Modelling & Predicting all 4 tiers of English League Football

Gareth Ridall[1] and Conor Higgins[1]

[1]Lancaster University, Lancaster, UK

Abstract

This paper aims to predict the outcomes of English league football. We evaluate a multitude of models ranging from a univariate Poisson to a dynamic state-space sequentially updating mixture model and others along the way. The models are all assigned attacking and defending parameters, whilst also being assigned a common home ground advantage. In the dynamic model we allow these parameters to vary throughout the season in order to accurately represent the team's form. This model uses a Bayesian methodology with Gamma priors assigned to each of the parameters. We deal with the problem of relegation and promotion between divisions, and utilise forgetting to capture the current form of a team. In order to optimise all of the values within the model we use scoring methods such as the rank probability score in order to decipher which of our predictions are best. We are then able to produce graphs which depict each team's journey over the past 22 seasons, and therefore make inference on why some of the peaks and troughs in the graph that occur. This allows us to better see what can affect a football club within and between seasons.

Contents

1 Introduction

1.1 Motivation

Football is one the most universally loved sports on the planet. It is a contest of skill, strength and willpower that brings adulation, heartbreak and hope to fans everywhere. It is a sport in which more than half of the world (3.572 billion viewers) watched the most recent FIFA Men's World Cup Final in 2018, FIFA (2018) [1]. With a sport this large comes the opportunity to place a bet on almost anything. This has created a betting culture in the UK, with it now being the world's largest online regulated gambling market according to the Gambling Commission (2018) [2].

In recent years the rise of statistics within football has been nothing short of astronomical, with many outlets now using measures such as *Expected Goals* (xG) or *Non-Shot Expected Goals* (NSxG) along with much more. The likes of bookmakers and exchanges now spend millions of pounds and countless hours using these statistics to best predict what will happen in any given game of football. However, not everyone has this luxury or wealth of information. Therefore, we must use the simple information which is widely available to try and gain our own advantage. We want to be able to produce accurate predictions within a short time span so we must also try to reduce the use of time consuming Monte Carlo Markov Chain methods whilst still maintaining the accuracy that these methods provide.

1.2 Aims

The overall aim of this paper is to produce accurate predictions for association football. However, along the way we also see if we can notice trends to other footballing events such as managerial sackings or trends towards relegation battles for certain teams.

In the first half of the paper we will use classical methodology and modelling to produce stationary parameters of which we use to model the number of goals a team is likely to score. We explore models created by others whilst also designing our own. We then test the validity of these models using diagnostic techniques and use scoring methods to determine the best models.

In the second part of the paper we create a dynamic state space model and use Bayesian methodology to sequentially update the state space. We can therefore update our parameters in order to have a non-stationary model. This will allow us to better represent teams throughout the season. We then extend this model to multiple different English tiers of football in order to increase the accuracy of our predictions at the beginning of the season.

1.3 Data

Data from every Football League game in England was collected from `http://football-data.co.uk/`. The data consists of 22 seasons spanning from 1997/98 all the way to the most recent season 2018/19. There is a vast wealth of simple data available from these seasons, including which referee took charge of the game and what the bookmakers odds of result were before kick off. However, as previously mentioned, we focus on the simple variables available to us. These are: Division (e.g. Premier League); Match Date; Home Team; Away Team; Home Team Goals and Away Team Goals. We have also created a data set for the coming seasons fixtures so that we can test our predictions in future games. In total there are 2036 fixtures played each season over four divisions each year, meaning that we have over 40,000 points of data to analyse for 111 different league teams.

2 Univariate Poisson

The univariate Poisson model is the simplest model that we consider for our analysis. A univariate Poisson model must take into account:

- The ability of both the home and away teams,
- The idea that a teams ability is best split into an Attacking Strength and a Defensive Strength,
- The Home Ground Advantage (HGA),

to give an accurate prediction of the outcome of a football game according to Dixon and Coles (1997) [3].

2.1 Model Assumptions & Formulation

The assumptions of our univariate Poisson are also simple, we assume that the number of goals scored by a home team is independent of the number of goals scored by an away team for any given match. Another assumption is that the rate at which a team scores goals based on their attacking and defensive strength is fixed over an entire season. These assumptions lead to our univariate Poisson model having conditionally independent random variables for both home and away goals.

To formulate our model we define α to be the Attacking Strength, β to be the Defensive Strength and γ to be the home ground advantage. The rate at which a home and away team scores goals is therefore

$$\lambda_t = \alpha_i \beta_j \gamma, \quad \text{and} \quad \mu_t = \alpha_j \beta i,$$

respectively. Where i indexes the home side and j indexes the away side in game t. It is important to note that a lower β indicates a better defensive strength and that a transformation of $\frac{1}{\beta}$ will be used to plot defensive strength such that a higher defensive strength is better, this is done for ease of interpretation.
Our univariate Poisson model is therefore given by

$$X_t \sim \text{Poisson}\left(\lambda_t\right), \quad Y_t \sim \text{Poisson}\left(\mu_t\right), \tag{2.1}$$

where λ_t and μ_t are as above and $\alpha_i, \alpha_j, \beta_i, \beta_j, \gamma > 0 \ \forall \ i, j$. X_t and Y_t now denote the number of expected home and away goals in game t.

2.2 Identifiability

Identifiability is a property that must be satisfied in order for accurate inference. In the case of the univariate Poisson we must solve the identifiability problem in order to ensure that we can find a correct and unique maximum likelihood estimate (MLE). We can use multiple methods to solve the identifiability problem.

The first method of countering the identifiability problem is to set one of the 20 Premier League teams to a baseline score of 1 for both attacking and defensive strength - this idea was posed in Dixon and Coles (1997) [3]. For simplicity, and the fact that they have been ever present in the Premier League, we set Arsenal to be this team. The problem with this constraining method is that all team strengths are relative to Arsenal, which would not be a true reflection of their strengths. We would also encounter another problem if the model were to be extended to the Championship, since there is no team that has been present in each of the last 22 Championship seasons. This means we have no team to compare others against and analysis of the parameters would be much harder and extremely tedious. This will prompt a new way of constraining parameters.

The second method of dealing with identifiability places a constraint on the defensive strength such that

$$\sum_{k=1}^{n} \log \beta_k = 0,$$

where n is the number of teams in the division. This is achieved by setting $-\sum_{k=2}^{n} \log \beta_k = \log \beta_1$. This method provides us with the unique identifiability that we desire whilst also giving a true reflection of the parameters as we no longer have a score relative to another club. With this method we are now also able to extend the idea to other division across England and even into Europe. Therefore we proceed by using the second method of identifiability along with Equation 2.1 for our univariate model, we establish this as Model 1. The model has 20 attacking parameters, 19 defensive parameters and the home ground advantage for a total of 40 parameters.

2.3 Other Rejected Models

Other models were formulated under the univariate Poisson assumptions, but these models are ultimately rejected for various reasons. Firstly, one model combined both attacking and defensive strengths into one score in an attempt to reduce the number of parameters in the model but failed to correctly portray the data. Whilst another model assigned a unique home ground advantage to each team, but with a total of 59 parameters the model was over-fitting. Over-fitting is a problem in which the model fits the data "too well" and struggles to correct predict results, this will be further explained in Section 4.

2.4 Skellam Distribution

The Skellam distribution is a method of modelling the difference between two conditionally independent Poisson distributed random variables and was first discovered by Skellam (1946) [4]. The distribution has been utilised in football predictions for numerous years as it is able to give a probability for the difference in goals in the game. Utilising the skellam package within R we can use the expected difference in goals to calculate a Home Win, Draw and Away Win probability for any given game. For example, with the univariate model if Manchester City play Huddersfield at home in the 18/19 season are expected probabilities would be 0.95, 0.04, 0.01 for each result respectively.

For the Skellam distribution we set $K_t = X_t - Y_t$ and let the means of the two Poisson random variables be as earlier $\lambda_t = \alpha_i \beta_j \gamma$ and $\mu_t = \alpha_j \beta_i$. The formula for the Skellam distribution is hence

$$p\left(k; \lambda_t, \mu_t\right) = \mathbb{P}\{K_t = k\} = e^{-(\lambda_t + \mu_t)} \left(\frac{\lambda_t}{\mu_t}\right)^{k/2} I_k\left(2\sqrt{\lambda_t \mu_t}\right), \tag{2.2}$$

where I_k is the modified Bessel function of the first kind [5].

2.5 Delta Method

The Delta method is a calculation of the variance of a transformation, it can be defined by letting a random variable $B = f(A)$ where f is a differentiable function. By taking the Taylor series expansion of A around a "small neighbourhood" μ we can say that

$$B - f(\mu) \approx f'(\mu)(A - \mu),$$

we then square this and take expectations such that

$$\mathbb{E}\left[(B - f(\mu))^2\right] \approx f'(\mu)^2 \mathbb{E}\left[(A - \mu)^2\right].$$

We can thus state that $\mathrm{Var}(B) = \left(g'(\mu)\right)^2 \mathrm{Var}(A)$ and the standard error will therefore be the square root of this. In the case of the univariate Poisson we calculate all standard errors on the logarithmic scale. For example, to calculate the standard error of β we call the logarithmic standard error b we can then use the delta method as such

$$\mathrm{se}(\beta) = \exp(\hat{b}_i) \times \mathrm{se}\left(b_i\right).$$

This method can then be applied to all parameters to provide the standard error.

2.6 Results

In Table 1 we see the univariate Poisson model parameter estimates from the "Big 6" teams in the Premier League in the 2019 season. It is clear to see a gulf in class between the rest of the clubs when compared against Liverpool and Manchester City, especially in the defensive department. This was also the case throughout the entire 18/19 season with Manchester City and Liverpool keeping 41 clean sheets between them and conceding just 45 goals. These two clubs also ran away from the rest of the league in the points tally, with third place Chelsea 25 points adrift of second place Liverpool.

Team	Attack	Defence
Arsenal	1.63	0.98
Chelsea	1.39	1.30
Liverpool	1.93	2.24
Man City	2.06	2.13
Man United	1.45	0.93
Tottenham	1.47	1.29

Table 1: Model parameter estimate for attacking and defensive strengths.

Figure 1 displays the attacking and defensive strengths of the entire league. As predicted, Huddersfield, Cardiff and Fulham were all relegated with each of the clubs falling into the bottom three scores for attacking strength. The red bars on the graph indicate the standard errors of the parameters which were calculated using the Delta method in Section 2.5. We can see that most parameters have similar standard errors, but Liverpool and Manchester City have large standard errors for their defensive score, this could be due to the constraint on the parameter but is no cause for concern.

2.7 Model Diagnostics

Model Diagnostics are an integral part of any statistical analysis. Diagnostics are a way of testing model assumptions and model fit, and must be rigorously carried out in order to ensure the best model has been selected. There are multiple different methods used to diagnose a model and these will be explored within the following section.

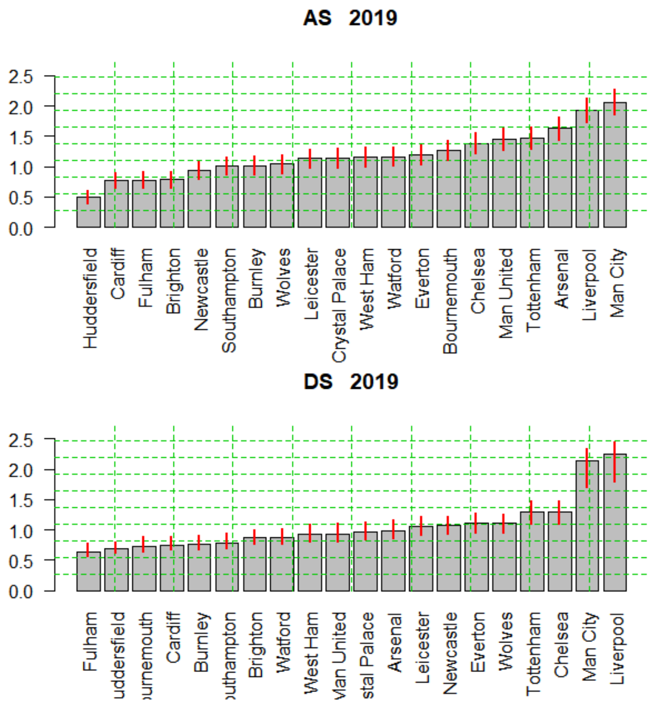

Figure 1: Univariate model Attacking and Defending Strengths.

2.7.1 Correlation

To validate the assumption of independence within our univariate Poisson model we test the correlation of the Pearson residuals of the data. Pearson residuals are a way of measuring the difference between the expected scoreline and what we actually observed. The Pearson residuals for home and away goals are given by

$$PR_t^X = \frac{O_t^X - \lambda_t}{\sqrt{\lambda_t}} \quad \text{and} \quad PR_t^Y = \frac{O_t^Y - \mu_t}{\sqrt{\mu_t}}, \tag{2.3}$$

respectively. Where O_t^X and O_t^Y are the observed number of home and away goals in game t. λ_t and μ_t are as earlier, the expected number of goals scored by a home and away team.

It seems likely that our assumption of independence may not hold since there are external factors that can affect the number of goals scored in a football game. For example, if a team sets up defensively away from home in the hopes of coming away with a goalless draw this will surely have an impact on the number of goals either team will score? In the opposite case this could also be true. If a team at the top of the league is playing at home and winning 6-0, surely this makes the away side much less likely to score than if it were 0-0? Another factor we may need to take into consideration is the occurrence of own goals. This is a random event in a game but we do not allow for this in our analysis and instead allow a teams attacking strength to increase regardless of the occurrence of an own goal.

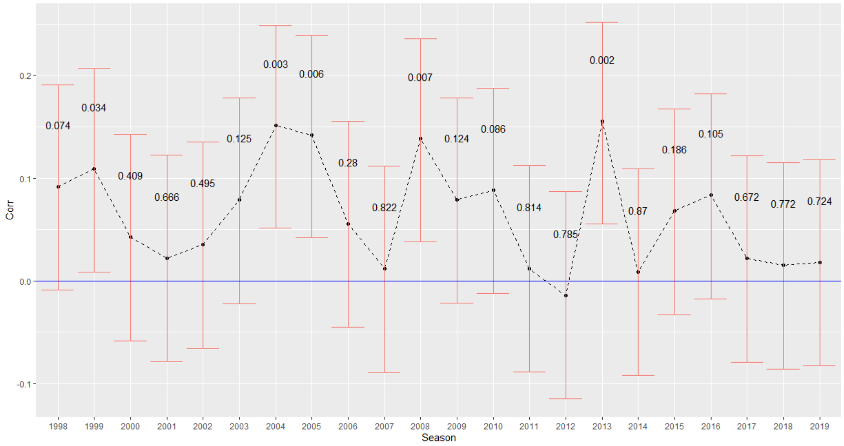

Figure 2: The Correlation of the Pearson residuals each season, with given p-values.

In Figure 2 we can see the correlation of the Pearson residuals in each of the last 22 seasons with a 95% confidence interval and given p-values. We can see that in 5 of the 22 seasons (1999, 2004, 2005, 2008, 2013) the p-value of correlation is < 0.05 and the confidence interval lies above 0. Therefore we can say that the correlation in fact is not 0 and that these seasons are deemed to not have independence between the number of goals scored by a home and away side. In the most extreme case of 2013, we find the correlation to be a staggering 0.15. We can also see that one of the seasons (2012) is slightly negatively correlated. However, the p-value of this correlation is 0.785 and is no real cause for concern.

2.7.2 Overdispersion

Overdispersion is the presence of extreme variability within a statistical model when compared to data. We can test for overdispersion using the Pearson residuals calculated in Section 2.7.1. Using these residuals we can take the sum of squares to calculate a χ^2 test statistic, this test for overdispersion is considered in Dean (1989) [6]. Our test statistic is therefore given by

$$\chi^2 = \sum_{t=1}^{380} \left(\frac{O_t^X - \lambda_t}{\sqrt{\lambda_t}} \right)^2 + \sum_{t=1}^{380} \left(\frac{O_t^Y - \mu_t}{\sqrt{\mu_t}} \right)^2, \tag{2.4}$$

This value can then be compared against the χ^2 distribution, which at large values of n can be approximated by the normal distribution using the central limit theorem

Gyu and Hae (2017) [7]. In our case the χ^2 distribution will have $760 - 40 = 720$ degrees of freedom since we have 760 residuals and 40 model parameters. At a 95% confidence level the χ^2 distribution gives a confidence interval of [647.5, 796.3].

We find that for our univariate Poisson model 20 of the 22 seasons lie within this interval. The 1998 season was found to have overdispersion with a test statistic of 809.4. The 2007 season was found to be underdispersed with a test statistic of 639.9. Since almost all of the seasons lie within the confidence interval, it is reasonable to say that the Poisson distribution is a good fit for the data.

2.7.3 Residual Analysis

Analysis of residuals is an important part of model diagnostics as it helps us identify possible outlying results whilst also being able to look at the residuals in more detail.

Figure 3: Pearson residuals for the 98/99 and 03/04 seasons.

Figures 3, 4 and 5 show the six seasons which we mentioned in Section 2.7.2. Firstly, we can see that in the 1998/99 season (Figure 3) there are a large number of residuals we deem to be unusual (residuals that exceed the value 3). It is worth noting that all of these games are high scoring results with the lowest scoring game still having 5 goals between the two teams.

Expanding our view to all 3 Figures we can see that the “Big 6” teams mentioned earlier appear in 16 of the 37 unusual results or 43% of the matches. Considering these teams make up only 30% of the league this is not something we would expect, not to mention the fact that some of these games are played between two “Big 6”

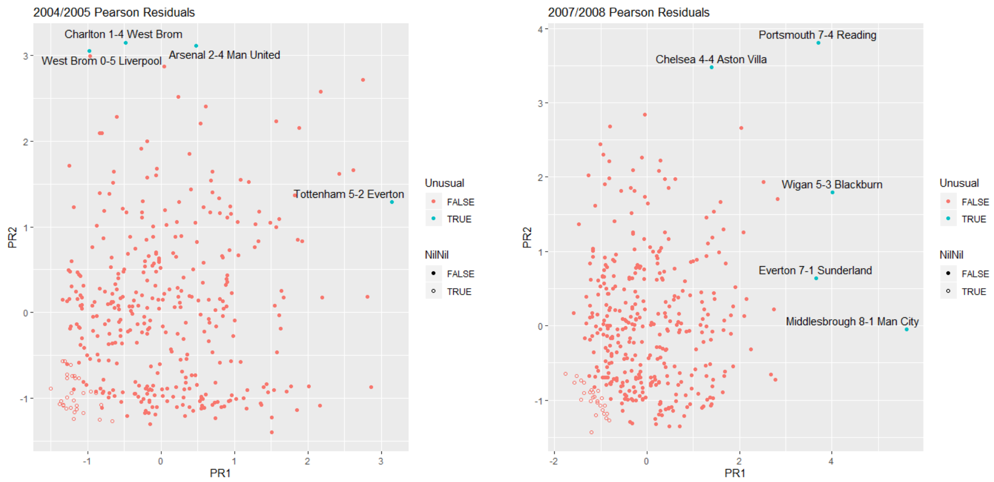

Figure 4: Pearson residuals for the 04/05 and 07/08 seasons.

clubs. This would point toward the fact that these big teams are more likely to cause a high scoring game. A possible solution could be to truncate these clubs at a maximum of 4 goals so that we can restrict the size of the Pearson residuals. The problem with this idea is that if a side were to score 4 against one of the truncated clubs, whilst one of those clubs scored 5 goals, we would not accurately represent the actual scoreline. We would assume the game was a draw when in fact it was not.

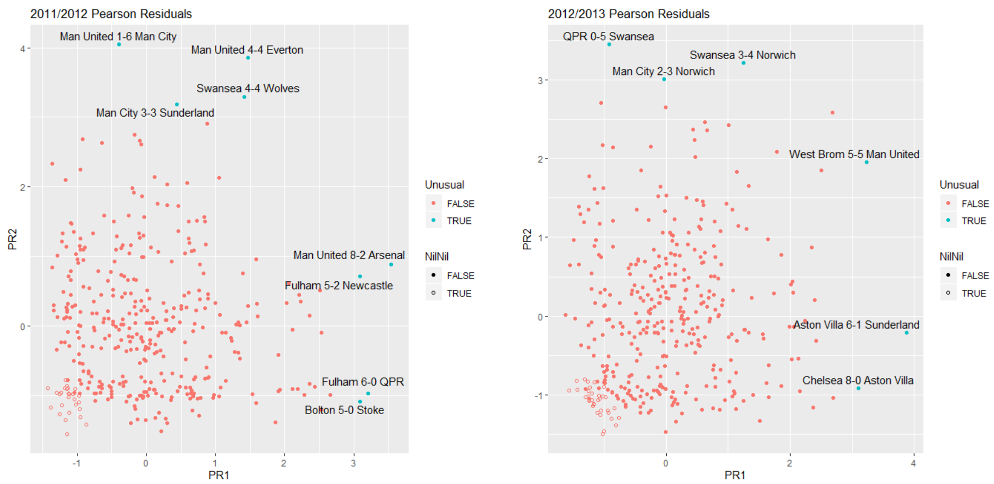

Figure 5: Pearson residuals for the 11/12 and 12/13 seasons.

It is also important to note the 0-0 games. All of these games fall into negative Pearson residuals and are proof that we are under-predicting the number of times a team will score 0 goals. A possible solution for this could be a zero-inflated Poisson distribution, this however is not discussed in the scope of this paper.

2.7.4 Prediction

Prediction is an integral part of any football model, if the model struggles to predict the result of games it is rendered almost useless. For the prediction within our univariate Poisson model we can use the Skellam distribution from Section 2.4 to calculate the total number of predicted: Home Wins, Draws and Away Wins over a given season. We can then look at the actual number of these results and take the difference to see how well we predict each result.

In Table 2 we can see these differences for the past 11 seasons. The averages give us a rough idea at how well the model performs. It is clear to see that we are under-predicting the number draws in any given season, with the worst case occurring in the 10/11 season in which we under-predicted draws by 22.19. This under-prediction of draws in turn causes an over-prediction of Away wins over a given season, giving away sides more credit than they deserve. The model seems to accurately predict the number of home wins on average with the error margin being an average of 1.11.

Season	Home Win	Draw	Away Win
08/09	-4.52	-4.08	8.60
09/10	-1.61	-12.44	14.05
10/11	1.69	-22.19	20.50
11/12	2.94	-5.77	2.83
12/13	3.98	-19.75	15.77
13/14	-5.19	8.71	-3.51
14/15	1.70	-0.19	-1.51
15/16	9.85	-16.58	6.72
16/17	-11.36	0.23	11.13
17/18	0.58	-11.08	10.50
18/19	-10.27	11.54	-1.27
Average	**-1.11**	**-6.51**	**7.62**

Table 2: Difference tables between Expected and Observed results.

2.7.5 Conclusion

Using the results throughout the section, we can now make inference on the fit and ability of the univariate Poisson. The fact that correlation was found within five of the seasons leads towards a bivariate Poisson, which allows for the dependency of the home and away goals. Over dispersion being found in one of the seasons also leads towards a bivariate negative binomial model being considered. However, overall the univariate Poisson is a well fitting model for football predictions, for how simple the model is performed reasonably well, there is certainly room for improvement though.

3 Bivariate Poisson & Negative Binomial

The models in the following section are explored as a result of the inadequacies of the univariate Poisson. The bivariate Poisson can be used to counteract the problems of correlation, whilst the negative binomial can solve the problem of overdispersion.

3.1 Bivariate Poisson

3.1.1 Model Assumptions & Formulation

The bivariate Poisson can be thought of as an extension of the univariate Poisson from Section 2. We continue using the assumption that the model parameters are fixed over the entire season. However, we relax the assumption of independence between the number of home and away goals scored.

The first bivariate Poisson model explored is from Karlis and Ntzoufras (2003) [8], who designed a bivariate Poisson model that allows for the dependence between the two random Poisson variables. We now refer to this model as Model 2 in our analysis. Consider the random variables X_κ where $\kappa = 1, 2, 3$ which follow independent Poisson distributions with parameters $\lambda_\kappa > 0$. Random variables $X = X_1 + X_3$ and $Y = X_2 + X_3$ follow a bivariate Poisson distribution $\mathrm{BP}\left(\lambda_1, \lambda_2, \lambda_3\right)$. We can think of X_1 and X_2 as the scoring parameters for the home and away side, whilst X_3 measures the extent of external factors that can contribute to the score of a game. This model can be given with probability function

$$\begin{aligned}\mathbb{P}_{X,Y}(x, y) &= \mathbb{P}(X = x, Y = y) \\ &= \exp\left\{-\left(\lambda_1 + \lambda_2 + \lambda_3\right)\right\} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{k=0}^{\min(x,y)} \begin{pmatrix} x \\ k \end{pmatrix} \begin{pmatrix} y \\ k \end{pmatrix} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k,\end{aligned} \tag{3.1}$$

such that $\mathbb{E}(X) = \lambda_1 + \lambda_3$ and $\mathbb{E}(Y) = \lambda_2 + \lambda_3$ and $\mathrm{cov}(X, Y) = \lambda_3$. λ_1 and λ_2 are thought of as the scoring parameters of each team. Hence λ_3 is a measure of dependence between the two random variables. If $\lambda_3 = 0$, then we can say that there is independence between the scoring parameters of each team.

The second bivariate Poisson we explore is one of Dixon and Coles (1997) [3], who continued the work suggested by Maher (1982) [9] and tackled the assumption of independence in the Univariate Poisson, by designing the following Bivariate Poisson model which we now treat as Model 3. The model aims to counteract the fact that "low-scoring" games (0-0, 1-0, 0-1, 1-1) seem to be some of the most correlated fixture results. Dixon and Coles do this by introducing ρ, which acts as a dependence parameter. If both the home and away scorelines exceed 1 we set $\rho = 1$ and treat the two scorelines as independent. The joint density function is then given by

$$\mathbb{P}\left(X_t = x, Y_t = y\right) = \tau_{\lambda,\mu}(x,y)\frac{\lambda_t^x \exp(-\lambda_t)}{x!}\frac{\mu_t^y \exp(-\mu_t)}{y!}, \tag{3.2}$$

where $\lambda_t = \alpha_i \beta_j \gamma$, $\mu_t = \alpha_j \beta_i$, as previous and

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

3.2 Bivariate Negative Binomial

3.2.1 Model Assumptions & Formulation

As a model, the bivariate binomial has the potential to solve both the correlation and overdispersion problems of the univariate Poisson. The model again relaxes the assumption of independence and assumes that there is a shared random effect between the home and away goals.

This model was designed by Ridall (2019), and the likelihood marginalised over the shared random effect $\epsilon_t \sim \text{Gamma}(\kappa, \kappa)$ is given by

$$
\begin{aligned}
L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \kappa) &= \int_{\boldsymbol{\epsilon}} f(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \kappa) p(\boldsymbol{\epsilon} \mid \kappa) d\epsilon \\
&= \prod_{t=1}^{380} \frac{1}{x_t! y_t!} \int_{\epsilon_t} e^{-\epsilon_t(\mu_t + \lambda_t)} \, (\mu_t \epsilon_t)^{y_t} (\lambda_t \epsilon_t)^{x_t} \times \frac{\kappa^\kappa}{\Gamma(\kappa)} \epsilon_t^{\kappa - 1} e^{-\epsilon_t \kappa} d\epsilon_t \\
&= \prod_{t=1}^{380} \frac{\kappa^\kappa}{\Gamma(\kappa)} \frac{1}{x_t! y_t!} \mu_t^{y_t} \lambda_t^{x_t} \int_{\epsilon_t} e^{-\epsilon_t(\kappa + \mu_t + \lambda_t)} \, \epsilon_t^{\kappa + x_t + y_t - 1} d\epsilon_t \\
&= \prod_{t=1}^{380} \frac{\Gamma(\kappa + x_t + y_t)}{\Gamma(\kappa)\Gamma(x_t + 1)\Gamma(y_t + 1)} \kappa^\kappa \mu_t^{y_t} \lambda_t^{x_t} (\kappa + \mu_t + \lambda_t)^{-(\kappa + x_t + y_t)} \\
&= \prod_{t=1}^{380} \frac{\Gamma(\kappa + x_t + y_t)}{\Gamma(\kappa)\Gamma(x_t + 1)\Gamma(y_t + 1)} p_t^{y_t} q_t^{x_t} (1 - p_t - q_t)^{\kappa_t}
\end{aligned}
$$

where $p_t = \frac{\mu_t}{\kappa + \mu_t + \lambda_t}$ and $q_t = \frac{\lambda_t}{\kappa + \mu_t + \lambda_t}$. This gives the results

$$
\begin{aligned}
\text{Var}(X_t) &= \mathbb{E}\left(\text{Var}[X_t \mid \lambda_t, \epsilon_t]\right) + \text{Var}\left(\mathbb{E}[X_t \mid \lambda_t, \epsilon_t]\right) \\
&= \mathbb{E}\left[\epsilon_t \lambda_t\right] + \text{Var}\left[\epsilon_t \lambda_t\right] \\
&= \lambda_t + \frac{\lambda_t^2}{\kappa} \\
\text{Cov}(X_t, Y_t) &= \mathbb{E}\left(\text{Cov}[X_t, Y_t \mid \mu_t, \lambda_t, \epsilon_t]\right) + \text{Cov}\left(\mathbb{E}[X_t, Y_t \mid \mu_t, \lambda_t \epsilon_t]\right) \\
&= 0 + \text{Cov}[\epsilon_t \mu_t, \epsilon_t \lambda_t] \\
&= \mu_t \lambda_t \text{Var}[\epsilon_t] \\
&= \frac{\mu_t \lambda_t}{\kappa}.
\end{aligned}
$$

Where λ_t is the number of expected goals scored by a home team in game t. This model will now be treated as Model 4 within our analysis.

3.3 Conclusion

Both the bivariate Poisson and the bivariate negative binomial are designed to improve on the inadequacies of the univariate Poisson as mentioned earlier. In this paper we will not perform model diagnostics on these models as they are difficult, time consuming and computationally intensive. Instead, we compare the models against the univariate Poisson in order to check if they actually improve on our base model before diagnostics are carried out. Using the comparison made in the following section, we are able to decipher if these models have indeed solved the problems for which they were designed.

4 Model Selection

Model selection is an extremely important method for any statistical process. We must be able to notice which of our designed models is the most effective in order to use it going forward. We must find the correct balance between a good model fit and the complexity of the model. Over-fitting, caused by having too many parameters, can be an extremely serious problem for a model. It can lead to extremely poor predictions as the model fits the data "too well". The same can be said for under-fitting. If the model is not complex enough, it will not truly represent the data and will likely be too vague. Selection can be done in multiple different ways, some of which are discussed in this section.

4.1 AIC

Akaike's information criterion (AIC) is the first classical approach that we take and is given by

$$\text{AIC} = 2K - 2\hat{\ell}, \tag{4.1}$$

where K is the number of parameters in the model and $\hat{\ell}$ is the log likelihood at it's maximum value, Bozdogan (1987) [10]. One can think of the number of parameters as the complexity of the model and the MLE of the log likelihood as the model fit.

Lower AIC indicates a better score than that of a high AIC. Adding parameters to the model will raise the AIC score, and can thereby find a way of punishing the model for the number of parameters it contains. AIC alone, has no real information on how good a model fits: it is a relative score and we therefore compare all models to a baseline model.

4.2 BIC

Bayesian information criterion (BIC), is the second classical approach we use for model selection and is given by

$$\text{BIC} = K\log(n) - 2\hat{\ell}, \tag{4.2}$$

where K and $\hat{\ell}$ are as before and n is the number of observations (In our case 380 games). The BIC punishes the model for adding new parameters harsher than the AIC, Kuha (2004) [11]. However, each of our 22 seasons all contain 380 games and the BIC will now just be a transformation of the AIC. We therefore proceed by using the AIC as our model selection method.

4.3 AIC Comparison

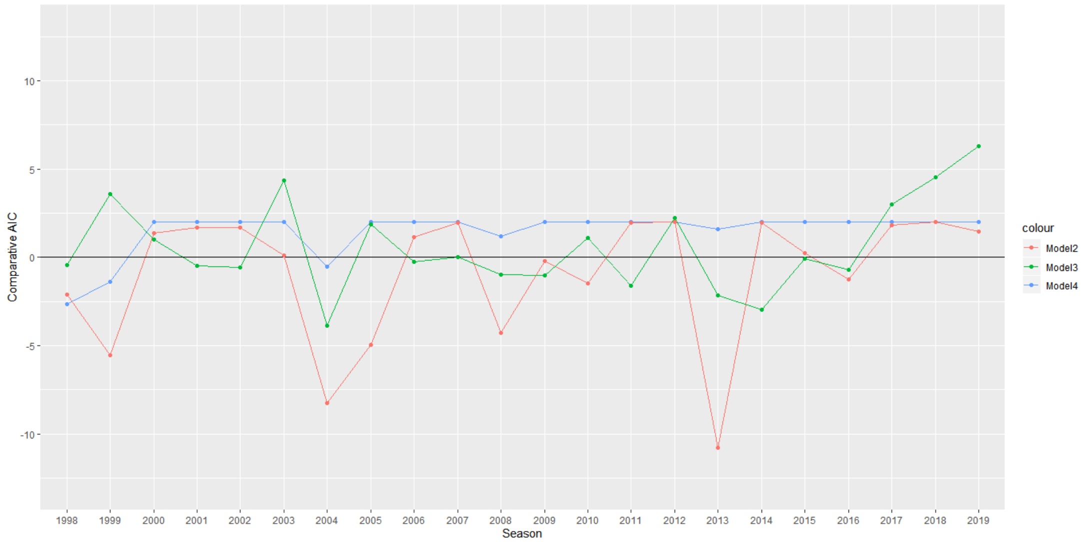

Figure 6: Relative AIC of all models.

From Figure 6 we can see each of the previous model's AIC scores over each of the last 22 seasons: where Model 2 represents the Karlis and Ntzoufras bivariate Poisson [8]; Model 3 represents the Dixon and Coles bivariate Poisson [3] and Model 4 represents the bivariate negative binomial. These models have been compared to the baseline of Model 1 (univariate Poisson) and therefore the black line along the y-axis represents Model 1. Any AIC score that falls below the black line means that this model outperformed Model 1 in that given season.

We can see that Model 4 only outperforms the univariate Poisson of Model 1 in 2 of the 22 seasons. We therefore prefer Model 1 over Model 4 for the analysis of football predictions and can remove Model 4 from future analysis. Additionally, we see in Figure 6 that Model 2 outperforms Model 1 in 9 of the 22 seasons. Although this number is less than half of the total number of seasons, the seasons in which the bivariate Poisson does fit better, the improvement of the AIC against Model 1 is drastic. The seasons in which Model 2 does not perform as well as the univariate

Poisson the AIC change in only relatively small and this bodes well for Model 2. The bivariate Poisson (Model 2) solves the problem for which it was designed. In the 9 seasons that it does improve on Model 1, all 5 of the correlated seasons fall in this category. Model 2 therefore does perform better when we know that there is correlation in the Pearson residuals. However, we will only know if this is the case after the end of the season and is something we cannot predict.

Model 3 is more desirable than Model 1 in 13 of the 22 previous seasons, meaning that Model 3 is better more than half of the time. However, the model AIC is much more volatile than Models 2 & 4. Model 3 has also performed rather poorly over the last 3 seasons, with it being the worst peforming model of all. Whilst the univariate Poisson has performed best in the 3 previous seasons and could be an indication that this is a preferred model going forward.

A case can be made for the selection of Model 1, 2 or 3, with each having their own upside. Model 2 and 3 both improve on the univariate Poisson in different ways over the span of analysis. All 3 models could be considered as favourable for analysis going forward, but we try to improve these models by relaxing the assumption of a constant scoring rate over a season and trying to better represent team form in Section 7.

5 Scoring Methods

Scoring methods are a way of measuring the accuracy of model predictions. There are multiple different methods of scoring our predictions, and in this paper we explore the most relevant of these methods. Since we use the Skellam distribution that measures the difference between two independent random Poisson variables we may only calculate scores for the univariate Poisson model.

5.1 Brier Scoring Method

The Brier Scoring method is the first way of measuring the accuracy of probabilistic predictions we explore. For any given football match there are 3 outcomes, Home Win, Draw or Away Win. How accurately we can predict each of these is extremely important in the context of this paper. The Brier score is given by

$$BS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}(f_{ti} - o_{ti})^2 \tag{5.1}$$

where f_{ti} is the probability that was predicted and o_{ti} is the actual outcome of the event at instance t on a binary scale of [0,1]. N is the total number of outcomes in a game (in our case 1), R is the number of possible classes in which the event can fall. Meaning that the Brier score for a football match is on a [0,2] scale, with 0 being the best possible prediction. An example of a Brier score in this context is to set the home, draw and away probabilities to 0.6, 0.3 and 0.1 respectively and the true outcome of the game was a home win. The Brier score would therefore be $(0.6 - 1)^2 + (0.3 - 0)^2 + (0.1 - 0)^2 = 0.26$.

5.1.1 Brier Score Results for the Univariate Poisson

In Figure 7 we can see the average Brier score over each of the previous 22 seasons. It is important to remember that a lower score represents a better predictions. We must also remember that these predictions are "in-sample" since we use all of the data over a season to calculate the scoring rate and then carry out predictions for each game. This type of prediction is unreliable as it can produce better scores than we would usually expect and we cannot make predictions ahead of time. We see that the average score lies between 0.5 and 0.6 with variation between each season as we would expect. The problem with Figure 7 is that the Brier score is on a [0,2] scale. However, the nature of the Brier score is to have a bias towards the lower end of this scale since we almost never have a team near a probability of 1 of winning. It is therefore difficult to interpret the Brier score without a comparison.

5.1.2 Disadvantages of the Brier Score

The Brier score has it's downfalls which are well documented in Constantinou and Fenton (2012) [12], who point out the problems with the Brier score. The article outlines five different scenarios and pits multiple scoring methods against one another. The Brier score performs well for three of these scenarios, but falls short in the other two. The article also poses questions about better methods to score football predictions and points towards the Rank Probability Score (RPS).

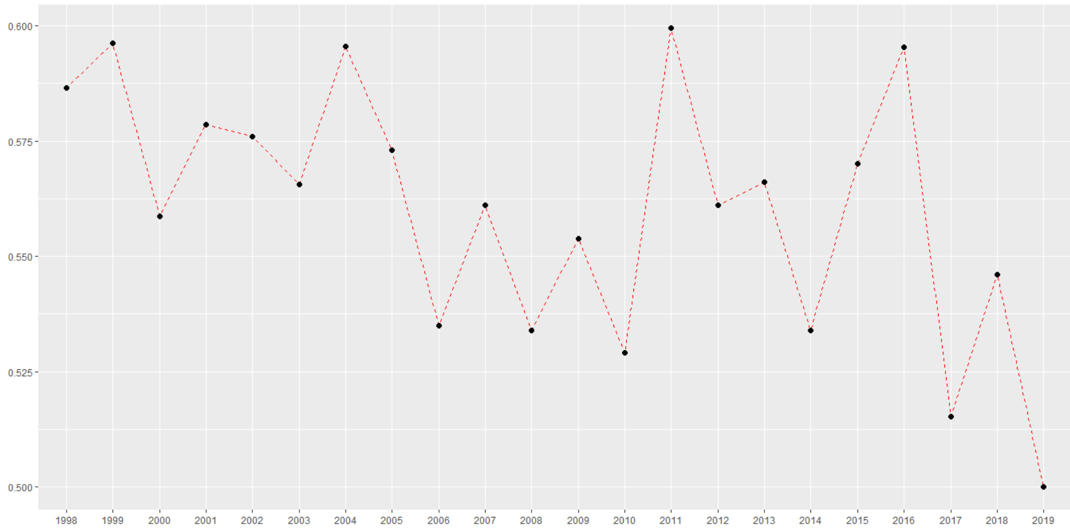

Figure 7: Average Brier Scores per game over the past 22 seasons.

5.2 Rank Probability Scoring Method

The RPS was first published in Epstein (1969) [13] and is an appropriate scoring method for probabilities of ordered variables, which is the case in football predictions. The Rank Probability takes into account the probabilities of incorrect predictions to a further extent than that of the Brier score. If we predict the probabilities of 0.6, 0.3, and 0.1 for each result respectively and the home team win in the Brier score. This would be the same as predicting 0.6, 0.1, and 0.3, which is not as accurate of a prediction. The RPS takes into account the order of these probabilities and is given by

$$RPS = \frac{1}{r-1} \sum_{j=1}^{R} \left(\sum_{i=1}^{j} p_i - \sum_{j=1}^{j} e_i \right)^2 \tag{5.2}$$

where r is the number of potential outcomes, and p_i and e_i are the forecasts and observed outcomes at position i, which is of length 3 (Home, Draw and Away predictions respectively). The RPS penalises the model harshly for the distance of it's incorrect predictions so it will give a true gauge on how well we predict games.

5.2.1 Rank Probability Results for the Univariate Case

In Figure 16 we can see the average RPS score per game over the past 22 seasons. We can see that the general trend is similar to that in Figure 7. Unfortunately, we cannot compare the Brier Score against the RPS since they are on different scales. We again note that the predictions are in sample and are therefore not a great measure of prediction overall. We will however use RPS later in the paper for out of sample predictions.

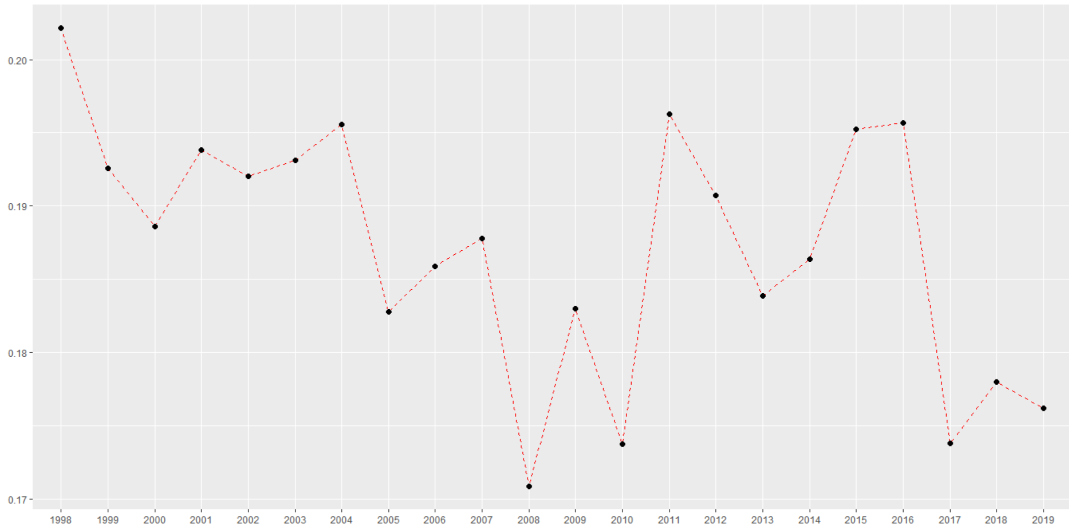

Figure 8: Average Rank Probability Score per game over the past 22 seasons.

6 The Sequential Stationary Approach

6.1 The Sequential Univariate Poisson Model

The sequential stationary approach still assumes that the model parameters are fixed over an entire season - but, these model parameters are now calculated using a Bayesian method initialising with non-informative priors. The priors are set to be non-informative so that we can fully cover the data whilst allowing a large variance in order for parameters to take their true value faster. The model is the same as Model 1 and is as such

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} \sim \text{Poisson} \begin{pmatrix} \alpha_i \beta_j \gamma \\ \alpha_j \beta_i \end{pmatrix}, \tag{6.1}$$

where α, β and γ are also as previously defined. The priors for each of the model parameters are given as

$$
\begin{aligned}
\alpha_i &\sim \text{Gamma}(p_i^{\alpha_i}, q_i^{\alpha_i}), \\
\alpha_j &\sim \text{Gamma}(p_j^{\alpha_j}, q_j^{\alpha_j}), \\
\beta_i &\sim \text{Gamma}(p_i^{\beta_i}, q_i^{\beta_i}), \\
\beta_j &\sim \text{Gamma}(p_j^{\beta_j}, q_j^{\beta_j}), \\
\gamma &\sim \text{Gamma}(p^{\gamma}, q^{\gamma}),
\end{aligned}
$$

where p, q are the sufficient statistics we use to update the prior and posterior. p, q are both set to be 10 for the first game of each season this also maintains a mean of 1, since the mean of the Gamma distribution will be $\frac{p}{q}$. The priors are found to have little influence on the model parameters, this however does not affect the predicting abilities of the model. Using the priors for each of the model parameters, which we now denote as $\theta = (\alpha_i, \alpha_j, \beta_i, \beta_j, \gamma)$, we are able to calculate a joint posterior which can be given by

$$
\begin{aligned}
\pi\left(\theta|\mathbf{x},\mathbf{y}\right) \propto\; & \underbrace{\exp\left[-\left(\alpha_i\beta_j\gamma + \alpha_j\beta_i\right)\right] \times \left[\alpha_i\beta_j\gamma\right]^{x_t} \times \left[\alpha_j\beta_i\right]^{y_t}}_{\text{Likelihood}} \\
& \times \underbrace{\alpha_i^{p_i^{\alpha_i - 1}} \exp\left(-q_i^{\alpha_i}\alpha_i\right) \times \alpha_j^{p_j^{j}-1} \exp\left(-q_j^{\alpha_j}\alpha_j\right)}_{\text{Attacking Priors}} \\
& \times \underbrace{\beta_i^{p_i^{\beta_i}-1} \exp\left(-q_i^{\beta_i}\beta_i\right) \times \beta_j^{\beta_j^{\beta_j}-1} \exp\left(-q_j^{\beta_j}\beta_j\right)}_{\text{Defensive Priors}} \\
& \times \underbrace{\gamma^{p^{\gamma}-1} \exp\left(-q^{\gamma}\gamma\right)}_{\text{HGA Prior}}.
\end{aligned}
$$

Using this equation, we can find the conditional posteriors of each model parameter, which is done by taking the values proportional to each parameter from the joint posterior. Once we have found the conditional posterior, we are able to calculate the conditional posterior distribution. We must note that we are using the posterior mean as our parameter estimates. The hat denotation on top of the parameter estimates represent the mean estimates at the previous iteration. This is used because we do not know the true parameter estimate at this time point. We use the conditional posterior in a different way to Gibbs Sampler. We update the posterior just once, rather than repeating the process until convergence with Gibbs Sampler. The latter is computationally intensive and performing this for all model parameters would be extremely time consuming. The conditional posterior and its distribution for the home attacking strength are hence given by

$$\begin{aligned}\pi\left(\alpha_i|\theta_{-\alpha_i},\mathbf{x},\mathbf{y}\right) &\propto \alpha_i^{p_i^\alpha-1}\exp\left(-q_i^\alpha\alpha_i\right)\times\alpha_i^{x_t}\exp\left(-\alpha_i\beta_j\gamma\right)\\ &=\alpha_i^{p_i^\alpha+x_t-1}\exp\left(-\alpha_i\left(\beta_j\gamma+q_i^\alpha\right)\right),\end{aligned}$$

and

$$\alpha_i|\theta_{-\alpha_i},\mathbf{x},\mathbf{y}\sim\text{Gamma}\left(x_t+p_i^{\alpha_i},\hat{\beta}_j\hat{\gamma}+q_i^{\alpha_i}\right),$$

respectively. One can then do the same for each of the other four model parameters. We can then update the model using the sufficient statistics. In the case of α_i, the sufficient statistics are $x_t+p_i^{\alpha_i}$ and $\hat{\beta}_j\hat{\gamma}+q_i^{\alpha_i}$. The updates of the model statistics are given below

$$\begin{aligned}
p_t^{\alpha_i} &\leftarrow p_{t-1}^{\alpha_i}+x_t, & q_t^{\alpha_i} &\leftarrow q_{t-1}^{\alpha_i}+\hat{\beta}_{j,t-1}\hat{\gamma}_{t-1},\\
p_t^{\alpha_j} &\leftarrow p_{t-1}^{\alpha_j}+y_t, & q_t^{\alpha_j} &\leftarrow q_{t-1}^{\alpha_j}+\hat{\beta}_{i,t-1},\\
p_t^{\beta_i} &\leftarrow p_{t-1}^{\beta_i}+y_t, & q_t^{\beta_i} &\leftarrow q_{t-1}^{\beta_i}+\hat{\alpha}_{j,t-1},\\
p_t^{\beta_j} &\leftarrow p_{t-1}^{\beta_j}+x_t, & q_t^{\beta_j} &\leftarrow q_{t-1}^{\beta_j}+\hat{\alpha}_{i,t-1}\hat{\gamma}_{t-1}\\
p_t^{\gamma} &\leftarrow p_{t-1}^{\gamma}+x_t, & q_t^{\gamma} &\leftarrow q_{t-1}^{\gamma}+\hat{\alpha}_{i,t-1}\hat{\beta}_{j,t-1},
\end{aligned}$$

where the hat notation is as previous. This process is then iterated for each of the 380 league games in a Premier League season to find the model parameter estimates in the final game. This methodology is preferred to that of the classical approach, it allows us to use prior information whilst retaining accuracy and is also much faster than letting the model converge every season like the classical approach. Whilst we have created a model more beneficial than the classical approach we have yet to represent a team's current form, by updating the model throughout the seasons, this is something we will explore with the Dynamic Model.

7 The Dynamic Model

The Dynamic model will extend on the previous stationary model in order to improve the predictions we create. We no longer assume that model parameters are fixed over an entire season, and therefore allow them to change between games. This should more accurately represent the current form of any given team. We also continue with the Bayesian approach. Using the work of Gamerman (2013) [14], we create a state space model - and sequentially update it to give dynamic estimates of model parameters. Model parameters will be similar to those in Section 2, but will now depend on time - in order for them to be updated at each time point. We therefore let

- $\alpha_{i,t}$ be the attacking strength of team i in game t,
- $\alpha_{j,t}$ be the attacking strength of team j in game t,
- $\beta_{i,t}$ be the defensive strength of team i in game t,
- $\beta_{j,t}$ be the defensive strength of team j in game t,
- γ_t be the common home ground advantage.

where i, j represent the home and away teams respectively. Our dynamic model is then given by

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} | \kappa \sim \text{Poisson} \begin{pmatrix} \alpha_{i,t}\beta_{j,t}\gamma_t\epsilon_t \\ \alpha_{j,t}\beta_{i,t}\epsilon_t \end{pmatrix}, \tag{7.1}$$

where X_t represents the expected number of home goals in game t, Y_t represents the expected number of away goals in game t and $\epsilon_t \sim \text{Gamma}(\kappa, \kappa)$ is the shared random effect for game t.

7.1 Forgetting

Since we are now allowing our parameters to constantly change with time we will have a lot more results constantly changing between game days. This motivates the use of a forgetting parameter. Forgetting is a way to "forget" results that are in the distant past by lowering the weighting of the use of these results. This can help us better represent the current form more accurately since we are able to forget what happened earlier in the season. We apply forgetting to each of our sufficient statistics and this will be denoted by ω which is bounded by $0 \leq \omega \leq 1$. We later explain how optimal ω is found for each parameter.

7.2 Dynamic Model Priors

As with the stationary model, we can now assume the priors for the dynamic model. Since we are constantly updating each of the parameters, the priors will be denoted with t such that we know which prior represents which game The priors for game t

are given by

$$\alpha_{i,t} \sim \text{Gamma}(\tilde{p}^{\alpha}_{i,t}, \tilde{q}^{\alpha}_{i,t}),$$
$$\alpha_{j,t} \sim \text{Gamma}(\tilde{p}^{\alpha}_{j,t}, \tilde{q}^{\alpha}_{j,t}),$$
$$\beta_{j,t} \sim \text{Gamma}(\tilde{p}^{\beta}_{j,t}, \tilde{q}^{\beta}_{j,t}),$$
$$\beta_{i,t} \sim \text{Gamma}(\tilde{p}^{\beta}_{i,t}, \tilde{q}^{\beta}_{i,t}),$$
$$\gamma_{t} \sim \text{Gamma}(\tilde{p}^{\gamma}_{t}, \tilde{q}^{\gamma}_{t}),$$
$$\epsilon_{t} \sim \text{Gamma}(\kappa, \kappa).$$

As the priors update they are given by the previous posterior. This is denoted by the ˜ above each of the model parameters. Using the forgetting parameters with the previous posterior, these priors are calculated as such. This is known as the ***extension*** step.

$$\begin{aligned}
\tilde{p}^{\alpha}_{i,t} &= \omega p^{\alpha}_{i,t-1}, & \tilde{q}^{\alpha}_{i,t} &= \omega q^{\alpha}_{i,t-1},\\
\tilde{p}^{\alpha}_{j,t} &= \omega p^{\alpha}_{j,t-1}, & \tilde{q}^{\alpha}_{j,t} &= \omega q^{\alpha}_{j,t-1},\\
\tilde{p}^{\beta}_{j,t} &= \omega p^{\beta}_{j,t-1}, & \tilde{q}^{\beta}_{j,t} &= \omega q^{\beta}_{j,t-1},\\
\tilde{p}^{\beta}_{i,t} &= \omega p^{\beta}_{i,t-1}, & \tilde{q}^{\beta}_{i,t} &= \omega q^{\beta}_{i,t-1},\\
\tilde{p}^{\gamma}_{t} &= \mathrm{w} p^{\gamma}_{t-1}, & \tilde{q}^{\gamma}_{t} &= \mathrm{w} q^{\gamma}_{t-1},
\end{aligned}$$

where w denotes the fact that the forgetting parameter for the home ground advantage is different to that of the other parameters. As previous, for the first game of the season we set the sufficient statistics of p, q to be 10, and set the value of κ equal to 50. We now denote each of the updating model parameters as $\boldsymbol{\theta}_t = \{\alpha_{j,t}, \alpha_{i,t}, \beta_{i,t}, \beta_{j,t}, \gamma_t, \epsilon_t\}$.

The second part of the sequential update of the dynamic model is the ***update*** step. We denote the fixed parameters (ω, κ) by $\boldsymbol{\phi}$ and, as before, we use the joint posterior which is now given as

$$\begin{aligned}
\pi\left(\boldsymbol{\theta}_t|\mathbf{x},\mathbf{y},\boldsymbol{\phi}\right) \propto{}& \exp -\left[\epsilon_{\boldsymbol{t}}\left(\alpha_{i,t},\beta_{j,t}\boldsymbol{\gamma}+\alpha_{j,t}\beta_{i,t}\right)\right] \times \left[\epsilon_t\alpha_{i,t}\beta_{j,t}\right]^{\mathbf{x}_t} \times \left[\epsilon_t\alpha_{j,t}\beta_{i,t}\right]^{\mathbf{y}_t}\\
&\times \alpha_{i,t}^{\tilde{p}^{\alpha}_{i,t}-1}\exp\left(-\tilde{p}^{\alpha}_{i,t}\alpha_{i,t}\right)\alpha_{j,t}^{\tilde{p}^{\alpha}_{j,t}-1}\exp\left(-\tilde{p}^{\alpha}_{j,t}\alpha_{j,t}\right)\\
&\times \beta_{i,t}^{\tilde{p}^{\beta}_{i,t}-1}\exp\left(-\tilde{p}^{\beta}_{i,t}\beta_{i,t}\right)\beta_{j,t}^{\tilde{p}^{\beta}_{j,t}-1}\exp\left(-\tilde{p}^{\beta}_{j,t}\beta_{j,t}\right)\\
&\times \gamma_t^{\tilde{p}^{\gamma}_{t}-1}\exp\left(-\tilde{q}^{\gamma}_{t}\gamma_t\right)\\
&\times \epsilon_t^{\kappa-1}\exp\left(-\kappa\epsilon_t\right)
\end{aligned}$$

Using this posterior we can find the conditional posterior distributions for each of the model parameters to perform updates to the sufficient statistics, these updates are given by

$$\begin{aligned}
p^{\alpha}_{i,t} &\leftarrow \tilde{p}^{\alpha}_{i,t} + x_t, & q^{\alpha}_{i,t} &\leftarrow \tilde{q}^{\alpha}_{i,t} + \hat{\gamma}_t \hat{\beta}_{j,t},\\
p^{\alpha}_{j,t} &\leftarrow \tilde{p}^{\alpha}_{j,t} + y_t, & q^{\alpha}_{j,t} &\leftarrow \tilde{q}^{\alpha}_{j,t} + \hat{\gamma}_{i,t},\\
p^{\beta}_{i,t} &\leftarrow \tilde{p}^{\beta}_{i,t} + y_t, & q^{\beta}_{i,t} &\leftarrow \tilde{q}^{\beta}_{i,t} + \hat{\alpha}_{j,t},\\
p^{\beta}_{j,t} &\leftarrow \tilde{p}^{\beta}_{j,t} + x_t, & q^{\beta}_{j,t} &\leftarrow \tilde{q}^{\beta}_{j,t} + \hat{\gamma}_t \hat{\alpha}_{i,t},\\
p^{\gamma}_{t} &\leftarrow \tilde{p}^{\gamma}_{t} + x_t, & q^{\gamma}_{t} &\leftarrow \tilde{q}^{\gamma}_{t} + \hat{\alpha}_{i,t}\hat{\beta}_{j,t},\\
p^{\epsilon}_{t} &\leftarrow \kappa + x_t + y_t, & q^{\epsilon}_{t} &\leftarrow \kappa + \hat{\gamma}\hat{\alpha}_{i,t}\hat{\beta}_{j,t} + \hat{\alpha}_{j,t}\hat{\beta}_{i,t}.
\end{aligned}$$

We then iterate the ***extend*** and ***update*** step for the entirety of the season whilst also storing the previous information in order to be able to produce plots for the information over an entire season.

7.3 Model Averaging

Since we do not know what the optimum value for ω is at each time point we can create a mixture model of multiple different ω and assign weights to each in order to create a more accurate representation of form within teams. We determine the weighting of this model using the cumulative evidence. The cumulative evidence until time point t is calculated by

$$Z_{k,t} = \prod_{t^*=1}^{t} p\left(x_{t^*}, y_{t^*} | \mathbf{x}_{1:t^*-1}, \mathbf{y}_{1:t^*-1}, \phi_k\right), \tag{7.2}$$

where $k = 1, 2, \ldots, K$ represents the number of different ω we design, $Z_{k,t}$ represents the cumulative evidence and $p\left(x_{t^*}, y_{t^*} | \mathbf{x}_{1:t^*-1}, \mathbf{y}_{1:t^*-1}, \phi_k\right)$ represents the posterior predictive at each time point 1 to $t-1$. The posterior predictive at each time point until $t-1$ can be calculated as

$$\begin{aligned}
p\left(x_t, y_t | \mathbf{x}_{1:t-1}, \mathbf{y}_{1:t-1}, \phi\right) = {} & \frac{1}{x_t! y_t!} \times \frac{\Gamma\left(p^{\alpha}_{i,t}\right)}{\Gamma\left(\tilde{p}^{\alpha}_{i,t}\right)} \frac{(\tilde{q}^{\alpha}_{i,t})^{\tilde{p}^{\alpha}_{i,t}}}{(q^{\alpha}_{i,t})^{p^{\alpha}_{i,t}}} \times \frac{\Gamma\left(p^{\alpha}_{j,t}\right)}{\Gamma\left(\tilde{p}^{\alpha}_{j,t}\right)} \frac{(\tilde{q}^{\alpha}_{j,t})^{\tilde{p}^{\alpha}_{j,t}}}{(q^{\alpha}_{j,t})^{p^{\alpha}_{j,t}}}\\
& \times \frac{\Gamma\left(p^{\beta}_{i,t}\right)}{\Gamma\left(\tilde{p}^{\beta}_{i,t}\right)} \frac{(\tilde{q}^{\beta}_{i,t})^{\tilde{p}^{\beta}_{i,t}}}{(q^{\beta}_{i,t})^{p^{\beta}_{i,t}}} \times \frac{\Gamma\left(p^{\beta}_{j,t}\right)}{\Gamma\left(\tilde{p}^{\beta}_{j,t}\right)} \frac{(\tilde{q}^{\beta}_{j,t})^{\tilde{p}^{\beta}_{j,t}}}{(q^{\beta}_{j,t})^{p^{\beta}_{j,t}}}\\
& \times \frac{\Gamma\left(p^{\gamma}_{t}\right)}{\Gamma\left(\tilde{p}^{\gamma}_{t}\right)} \frac{(\tilde{q}^{\gamma}_{t})^{\tilde{p}^{\gamma}_{t}}}{(q^{\gamma}_{t})^{p^{\gamma}_{t}}} \times \frac{\Gamma\left(p^{\epsilon}_{t}\right)}{\Gamma(\kappa)} \frac{(\kappa)^{\kappa}}{(q^{\epsilon}_{t})^{p^{\epsilon}_{t}}}
\end{aligned}$$

Using the cumulative evidence calculated in Equation 7.2 we can calculate the weight of each model in the mixture. We begin with the model having an average of $\frac{1}{K}$ for each component so that they are equal. We then calculate the updated weighting at each time point as such

$$\Omega_{k,t} = \frac{Z_{k,t}}{\sum_{k=1}^{K} Z_{k,t}}, \tag{7.3}$$

where Ω now represents the updated weighting of each component. Using this fact we can now represent each of the attacking and defending parameters as a mixture of gamma distributions as such

$$\boldsymbol{\theta}_t|\boldsymbol{\phi} = \sum_{k=1}^{K} \Omega_{k,t}\boldsymbol{\theta}_{k,t}. \tag{7.4}$$

Our model of choice has 13 components of ω, ranging from 0.97 to 1, this is deemed as the best way to accurately represent form when using the RPS method. We also set w, the forgetting for the HGA to 1, in order to forget no information with regards to the HGA. These specific values were calculated using the Rank Probability Score, since we are predicting ahead of time we can use this score to find which mixture of values will provide us with the best prediction over every season and then apply these values as above.

$$Z_t^* = \prod_{t^*=1}^{t} \sum_{k=1}^{K} \Omega_{k,t-1} p\left(x_{t^*}, y_{t^*} | x_{1:t^*-1}, y_{1:t^*-1}, \phi_k\right) \tag{7.5}$$

The cumulative evidence for the mixture of models given in equation 7.5 is the sum of the predictive distributions weighted by the cumulative evidence up until the most recent observation. This mixture performs much better than any of the single forgetting values alone and we therefore proceed with the model averaging techniques.

7.4 Between Season Forgetting

If we were to analyse each season individually we would have to set p, q equal to 10 to form a Gamma$(10, 10)$ prior each season. Doing this causes inaccurate predictions in the early part of each season as we eluded to earlier, this is due to the fact that the best teams are currently ranked the same as the worst. To improve the model we can utilise the final posterior from the end of the previous season. Using this we can create a prior for the first game of the new season. This means that we now only have to assign a Gamma$(10, 10)$ prior to the newly promoted Premier League sides.

With the volatility of the Premier League and millions of pounds being spent each year, teams can change quite drastically between seasons. This can be caused by either the transfer market or with a change of manager. Meaning that we cannot be 100% confident that a team is as good as they were previously at the end of last season, or they could in fact be better. In order to allow for this change in ability we want to reduce the value of p, q whilst maintaining the same mean. We can do this by multiplying the sufficient statistics by a value, namely δ, which we now refer to as the between season forgetting. This allows us to increase the variance of the model parameters. We can also apply this theory to the home ground advantage and this change will be denoted by δ_{HGA}, which is now the between season forgetting for the HGA. To determine the optimum values of δ, δ_{HGA} we use the rank probability score from Section 5.2. RPS allows us to measure the accuracy of our predictions. Since the predictions for the dynamic model are calculated before the information is gained we can use this to compare multiple different values in order to optimise the amount of out of season forgetting, how we do this is later presented in Section 8. The optimum values for δ and δ_{HGA} are found to be 0.60 and 1 respectively. This eludes to the fact that the model wants to forget nearly half of the information that has previously been gained for the teams every year. However, the home ground advantage does not want to be forgotten between seasons, meaning that this value is much more likely to be constant over the 22 seasons of the Premier League.

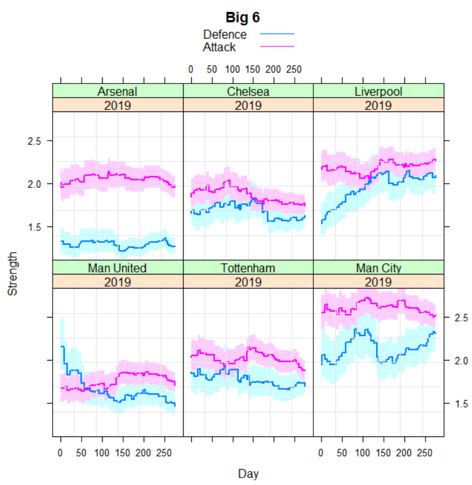

Figure 9: Parameter updates for the 2019 season of the “Big 6” teams.

In Figure 9 we can see the model updates for the 2019 season of the earlier mentioned Big 6 teams. We can see that the information from previous seasons is used to inform the prior at the first game of the season since none of the sides start at a score of 1 for either Attack or Defence. These graphs are not only great at representing a teams strengths and weaknesses, they are also very good at spotting trends within the footballing world. For instance, we can spot the point at which José Mourinho began to struggle as Manchester United manager, with his typical defensive style starting to fail and United's defensive score plummeting as a consequence. We can also see Manchester City's mid season blip in which they lost 3 in 4 games shipping 8 goals in the process around day 140 of the season. City managed to recover from this slip and went on to win the final 14 of their Premier League fixtures, whilst keeping 10 clean sheets in these games.

7.5 Results

Using all of the information we have gathered in the Dynamic model, we are now able to plot the entirety of the 22 Premier League seasons for any team that has played in the division. We utilise ggplot2 to create trellis plots for each team we choose to analyse.

Figure 10: Manchester United and Chelsea parameter updates over 22 seasons.

In Figure 10 we can see the parameter updates for Manchester United and Chelsea over the past 22 seasons whilst using the Dynamic model. In regards to Chelsea, we can see an extremely large peak between 2005 and 2006 that actually exceeds the graph limit of 3 for a couple of games. This peak coincides with the season in which Chelsea set the record for the least number of goals conceded in a season by a team in the Premier League. Under the masterful work of José Mourinho the team became an English power house, winning back-to-back Premier League titles in the mentioned seasons. Mourinho's work resonates through the squad for numerous years after his departure, with the squad still being a rigid, tough to score against side.

As mentioned previously, Mourinho also managed Manchester United, he managed to also improve a shaky United defence in 2018 before receiving the sack in his third season as manager in mid 2019. However, we cannot mention Manchester United without talking about Sir Alex Ferguson. Often regarded as the best manager to work in the Premier League, Ferguson took a struggling United side and transformed them in 13-time champions. In 2009 we can see a defensive spike for Manchester United - in this time period Edwin van der Sar kept 14 consecutive clean sheets, a record that still stands today. However, United have struggled as of late and this is clear to see with the defensive score dropping, whilst also not being able to score enough at the same time.

Figure 11: Liverpool and Manchester City parameter updates over 22 seasons.

In Figure 11 we can see the trellis plot for the title contenders of the 2019 season, Liverpool and Manchester City. Liverpool, now considered the club to copy in terms of the financial aspects behind a football team, have had an astronomical rise in ability within the past 3 seasons. Since the addition of Jürgen Klopp as manager the club has transformed and become one of the best in world football - recently winning the 2019 Champions League. We can see a huge rise in the defensive strength in the 2019 season. This is largely down to Liverpool defender Virgil van Dijk. Since he arrived at the club in January 2018 the defence has transformed and Liverpool have the best defensive record in the league over the last year and he is the bookies favourite to win the most prestigious individual prize in football, the Balon d'Or.

Focusing on Manchester City, not only can we see the astronomical rise in the club since the Saudi takeover in 2008, a process which saw Sheikh Mansour take over the club instantly making them billionaires, we can also see a lack of information for select seasons. The reason for this lack of information is because in these seasons City were relegated to the Championship and eventually League One. This lack of data paired with the fact that we would have to reset the prior when a new team gets promoted to the Premier League triggers the idea to extend the model to different leagues and use the data from these seasons to better inform the beginning of the next season in which the club restores it's Premier League status.

8 Extending the Model to Multiple Divisions

Motivated by the lack of information in some of Manchester City's seasons in Figure 11, we now extend our model to all 4 tiers of English football (Premier League, Championship, League One and League Two). The basis of our model will remain the same, but we are now able to produce predictions for each of the 92 football league sides. Since we are adding new leagues we will also define new parameters for the forgetting between seasons. We let

- δ_{PL} be the between season forgetting in the Premier League,
- δ_{CH} be the between season forgetting in the Championship,
- δ_{L1} be the between season forgetting in League One,
- δ_{L2} be the between season forgetting in League Two,
- $\delta_{HGA_{PL}}$ be the between season forgetting for the HGA in the Premier League,
- $\delta_{HGA_{CH}}$ be the between season forgetting for the HGA in the Championship,
- $\delta_{HGA_{L1}}$ be the between season forgetting for the HGA in League One,

- $\delta_{HGA_{L2}}$ be the between season forgetting for the HGA in League Two.

To decipher the optimal forgetting value for the between season forgetting of the teams and the home ground advantage, we find the value that minimises the Rank Probability Score as this indicates the best possible predictive ability of the model. We will find the optimal value to 2 d.p. as the tuning runs to find these values are computationally intensive and time consuming.

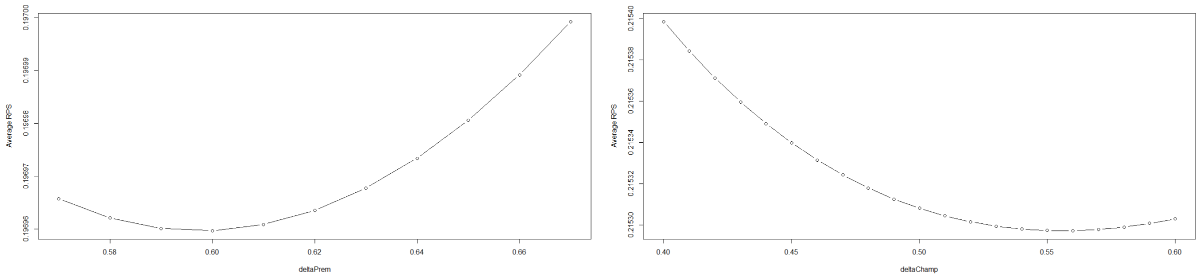

Figure 12: RPS plotted against forgetting value for δ_{PL} & δ_{CH}

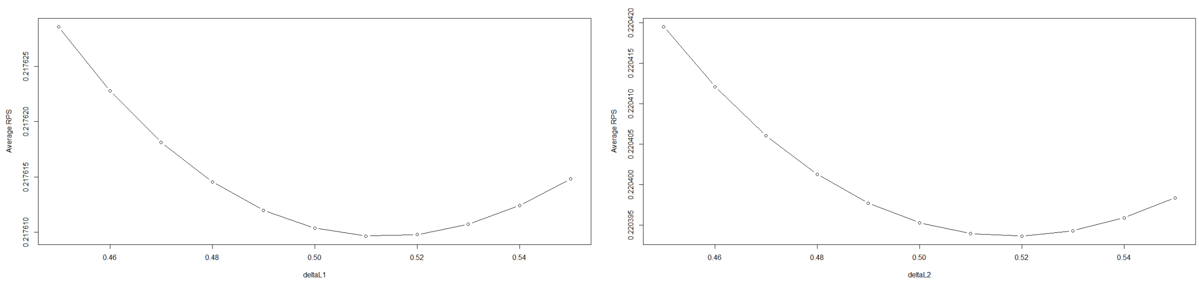

Figure 13: RPS plotted against forgetting value for δ_{L1} & δ_{L2}

In Figures 12 and 13 we can see the curves produced when plotting the RPS against the optimal forgetting value for δ_{PL}, δ_{CH}, δ_{L1}, δ_{L2}. We are looking for the minima of the curve as this is the point in which the RPS is minimised. The nature of the forgetting value is that there is an optimal point which produces the best predictions for each team overall, and above or below this point produces worse predictions. We are able to use the RPS as a measurement tool since our predictions are out of sample and ahead of each game, meaning that we are not informed before we make a prediction.

In Figures 14 and 15 we can see similar plots for $\delta_{HGA_{PL}}$, $\delta_{HGA_{CH}}$, $\delta_{HGA_{L1}}$, $\delta_{HGA_{L2}}$. We can again find the optimal value for each of these parameters using these graphs. It is important to notice Figure 14 which contains the optimal forgetting value for the HGA of the Premier League. This Figure does not actually find the minima of

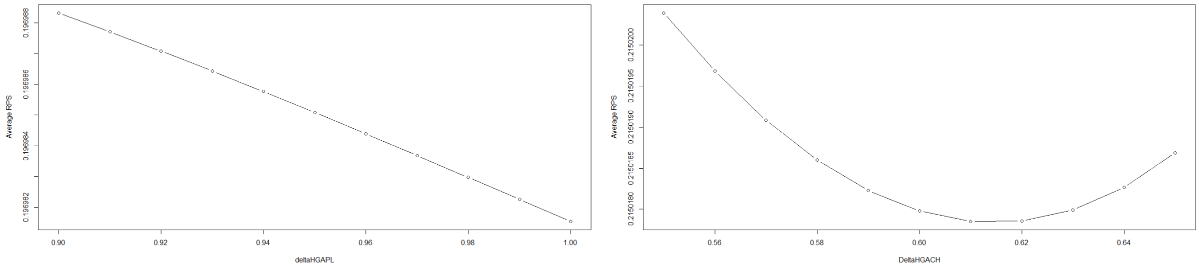

Figure 14: RPS plotted against forgetting value for $\delta_{HGA_{PL}}$ & $\delta_{HGA_{CH}}$

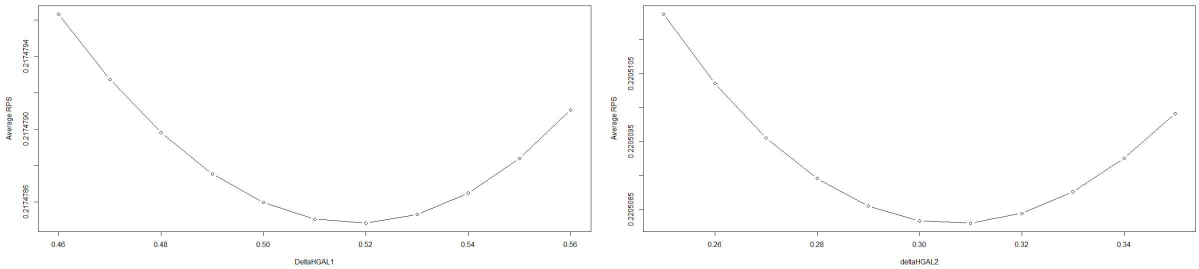

Figure 15: RPS plotted against forgetting value for $\delta_{HGA_{L1}}$ & $\delta_{HGA_{L2}}$

the curve, but instead finds the maximum value that the forgetting can take (1). Since the figure shows that the parameter is strictly decreasing we will continue by treating the optimal value as 1. However, it does indicate the fact that the home ground advantage in the Premier League is unlikely to change over each of the 22 previous seasons.

Parameter	**Optimum Forgetting**
δ_{PL}	0.60
δ_{CH}	0.56
δ_{L1}	0.51
δ_{L2}	0.52
$\delta_{HGA_{PL}}$	1
$\delta_{HGA_{CH}}$	0.61
$\delta_{HGA_{L1}}$	0.52
$\delta_{HGA_{L2}}$	0.31

Table 3: Results for the Optimum Forgetting of each of the between season parameters.

From Table 3 we can see the optimum forgetting values as defined from the previous Figures, these estimates are calculated by finding the optimal value for each parameter and iterating this process until we find convergence of the estimates. We can see that the forgetting between each of the 4 divisions is relatively similar with the Premier League retaining the most information. This does not come as a surprise as the top flight is home to the richest clubs in English football who tend to stay around the same ability, whilst other teams in lower divisions tend to change extremely often. However, the difference in the home ground forgetting between seasons is quite large. League Two forgets almost all of the information in regards to HGA with the optimal value being 0.31, whilst the Premier League chooses to forget none of the information between seasons. This could be due to the fact that teams usually don't have long stays in League Two with many teams 'yo-yoing' between divisions either above or below League Two.

We now deal with the problem of relegation and promotion between each of the divisions. In order for the model to work we must calculate each division seperately, doing this means that a team which is at the top of League 2 will have a similar Attacking and Defensive strength to a team that is at the top of the Premier League. Of course, this is not actually the case, in reality there would be an extreme difference in the ability of these two sides. Therefore, we must determine a way to alter the Attacking and Defensive strength of a side when that team is promoted or relegated from a division. To do this we will multiply the sufficient statistic p for both the attacking and defensive strength. We will therefore let

- $ATT_{PL,CH}$ be the transformation to the attacking strength of a team swapping between the Premier League and the Championship,
- $ATT_{CH,L1}$ be the transformation to the attacking strength of a team swapping between the Championship and League One,
- $ATT_{L1,L2}$ be the transformation to the attacking strength of a team swapping between League One and League Two,
- $DEF_{PL,CH}$ be the transformation to the defensive strength of a team swapping between the Premier League and the Championship,
- $DEF_{CH,L1}$ be the transformation to the defensive strength of a team swapping between the Championship and League One,
- $DEF_{L1,L2}$ be the transformation to the defensive strength of a team swapping between League One and League Two,

where each of the above represent the transformation to p for the respective division, we not that teams changing division also still get mulitplied by the between season forgetting of the league they were in. To decipher the optimal value for these transformations we again use the RPS and the minima of the curve that it creates.

At this point it is important to note again that the lower a defensive strength the better and that multiplying this value by a number greater than 1 would indeed make the defensive strength worse. Teams promoted to the Premier League from the Championship will be mulitplied by this parameter such that their ability decreases, whilst teams relegated to the Championship from the Premier League will have their sufficient statistic divided by this value such that their ability increases. This idea comes from the fact that a team at the bottom of a division will usually become one of the better teams in the division below. The optimal values are given in Table 4.

Parameter	**Optimum Value**
$ATT_{PL,CH}$	0.63
$ATT_{CH,L1}$	0.81
$ATT_{L1,L2}$	0.91
$DEF_{PL,CH}$	1.32
$DEF_{CH,L1}$	1.15
$DEF_{L1,L2}$	1.08

Table 4: Optimal Values for the transformation of the sufficient statistic p

From Table 4 we can now make inference on how difficult the step between divisions really is. We can see that the step between the Championship and the Premier League is in fact the largest. This is to be expected as the top division of English football is regarded to be one of, if not the, best division in the world. We can also see that the change between League One and Two is not large, meaning that there is not much change in ability between teams playing in the respective leagues.

We can now compare the previous Dynamic model which contains just the Premier League to the new Dynamic model containing each of the divisions. We will compare the predictive ability of each model throughout the last 22 seasons using the RPS. We would like the RPS to be lower in the new Dynamic model, in Figure 16 we can see that this is in fact the case. In the early seasons the new dynamic model performs a lot better than that of the model before extension. This gap slowly decreases over time, but the extended model still performs better in the majority of the seasons in which we have analysed. The fact that the gap is so large at the start and smaller at the end may point towards us including too many seasons in our analysis. This would mean that we would use less information from 20 years ago, which may be wise as teams tend to change a lot within a 20 year period.

In Figure 17 we can see the average RPS per game comparison between each of the

Figure 16: RPS comparison of the dynamic model before and after league extension.

4 divisions over the previous 22 years. We can see that the "easiest" division for the model to predict is the Premier League, whilst the other 3 divisions are relatively similar. The 2010 season seems to be one of the best predicted seasons over the 4 divisions overall, with only League Two results being relatively difficult to predict in comparison to others. For a betting strategy, this would indicate that the Premier League is the best to predict, whilst we should probably avoid League Two.

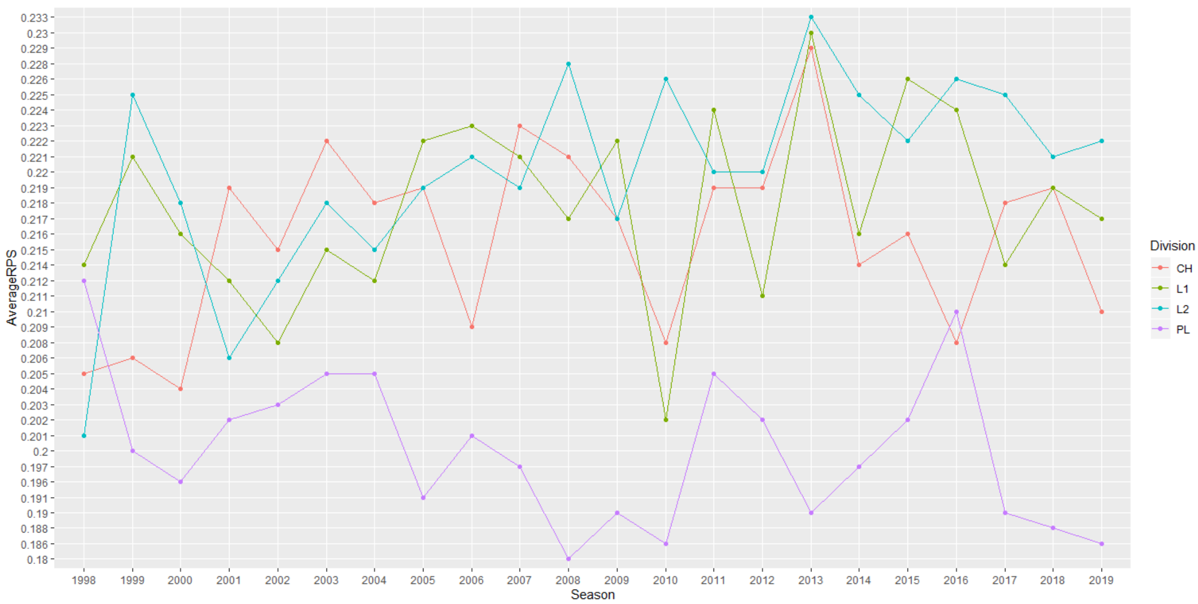

Figure 17: RPS Comparison between divisions.

We can now repeat the Liverpool and Manchester City plot from Figure 11 and predict the seasons in which they weren't in the top flight. In Figure 18 we can see this updated plot. To recognise plots from the new extended dynamic model by having the plots be coloured in the home strip colours of each respective side. We can see that City spent a year in League One as recently as 1999, just 20 years ago, they also struggled in the Premier League for multiple years after their return. In the 2002 season City were relegated to the Championship but managed to get promoted back to the Premier league after maintaining an attacking strength of around 2 for the entirety of the season. These new plots will not only be useful for plotting teams who are in the Premier League, we can now visualise the astronomic rise of clubs through the tiers of English football as well as some of the lows clubs may face.

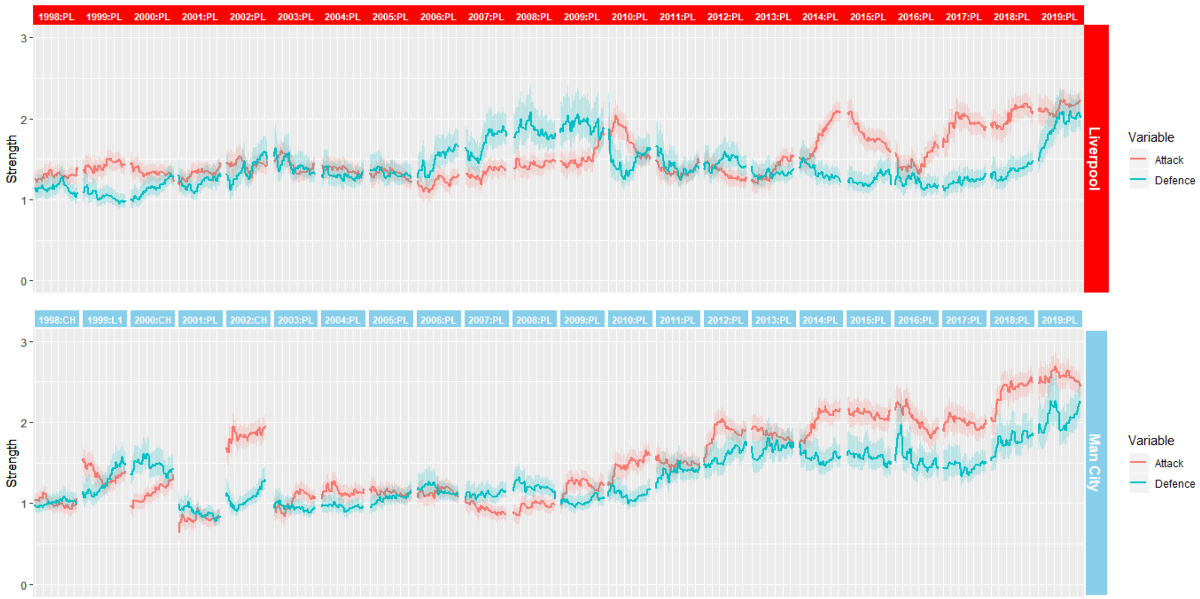

Figure 18: Liverpool and Manchester City parameter updates over 22 seasons with the new dynamic model.

Two perfect examples of this are Bournemouth and Leicester City. Bournemouth, now considered the model to follow for smaller clubs looking to rise through the divisions were in League Two as recently as 2010. In Figure 19 we can visualise this astounding run of Bournemouth under revolutionary young manager Eddie Howe. Even though the club is constantly changing division, we can see the consistency of the club to maintain both it's attacking and defending strength. Partnered with the outrageously high scoring season of 2015 in the Championship this was enough to see Bournemouth establish themselves as a Premier League side. Bournemouth have gone on to become a great Premier League side no longer toying with relegation and Eddie Howe looks like a man about to be head-hunted into a "bigger" job.

Another fairytale story is that of Leicester City. The club has "chopped and changed" between the Championship and the Premier League throughout the past 22 seasons. However, the unthinkable happened in 2016, just one season after their top flight status was restored the club managed to lift the league title. The club were 5000/1 to win the league at the beginning of the season and have since changed betting, with many companies offering no higher than 2000/1 on league winner bets. We can see from Figure 19 that Leicester weren't in fact great scorers or even great at keeping clean sheets, but their ability to grind out 1-0 victories and a remarkable goal scoring run from Jamie Vardy who scored in 11 consecutive games was enough to lead them to the elusive Premier League title.

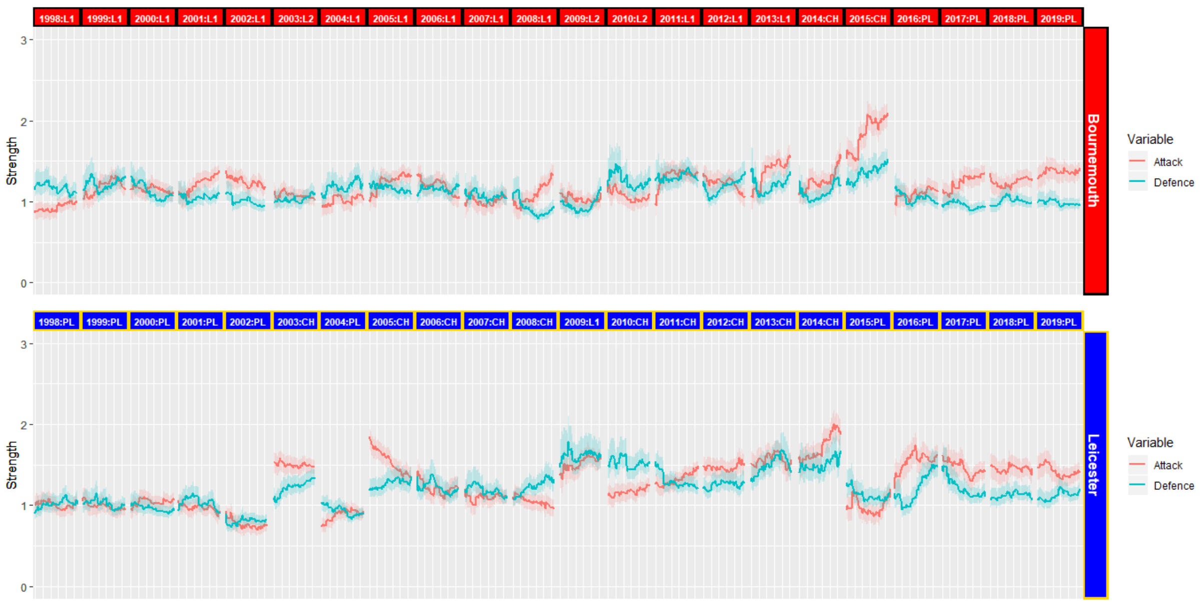

Figure 19: Bournemouth and Leicester City parameter updates over 22 seasons with the new dynamic model.

9 Discussion

The statistical processes within the paper have been carried out to a high and meticulous standard throughout. However, this level of detail takes a long time and we are not fully able to undergo every process that we would like. Given further study, the following ideas should be considered:

1. Comparison of model predictions against bookmakers betting odds and possibly devise a betting strategy.

2. Try to find the optimal number of seasons to use within our analyses.

3. Identify why the home ground advantage of the Premier League in the extended model is not a minima of a curve.

4. Compare RPS between each of our models and others in Literature to determine which performs best.

5. Perform more in depth analysis on the bivariate Poisson and bivariate negative binomial.

6. Explore other factors that could effect the number of goals a team scores e.g. own goals.

10 Conclusion

Firstly, we will discuss the models which utilised a classical approach. The univariate Poisson from Section 2 is a simple model which helped begin our analysis. For it's simplicity it is an extremely effective model and a helpful starting point for anyone who would like to enter the world of sports modelling. It is a simple enough model for someone new to code and understand. However, this simplicity of course means that there is room for improvement. This improvement comes in the form of the bivariate Poisson in Section 3. This model helped us better represent the correlation between home and away goals, by introducing a new parameter to allow for this dependence. It also managed to improve on the univariate Poisson, but did not accurately represent the current form of a team.

To better represent the current form of a team we expanded our view to a Bayesian methodology and a dynamic model. We have managed to create a model that utilises state-spaces to sequentially update parameters over a given season. This model is a reasonably accurate one as it produces predictions with a good rank probability score, and is certainly an improvement on the earlier models. We were able to produce plots of the last 22 seasons for any football club in the English system - which allowed us to make inference into why certain peaks and troughs occur. We can now produce predictions for each game week in every division of English football, and after more work we will be able to create a betting strategy. Although we are not able to say whether this strategy would be profitable, we have achieved our goal of being able to make these reasonably accurate predictions. Ultimately, with an ever-changing football landscape there is always more research to be done.

References

[1] FIFA. More than half the world watched record-breaking 2018 world cup. https://www.fifa.com/worldcup/news/more-than-half-the-world-watched-record-breaking-2018-world-cup, 2018.

[2] Gambling Commission. Review of online gambling. http://www.gamblingcommission.gov.uk/PDF/Online-review-March-2018.pdf, 2018.

[3] M. J. Dixon and S. 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.

[4] D. Karlis and I. Ntzoufras. Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, 20(2):133–145, 09 2008.

[5] D. Jankov Maširević, R. K. Parmar, and T. K. Pogány. (p, q)-extended bessel and modified bessel functions of the first kind. *Results in Mathematics*, 72(1):617–632, Sep 2017.

[6] C. B. Dean. Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457, 1992.

[7] K. S. Gyu and K. J. Hae. Central limit theorem: the cornerstone of modern statistics. *Korean J Anesthesiol*, 70(2):144–156, 2017.

[8] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

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

[10] H. Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, Sep 1987.

[11] J. Kuha. Aic and bic: Comparisons of assumptions and performance. *Sociological Methods & Research*, 33(2):188–229, 2004.

[12] A. C. Constantinou and N. E. Fenton. jqas. 8, 2019 2012. 1.

[13] E. S. Epstein. A scoring system for probability forecasts of ranked categories. 8(6):985–987, 1969.

[14] D. Gamerman, T. R. dos Santos, and G. C. Franco. A non-gaussian family of state-space models with exact marginal likelihood. *Journal of Time Series Analysis*, 34(6):625–645, 2013.