An exploration of predictive football modelling

Article *in* Journal of Quantitative Analysis in Sports · December 2019

DOI: 10.1515/jqas-2019-0075

3 authors, including:

Glen Livingston Jr
The University of Newcastle, Australia
35 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Robert King
The University of Newcastle, Australia
91 PUBLICATIONS **1,352** CITATIONS
SEE PROFILE

1 Introduction

Given the popularity of association football; coaching staff, spectators, and football pundits have an interest in predicting the outcomes of matches as well as an interest in what factors influence results. As a consequence of this curiosity, several studies have been published attempting to quantify the game and develop models that can accurately predict the result.

The first predictive football model was proposed by Maher (1982) using an independent Poisson model with means representing attacking and defensive strengths for each team. Maher additionally allows for a home ground advantage and determines that it is statistically significant.

In Dixon and Coles (1997), the authors introduce two adjustments to improve the independent Poisson model proposed by Maher (1982). Firstly, they allow for dependence between teams via an ad hoc correction which adjusts the probabilities for low scoring matches. Secondly, a decaying time-weighting function is included which causes matches further in the past to have less influence than most recent matches on the parameter estimates.

Rue and Salvesen (2000) adopt a Bayesian framework to study the variation of team strengths over time. Incorporating the framework of both Maher (1982) and Dixon and Coles (1997), they also introduce an adjustment which accounts for stronger teams underestimating weaker opponents and conversely, weaker opponents preparing more when facing a stronger team in the league. Similarly, Owen (2011) also utilises Bayesian dynamic generalised linear models to allow the abilities of each team to vary over time. Owen performs his analysis on data from the Scottish Premier League finding improvements in comparison to the performance of non-dynamic models.

The bivariate Poisson distribution suggested by Maher (1982) is employed by Koopman and Lit (2015). Utilising a state space model, the authors develop a statistical model where the strength parameters for each team are allowed to vary stochastically with time. Most recently, Boshnakov et al. (2017) propose an alternative distribution to the Poisson. The authors employ a count process derived when the inter-arrival times are assumed to follow an independent and identically distributed Weibull distribution. A copula is utilised to generate a bivariate distribution to account for dependence between the home and away teams. This is compared to both the independent and copula-induced bivariate Poisson model and is found to provide superior results.

Our objectives in this paper are to implement and compare the results of the Independent Poisson Model (Maher, 1982), the Dixon-Coles Model (Dixon and Coles, 1997), and the Bivariate Weibull Count Model (Boshnakov et al., 2017). The models are applied to the top five European leagues, namely: the English Premier League (EPL); French Ligue 1; German Bundesliga; Italian Serie A; and, Spanish La Liga. A total of six seasons from 2012 - 2018 are utilised from each league to perform the study. For the sake of brevity, for each model only the estimation results for the EPL are presented. We also propose factoring in the effects of travelling in football and highlight this using data from the Australian A-League and Russian Premier League.

The remainder of this paper is structured as follows. Section 2 presents the independent Poisson, Dixon-Coles and bivariate Weibull count models in detail. Implementation of these models using the statistical software R is demonstrated. In Section 3 the rank probability score is introduced and utilised to compare the three models across the five league's respective out-of-sample validation sets. Additionally, the rank probability score is proposed to improve the optimisation of the Dixon-Coles time-weighting parameter. Section 3.3 aims to further quantify the home advantage parameter by accounting for the distance travelled by the away team. Finally, Section 4 concludes and provides potential research to further improve the current models in the literature.

2 The Models

2.1 Independent Poisson Model

2.1.1 The Model

In what is considered the seminal paper in predictive football modelling, Maher (1982) proposed the independent Poisson model where in a given match, the number of goals scored by both the home and away team follow independent Poisson distributions. While the assumptions for this model are debatable the Poisson distribution provides a reasonable approximation for the number of goals scored in a football match and has been adopted by numerous researchers.

In Maher's model, the rate at which a team is expected to score is a function of both their own offensive ability and their opponent's defensive ability. Specifically, in a match between the home team i and the away team j, let $(X_{i,j}\ ,\ Y_{i,j})$ represent the number of goals scored by the home and away team respectively. Then;

$$\begin{aligned} X_{ij} &\sim \text{Poisson}(\lambda = e^{\alpha_i+\beta_j+\gamma}) \\ Y_{ij} &\sim \text{Poisson}(\mu = e^{\alpha_j+\beta_i}) \end{aligned} \tag{2.1}$$

where α_i and β_i are measures of the attacking and defensive strengths of the i^{th} team respectively whilst γ is a home advantage parameter. In this implementation the exponential is analogous to a link function, ensuring that both λ and μ remain non-negative. Observing (2.1), it is evident that smaller values of β represent stronger defences since larger values of β result in an increase of the expected number of goals for the opponent. This model can be implemented via a Poisson regression where the attacking and defensive ability along with home advantage are explanatory variables and the response variable is the number of goals scored.

2.1.2 Results

Figure 1 presents the estimated strength parameters for all teams in the EPL. Note that the defensive strength parameters have been multiplied by negative one to ensure the stronger teams are shown in the first quadrant of the plot. As expected the notable teams in the league are rated towards the upper right hand part of the plot indicating strong attack and defence.

Table 1 contains the home advantage parameter for each of the five leagues also. There appears to be a positive home advantage for all five leagues. This is consistent with the research of Pollard (1986) and Clarke and Norman (1995) that a home advan-

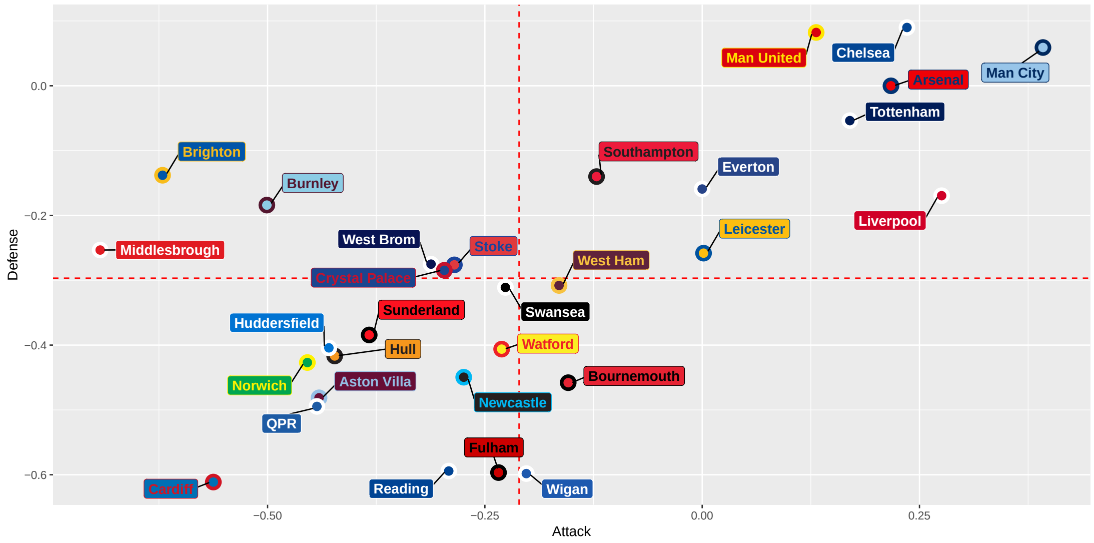

Figure 1: Scatter Plot of the Independent Poisson Strength Parameters for English Premier League Teams

tage exists in football. Performing an analysis of deviance on these models reveal that the three explanatory variables utilised in this analysis are statistically significant for all five leagues.

	England	France	Germany	Italy	Spain
γ	0.2526847	0.3055232	0.2485644	0.2486163	0.3278102

Table 1: Home Advantage Parameters for all Leagues

2.2 Dixon-Coles Model

2.2.1 The Model

Dixon and Coles (1997) analyse data from 1992-1995 consisting of all matches from the top four tiers of English football. For this data set, they assess the assumption of independence by observing the ratio of the joint empirical probability function and the product of the home and away marginal empirical probability functions. Specifically;

$$\frac{\tilde{f}(i,j)}{\tilde{f}_H(i)\cdot\tilde{f}_A(j)} \tag{2.2}$$

where $i = 0, ..., 6$, $j = 0, ..., 5$ and $\tilde{f}$, $\tilde{f}_H$ and $\tilde{f}_A$ are the joint and marginal empirical probability functions for home and away scores respectively. Observing these ratios for the specified number of home and away goals, they find that the independence assumption is valid with the exception of low scoring games; 0-0, 1-0, 0-1 and 1-1. Particularly, the draws 0-0 and 1-1 appear to be underestimated whilst the scores 1-0 and 0-1 are overestimated.

With this in mind, Dixon and Coles (1997) decided to model the departure from independence for low scoring games. This modification to the independent model is as follows:

$$\Pr(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\,\mu}(x, y)\frac{\lambda^x e^{-\lambda}}{x!}\frac{\mu^y e^{-\mu}}{y!} \tag{2.3}$$

where $\lambda = e^{\alpha_i+\beta_j+\gamma}$, $\mu = e^{\alpha_j+\beta_i}$, and

$$\tau_{\lambda,\,\mu}(x, y) = \begin{cases} 1 - \lambda\mu\rho & \text{if } x = y = 0, \\ 1 + \lambda\rho & \text{if } x = 0,\ y = 1, \\ 1 + \mu\rho & \text{if } x = 1,\ y = 0, \\ 1 - \rho & \text{if } x = y = 1, \\ 1 & \text{otherwise.} \end{cases} \tag{2.4}$$

The ad hoc correction $\tau_{\lambda,\,\mu}(x, y)$ is reliant on the dependence parameter ρ which determines the extent of how the low scoring probabilities change, with $\rho = 0$ representing the independent Poisson model in Maher (1982).

Upon observation of (2.3), there are now $2n + 2$ parameters to be estimated where n denotes the number of teams. Specifically, there are n attack parameters $(\alpha_1, ..., \alpha_n)$, n defence parameters $(\beta_1, ..., \beta_n)$ the home advantage parameter γ and lastly the dependence parameter ρ. To ensure that the model is not over-parametrised, the following constraint is included;

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i = 1. \tag{2.5}$$

To estimate these parameters, maximum likelihood estimation is utilised.

A further extension of the independent Poisson model from Dixon and Coles (1997) is to account for the fluctuation of team performance throughout each season. To allow for time-varying team strengths, a time decay factor is introduced into the likelihood function such that more recent matches are weighted greater than those results further in the past. This adjustment results in the following pseudo-likelihood for each time point t,

$$L_t(\alpha_i, \beta_i, \rho, \gamma; i = 1, ..., n) = \prod_{k \in A_t} \left(\tau_{\lambda_k, \mu_k}(x_k, y_k) e^{-\lambda_k} \lambda_k^{x_k} e^{-\mu_k} \mu_k^{y_k} \right)^{\phi(t-t_k)} \tag{2.6}$$

where t_k is the time that match k was played, $A_t = \{k : \ t_k < t\}$ and $\phi(t) = e^{-\xi t}$. The time weighting function, $\phi(t)$, is a non-increasing function. It is dependent on parameter $\xi \geq 0$ with $\xi = 0$ representing no time weighting. As stated in Dixon and Coles (1997), optimising ξ is problematic since (2.6) defines a sequence of non-independent likelihoods and can be trivially maximised by simply increasing ξ arbitrarily. Therefore, ξ is instead chosen such that the predictive capability of the model is maximised. Dixon and Coles (1997) define the following equation which corresponds to a predictive profile log-likelihood;

$$S(\xi) = \sum_{k=1}^{N} \left(\delta_k^H \log p_k^H + \delta_k^D \log p_k^D + \delta_k^A \log p_k^A \right) \tag{2.7}$$

where p_k^H, p_k^D and p_k^A are the maximum likelihood estimates of a home win, draw and away win from (2.6) respectively. The δ_k's are binary variables indicating the true result of match k. For instance, $\delta_k^H = 1$ if match k is a home win and $\delta_k^H = 0$ otherwise. For their analysis, the authors utilised half-weeks as their unit of time and found that the value $\xi = 0.0065$ maximised equation (2.7).

2.2.2 Results

The time-weighting parameter ξ cannot be optimised along with the other parameters. Alternatively, ξ is held constant while the remaining parameters are estimated. To determine ξ for each league, a finite range of ξ values are chosen and the value of ξ that maximises the objective function (2.7) is used.

To assess the predictive capability of the model, cross-validation is utilised such that the parameters of the model are optimised given a training dataset and $S(\xi)$ is evaluated on a validation set independent of the optimisation. This process proves to be computationally demanding. This is because the parameters of model (2.6) need to be estimated for each time point t that a prediction is made using only match scores prior to that time. Thus, the model will be iteratively optimised for each value of t in the chosen validation set to calculate $S(\xi)$. To surmount the computational burden of this procedure, the model was optimised for each league across a range of ξ values utilising parallel computing on The University of Newcastle Research Computer Grid.

To perform the cross-validation, the first one and a half seasons of each league were utilised as an initial training set for which the parameter estimates of model (2.6) were obtained. The second half of seasons two through to five as well as the second quarter of the sixth season were then utilised as a validation set. In contrast to Dixon and Coles (1997), we have utilised days as the time unit of t rather than half-weeks. Hence, the parameter estimates acquired from the initial training set are used to predict the matches played on the first date of the validation set. Once $S(\xi)$ has been calculated for that set of matches, these observations are then included in the training set and the model is re-evaluated and used to predict the subset of fixtures corresponding to the next date in the validation set. The first half of each season (and first quarter with respect to the sixth season) have been omitted from the validation set to avoid issues pertaining to newly promoted teams with absence of data. Figure 2 is a plot of the predictive profile log-likelihood as it varies across the relevant ranges of ξ for each league.

Initially, $S(\xi)$ for each league was calculated for $0 \leq \xi \leq 0.005$ in increments of 0.0001. Observing Figure 2 shows that a much wider range of ξ values were investigated for the EPL. This is because within the range $0 \leq \xi \leq 0.005$, $S(\xi)$ reached its maximum at 0.005. Hence, the range of ξ values was extended to 0.01. The optimisation algorithm failed to converge for many values beyond 0.0083 and thus, results were truncated at $\xi = 0.0083$. This value is suspiciously large in comparison to the values identified for the other leagues, therefore this truncation seems to be reasonable. Extremely large values of ξ could leave the model prone to over fitting and hence, poor predictive capability.

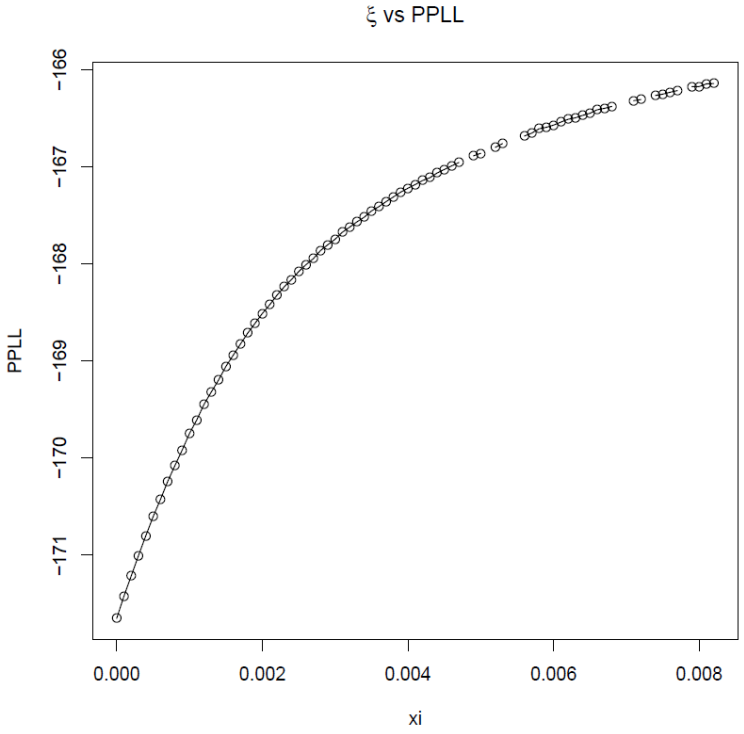

Figure 2: ξ vs PPLL for English Premier League Teams

The problem of estimating ξ for the EPL data is addressed later by using the rank probability score for choosing the optimal value.

Table 2 presents the parameter estimates of γ, ξ, and ρ for the five leagues. Interestingly, $S(\xi)$ is maximised at $\xi = 0$ for both the Italian Serie A and Spanish La Liga. This implies that the static model has the best predictive capability and exponentially down-weighting past performances makes no improvement to either model. These results are perhaps indicative of each league's dynamics since both leagues have had rather stable tiers in terms of team strengths over the six seasons of data. However, it is expected that at least some variation in team strengths should occur over time, especially since there are six seasons worth of data. The issue relating to these ξ values is discussed further in section 3.2 where an improvement regarding their optimisation is proposed.

The parameter estimates for the Dixon-Coles model for the EPL are presented in Figure 3. Comparing this with Figure 1, some notable changes in positions are for

	England	France	Germany	Italy	Spain
γ	0.20655	0.31597	0.29919	0.24857	0.32773
ξ	0.0083	0.0031	0.0037	0.00000	0.00000
ρ	-0.03832	-0.01572	-0.09661	-0.05628	0.00458

Table 2: Home Advantage, Time-Weighting and Low Scoring Dependence Parameters for all Leagues

Arsenal, Cardiff and Southampton. Looking at the performance of both Arsenal and Southampton in the later years analysed their final ladder positions fell. As the Dixon-Coles model provides greater weighting to the more recent seasons, we see that the positions for both of these teams moves towards the third quadrant compared to the Independent Poisson model's results. This is particularly noticeable for Southampton who finished 17th in the last season compared to 8th and 6th the two seasons before that. Cardiff was only in the EPL for one season (2013-2014). As a consequence there is a significant amount of uncertainly surrounding the parameter estimates and this large change is a reflection of that.

Observing the ρ parameter for each league in Table 2, with the exception of the Spanish La Liga all leagues have negative values. This indicates that these leagues have an inflation of 1-0 and 0-1 scores in comparison to 0-0 and 1-1 draws. The Spanish La Liga yields a positive ρ parameter, however, is very small indicating only a small departure from independence in the low scoring games.

2.3 Bivariate Weibull Count Model

2.3.1 The Model

The final model analysed is the bivariate Weibull count model discussed in Boshnakov et al. (2017). The probability mass function for the Weibull count model is as follows;

$$\Pr(X(t) = x) = \sum_{j=x}^{\infty} \frac{(-1)^{x+j}(\lambda t^c)^j \alpha_j^x}{\Gamma(cj+1)} \tag{2.8}$$

where $\alpha_j^0 = \frac{\Gamma(cj+1)}{\Gamma(j+1)}$ for $j = 0, 1, 2, ...$, and $\alpha_j^{x+1} = \sum_{m=x}^{j-1} \alpha_m^x \frac{\Gamma(cj-cm+1)}{\Gamma(j-m+1)}$ for $x = 0, 1, 2, ...$ and $j = x+1, x+2, x+3, ...$ In (2.8), λ serves as the rate parameter while c is the

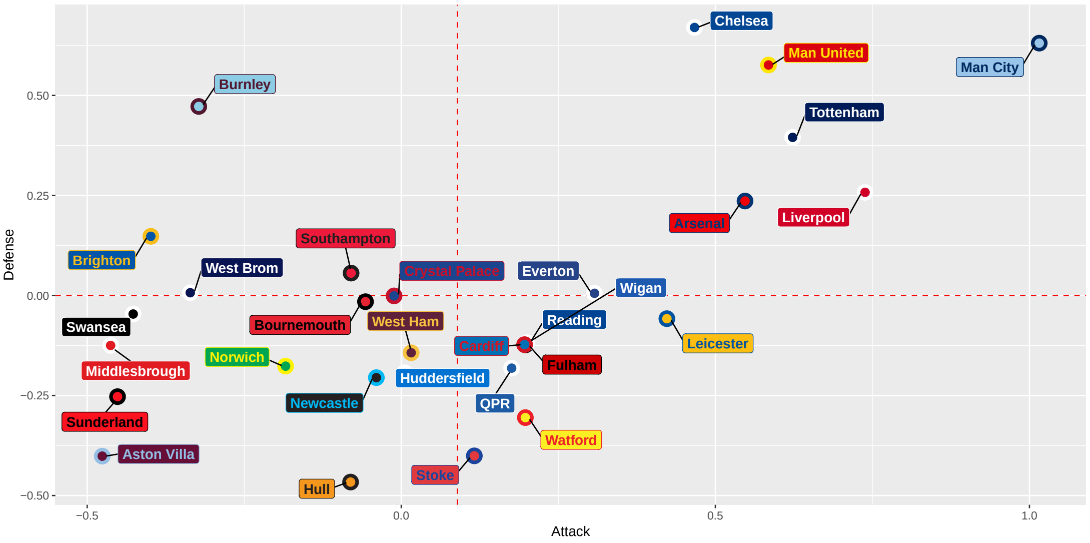

Figure 3: Scatter Plot of the Dixon-Coles Strength Parameters for English Premier League Teams

shape parameter of the distribution.

In the context of football modelling, each match represents an observation which is assumed to have a duration of 1 time unit. Thus, λ can be interpreted as the rate at which goals are scored per match. As in both Maher (1982) and Dixon and Coles (1997), the rate parameter for the home team i playing against the away team j is $\lambda = e^{\alpha_i+\beta_j+\gamma}$ and similarly, the away team's rate parameter is $\mu = e^{\alpha_j+\beta_i}$.

An advantage that the Weibull count distribution has over the Poisson is that it more flexible due to its shape parameter c. Via $c < 1$ or $c > 1$, the distribution is able to capture over-dispersed and under-dispersed data respectively. For $c = 1$, the Weibull count distribution is in fact equivalent to the Poisson and hence, represents equi-dispersed data.

The bivariate distribution is generated through the use of a copula. According to Sklar's theorm (Sklar, 1973), the joint cumulative distribution function F of any pair of random variables (X, Y) can be written as;

$$F(x, y) = C\big(F_1(x), F_2(y)\big), \quad (x, y) \in \mathbb{R}^2, \tag{2.9}$$

where F_1 and F_2 are the respective marginal cumulative distribution functions of x and

y and C is a copula.

Frank's copula is employed due to it's ability to allow for both positive and negative dependence. This is defined as follows

$$C(u,v) = -\frac{1}{\kappa}\log\left(1 + \frac{(e^{-\kappa u}-1)(e^{-\kappa v}-1)}{e^{-\kappa}-1}\right), \tag{2.10}$$

where $\kappa \in \mathbb{R}\backslash\{0\}$ is the dependence parameter. Thus, utilising Frank's copula $C(u,v;\kappa)$ to combine the two cumulative Weibull count distribution functions $F_1(x;\lambda,c_H)$ and $F_2(y;\mu,c_A)$, the likelihood function for the parameter vector $(\lambda,\mu,c_H,c_A,\kappa)$ for the i^{th} match (x_i,y_i) is:

$$\begin{aligned} L(\lambda,\mu,c_H,c_A,\kappa;x_i,y_i) &= \Pr(X=x_i,Y=y_i) \\ &= C(F_1(x_i),F_2(y_i)) \\ &- C(F_1(x_i-1),F_2(y_i)) \\ &- C(F_1(x_i),F_2(y_i-1)) \\ &+ C(F_1(x_i-1),F_2(y_i-1)), \end{aligned} \tag{2.11}$$

where c_H and c_A represent the home and away shape parameters respectively.

It then follows that the log-likelihood function over a sample of n matches is simply $\ell(\lambda,\mu,c_H,c_A,\kappa;\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\log(L)$. For the bivariate Weibull count model there are $2n+4$ parameters to be estimated. Specifically, there are n attack parameters $(\alpha_1,...,\alpha_n)$, n defence parameters $(\beta_1,...,\beta_n)$ the home advantage parameter γ, the dependence parameter κ and lastly the home and away shape parameters c_H and c_A. Much like Dixon and Coles (1997), Boshnakov et al. (2017) use numerical maximisation to maximise the (log-) likelihood function and find the required parameter estimates.

Similar to Dixon and Coles (1997), Boshnakov et al. (2017) utilise a time weighting function to exponentially down-weight matches further in the past. Due to the computational burden of estimating the time weighting parameter ξ for the bivariate Weibull count model, this parameter is not estimated. The optimum value from the Dixon-Coles model is used.

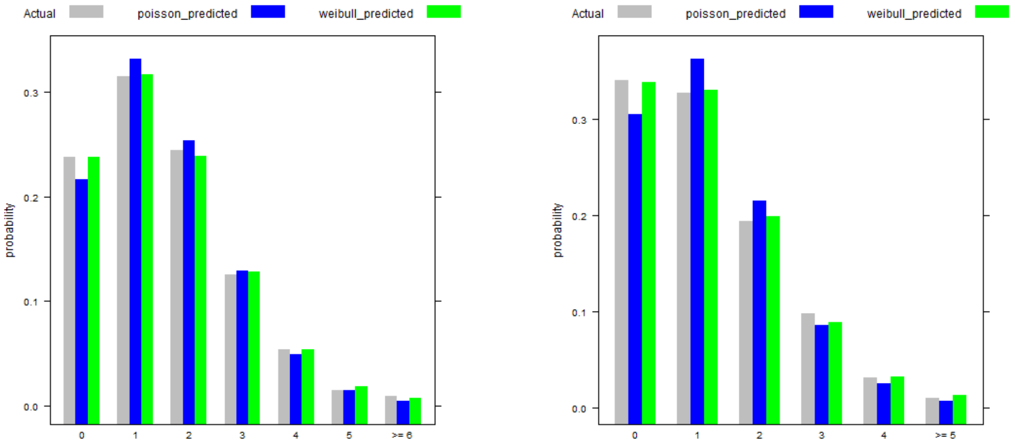

Figure 4: Histrograms for English Premier League home (left) and away (right) goals with the fitted Poisson and Weibull count models

	Home Goals			**Away Goals**		
Model	DF	χ^2	P-value	DF	χ^2	P-value
Poisson	5	11.87	0.03662	5	28.866	0.00002464
Weibull Count	5	3.438	0.6328	5	6.5045	0.2602

Table 3: Chi-squared goodness-of-fit test for the Poisson and Weibull count models on the English Premier League

2.3.2 Results

A comparison of the Poisson and Weibull count distributions to the distribution of goals scored by the home and away team in the EPL is shown in Figure 4. Observing the clustered bar charts it appears that the Weibull count does indeed provide a superior fit to the distribution of both home and away goals in comparison to the Poisson for the EPL. A formal chi-square goodness-of-fit test was performed on both models with the results summarised in Table 3. At a 0.05 significance level we can see that the Poisson model is rejected and hence, is not a good fit for either the home or away goals in the EPL. Alternatively, there is not enough evidence to reject the null hypotheses that the home and away goals follow a Weibull count distribution respectively. These results are consistent with the observations made via the analysis of the histograms.

For the other four major leagues, there was enough evidence at 0.05 significance to reject the Poisson distribution for away goals with the Spanish La Liga also rejecting the Poisson distribution for the home goals. Overall, it appears that the Weibull

	England	France	Germany	Italy	Spain
γ	0.35995	0.40661	0.43220	0.28976	0.35173
κ	0.17572	-0.64826	0.32670	0.61312	0.25265
c_H	1.19634	1.13258	1.19390	1.11870	1.07045
c_A	0.96795	0.99603	0.98973	1.06636	1.04741

Table 4: Home Advantage, Dependence and Shape Parameters for all Leagues

count distribution is suitable for modelling the number of goals scored by a team in football. For the five leagues studied the Weibull count was only rejected once at a 0.05 significance level and in all cases, provided a superior fit to the Poisson distribution.

The bivariate Weibull count model is significantly more expensive computationally than the Dixon-Coles model. Thus, for this model the optimisation of ξ has been neglected and instead, the ξ values calculated for the Dixon-Coles model have been adopted. The parameter estimates for the EPL have been summarised in Figure 5.

As shown in Table 4, all leagues have positive κ values except the French Ligue 1. Positive values of κ correspond to positive values of Kendall's τ suggesting a positive dependence in these leagues. On the contrary, the French Ligue 1 appears to have a negative dependence between teams given it's negative value of κ. As expected, the bivariate Weibull count model has also determined positive home advantage parameters for all five leagues.

3 Predictive Capability

3.1 Rank Probability Score

The rank probability score was originally proposed by Epstein (1969) as a scoring rule for probabilistic predictions of ordered variables. This scoring rule is advantageous because it not only considers the prediction accuracy of the observed outcome, but it is also sensitive to distance, that is predictions that concentrate their probability about the observed outcome receive better scores. For a single observation, the rank

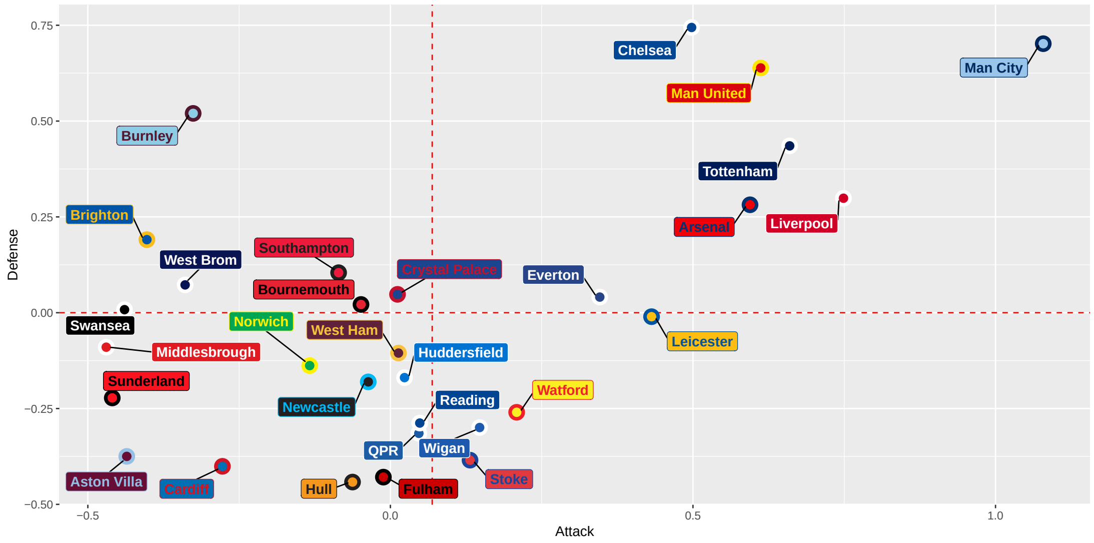

Figure 5: Scatter Plot of the Bivariate Weibull Strength Parameters for English Premier League Teams

probability score is defined as

$$\text{RPS} = \frac{1}{r-1}\sum_{i=1}^{r}\left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j\right)^2 \tag{3.12}$$

where r is the number of potential outcomes and p_j and e_j are the model probabilities and observed outcomes at position j respectively. Since the rank probability score represents the difference between the cumulative distributions of the predictions and observations, lower scores indicate more accurate predictions.

In Constantinou and Fenton (2012), it is argued that the rank probability score should be the preferred scoring rule when evaluating predicted probabilities of a football match. The basis of this recommendation stems from the observation that the outcomes of a football match follow an ordinal scale. As the authors mention, the outcome of a draw is closer to a home win in comparison to an away win. Under this assumption one would hope that if the result of a football match was a home win, a suitable scoring rule should apply a harsher penalty to the probability assigned to an away win compared to that of a draw. However, many studies on predictive football modelling have utilised scoring rules that assume the outcomes of a football match are nominal and in doing so have not adequately assessed the predictive capability of their model.

Model	**EPL**	**FL1**	**GB**	**ISA**	**SLL**	**Total**
Independent Poisson	37.2673	35.5844	32.9243	37.1353	39.2251	182.1364
Dixon-Coles	36.8437	36.2868	31.6165	37.1773	39.3028	181.2271
Bivariate Weibull Count	36.7900	36.2718	31.7603	37.1796	39.3442	181.3459

Table 5: Rank Probability Scores for the three models across all leagues

3.2 Model Comparison via Rank Probability Score

Utilising the rank probability score, the three models analysed in Section 2 will be compared with each other across the five different football leagues. This comparison will be accomplished by testing each model on the out-of-sample validation set from the latter half of the 2017-2018 seasons. The rank probability score for each model across all leagues have been summarised in Table 5.

Interestingly, the results vary considerably among the five leagues. For the English Premier League, the results are perhaps as one would expect. The independent Poisson model performs the worst yielding the highest rank probability score. Improvement is achieved employing the Dixon-Coles model whilst the bivariate Weibull count model contributes a further reduction in rank probability score. For the German Bundesliga, the Dixon-Coles model is deemed to produce the most accurate predictions followed by the bivariate Weibull count model. Perhaps surprisingly, the independent Poisson model yields the lowest rank probability score for the French, Italian and Spanish leagues. Even though both the Dixon-Coles and bivariate Weibull count model account for both dependence between teams as well the inclusion of a time-weighting function, they fail to outperform the simple independent Poisson model for three of the five leagues. Aggregating the rank probability scores across the five leagues we find that the Dixon-Coles model performs marginally better than the bivariate Weibull count model. Both of these models outperform the independent Poisson model.

3.2.1 Updating Time Weighting Parameter

In Section 2 the time time-weighting parameter ξ is determined through use of the predictive profile log-likelihood function. This function only considers the observed

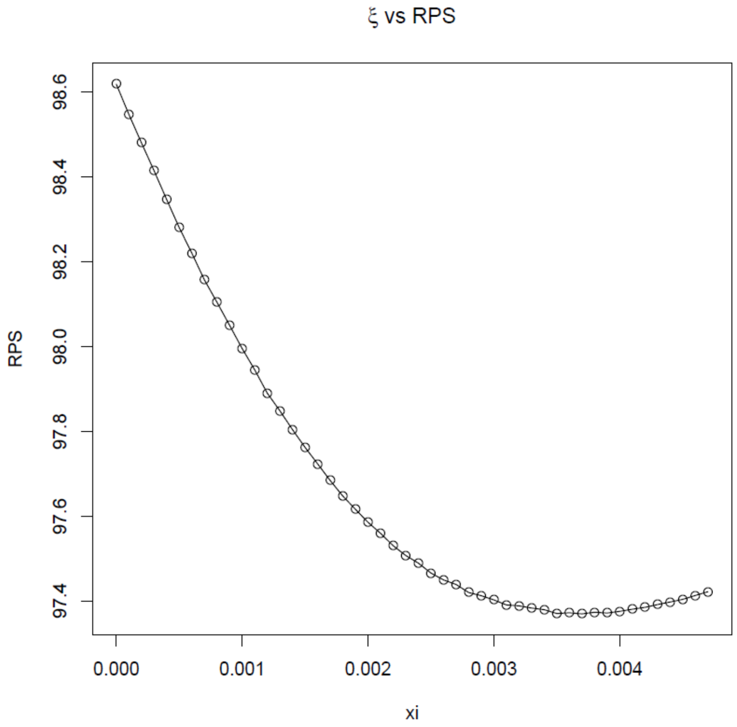

Figure 6: ξ vs RPS for English Premier League Teams

outcomes of each match and as Constantinou and Fenton (2012) argues, may not represent the most accurate assessment of the predictive capability of the model.

We recalculate the Dixon-Coles and bivariate Weibull count models utilising the rank probability score rather than the predictive profile log-likelihood to determine the optimal values of ξ. This procedure is analogous to the previous implementation where ξ is determined utilising cross-validation. This procedure is once again only performed on the Dixon-Coles model due to computational limitations. The optimal ξ value obtained is then used to recalculate the parameter estimates for both models.

A plot of the rank probability score as ξ varies for the EPL is shown in Figure 6. It appears that the values are more sensible compared to those obtained via the predictive profile log-likelihood. In particular, the optimal ξ value for the English Premier League is determined to be 0.0037, a significantly smaller value compared to the one previously utilised.

The estimates of ξ for the other leagues are shown in Table 6. Comparing these

	England	France	Germany	Italy	Spain
γ	0.23580	0.31451	0.26481	0.22047	0.27661
ξ	0.0037	0.0021	0.0006	0.0019	0.0035
ρ	-0.04028	-0.03308	-0.08215	-0.05933	0.06621

Table 6: Home Advantage and Low Scoring Dependence Parameters for all Leagues (Using RPS to choose ξ)

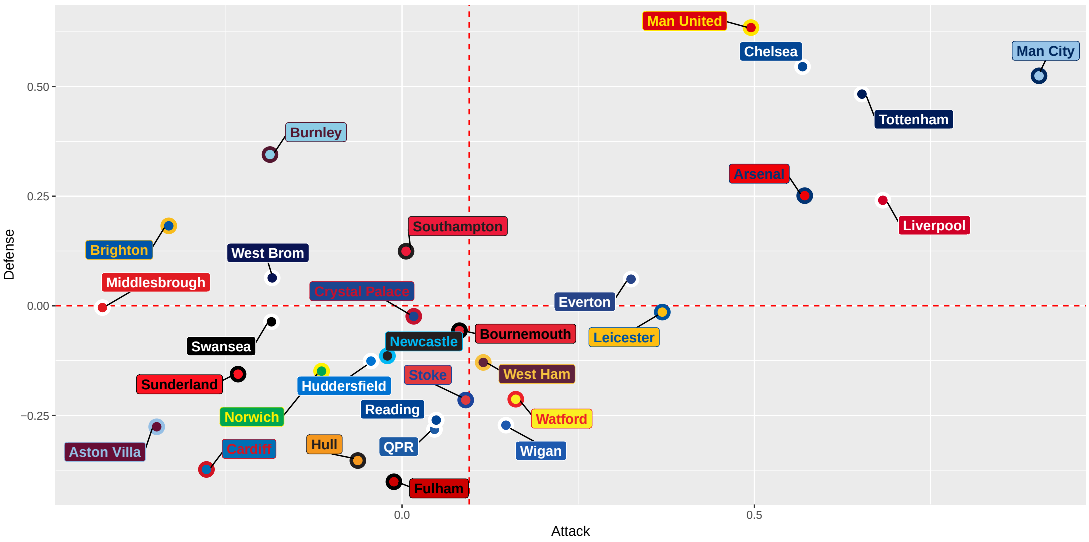

Figure 7: Scatter Plot of the Dixon-Coles Strength Parameters for English Premier League Teams (Using RPS to choose ξ)

with the values estimated using the predictive profile log-likelihood function presented in Table 2 it is apparent that the optimal ξ values for the Italian Serie A and Spanish La Liga are now non-zero. This indicates that these leagues do indeed benefit from exponentially down-weighting previous results. When comparing ξ values across different leagues, it appears that the English Premier League and Spanish La Liga have the highest values, indicating historical data is of less importance in these competitions.

The attacking and defensive strength parameter estimates for the Dixon-Coles model applied to the EPL data utilising the new ξ values are summarised below in Figure 7.

The sign of the ρ parameters in Table 6 remain the same in comparison to the ρ's in Table 2. A notable difference, however, is the fact that the ρ parameter for the Spanish La Liga is now greater, indicating that perhaps there is a significant departure from independence in low scoring matches.

	England	France	Germany	Italy	Spain
γ	0.35439	0.40117	0.33139	0.23104	0.33245
κ	0.10789	-0.47394	0.15272	0.62009	0.24867
c_H	1.12213	1.11150	1.08890	1.11650	1.11974
c_A	0.93591	0.97397	0.98703	1.11902	1.05089

Table 7: Home Advantage, Dependence and Shape Parameters for all Leagues (Using RPS to choose ξ)

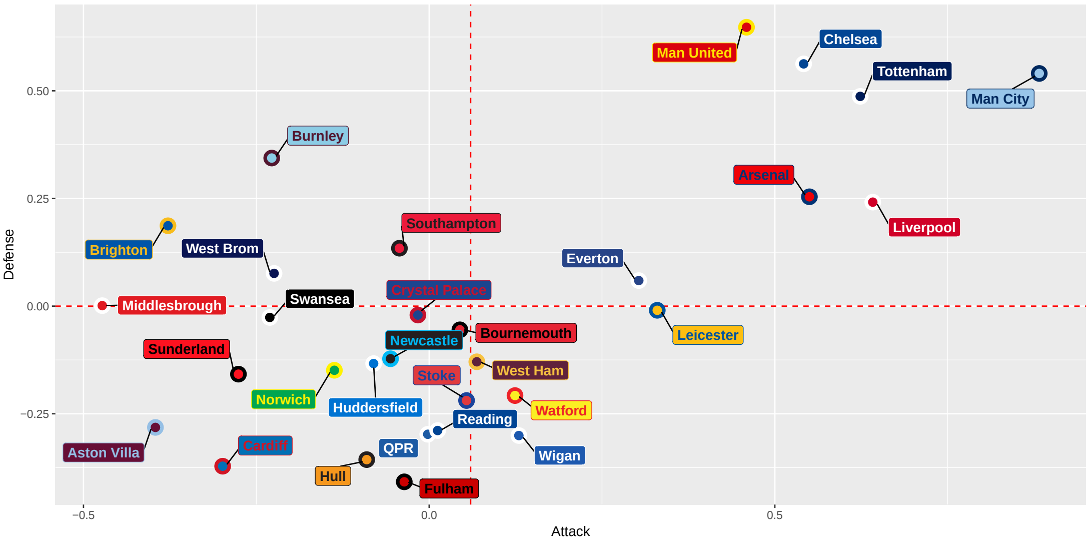

Figure 8: Scatter Plot of the Bivariate Weibull Strength Parameters for English Premier League Teams (Using RPS to choose ξ)

Model	**EPL**	**FL1**	**GB**	**ISA**	**SLL**	**Total**
Independent Poisson	37.2673	35.5844	32.9243	37.1353	39.2251	182.1364
Dixon-Coles	36.2717	35.8551	32.4760	36.6161	37.9672	179.1861
Bivariate Weibull Count	36.1982	35.8774	32.6408	36.6450	37.8859	179.2473

Table 8: Rank Probability Score for the three models across all leagues (using RPS to optimise ξ)

The attacking and defensive strength parameter estimates for the bivariate Weibull count model model applied to the EPL data utilising the new ξ values are summarised in Figure 8. Observing Table 7, the κ parameters are still of the same sign indicating no change in the dependence structure of each league.

Utilising these new parameter estimates, the rank probability scores have been recalculated on the out-of-sample validation sets.

Analysing the results in Table 8, it is evident that utilising the new ξ values improved the rank probability score for all leagues, with the exception of the German Bundesliga. ~~However, it appears that the Dixon-Coles model is still deemed to perform the best in both scenarios.~~ Accumulating the rank probability scores across the five leagues we again find that the Dixon-Coles model performs marginally better than the bivariate Weibull count model with both of these models outperforming the independent Poisson model. The bivariate Weibull count model still yields the lowest rank probability score for the English Premier League and with the newly derived parameter estimates, also performs the best for the Spanish La Liga. Despite the reduction in rank probability score for both the Dixon-Coles and bivariate Weibull count models, it appears that the independent Poisson model is still superior for the French Ligue 1. Lastly, when analysing the Italian Serie A it is clear that the reduction in rank probability score from the updated parameter estimates has resulted in the Dixon-Coles model now being deemed the best.

Given the inclusion of a distribution providing a better fit and copula based dependence structure, it is perhaps surprising that the bivariate Weibull count model does not outperform the other two models. For this analysis, these reasons could be twofold. Firstly, due to computational limitations, the time-weighting parameter was not calculated directly for the bivariate Weibull count model. As a result of adopting the ξ parameter from the Dixon-Coles model, this could potentially return a suboptimal rank probability score in the context of the bivariate Weibull count model. Alternatively, this could be simply due to the validation sets utilised being 'noisy' and better accustomed to the Poisson. As is most often the case in statistics, no one model dominates all others across various data sets. Despite it's additional improvements, it is to be expected that for particular data sets the bivariate Weibull count model will not always be the best performing model.

3.3 Further Quantification of Home Advantage

The existence of a home ground advantage in football has been well established in the literature. Extensive research has been performed (Pollard, 1986; Clarke and Norman, 1995) to determine it's causes and influence on the outcome. In all predictive football models proposed since the seminal paper Maher (1982), the home advantage has been accounted for given it's importance in determining the outcome.

In researching the home ground advantage in the four major American sports: baseball; ice hockey; basketball; and, American football, Schwartz and Barsky (1977) identified that there are tactical, physiological and psychological factors that affect it.

It is intuitive to assume that an away team playing a local derby against a team in the same city is at less of a disadvantage than an away team that is required to travel a significantly long distance. As mentioned in Pollard (1986), this can be due to numerous factors such as players experiencing travel fatigue as well as a lack of crowd support. With this in mind, we test the psychological factor influencing the home ground advantage by assessing if distance travelled is a significant predictor in modelling football scores.

Due to the lack of long distance trips and reasonably small proximity among clubs, it is difficult to determine whether travel distance is significant in any of the top European leagues that have been studied thus far. Hence, to analyse the impact of travel distance on the away team, data has been collected for two leagues which have some significantly long away trips and spatial disparity between teams. These two leagues are the Hyundai

Australia		**Russia**	
Coefficient	Estimate	Coefficient	Estimate
Region: Close	-0.131095	Region: Close	-0.142253
Region: Far	-0.244638	Region: Far	-0.281670

Table 9: Parameter Estimates for ‘Close‘ and ‘Far’ in the Independent Poisson model for both the A-League and Russian Premier League

A-League and the Russian Premier League.

To analyse whether travel distance has a statistically significant effect on the outcome of a match, the coordinates of each team’s home stadium were collected for both leagues. These were then utilised to group the teams into regions via hierarchical clustering analysis. The home advantage explanatory variable was then reclassified as the variable ‘Region’ with three levels; ‘Home’, ‘Close’ and ‘Far’ where the last two levels represent playing a team within and outside of your region respectively. The independent Poisson model was fitted to both leagues with and without this modified variable to see whether it provided any improvement in the predictive capability of the model. It was decided to fit the independent Poisson model due to it’s efficiency computationally and easily interpretable output from the `glm()` function.

For the hierarchical clustering, complete linkage was utilised as the distance measure for the clustering algorithm. The resulting dendrograms are plotted in Figures 9 and 10. Observing these figures, the decision was made to cluster the teams of both leagues into six regions. The red dashed line indicates the cut which produces each league’s respective clusters.

Parameter estimates from the fitted models for both ‘Close’ and ‘Far’ are summarised in Table 9. Observing this output, the estimates are as expected for both leagues. Since ‘Home’ is the reference level, it’s estimate is 0. Interpreting the remaining estimates, it is clear that the away side is at more of a disadvantage when playing a team outside of their region in comparison to a team within their region. Additionally, when producing an analysis of deviance table, the ‘Region’ variable is deemed statistically significant for any relevant significance level.

Finally, when implementing both versions of the independent Poisson model on each

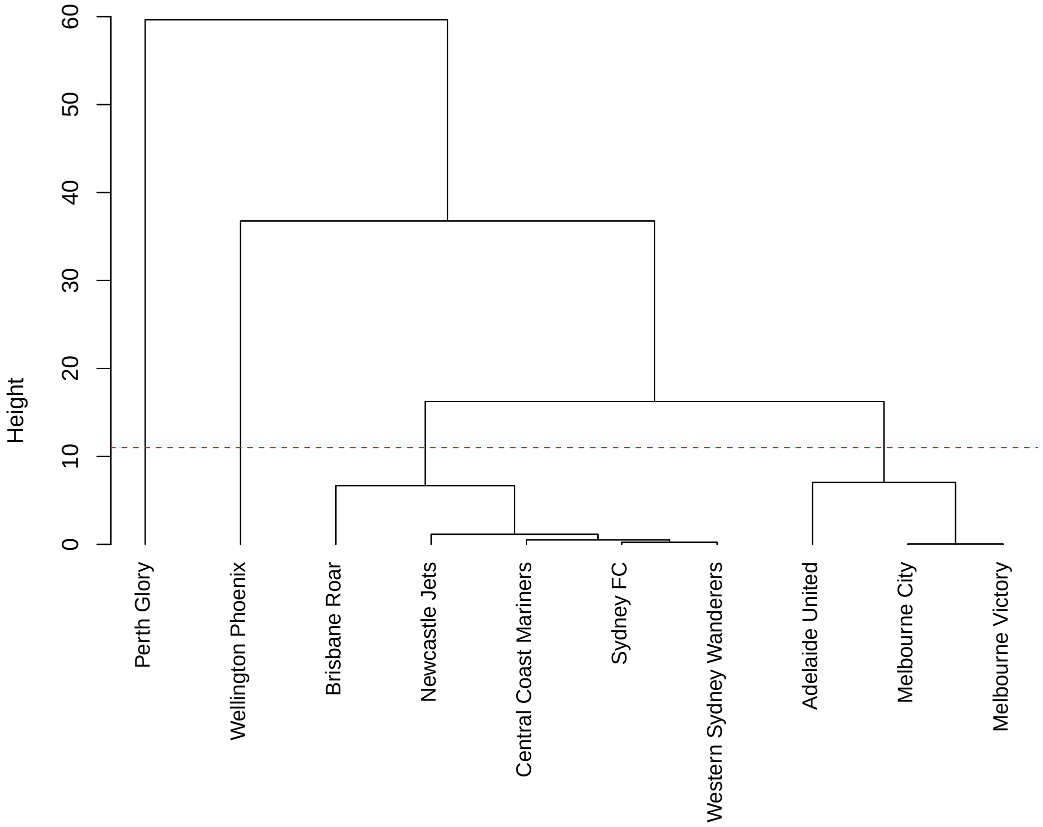

Figure 9: Dendrogram for clustering Australian teams via complete linkage

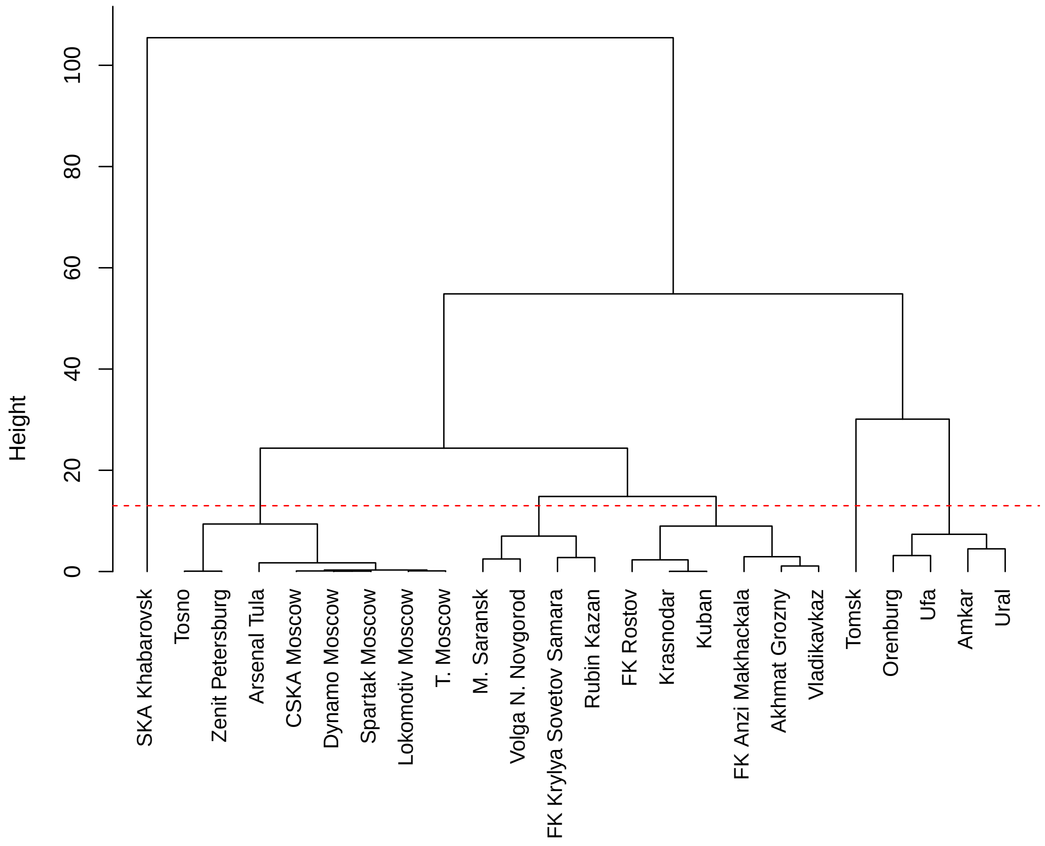

Figure 10: Dendrogram for clustering Russian teams via complete linkage

league's out-of-sample validation set, the inclusion of the 'Region' variable yields a lower rank probability score for both competitions. This provides evidence to suggest that when modelling matches from these leagues, it is important to account for the distance travelled by the away team. These results also indicate that the travel distance should at least be considered before implementing a model on any league, especially for leagues with significant spatial disparity such as; The Major Soccer League in America, the Campeonato Brasileiro Série A in Brazil or the Chinese Super League.

Australia		**Russia**	
Model	RPS	Model	RPS
Home Variable	14.4039	Home Variable	23.3932
Region Variable	14.2888	Region Variable	23.3917

Table 10: Rank Probability Score for the Independent Poisson Model with and without the 'Region' variable

4 Conclusion

In this paper we have analysed and compared three predictive football models from the literature on the top European leagues focusing our presentation on the EPL. A benefit of such analysis is not only to independently determine teams' strengths within leagues but also in the prediction of international club games such as those in The UEFA Champions League.

A comparison of each model's predictive capability has been made using the rank probability score. It was determined that there was not one superior model among the five leagues. Despite the Weibull count providing an improved fit to the distribution of goals in comparison to the Poisson, the bivariate Weibull count model was only deemed the best for two out of the five leagues. It is noted that this could potentially be due to the fact that the time-weighting parameter from the corresponding leagues' Dixon-Coles model was used rather than deriving it for the model itself. However, each league is different and perhaps particular models are simply better suited to specific leagues. Across all five leagues, the Dixon-Coles model performs best, followed by the bivariate Weibull count model, then the independent Poisson model according to the rank probability scores. However, the difference between the Dixon-Coles model and the bivariate Weibull count model was only very small.

An adjustment to the home advantage to account for travel distance of the home team was also investigated. Both the Australian and Russian football leagues are shown to experience improvement in rank probability score with the inclusion of the proposed adjustment. This provides evidence to suggest that when modelling football matches in leagues where there is significant disparity in travel distance it should be included in the model.

With evidence to support the significance of travel distance in the Australian and Russian leagues, including this adjustment in more complicated models such as the Dixon-Coles and bivariate Weibull models should be investigated to determine whether any further improvements can be achieved. Additionally, analysis on tactical and psychological factors of the home ground advantage could be accomplished. If quantified, these factors could too be added to existing models in order to further improve predictive capability.

References

Georgi Boshnakov, Tarak Kharrat, and Ian G McHale. A bivariate weibull count model for forecasting association football scores. International Journal of Forecasting, 33 (2):458–466, 2017.

Stephen R Clarke and John M Norman. Home ground advantage of individual clubs in english soccer. The Statistician, pages 509–521, 1995.

Anthony Costa Constantinou and Norman Elliott Fenton. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. Journal of Quantitative Analysis in Sports, 8(1), 2012.

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. ISSN 00359254, 14679876.

Edward S Epstein. A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology, 8(6):985–987, 1969.

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.

Michael J Maher. Modelling association football scores. Statistica Neerlandica, 36(3): 109–118, 1982.

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.

Richard Pollard. Home advantage in soccer: A retrospective analysis. Journal of sports sciences, 4(3):237–248, 1986.

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.

Barry Schwartz and Stephen F Barsky. The home advantage. Social forces, 55(3): 641–661, 1977.

Abe Sklar. Random variables, joint distribution functions, and copulas. Kybernetika, 9(6):449–460, 1973.