG, HTAG, AS, ASOT, AC, AF, AY and AR represent the corresponding values for the away team. In addition the data includes home win, draw and away win odds collected from

a variety of bookmakers, and the most recent seasons include maximum and average odds across the different bookmakers as calculated by www.betbrain.com. This will prove useful for creating and testing an effective betting strategy.

I am interesting in finding out which data have the strongest correlations to scoring and conceding goals, which again leads to winning or losing games. I have chosen to forecast match results by predicting the amounts of goals scored instead of directly predicting the result, so I will not be looking for any connections directly leading to a team winning, nor will I look at the statistics recorded at Half Time.

Effect on defense from committing fouls A foul gives the opponent a free kick, so it's natural to think that committing fouls has a negative effect on your defense. To look at this effect I make three linear regression models where goals allowed, shots allowed and SOT allowed are explained by fouls.

$$\text{ShotsAllowed} = 13,435 + 0,0004 \cdot \text{Fouls}$$

$$\text{SOTAllowed} = 4,477 - 0,0014 \cdot \text{Fouls}$$

$$\text{GoalsAllowed} = 1,362 + 0,0021 \cdot \text{Fouls}$$

First of all these effects very small. On average a team commits 10.77 fouls per game, which translates to about 0.02 goals per game. Secondly none of the effects appear to be significant. I safely reject that fouls committed have any real impact on a match result.

Effect on defense from receiving yellow or red cards If a team receives a yellow card (YC) they have to play more carefully, while receiving a red card (RC) will often directly weaken their defense, so I am interested in how they affect the defensive statistics (goals allowed, shots allowed and SOT).

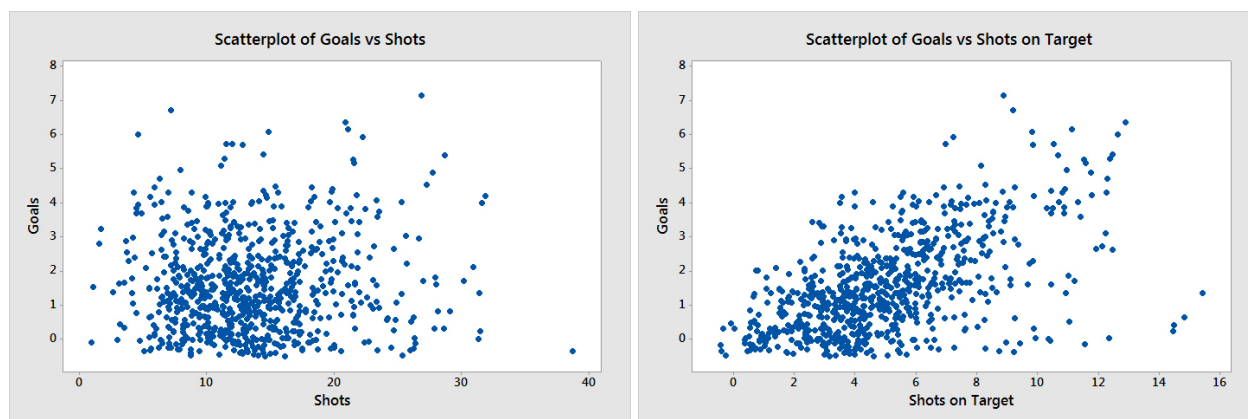
$$\text{GoalsAllowed} = 1,1529 + 0,1214 \cdot \text{YC} + 0,540 \cdot \text{RC}$$

$$\text{ShotsAllowed} = 12,678 + 0,325 \cdot \text{YC} + 3,482 \cdot \text{RC}$$

$$\text{SOTAllowed} = 4,093 + 0,1870 \cdot \text{YC} + 1,012 \cdot \text{RC}$$

Every one of these effects are significant ($p < 0.05$). To put them in perspective, the average team receives 1.59 YCs and 0.070 RCs per game. The three most usual amounts of YCs handed out to a team in a game are 1, 2 and 0 in that order, and this amounts to about 80% of the matches. Overall the effects from YCs are not that strong, and more importantly the expectation is that a team will receive one or two YCs in a match. As a team is expected to receive one or two YCs, the effect of those cards can be included in the expected goals/shots/SOT allowed, meaning YCs will only have a real impact when it is abnormally high. A team receiving 6 YCs (unusually high) in a game will only be about 4 over the expected, and this should lead to maybe 1 extra SOT, which is not that big of an impact.

RCs on the other hand have a very strong effect in all categories, which is to be expected. RCs appear in only about 5% of all matches, but when they do they make a clear impact. It's interesting that receiving an RC increases the $\frac{\text{GoalsAllowed}}{\text{SOTAllowed}}$ ratio, and this might be because the RC is sometimes received in combination with giving away a penalty or a free kick in a dangerous position. Based on all this I think ignoring YCs and focusing on RCs is a good strategy, and I want to keep the effect of an RC on both goals and shots/SOT.



(a) Scatterplot showing the connection between shots and goals scored. (b) Scatterplot showing the connection between SOT and goals scored.

Effect on offense from receiving red cards Receiving a YC shouldn't have any real impact on the offensive strength of a team, but if a team is reduced to 10 players their offensive power definitely takes a hit. Using the same approach as for the defense I have these formulas for explaining the impact of red cards on the offense of a team:

$$GoalsScored = 1,4285 - 0,635 \cdot RC$$

$$ShotsFired = 13,637 - 2,832 \cdot RC$$

$$SOTFired = 4,5695 - 1,544 \cdot RC$$

where again all the effects are significant ($p < 0.05$). As expected, receiving an RC makes a big negative impact, and again the RC seems to have two separate effects on goals and SOT. The $\frac{GoalsScored}{SOTFired}$ ratio decreases, which could be a result of attackers either being sent off or being substituted with a defender to compensate for a defender being sent off. Not only does the team fire fewer shots, but the mixture of players will now also on average be worse at shooting, so again I want to keep the effect on both goals and shots/SOT.

Shots/SOT and goals Shots and SOT both have the obvious connection to goals scored that (ignoring own goals) you can't have a goal without both a shot and an SOT. Figures 3.1a (goals vs shots) and 3.1b (goals vs SOT) are both plotted with noise of (+/- 0.5) to avoid points being stacked on top of each other. Goals vs Shots has a cloud centered around the two averages ($\mu_s = 13.5$ and $\mu_g = 1.38$), but there is no clear indication that a lot of shots lead to a lot of goals. Goals vs SOT has a more distinctive trend where, as expected, more SOT is associated with more goals scored. Both will need to be assessed further, but SOT looks to be the more useful statistic.

Correlation between shots/SOT and shots/SOT conceded A team with a good offense typically has a good defense, so a team getting a lot of chances should not allow as many chances. And if a team gets a chance then there is a significant interval of time where the other team could not have gotten a chance as they cannot happen at the same time.

I test the correlation using Spearman's r and Pearson's ρ .

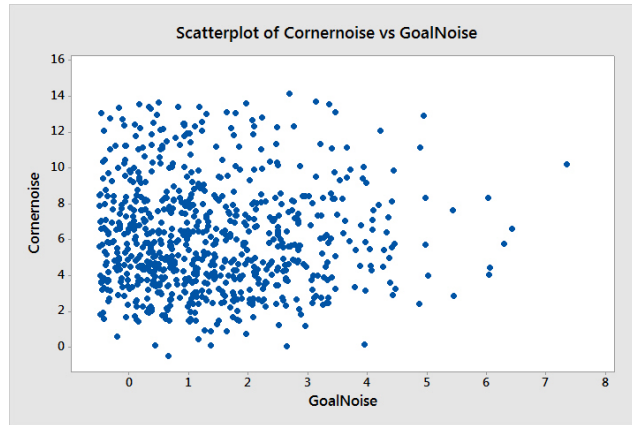


Figure 3.2: Scatterplot showing the connection between corners and goals scored.

Shots and shots allowed have correlations $r = -0.480$ and $\rho = -0.501$, significant with $p < 0.001$. SOT and SOT allowed have correlations $r = -0.280$ and $\rho = -0.273$, significant with $p < 0.001$.

This shows that there is a clear positive correlation between $\hat{\alpha}$ and $\hat{\beta}$.

Corners I don't have access to goals scored by corners, so I'll have to look at the indirect effect of the total goals scored. Figure 3.2 shows goals scored (with noise) plotted against corners (with noise). From this it's pretty clear that the amount of corners a team gets has basically no impact on goals scored.

The Home Field Advantage (HFA) I'm interested in how the home field advantage changes goals scored (HFA_g), Shots Fired (HFA_s) and SOT (HFA_{SOT}). Testing this formally is difficult because the data are taken from matches with different teams so even if goals scored follows some distribution the sample values would be drawn from 380 similarly shaped distributions with different means. Instead I'll simply compare the means of the data and do a graphical analysis. Figure 3.4 shows the frequency of shots on target grouped by home and away teams. Clearly home teams have some advantage in creating chances.

To examine HFA_g I'll look at how the goal:shot and goal:SOT ratios change by being home or away.

$\mu_{G:shot,H}$ and $\mu_{G:shot,A}$ are the average goal:shot ratios for home and away teams, while $\mu_{G:SOT,H}$ and $\mu_{G:SOT,A}$ are the average goal:SOT ratios.

For my data sample I find that $\mu_{G:shot,H} = 0.110$, $\mu_{G:shot,A} = 0.102$, $\mu_{G:SOT,H} = 0.315$ and $\mu_{G:SOT,A} = 0.294$. There appears to be some positive effect on the ratios from being at home, but the effect is so minuscule that I choose not to include it further.

The distribution of shots and SOT is of interest as I want to estimate the expected amount of chances created by a team in a game, and it would be very convenient if they were to follow the Poisson distribution. I can't actually test for this formally as like earlier every match would have a unique mean to their distribution, but I can plot the count data for shots and SOT against the expected

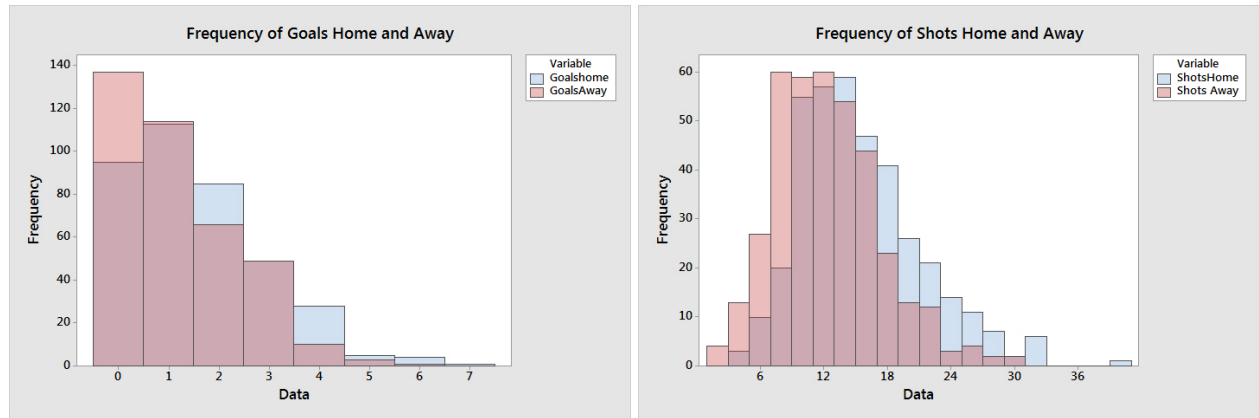


Figure 3.3: Histogram of goals (left) and shots (right) by home and away teams

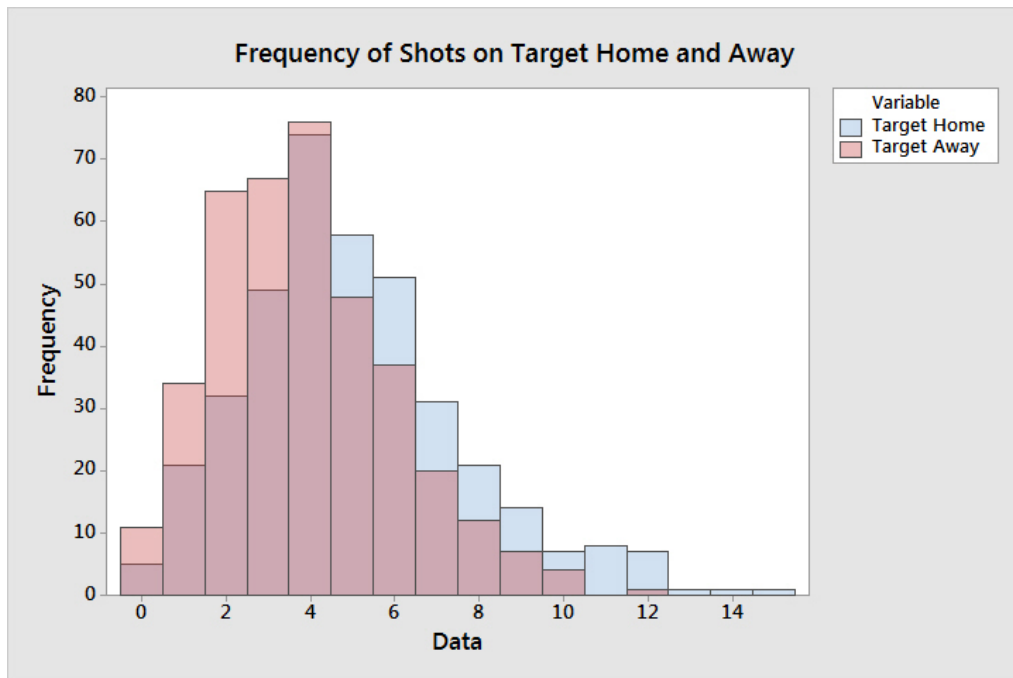


Figure 3.4: Histogram of SOT by home and away teams

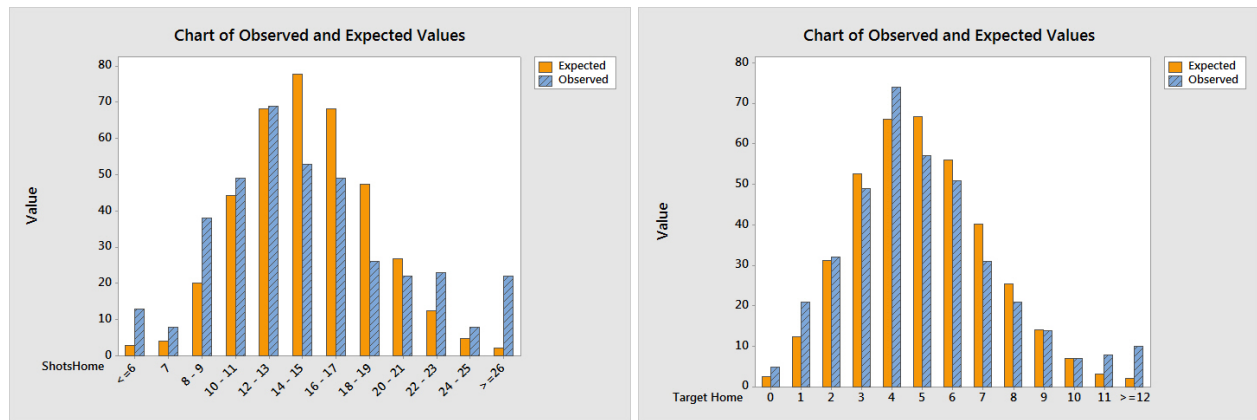


Figure 3.5: Observed vs Expected values for shots (left) and SOT (right) for home teams

values to see if it looks plausible. To prevent the HFA from disrupting the data I only look at the data for home teams.

By just looking at the graphs it seems like both *could* follow the Poisson distribution, and it seems as though SOT has a better fit.

To recapitulate, red cards seem to affect every part of both defense and offense negatively. SOT seems to more consistently be converted into goals than just shots, so it could be useful for describing a teams ability to create goal-scoring opportunities. The home field advantage has a clear, positive impact on goals scored, shots and SOT. The other statistics (fouls and corners) do not seem to be of any big importance. Lastly, both shots and SOT seem to fit decently to the Poisson distribution.

Chapter 4

Quality Assessment of the models

Prior to introducing the actual model, I need a way to gauge how well it works. To measure the accuracy of the models I'll use three tools; DIC and WAIC - two related tools for comparing goodness of fit while factoring in model complexity, and Second Half Pseudo-Likelihood - used for comparing mid-season predictive power. Both WAIC and DIC are readily available in the INLA package for R, while the Pseudo-Likelihood is implemented specifically for this project.

4.1 Likelihood

I won't be using the likelihood directly, but it plays a big part in calculating the WAIC and DIC so it's natural to include a brief explanation. Essentially the likelihood is how likely an explanation is, or in this case, how likely it is that the data could have been produced by the suggested model. Calculating the likelihood means making a guess as to how, i.e. by what distribution, the data was generated, and then taking the product of the probability mass function of all the data. So if my observed data is X_1, \dots, X_n and I believe this is generated from the Poisson distribution with mean λ then the likelihood function is $L(\lambda; X) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$. Typically the log-likelihood $l(\lambda; X) = \log(L(\lambda; X))$ is used as it's easier to maximize.

4.2 Deviance Information Criterion (DIC)

The DIC [2] is a generalization of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), designed to compare the goodness of fit (deviance) of two models while penalizing over-parametrization. Ideally I would want to look at the predictive capabilities on out-of-sample data. The problem is that we don't have access to the actual data-generating model, so the data becomes increasingly sparse the bigger the out of sample becomes. Leave-One-Out Cross Validation is an option, but it is very computationally expensive. DIC is an attempt to work around this by giving an adjusted within-sample predictive accuracy.

The deviance of a fit of a fit of a fit is defined as the double log-likelihood ratio between the model and the full model where every observation has a parameter giving it perfect fit, $D(y) =$

$-2(\log(p(y|\theta_0)) - \log(p(y|\theta_s)))$, where a smaller deviation means a better fit. θ_0 is the posterior parameters estimated by maximum likelihood, θ_s is the fitted parameters for the full model and y is the observed effects. This is an insufficient statistic as it can get arbitrarily small by simply adding more parameters, so there has to be a term that punishes having too many parameters.

AIC is an attempt to fix this by adding a linear penalty for adding parameters, meaning they have to significantly contribute to an increased fit to be included. $AIC = D(y) + 2k$, where k is the number of parameters in the model and $D(y)$ is the deviance as defined above. This is sufficient for very simple models, but having informative priors tends to (1) reduce the amount of over-fitting (meaning the $+2$ punishment per parameter is too strict), and (2) reduces the "effective number of parameters".

DIC makes two changes to the AIC. It replaces the Maximum Likelihood with the posterior mean θ_{Bayes} in the Deviance definition, and changes out k with a data-based correction term.

The replacement for k is referred to as the effective number of parameters, defined as $p_{DIC} = 2(\log(p(y|\theta_{Bayes})) - E_{post}(\log(p(y|\theta))))$, where the second term is the average of the posterior parameters calculated through simulations. The posterior mean θ_{Bayes} is the mean of the posterior distribution with mean square error used as risk.

The actual equation becomes $DIC = -2(\log(p(y|\theta_{Bayes})) + 2p_{DIC})$, where a lower DIC is better.

4.3 WAIC - Watanabe Akaike Information Criterion

WAIC[20] is defined as $WAIC = -2(lppd - p_{WAIC})$ where $lppd$ is the log pointwise predictive density calculated as $\sum_{i=0}^n \log(\frac{1}{S} \sum_{s=1}^S p(y_i|\theta^s))$ for S simulations of the posterior density. p_{WAIC} is the WAIC effective number of parameters $p_{WAIC} = \sum_{i=1}^n var_{post}(\log(p(y_i|\theta)))$, and is a way of expressing the amount of unconstrained parameters. Parameters having no prior information and no constraints will count as 1, while parameters with complete prior information will count as 0.

WAIC is based on pointwise calculations, with the ambition of estimating a LOO-CV test (as this is too computationally demanding). This is important because it means that WAIC is evaluating predictions of actual not-seen data, a property not found in AIC and DIC. [7] I'm using the version of WAIC implemented in the INLA package.

4.4 Second Half Pseudo-Likelihood

I want to directly test the predictive power of the models, meaning I have to attempt out of sample prediction. Pseudo-Likelihood (PL) refers to the geometric mean of the estimated probabilities for the actual results, as designed by Rue and Salvesen (2000)[18] and later used by Goddard (2005)[8]. For each match M_i the prediction model gives a probability distribution for the outcomes H (home victory), D (draw) and A (away victory), for instance $P(H) = 0.25$, $P(D) = 0.25$, $P(A) = 0.5$. Over a large amount of games, the model that gives me the highest probability prediction for the actual result is the best one. Put mathematically, where $R(M_i)$ is the result that can either be H, D or A, and that the actual outcome for that match is denoted by r :

$$PL = \sqrt[N]{\prod_{i=0}^N P(R(M_i) = r)} \quad (4.1)$$

For this to test actual predictive power, I'll only start the prediction after half the season. A season has 38 rounds, so the first to be predicted is round 20 where the first 19 rounds are included in the model as history, the second to be predicted is the 21st round where the first 20 rounds are included as history, and so on. If the PL gets substantially higher by increasing the complexity of the model, then that is a good indication that the added complexity is worth it.

Finding the probability of the outcomes H, D and A is not trivial, an estimation based on simulation is my best option. I'll have to generate valid samples of the parameters and use them to simulate matches to get a general picture of the probability distribution. In all the models we have that $X \sim \text{bin}(\hat{X}, p)$, $Y \sim \text{bin}(\hat{Y}, q)$ and $\hat{X} \sim \text{Po}(\hat{\lambda}_x)$, $\hat{Y} \sim \text{Po}(\hat{\lambda}_y)$.

The parameters are not known to an exact degree, they are estimated to best ability with an accompanying precision matrix quantifying the uncertainty. I can generate a large number of samples from the posterior distribution, and for each of these I can simulate each game a large number of times each time recording the result (x,y). After enough simulations I'll have a good overview of the probabilities for each score and by extension the result of the game. Equation 4.1 is valid for measuring both of these predictive capabilities. PL(score) is the Pseudo-Likelihood for the match scores (i.e. (3,2), (1,0) etc), while PL(result) is the Pseudo-Likelihood for the match result (i.e. H, D, A).

Chapter 5

Designing a model for prediction

To make the approach apprehensible I want to start off with a simple model and gradually increase the complexity. This will also allow me to verify that the accuracy of the model is improving. For the simulations and testing in this chapter, what is referred to as a chance is always a shot on target (SOT). This is done because running simulations with both versions would be too time consuming, so I will first focus on finding a good model and leave the shots vs. SOT decision for later.

The models are implemented using the R package INLA.

5.1 Model 1: Chances Poisson distributed and constant p of conversion

I start with a model where each team (home team i and away team j) has two strengths ($\hat{\alpha}$ and $\hat{\beta}$) describing their chance creating and chance preventing abilities, and say that chances created in a match ($\hat{X}_{i,j}$ and $\hat{Y}_{i,j}$) are independently distributed Poisson processes conditioned on these strengths with means $\hat{\lambda}_{i,j}$ and $\hat{\mu}_{i,j}$. There is also a constant home field advantage $\hat{\delta}$ that works in the favor the home team. The estimated means are calculated by maximum likelihood estimation of the regressions $\ln(\hat{\lambda}_{i,j}) = \hat{\alpha}_0 + \hat{\alpha}_i - \hat{\beta}_j + \hat{\delta}$ $\ln(\hat{\mu}_{i,j}) = \hat{\alpha}_0 + \hat{\alpha}_j - \hat{\beta}_i$

where $\hat{\alpha}_0$ is the intercept term.

As $\hat{X}_{i,j}$ and $\hat{Y}_{i,j}$ are both Poisson distributed with means $\hat{\lambda}_{i,j}$ and $\hat{\mu}_{i,j}$, together chances created by both teams are distributed as

$$P(\hat{X}_{i,j} = x, \hat{Y}_{i,j} = y | \lambda; \mu) = e^{-(\lambda+\mu)} \frac{\lambda^x \mu^y}{x! y!}, \quad (5.1)$$

where $\lambda = \hat{\lambda}_{i,j}$ and $\mu = \hat{\mu}_{i,j}$.

Goals scored ($X_{i,j}$ and $Y_{i,j}$) are Binomially distributed conditioned on chances created ($\hat{X}_{i,j}$ and $\hat{Y}_{i,j}$) and probability of conversion p , where p is assumed equal for all teams. For home team i , $X_{i,j} \sim \text{bin}(\hat{X}_{i,j}, p)$, so for $n = \hat{X}_{i,j}$ we have that $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$.

I impose the constraint that $\sum_{i=0}^J \alpha_i = 0$, $\sum_{i=0}^J \beta_i = 0$, $\sum_{i=0}^J \hat{\alpha}_i = 0$, $\sum_{i=0}^J \hat{\beta}_i = 0$, meaning that a match result is always equally good for one team as it is bad for the other team.

5.1.1 Priors

For model 1 I have to choose suitable priors for the chance-related strengths ($\hat{\alpha}$ and $\hat{\beta}$), the league-wide constant probability of conversion p and the home field advantage $\hat{\delta}$. Using Gaussian priors are the most reasonable choice, so I have to give expected values and precision for these priors. I do this by running the model over 14 seasons using uninformative priors and comparing the posterior. The mean makes an appropriate expected value, and I can calculate the sample variance to see how much the values vary. Precision is defined as $\text{Prec}(X) = \frac{1}{\text{Var}(X)}$, but I'll be using $\text{Prec}(X) = \frac{1}{2\text{Var}(X)}$ as a conservative choice as 14 samples is not enough to really judge variance.

For $\hat{\alpha}$ and $\hat{\beta}$ I just collect all the offensive and defensive strengths observed, giving me $20 * 14 = 280$ values of each. By design they will be centered around 0, and both the offensive and defensive mean have sample variances around 0.035. So my priors for $\hat{\alpha}$ and $\hat{\beta}$ are both Gaussian with mean 0 and precision 14.

$\hat{\delta}$: Sample mean 0.28, sample variance 0.001363. A variance this low would lead to the posterior being completely dominated by the prior, so in this case I simply say that the prior for $\hat{\delta}$ is Gaussian with mean 0.28 and precision 50.

p : I'm actually looking for the variance and mean of $\text{logit}(p)$. Sample mean is -1.267 and sample variance is 0.031. I'll let the intercept be decided within the model, so I'm only modeling the error term. Because of this my prior is Gaussian with mean 0 and precision 15.

5.1.2 Performance with data

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
Model1DIC	5388.88	5418.70	5764.27	5855.31	5644.97	5632.78	5652.16	5784.30	5815.74	5901.10	5905.48	5973.16	5963.57	5216.53
Model1WAIC	5389.38	5420.26	5772.83	5865.00	5653.14	5638.60	5660.89	5793.94	5825.73	5908.88	5912.97	5982.81	5972.71	5216.12

Table 5.1: WAIC and DIC for Model 1

WAIC and DIC for a model alone is not very useful, this is for comparison with future models.

5.2 Model 2: Unique probability of conversion for each team

Some players are good finishers and others are bad finishers, so as a natural extension of Model 1 I here allow for teams to have a unique probability of converting a chance into a goal. This means every team is described by three strengths ($\hat{\alpha}$, $\hat{\beta}$ and p). Chances are predicted identically to Model 1 and goals are Binomially distributed with $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$, where p is unique for every attacking team and $n = \hat{X}_{i,j}$.

5.2.1 Priors

$\hat{\alpha}$, $\hat{\beta}$ and $\hat{\delta}$ are treated identically to model 1. Here every team has a unique parameter $\text{logit}(p)$, and I'm looking at the variation of the values(+ intercept) across the 14 seasons. The sample mean is at -1.29 and sample variance is 0.0767. The intercept is again decided within the model, I'm only interested in the variance term. Therefore my prior for p is Gaussian with mean 0 and precision 6.

5.2.2 Performance with data

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
Model2DIC	5398.08	5402.98	5762.15	5862.17	5649.14	5629.84	5650.69	5773.62	5810.38	5881.42	5920.98	5980.19	5964.18	5202.39
Model2WAIC	5399.62	5406.59	5773.08	5877.24	5659.96	5640.43	5662.85	5787.51	5823.87	5893.31	5931.33	5992.61	5976.16	5203.97

Table 5.2: WAIC and DIC for Model 2

Table 5.2 WAIC and DIC results for Model 2. As mentioned earlier they have no use on their own, so table 5.3 contains DIC(Model2) - DIC(Model1) and WAIC(Model2) - WAIC(Model1). As the goal is a low DIC and WAIC, a negative number means that Model 2 is better.

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
DIC	9.20	-15.72	-2.12	6.86	4.17	-2.94	-1.47	-10.67	-5.35	-19.68	15.50	7.03	0.60	-14.14
WAIC	10.25	-13.68	0.26	12.24	6.82	1.83	1.96	-6.44	-1.86	-15.58	18.36	9.81	3.45	-12.15

Table 5.3: Difference in WAIC and DIC between Model 1 and Model 2. Negative number means Model 2 performs better.

Overall the results indicate that Model 2 is better than Model 1, though for some seasons the increase in complexity is not worth it.

5.2.3 Validity of Model 2 over Model 1

I'm interested in the validity in going from Model 1 to Model 2 - going from a single value p that is equal for all teams to letting each team have their own p .

In Model 1, the average $\text{logit}(p)$ value is -1.267, which translates to a scoring probability of $p = 0.2197399$ for every team. The sample variance is 0.031.

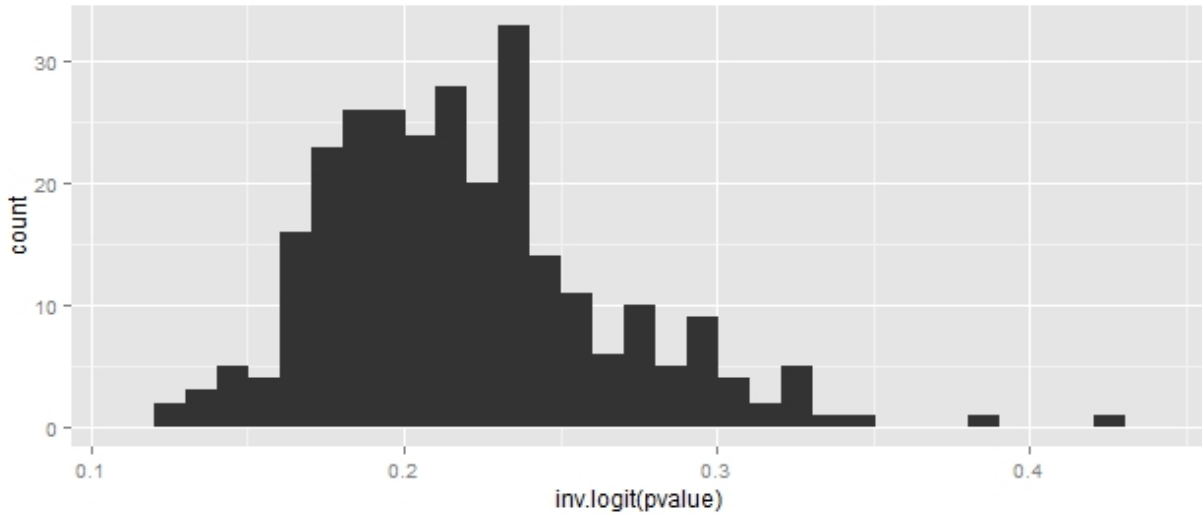


Figure 5.1: Histogram showing the range of probability values in Model 2 for all seasons

Now for Model 2, the average $\text{logit}(p)$ is -1.287 , so pretty much the same. The sample variance here is 0.067 , so more than double. Looking at the max/min values, we have $\max(\text{logit}(p)) = -0.3182$ and $\min(\text{logit}(p)) = -1.9869$ which corresponds to probabilities $\max(p) = 0.42$ and $\min(p) = 0.12$. This means teams that convert about 12% and 42% of their shots would both be treated as converting 22% of them. Because of this it seems very justified to let every team have their own probability value. Figure 5.1 shows the range of different probability values observed.

5.3 Model 3: Probability of conversion depends on the opposition

Similarly to how some players are good finishers and some are bad, not all keepers hold the same quality. I want the probability of conversion to be dependent on both the finishing-strength of the attacking team and the shotstopping-strength of the defensive team. Each team is therefore described by four strengths ($\hat{\alpha}$, $\hat{\beta}$, the goal converting strength α , and the goal preventing strength β). The probability of conversion for the home team in a match is $p_{i,j} = \frac{1}{1+e^{\alpha_0-\alpha_i+\beta_j}}$ and for the away team $q_{i,j} = \frac{1}{1+e^{\alpha_0-\alpha_j+\beta_i}}$, where α_0 is the model offset.

5.3.1 Priors

$\hat{\alpha}$, $\hat{\beta}$ and $\hat{\delta}$ are again treated identically to model 1 and 2. Here every team has two unique parameter α and β and I'm looking at the variation of the values(+ intercept) across the 14 seasons. For α the sample mean is at -1.30 while the sample variance is 0.0776. For β the sample mean is -1.32 and the sample variance is 0.0740. The variances are basically the same, and as again I'm only interested in the variance term I can use the same prior for both parameters. Therefore my prior is Gaussian with mean 0 and precision 6.

5.3.2 Performance with data

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
Model3DIC	5415.15	5413.12	5773.76	5861.77	5652.99	5643.01	5647.89	5765.94	5805.30	5866.15	5928.82	5976.21	5955.25	5213.31
Model3WAIC	5418.47	5419.51	5788.26	5882.05	5667.51	5659.22	5663.85	5783.50	5822.44	5881.51	5942.59	5991.92	5969.75	5217.41

Table 5.4: WAIC and DIC for Model 3

Table 5.4 shows the WAIC and DIC scores for Model 3 across the seasons.

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
DIC	17.07	10.13	11.61	-0.40	3.85	13.17	-2.80	-7.68	-5.08	-15.27	7.85	-3.98	-8.93	10.92
WAIC	18.85	12.93	15.17	4.81	7.55	18.79	0.99	-4.01	-1.43	-11.80	11.26	-0.70	-6.41	13.44

Table 5.5: Difference in WAIC and DIC between Model 2 and Model 3. Negative number means Model 3 performs better.

Table 5.5 shows the change in WAIC and DIC when moving from Model 2 to Model 3, where a negative number means an improvement. Overall the benchmark seems to favor Model 2, though 5 of the 14 seasons have an improved WAIC for Model 3.

5.3.3 Validity of Model 3 over Model 2

I'm interested in the validity in going from Model 2 to Model 3 - from each team having a unique probability p of conversion to letting p be unique to each match - decided by the parameters α and β

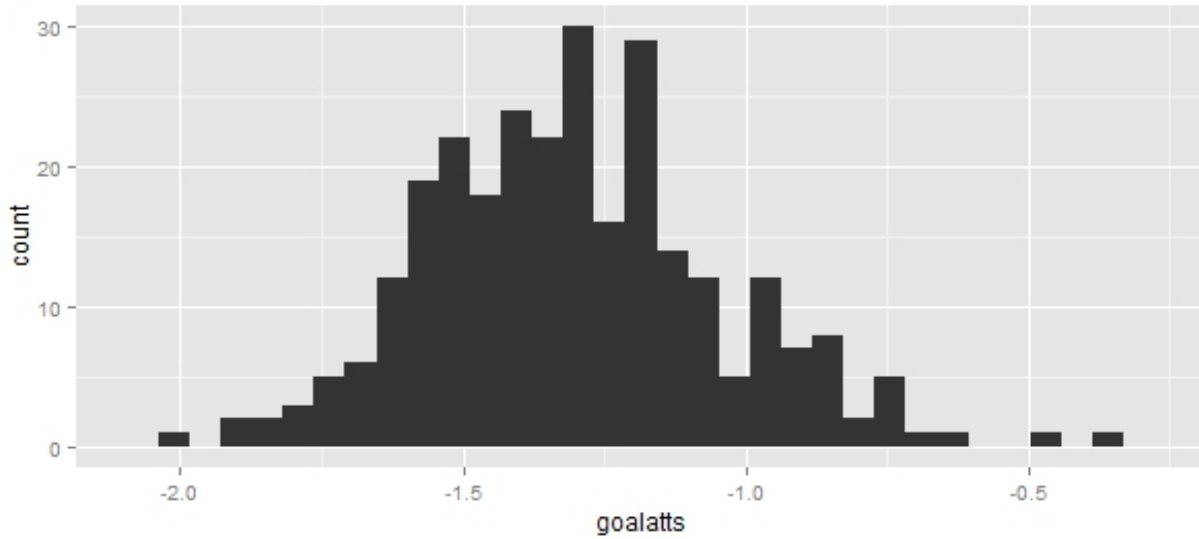


Figure 5.2: Range of goal converting strength α in Model 3 all seasons

of the attacking and defending team respectively.

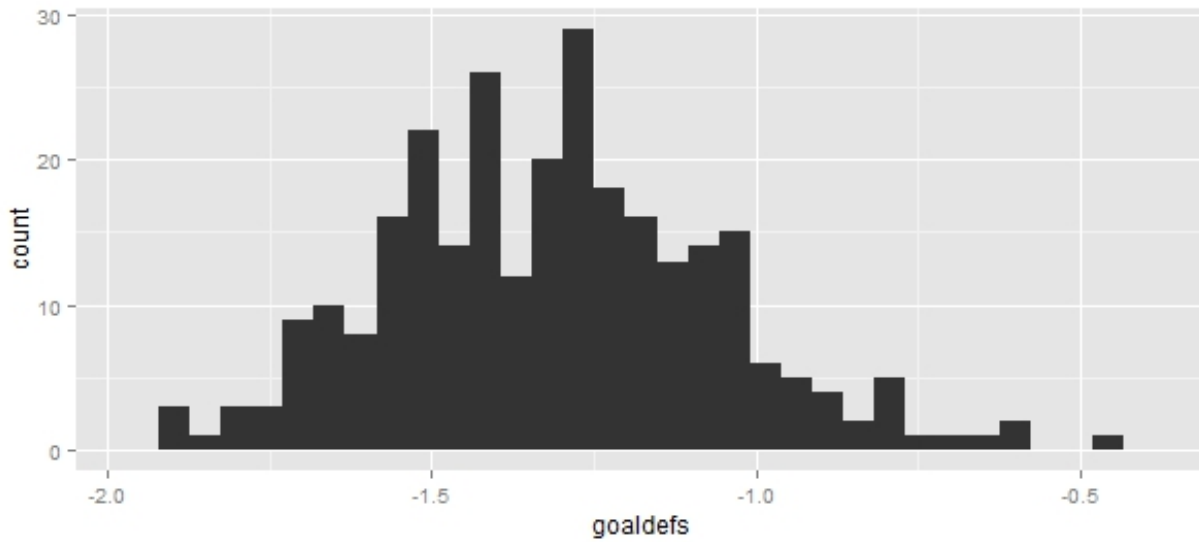


Figure 5.3: Range of goal defending strength β in Model 3 all seasons

As shown in figures 5.2 and 5.3, both parameters are estimated to have widely different values for different teams, so there is definitely a case for having both parameters for each team.

Summary Model 1-3

Figures 5.4 to 5.7 demonstrate how the models compare to each other. DIC prefers Model 3 in the majority of the cases. WAIC seems to value the 3 models rather evenly. Season 14 (actually season 13/14) seems to be an something of an outlier in that it was very consistent and therefore gets low scores in WAIC and DIC.

For the Pseudo-Likelihood I ran prediction for seasons 05/06 until 11/12. I run 1000 samples from the posterior and for each of these the games are simulated 1000 times. This means that running one season with one model takes about 20 minutes, but it is necessary to make sure every possible outcome of every match is being simulated, especially for looking at the Pseudo-Likelihood of the score.

The results of the prediction testing is not really conclusive. For PL(result) - Model 3 does the best of all models in 4 of the 7 seasons, but also performs the worst of all for season 10/11. For PL(score) - Model 3 is only the best in two of the seasons, and Model 2 is generally the worst.

Overall model 3 seems to be an improvement on Model 1 and especially Model 2, but I think the Model can be improved upon significantly.

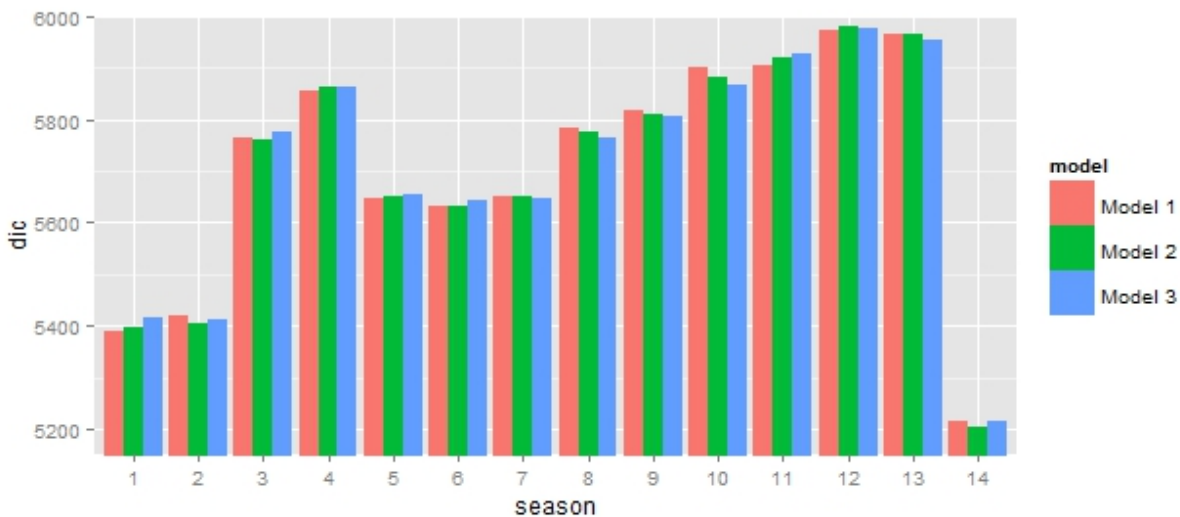


Figure 5.4: DIC for all seasons Model 1-3

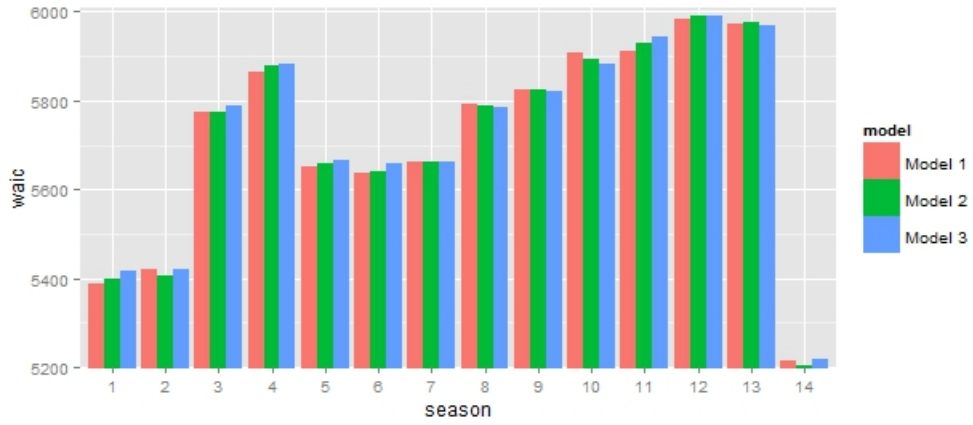


Figure 5.5: WAIC for all seasons Model 1-3

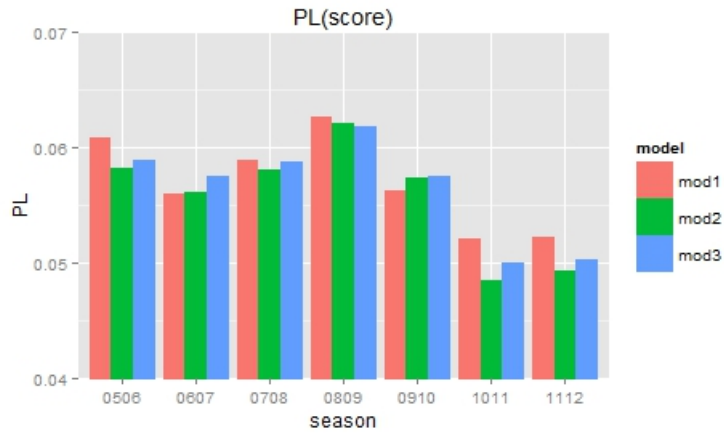


Figure 5.6: PL(score) (y-axis) seasons 03/04 and 08/09 Model 1-3

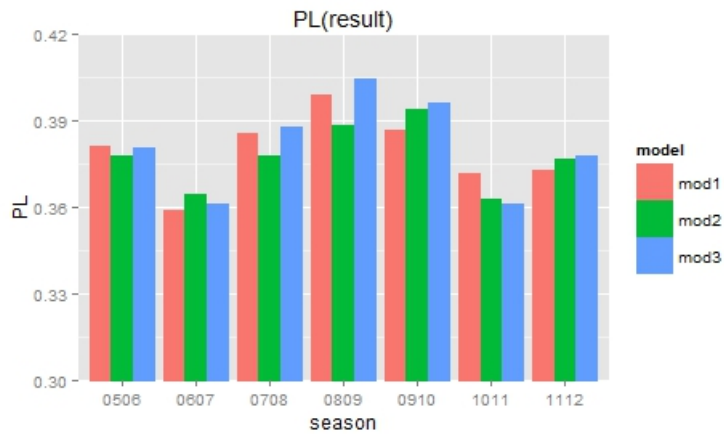


Figure 5.7: PL(result) (y-axis) seasons 03/04 and 08/09 Model 1-3

Further Improvements on Model 3

5.3.4 Correlation between parameters

So far the models have assumed that all parameters are uncorrelated. A natural assumption is that, while a team does not have to be equally good at producing shots and at converting shots into goals, a good offense is most likely good at both. Also, a team with a good offense most likely has a good defense, as good teams tend to be good in all aspects of the game and conversely for bad teams. This does not mean that I can reduce the amount of parameters, but if α , β , $\hat{\alpha}$ and $\hat{\beta}$ are correlated to some degree then that puts constraints on the values that reduces variance and the number of effective parameters.

I am looking for correlations between α and $\hat{\alpha}$, between β and $\hat{\beta}$, and between $\hat{\alpha}$ and $\hat{\beta}$. This is because $\hat{\alpha}$ and $\hat{\beta}$ both refer to the overall defensive and offensive skills of the team. The most natural way of exploring the correlations is to plot the values against each other and see graphically if there is a connection.

Figures 5.8, 5.9 and 5.10 show consistent positive correlations between these parameters. The correlation is especially strong between $\hat{\alpha}$ and $\hat{\beta}$.

In addition to this, as shown by figure 5.11 there is a small positive correlation observed between α and β . These are more player-specific skills referring to the shot-stopping of the keeper and the quality of shots made by the forwards. There is no reason to think that these would be strongly correlated, except for the general idea that good teams have good players.

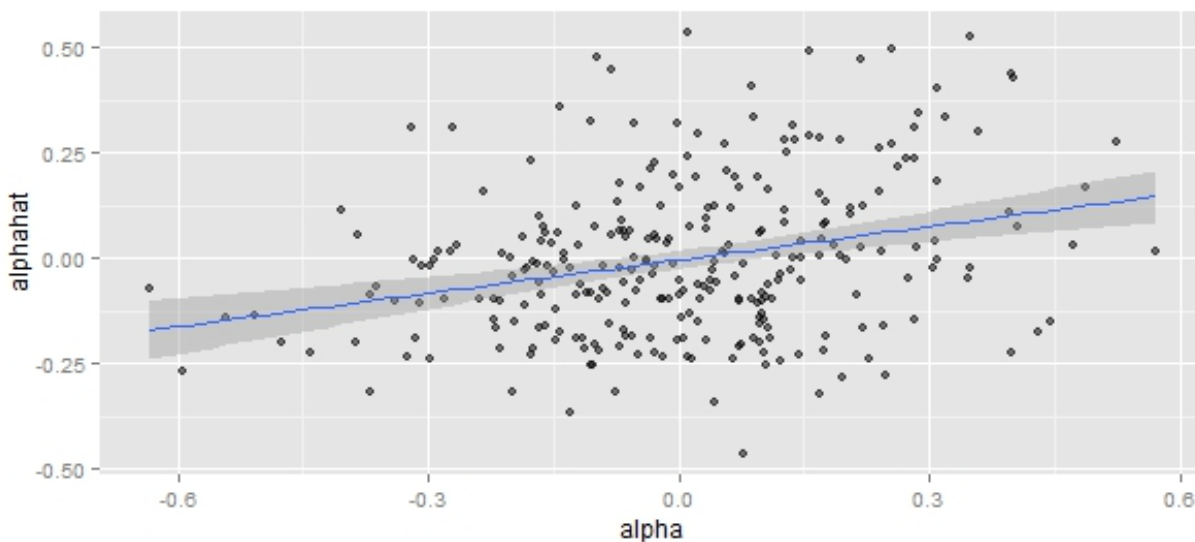


Figure 5.8: Scatter plot α against $\hat{\alpha}$

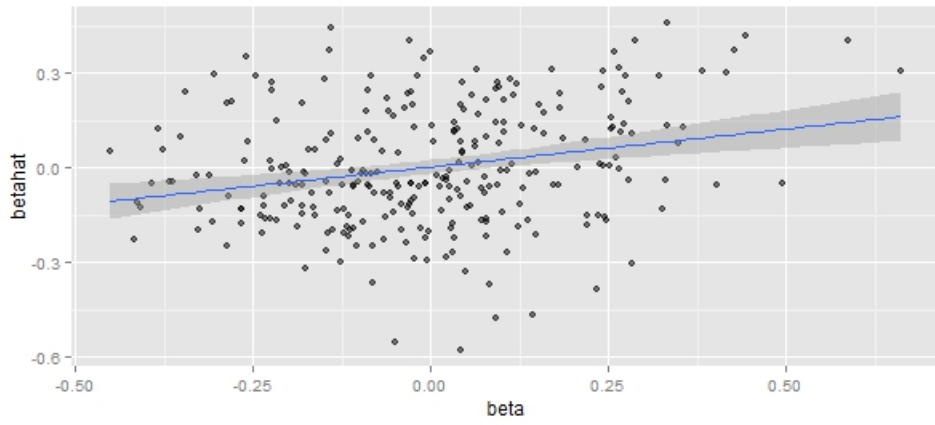


Figure 5.9: Scatter plot β against $\hat{\beta}$

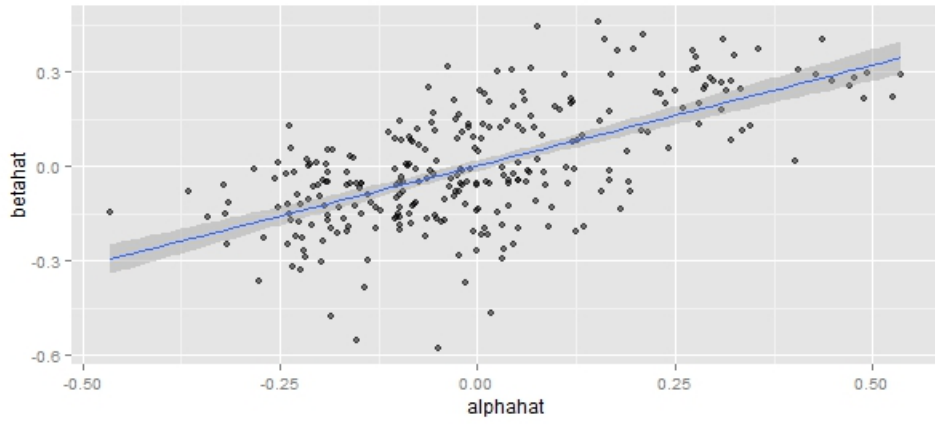


Figure 5.10: Scatter plot $\hat{\alpha}$ against $\hat{\beta}$

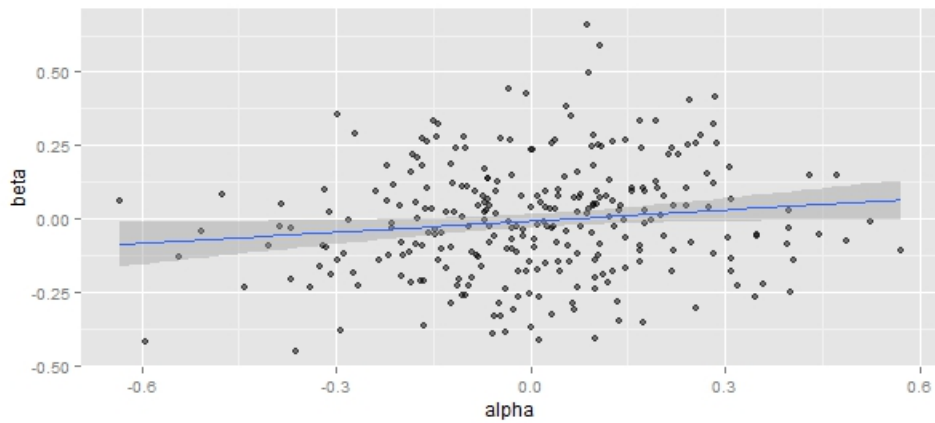


Figure 5.11: Scatter plot α against β

5.3.5 Parameters change over time

A natural progression would be to let the parameters change over time. This would allow for changes in short-term form, such as a player getting injured or suspended, a tightly packed schedule or other temporary changes to a teams performance. It would allow for the recent history of games to be weighted more heavily than those from a long time ago. This would allow for more long-term changes, as when a team buys/sells a player. Letting parameters change a lot during longer breaks should make so I can look at multiple seasons together instead of always starting fresh when a new season begins.

Figure 5.12 shows how Arsenal's parameters change as the season goes on. The calculations start after half the season is finished, but still the values change drastically going through tops and dips. I clearly need a model that accounts for these changes to predict accurately.

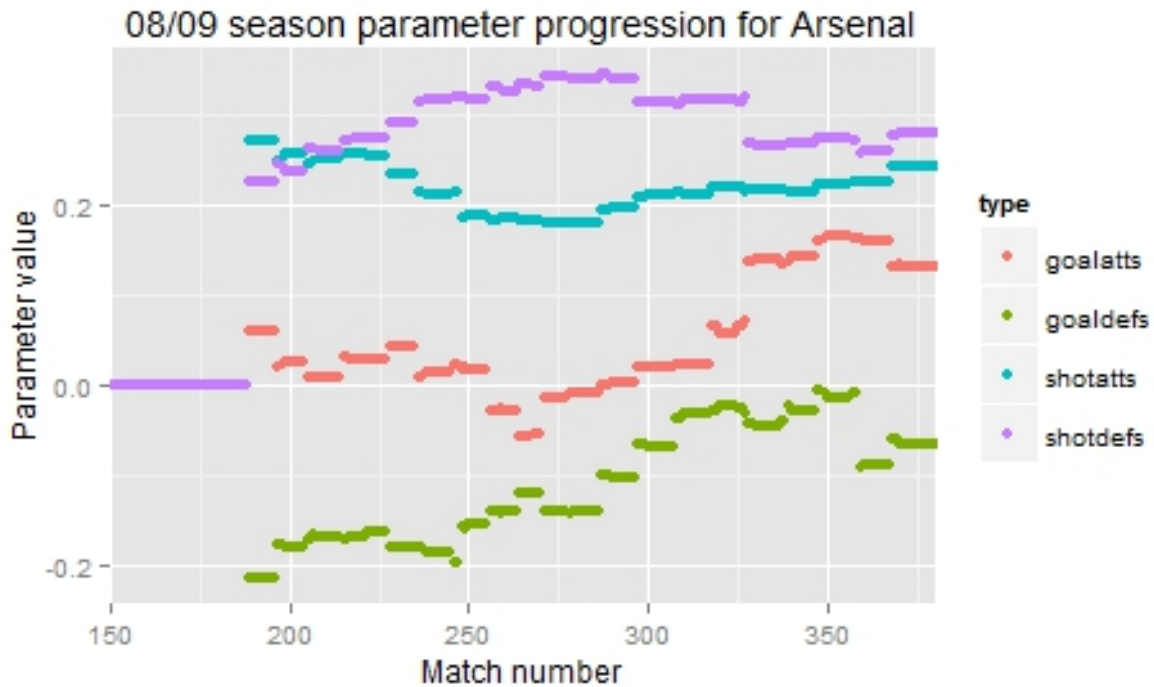


Figure 5.12: Arsenal's parameters over time for season 08/09

5.4 Model 4: Team-specific strengths change over time

As mentioned in the previous chapter, a natural extension of model 3 would be to let the parameters be time-variant. This makes the computation a lot more time demanding, so to keep it at a workable level I find it necessary to introduce certain simplifications. Firstly, the time unit will be the week the match takes place, counting from 0 for the first match of the season. Secondly I no longer impose the constraints $\sum_{i=1}^{20} \alpha_i = 0$, $\sum_{i=1}^{20} \hat{\alpha}_i = 0$, $\sum_{i=1}^{20} \beta_i = 0$ and $\sum_{i=1}^{20} \hat{\beta}_i = 0$. The latter in particular reduces the time of one model fitting from half an hour to about a minute.

The strengths α , $\hat{\alpha}$, β and $\hat{\beta}$ are now modeled as 1st order autoregressive parameters. This means that $\alpha_i = \rho_t \alpha_{i-1} + \epsilon_i$, with $|\rho_t| < 1$ being the time correlation, and ϵ_i the error term $\sim N(0, \tau_t^{-1})$ with precision τ_t , and similarly for the other strengths.

5.4.1 Priors

For the home advantage I keep the same prior as for the previous models, Gaussian with mean 0.28 and precision 50.

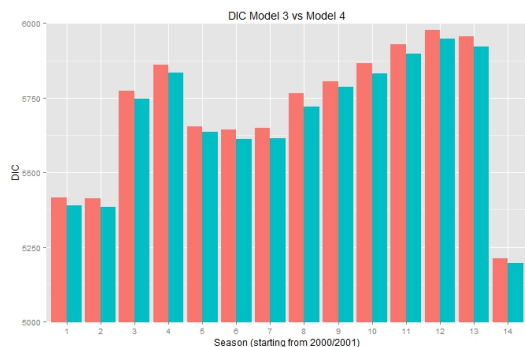
For the parameters that now are autoregressive I use the PC prior for correlation ρ_t , with $(\text{mean}, \alpha) = (0.5, 0.75)$. The previous models have been with $\rho_t = 1$, meaning constant through time, so this prior suggests that $\rho_t > 0.5$ with a 75% certainty. This seems reasonable, as it is to be expected that at least half of the quality in a team is constant within a season. The home advantage remains unchanged, so I use the Gaussian prior with parameters $(\text{mean}, \alpha) = (0.28, 50)$.

5.4.2 Performance with data

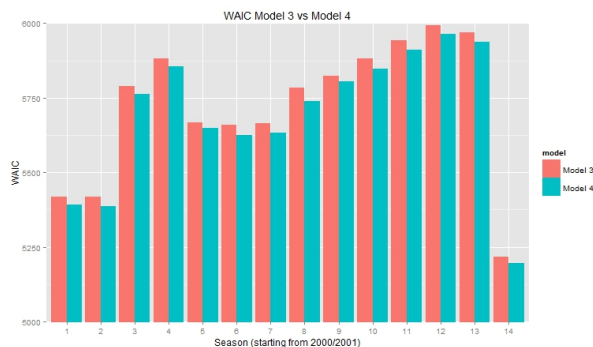
Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
Model4DIC	5390.55	5384.52	5747.46	5834.44	5634.92	5610.59	5615.65	5720.39	5786.37	5831.13	5898.37	5946.69	5920.56	5195.54
Model4WAIC	5391.76	5386.87	5761.46	5853.94	5648.98	5624.09	5632.75	5739.54	5804.29	5847.12	5910.31	5964.49	5936.89	5195.83

Table 5.6: WAIC and DIC for Model 4

The WAIC and DIC values for model 4 are much better than either of the previous models. Table 5.7 shows how much lower the values are for model 4 compared to model 3.



(a) Comparison of DIC values for model 3 and 4



(b) Comparison of WAIC values for model 3 and 4

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
DIC	-24.60	-28.60	-26.30	-27.32	-18.07	-32.42	-32.25	-45.55	-18.93	-35.03	-30.45	-29.52	-34.69	-17.77
WAIC	-26.71	-32.64	-26.80	-28.11	-18.52	-35.14	-31.10	-43.95	-18.15	-34.38	-32.28	-27.43	-32.86	-21.58

Table 5.7: Difference in WAIC and DIC between Model 3 and Model 4. Negative number means Model 4 performs better.

Looking at the values for PL(score) and PL(result), the results are not so clear. Figure 5.17 shows the pseudo likelihood for predicting the correct scoreline, and model 4 is mostly a clear improvement even if it gets beaten by model 1 for some of the seasons. Figure 5.18 shows the pseudo likelihood for predicting the correct result, and in the seven season sample model 4 is only an improvement on models 1-3 in two of the seasons.

5.4.3 Validity of model 4

I've chosen to look at the 2013/2014 season for Liverpool FC, to see if the variations in the strengths agree with the performance of the team. This was an interesting Liverpool season for a lot of reasons. They finished second in the league, scoring 101 goals in the process - the most goals ever scored by a Premier League runner-up. This means they had the second best attack in the league, while also conceding 50 goals making them the leagues 13th worst defense. And interestingly for this analysis, Liverpool went streaks without some of their most important players (Daniel Sturridge and Luis Suarez) because of injuries and bans.

Figure 5.14 shows how the strengths develop in time. The first interesting thing is that the chance creating strength has a steady decline. This could be explained by the fact that Liverpool started the season without Luis Suarez because of a 10-game ban, and this meant they were playing with more traditional midfielders than real forwards. This also explains why the goal converting strength is increasing with time, as having two forwards typically will give you a better finishing rate but lower possession and thereby lower chances created.

Simon Mignolet played as a goalkeeper every single minute of the season, so that explains the goal defending strength being basically constant through time. The chance preventing skill however got much better toward the end of the season. The defense changed a lot during the season, but one clear change is that Jon Flanagan was included as fullback (rotating between a right back and a left back depending on the rest of his team) for most of the games from gameweek 11, and this coincides with the positive change in trend.

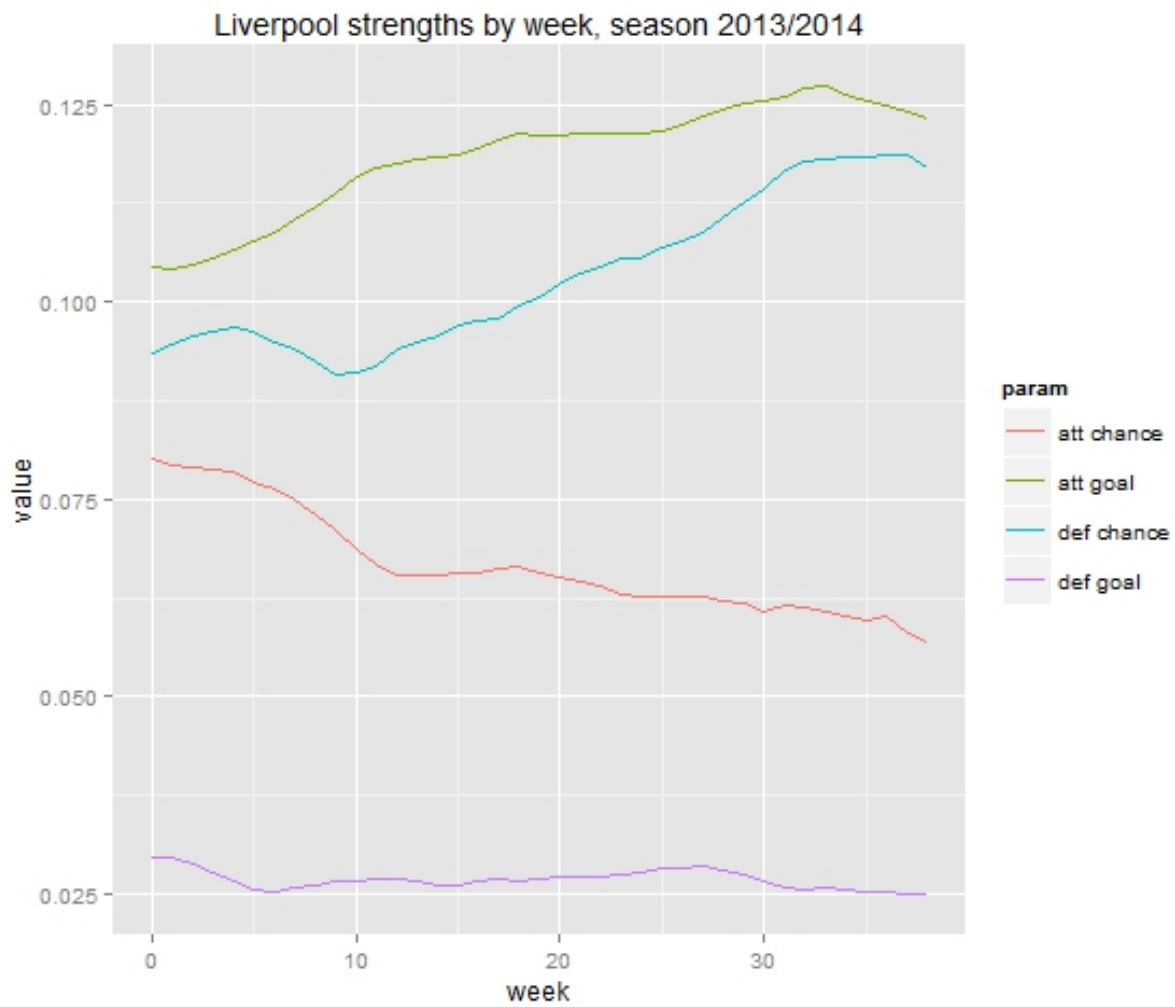


Figure 5.14: Liverpool strengths for the season 2013/2014

5.5 Model 5: Correlation between parameters

Just from looking at the final table it is easy to see that a team with a good offense typically also has a good defense, as these qualities depend on the players of the team which are typically bought from other clubs. As rich clubs tend to do well in all aspects of the game, it is natural to assume that the teams' strengths are all correlated in some way. By saying that the strengths are correlated I'm essentially making them less random by tying them loosely together. This should lower the effective number of parameters in the model.

The parameter correlation ρ is implemented so that for a team A with generic team parameters e and e' (being either α , β , $\hat{\alpha}$ or $\hat{\beta}$) and time t we have that

$$\text{Corr}((e_A, t), (e'_A, t')) = \rho \rho_t^{(t-t')},$$

where $t > t'$.

This means that the correlation ρ is constant across all parameters, so the correlation between $(\beta, \hat{\beta})$ is assumed to be the same as, for instance, (α, β) . This is not ideal, as the former correlation seems to be much stronger than the latter, but helps with keeping computation times down. I've experimented with only having a correlation between $\hat{\alpha}$ and $\hat{\beta}$, as this was supposed to be the strongest of the correlations, but the WAIC/DIC remain basically unchanged and it effectively doubles the runtime while increasing the effective number of parameters.

Looking specifically at the 2008/2009 season estimated by models 4 and 5, the main difference is that model 5 has an extra hyperparameter ρ . The runtime increases from 30 seconds to 90 seconds, while the number of effective parameters according to the WAIC is reduced from 108.62 to 97.5. The correlation is estimated at $\rho = 0.3785$ with a standard deviation of 0.1458, so the correlation is different from zero significant on a 95% level, but the variation is still a bit high.

5.5.1 Performance with data

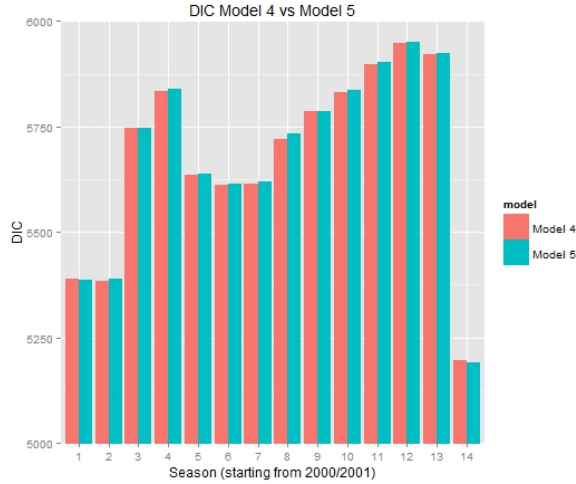
Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
Model5DIC	5388.10	5389.43	5747.65	5838.26	5637.86	5614.03	5620.89	5733.50	5786.78	5837.54	5901.85	5950.37	5923.39	5190.52
Model4WAIC	5389.31	5392.31	5759.93	5856.25	5651.06	5625.65	5637.25	5749.55	5803.26	5852.70	5913.36	5967.41	5939.81	5190.99

Table 5.8: WAIC and DIC for Model 5

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
DIC	-2.44	4.91	0.18	3.82	2.94	3.44	5.25	13.11	0.41	6.42	3.48	3.68	2.84	-5.02
WAIC	-2.45	5.45	-1.54	2.31	2.07	1.56	4.51	10.01	-1.03	5.57	3.05	2.93	2.91	-4.84

Table 5.9: Difference in WAIC and DIC between Model 4 and Model 5. Negative number means Model 5 performs better.

Compared to model 4, model 5 performs at the same level or slightly worse in WAIC and DIC. Figures 5.17 shows that model 5 outperforms model 4 at predicting the correct score every season. For predicting the result, figure 5.18 shows that model 5 is the better model at 4 out of 7 seasons, and on average it does much better.



(a) Comparison of DIC values for model 4 and 5



(b) Comparison of WAIC values for model 4 and 5

5.6 Model 6: Effect of red cards included

Receiving a red card puts a team at a serious disadvantage, so it's natural to assume that this would improve the descriptive and predictive properties of my model. As a red card only appears in about one out of twenty matches, I've chosen to not use it for simulation and simply assume that the game being simulated will not have any red cards in it.

To do this I simply add to the model four new parameters describing the effects of red cards on the defense and offense, and I let these be constant across all teams. The model then becomes

$$\ln(\hat{\lambda}_{i,j}) = \hat{\alpha}_0 + \hat{\alpha}_i - \hat{\beta}_j + \hat{\theta}_\alpha r_i + \hat{\theta}_\beta r_j + \hat{\delta}$$

$$\ln(\hat{\mu}_{i,j}) = \hat{\alpha}_0 + \hat{\alpha}_j - \hat{\beta}_i + \hat{\theta}_\alpha r_j + \hat{\theta}_\beta r_i$$

$$p = \frac{1}{1 + \exp(-\alpha_0 - \alpha_i + \beta_j - \theta_\alpha r_i - \theta_\beta r_j)}, \quad q = \frac{1}{1 + \exp(-\alpha_0 - \alpha_j + \beta_i - \theta_\alpha r_j - \theta_\beta r_i)},$$

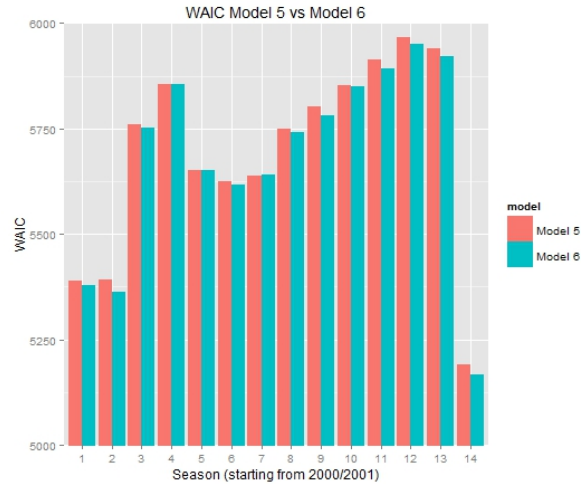
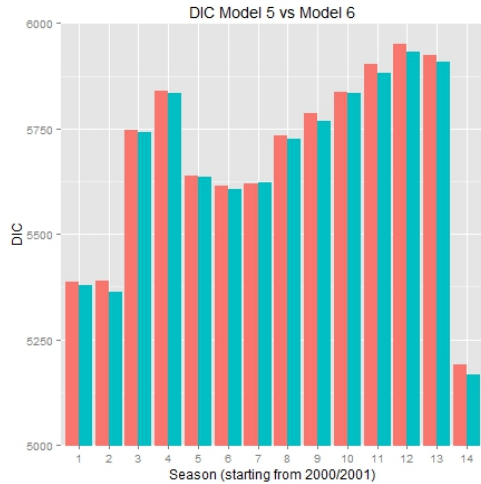
where $\hat{\theta}_\alpha$ is the effect on the chance creating ability, $\hat{\theta}_\beta$ is the effect on the chance defending ability, θ_α is the effect on the goal converting ability and θ_β is the effect on the goal stopping ability. r_i and r_j are the red cards received by the home team and the away team, respectively.

5.6.1 Performance with data

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
Model6DIC	5378.55	5362.83	5740.52	5835.15	5636.48	5605.61	5623.52	5726.18	5766.76	5834.32	5881.11	5933.20	5908.36	5166.92
Model6WAIC	5378.98	5363.90	5752.68	5855.00	5650.55	5617.15	5640.58	5742.75	5781.14	5849.60	5891.24	5949.69	5922.13	5166.59

Table 5.10: WAIC and DIC for Model 6

For WAIC and DIC, model 6 is an improvement on every other model for every season except for the 2006/2007 season where model 4 does even better. Despite this, these are very strong results for model 6.



(a) Comparison of DIC values for model 5 and 6

(b) Comparison of WAIC values for model 5 and 6

Season	00/01	01/02	02/03	03/04	04/05	05/06	06/07	07/08	08/09	09/10	10/11	11/12	12/13	13/14
DIC	-9.56	-26.60	-7.13	-3.11	-1.38	-8.41	2.63	-7.32	-20.03	-3.23	-20.74	-17.18	-15.03	-23.60
WAIC	-10.33	-28.41	-7.25	-1.25	-0.50	-8.49	3.33	-6.80	-22.12	-3.10	-22.12	-17.72	-17.68	-24.41

Table 5.11: Difference in WAIC and DIC between Model 5 and Model 6. Negative number means Model 6 performs better.

For PL(score), model 6 does generally well except for season 2005/2006 where model 1(!) performs the best. All other seasons model 6 is either the best or among the best.

For PL(result), model 6 again has strong results but is inexplicably the worst model in 2011/2012. Overall though it seems to be one of the best, if not the best.

Summary Model 4-6

Figures 5.20 and 5.19 are exactly what I was hope to achieve. The models are consistently getting better WAIC and DIC results for all seasons as I make them more complex, and is a good indication that using all this extra data is actually worth it.

Figures 5.17 and 5.18 are not as easy to interpret. For some reason model 1 does really well for a lot of the seasons, especially season 2010/2011 where it is the clear winner at predicting results. Model 3 does very well some seasons and very poorly in others. And the last three models are rated similarly, though model 6 seems to come out on top.

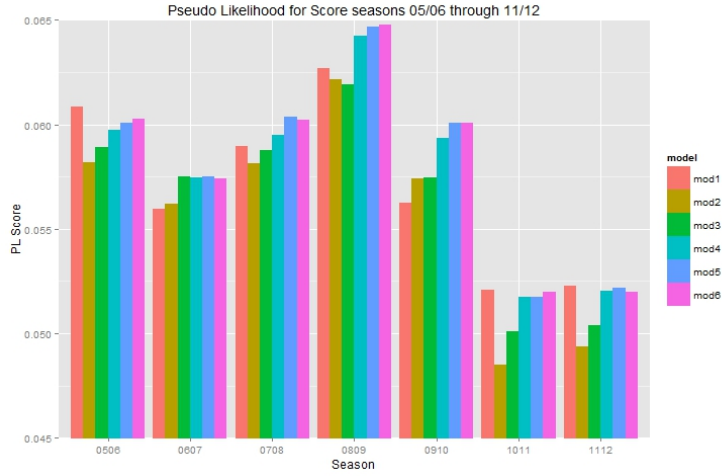


Figure 5.17: PL(score) (y-axis) seasons 05/06 through 11/12 all models

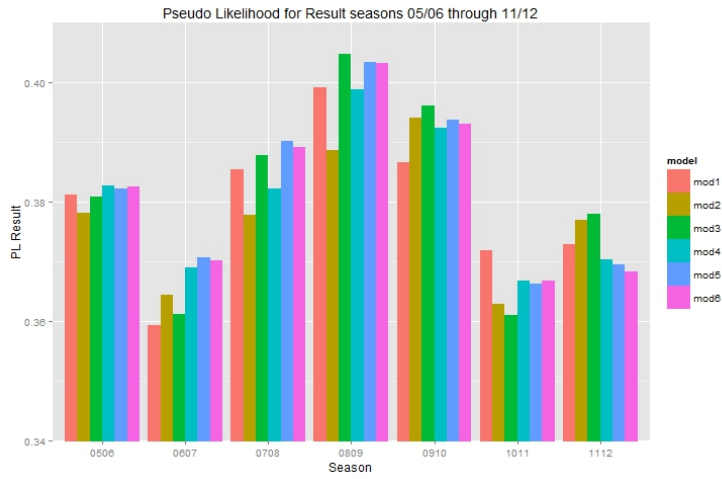


Figure 5.18: PL(result) (y-axis) seasons 05/06 through 11/12 all models

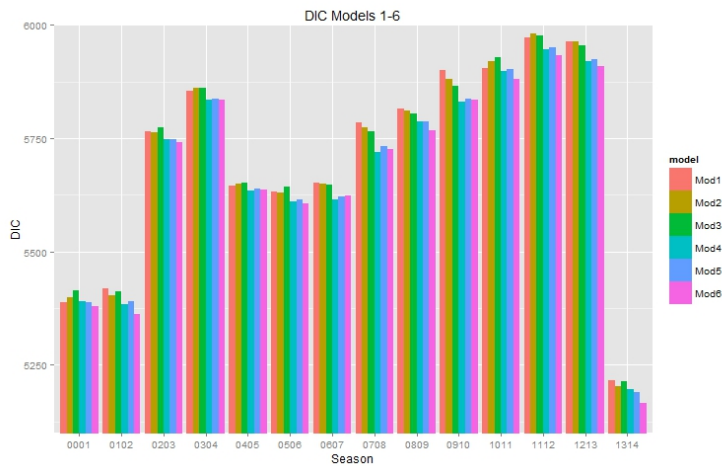


Figure 5.19: DIC values for all models, all seasons

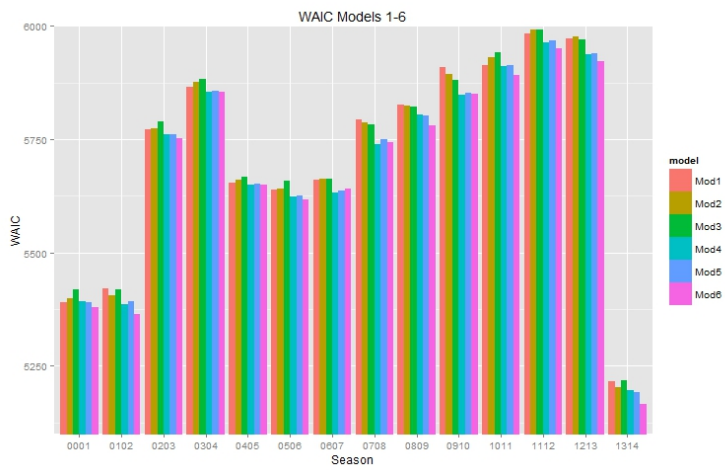


Figure 5.20: WAIC values for all models, all seasons

Chapter 6

The prediction model

The previous two sections have concluded with model 6 being the best at both describing the data (WAIC/DIC) and predicting unseen results (PL(score) and PL(result)). In this section I want to describe the model as a whole and then determine whether or not using SOT gives better results than shots.

6.1 Chance model

Chances are Poisson distributed with interaction models

$$\ln(\hat{\lambda}_{i,j}) = \hat{\alpha}_0 + \hat{\alpha}_i - \hat{\beta}_j + \hat{\theta}_\alpha r_i + \hat{\theta}_\beta r_j + \hat{\delta}$$

$$\ln(\hat{\mu}_{i,j}) = \hat{\alpha}_0 + \hat{\alpha}_j - \hat{\beta}_i + \hat{\theta}_\alpha r_j + \hat{\theta}_\beta r_i$$

Together they follow a Poisson distribution,

$$P(\hat{X}_{i,j} = x, \hat{Y}_{i,j} = y | \lambda; \mu) = e^{-(\lambda+\mu)} \frac{\lambda^x \mu^y}{x! y!}, \quad (6.1)$$

for $\lambda = \hat{\lambda}_{i,j}$ and $\mu = \hat{\mu}_{i,j}$.

As this is a Poisson distribution we have that $E(\hat{X}_{i,j}) = Var(\hat{X}_{i,j}) = \hat{\lambda}_{i,j}$ and $E(\hat{Y}_{i,j}) = Var(\hat{Y}_{i,j}) = \hat{\mu}_{i,j}$.

6.2 Goal model

Goals are Binomially distributed with interaction models

$$p_{i,j} = \frac{1}{1+\exp(-\alpha_0-\alpha_i+\beta_j-\theta_\alpha r_i-\theta_\beta r_j)}, \quad q_{i,j} = \frac{1}{1+\exp(-\alpha_0-\alpha_j+\beta_i-\theta_\alpha r_j-\theta_\beta r_i)}$$

$X_{i,j}$ and $Y_{i,j}$ are two independent Binomially distributed random variables,

$P(X_{i,j} = x | n = \hat{X}_{i,j}) = \binom{n}{x} p^x (1-p)^{n-x}$, and likewise $P(Y_{i,j} = y | n = \hat{Y}_{i,j}) = \binom{n}{y} q^y (1-q)^{n-y}$ with

$$E(X_{i,j}) = \hat{X}_{i,j}p_{i,j}, \text{Var}(X_{i,j}) = \hat{X}_{i,j}p_{i,j}(1 - p_{i,j})$$

and

$$E(Y_{i,j}) = \hat{Y}_{i,j} \cdot q_{i,j}, \text{Var}(Y_{i,j}) = \hat{Y}_{i,j} \cdot q_{i,j}(1 - q_{i,j}).$$

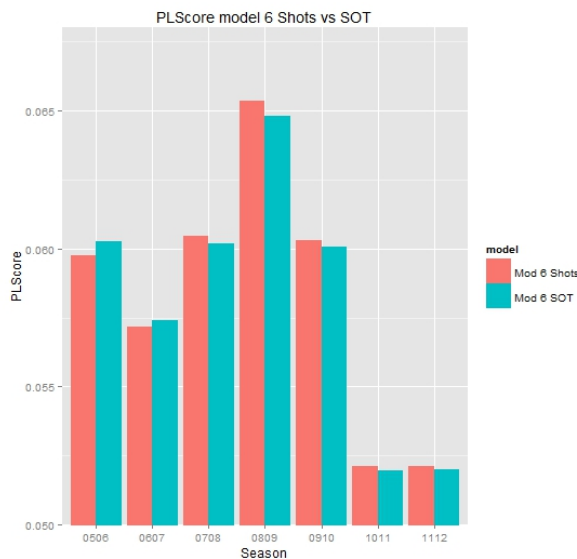
6.3 Parameter Properties

The parameters $\hat{\alpha}$, $\hat{\beta}$, α and β are first order auto-regressive with time correlation ρ_t , and group correlated with correlation ρ . The fixed effects $\hat{\delta}$, $\hat{\alpha}_0$, $\hat{\theta}_\alpha$, $\hat{\theta}_\beta, \alpha_0$, θ_α and θ_β are determined in the model. The model has four hyperparameters - ρ , ρ_t , its precision τ_t , and the precision for the home field advantage τ_δ .

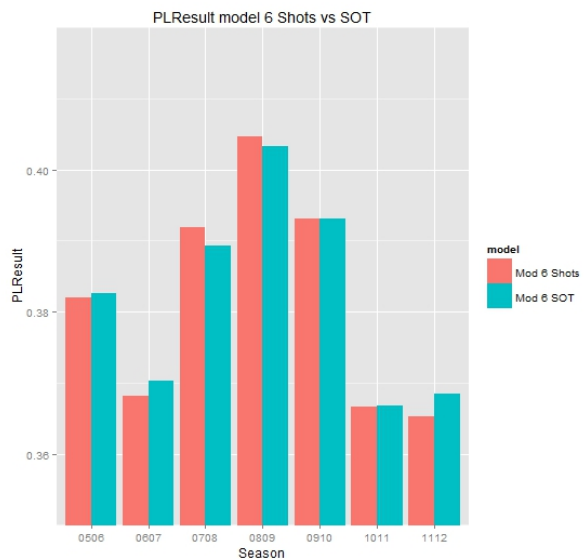
6.4 Chances - Shot or SOT?

The last thing to decide is whether to use shots on SOT as a chance. There are more shots than shots on target, so that gives us more data and probably lower uncertainty. But shots can be taken from anywhere and may have an incredibly small chance of becoming a goal depending on where it is taken from. A SOT thereby probably has a more consistent connection to goals scored.

I've run the second half Pseudo-Likelihood prediction analysis for model 6 using shots and SOT. Figures 6.1a and 6.1b show the results where the model with SOT is in blue and shots is in red. SOT does better at PL(result) in 5 of 7 seasons, while shots does better at PL(score) in 5 of 7 seasons. This is a somewhat ambiguous, but as I'm primarily interested in predicting match results it seems like using SOT is the best option.



(a) PL(Score) for model 6 with shots and SOT



(b) PL(result) for model 6 with shots and SOT

Chapter 7

Prediction model vs betting companies

The best way of testing a prediction model is to see if it's able to gain a profit on the betting market. This would mean that the model is significantly better at predicting than the models used by the betting companies, and most importantly if it consistently beats the odds then it would give you a potentially bottomless source of income.

For gamblers there is a myriad of options for placing bets - this could be anything from betting on which team will receive the most yellow cards to specific players scoring from specific positions or using specific body parts. Most frequently the bets revolve around predicting whether the game will be a home victory (H), a draw (D) or an away victory (A). The betting company will decide their odds based on what they consider the probabilities of the different outcomes to be, and they will make the odds slightly worse so that their expected return on any bet is above zero (a betting margin). For a given match between two teams with odds (H,D,A), the number $\frac{1}{H}$ is essentially the probability that the home team will win according to the model of the betting company, though with a small padding. For example, a game between Chelsea and Burnley is given the odds $(H, D, A) = (1.20, 6.75, 17.00)$. If this bet were "fair" then $\frac{1}{H} + \frac{1}{D} + \frac{1}{A} - 1$ would be exactly 0. In this case $\frac{1}{1.20} + \frac{1}{6.75} + \frac{1}{17} - 1 = 0.0403$, meaning the company has chosen odds such that if their probabilities of (H,D,A) are correct then their expected return should be about 4%. For the season of 2013-2014 the average betting margin for all games across more than 35 bookmakers was 5.5%, using data collected by www.betbrain.com.

If there were no betting margin a bettor could place money on whichever result and the expected return would be 0. The goal of the bettor is therefore to calculate the probabilities of each outcome that much better than the bookmaker such that the expected return is above zero. To make this slightly safer, I'll require my expected return to be above a certain cutoff ω , and I'll experiment with different values of this cutoff to maximize the return.

Of course if the bettors largely favor the odds of one of the options then the betting company could still end up losing money. As a result the odds are typically adjusted over time so that the betting company ends up winning no matter what result comes in. This is actually in the favor of the bettor as it can skew the odds so that they are no longer optimally decided.

Luckily a bettor is not constricted to dealing with only one betting company, which lets us bet using the highest odds we can find. There are also betting promotions that may be exploited for further improved returns, and various interesting betting strategies could conceivably be profitable.

7.1 The betting model

My betting model is fairly simple; simulations are run in the same way as for calculating the second half Pseudo-Likelihood for predictions, using the first half of the season only for training the model. This gives me a matrix with values (H,D,A) denoting my simulated probabilities of the different outcomes for each match. For cutoff value ω , I put money on every bet that satisfies the requirement

$$p(r) * odds(r) - 1 > \omega, \quad (7.1)$$

where $p(r)$ is my estimated probability of a result occurring and $odds(r)$ is the odds given by a betting company for a result.

The next question is how much money to bet - the risk strategy. One option (flat) is to just put the same amount on every single bet, but this ends up being very volatile as a lot of the bets that qualify the above requirement are bets with very high odds. Koopman, Lit(2012)[10] suggested lowering the bet from 1.0 to 0.3 for all bets with odds higher than 7. My solution (scaled) has been to always bet $\frac{1}{odds(r)}$, as this risk adjustment scales with increasing odds, and also has the attractive property that winning a bet always rewards 1.0 "money". In the long run any betting strategy with an expected return > 0 should be fruitful, but this lowers the chance of bankruptcy which would be an absorbing state for the bettor. I've decided to also compare how methods flat and scaled differ in results.

7.2 Results from betting

I've run betting for all seasons from 07/08 until 11/12 with both shots and SOT used as chances, with both risk strategies flat and scaled, and with cutoff values ω from 0 to 1 with 0.01 increments. I had to divide this all into 5 different figures, so figures 7.1 to 7.5 all show the betting results for the seasons from 07/08 until 11/12. The colors represents the combination of risk strategy and whether shots or SOT are being used - for instance the purple line (scaledSOT) shows the betting results using SOT and the scaled risk strategy. As the scaled versions have variable bet size, I've normalized the results for those models so that the mean bet size is the same for all models (=1).

If you want to use this model on the betting market you can't retroactively choose the options that give the best results, so we need to make a selection that is consistently among the best for all seasons. This is surprisingly hard as different choices seem to do better for different seasons. In general, Scaled-SOT and Scaled-shots do the best, with Scaled-SOT being the overall winner. As for the choice of cutoff value there are a few candidates. Figure 7.6 shows how the final balance varies depending on the choice of ω . A higher cutoff value means placing bets that the model feels are safer, which reduces the amount of bets placed and centralizes the balance around zero. There are still a few standout candidates for the best cutoff value - which are even clearer from looking at figure 7.7 showing the mean balance by cutoff value. The two best values seem to be $\omega = 0.11$ and $\omega = 0.34$.

Figure 7.8 shows how the amount of bets placed by the models (SOT and shot) in season 10/11 depends on the cutoff value chosen. For $\omega = 0$ with shots, a bet is placed on essentially every possible match (203 out of 208). For $\omega = 0.10$ this is lowered to 150 bets placed. For $\omega = 0.60$ about 30 bets are placed. For $\omega = 0.96$ and above, only 9 bets are placed. Shots and SOT will typically choose the same candidate, but they value the bets slightly differently which keeps the two graphs from overlapping perfectly.

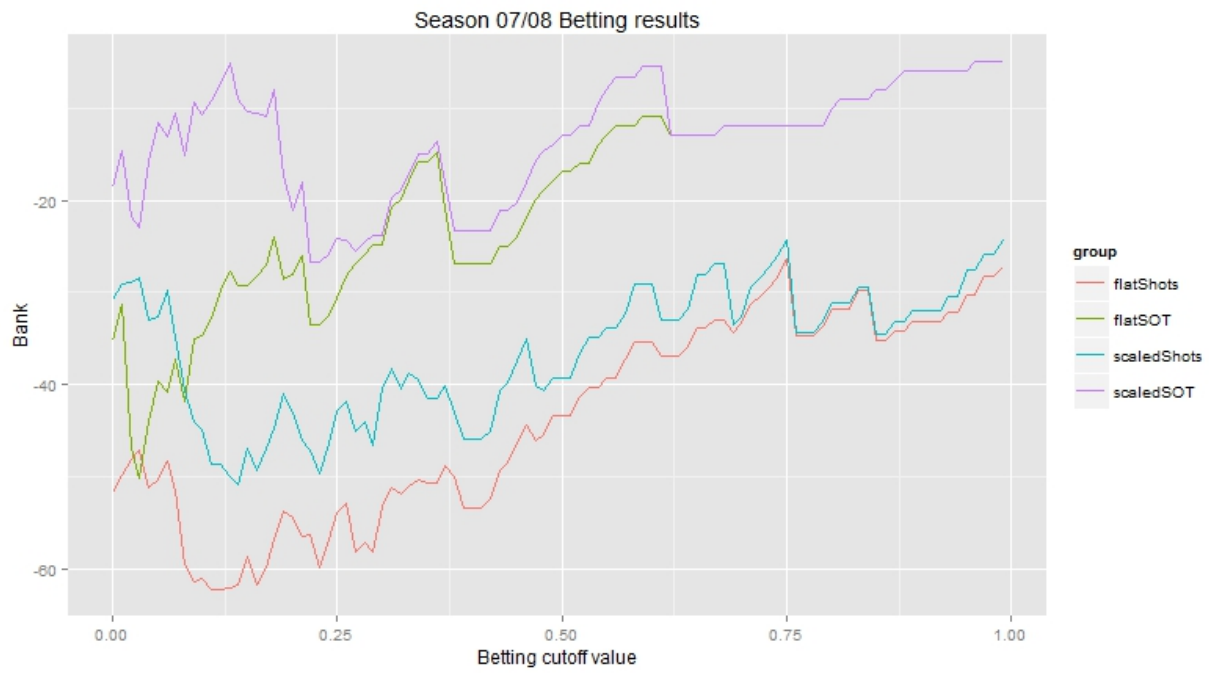


Figure 7.1: Betting results season 07/08

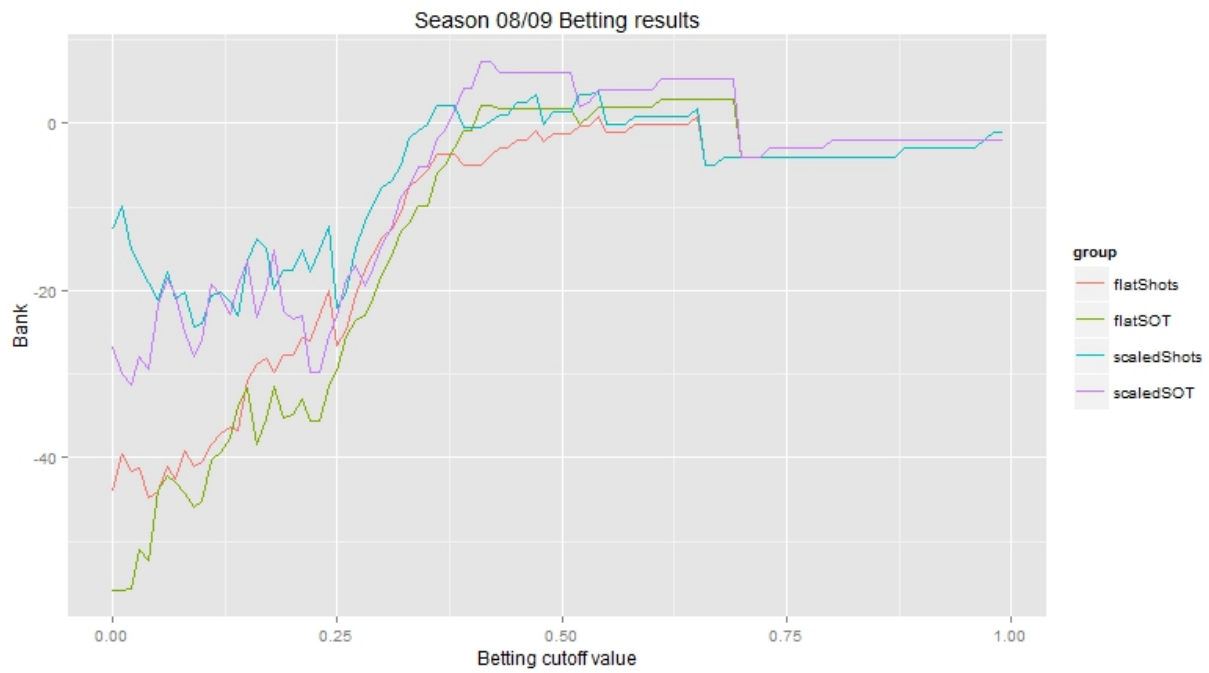


Figure 7.2: Betting results season 08/09

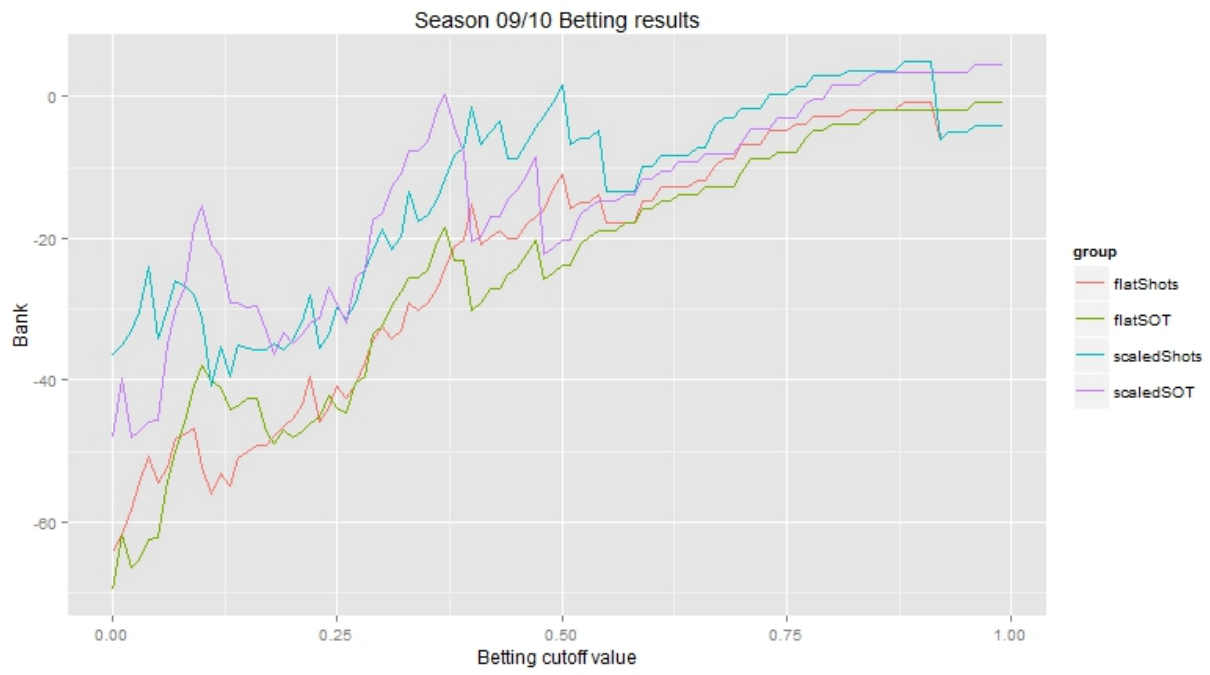


Figure 7.3: Betting results season 09/10

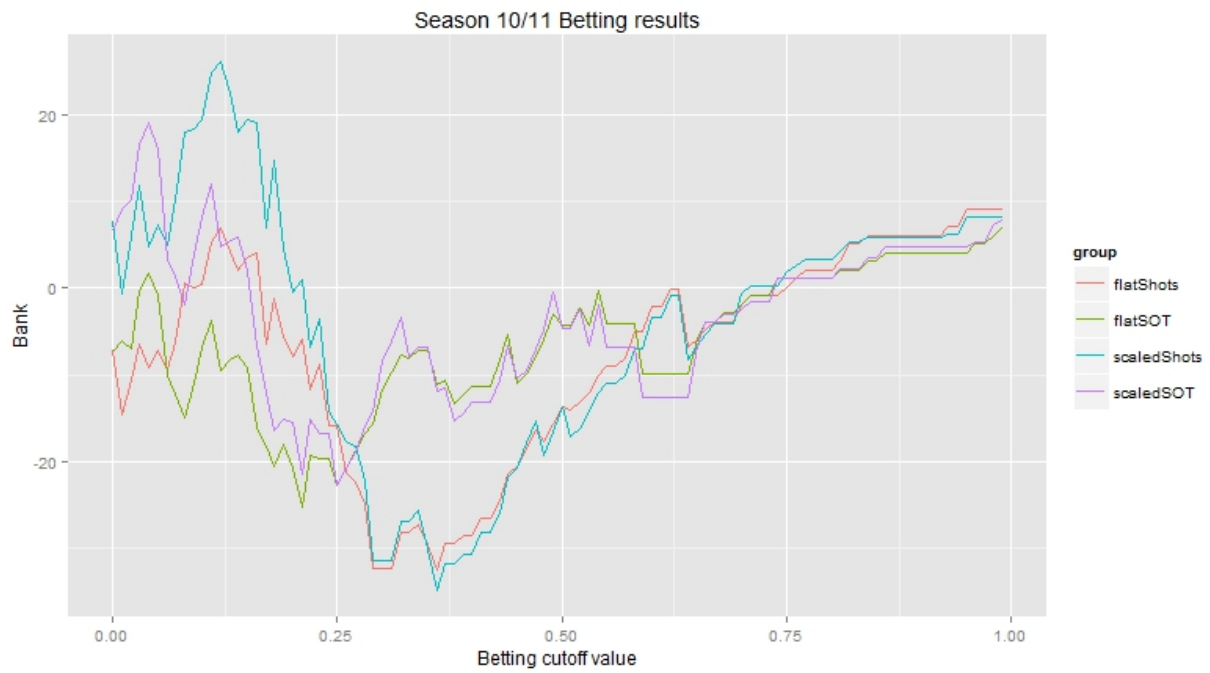


Figure 7.4: Betting results season 10/11

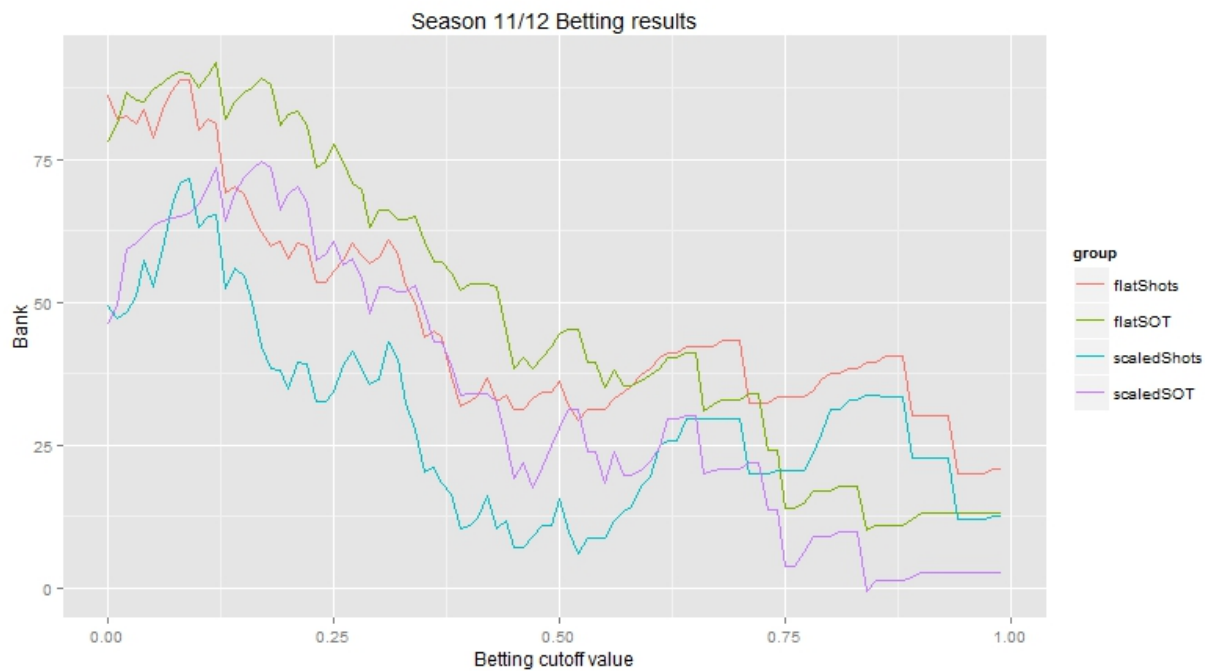


Figure 7.5: Betting results season 11/12

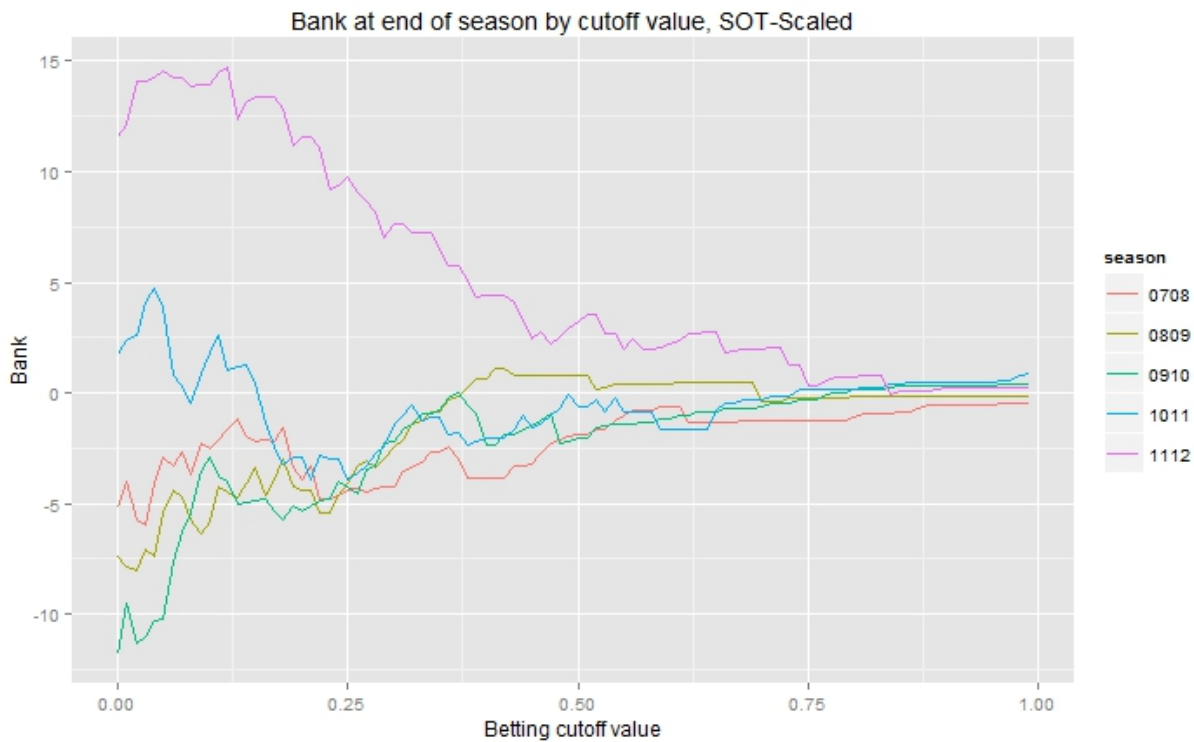


Figure 7.6: End of season balance for different cutoff values ω , all seasons, Scaled-SOT

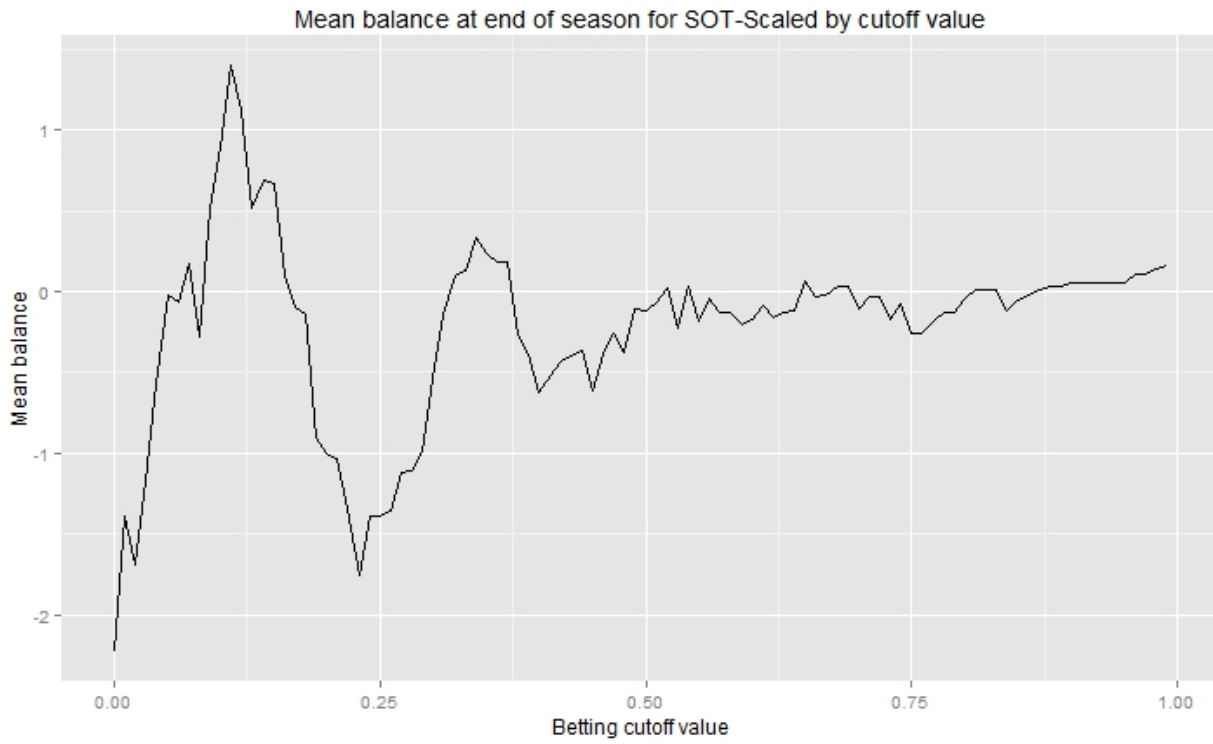


Figure 7.7: Mean end of season balance for all seasons, Scaled-SOT by cutoff value ω

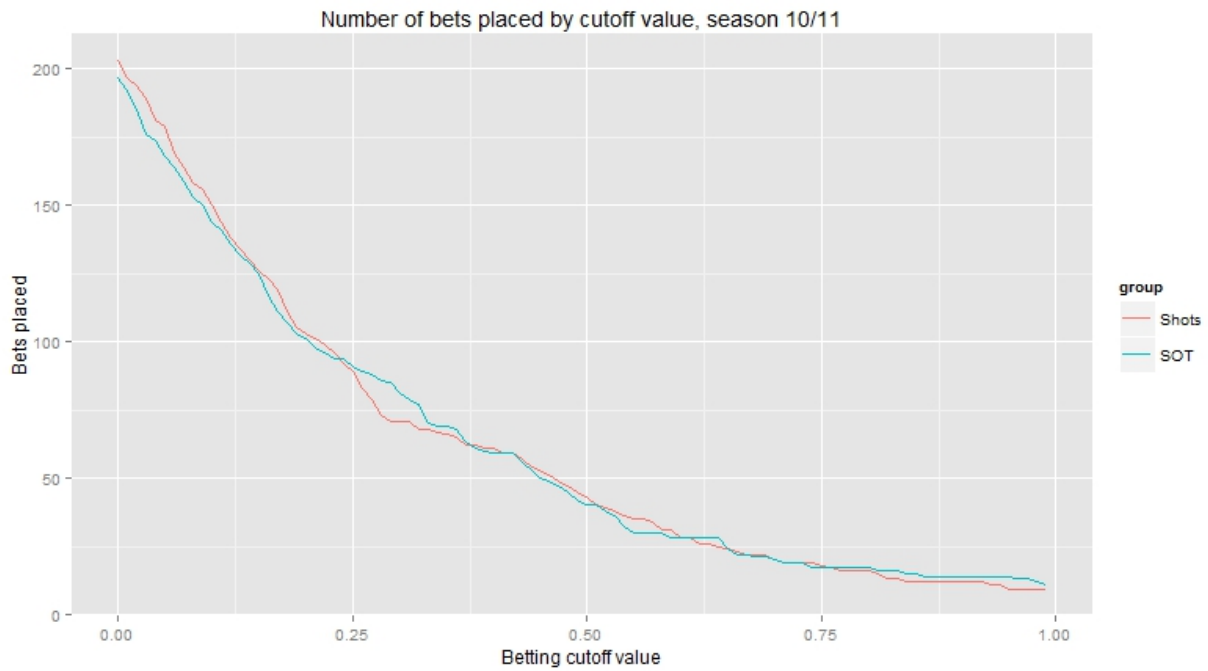


Figure 7.8: Number of bets placed by cutoff value, season 10/11

Chapter 8

Results and discussion

8.1 Results

For the results section I thought it would be interesting to look at an entirely unseen season: the 2014/2015 season of Premier League. This season saw a dominant Chelsea at times struggle with creativity, often scraping by with narrow 1 goal wins, but only losing three games. Southampton started off looking like a surprise candidate for placing top 4, but slowly drifted down to seventh place being replaced by the usual suspects. Manchester United returned to at least some of its old glory after a disastrous season 13/14. And Liverpool struggled for consistency despite spending more than 110 million GBP on new players.[21]

After all teams had played half of their games the table looked like figure 8.1. Figures 8.2a and 8.2b show the predicted final table and the actual final table. The two tables are quite similar, but according to the model teams like Queens Park Rangers and Aston Villa underperformed while Leicester City and Crystal Palace got more points than they deserved. Within the top 8 the prediction is largely in line with reality, the exception being West Ham United that got 10 points less than they "deserved" and lost 4 places because of that.

Figure 8.3 shows how the model does with betting on season 14/15 for cutoff values $\omega = 0.11$ and $\omega = 0.34$. $\omega = 0.11$ places 115 bets and is the only one that actually has a positive balance at some point, but $\omega = 0.34$ placing 36 bets does better in the end.



POS		LP	CLUB	P	W	D	L	GF	GA	GD	PTS
1	—	(1)	Chelsea	18	14	3	1	40	13	27	45
2	—	(2)	Manchester City	18	13	3	2	39	15	24	42
3	—	(3)	Manchester United	18	10	5	3	33	19	14	35
4	▲	(5)	Southampton	18	10	2	6	31	14	17	32
5	▼	(4)	West Ham United	18	9	4	5	29	21	8	31
6	—	(6)	Arsenal	18	8	6	4	32	22	10	30
7	—	(7)	Tottenham Hotspur	18	9	3	6	24	24	0	30
8	—	(8)	Swansea City	18	8	4	6	23	19	4	28
9	▲	(10)	Liverpool	18	7	4	7	22	24	-2	25
10	▼	(9)	Newcastle United	18	6	5	7	19	26	-7	23
11	▲	(13)	Stoke City	18	6	4	8	19	23	-4	22
12	▼	(11)	Everton	18	5	6	7	27	28	-1	21
13	▼	(12)	Aston Villa	18	5	5	8	11	22	-11	20
14	—	(14)	Sunderland	18	3	10	5	16	27	-11	19
15	—	(15)	West Bromwich Albion	18	4	5	9	18	26	-8	17
16	—	(16)	Queens Park Rangers	18	5	2	11	21	34	-13	17
17	▲	(19)	Hull City	18	3	7	8	18	25	-7	16
18	▼	(17)	Crystal Palace	18	3	6	9	20	30	-10	15
19	▼	(18)	Burnley	18	3	6	9	12	27	-15	15
20	—	(20)	Leicester City	18	2	4	12	16	31	-15	10
POS		LP	CLUB	P	W	D	L	GF	GA	GD	PTS

Figure 8.1: Table after half of season 14/15. Source: www.premierleague.com

teams	points
1 Chelsea	84.09
2 Man City	81.69
3 Man United	68.27
4 Arsenal	67.29
5 Southampton	66.41
6 Tottenham	58.27
7 Liverpool	57.24
8 West Ham	57.05
9 Swansea	52.01
10 Everton	50.41
11 Newcastle	48.94
12 Stoke	48.47
13 Aston Villa	42.17
14 Crystal Palace	40.89
15 Hull	39.57
16 West Brom	39.06
17 Sunderland	38.99
18 Queens Park Rangers	37.48
19 Burnley	33.70
20 Leicester	32.83

POS	LP	CLUB	P	W	D	L	GF	GA	GD	PTS
1	(1)	Chelsea	38	26	9	3	73	32	41	87
2	(2)	Manchester City	38	24	7	7	83	38	45	79
3	(3)	Arsenal	38	22	9	7	71	36	35	75
4	(4)	Manchester United	38	20	10	8	62	37	25	70
5	(6)	Tottenham Hotspur	38	19	7	12	58	53	5	64
6	(5)	Liverpool	38	18	8	12	52	48	4	62
7	(7)	Southampton	38	18	6	14	54	33	21	60
8	(8)	Swansea City	38	16	8	14	46	49	-3	56
9	(9)	Stoke City	38	15	9	14	48	45	3	54
10	(12)	Crystal Palace	38	13	9	16	47	51	-4	48
11	(10)	Everton	38	12	11	15	48	50	-2	47
12	(11)	West Ham United	38	12	11	15	44	47	-3	47
13	(13)	West Bromwich Albion	38	11	11	16	38	51	-13	44
14	(14)	Leicester City	38	11	8	19	46	55	-9	41
15	(17)	Newcastle United	38	10	9	19	40	63	-23	39
16	(15)	Sunderland	38	7	17	14	31	53	-22	38
17	(16)	Aston Villa	38	10	8	20	31	57	-26	38
18	(18)	Hull City	38	8	11	19	33	51	-18	35
19	(19)	Burnley	38	7	12	19	28	53	-25	33
20	(20)	Queens Park Rangers	38	8	6	24	42	73	-31	30

(a) Predicted final table, season 14/15

(b) Actual final table, season 14/15. Source: www.premierleague.com

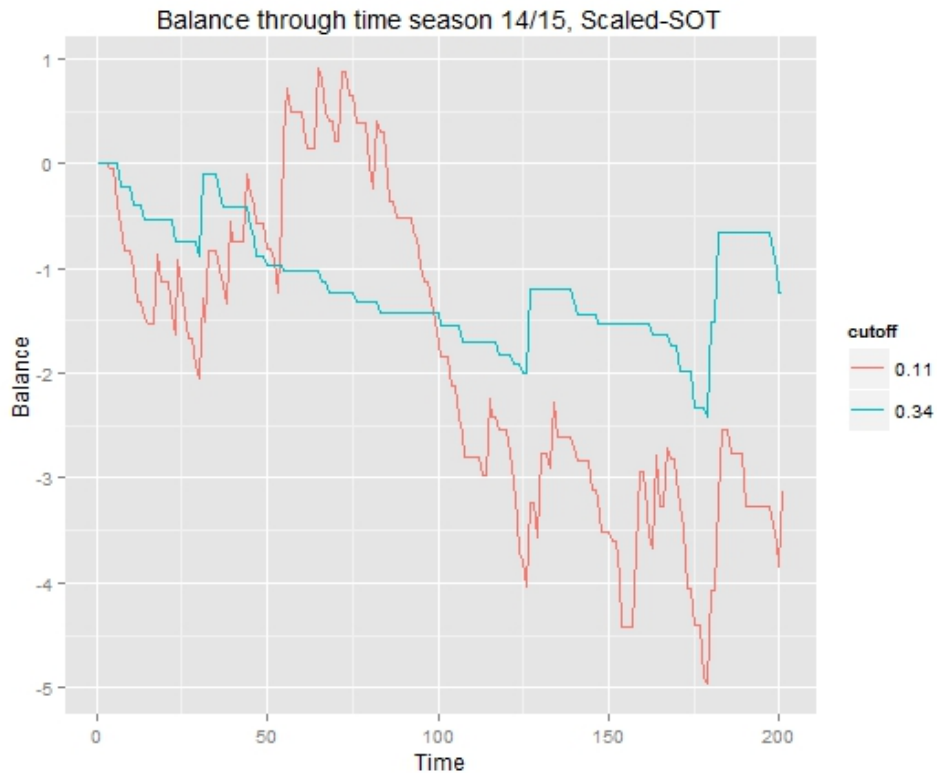


Figure 8.3: Season 14/15 Balance through time, Scaled-SOT with $\omega = 0.11$ and $\omega = 0.34$

8.2 How to improve the model

8.2.1 Initial value for teams

I mentioned in chapter 7 how increasing time used to train the model further could improve the results. This was because the model seemed to be better at betting after about three quarters of the season had been played. The problem with this is that it reduces the amount of games we are able to bet on. A more sensible approach is to include an initial value for the teams in the informative priors, so that the model already has an idea to what the strengths of the teams are relative to each other. Different ways of implementing this was mentioned in my summary of Owen (2011) [15], but because of time limitations I was not able to implement either of these. This could potentially make a huge improvement, especially if the model were to be used before the halfway of the season had been played.

8.2.2 Home Field Advantage for goal converting or goal preventing

Playing on home ground appeared to have some positive effect for converting shots and SOT into goals, as the goal:shot and goal:SOT ratios are slightly higher for the home team. I chose to ignore this effect as it seemed not to be very large, but including it in the model and running prediction is the only way to test if it actually could improve performance. As the models becomes increasingly accurate, an effect like this that at first seemed irrelevant could turn out to be significant enough to warrant including, and it would only add one more parameter so the decrease in performance is negligible.

8.2.3 Different correlations between parameters

A limitation brought up in chapter 5 is that in moving from model 4 to model 5 I enforce the same correlation between all parameters - despite the fact that certain parameters were showed to have stronger relations than others in chapter 3. In adding the correlation I limit how freely two parameters are allowed to move away from each other, which is great for the connection between $\hat{\alpha}$ and $\hat{\beta}$ which clearly are strongly correlated, but not so great when non-correlated parameters like α and β are influenced to move in the same direction. I attempted to keep the correlations specific between certain parameters, but the implementation proved problematic and the result was mostly just a big increase in computation time. If anyone were to properly implement this, however, I believe it could make a big improvement for the model.

8.2.4 Including non-quantitative information

The fact that this model only used quantitative data for prediction is both a strength and a weakness. In one way it keeps the model objective and unbiased, and any long term changes in a team should anyway be caught up in the changing parameters. However, some times qualitative information could greatly impact how the odds are seen. For instance, a team totally reliant on a single player the way

Liverpool was dependent on Suarez for the 2013/2014 season would be severely crippled if that player were to get injured or suspended, and the model would not pick up on that instantly.

As a hypothetical example, say that Wayne Rooney (an important player for Manchester United, a strong team) gets a red card in the very end of a game versus Arsenal (another strong team) for violent misconduct - resulting in a three game ban. Now if Manchester United were to win this game, their strengths would probably be valued very highly as they beat a strong team despite receiving a red card. In reality their value should be lowered as they just lost their best player for the next three games. So for these games, Manchester United will probably struggle and the model will slowly catch up and eventually lower their strengths. Now after these three games of struggling Wayne Rooney will be back which should give them a big boost, but they are now rated poorly and will probably perform at a level better than what the model considers them, and again the model will spend a few games to catch up.

This whole period the model is rating Manchester United wrongly, and because of that it will recommend bets based on faulty information, bets you are likely to lose. Including qualitative information like that directly into the model seems tricky, but I believe it is important to look at what the model provides in the context of the real world instead of blindly following its recommendations.

8.2.5 Other risk strategies for betting

I chose here to go with a scaled betting model, a strategy where the bets placed are inversely proportional to the odds. The ideal goal is to have steady growth, not volatile spikes in balance. Koopman, Lit (2012) [10] lowered the bets from 1.0 to 0.3 for all bets with odds higher than seven, but there are other possible approaches. Betting more on bets where the expected return is greater seems like a logical strategy, though it might favor high-odds bets too much as they tend to be unnaturally high to draw in bettors. Perhaps a combination of the strategies that increase with the expected return and decreases with increasing odds could be a good idea.

Chapter 9

Conclusion

In this project I have given a review of existing literature in Soccer result-prediction, and I have expanded on existing models by including data that previously were not considered. I have developed 6 different models for prediction with increasing complexity. All models have been tested in how well they fit already seen data and how well they predict unseen, out-of-sample data. Lastly, the final model has been applied in a betting scheme and tested across several seasons.

As we are no longer limited to knowing only goals scored by each team in a match, including more interesting data was a natural way to evolve prediction models. I did this by foregoing the standard model of goals following the Poisson distribution, and instead focused on a model where goals are a result of chances created which in turn are Poisson distributed.

Assuming that a chance becoming a goal can be modeled as a Bernoulli trial, I was able to describe teams with 4 different parameters each having a real world interpretation: the offensive players' ability to create chances, the offensive players' ability to convert chances into goals, the defensive players ability to prevent the opposition from getting to chances, and the defensive players' ability to prevent chances from becoming goals. It turns out that while these skillsets are often correlated - they are not the same, and this allows teams to stand out in different ways in the model.

By letting parameters change over time and introducing a correlation between them, the fit and predictive ability of the model was further improved. Lastly, including red cards gave the model a possible explanation for outliers where good teams struggled versus bad.

The model was tested on the betting market using both shots and shots on target to predict goals, and the conclusion is that shots on target have a more consistent connection to goals scored and as such works better in prediction. The model somewhat disappoints in the betting market - with the right choice of betting cutoff the model breaks even or makes a small profit, though the results are a little generous as the betting cutoff is chosen to maximize profits. Furthermore the results were helped a lot by a very successful season 2011/2012.

So while the model still isn't ready to consistently make money on the betting market, the results as a whole are very encouraging. By gradually increasing the complexity of the model, I have shown that incorporating shots on target and red cards significantly improves predictive accuracy and describes the data better than modeling goals scored directly.

Working on this project has been challenging, but ultimately very rewarding. I hope the capabilities

of the model are further improved by implementing modifications such as pre-season initial values for parameters and a more robust betting strategy.

Bibliography

- [1] The Football Association. Premier League Season Review 2013-2014. [Mhttp://review.premierleague.com/2013-14/](http://review.premierleague.com/2013-14/), 2014. [Online; accessed 23-June-2015].
- [2] Andreas Berg, Renate Meyer, and Jun Yu. Deviance information criterion for comparing stochastic volatility models. *Journal of Business & Economic Statistics*, 22(1):107–120, 2004.
- [3] Mark Dixon and Michael Robinson. A birth process model for association football matches. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):523–538, 1998.
- [4] 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.
- [5] Youyi Fong, Håvard Rue, and Jon Wakefield. Bayesian inference for generalized linear mixed models. *Biostatistics*, page kxp053, 2009.
- [6] football data.co.uk. [Mhttp://www.football-data.co.uk/englandm.php](http://www.football-data.co.uk/englandm.php), 2015. [Online; accessed 23-June-2015].
- [7] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.
- [8] John Goddard. Regression models for forecasting goals and match results in association football. *International Journal of forecasting*, 21(2):331–340, 2005.
- [9] ID Hill. Association football and statistical inference. *Applied statistics*, pages 203–208, 1974.
- [10] Siem Jan Koopman and Rutger Lit. A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):167–186, 2015.
- [11] Alan J Lee. Modeling scores in the premier league: is manchester united really the best? *Chance*, 10(1):15–19, 1997.
- [12] Mike J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.
- [13] J. Moroney. *Facts from figures*. Pelican books ; A236. Penguin Books, 1951.

- [14] Joel Oberstone. Differentiating the top english premier league football clubs from the rest of the pack: Identifying the keys to success. *Journal of Quantitative Analysis in Sports*, 5(3), 2009.
- [15] Alun Owen. Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, 22(2):99–113, 2011.
- [16] C Reep, R Pollard, and B Benjamin. Skill and chance in ball games. *Journal of the Royal Statistical Society. Series A (General)*, pages 623–629, 1971.
- [17] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.
- [18] 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.
- [19] Aditya Srinivas Timmaraju, Aditya Palnitkar, and Vikesh Khanna. Game on! predicting english premier league match outcomes. 2013.
- [20] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594, 2010.
- [21] www.lfchistory.net. Players bought by brendan rogers. [Mhttp://www.lfchistory.net/Transfers/ByManager/25-1](http://www.lfchistory.net/Transfers/ByManager/25-1). [Online; accessed 25-June-2015].

Appendix A

INLA

Integrated Nested Laplace Approximation (INLA) is an approach for doing inference by the use of latent Gaussian Markov Random Fields. As an alternative to Markov Chain Monte Carlo (MCMC) methods, its main advantage is its speed, though the error can not be made arbitrarily small like MCMC allows for. This will provide a superficial explanation of the method, for further details see Rue et al. (2009) [17].

A Latent Gaussian Model is a hierarchical structure consisting of three stages. The first stage is a likelihood function with the response \mathbf{y} assumed conditionally independent on the latent parameters \mathbf{z} and additional parameters $\boldsymbol{\theta}$,

$$\pi(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) = \prod_j \pi(y_j|\eta_j, \boldsymbol{\theta})$$

The second stage is formed by the latent Gaussian field by attributing a Gaussian distribution on the latent field η with mean $\mu(\boldsymbol{\theta})$ and precision matrix $Q(\boldsymbol{\theta})$,

$$z_t|\boldsymbol{\theta} \sim N(\mu(\boldsymbol{\theta}), Q^{-1}(\boldsymbol{\theta})).$$

The third stage is formed by the prior distribution assigned to the hyperparameters,

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

The INLA package in R provides an interface for this that supports use of different models, likelihoods and priors in a language akin to the GLM package for Generalized Linear Models. The response is stored in a vector \mathbf{Y} , the formula contains the latent parameters with informative or uninformative priors, and the `inla()` call contains the family and link function used. The output is the posterior marginal distribution of each parameter with expected value, quantiles and covariance matrix.

Appendix B

R Code

```
## Load packages and set correct work directory
library(INLA)
library(boot)
setwd("C:/Users/jihad/Documents/inlaskit")

## Load the data from the season I want to work with
source("datareader.R")
r = getSeason("0809")

## Useful values
## n.times finds gameweek number
n.times = (max(r$Date)%/%7)
teams = unique(sort(c(r$HomeTeam, r$AwayTeam)))
n.teams = length(teams)
n.matches = nrow(r)

## Set priors
hyperh = list(prec = list(prior = "gaussian", param = c(0.28, 50)))
hypert = list(rho = list(prior = "pc.rho1", param = c(0.5, 0.75)))

## Formula for the model.
## time and time2 hold the 4 parameters per team
## prop have attacking (1) and defending (2) chance strengths
## prop2 have attacking (3) and defending (4) goal strengths
## team is always attacking team, team2 always defending
## red.a, red.d are attacking and defensive goal-impact
## of receiving a red card
## red.a.c, red.d.c – same but for chances
formulal = Y ~
  -1 + intercept1 + intercept2 +
  red.a+red.d+ red.a.c+red.d.c+
```

```

f(idx.home.advantage, model="iid", hyper = hyperh) +
f(time2, weight, model="ar1", replicate = team2, group = prop2,
values = 0:n.times, control.group = list(model = "exchangeable"),
hyper = hypert)+
f(time, weight1, copy="time2", replicate = team, group = prop)

## Building the parts of
## Y1 poisson - chances
## Y2 binomial - goals
## .t suffix means temporary, mostly just semantics
Y1 = matrix(NA, n.matches * 4, 1)
Y2 = matrix(NA, n.matches*4, 1)
Ntrials.t = c()
idx.home.advantage.t = c()
weight.t = c()
time.t = c()
team.t = c()
prop.t = c()
prop2.t = c()
team2.t = c()
red.a.t = c()
red.a.c.t = c()
red.d.t = c()
red.d.c.t = c()
intercept1.t = c()
intercept2.t = c()
weight1 = rep(1, 4*n.matches)

count = 0
## Each match gets four lines in Y1/Y2.
## Att-chance - Def-chance - Att-Goal - Def-Goal
for(g in 1:n.matches) {
## cx/cy are chances. HST means Shots on Target.
## Replace with HS if you want just shots
x = r$FTHG[g]
y = r$FTAG[g]
cx = r$HST[g]
cy = r$AST[g]
rx = r$HR[g]
ry = r$AR[g]
home = r$HomeTeam[g]
away = r$AwayTeam[g]
date = r$Date[g]

## cx

```

```

count = count + 1
Y1[count] = cx
idx.home.advantage.t = c(idx.home.advantage.t, "home.advantage")
weight.t = c(weight.t, -1)
Ntrials.t = c(Ntrials.t, NA)
time.t = c(time.t, date)
team.t = c(team.t, which(teams == home))
prop.t = c(prop.t, 1)
team2.t = c(team2.t, which(teams == away))
prop2.t = c(prop2.t, 2)
red.a.t = c(red.a.t, NA)
red.a.c.t = c(red.a.c.t, rx)
red.d.t = c(red.d.t, NA)
red.d.c.t = c(red.d.c.t, ry)
intercept1.t = c(intercept1.t, 1)
intercept2.t = c(intercept2.t, NA)

```

```
## cy
```

```

count = count + 1
Y1[count] = cy
idx.home.advantage.t = c(idx.home.advantage.t, NA)
weight.t = c(weight.t, -1)
Ntrials.t = c(Ntrials.t, NA)
time.t = c(time.t, date)
team.t = c(team.t, which(teams == away))
prop.t = c(prop.t, 1)
team2.t = c(team2.t, which(teams == home))
prop2.t = c(prop2.t, 2)
red.a.t = c(red.a.t, NA)
red.a.c.t = c(red.a.c.t, ry)
red.d.t = c(red.d.t, NA)
red.d.c.t = c(red.d.c.t, rx)
intercept1.t = c(intercept1.t, 1)
intercept2.t = c(intercept2.t, NA)

```

```
## x
```

```

count = count + 1
Y2[count] = x
idx.home.advantage.t = c(idx.home.advantage.t, NA)
weight.t = c(weight.t, -1)
Ntrials.t = c(Ntrials.t, cx)
time.t = c(time.t, date)
team.t = c(team.t, which(teams == home))
prop.t = c(prop.t, 3)
team2.t = c(team2.t, which(teams == away))

```

```

prop2.t = c(prop2.t, 4)
red.a.t = c(red.a.t, rx)
red.d.t = c(red.d.t, ry)
red.a.c.t = c(red.a.c.t, NA)
red.d.c.t = c(red.d.c.t, NA)
intercept1.t = c(intercept1.t, NA)
intercept2.t = c(intercept2.t, 1)

## y
count = count + 1
Y2[count] = y
idx.home.advantage.t = c(idx.home.advantage.t, NA)
weight.t = c(weight.t, -1)
Ntrials.t = c(Ntrials.t, cy)
time.t = c(time.t, date)
team.t = c(team.t, which(teams == away))
prop.t = c(prop.t, 3)
team2.t = c(team2.t, which(teams == home))
prop2.t = c(prop2.t, 4)
red.a.t = c(red.a.t, ry)
red.d.t = c(red.d.t, rx)
red.a.c.t = c(red.a.c.t, NA)
red.d.c.t = c(red.d.c.t, NA)
intercept1.t = c(intercept1.t, NA)
intercept2.t = c(intercept2.t, 1)
}

Ntrials = Ntrials.t
time.t = time.t %% 7

## The call to inla, fits the data to a model, estimates parameters
result1 = inla(formula1,
               data = list(
                 Y = cbind(Y1,Y2),
                 idx.home.advantage = as.factor(idx.home.advantage.t),
                 time = time.t,
                 team = team.t,
                 prop = prop.t,
                 time2 = time.t,
                 team2 = team2.t,
                 prop2 = prop2.t,
                 weight = weight.t,
                 red.a = red.a.t,
                 red.d = red.d.t,
                 red.a.c = red.a.c.t,

```

```

    red.d.c = red.d.c.t,
    intercept1 = intercept1.t,
    intercept2 = intercept2.t),
family = c("poisson", "binomial"),
Ntrials = Ntrials,
control.compute = list(dic = TRUE, waic = TRUE)
)

## Extracting parameter values:
## Sorted by team -> parameter -> time
times = result1$summary.random$time2$mean
## To find team a, parameter b, time c:  $392*(a-1) + (b-1)*98 + c$ 
perLag = split(times, ceiling(seq_along(times)/(4*(n.times+1))))

```