An Introduction to Football Analytics and Modelling

Dimitrios Kotsis

2066447

1 Introduction

Association Football is one of, if not the, most popular and prestigious sports in the world. It is estimated that there are over 3.5 billion football fans around the world Allianz (2022). Major football events like the World Cup and the Champions League concentrate the interest of the globe. The sport betting market has grown substantially the last few years. In 2013, the industry was estimated to be worth up to £625 billion a year, with 70% of that being traded on football (Keogh and Rose, 2013). To this direction, modelling the characteristics of the game using appropriate statistical models and/or machine learning methods have found increasing appeal in a wide range of stakeholders. Such models relate to almost every aspect of the football itself, including modelling the score of the match, the optimal time for substitutions, predictions about injuries, pricing of football players, to name few.

Modelling in football dates back to Reep and Benjamin (1968), where they investigated the likelihood of a successful n_{th} pass, namely to have n successive passes from the same team. Since then, we have seen a lot of progress and many new models, while the data availability has also increased tremendously from simple observation as in Reep and Benjamin (1968) to recent tracking technologies that can provide info about the players and the ball 20 times per second.

Statistical models for football are important for a series of reasons including:

- prediction of the outcome of the game,
- explaining what are the drivers for success, perhaps after the game,
- using the model as background/adjustment in order to investigate particular aspects of the game, as for example, the effect of the playing ground, the effect (if any) of the referee etc.

The huge literature on modeling the football outcome can be divided in two big categories (Scarf and Rangel Jr, 2017):

- Direct Models: models that try to model/predict the outcome (win/draw/loss) of a football game using appropriate techniques
- Indirect Models: models that attempt to predict the score of the game, namely how many goals each team will score. Of course modeling the score, you can calculate many more things including the win/draw/loss.

One can notice that the data needed for each category of models are different. In this thesis we will use models that try to predict the score of the game and in particular we will try to model the number of goals scored by each team in a match.

Among such models, modeling the score or the winner of a game have found tremendous interest in academia, in business and in society. In this thesis we concentrate on this problem, how one can predict the outcome of a match before the game starts. Note that there is also a number of other types of models one can use:

- Model the difference in score and not the score itself. A great example is the Skellam Model (Karlis and Ntzoufras, 2009).
- Survival models: model the time to goal scoring (Dixon and Robinson, 1998).
- In-play modelling: model the final outcome conditional on some information during the game. For example, what is the probability the away team wins if the score is 1-1 at 60 minutes.

Due to space restrictions we do not pursue such models in this thesis.

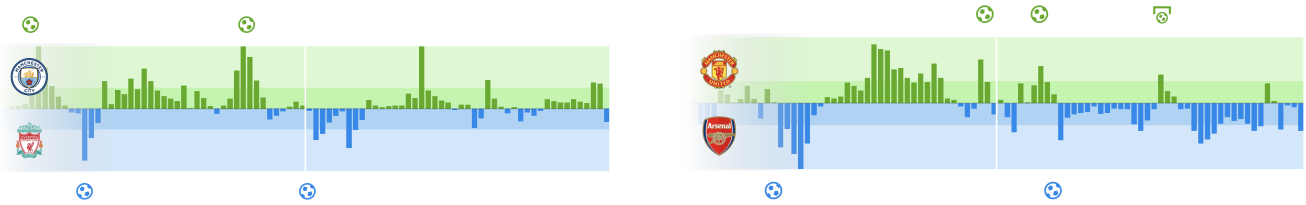
Football is a low-scoring sport, so it is very difficult to predict the outcome of a game (Scarf et al., 2021). There can be a lot of surprises and changes during a match.This explains why it is extremely important to understand the impact of some elements of the game and how one can use them to improve the accuracy of the models. Before presenting the models this paper is going to focus on, let us have a look at some of these aspects of the game in section 2. The models used in the dissertation will be described in Section 3 while the application to the English Premier League data from 2017-2018 is provided in Section 4. Final conclusion re given in Section 5.

2 Football Data

Each team and coach has a different philosophy on how to approach the game. But if we wanted to break the game into simpler terms, the objective is to outscore the opposing team. To do so, you need a good offense (scoring goals) and a good defense (not conceding goal). Ball possession, passing, and team and player movement on the pitch have huge impacts to the overall performance of the team. In order to create the most optimal model, it is important to collect sufficient data for all of the above and understand the significance of each one.

As an example, attack momentum is very helpful to understand the offensive performance of a team. It is a visual representation of an algorithm that measures the swing of a match and which team is creating more threatening situations at certain points in time (Whitmore, 2021). Ball possession, chances created and shots are the key elements of the game taken into consideration to calculate the momentum. We can find these graphs from a number of websites online. Here, we will be looking at the Attacking Momentum of two different games of the English Premier League that were played in the 2021/2022 season.

The graph is simple: the green part is the home team and the blue the away team. Each minute produces a spike; the bigger the pressure, the higher the spike. Notice how, in Figure 1a, both goals from Manchester City and the first goal from Liverpool were scored at some peaks. Same goes for Figure 1b and the first goal by Arsenal. That shows the impact these aspects have on the game and the scoring rates of the team. However, there's another observation we can make. Some of the goals, as seen in Figure 2, were not scored in the moments of highest, if any, pressure. That



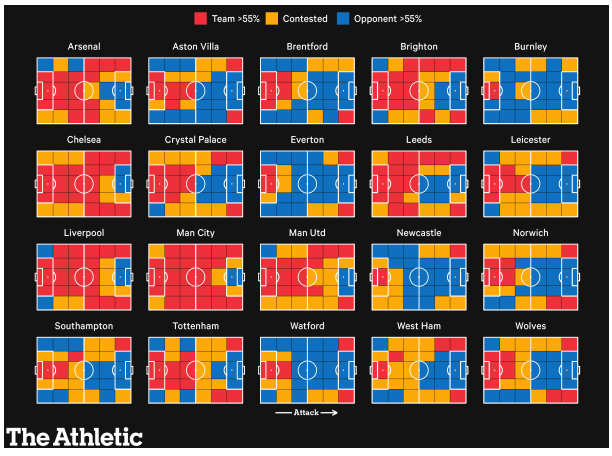
(a) Manchester City 2 - 2 Liverpool

(b) Manchester United 3 - 2 Arsenal

Figure 1: Attack Momentum for two matches from Premier League 2021-2022SofaScore (2022)

bring us back to our point in Section 1. Football can be full of surprises. Being the most dominant team, having the most shots, the highest ball possession, does not guarantee goals and wins.

Having a high ball possession and generally more control of the game does not only impact offence but also defense. First, let us divide the pitch into sections, territories. By taking the average of the ball possession of each team, for each territory, we can create a visual representation of possession, as seen in Figure 2a. The table in Figure 2b shows us the number of goals conceded by each team, with the top of the list being the team with the least number of goals against. By comparing the two figures we can easily see that the team with the least control over the most territories, especially in the defensive end, has the most goals conceded. By allowing the opposing team "own" a territory near your goal area, lets them create more chances and increase the likelihood of a goal. Let us look at Liverpool, Manchester City and Chelsea. They dominate possession throughout and have conceded the least number of goals. On the other hand, Everton, Newcastle United, Watford struggle in that aspect of the game and hence, they are suffering the consequences. However, there are of course exception to this. For example, Leeds United seems to have a good ball possession in the defensive end of the pitch, yet has the most goals scored against them.



(a) Possession by zone in the 2021-22 Premier League Muller (2022)

	Team	Goals against	Avg per game	Clean Sheets
1	Manchester City	20	0.63	19
2	Liverpool	22	0.69	19
3	Chelsea	27	0.87	14
4	Wolverhampton Wanderers	28	0.88	11
5	Tottenham Hotspur	38	1.19	12
6	Arsenal	39	1.22	13
7	Brighton & Hove Albion	40	1.21	9
8	Crystal Palace	41	1.28	9
9	West Ham United	43	1.30	7
10	Burnley	45	1.41	8
11	Aston Villa	46	1.48	9
12	Manchester United	48	1.45	7
13	Brentford	49	1.48	7
14	Leicester City	51	1.65	5
15	Everton	53	1.71	6
16	Southampton	54	1.64	8
17	Newcastle United	55	1.67	6
18	Watford	62	1.94	3
19	Norwich City	66	2.06	6
20	Leeds United	68	2.13	4

(b) Number of goals conceded SportsMole (2022)

Figure 2: Impact of ball possession to defence

Other than goals scored and goals conceded, there are many more variables to the game that can have a major impact, such as home advantage, chance of injury, weather, current form, tiredness

etc. Some of these can and should be included into the models to improve the accuracy of our predictions. We will discuss further the home advantage in section 3.3 and current form in section 3.5.

3 The Models

3.1 Basic Characteristics

The literature on statistical models for football is quite large and contains a lot of models based on different assumptions and complexity. One of the basic questions that need to be answered is whether pure chance dominates the game, i.e. the goals are events that occur due to randomness only. Such an assumption leads to the Poisson distribution in order to describe the number of goals for each team. A Poisson distribution has the property that its mean is equal to its variance, a property also known as *equidispersion*. Real data on football typically show a small overdispersion, i.e. the observed variance is larger than the observed mean. This can be explained due to the fact that each team has a different scoring ability and hence what we observe is not a single Poisson distribution but several ones.

A second important question is whether the number of goals scored by each team are uncorrelated. Intuitively, since the two teams compete together, scoring a goal may initiate an increased attempt of the other team to score against and this can be considered as introducing correlation. Observed data have shown a small amount of correlation and hence one would like to be able to account for this.

A final important question is which information could be useful in order to better predict the game result. Of related nature is whether this information is known prior to the game and whether we can measure its impact to the final result. This is typically a very refined procedure, not always possible to take into account due to missing of the information. For example, the player availability is not always known, since some injury may lead to the absence of some player. Or the weather conditions can be only guessed until the very last minute of the game.

3.2 The Poisson Distribution

To start with the description of the models used in this dissertation, we begin with the basic assumption about the underlying stochastic mechanism. As already said, a natural debate is whether the Poisson distribution can do the job. Poisson assumption implies that only chance governs the sport; this is clearly very unrealistic. In practice, models based on Poisson distribution and some covariate information to account for the in-homogeneity, can capture well the dynamics.

Consider the random variable X that counts the number of goals that one team will score in the match. Then, assuming that X follows a Poisson distribution, the probability that the team will score k goals is given by

$$P(X = k) = \frac{\exp(-\lambda)\lambda^k}{k!}, \quad \lambda > 0, \quad x = 0, 1, 2, \ .$$

The Poisson model implies that the mean coincides with the variance, namely $E(X) = Var(X) = \lambda$, where λ can be interpreted as the expected number of goals for this team.

Assume that the number of goals scored by each team follows a Poisson distribution. We take λ to be the scoring rate of a team, which can be reasonably estimated by:

$$\lambda = \frac{\text{Number of goals scored}}{\text{Number of games played}}.$$

Based on this, we can now calculate the probability of a this team to score x number of goals. For example, in the 2021/2022 season, Tottenham has scored 56 goal in their first 32 games. Hence, $\lambda = 1.75$ and:

$$P(\text{Tottenham scores 2 goals}) = \frac{1.75^2 \cdot e^{-1.75}}{2!} = 0.266091 \text{ or } 26.61\%.$$

This model is a good starting point, however, we need to take into consideration the offensive ability of the team which also relates to the defensive performance of the other team. We need λ to represent the strength of a team given also the quality of the opposition. So, a natural question is how to model λ? Typically this is a Poisson regression model that assume that $\log \lambda$ relates to some external information by a typical log-link function namely

$$\log \lambda_i = \beta_0 + \sum_{j=1}^{p} \beta_j z_{ij}$$

where z_{ij}, $j = 1, \ldots, p$ are some covariate information for the i-th team and the β's are regression coefficients to be estimated.

We can now define the model we are going to use in this dissertation.

3.3 Double Poisson Regression

Consider a football match against two teams. We want to model the number of goals say X and Y that the two teams will score. Typically X relates to the home team and Y to the away team. Note that by such a model $P(X > Y)$ gives the probability that the first team will win, ($P(X = Y)$ is the probability of a draw and $P(X < Y)$ the probability that the second team will win. The first model, assumes that each of the two variables follows a Poisson distribution and that they are not correlated. This assumption will be relaxed later. It is known as Double Poisson (see, e.g. Lee, 1997).

We assume for the i-th match, $i = 1, \ldots, n$, that the pair (X_i, Y_i), i.e. the number of goals for each of the two teams, are modeled as two independent Poisson distributions. Then the model takes the form as:

$$\begin{aligned} X_i|\lambda_{1i} &\sim \text{Poisson}(\lambda_{1i}), \\ Y_i|\lambda_{2i} &\sim \text{Poisson}(\lambda_{2i}), \\ \log(\lambda_{1i}) &= \mu + \text{home} + \text{att}_{h_i} + \text{def}_{a_i}, \\ \log(\lambda_{2i}) &= \mu + \text{att}_{a_i} + \text{def}_{h_i}, \end{aligned} \tag{1}$$

where $\lambda_{1i}, \lambda_{2i}$ represent the *scoring rates* for the two teams, i.e. the expected number of goals for the home and away team, respectively; the parameters att_k and def_k encapsulate the offensive (or attacking) and defensive performances of team k, respectively, for each team k, $k = 1, \ldots, n_t$; the nested indexes $h_n, a_n = 1, \ldots, t$ denote the home and the away team playing in the i-th game,

respectively; μ represents the constant intercept and home represents the *home-advantage*. The last parameter is of particular importance since it is well known that football has a significant home advantage Pollard (1986), i.e. the teams playing at their own field show a better performance than when they play away from their field. This has been widely documented and can be attributed in a series of factors including the effect of the spectators of the team, better knowledge of the field, less fatigue due to no traveling, and some more psychological factors.

To achieve identifiability in the above model (i..e to be able to estimate all the parameters) we assume

$$\sum_{k=1}^{n_t} \text{att}_k = 0, \ \sum_{k=1}^{n_t} \text{def}_k = 0, \tag{2}$$

or equivalently a 'corner' constraint by imposing a *baseline* team whose teams abilities are set to zero, and the other ones are incremental with respect to the baseline:

$$\text{att}_1 = 0, \text{def}_1 = 0 \tag{3}$$

$$\sum_{k=2}^{n_t} \text{att}_k = 0, \ \sum_{k=2}^{n_t} \text{def}_k = 0, \tag{4}$$

Other constraints are possible (offering different interpretation).

The above is a basic model that relates the outcome of the game to the strength of the teams and some home advantage. Of course one may add more covariates in the λ's to account for any other factors. Such factor and in order to be used for predictive purposes need to be known in advance. Such candidate covariates can be

- the shape of the team measured by the results in the last games
- the motivation of the teams, whether they still run after some target (e.g. participation in some European championship or to avoid relegation)
- team composition, if an important player is absent due to injury for example
- fatigue due to previous matches
- many others including expert opinions about the match itself as they can be incorporated for example with betting odds given by some bookmaker.

Some comments apply to this basic model. As already mentioned the Poisson assumption is perhaps too limited. We may built similar models for other distributional assumptions, like the negative binomial distribution that accounts for overdiseprsion.The model does not assume correlation between the two variables. Estimation of the model can be done using standard statistical approaches like maximum likelihood method. If one wants to down-weight matches further away in time, he may use some time function to down-weight matches played before a long time and give more weight to the more recent matches. Finally, note that the above model assumes the same home advantage for all teams in the league. One may think that different teams have different home effects and transform the model to account for this by assigning more home effect parameters.

3.4 Bivariate Poisson Distribution

The shortcoming of the previous model is that it assumes that there is no correlation between the number of goals scored by the two teams. this is perhaps restrictive. The next step is to create a model that allows for correlation. Such a model is known as the bivariate Poisson distribution which creates a bivariate distribution that has Poisson marginal distributions that are correlated.

Consider random variables $X_r, r = 1, 2, 3$, which follow independent Poisson distributions with parameters $\lambda_r > 0$. Then, the random variables $X = X_1 + X_3$ and $Y = X_2 + X_3$ follow jointly a bivariate Poisson distribution denoted as $\text{BP}(\lambda_1, \lambda_2, \lambda_3)$, with joint probability function

$$\begin{aligned} P_{X,Y}(x, y) &= P(X = x, Y = y) \\ &= \exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\}\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!}\sum_{k=0}^{\min(x,y)}\binom{x}{k}\binom{y}{k}k!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k. \end{aligned} \tag{5}$$

Marginally each random variable follows a Poisson distribution with $\text{E}(X) = \lambda_1 + \lambda_3$, $\text{E}(Y) = \lambda_2 + \lambda_3$, and $\text{cov}(X, Y) = \lambda_3$; λ_3 measures the dependence between the goals scored by the two competing teams. If $\lambda_3 = 0$ then the two variables are independent The Bivariate Poisson model aims at accounting for the correlation between the numbers of goals (Karlis and Ntzoufras, 2003). Now we assume that

$$\begin{aligned} (X_i, Y_i)|\lambda_{1i}, \lambda_{2i}, \lambda_{3i} &\sim \text{BP}(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}) \\ \log(\lambda_{1i}) &= \mu + \text{home} + \text{att}_{h_i} + \text{def}_{a_i}, \\ \log(\lambda_{2i}) &= \mu + \text{att}_{a_i} + \text{def}_{h_i}, \end{aligned} \tag{6}$$

where for the covariance parameters λ_{3i} we may assume the general form:

$$\log(\lambda_{3i}) = \beta_0 + \beta_{h_i}^{\text{home}} + \beta_{a_i}^{\text{away}} + \boldsymbol{\beta} w_i,$$

where β_0 is a constant parameter, $\beta_{h_i}^{\text{home}}$ and $\beta_{a_i}^{\text{away}}$ are parameters that depend on the home and away team respectively, w_i is a vector of covariates for the i-th match used to model the covariance term and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients.

The interpretation remains pretty unchanged with respect to the double Poisson model, Now an explicit *positive* correlation between the teams' scores is introduced. The same constraints for identifiability proposed for the Double Poisson apply here. In the application section that follows we do not assume any covariates at λ_3 which is assumed to be constant across all matches.

We note that this distribution is a natural generalization of the Poisson to higher dimensions. It has a lot of applications in many other disciplines.

3.5 Dixon and Coles Model

Dixon and Coles (1997) showed that the Double Poisson model fails at particular scores and especially for low scores like 0-0, 1-0, 0-1 and 1-1, which appear more or less frequently that that predicted by a Double Poisson model. Figure 3 is a heat map from Sheehan (2018) that shows the average difference between actual match outcomes and the relative BP model predictions,for the 2005/06 season all the way up to the 2017/18 season. Green indicates underestimation from the model and red overestimation. One can see that 0-0 and 1-1 are the scores that are represented poorly by the model.

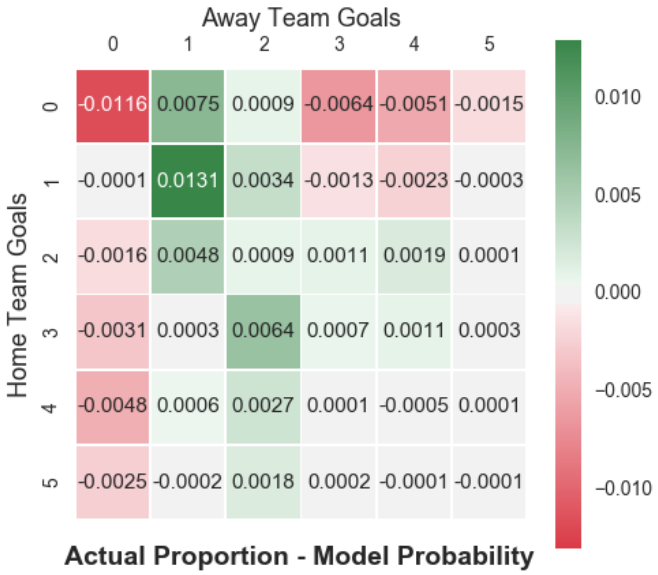


Figure 3: Heat Map for Scores: Actual vs Predictions (Sheehan, 2018)

Dixon and Coles (1997) proposed the introduction of a correction term to account for the under and overestimation of low scoring games of the BP model. They mathematically defined the model as:

$$P_{\mathrm{X,Y}}(x,y) = P(X=x, Y=y) = \tau_{\lambda_1,\lambda_2}(x,y) \times \frac{\lambda_1^x \cdot \exp(-\lambda_1)}{x!} \times \frac{\lambda_2^y \cdot \exp(-\lambda_2)}{y!},$$

where τ measures the correlation between the scores:

$$\tau_{\lambda_1,\lambda_2}(x,y) = \begin{cases} 1-\lambda_1\lambda_2\rho, & \text{if } x=y=0 \\ 1+\lambda_1\rho, & \text{if } x=0, y=1 \\ 1+\lambda_2\rho, & \text{if } x=1, y=0 \\ 1-\rho, & \text{if } x=y=1 \\ 1, & \text{otherwise} \end{cases},$$

and

$$\begin{aligned} \lambda_1 &= \alpha_{\mathrm{i}}\beta_{\mathrm{j}}\gamma, \\ \lambda_2 &= \alpha_{\mathrm{j}}\beta_{\mathrm{i}}. \end{aligned}$$

Here, ρ satisfies:

$$\max\left(\frac{-1}{\lambda_1}, \frac{-1}{\lambda_2}\right) \leq \rho \leq \min\left(\frac{1}{\lambda_1\lambda_2}, 1\right).$$

Similarly to the BP, ρ measures the dependence of the scores, and thus, $\rho = 0$ corresponds to independence, but otherwise the independence distribution is perturbed for events with x ≤ 1 and y ≤ 1. Also note that, marginal distributions are still Poisson (Dixon and Coles, 1997). Parameters α_j can be interpreted as the attacking parameters, β_j as the defensive parameters and γ as the home effect.

In this model, and assuming n teams, there are $2n+2$ parametersto be estimated. To prevent the model being over-parameterized, the following constraints are imposed:

$$n^{-1}\sum_{i=1}^{n}\alpha_{\mathrm{i}} = 1 \text{ and } n^{-1}\sum_{i=1}^{n}\beta_{\mathrm{i}} = 1.$$

which are similar to those used for the Double Poisson model.

3.6 Estimating the models

In the theory of statistics, there are multiple methods to estimate parameters of a model. Here, we will use the maximum likelihood estimator (MLE). First, we need to introduce the likelihood function. For discrete cases, the likelihood function of a parameter vector θ given an independent and identically distributed sample $X_1, X_2, ..., X_n$ is defined by:

$$\ell(\theta; X_1, X_2, ..., X_n) = \prod_{i=1}^{n} p(X_i; \theta).$$

The maximum likelihood estimator of θ is the value $\hat{\theta}$ that maximises the likelihood function $\ell(\theta)$. It is often easier to maximise its logarithm $L(\theta) = \log \ell(\theta)$ If $L(\theta)$ is twice differentiable, then the MLE $\hat{\theta}$ satisfies:

$$\left.\frac{\partial L}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \text{ and } \left.\frac{\partial^2 L}{\partial \theta^2}\right|_{\theta=\hat{\theta}} < 0$$

Standard errors can be obtained by inverting the Hessian matrix containing the second derivatives of the log-likelihood.

3.7 Simulating the League

In order to assess the goodness of fit of the model, a typical approach can be based on simulating the League based on the fitted model (Leitner et al., 2010). The simulated Leagues can be the basis to calculate the final standings and the ranking of the teams which can act as a goodness of fit if compared to the actual ones.

For the simulation we used the following approach:

Step 1: Based on the estimated parameters from the selected model (the Dixon and Coles model in our case), we calculate for each match the expected number of goals per team and the correlation parameter τ (if any) (this is the same for the entire league)

Step 2: Then we simulate a score for each match from this bivariate distribution based on table look-up method. The expected number of goals are the parameter of the two marginal Poisson models.

Step 3: This way we simulated the entire League repeating steps 1 and 2, for all 380 matches using the relevant information for each match.

Step 4: For this replicated league we calculated the points for each team. We applied as a tie breaking rule the random choice to avoid complicating the code.

Step 5: Then we replicated the League $B = 10000$ times, by repeating steps 1-4. For each League we got the number of points per team and the ranking of the team

At the end of this simulation we have replicated the League B times and hence we can find quantities as the expected number of points per team, the average ranking of the team etc.

4 Application

4.1 About the Data

The Premier League, also known as the English Premier League (EPL) is the top level of the English football league system. Contested by 20 clubs, it operates on a system of promotion and relegation with the English Football League (EFL). Seasons run from August to May with each team playing 38 matches (playing all 19 other teams both home and away). Simple arithmetic shows that there are $380 = 20 \times 19$ games in the season. A win gets a team three points and a draw one point. The first four teams are qualified for the prestigious Champions League competition, a European tournament of high value for the football companies. The last three teams relegate to the Football League.

We are using data from the season 2017-2018. The data have been obtained from `www.football-data.co.uk`. In fact the dataset contains much more information including the bookmaker odds from several companies as well as box-score statistics from each match. Since the box-score statistics relate to the match itself they lack any predictive ability and they will not be used here. Our main purpose is to examine whether the simple but quite common models described in section 3 are sufficient to capture the dynamics of the season. This year Manchester City won the championship with 100 points, far away from the second team which was Manchester United with 81 points. It is of interest to examine the uncertainty around the ranking of the teams.

To work with we use two different approaches. The first one, acting as a goodness of fit, fits the different models and then examines whether they can account for the observed final standings. The second approach, makes use of the first 25 days of the championship and tries to predict the rest 13 days aiming at revealing the predictive potential of the model.

All calculations were done using `R` software. We have used the package `bivpois` (Karlis and Ntzoufras, 2005) for the bivariate Poisson and the Double Poisson models and code from `http://opisthokonta.net/?cat=48`. Simulations of the League were run using our own code in `R`.

4.2 Some descriptive analysis

To start with it is useful to mention that home teams scored on average 1.53 goals while the away teams scored 1.14. Figure 4 is useful to see the overdispersion for all teams. The plot depicts the mean number of goals per team in the x-axis and the variance in the y-axis. Different colors indicate home and away matches for each team. Under a Poisson assumption, we expect to see numbers close to the diagonal line. Indeed we do not see serious departures from this line; any departures can be attributed to randomness. This is good indication that a model based on the Poisson distribution is sufficient to describe the case. Interesting enough is the point that the Pearson correlation between the number of goals for the home and the away team is -0.13.,i.e. slightly negative one.

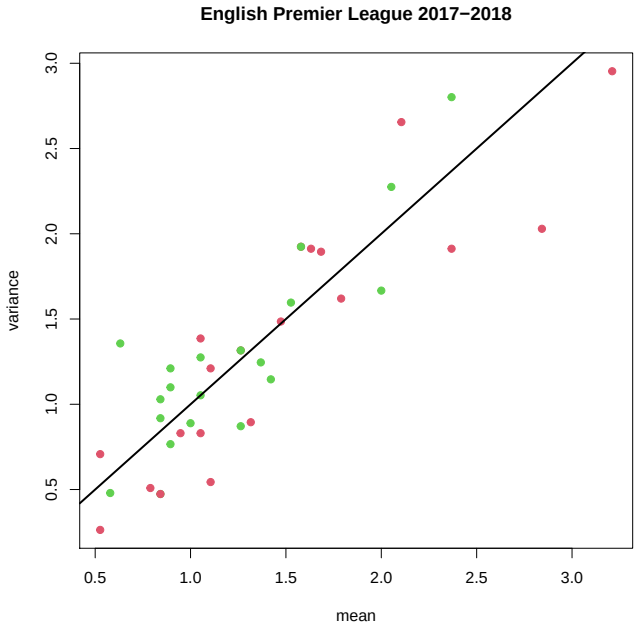


Figure 4: Mean versus variance for the number of goals scored by each team in the EPL 2017-2018. Red points are the home goals and green the away goals. The line indicates the diagonal, i.e. mean equals the variance which is a strong indication for a Poisson model. The departures from this lines can be attributed to randomness.

Model	Log-likelihood	AIC
Double Poisson	-1052.338	2184.776
Bivariate Poisson	-1052.338	2186.776
Dixon and Coles	-1050.801	2183.602

Table 1: Fitted models

4.3 Fitting the models

We have fitted the three models described in Section 3 using maximum likelihood method. All 380 available matches have been used. The fitted model assumed a common home effect, the same for all the teams, and offensive and defensive parameters for each team as indicated in the previous section. Table 1 provides the results from the fitted models. One can see that the Bivariate Poisson model has the same log-likelihood with the Double Poisson. The reason is that the data showed negative correlation and this model allows only for positive correlation, so the correlation parameter was estimated as 0, which makes the model the same with the Double Poisson. On the other hand the Dixon and Coles model improved the likelihood slightly since it can have negative correlation.

Using the Akaike Information Criterion (AIC), the Dixon and Coles is slightly preferable. We use the formula

$$AIC_m = -2L_m + 2d_m$$

where for model m, L_m is the maximized log-likelihood and d_m the number of parameters. Note that using Bayesian Information Criterion (BIC) that uses larger penalty for each additional pa-

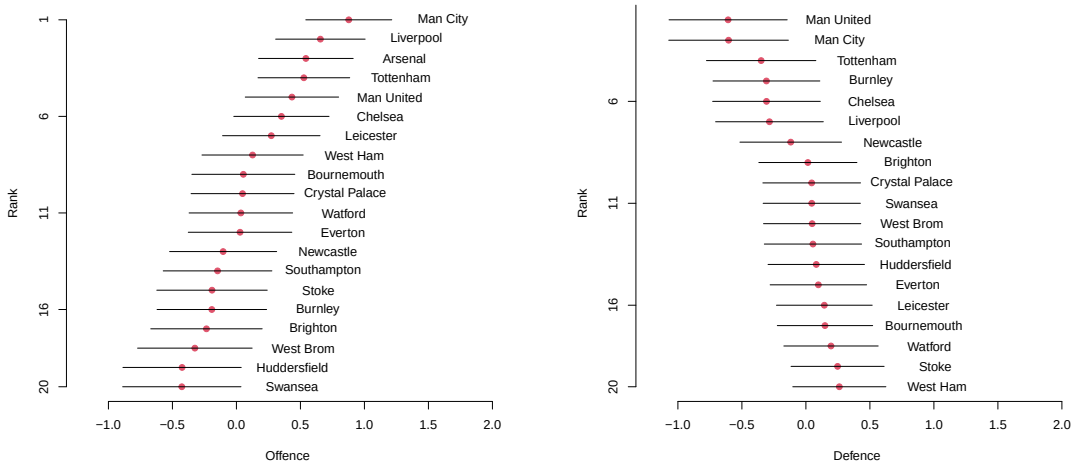


Figure 5: Offensive and Defensive parameters estimated by the Dixon and Coles model.

rameter, the Double Poisson model is the selected one. However we proceed with the Dixon and Coles as it can capture the small correlation we have seen in the data, providing a slight improvement with respect to the Double Poisson model.

Figure 5 shows the estimated offensive and defensive parameters for each team together with their 95% confidence intervals. Manchester United is performing very well in both. Manchester United has the better defence. Some interesting points are for example Swansea that while has the weakest defence, has a very good defence.

It is interesting that the home effect parameter was estimated as 0.29, which implies that the home team has an additional bonus of 0.29 goals per match. The 95% confidence interval equals (0.17,0.41) which implies that the effect is present and statistically significant. To have an indication about the magnitude of such a home effect, consider the case that two teams of equal offensive and offensive ability are playing one against the other. Suppose further that on average they can score 1.34 goals (this is the observed mean for the data we have). The additional effect of the home team, will increase the probability that the home team will win from 0.37 to 0.44 which is not negligible.

4.4 Goodness of fit - Replicating the League with simulation

Table 2 shows the actual versus the expected points for each team together with ranks based on the simulations, namely 10000 replications from the Dixon-Coles model (see Section 3.7. We report also the mean rank for each team, the probability of winning the EPL, the probability of being at the 4st top places leading to Champions League (CL) and finally the probability of relegation.

Also Figure 6 shows the points expected per team based on the 10000 simulations. The actual values of points are the black dots and the red ones are the means over the 10000 replications. The line indicates a 95% confidence interval based on the observed 2.5% percentile and the 97.5% percentile of the simulated values. We see some differences between the expected and actual points but it seems that the model fits well the data and the reproduced value are close to the observed ones.

Some interesting findings are the following:

		Points		Mean	Probability for		
	Team	Actual	Expected	Rank	1st	CL	Releg.
CL	Man City	100	93.94	1.14	0.88	1.00	0.00
	Man United	81	77.69	3.29	0.03	0.83	0.00
	Tottenham	77	76.26	3.51	0.03	0.78	0.00
	Liverpool	75	79.09	3.02	0.05	0.87	0.00
	Chelsea	70	67.74	5.16	0.00	0.28	0.00
	Arsenal	63	66.44	5.40	0.00	0.22	0.00
	Burnley	54	49.28	9.91	0.00	0.00	0.02
	Everton	49	44.38	12.09	0.00	0.00	0.07
	Leicester	47	51.26	9.27	0.00	0.01	0.01
	Newcastle	44	46.30	11.23	0.00	0.00	0.04
	Crystal Palace	44	46.58	11.15	0.00	0.00	0.04
	Bournemouth	44	44.12	12.27	0.00	0.00	0.08
	West Ham	42	42.83	12.80	0.00	0.00	0.11
	Watford	41	41.04	13.69	0.00	0.00	0.17
	Brighton	40	39.47	14.45	0.00	0.00	0.19
	Huddersfield	37	32.68	17.36	0.00	0.00	0.59
	Southampton	36	40.28	14.04	0.00	0.00	0.19
Rel	Swansea	33	33.68	17.02	0.00	0.00	0.52
	Stoke	33	33.34	17.08	0.00	0.00	0.55
	West Brom	31	35.90	16.08	0.00	0.00	0.41

Table 2: Actual versus expected points values together with ranks. Namely we can see the actual number of points together with the expected ones by the simulations based on 10000 replications from the Dixon-Coles model. We report also the mean rank for each team, the probability of winning the EPL, the probability of being at the 4st top places leading to Champions League and finally the probability of relegation

- Manchester City won somewhat more points than expected.
- Liverpool finishing at the 4-th place could have been better, its expected points are more, very close to Manchester United. In fact from Table 2 one can see that Liverpool had more chances to finish at the second place.
- At the bottom of the table we can see that West Bromwich had less points than expected and actually their expected points were not very different than teams that avoided relegation.

4.5 Prediction

In this subsection we use the model for prediction purposes. Suppose that we are in the end of January 2018. The day 25 of the season has ended and we have the results up to this time point. At this particular time point we estimate the model and then we simulate the rest of the season to predict as early as possible the final winner. We emphasize that we estimate the parameters of the model with only the data up to day 25. We report similar quantities as in Table 2.

From Table 3 one can see some interesting points. Very few teams performed differently for the remaining part of the season. Leicester is 8.5 points below than expected, Chelsea has less

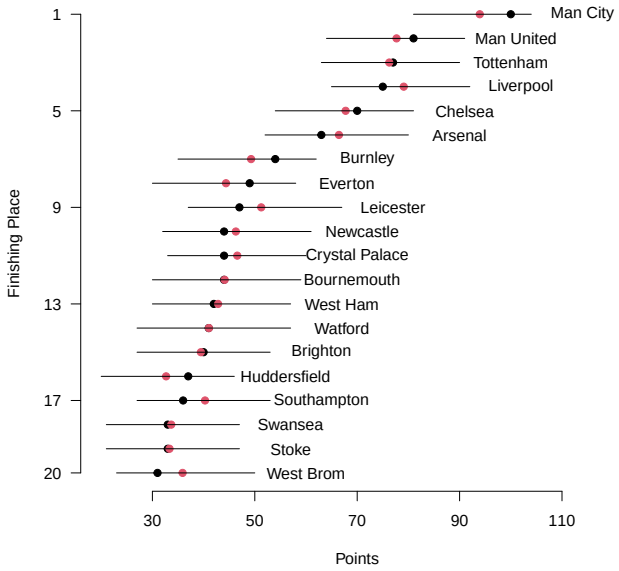


Figure 6: Actual number of points for each team (black dot) together with expected points in red dots. The line is a 95% confidence interval for the points created based on the simulated replications of the League.

5 points than what was expected after day 25; at that point Chelsea was in a better position than Tottenham. In general, Tottenham had a better performance and this led to the Champions League position. Also Manchester City was already almost sure champion and he did what was expected as far as the number of final points. Newcastle did very well in the remaining matches and they gain almost 8 more points than expected, moving away from the relegation zone.

5 Summary

In this dissertation we have applied some typical model for modelling soccer games. Modelling soccer has found increasing interest from a wide range of stakeholders and it is an active research area. We have applied some models to data from EPL of the season 2017-2018 aiming at showing the predictive ability of the models. Results show that the models are quite relevant to reproduce the dynamics of the championship. We acknowledge that more sophisticated models can be used if more detailed data are available. Here we have use rather simple models that take into account the teams ability and the home effect. We have also checked whether other variables like the shape of the team measured by the last 5 matches can have a predictive ability, but we found that this is neither statistically significant nor improves the predictive ability of the models. We do not present such models here due to space restrictions.

References

Allianz (2022). Allianz and football. *https://www.allianz.com/en/about-us/sports-culture/football/allianz-football.html*.

Team	Final	Until Day 25	Expected	Mean Rank	Probability 1st	CL	Releg.
Man City	100	68	99.47	1.00	1.00	1.00	0.00
Man United	81	53	81.40	2.50	0.00	0.97	0.00
Tottenham	77	48	73.84	4.29	0.00	0.50	0.00
Liverpool	75	50	77.49	3.38	0.00	0.84	0.00
Chelsea	70	50	75.13	3.96	0.00	0.66	0.00
Arsenal	63	42	64.56	5.99	0.00	0.02	0.00
Burnley	54	35	52.65	8.02	0.00	0.00	0.00
Everton	49	31	47.66	9.56	0.00	0.00	0.00
Leicester	47	34	55.49	7.41	0.00	0.00	0.00
Newcastle	44	24	36.16	15.47	0.00	0.00	0.26
Crystal Palace	44	26	39.34	13.43	0.00	0.00	0.08
Bournemouth	44	28	45.07	10.49	0.00	0.00	0.00
West Ham	42	27	40.31	12.83	0.00	0.00	0.04
Watford	41	27	41.14	12.37	0.00	0.00	0.04
Brighton	40	24	35.52	15.96	0.00	0.00	0.32
Huddersfield	37	24	33.98	16.93	0.00	0.00	0.49
Southampton	36	23	36.01	15.60	0.00	0.00	0.29
Swansea	33	23	35.21	16.17	0.00	0.00	0.36
Stoke	33	24	34.08	16.88	0.00	0.00	0.48
West Brom	31	20	32.64	17.76	0.00	0.00	0.64

Table 3: Final standings and prediction based on the first 25 matches only. As in the previous Table we report several quantities using only data up to day 25

Dixon, M. and M. Robinson (1998). A birth process model for association football matches. *Journal of the Royal Statistical Society: Series D (The Statistician) 47*(3), 523–538.

Dixon, M. J. and S. G. Coles (1997). 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.

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

Karlis, D. and I. Ntzoufras (2005). Bivariate Poisson and diagonal inflated bivariate Poisson regression models in `R`. *Journal of Statistical Software 14*, 1–36.

Karlis, D. and I. Ntzoufras (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics 20*(2), 133–145.

Keogh, F. and G. Rose (2013). Football betting - the global gambling industry worth billions. *https://www.bbc.co.uk/sport/football/24354124*.

Lee, A. J. (1997). Modeling scores in the Premier League: is Manchester United really the best? *Chance 10*(1), 15–19.

Leitner, C., A. Zeileis, and K. Hornik (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the EURO 2008. *International Journal of Forecasting 26*(3), 471–481.

Muller, J. (2022). Visualising possession: Where is every Premier League team having more of the ball than their opponents? *https://theathletic.com/3118980/2022/02/09/visualising-possession-where-is-every-premier-league-team-having-more-of-the-ball-than-their-opponents/*.

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

Reep, C. and B. Benjamin (1968). Skill and chance in association football. *Journal of the Royal Statistical Society: Series A (General) 131*(4), 581–585.

Scarf, P., A. Khare, and N. Alotaibi (2021). On skill and chance in sport. *IMA Journal of Management Mathematics 33*(1), 53–73.

Scarf, P. and J. S. Rangel Jr (2017). Models for outcomes of soccer matches. *Handbook of Statistical Methods and Analyses in Sports Boca Raton, FL: Chapman and Hall/CRC*, 341–354.

Sheehan, D. (2018). Predicting football results with statistical modelling: Dixon-coles and time-weighting. *https://dashee87.github.io/football/python/predicting-football-results-with-statistical-modelling-dixon-coles-and-time-weighting/*.

SofaScore (2022). Attack momentum. *https://www.sofascore.com*.

SportsMole (2022). Premier League defence stats. *https://www.sportsmole.co.uk/football/premier-league/best-defence.html*.

Whitmore, J. (2021). What is match momentum? *https://theanalyst.com/eu/2021/11/what-is-match-momentum/*.

Appendix

Log-likelihood function for the Dixon and Coles model

Consider the Dixon and Coles model. For a number of matches $k = 1, \ldots, n$, and n_t teams we have:

$$\begin{aligned} \ell(\alpha_{\mathrm{i}}, \beta_{\mathrm{i}}, \rho, \gamma; i = 1, ..., n_t) &= \prod_{k=1}^{n} \{\tau_{\lambda_k,\mu_k}(x_k, y_k) \frac{\exp(-\lambda_k)\lambda_k^{x_k}}{x_k!} \frac{\exp(-\mu_k)\mu_k^{y_k}}{y_k!}\} \\ &\propto \prod_{k=1}^{n} \{\tau_{\lambda_k,\mu_k}(x_k, y_k) \exp(-\lambda_k)\lambda_k^{x_k} \exp(-\mu_k)\mu_k^{y_k}\}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} \lambda_k &= \alpha_{i(k)}\beta_{j(k)}\gamma, \\ \mu_k &= \alpha_{j(k)}\beta_{i(k)}. \end{aligned}$$

By taking the logarithm of equation 7, we end up with:

$$\begin{aligned} L(\alpha_{\mathrm{i}}, \beta_{\mathrm{i}}, \rho, \gamma; i = 1, ..., n_t) &= \sum_{k=1}^{n} \log\left(\tau_{\lambda_k,\mu_k}(x_k, y_k) \exp(-\lambda_k)\lambda_k^{x_k} \exp(-\mu_k)\mu_k^{y_k}\right) \\ &= \log\left(\tau_{\lambda_k,\mu_k}(x_k, y_k)\right) + \log\left(\exp(-\lambda_k)\right) + \log\left(\lambda_k^{x_k}\right) + \log\left(\exp(-\mu_k)\right) + \log\left(\mu_k^{y_k}\right) \\ &= \log\left(\tau_{\lambda_k,\mu_k}(x_k, y_k)\right) - \lambda_k + x_k \log(\lambda_k) - \mu_k + y_k \log(\mu_k) \end{aligned}$$

This cannot be maximized without numerical methods. We have used `optim` function in `R` to maximize it. We use as initial values the offensive and defensive parameters from a Double Poisson model and set $\rho = 0$.

However, as previously mentioned, a team's performance tends to be dynamic, varying from one time period to another Dixon and Coles (1997) considered a time-decaying weighting function for the log-likelihood to down-weight the importance of the matches further in the past. No down-weighting is applied here. Setting $rho = 0$ we get the Double Poisson model.

Example for the calculation of score probabilities

Consider the following Table with the defensive and offensive parameters for some teams.

	Coefficients
Constant	0.427229
Off. Chelsea	-0.175425
Off. Liverpool	0.141749
Off. Man City	0.287016
Def. Chelsea	-0.293312
Def. Liverpool	-0.518996
Def. Man City	-0.687380
Home effect	0.255933

Table 4: Assumed parameter values

For each of the three teams Chelsea, Liverpool and Manchester City, we have the attacking and defensive performances of the corresponding team, respectively and also the home advantage factor. Let us look at an example. Assume that Liverpool is playing Manchester City at home. Then:

$$\begin{aligned}\lambda_{\text{home}} &= \beta + \beta_{\text{home}} + \text{att}_{\text{home}}(\text{Liverpool}) + \text{def}_{\text{away}}(\text{Man City}) \\ &= \exp(0.427229 + 0.255933 + 0.141749 - 0.687380) \\ &= 1.1474.\end{aligned}$$

and

$$\begin{aligned}\lambda_{\text{away}} &= \beta + \text{att}_{\text{away}}(\text{Man City}) + \text{def}_{\text{home}}(\text{Liverpool}) \\ &= \exp(0.427229 + 0.287016 - 0.518996) \\ &= 1.2156.\end{aligned}$$

So, if Liverpool and Manchester City played each other multiples times, on average, Liverpool would score 1.1474 goals and Manchester City 1.2156 goals. The probability of Liverpool scoring μ number of goals and Manchester City to score ν goals, from the Double Poisson model is given by:

$$\frac{\lambda_{home}^{\mu} \cdot e^{-\lambda_{\text{home}}}}{\mu!} \times \frac{\lambda_{away}^{\nu} \cdot e^{-\lambda_{\text{away}}}}{\nu!}.$$

We have calculated the matrix of the probabilities for certain combinations of scores Table 5 shows the probability for all scores up to 6-6. The diagonal of the above table gives us the probabilities that the two teams draw. The upper diagonal entries correspond to the probabilities of an away win (Manchester City wins) whilst the lower diagonal entries to a home win. For example, the probability Liverpool wins 2-1 against Manchester City is 7.53%. Overall we have that Manchester City will win with probability 0.377, Liverpool will win with probability 0.343 and we will have a draw with probability 0.278.

		Manchster City						
		0	1	2	3	4	5	6
Liverpool	0	0.0941	0.1145	0.0696	0.0282	0.0086	0.0021	0.0004
	1	0.1080	0.1313	0.0798	0.0324	0.0098	0.0024	0.0005
	2	0.0619	0.0753	0.0458	0.0186	0.0056	0.0014	0.0003
	3	0.0237	0.0288	0.0175	0.0071	0.0022	0.0005	0.0001
	4	0.0068	0.0083	0.0050	0.0020	0.0006	0.0002	0.0000
	5	0.0016	0.0019	0.0012	0.0005	0.0001	0.0000	0.0000
	6	0.0003	0.0004	0.0002	0.0001	0.0000	0.0000	0.0000

Table 5: Probabilities of certain scores for the example Liverpool versus Manchester City