DEPARTMENT OF STATISTICS
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Bringing Order to the Chaos of Football

Leonidas Ntrekos

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of B.Sc. in Statistics

Athens
2026

Ευχαριστίες

Κύριο μου μέλημα είναι να ευχαριστήσω τον κύριο Καρλή, κατ' αρχήν για την υπομονή που είχε να απαντήσει όλες τις φορές που τον ενόχλησα, την κατανόηση που έδειξε για να λύσει όλες τις απορίες που μου δημιουργήθηκαν και τον χρόνο που αφιέρωσε για να διεκπεραιωθεί τελικά η εργασία. Θέλω επίσης να τον ευχαριστήσω για όλα τα πράγματα που μου έμαθε και τη γνώση που μου μετέφερε όλα αυτά τα χρόνια. Επιπλέον, αισθάνομαι ιδιαίτερα ευγνώμων για την επιμονή του να πραγματοποιηθεί εν τέλει η πτυχιακή, καθώς από αυτήν τη διαδικασία αποκόμισα πράγματα τα οποία δεν φανταζόμουν όταν ξεκινούσα.

Περαιτέρω, θα ήθελα να εκφράσω τις ευχαριστίες μου προς όλους τους καθηγητές, καθώς οι διαλέξεις τους υπήρξαν καθοριστικές στη διαμόρφωση της σκέψης μου.

Τέλος για τους γονείς, τα αδέρφια και τους φίλους μου δεν θα εκφράσω κάποια ευχαριστία κάθως αυτήν προσπαθώ να την δείχνω με την κάθε ημέρα.

Στην γιαγιά μου... Την Μαρία,
Συγγνώμη αλλά ακόμη δεν ξέρω γεωμετρία

I have nothing, I know nothing, I can
do nothing, I have learned nothing.
How wondrous this is!
— **Hermann Hesse, Siddhartha**

Bringing Order to the Chaos of Football

Abstract

Sports data analysis, and more specifically data related to football, is able to quantify the events that occur on the field, explain the mechanics of the game, give substance to what we perceive empirically, and ultimately predict future outcomes. It essentially constitutes the tool through which we can transform data — that is, facts we already know — into information, namely meaningful knowledge that can have a decisive impact on the sports industry, from the decisions made by coaches and management, to the financial aspects of competitions and the betting markets. In this thesis, we apply a range of statistical methods, with the ultimate goal of constructing predictive models for the final outcome of a match, or, more abstractly put, of bringing order to the chaos of football.

Specifically, we model a team's scoring intensity based on the Poisson distribution, the negative binomial distribution, as well as a hierarchical (or compound) model in which one component is a Poisson-distributed variable and the other a binomial-distributed variable. Subsequently, we model the goal difference using the Skellam distribution, and extend this approach by incorporating correlation between the two halves via copulas. Finally, we revisit the Poisson distribution by constructing a four-dimensional model based on the two halves, and attempt to capture the resulting correlations once again through copulas.

As we present the various approaches used, we outline the theoretical foundations on which they rely, explain the interpretations behind their results, and discuss their limitations. Finally, we apply the most suitable models we developed in order to predict the remainder of the current season (at the time of writing, we are approximately halfway through) of the five major European football leagues. The present thesis may serve as a small and introductory guide to the construction of statistical models for predicting the outcomes of football matches.

Βάζοντας τάξη στο χάος του ποδοσφαίρου

Περίληψη

Η ανάλυση αθλητικών δεδομένων και, πιο συγκεκριμένα, δεδομένων που αφορούν το ποδόσφαιρο, δύναται να ποσοτικοποιήσει τα γεγονότα που συμβαίνουν στο γήπεδο, να εξηγήσει πώς παίζεται το παιχνίδι, να δώσει υπόσταση σε αυτό που αντιλαμβανόμαστε εμπειρικά, και ακόμη να προβλέψει τι εν τέλει θα συμβεί. Αποτελεί, ουσιαστικά, το εργαλείο μέσω του οποίου μπορούμε να μετατρέψουμε δεδομένα, δηλαδή γεγονότα που ήδη γνωρίζουμε, σε πληροφορία — σε ουσιαστική γνώση η οποία μπορεί να έχει καθοριστική επίδραση στη βιομηχανία του αθλήματος, από τις αποφάσεις που λαμβάνουν προπονητές και διοικήσεις, μέχρι τα οικονομικά των διοργανώσεων και τις στοιχηματικές αγορές. Στην πτυχιακή αυτή εφαρμόζουμε ένα πλήθος στατιστικών μεθόδων, με τελικό στόχο να κατασκεύασουμε προβλεπτικά μοντέλα που αφορούν την τελική έκβαση ενός αγώνα ή, αν το θέσουμε πιο αφηρημένα, να βάλουμε τάξη στο χάος του ποδοσφαίρου.

Συγκεκριμένα, μοντελοποιούμε τα γκολ που σκοράρει μία ομάδα με βάση την κατανομή Poisson, την αρνητική διωνυμική, καθώς και μέσω ενός ιεραρχικού μοντέλου, στο οποίο το ένα μέρος βασίζεται στην κατανομή Poisson και το άλλο στην διωνυμική κατανομή. Στη συνέχεια, μοντελοποιούμε τη διαφορά των γκολ με βάση την κατανομή Skellam, ενώ επεκτείνουμε αυτήν την τεχνική ενσωματώνοντας συσχέτιση μεταξύ των ημιχρόνων μέσω copula. Τέλος, επανερχόμαστε στην κατανομή Poisson, κατασκευάζοντας ένα τετραπλό μοντέλο με βάση τα ημίχρονα και επιχειρούμε να αποτυπώσουμε τις συσχετίσεις που προκύπτουν, και πάλι μέσω copula.

Κατά την παρουσίαση των διαφόρων προσεγγίσεων που εφαρμόστηκαν, αναλύουμε το θεωρητικό υπόβαθρο στο οποίο βασίζονται, ερμηνεύουμε τα αποτελέσματά τους και παράλληλα συζητάμε τους περιορισμούς τους. Και, τέλος, εφαρμόζουμε τα πλέον καταλληλότερα μοντέλα που κατασκευάσαμε, προκειμένου να προβλέψουμε το υπόλοιπο της τρέχουσας αγωνιστικής περιόδου (τη στιγμή συγγραφής βρισκόμαστε περίπου στα μισά) των πέντε μεγάλων ευρωπαϊκών πρωταθλημάτων ποδοσφαίρου. Η παρούσα πτυχιακή εργασία μπορεί να αποτελέσει ένα μικρό και ίσως εισαγωγικό εγχειρίδιο για την κατασκευή στατιστικών μοντέλων που προβλέπουν την έκβαση ποδοσφαιρικών αγώνων.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Mathematics and Statistics cannot compete with Football. The sport can unite whole nations for the same purpose, like nothing else. It can offer the highest quality drama that even the best Hollywood directors couldn't think of. It combines every emotion an individual can feel. Heartbreak when your team relegates, utterly Joy when the team wins the league; envy when the local opponent has the slightest victory. It is full of remarkable stories of both triumph and failure, at both individual and team levels. It speaks to people of all kinds, across every country and social class, and that is why thousands fill up stadiums every weekend - and middays- and billions are captured in front of the television to watch the World Cup. It is a market and a business industry that is increasing tremendously, offering great opportunities to people.

Football also has a shady side. Corruption follows the game at every tier, from grassroots divisions all the way up to the Champions League. Some clubs have grown so large that they are likened to entire governments. Powerful and dangerous individuals use football to wash their image and as a leverage to influence for their own benefit.

Compare this world with mathematics and statistics. Obscure academic journals, textbooks and notes lie unread in half-empty libraries collecting dust. Seminars and paper presentations are attended by a couple kind professors followed -from obligation- by a small group of PhD students. There is no competition between the two.

Yet, there is a narrow frontier where these two worlds collide. Given the tools passed from mathematics, football can be measured, quantified, and forced into a defined shape. Football analytics is precisely the intersection where randomness, chaos, action, and emotion confront the strict structure and rigidity of statistical models, the pursuit of explanation, and the theoretical curiosity that drives mathematics. This collision resembles one of the oldest questions in science: what happens when an immovable object meets an unstoppable force?

In football, the events unfold not in equations, but in ninety minutes of play, where probability, strategy, and human behaviour converge in real time. Football analytics is not a replacement for the sport, nor an attempt to strip away its magic; it is the effort to understand, explain, and illustrate the patterns within the chaos, to define why a moment happens the way it does, and to explore whether mathematics can ever truly capture the unpredictability of the game. This thesis is precisely that, an attempt to bring order to the chaos of football through statistical models.

In the past, numerous researchers and practitioners have concerned themselves with modelling and analysing football data. As early as the mid-fifties, Moroney (1951) showed that even though satisfactory results were yielded from fitting a Poisson Distribution, they could still be improved from fitting an alternative, the Negative Binomial Distribution. In the same notion, Reep et al. (1971) examined the fit of the same distribution to scores from football matches.

A first approach to estimating a model that accounts for the different performance qualities of each team using the Poisson distribution is presented by Maher (1982). Later, Lee (1997) and Karlis and Ntzoufras (2000) observed that there is a relatively minor but still existing correlation between the number of goals scored by the opposing teams. This has been ignored in most approaches as it demands more sophisticated modelling techniques. This independent Poisson model was modified by Dixon and Coles (2002), who incorporated an inflation parameter for the low scores, and later extended by Karlis and Ntzoufras (2003), who introduced the bivariate Poisson model capturing the dependence of oppositions. This was later further extended in Karlis and Ntzoufras (2005) by accounting for the surplus of draws exhibited compared to what the bivariate Poisson predicted. Additionally, McHale and Scarf (2011) modelled the dependence of goals scored by opposing teams using copulas.

Different modelling approaches have also been proposed. In Karlis and Ntzoufras (2006), the Skellam distribution was used to directly model the score differences through a Bayesian approach rather than the teams' scoring rate separately. More recently, in Karlis et al. (2024), the same logic was applied in German handball with the addition of accounting for the dependence between halves through copulas.

Additionally, football (along with other sports as well) analysis/modelling does not only concern scoring patterns and predicting final outcomes, but also describing the game and evaluating teams through numbers. For example, Reep and Benjamin (1968) modelled the number and type of passing moves within a game. In the same notion, Pelechrinis and Winston (2020) used a Skellam regression model to determine what factors determine team performance. Clarke and Norman (1995) investigated the advantage of playing at home, as well as Pollard (1986).

Other researchers have looked into other aspects of the sport as well. For example, Dawson et al. (2007) and Sutter and Kocher (2004) have looked at the referee decision-making and tried to determine whether there is bias or not. Even Garicano et al. (2001) studied their behaviour under social pressure, which is applied by the crowd watching the game. For an overall overview with great detail, applications, examples, and even code that can be reproduced, one can refer to the excellent work of Sumpter (2016) as well as the recent publication of Egidi et al. (2025).

Generally, as one can imagine, sports analytics and especially football analytics are an interesting field of research. It offers the opportunity to apply theoretical tools to a subject that concerns a great proportion of the population and where a great amount of wealth is invested. The works of researchers we have reported are a tiny glimpse of the actual body of research that exists out there.

In this work, we concentrate on modelling football outcomes, interpreting these results, and examining the intuitive nature of the models themselves. In the following chapters, we present the theoretical background used for modelling, which is probability theory. We examine how modelling focused on explanation differs from modelling aimed

at prediction. We review the most well-established work on predictive modelling and examine the limitations these approaches may encounter. In the later sections, we present some more sophisticated modelling techniques and compare the results with more straightforward approaches. And of course, do not miss out on the final chapter, where we make predictions for the current season 2025-2026 of the 5 Major European Championships (the data available extend up to January 2026, which is both when the predictions will be generated and when this thesis will be completed), after having concluded on the most suitable options based on the findings of the preceding chapters.

If the reader had actually the kindness to read the introduction, first of all, thank you, and can jump directly to the predictions on chapter 4 to check whether the results are accurate (if by the time you read this, season 2025-2026 has been completed). Had the results been satisfactory, the reader could then explore the various methods discussed and applied in Chapter 3. If the reader has no previous experience in statistics, forecasting, or generally modelling data, we advise them to start from Chapter 2 for a gentle introduction to probability theory.

In the pursuit of setting order to football's Chaos one must rely on a perfectly working compass. So all models, tables, figures, and generally results produced in this thesis are generated based on data concerning the 5 "Big" European Championships, that is, the English Premier League, the Italian Serie A, the Spanish La Liga, the German Bundesliga, and the French Ligue 1 during the 5 seasons from 2020 to 2025. Without further ado, let the data exploration begin, but first, some pre-work must be done.

Chapter 2

Some Probability Theory

"We can see that probability theory is, at its foundation, nothing more than pure logic and common sense, translated into numerical calculations that allow us to precisely estimate what a regular mind senses intuitively, even if it cannot always articulate it.... It is remarkable that this discipline, which originated from the study of gambling, may become the most fundamental area of human knowledge. The most important questions are, for the most part, problems of probability." These are the words of the famous French mathematician and astronomer Pierre-Simon, Marquis de Laplace, as quoted in Ross (1998). Although this claim may sound somewhat extravagant, there is no doubt that probability theory has evolved into a powerful instrument for most scientists. Indeed, an informed individual does not ask, "Is this true?" but rather states the question as, "What is the probability that this is true?"

Sports analytics—and in particular football analytics, which is the focus of this thesis—are no exception. Before we turn to models, diagrams, results, predictions, and conclusions, we must first establish the basis upon which we will discuss, the foundation from which all approaches discussed stem. It is essential to present the theoretical ground that will be used to build the work that follows.

Although the subjects addressed in this chapter may appear too mathematical or theoretical, and that is not related to football, they should not be dismissed. These ideas constitute the basis on which more sophisticated models and methods are constructed. Without a solid understanding of them, it is impossible to fully comprehend the practical modelling techniques.

2.1 Random Variable

A standard approach when analysing random experiments is to concentrate on numerical characteristics that describe and summarize their outcomes. It is usually convenient to link each possible outcome with a specific numerical value. For instance, in a football match, we might keep count of the number of goals scored, total shots, shots on target attempted, the number of corner kicks, fouls rewarded and many other related metrics. When we treat the whole match as a random experiment, each of these numbers provides partial insight into what occurred during the time of play. Quantities of this nature are referred to as **random variables**.

A random variable is a quantity whose outcome is determined through the result of a random process. Mathematically, we say that a random variable X is a function

$$X : \Omega \to \mathbb{R},$$

for some probability space Ω.

The total number of possible values a random variable can take is called **range of values** and is denoted as S_X. Random Variables can be either continuous or discrete. The difference between the two is that the first can take a finite or an infinite, but countable set of numbers. Whereas a continuous variable takes values in an uncountable set. Therefore, we see that the type of each random variable is determined by the structure of its range.

If X is a discrete random variable, then its range S_X is made up of countably distinct values, which we denote by S_X. Equivalently, we can express this as

$$S_X = \{x_1, x_2, x_3, \ldots\}.$$

Observe that $x_1, x_2, x_3, \ldots$ denote specific values, or realizations, of the random variable. By standard notation, random variables are written with uppercase letters (e.g., X), whereas their particular observed values are written in lowercase (e.g., x). For a discrete random variable X, we are interested in finding the probabilities of events of the form $X = x_k$. For example, the event $A = \{X = x_k\}$ consists of all outcomes $s \in S_X$ such that the random variable assigns the value x_k to s, that is

$$A = \{\omega \in \Omega | X(\omega) = x_k\}.$$

The range of a random variable X, shown by S_X, is the set of possible values of X. A continuous random variable may assume any real value, that is, any value in $\mathbb{R} = (-\infty, +\infty)$. Since most football-related measures (such as goals, fouls, or passes) are naturally count-based, this thesis focuses exclusively on discrete random variables. For the remainder, we must keep in mind that

X is a discrete random variable if its range is countable.

2.2 Probability Mass Function

The probabilities of events $\{X = x_k\}$ are formally shown by the **probability mass function (pmf)** of X. Let X be a discrete random variable with a set of values S_X, then the function $P : S_X \to [0, 1]$ which is defined as

$$P_X(x) = P(X = x), \ \forall \ x \in S_X,$$

is called the **pmf** of X.

Thus, the pmf is a probability measure that assigns probabilities to the possible values of a random variable. Probability, intuitively, is the frequency of an experiment's or a process's outcome when repeated n times under the same conditions, where n must be a very large number close to infinity.

The probability mass P of a discrete random variable X with a set of values S_X has the following properties:

1. **Non negativity** :
$$P_X(x) = P(X = x) \geq 0 \ \forall x \in S_X$$
2. **Normalization** $x \in S_X$
$$\sum_{x \in S_X} P_X(x) = 1$$
3. **Additivity**:
$$P(X \in T) = \sum_{x \in T} P_X(x), \text{ for any T} \in S_X$$

When we observe a random variable X, we are usually interested in its realizations, a sequence of numbers $x_1, x_2, \ldots, x_N$. For that sequence of numbers, we would like to know the general area or space that they are concentrated in. Based on the **average** or **mean** of these values can can obtain the general tendency of the random variable we observe.

The **mean value** (or expected value) of discrete random variable X with set of values S_X and probability mass function $P(x)$, is defined as

$$\mu = \mathbb{E}[X] = \sum_{x_k \in S_X} x_k P(X = x_k) = \sum_{x_k \in S_X} x_k P_X(x_k).$$

The expected value $\mathbb{E}[X]$ is a weighted average of all the possible values, with each value weighted by its probability. Conceptually, it expresses that the values of the random variable X tend to vary around the value μ. Similarly, the variance of a random variable X provides a quantitative indication of the degree of that variation around the value μ. The variance of a discrete random variable X with mean value μ is

$$Var(X) = \mathbb{E}[(X - \mu)^2],$$

or with equivalent written

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

The interpretation of the quantity above can be that the higher the variance, the greater the dispersion around the mean. As a result, if X is always equal to $\mathbb{E}[X]$, then $Var(X) = 0$. Variance is a key metric because it quantifies the information contained in a random variable. For example, as mentioned above, if $Var(x) = 0$, no information can be extracted from that variable X as it always results in $\mathbb{E}[X]$. On the contrary, a random variable with high Variance corresponds to a random variable that results in many different outcomes, and so a phenomenon worth observing. The above were adapted from Pishro-Nik (2014), Ross (1998) and mainly from Κοντογιάννης and Τουμπής (2015).

2.3 Bernouli and Binomial

Suppose that we undergo a trial or an experiment, and that each outcome can be classified as either a success or a failure. If we define that $X = 1$ counts as a success and that $X = 0$ counts as a failure, the pmf of the random variable X is defined as,

$$\mathrm{P}\left(X_i = x_i\right) = \begin{cases} p, & \text{if } x_i = 0, \\ 1 - p, & \text{if } x_i = 1, \end{cases} \tag{2.1}$$

where p, $0 \leq p \leq 1$, is the probability of success. Such a random variable X is called a **Bernoulli random variable** if its **pmf** is given by (2.1) for some $p \in (0, 1)$, and is denoted by

$$X \sim \text{Bernoulli}(p).$$

Suppose now that we perform n independent and identically distributed trials, each resulting in a success with probability p and a failure with probability $1 - p$. If we denote by X the number of successes that occur during the n trials, then X is called a **Binomial** random variable with parameters (n, p) and is denoted by

$$X \sim \text{Binomial}(n, p),$$

and its **pmf** is given by

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n. \tag{2.2}$$

We can verify this by first observing that, under the assumed independence of the trials, the probability of any given sequence of n outcomes with x successes and $n - x$ failures is $p^x(1 - p)^{n-x}$. From that, equation (2.2) follows, since there are $\binom{n}{x}$ different sequences of n outcomes that yield x successes and $n - x$ failures. This is easier to understand by noting that there are $\binom{n}{x}$ ways to choose x trials that succeed.

2.4 Poisson

A widely used distribution in sports analytics is the Poisson distribution (see, for example Dixon and Coles (2002), Karlis and Ntzoufras (2000), Maher (1982)). The Poisson distribution models the number of events occurring in a fixed interval when events happen at a constant mean rate λ. We symbolise that a random variable x follows a Poisson distribution as

$$X \sim \mathcal{P}(\lambda) \ , \ \lambda > 0,$$

and its **pmf** is given by

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2 \ldots \tag{2.3}$$

The distribution has both mean and variance equal to $\mathbb{E}[X] = Var[X] = \lambda$. This characteristic will be important in the discussion that follows, when we apply it to modelling team goals. The Poisson random variable has wide-ranging applications beyond

sports analytics, since it can serve as an approximation to a binomial random variable with parameters (n, p), as described in Ross (1998), when n is large and p is sufficiently small so that the product np remains of moderate size.

Let X be a binomial random variable with (n, p) parameters, and let $\lambda = np$. Then,

$$\begin{aligned} P(X = x) &= \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} \\ &= \frac{n!}{(n-x)!x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)\dots(n-x-1)}{n^x} \frac{\lambda}{x} \frac{\left(1-\lambda/n\right)^n}{\left(1-\lambda/n\right)^x} \end{aligned}$$

Now, for large values of n and moderate values of λ,

$$\left(1 - \frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \quad \frac{n(n-1)\dots(n-x-1)}{n^x} \approx 1, \quad \left(1 - \frac{\lambda}{n}\right)^x \approx 1.$$

Therefore, for large values of n and moderate values of λ

$$P(X = x) \approx e^{-\lambda} \frac{\lambda^x}{x!}.$$

In summary, if we consider n independent trials, each yielding a success with probability p, then in the regime where n is large and p is sufficiently small so that the product np remains of moderate size, the total number of successes can be approximated by a Poisson random variable with parameter $\lambda = np$.

2.5 Negative Binomial

An alternative to the Poisson that accounts for cases where the mean is not equal to the variance (see, for example, Dawson et al. (2007), who modelled yellow cards) is the **Negative Binomial**. The negative binomial distribution models the number of failures before the rth success in a sequence of independent Bernoulli trials, where each trial has only two possible outcomes: success or failure with probability p and $1 - p$ respectively. We denote that a random variable X follows a negative binomial as

$$X \sim NB(r, p),$$

with its **pmf** given by,

$$P(X = k) = \binom{k + r - 1}{k} p^r (1-p)^k, \quad k = 0, 1, 2, \dots$$

Its mean is equal to $\mathbb{E}[X] = r(1-p)/p$ and its variance is equal to $Var[X] = r(1-p)/p^2$. If $r = 1$, the negative binomial becomes a **geometric distribution**, but that is a different discussion. Equivalently, the **pmf** can be written as,

$$P(X = k) = \binom{r}{r + \mu}^r \frac{\Gamma(r + k)}{k!\Gamma(r)} \left(\frac{\mu}{r + \mu}\right)^k. \tag{2.4}$$

Here, the mean is given by $\mathbb{E}[X] = \mu$, while the variance is $Var[X] = \mu + \mu^2/r$. In the limit as $r \to \infty$, this expression reduces to the Poisson model.

As mentioned earlier, the rationale for using the negative binomial distribution instead of the Poisson in sports analytics is to account for the overdispersion that frequently occurs in sports data (as discussed in Moroney (1951)). We would like to remind, that overdispersion means the observed variance exceeds the observed mean. Accordingly, under-dispersion implies that the observed variance is less than the observed mean. In the equation (2.4), t measures the degree of overdispersion.

2.6 Compound Poisson

The Compound Poisson is a hierarchical model consisting of a Binomial Distribution and a Poisson distribution. The main idea is that we describe the probability p of an event occurring, given that the number of trials is a Poisson-distributed random variable. In a football setting, for instance, N might represent the number of shots or shots on target, and p the chance that any given shot results in a goal (though many similar scenarios could serve as examples).

Let N be the number of attempts that follows a Poisson distribution with parameter λ:

$$N \sim \mathcal{P}(\lambda),\ \lambda > 0,$$

and the **pmf** given by (2.3) and then let X be the outcome of each attempt to turn into success if $X = 1$ with a probability p and failure if $X = 0$ with probability $1 - p$, which follows a Bernoulli distribution as

$$X \sim Bernoulli(p),$$

and its **pmf** is given by (2.1) where $0 < p < 1$. Now let $X = \sum_{i=1}^{N} X_i$ be the number of successes in each attempt. If X_i's are assumed to be mutually independent, then the conditional probability function of X given $N = n$ is modeled as

$$X \mid N \sim Bin(n, p)$$

with a **pmf** given by (2.2).

Thus, the conditional mean and variance are given by $E\left(X \mid N = n\right) = np$ and $Var(X \mid N = n) = np(1 - p)$. Therefore, the joint distribution of the total number of attempts N and of the total number of successes X shown from equations (2.3) and (2.2) is written as,

$$\begin{aligned} \mathrm{P}(N = n, X = x) &= \mathrm{P}(X \mid N = n)\,\mathrm{P}(N = n) \\ &= \frac{\lambda^n p^x (1-p)^{n-x}}{x!(n-x)!} \exp(-\lambda), \end{aligned} \tag{2.5}$$

for $n = 0, 1, \ldots$ and $x = 0, 1, \ldots, n$. From the joint function given by equation (2.5), the **pmf** of the number of successes is given by

$$P(X = x) = \sum_{n=x}^{\infty} P(X = x|N = n)P(N = n),$$

which is written as,

$$\begin{aligned} P(X=x) &= \sum_{n=x}^{\infty} \binom{n}{x} p^x (1-p)^{n-x} \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \sum_{m=0}^{\infty} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \sum_{m=0}^{\infty} \frac{p^x (1-p)^{n-x} \lambda^n e^{-\lambda}}{x!(n-x)!}, \end{aligned}$$

and by setting $m = n - x \Leftrightarrow n = m + x$ we get,

$$P(X=x) = \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{n=x}^{\infty} \frac{\left[\lambda(1-p)\right]^m}{m!},$$

where,

$$\sum_{n=x}^{\infty} \frac{\left[\lambda(1-p)\right]^m}{m!} = e^{\lambda(1-p)},$$

we finally get,

$$P(X=x) = \frac{(\lambda p)^x e^{-\lambda p}}{x!}, \ x = 0, 1, 2, \ldots, \tag{2.6}$$

For which the mean and variance are the same and given by

$$\mathbb{E}(X) = Var(X) = \lambda p \tag{2.7}$$

The probability function in equation (2.6) is derived to be a Poisson Distribution with parameter λp. Under the proposal above, if the total number of shots or shots on target are assumed to be Poisson with mean λ, then the total number of goals scored is again Poisson, with mean λp, which is the product of the mean of the total number of shots or shots on target and the probability of that shot being converted into a goal. This seems reasonable; thus, the mean of goals scored is a proportion of the total number of goals scored. This approach allows us to model two different things at the same time: quantity and quality.

2.7 Skellam

Another method for modelling football data is to model the final goal difference of each match directly (see, for example, Karlis and Ntzoufras (2006) and Karlis et al. (2024)). The score difference is always a positive or negative integer, which is precisely an element of $\mathbb{Z}$. One of the most established distributions in $\mathbb{Z}$ is the Skellam distribution. In Irwin (1937), the probability mass function of the differences of two independent variables following the same Poisson distribution was derived. Later Skellam (1946) extended to different Poisson populations, precisely meaning different λ parameters.

Let the random variables X_1 and X_2 follow independent Poisson distributions with parameters λ_1 and λ_2, respectively. Then the random variable $Z = X_1 - X_2$ has a pmf

given by,

$$P(Z = z \mid \lambda_1, \lambda_2) = e^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_{|z|}\left(2\sqrt{\lambda_1\lambda_2}\right), \quad z \in \mathbb{Z},\ \lambda_1, \lambda_2 > 0, \tag{2.8}$$

Where $I_r(x)$ is the modified Bessel function of order r as defined in Abramowitz and Stegun (1974),

$$I_r(x) = \left(\frac{x}{2}\right)^r \sum_{m=0}^{\infty} \frac{\left(\frac{x^2}{4}\right)^m}{m!\,\Gamma(r+m+1)}.$$

This distribution will be denoted as the Skellam(λ_1, λ_2) distribution. For the particular case, where $\lambda_1 = \lambda_2 = \lambda$ Skellam (1946) confirmed Irwin (1937) that,

$$P(Z = z \mid \lambda) = e^{-2\lambda} I_{|z|}(2\lambda), \quad z \in \mathbb{Z},\ \lambda > 0.$$

Later, Karlis and Ntzoufras (2006) generalized the derivation of the Skellam distribution to differences of distributions other than Poisson. Also, they stated that the odd cumulants are equal to $\mu = \lambda_1 - \lambda_2$ while the even cumulants are equal to $\sigma^2 = \lambda_1 + \lambda_2$. It is also highlighted that for large values of $\mu = \lambda_1 + \lambda_2$, it can be sufficiently approximated by the normal distribution. While if λ_2 tends to zero, the distribution tends to a Poisson distribution. Further findings for the Skellam and its various applications can be found in Tomy and Veena (2022).

Skellam's distribution mean and variance are equal to $\mathbb{E}(Z) = \lambda_1 - \lambda_2$ and $\mathrm{Var}(Z) = \lambda_1 + \lambda_2$ respectively. So a reparametrised version of the distribution with mean equal to $\mu = \lambda_1 + \lambda_2$ and variance equal to $\sigma^2 = \lambda_1 + \lambda_2$ is applied in Koopman et al. (2017) to study intraday stochastic volatility, allowing for a clearer interpretation. Karlis et al. (2024) denoted this distribution as Skellam2(μ, σ) and was used to model handball outcomes. For a broad review of Skellam and related distributions, see Karlis and Mamode Khan (2023).

2.8 Bivariate Poisson

Another extension of the regular Poisson is the Bivariate Poisson Distributions introduced by Karlis and Ntzoufras (2003). Consider the random variables X_k, $k = 1, 2, 3$, which follow independent distributions with parameters λ_k, respectively. Then the random variables $X = X_1 + X_3$ and $Y = X_2 + X_3$ follow jointly a bivariate Poisson distribution denoted as $BP(\lambda_1, \lambda_2, \lambda_3)$, with joint probability function

$$\begin{aligned} f_{BP}(x, y \mid \lambda_1, \lambda_2, \lambda_3) &= P_{X,Y}(x, y) = P(X = x, Y = y) \\ &= e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^x}{x!} \frac{\lambda_2}{y!} \sum_{i=0}^{min(x,y)} \binom{x}{i} \binom{y}{i} i! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^i \end{aligned} \tag{2.9}$$

This bivariate distribution formulation accommodates positive dependence between the two random variables. Marginally, each random variable follows a Poisson distribution with $\mathbb{E}[X] = \lambda_1 + \lambda_2$ and $\mathbb{E}[Y] = \lambda_1 + \lambda_2$. Moreover, $Cov[X, Y] = \lambda_3$, and hence λ_3

measures the dependence between the two random variables. If $\lambda_3 = 0$, then the two random variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions(referred to as the double Poisson distribution).

2.9 Zero Inflation

In many instances, and not only in sports or football (see, for example, Lambert (1992), Karlis and Ntzoufras (2003)), zero is observed more frequently than what the assumed distribution predicts. For that matter, an alternative distribution may be considered to account for the shift required. Let's consider a Poisson random variable X with parameter λ, which is the Zero-Inflated distribution, which accounts for excess zeros by introducing a two-part process, and has a pmf written as

$$P_X(X = k \mid \lambda, p) = \begin{cases} p + (1-p)\,\mathrm{P}(X = k \mid \lambda, \theta), & k = 0, \\ (1-p)\,\mathrm{P}(X = k \mid \lambda, \theta) & k \neq 0 \end{cases} \tag{2.10}$$

In the same essence, if a random variable Z follows a Skellam distribution with parameters λ_1 and λ_2, then the pmf of the Zero-Inflated Skellam distribution is given by

$$P_Z(Z = z \mid \theta_1, \theta_2, p) = \begin{cases} p + (1-p)\,P(Z = z \mid \theta_1, \theta_2), & \text{if } z = 0, \\ (1-p)\,P(Z = z \mid \theta_1, \theta_2), & \text{if } z \neq 0, \end{cases} \tag{2.11}$$

with $p \in [0, 1)$ and $P(\cdot)$, being the regular pmf of each distribution. The Zero-Inflated version of any distribution shifts some probability from the other values towards zero.

2.10 Copulas

Copulas provide a way to separate the dependence structure from a multivariate distribution. In particular, as shown in Nelsen (2010), any multivariate distribution can be built by specifying its marginal distributions and an appropriate copula. In essence, copulas are functions that link or "couple" a multivariate distribution function with its one-dimensional marginal distribution functions. They allow us to disentangle the marginal behavior from the dependence pattern of a multivariate distribution, and they are especially useful for revealing and clarifying the limitations and misconceptions associated with correlation.

More formally, a copula is a multivariate distribution with all univariate marginal distributions being uniformly distributed on the unit interval, $[0, 1]$; hence, C is the distribution function of a multivariate uniform random vector. For a bivariate distribution F with margins F_1 and F_2 and some parameter θ, the copula associated with F is a distribution function $C : [0, 1]^2 \to [0, 1]$ that satisfies

$$F(x, y; \theta) = C\{F_1(x), F_2(y); \theta\}, \quad (x, y) \in \mathbb{R}^2 \tag{2.12}$$

The copula C is uniquely determined on the unit square whenever F_1 and F_2 are continuous. The copula itself characterises the dependence between the random variables X and Y with marginal distributions F_1 and F_2. Thus, the copula representation 2.12 resolves the joint distribution into the marginals F_1 and F_2 and the dependence structure C (McHale and Scarf (2011)).

In the discrete framework that primarily concerns us when modelling football data, the joint pmf is derived by taking differences. For instance, as illustrated in Karlis et al. (2024), in the bivariate case with marginal distributions $F_1(x_1)$ and $F_2(x_2)$, the joint pmf can be derived as

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2) = C(F_1(x_1), F_2(x_2); \theta) - C(F_1(x_1 - 1), F_2(x_2); \theta) \\ - \; C(F_1(x_1), F_2(x_2 - 1); \theta) + C(F_1(x_1 - 1), F_2(x_2 - 1); \theta) \end{aligned} \tag{2.13}$$

We are going to use marginal distributions that are linked through the bivariate Frank copula, which is given by

$$C(u, v; \theta) = -\frac{1}{\theta} \log \left[1 + \frac{\left(e^{\theta u} - 1\right)\left(e^{\theta v} - 1\right)}{e^{\theta} - 1} \right] \tag{2.14}$$

We select this particular copula to be as flexible as necessary since it allows for both negative and positive correlation. That way, we have the freedom to explore the actual relationship between the two random variables without imposing restrictions. However, of course, any other copula could also be used.

In the multivariate case, where the number of variables p is greater than 2, a very common choice of copula is the Gaussian copula. Now let $\mathbf{X} \sim \mathcal{MN}_d(\mathbf{0}, \mathbf{R})$, where $\mathbf{R}$ is the correlation matrix of $\mathbf{X}$. Then the corresponding Gaussian copula is defined as

$$C_{\mathbf{P}}^{\text{Gauss}}(\mathbf{u}) := \Phi_{\mathbf{R}}\big(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\big)\,, \tag{2.15}$$

where $\Phi(\cdot)$ denotes the standard univariate normal cumulative distribution function and $\Phi_{\mathbf{P}}(\cdot)$ denotes the joint cumulative distribution function of $\mathbf{X}$.

Chapter 3

Modelling

When building a model, essentially constructing an equation that describes a phenomenon, there are two main, often overlapping, objectives. Those are **Inference** and **Prediction**

- **Inference**, is when we wish to examine if certain hypotheses, assumed, are true. Additionally, to check whether specific variables are related to the response variable, to measure that relation, and to account for the uncertainty surrounding this relation and the phenomenon under investigation as a whole.
- **Prediction**, is when we wish to tell what is going to happen in the future. In statistical language, this translates to building a model that can be used to make estimations based new data or observations not yet seen.

These two approaches to the seemingly same procedure differ substantially from one another. Inference is always based on a correctly specified model. Otherwise, significant bias might be included in the coefficients, standard errors might be completely inaccurate, and worst of all, interpretation and conclusions might be completely wrong. Whereas, for prediction, the truth is that a misspecified model is not an obstacle to using it if its accuracy is satisfactory. This idea is well captured by the well-known aphorism that *all models are wrong, but some are useful*, emphasizing that the value of a predictive model lies not in its literal correctness, but in its ability to make reliable predictions.

A major consideration in predictive modelling is the availability of explanatory variables, prior to the event of interest. For example, ball possession may help explain the outcome of a football match retrospectively, but it cannot be used for prediction since it is not known before the game.

Variable selection is a related, but not identical, issue. In particular, it concerns choosing among non-nested models, meaning models where one cannot be obtained from another simply by omitting terms or fixing certain parameters to specific values. For example, Poisson versus negative binomial regression (a case we examine further below), or choosing between models that use different sets of explanatory variables.

Typically, **Akaike Information Criterion (AIC)** or **Mean Squared Error (MSE)** or related information-based measures are used for model selection.Remember that, in general, simpler models are preferable, in line with Occam's Razor. Interpretability is equally important, especially in applied settings.

3.1 The Main Idea

When we go under the endeavour of developing a model, essentially, we are interested in understanding, explaining, or more generally exploring the bivariate relationship between two variables, a variable Y and a variable X. Specifically, our objective is to succeed in attributing or capturing part of the variation in the variable of interest, usually denoted as Y, based on the knowledge we have of another variable X.

The variable Y (whose variability we are interested in controlling) is called the **response** or **dependent** variable, while the variable X (which is used to explain Y) will be called **predictor** or **independent** variable. From a statistical point of view, our interest is in studying the conditional distribution of $Y \mid X$ (Y given X), so that we will not only be able to understand the way Y varies for different values of $X = x$, but we will also be able to provide predictions of Y for given (or known) values of X.

However, the joint distribution of (X, Y) will not always be available. For this reason, we will attempt to at least model the expected value of $Y \mid X$, i.e, the $\mathbb{E}[Y \mid X]$. Most of the times a model cannot perfectly predict the outcomes; there is always going to be a margin of error that we will try to minimize each time. We generally prefer to make assumptions about the produced error to describe the general behaviour of the phenomenon that we study.

The simplest idea in modelling is to try fitting, or more simply put, applying a line to our data each time. The equivalent assumption is that

$$Y_i \mid X_i = x_i \sim N(a + \beta x_i, \sigma^2),$$

for all $i = 1, \ldots, n$ and the observations are independent. The main idea is that we fit a linear function in the mean of a random variable Y for which the uncertainty is described via a normal distribution. This is a really straightforward idea, which, many times in real-world applications, variables exhibit a more complicated relationship than a simple linear one. As a result, we wish to generalize this to more complex relationships and to variables that may have constraints.

The rationale is that the linear model is based on some assumption of normality, which implicitly assumes data on $(-\infty, \infty) = \mathbb{R}$. What if the data are discrete/counts or binary? For example, we wish to model the probability that player A will win over player B in a tennis match. As a probability, it shall be in the $(0, 1)$ interval on which occasion the response is a binary (win/loss) variable. Another one can be the time until the next goal; we have to consider that time cannot be negative. Additionally, the dynamics of a given match need to be taken into account. The number of goals/fouls committed or asses scored in a tennis match, all of them taking values $0, 1, \ldots = \mathbb{N}$. For that matter, a general framework needs to be established that is going to be equipped with the suitable tools to capture occasions such as the aforementioned ones and even more.

Generally, modelling in sports analytics plays an important role across a wide range of applications. For instance, projecting how players will perform by estimating different metrics that reflect a player's skill or overall performance. Another important issue for teams is the fatigue level of players, which can be evaluated using physiological indicators that capture the impact of training programs, match involvement, and congested schedules.

3.2 Generalized Linear Models

Generalized Linear Models extend the rather simple but robust idea of fitting a straight line to describe and explain the data or phenomenon that we work with each time, that is, the linear model, to other types of data. The main idea is that we assume that the response Y follows some suitable distribution, say $f(\cdot)$ (most models belong to the so-called exponential family, but the theory can be extended beyond this), with its mean being equal to

$$\mathbb{E}[Y_i] = \mu_i$$

and the linear predictor being

$$g(\mu_i) = \eta_i$$

for some function $g(\cdot) = \eta_i$. And the regression part is

$$\eta_i = \sum x_{ij}\beta_j$$

Based on the form of the data, we may assume a suitable probability model (not necessarily normal) that allows us to account for the uncertainty that we deal with. A suitable choice of $g(\cdot)$ ensures that the mean remains in the admissible range. A linear model is a special case of the normal assumption. So, we can fit models to a large range of datasets, experiments, and phenomena. Generalised linear models are made up of 3 components.

- **Random component**: Identifies the dependent variable Y and its probability distribution.
- **Systematic Component**: Identifies the set of explanatory variables $(X_2, \ldots, X_k)$.
- **Link Function**: Identifies a function of the mean that is a linear function of the explanatory variables

3.3 Adding Covariates

In order to introduce categorical predictors (covariates), we need to use dummy variables. These are one-hot encoded variables, which means that they consist of 0 or 1 indicators denoting the presence or absence of a given category. For example, if we wish to introduce teams as predictors X (covariates) in a model, we construct a vector containing all teams in a given league and represent each observation by a set of dummy variables, where each dummy takes the value 1 if the corresponding team is involved and 0 otherwise.

Simple and logical it might seem, but it hides many problems underneath. For a categorical variable with k levels, we need $k - 1$ dummies. There are several different ways to express the dummies, with the most common one being as binaries, which is what we described above. Dummy variables can be very useful to represent categorical variables and define more refined models if we must. The drawback is that they may increase the number of parameters to estimate by a lot and perhaps create over-fitting models (something that we will discuss a bit later).

In the upcoming sections, we describe the framework from which we apply the aforementioned distributions in Chapter 2 to construct suitable models for predicting final results. More specifically, our goal is to build models that **estimate** outcomes of football matches, and, at each step, we will also examine how to interpret the models' coefficients. We fit and evaluate a wide range of models, including both widely used approaches and several extensions of them, across the five major European leagues for the 2020–2025 seasons.

3.4 Double Poisson Model - the Classics

When attempting to model football team scoring behaviour, based on Chapter 2, the first distribution that should come to mind is the Poisson distribution. We assume that the number of goals scored by a particular team, denoted by X, follows a Poisson distribution with parameter λ, with probability mass function given by equation (2.3).

As described in Section 2.4, the parameter λ represents both the mean and the variance of a Poisson-distributed random variable and, in the present context, characterizes a team's goal-scoring intensity in a given match. We aim to specify λ to capture as much of the variation in scoring behaviour as possible. This, as we can imagine, can depend on many different factors. One might argue that football metrics—such as possession, tackles, or similar in-game statistics—should be incorporated to explain variability in match outcomes. However, as discussed in the introduction of Chapter 3, such quantities do not contain predictive value, since they are not known before the match. Likewise, another one could argue that factors such as the psychological state of individual players, weather conditions, the political state of each country, and many more factors in a never-ending list should be taken into account. In practice, however, these factors are difficult to measure reliably. Moreover, guided by the principle of Occam's razor, discussed earlier, increased model complexity does not necessarily lead to improved predictive performance.

On the other hand, the performance of each team, in general, is somewhat consistent, with ups and downs throughout the season, depending on their form. So, for simplicity, but not lacking proper modelling strength, we may depict each team's scoring ability as $\lambda_i, i = 1, 2, \ldots, n$, where n is the total number of matches in a season of a given league, as the result between the team's attacking ability and its opponent's defensive strength. Furthermore, these abilities, which formally are referred to as covariates (for details see section 3.3), are known to depend on whether a team is playing at home or away, as documented in Clarke and Norman (1995), which indicates that generally teams have a slight advantage when playing at home.

In the language of mathematics, the paragraph above translates to the equation

$$\log(\lambda_{i,k,j}) = \gamma + \delta \cdot home_j + \beta_{att_{team_k}} + \beta_{def_{opp_k}} \tag{3.1}$$

for $j = 0, 1$, $k = 1, 2, \ldots K$ where K is the total number of teams in the league's season, and $i = 1, 2, \ldots, N$ where N is the total number of matches. This might seem a bit odd at first glance, but we will decompose it step by step. In each football match, two teams compete against one another, which we index by $j = 0, 1$ for home and away ground respectively. Matches, over a league's season, are represented by $i = 1, 2, \ldots, n$ where N is the number of matches.

Conceptually, we have two different (**independent**) λ that represent each competing team's scoring ability. Now, that particular intensity depends on the team's attacking ability $\beta_{att_{team_k}}$ and the opponent's defensive ability $\beta_{def_{opp_k}}$. Additionally, we introduce a home advantage term δ, which is incorporated through the dummy variable $\text{home}_j = 1$ if team j plays at home and 0 otherwise.

In a simple linear regression setting, the **intercept** γ represents the expected value of the **response** when all **predictors** are equal to zero. In the present framework, however, the predictors can not be 0, as this would indicate that the match was happening between no teams! To ensure identifiability and facilitate reasonable interpretation of the model parameters, we set a *baseline* (**reference**) team against which the effects of all other teams are measured. The intercept γ represents the **log-expected** goal intensity of two equally strong teams competing against each other.

Why do we equate $\log(\lambda)$ and not just regular λ to that equation, one might wonder. The reason is that the right side of the equation (3.1) can be both negative and positive, essentially meaning that it takes values in $\mathbb{R}$. In contrast, the Poisson parameter λ can only be positive, i.e. $\lambda \in \mathbb{R}_+$. Incorporating the logarithmic link function, we ensure this constraint is satisfied, since exponentiating the linear predictor guarantees a positive value for λ.

The interpretation of the parameters on the original scale is as follows. The quantity e^{γ} represents the expected number of goals scored by the away team in a match between two teams of equal strength. Similarly, $e^{\gamma+\delta}$ represents the expected number of goals by the home team in a game between two teams of equal strength. And lastly, e^{δ} represents the relative increase of the expected home goals in a game between two teams of equal strength.

If we wish for a nicer and more general interpretation of the team's effects, we impose sum-to-zero constraints

$$\sum_{k=1}^{K} \beta_{att_{team_k}} = 0 \quad \text{and} \quad \sum_{k=1}^{K} \beta_{def_{opp_k}} = 0$$

for both the attacking and defensive abilities of the teams. In this way, the comparison is consistently made against a team with average attacking or defensive strength, rather than against one particular baseline team. Additionally, in practice, we usually set the sum-to-zero dummies with $\beta_{att_{team_1}} = -\sum_{k=2}^{K} \beta_{att_{team_k}}$ and $\beta_{def_{opp_1}} = -\sum_{k=2}^{K} \beta_{def_{opp_k}}$. As a result, positive values of attacking abilities correspond to better teams than average in terms of attack, and negative values of defensive abilities correspond to better teams than average in terms of defence, respectively. Additionally, $e^{\beta_{att_{team_k}}}$ represents the relative increase/decrease in expected goals in comparison to an average team, and similarly $e^{\beta_{def_{opp_k}}}$ represent the relative decrease/increase in expected goals of the opponent team in comparison to an average team.

In that model specification, we have assumed that the home effect is common for all teams, as discussed in Lee (1997), Karlis and Ntzoufras (2000), which might seem an exaggerated simplification, since different teams may benefit from playing at home to varying degrees. In the case where the home effect is not introduced as a common factor for all teams, one would need to estimate separate attacking and defensive abilities for each team, depending on whether it is playing at home or away. The model based on equation (3.1), however, offers a nice interpretation of teams' abilities, as we estimate each

team's both attacking and defensive strength, and then add some additional advantage if that team is playing at home. Below is an example of such a fit for La Liga in season 2024-2025.

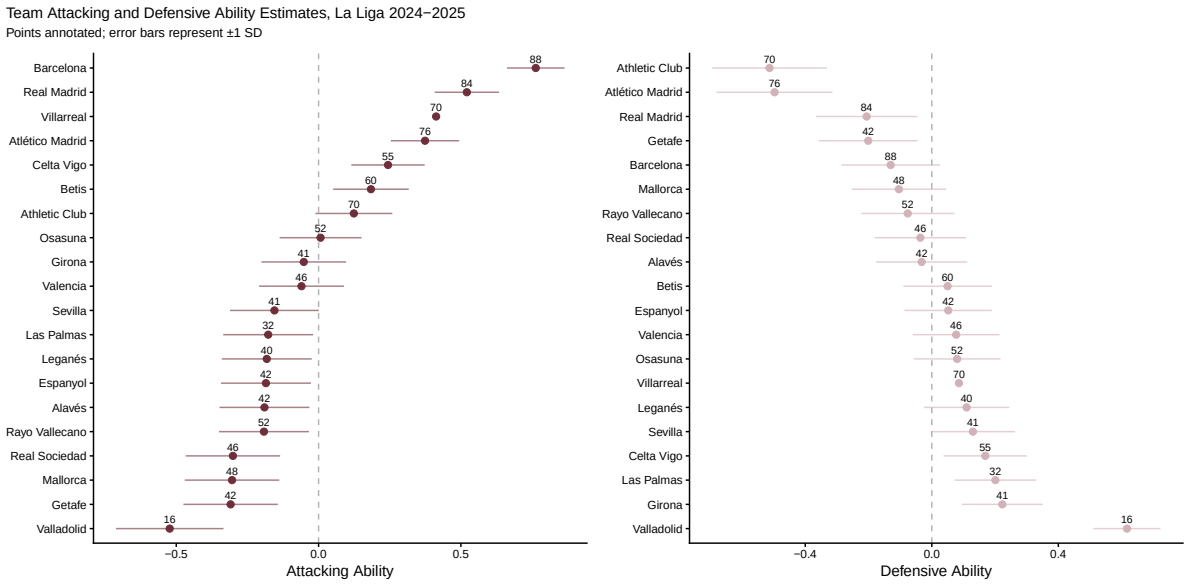


Figure 3.1: Estimates from a Poisson log-linear model with team-specific attacking and defensive strengths for La Liga 2024–2025, shown together with each team's actual points.

Barcelona, which won the league that season with 88 points, was estimated by the model to have the strongest attacking ability relative to the league average, while possessing only the fifth-strongest defensive ability, ranking just below Getafe, which finished 13$^{\text{th}}$ in the league with 42 points, as it had the 2$^{\text{th}}$ worst attacking performance. Both of these examples contradict the well-known aphorism that *attacks win matches, but defences win championships*. Atlético Madrid, a team well known for its strong physical style of playing and highly conservative tactical approach, has the 2$^{\text{nd}}$ strongest defensive ability according to the model, just below Athletic Club, with a really small margin of difference. At the bottom of both tables is placed Valladolid by quite a difference, indicating that among all teams, it simultaneously has both the weakest attacking and defensive abilities. This aligns with the observed outcome, as Valladolid finished last in the league and was relegated with only 16 points.

Such observations can be made for all teams, serving as a measure of comparison and evaluation of the performance they had throughout the season. There are times when the ranking or points of each individual team are heavily influenced by just pure luck or lack of effectiveness, and therefore do not allow for an objective conclusion of how well the play. Such models enable teams, analysts, head office staff, and any other related field to actually filter out noise and obtain a stable, comprehensive view of any team's performance, as well as to make meaningful comparisons with the other league's teams.

3.5 Poisson or Not Poisson

When modelling the number of goals scored by a team, a fundamental question is whether the Poisson distribution can accurately describe the 'true' underlying nature of goals. The construction of this model relies on a number of strong, silent assumptions. One of these is, for example, that each opposing scoring intensity λ follows an independent Poisson distribution. Essentially, meaning that we separately model each team's offence against the opponent's defence without taking any consideration of what is happening in between them. To make an illustration is like cutting the football pitch in half and looking only at what happens in each goalpost.

It has also been observed (see Lee (1997), Dixon and Coles (2002), Karlis and Ntzoufras (2003), Karlis et al. (2024)) that draws occur more frequently than predicted by the standard double Poisson models. Additionally, as discussed in Section 2.4, the Poisson distribution's λ parameter, which represents both the mean and variance, often leads to a phenomenon known as overdispersion—where the variance exceeds the mean (see Karlis and Ntzoufras (2000)).

All the assumptions concerning the Poisson distribution and effectively any distribution have to be evaluated and empirically verified for the constructed model to be valid. Otherwise, as discussed in the introduction to Chapter 3, various problems will arise.

3.5.1 Independence

One of the main questions that naturally arises in modelling soccer games is whether the number of goals scored by the two opponents is independent. For each championship out of the 5 major European leagues studied, a χ^2 test of independence in a contingency table was performed to examine possible dependencies. The table below displays the empirical correlation calculation for each league alongside the resulting p-value from the χ^2 independence test.

League	**Correlation**	**p-value**
Premier League	-0.131	0.001
Serie A	-0.085	0.002
La Liga	0.004	0.598
Bundesliga	-0.132	≤ 0.001
Ligue 1	-0.087	0.087

Table 3.1: Empirical correlations between average home and away goals in 5 major European leagues, with accompanying χ^2 test p-values for independence, during the 2025–2026 season.

In 4 out of the 5 major European championships, the assumption of independence is rejected, with the only exception being the Spanish championship, La Liga, in which the null hypothesis of independence was not rejected, with a correlation of just 0.004. We should also note that, given the very large sample size (five seasons per league, i.e., $5 \cdot 380 = 1900$ matches), the rejection of the null hypothesis in each league may partly reflect a sample-size effect rather than a practically meaningful deviation from the null.

The just slight negative correlation in each league is indicative of the aforementioned remark above. The exact nature of independence is more thoroughly examined further below.

The following table displays the observed joint and marginal frequencies of home and away goal counts in La Liga matches for the 2020-2025 seasons. This is the contingency table upon which we perform the χ^2 independence test (different in each corresponding league).

Home Goals \ Away Goals	0	1	2	3	4+	**Total**
0	167	166	99	37	15	484
1	243	245	125	36	20	669
2	131	167	91	28	15	432
3	71	60	37	17	6	191
4+	40	45	27	11	1	124
Total	652	683	379	129	57	1900

Table 3.2: Observed joint and marginal frequencies of home and away goal counts in La Liga matches (2020–2025 seasons).

To further investigate whether a dependence exists between two competing teams, we begin by examining the marginal frequencies, as these reflect the actual distribution of home and away teams. Across the 2020–2025 seasons, the typical home team in La Liga enjoys a slight edge over the typical away team, a trend that likewise appears consistently in all leagues. This stems from the fact that the average home team's distribution is slightly "shifted" towards higher scores compared to the away team's, which is better visualized with the barplot below,

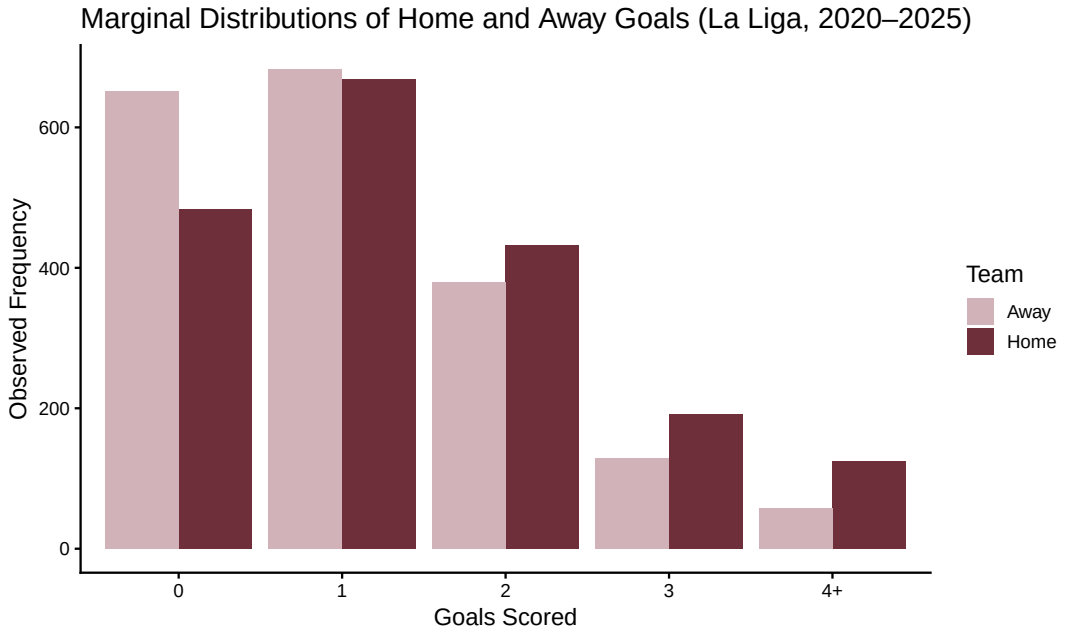


Figure 3.2: Marginal Distribution of average home and away team goal counts in La Liga for the 2020-2025 seasons.

The joint distribution is difficult to interpret when presented in tabular form, as illustrated in Table 3.2. To provide a clearer and more intuitive perspective, the plot below is presented, showing the expected number of goals—under the assumption of independence—for the average home and away teams in La Liga across the 2020–2025 seasons.

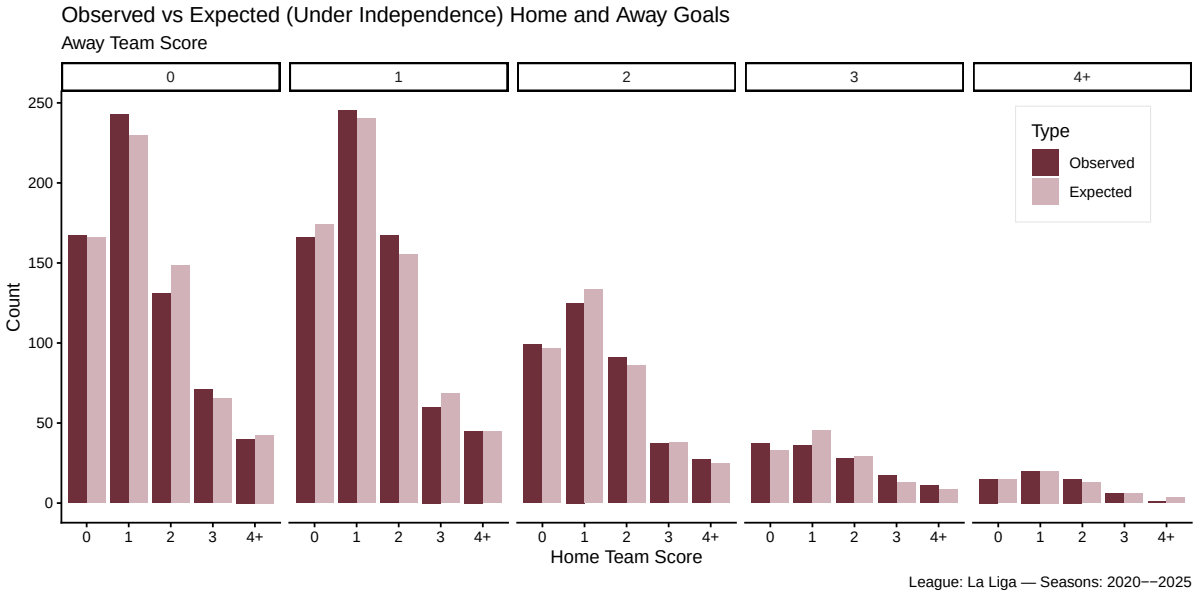


Figure 3.3: Comparison of observed and expected joint distribution of home and away goals in La Liga matches during the 2020–2025 seasons, where expected counts are derived under the assumption of independence.

The main takeaway is that the goal-scoring patterns do not change significantly throughout the different scores of each team in the Spanish League for the 2020-2025 seasons. This is probably clearer when comparing with figure B.1 in the appendix, which is the same plot for each of the remaining five major European leagues for which the independence hypothesis between the average home team and the average away team was rejected.

3.5.2 Draws

It has been widely observed that the number of draws is more likely to be empirically observed than supposed by the regular double Poisson model (see, for example, Dixon and Coles (2002), Karlis and Ntzoufras (2006)). So again, we need to investigate whether any abnormality is exhibited that would result in a significant deviation from the Poisson distribution.

In this case, we once more compute the mean number of goals for both home and away teams. Each mean serves as an estimate of the Poisson parameter λ and, in particular, corresponds to the **Maximum Likelihood Estimation (MLE)** (the $\bar{x}$) of this parameter for the average home and average away team, respectively. We then repeatedly simulate (with more repetitions being preferable) matches between the representative home team and the representative away team. This procedure is a *Monte Carlo* simulation, and its

outcomes allow us to calculate (or approximate) the probabilities of a home win, draw, or away win that the Poisson model predicts for games between the average home and away team in La Liga over the 2020–2025 seasons. More importantly, it enables us to estimate how probable the actually observed total number of draws (across all teams) is when compared with what would be expected from matches between these average home and away teams in La Liga during the 2020–2025 period. Below is the table that lists the observed and expected number of matches resulting in a draw under Poisson scoring for the average home and the average away team for each one of the big 5 European championships for the 2020-2025 seasons.

League	**Observed**	**Expected**	***p*-value**
Premier League	441	479.07	0.973
Serie A	528	494.85	0.037
La Liga	521	508.63	0.253
Bundesliga	386	364.71	0.076
Ligue 1	439	449.16	0.687

Table 3.3: Observed and expected number of matches resulting in a draw under Poisson scoring for the average home and the average away team for each one of the 5 Major European leagues, together with Monte Carlo p-values, for the 2020-2025 seasons.

We observe that in both the Premier League and Ligue 1, the observed number of draws was actually less, resulting in high p-values. In La Liga, despite the fewer observed draws, its p-value does not signal great deviance. Lastly, both Serie A and Bundesliga had fewer observed draws, and the p-value was small enough to indicate a statistically significant difference from what the Poisson distribution predicts between the average home and away teams. Below is the histogram of the Monte Carlo simulation illustrating the distribution of draws for the match between the average home and the average away team in La Liga for the 2020-2025 seasons. Accordingly, the same approach was applied to the remaining leagues, as can be seen in Figure B.2 of the appendix.

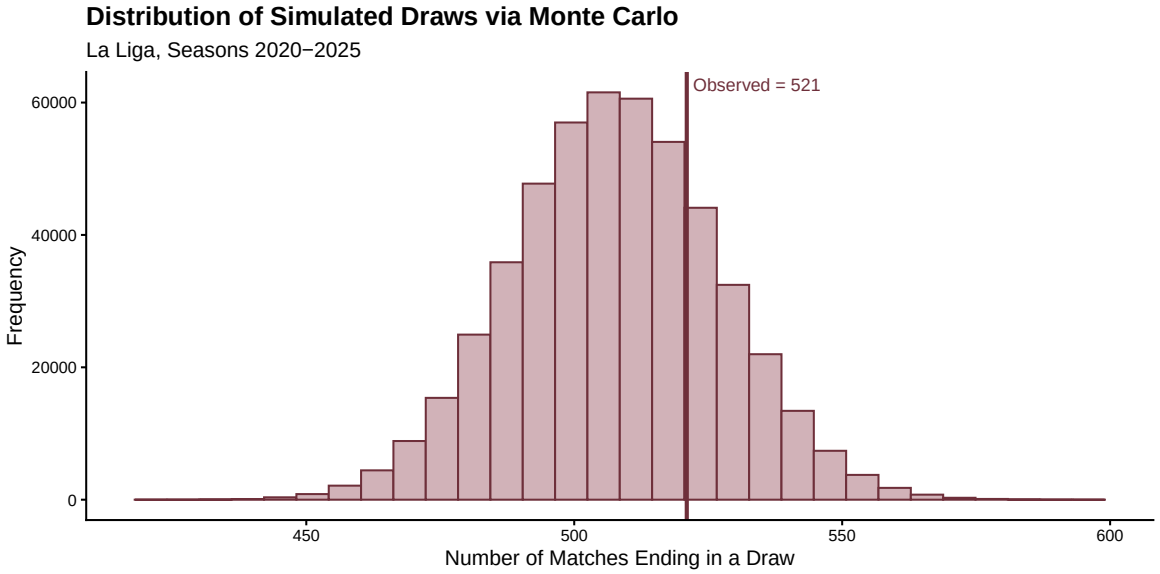


Figure 3.4: Histogram of Monte Carlo Simulated draws between the average home and average away team illustrating its Distribution for La Liga during the 2020-2025 seasons.

3.5.3 Overdispersion

The Poisson distribution has a formal theoretical basis and assumes that the mean is equal to the variance, denoted by the parameter λ, as was discussed in section 2.4. It is well documented that when applying the Poisson distribution to model various phenomena, this assumption is often violated. Overdispersion is the case when the variance is greater than the mean (see, for example, Karlis and Ntzoufras (2000), and less frequently occurring, under-dispersion is the case when the variance is less than the mean as is discussed in Karlis et al. (2024)).

In this case, as in the previous subsections, we compute the empirical mean and variance of goals scored for each league over the 2020–2025 seasons. As before, the estimated mean serves as the maximum likelihood estimator (MLE) for the scoring intensity of the average team in each of the five major European leagues, under the assumption of a Poisson distribution.

We then perform repeated simulations of the scoring performance of this estimated average team under the Poisson model, using several simulated observations equal to the total number of team–game appearances in a season. Since each match involves two teams, this corresponds to twice the number of league fixtures (e.g., 380 × 2 observations for a 380-match season). After each simulated season we calculate the variance of the goals scored. That way we get an estimation of what the expected variance of goal scoring would be if the average team was playing for each league in the seasons 2020-2025. Additionally, we can calculate how likely the actual, observed variance is compared to the simulated one which is the p-value.

Below the Table 3.4 contains the lower 2.5% quantile, the mean: 50% quantile, the upper 97.5% quantile, of the simulated goal scoring variance resulting from the average team, along with the total observed variance and the simulated p-value, for each one of the 5 major European championships.

League	2.5%	50%	97.5%	s^2	**p-value**
Premier League	1.333	1.405	1.480	1.61	≤ 0.001
Serie A	1.255	1.323	1.395	1.37	0.094
La Liga	1.177	1.242	1.310	1.34	0.002
Bundesliga	1.439	1.525	1.615	1.77	≤ 0.001
Ligue 1	1.289	1.361	1.438	1.46	0.008

Table 3.4: Monte Carlo simulation results for average Poisson variance across major European leagues (Seasons 2020-2025).

The results presented in the table indicate the presence of overdispersion in all leagues, as evidenced by the extremely small p-values. The only exception is the Italian championship, which resulted in a slightly higher p-value, but not so as to feel confident enough that there is no problem of overdispersion at all.

Below, in Figure 3.5 is an illustration of the Monte Carlo simulation of the average team's scoring variance (or dispersion) for the Spanish Championship in seasons 2020-2025. Figure B.3 displays the same plot for the rest of the major European leagues.

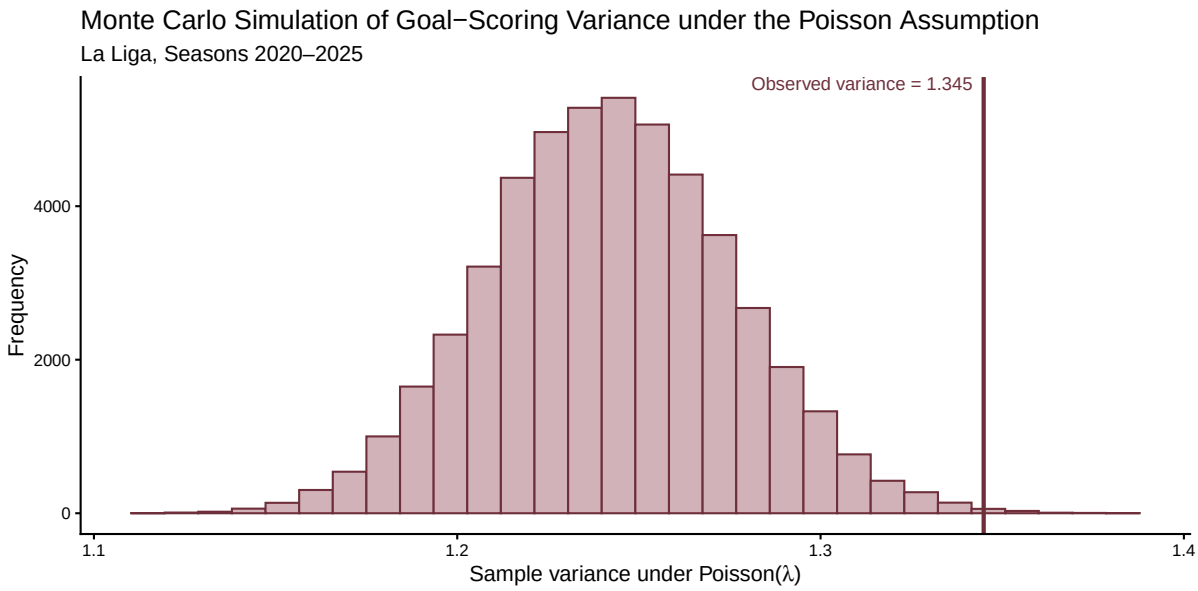


Figure 3.5: Monte Carlo Simulation of the average team's scoring variance for La Liga in 2020-2025 seasons

We have to note that these results, which indicate such high overdispersion across all 5 major European leagues, may be just artifacts. That is why we calculated the mean and the variance of goals scored by all the teams that played in each league during the 2020-2025 seasons. This alone has great dispersion, because as the Figure 3.1 below indicates, attacking and defensive abilities differ from team to team. As a result, from the simulation we did, we actually measured the deviation among the team's scored goals and not the actual overdispersion in the Poisson distribution.

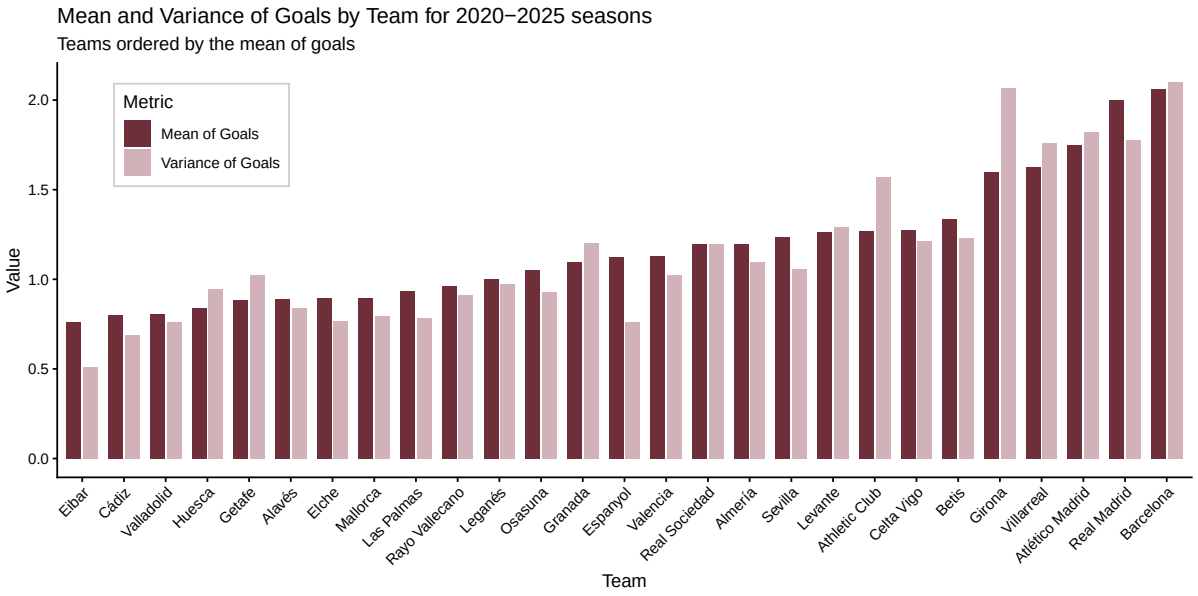


Figure 3.6: Mean and Variance of goals scored by each team, in La Liga during the 2020-2025 seasons.

A more thorough investigation is needed, actually, to assess the level of overdispersion present across all leagues. As a first step, we plot the mean and variance of goals scored by each team for all 5 major European championships during the seasons 2020-2025, as can be seen just above in Figure 3.6 for the Spanish League and in Figure B.4 of the Appendix for the rest of the European Leagues.

We observe that teams differ significantly from one another in terms of the mean goals they scored throughout the seasons. This is precisely the overdispersion that is detected by the tests and metrics we displayed above. Looking closer, however, we detect that each team does not have such a greater variance compared to the mean goals scored. As the Figure 3.6 and the Figure B.4 in the appendix verify, there are many cases where the variance is even less than the mean of goals scored by a team, see for example, Espanyol and Elbar.

To make things even clearer, below we plot with a darker shade the calculated mean goals on the x-axis and the corresponding calculated variance on the y-axis for each team during the seasons 2020-2025. Additionally, we plot the corresponding simulated variance given the mean of each team. Specifically, we estimate the mean of each team, and then simulate equally played matches to measure the variance produced. Moreover, we have added a 95% Confidence Interval for displaying the expected deviation of the variance, under the Poisson regime. The Straight line indicates the Poisson distribution. The Figure 3.6 below displays the results for the Spanish championship for the 2020-2025 seasons, and Figure B.4 in the Appendix displays the results for the rest of the leagues during the same time.

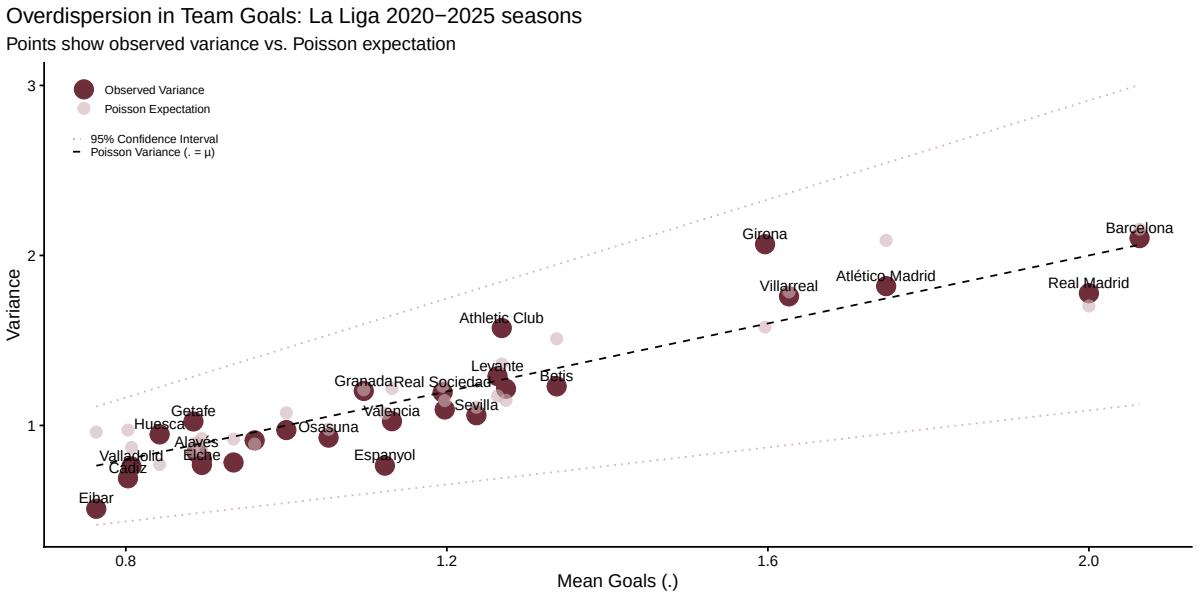


Figure 3.7: Scatter plot of Mean and Variance goal scoring. Each point represents a team; the darker ones are the observed and the lighter ones are simulated. The dashed line represents the Poisson expectation. Data are from La Liga during the 2020-2025 seasons.

All these plots we have seen suggest that there is no evidence of overdispersion in any of the leagues during the 2020–2025 seasons. As the observed image of each team's mean and variance has no significant difference from the simulated one. Generally, low-scoring teams are evenly distributed around the dark dashed line (which indicates the Poisson

expectation), and as the mean goals increase, teams tend to be placed above this line but without exceeding the upper limit set by the 95% confidence interval.

To properly address the problem of overdispersion, and actually be sure whether it is present in leagues, we model each league's season with the negative binomial distribution, which was discussed in 2.5. The model formulation slightly changes to the double Poisson model as was described in section 3.4 and according to Venables and Ripley (1994) can be written as

$$X_{ijk} \sim \mathcal{P}(\epsilon_{ijk}\lambda_{ijk}), \quad \epsilon_{ijk} \sim \Gamma(r, r)$$

where $\Gamma(\alpha, \beta)$ denotes the **Gamma** Distribution with mean α/β and variance α/β. The parameter r controls the overdispersion since we now have $\mathbb{E}(X_{ijk}) = \lambda_{ijk}$ and $Var[X] = \lambda_{ijk} + \lambda_{ijk}^2/r$. That way, we can actually test whether the Poisson assumption is violated by accounting for the different team abilities and the parameter r (the dispersion parameter) simultaneously. Large values of r imply low over-dispersion as $r \to \infty$ and $\lambda_{ijk}^2/r \to 0$, which means that the variance reduces to a Poisson distribution variance.

We fit the double Poisson model using the glm R function and the double negative binomial model using the glm.nb R function which belongs to the MASS library. In the table below, we display the model fitted each time, the number of parameters as Df, the log-likelihood as LogLik, the fitted probabilities of home win, draw, and away win, the home effect in exponential scale as μ, and the dispersion parameter as r.

La Liga

Season	**Model**	**Df**	**LogLik**	**Fit**			**Parameters**	
				Home	Draw	Away	μ	r
	Actual			155.00	117.00	108.00		
2020–2021	Double Poisson	40	-1009.28	160.89	96.60	122.51	1.22	
	Negative Binomial	40	-1009.28	160.90	96.50	122.60	1.22	17838.48
	Actual			167.00	105.00	108.00		
2021–2022	Double Poisson	40	-1041.37	167.89	97.41	114.71	1.31	
	Negative Binomial	40	-1041.38	167.96	97.30	114.75	1.31	9667.63
	Actual			180.00	94.00	106.00		
2022–2023	Double Poisson	40	-1001.57	173.67	96.16	110.17	1.39	
	Negative Binomial	40	-1001.58	173.72	96.16	110.12	1.39	17306.50
	Actual			167.00	103.00	110.00		
2023–2024	Double Poisson	40	-1037.39	165.52	94.57	119.90	1.26	
	Negative Binomial	40	-1037.39	165.61	94.60	119.79	1.26	15163.41
	Actual			163.00	102.00	115.00		
2024–2025	Double Poisson	40	-1011.51	163.15	94.41	122.44	1.23	
	Negative Binomial	40	-1011.51	163.03	94.58	122.39	1.23	20387.04

Table 3.5: Observed and expected numbers of home wins, draws, and away wins under double Poisson and negative binomial specifications across 2020-2025 seasons for La Liga.

We observe that for each separated season, the estimated r is so large that it essentially makes the model reduce to the double Poisson model. This is verified as both models yield almost equal log-likelihood (see Value column) and estimated fitting of

outcomes (see Fit column).

Based on these results and all the plots examined above, we can reasonably conclude that there is no overdispersion in La Liga during the 2020–2025 seasons. The same conclusion may be made for the rest of the leagues that we inspected, see Tables C.1, C.2, C.3, C.4, along with the corresponding Figures B.4 and B.5 of the Appendix.

When we tried to draw conclusions based on the average team scoring behaviour of each league, we received strong indications of overdispersion for every one of them. By looking a little closer, however, we determined that there is no such case. This is a straightforward, yet illustrative, example of the effectiveness of modelling and, by extension, statistical analysis. Just by incorporating covariates (effectively taking into consideration that teams have different abilities), we flipped the image that we first had.

Based on that, one could argue that the same false conclusions were drawn when performing the independence and draw inflation tests, and actually would be right. These assumptions, however, do not have such a straightforward and general solution, but nevertheless will be discussed in later sections of this chapter.

3.6 What is actually the home effect?

In sports analytics, a standard approach when modelling team sports is to incorporate a home effect. This effect actually captures the universal observation that has been documented throughout many sports (see, for example, Lee (1997), Karlis et al. (2024)) that usually teams perform better when playing at home. What factors come into play that affect teams in that way? Is it the social pressure that is being put on both players and the referee? (see for example, Boyko et al. (2007)) Is it that players who play at home know the pitch much better and are used to the same conditions? All these and many more? We could come up with speculative possibilities endlessly, without ever reaching a final list.

Home advantage is well established in various data-based studies, but also in more informal discussions among friends and football fans. It is documented that the home effect is stronger in 2^{nd} division leagues compared to 1^{st} division and that the same home effect is lower in derbies (see, for example, Leite and Pollard (2018)). In the study of Pollard (1986) was found that this effect was relatively stable in the English championship since 1888 and that roughly 64% of the points were collected from home teams for 1970-1981. Based on the data currently available, the proportions of outcomes are shown below.

League	Home	Draw	Away
Premier League	0.43	0.23	0.34
Serie A	0.40	0.28	0.32
La Liga	0.44	0.27	0.29
Bundesliga	0.44	0.25	0.31
Ligue 1	0.42	0.25	0.33

Table 3.6: Empirical probabilities (observed proportions) of home wins, draws, and away wins across major European leagues (seasons 2020–2025).

We observe that, generally, the average home team wins more frequently than the average away team. This observation, however, resembles a lot the assumptions that we tested above. We measure the home advantage by calculating the overall outcome results of each league. A more informative way to explore whether the home effect is actually true and estimate its degree is, of course, through modelling.

In Table 3.5 and Tables C.1, C.2, C.3, C.4 of the Appendix, the results from fitting a double Poisson model as discussed in section 3.4 to each corresponding league's season are displayed. In those tables, we have added the estimated home effect, and if we plot these, we get the image below,

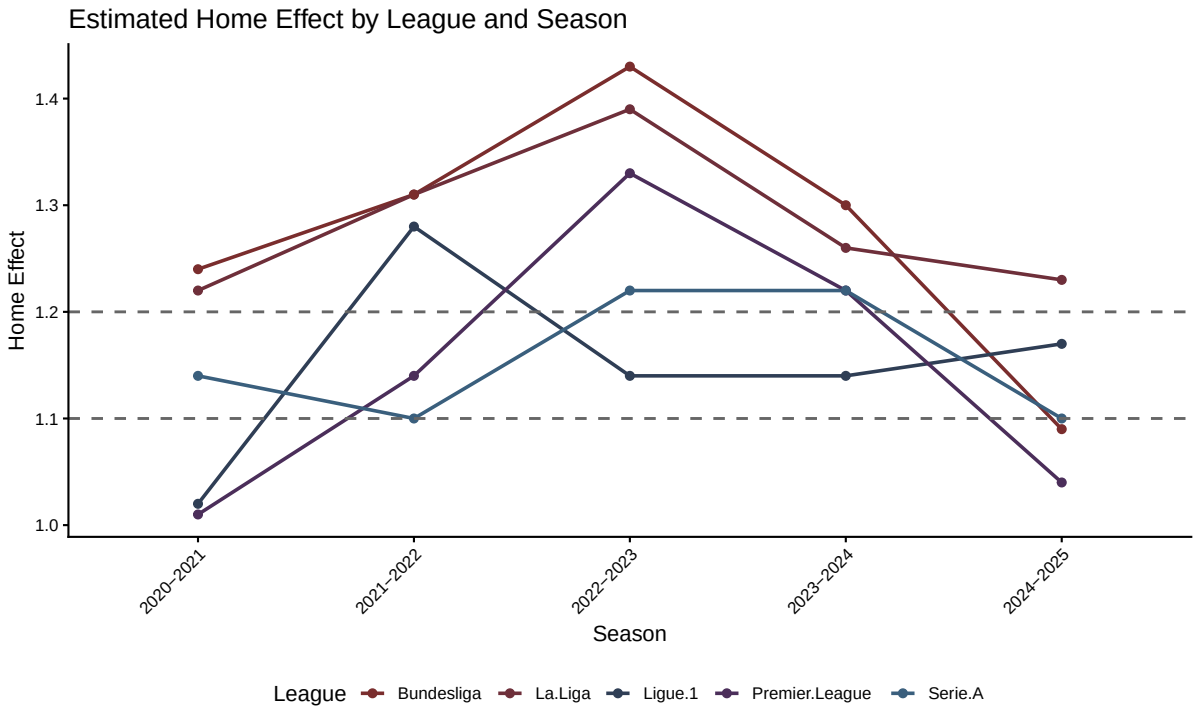


Figure 3.8: Estimated Home effect, under the double Poisson model of 5 major European leagues throughout the 2020-2025 seasons

The home effect is defined as the relative increase in expected goals scored by the home team when two teams of equal strength play against each other. For example, in the 2022–2023 Bundesliga season, the estimated home effect is equal to 1.43, indicating that the average home team is expected to score 43% more goals than the average away team, which was the highest recorded during those 5 seasons among the 5 major European Championships.

During the 2020-2021 season, which not so randomly happened to be the Covid-19 season with closed stadiums, all leagues were estimated to have a lower home effect compared to the next seasons, with the Premier League and the Ligue 1 home advantage being equal to just 1.01 and 1.02, respectively. Next on the 2022-2023 season, there was a peak of home effect for all leagues, which followed a downward trend ever since.

The interpretation of the factors behind this somewhat common behaviour among all leagues may vary. Covid-19 and its effect probably played a major role in the lows that were achieved that year, since games were played without fans on the stands and

teams were left without practising for a long time, bringing all teams closer in terms of strength. And now, probably the incorporation of VAR has brought balance to referee decisions, which have been found to usually be in favor of home teams (see, for example, Sutter and Kocher (2004), Garicano et al. (2001), Nevill et al. (2002)).

Moreover, further down below, we will explore the alternative of adding covariates to the home effect, specifically meaning that we will no longer assume a common home advantage for all teams in a league, but explicitly model it for each one of them. In the next section, we will explore an alternative and more informed way of how we can model a team's abilities.

3.7 A word about efficiency

In Section 2.6, we discussed the hierarchical model consisting of a Poisson distribution for modelling the number of trials and a binomial distribution for modelling the number of successes given the total number of those trials. We can then apply this regime to effectively model various phenomena relative to football. But first lets clarify what exactly that model describes.

Compound Poisson assumes the number of trials N is a Poisson-distributed variable with parameter λ and that $X \mid N$ follows a binomial distribution with probability p. Intuitively, we model the quantity of a given occurrence/phenomenon as N and the quality/likelihood of that specific event as p. For example, we can model the total number of crosses made by a team (quantity) and then by modelling the times those crosses turned into an actual goal (quality). Additionally, we can model the cards shown to a team and again model as a success if that card is actually red, accounting for the team's aggressiveness.

This approach provides a more thorough explanation (inference, as mentioned in the introduction of Chapter 3) while enabling estimation and prediction of outcomes. In this section, we concern ourselves with modelling shots and shots on target, examining what each model's estimates imply about teams' effectiveness and assessing the predictive performance of these models.

For the sake of convenience, equation (2.5) is presented in another form, so that covariates may be introduced later into the model. By equating $\lambda = \mu_n$ and (2.7) $p\lambda = \mu_x$, we obtain a "normalised" joint distribution of (N, X), as follows

$$\begin{aligned} P\left(N = n, X = x\right) &= \frac{\mu_n^n \left(\frac{\mu_x}{\mu_n}\right)^x \left(1 - \frac{\mu_x}{\mu_n}\right)^{n-x}}{x!(n-x)!} \cdot \exp(-\mu_n) \\ &= \frac{\mu_x^x (\mu_n - \mu_x)^{n-x}}{x!(n-x)!} \exp(-\mu_n), \end{aligned} \tag{3.2}$$

for $n = 0, 1, \ldots$ and $x = 0, 1, \ldots, n$, $\mu_n > 0$ and $0 < \mu_x < \mu_n$.

This probability function satisfies the condition that the marginal means should be given by $\mathbb{E}(N) = \mu_n$ and $\mathbb{E}(X) = \mu_x$ and thus is appropriate for introducing covariates. When covariates are considered, the log-likelihood is now proportional to

$$\ell\left(\beta_n, \beta_z; \tilde{n}, \tilde{x}\right) \propto \sum_{i=1}^{n} \left[x_i \log \mu_{x_i} + (n_i - x_i) \log\left(\mu_{n_i} - \mu_{x_i}\right) - \mu_{n_i}\right]$$

where

$$\mu_{N,i} = \exp(\omega_{N,i}^{\top}\beta_N),\ \mu_{X,i} = \mu_{N,i} \cdot \frac{\exp\left(\eta_{X,i}^{\top}\beta_X\right)}{1 + \exp\left(\eta_{X,i}^{\top}\beta_X\right)}. \tag{3.3}$$

Effectively, we have two different equations that we try to estimate a Poisson part and a Binomial part. Note that μ_X is a proportion p (which accounts for the number of successes) of μ_N.

3.7.1 Poisson

In this part, we let N : {Number of Shots / Shots on target} (from now on, we will just refer to them as shots) be a Poisson random variable with parameter $\mu_{N,i,j,k}$ for each team. Just like we discussed in section 3.4, the shooting intensity/frequency, denoted as $\lambda_{N,i,j,k} = \mu_{N,i,j,k}$ of each team, should capture the same team's ability to create opportunities for a shot as well as the opponent's ability to prevent these. Additionally, we would like to measure the overall home effect in the shooting intensity for a given league, if any exists. Without further ado, just as before, and based on the equation (3.3), we write the equation as

$$\log(\mu_{N,i,j,k}) = \gamma + \delta \cdot home_j + att_{team_k}^{(N)} + def_{opp_k}^{(N)}, \tag{3.4}$$

where $i = 1, 2, \ldots, G$ with G being the number of total matches in a league's season. We denote N as the number of trials, which in this case is the number of shots. Then, $j = 1, 2$ corresponds to the team playing home or away, and $k = 1, 2 \ldots T$ refers to each team as a covariate (for more details see section 3.4 and 3.3), with T being the total number of teams that play in each league's season.

The interpretation of each parameter is the same as in the Poisson model we thoroughly discussed in section 3.4, with the only exception being the response variable and therefore the context. Each team's covariate and home effect, in this case, does not measure the scoring intensity but the ability of a team to create shooting opportunities, when referring to the attacking one, and the opponent's ability to prevent that same team from creating such opportunities, when referring to the defensive one. Home effect, respectively, measures the general home effect that is exhibited in a given league's season, if one exists.

This way, we can actually quantify the attacking intensity or pressure a team puts on its opponent and, correspondingly, estimate the opponent's ability to restrict or minimise that pressure. This approach constitutes an indirect method to estimate any team's attacking and defensive abilities without accounting for its effectiveness. It provides an evaluation of the team's quantity in both creating and preventing opportunities, which we will tie to the evaluation of quality we will have in the next section.

3.7.2 Binomial

In the binomial part, we let X : {Number of Goals} and therefore that $X \mid N$ (X given N, where N is the number of shots) is a binomial distributed random variable with parameters n (the estimated shots from the Poisson part, see section 3.7.1) and p the

probability of each shot converting into a goal. As was discussed in previous sections, we would like to express the effectiveness of any team in a given match as the probability denoted by $p_{i,j,k}$, in relation to the team, its opponent, and whether it is at home or away. Based on the equation (3.3), we write the proportions of a team's shots converting to a goal as

$$logit(p_{i,j,k}) = \log\left(\frac{p_{i,j,k}}{1-p_{i,j,k}}\right) = \gamma + \delta \cdot home_j + att^{(p)}_{team_k} + def^{(p)}_{opp_k}, \tag{3.5}$$

where $i = 1, 2, \ldots, G$ with G being the number of total games in a league's season and $j = 1, 2$ referring to the home or away ground. Similarly, $k = 1, 2, \ldots, T$ expresses the teams as covariates where T is the total number of teams and $p_{i,j,k}$ denotes the k team's estimated conversion ability, attacking or defensive.

As we discussed in the subsection 3.7.1, the equation is similar to the previous cases we came across, but once again, the context changes. We essentially capture the likelihood, or conversion rate, of a team's shots resulting in a goal. The higher the attacking ability, the higher the effectiveness or quality of a team's shot, and from the other side, the lower the defensive ability, the lower the conversion rate that the opponent is conceding. Additionally, the home effect measures the overall advantage that teams exhibit when playing at home, in this case higher conversion rate, if any exists.

In this case, however, we use the *logit* link instead of the *log*. This results in a different interpretation of each parameter. The quantity $p_{i,j,k}/(1-p_{i,j,k})$ is also called odds and can take any value between 0 and ∞. Values of the odds close to 0 and ∞ indicate very low and very high probabilities or rates of goal conversion, respectively. By taking the logarithm of the odds, we arrive at the equation (3.5), the *logistic regression*, which is linear with respect to the coefficients.

Let's first consider the simplest case of logistic regression, where we want to estimate the probability p of an event Y happening with respect to a predictor X using the logistic regression as

$$p = \frac{e^{a+bX}}{1+e^{a+bX}} \iff \frac{p}{1-p} = e^{\alpha+\beta X} \iff \log\left(\frac{p}{1-p}\right) = \alpha + \beta X.$$

Recall from the introduction of chapter 3 and the section 3.4 that in linear regression, the parameter β (the coefficient of the predictor X) gives the average change in the response variable Y with one unit increase in X (for more details, see the book James et al. (2014)). Similarly, in a logistic regression model, increasing X by one unit changes the log odds by β. Equivalently, it multiplies the odds by e^{β}. The relationship between p and X, however, is not linear, β does not correspond to the change in p associated with a one-unit increase in X. But regardless of the value of X, if β is positive then increasing X will be associated with increasing p, and if β is negative then increasing X will be associated with decreasing p. The fact that there is not a straight-line relationship between p and X, and the fact that the rate of change in p per unit change in X depends on the current value of X, can also be seen by inspection of the Figure below, where we drew such a relationship.

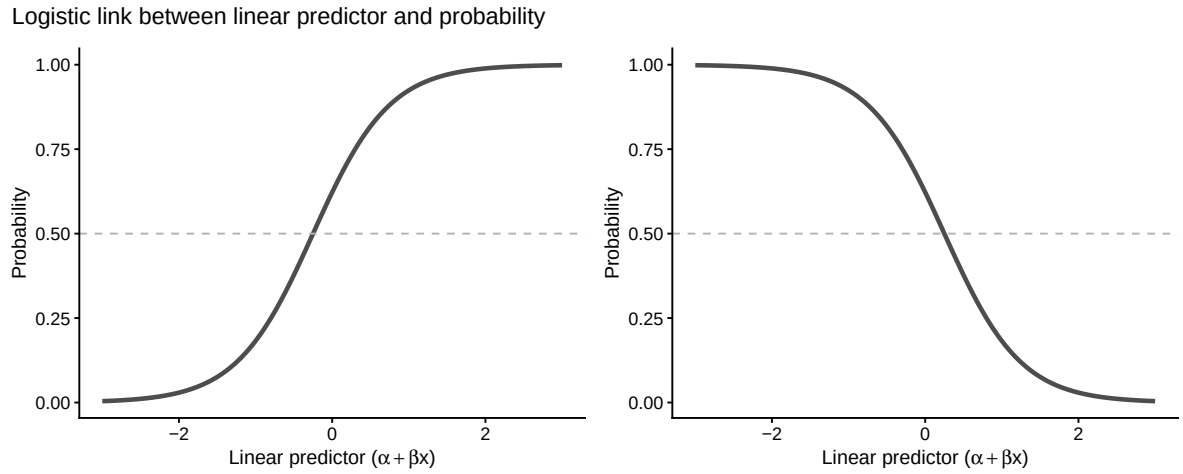


Figure 3.9: Logistic Regression example plot. Both plots have a common $a = 1/2$, the left one is with $\beta = 2$ and the right one with $\beta = -2$

What we discussed above applies when the predictor X is a continuous variable; in this particular case, however, we aim to estimate the goal conversion rate given the team, its opponents, and whether the game is played at home or away. All these represent categorical (covariates as are usually referred to, see section 3.3) variables that are incorporated into the model through dummy variables. Similarly, to the double Poisson model we discussed in section 3.4, for identifiability reasons, we set a baseline team (reference) against which the effects (abilities) of all other teams will be compared.

3.7.3 Combination of both

To allow for easier interpretation when discussing each team's efficiency, we impose sum-to-zero constraints on both the Poisson (3.4), and the Binomial part (3.5) as

$$\sum_{k=1}^{K} \beta_{att_{team_k}} = 0 \quad \text{and} \quad \sum_{k=1}^{K} \beta_{def_{opp_k}} = 0$$

for both the attacking and defensive abilities of the teams. That way, comparison is always made with a team of average (attacking or defensive) strength/ability rather than a specific baseline team for both models. Additionally, in practice, we usually set the sum-to-zero dummies with $\beta_{att_{team_1}} = -\sum_{k=2}^{K} \beta_{att_{team_k}}$ and $\beta_{def_{opp_1}} = -\sum_{k=2}^{K} \beta_{def_{opp_k}}$ just as we discussed in section 3.4.

When examining the Poisson model, positive values of attacking abilities correspond to teams that create more shooting opportunities than average, and negative values of defensive abilities correspond to teams that are better at preventing such opportunities from happening than average. Additionally, $e^{\beta_{att_{team_k}}}$ represents the relative increase/decrease in expected shots in comparison to an average team, and similarly $e^{\beta_{def_{opp_k}}}$ represent the relative decrease/increase in expected shots of the opponent team in comparison to an average team.

In the binomial model, likewise, positive values in attacking abilities indicate teams achieving a greater goals-to-shots ratio than the average team, and negative values for

defensive abilities indicate that the opponent has a worse goals-to-shot ratio compared to the average league's defence. Similarly, $e^{att^{(p)}_{team_k}}$ and $e^{def^{(p)}_{opp_k}}$ represent the relative increase/decrease in goal-to-shot conversion compared to the average team corresponding abilities (attacking or defensive).

In Figure 3.10 below, we present the estimated compound Poisson model, via the shots (and not shots on target), as introduced above, and display the corresponding team covariates. We used data from the Italian championship, Serie A, in the 2024-2025 season. On the y-axis, we have placed the Poisson effects, which correspond to shot volume for each team, which are represented on a log scale. On the x-axis, we have placed the binomial effects, which correspond to the conversion rate of each team, and they are represented on a logit scale. Additionally, we display with different shades the attacking and defensive abilities.

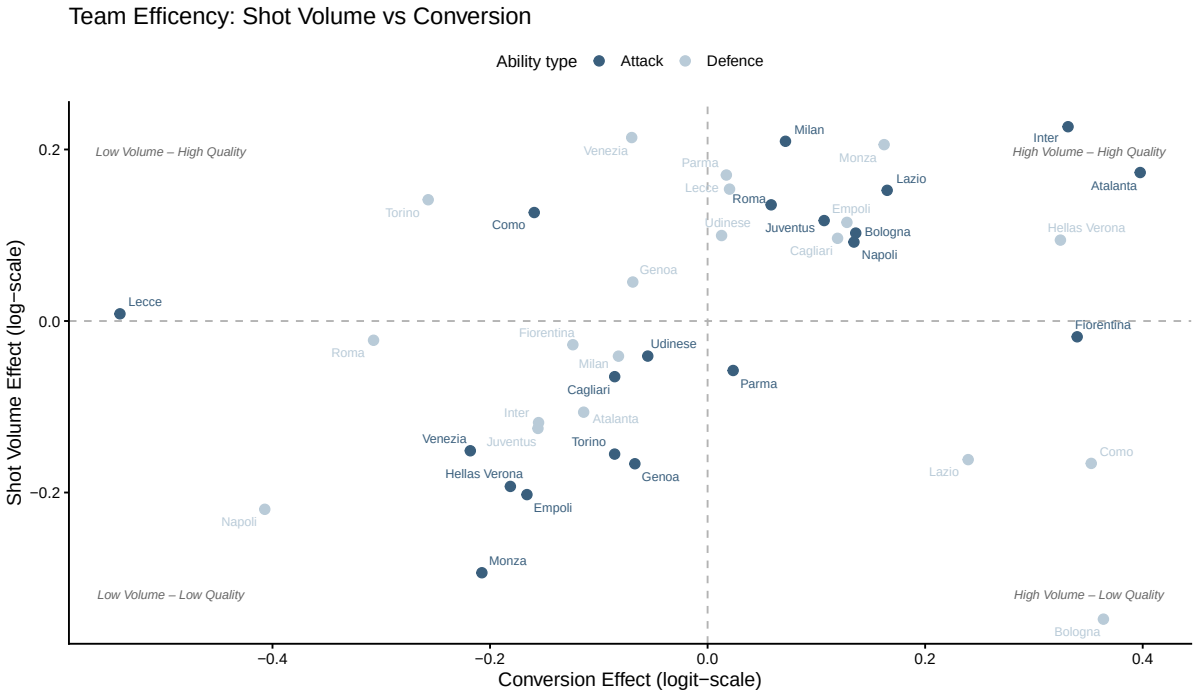


Figure 3.10: Map of team efficiency estimated from a Compound Poisson model, data concern the Italian Championship, Serie A, for the season 2024-2025.

The intuition behind the plot is that at the centre of the plot, at $(0, 0)$, the average team is placed. As we go west, we stumble across teams that are estimated to have a lower shot volume, but the interpretation depends on whether it refers to attacking or defensive ability, which is displayed via colour shade. For example, Lecce (left of the plot, darker shade) has the lowest shooting volume, and Gerona (just above the center of the axis, with lighter shade) concedes shots just below average, which is a good defence indication. Likewise, as we go north, we find teams that have a greater goal-to-shot conversion rate compared to the average team. For example, Inter was the most efficient attacking team (and the team with the greatest shot volume), and Bologna was the defence that allowed the shots with the worst quality despite being the team that conceded the most shots. Additionally, Napoli won the title that season even though their attack was only slightly above average in terms of both chance volume and conversion rate, because they allowed

the fewest shots in the league, and those shots were, at the same time, of the lowest quality.

Below in Table 3.7, we have fitted this model specification considering both shots and shots on target as the Poisson part for the Serie A during the 2020-2025 seasons. We have also added the results of the double Poisson to serve as a baseline. In the appendix, the corresponding tables can be found for the rest of the leagues during the same time.

Serie A

Season	Model	Df	LogLik	Fit			Parameters		
				Home	Draw	Away	μ	μ_n	μ_p
2020–2021	Actual			150.00	103.00	126.00			
	Double Poisson	40	-1092.90	161.78	83.27	133.95	1.14		
	Compound Poisson (Sh)	80	-11275.59	162.05	83.33	133.62		1.14	1.00
	Compound Poisson (SoT)	80	-458.82	162.25	83.16	133.59		1.14	1.00
2021–2022	Actual			147.00	99.00	134.00			
	Double Poisson	40	-1059.97	156.62	86.83	136.55	1.10		
	Compound Poisson (Sh)	80	-12564.13	156.26	86.66	137.08		1.18	0.92
	Compound Poisson (SoT)	80	-346.20	156.17	86.96	136.88		1.11	0.98
2022–2023	Actual			159.00	98.00	123.00			
	Double Poisson	40	-1016.58	162.63	94.08	123.29	1.22		
	Compound Poisson (Sh)	80	-12082.41	162.18	94.09	123.73		1.17	1.04
	Compound Poisson (SoT)	80	-515.97	161.73	93.98	124.29		1.14	1.09
2023–2024	Actual			157.00	114.00	109.00			
	Double Poisson	40	-1010.01	162.23	93.13	124.64	1.22		
	Compound Poisson (Sh)	80	-12114.36	163.21	93.23	123.57		1.23	1.00
	Compound Poisson (SoT)	80	-507.79	163.76	93.16	123.08		1.26	0.97
2024–2025	Actual			148.00	114.00	118.00			
	Double Poisson	40	-1011.26	152.35	94.26	133.39	1.10		
	Compound Poisson (Sh)	80	-11038.77	153.27	94.02	132.72		1.21	0.91
	Compound Poisson (SoT)	80	-593.99	151.75	94.23	134.02		1.15	0.94

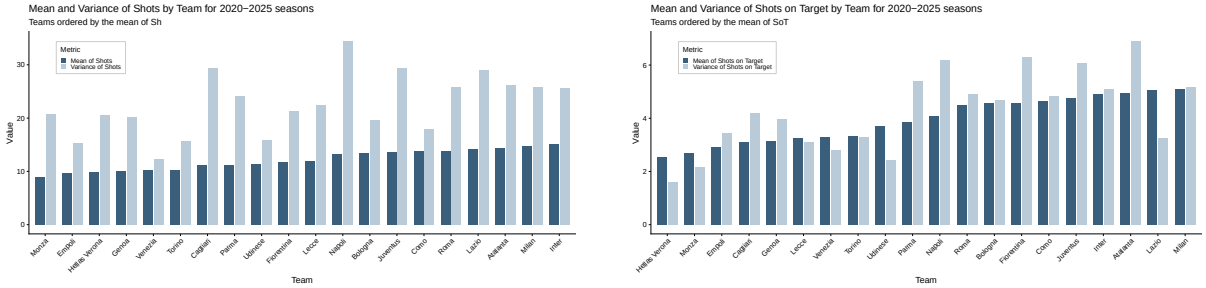
Table 3.7: Comparison of observed match outcome frequencies with expected values from double Poisson and compound Poisson models (Shots and Shots on Target) across Serie A during the 2020-2025 seasons.

All models yield very similar fitting results, as is expected (see Section 2.6). There is quite a difference, however, in the log-likelihood of each model since each one of them models different quantities (Poisson: goals, Sh: shots, goals, SoT: shots on target, goals) and additionally, the double Poisson estimates half the parameters of the other two.

The compound models offer, additionally, the option to estimate the number of shots or shots on target for each team, and the goal conversion rate on top of that. More detailed and informed conclusions can be drawn that way, based on this approach of modelling. We can quantify both attacking volume and shot quality for each team.

There is a little problem, however, with the Poisson part. As we discussed in section 3.5, Poisson has a very strict theoretical basis, which has to be practically validated for its results to be rightfully applicable and used. Considering the case of overdispersion, we

can see in Figure 3.11 that the shots of each team probably suffer from it. While shots on target seem to have a slight case of the phenomenon, they still require further investigation. This problem can probably be solved by using the negative binomial distribution instead of the Poisson, but this will be left for future work.



(a) Shots (b) Shots on target

Figure 3.11: Mean–variance relationship by team. The left panel shows shots, while the right panel shows shots on target. Data concern Serie A for the seasons 2020-2025

Additionally, one can argue that when we suppose that $X \mid N$ follows a binomial distribution, we also suppose that each trial is independent from the others. Of course, this is usually not true; however, we can make such a simplification and achieve satisfactory results because, in many cases, it turns out that the dependence is actually minimal and can be ignored. This is similar to when we model the scoring intensity as independent in the double Poisson model, but this is something that we will explore a bit later. For now, we will focus on modelling score differences.

3.8 Modelling Score Differences

An alternative approach to modeling football data is to model the final score or the goal difference of a match (Karlis et al. (2024), Karlis and Ntzoufras (2003)). Rather than jointly modelling the number of goals scored by each team, we shift our focus to the difference between the teams' goals. This strategy allows us to eliminate the impact of the correlation between the scoring performances of the two opposing teams, and the resulting model does not require Poisson marginal distributions, although based on the estimates of such a model, we will try to approximate the goals scored by each team.

The observed score differences are treated as realizations from the Skellam distribution, which is commonly employed to model the difference between two independent Poisson variables. These differences, however, (for a retrospective study on the Skellam distribution see Tomy and Veena (2022)) are shown not to be restricted to either independence or Poisson distribution. The main idea is that we model the differences in count data without considering the marginal distributions at all.

Yet because football results exhibit particular structural features, we address them directly in two ways. First, we incorporate a zero-inflated (ZI) Skellam model, which reallocates probability mass toward zero and lets us explore the excess of draws that is generally observed as we discussed in subsection 3.5.2, and we also consider a different parametrisation. Second, we examine the score differences of the first and second halves

separately, and we then model them jointly by constructing bivariate distributions through copulas. This approach allows us to estimate the conditional distribution of the second-half difference given the first-half difference. Copulas offer a flexible and transparent framework for building such bivariate models, as they permit separate specification of the marginal distributions while introducing dependence via the chosen copula.

3.8.1 Univariate Skellam

In this subsection, an example of fitting a Skellam distribution to model football differences is presented. We will fit both a "regular" Skellam and we will also fit the mean/variance parametrization of the Skellam2 model, which was described in 2.7. More precisely, in the first scenario, we employ the previously described distribution to represent the goal difference by defining the score difference in match i through the equation2.8 as

$$Z_i = G_{h,i} - G_{a,i} \sim Skellam(\lambda_{h,i}, \lambda_{a_k,i}),$$

where G_{h_k} and G_{a_k} correspond to home and away goals, respectively, in the i game. Concerning the model parameters $\lambda_{h,i}$, $\lambda_{a,i}$, we adopt a similar structure as in the double Poisson model we used in section 3.4 where we modelled the number of goals scored by each team, but with a subtle difference, that we do not assume a league overall home advantage. So we model the team covariates as

$$\log(\lambda_{h,i}) = \gamma + att_{h_{(i)}} + def_{a_{(i)}}$$
$$\log(\lambda_{\alpha,i}) = \gamma + att_{a_{(i)}} + def_{h_{(i)}}$$

Here, γ is a constant parameter, att denotes the attacking strength of the home or away team, and def denotes the defensive strength of the corresponding home or away team for each match i.. Note that, because it is not a standard home effect assumed, as in Lee (1997) and Karlis and Ntzoufras (2003), which we adopted in section 3.4, each team is assumed to have different attacking or defensive abilities when playing at home or away. In Figure 3.12, below these differences are displayed. We applied this model to the 2024–2025 Bundesliga season, using Augsburg as the reference team, and then visualized each club's attacking and defensive strengths, using different colours to distinguish home and away performance.

We observe that the distinguished home advantage has a large impact on the team's performance. The margin of differences appears to be somewhat common for most teams; there are cases, however, where that margin deviates a lot from the league's norm, especially when considering the defensive abilities. For example, Wolfsburg and Union Berlin are estimated to have a better attacking ability when playing away, and Dortmund is estimated to have almost equal defensive ability when playing home and away.

But how do these differences translate exactly into an overall home advantage? To measure the overall difference, we fit the match intensities $\lambda_{h,i}$ and $\lambda_{\alpha,i}$. We then calculate a league-level mean home effect as a multiplier $\mathbb{E}(\lambda_{h,i})/\mathbb{E}(\lambda_{a,i})$, but since we are in the log scale we calculate the logarithm $\log \mathbb{E}(\lambda_{h,i})/\log \mathbb{E}(\lambda_{a,i})$. The result is 0.097, and by transforming, we get 1.10, very close to the estimated home effect (1.09) by the double Poisson model (see Table C.3 of the Appendix).

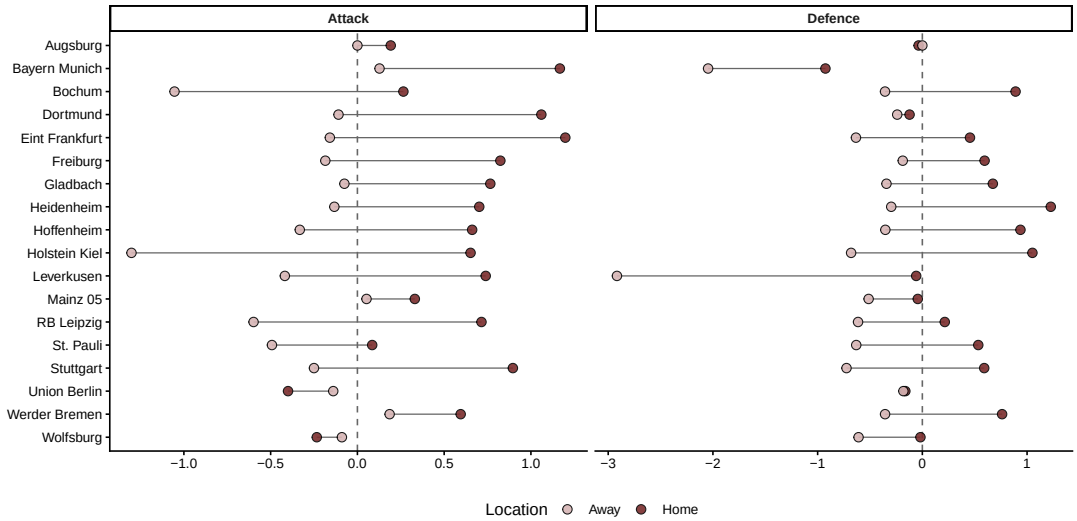


Figure 3.12: Estimated attacking and defensive team effects under home and away, obtained from a Skellam model for the 2024–2025 Bundesliga season.

On the other hand, in the Skellam2 parametrization, we assume that the score difference Z_{ij} for a given match between home team i and away team j follows a Skellam2(μ_{ij}, σ) as described in 2.7. More explicitly, we model the mean goal difference μ_{ij} between two teams as

$$\mu_{ij} = \gamma + h_i + a_j$$

Where i corresponds to the team playing at home and j to the opponent team playing away, each h_i represents the ability of the i-th team at home, while a_j is the ability away of the j-th team. For each model fit, a baseline team must be selected; for simplicity, the first team in alphabetical order is used. The interpretation of h_i is that when team i plays at home against the baseline team, the expected score difference is $\gamma + h_i$, while if the team plays away (in the baseline team's home), it is $\gamma + \alpha_j$. Both models explicitly assume a different home advantage for each team. All of the above were based on Karlis et al. (2024).

The key distinction between the two models is that the parameterized version assumes a common variance σ^2. More precisely, for each match we estimate the expected goal difference as μ_{ij}, while imposing a single variance σ^2 shared across the league. In contrast, in the earlier model specification, the variance was allowed to vary with the teams' covariates. This alternative parameterization produces parameters that are more straightforward to interpret. As an illustration, we again fit the model to the 2024–2025 Bundesliga season to compare the outcomes of both modeling approaches. Below in Table 3.8.

Again, as in the previous example, Augsburg is the Baseline team, so basically it is the fixed error γ. For instance, when Bayern Munich plays Augsburg at home, the expected score difference is $0.52 + 1.82 = 2.34$ for Bayern Munich, a result which corresponds to reality as Bayern Munich won the championship with 82 points and 13 points difference from the second, while Augsburg finished at mid table in 12^{th} ranking with 43 points. In the same motion, when Dortmund plays Eintracht Frankfurt at home, the expected score difference is $0.52 + 1.45 + -0.79 = 1.18$ for the home team, which is Dortmund. To illustrate a negative expected score, when Bochum plays Gladbach at

home, the expected score difference is $0.52 + -0.87 + -0.48 = -0.83$ in favor of the away team.

Team	Home Ability	Away Ability
Augsburg	0.52	
Bayern Munich	1.81	-2.21
Bochum	-0.87	0.03
Dortmund	1.45	-0.38
Eint Frankfurt	1.28	-0.79
Freiburg	0.31	-0.09
Gladbach	0.08	-0.48
Heidenheim	-1.06	-0.32
Hoffenheim	-0.51	-0.16
Holstein Kiel	-0.79	0.15
Leverkusen	0.86	-1.65
Mainz 05	0.32	-0.89
RB Leipzig	0.64	-0.34
St. Pauli	-0.51	-0.38
Stuttgart	0.39	-0.78
Union Berlin	-0.17	-0.02
Werder Bremen	-0.19	-0.93
Wolfsburg	-0.17	-0.91

Table 3.8: Home and away ability estimates for all Bundesliga teams for the season 2024–2025 based on a Skellam model.

Given that the estimated variance in the Skellam2 parametrization (see section 2.7) is $\hat{\sigma}^2$, we can also estimate the goals scored as

$$\hat{\lambda}_{h,i} = \frac{\hat{\sigma}^2 + \hat{\mu}_{i,j}}{2}, \quad \hat{\lambda}_{a,j} = \frac{\hat{\sigma}^2 - \hat{\mu}_{i,j}}{2}$$

That way, we may (arbitrarily) assume that the marginal distributions of the estimated score difference are two independent Poisson distributions with parameters $\lambda_{h,i}$ and $\lambda_{a,i}$ for the home and away team, respectively. We can then proceed as a regular double Poisson model to estimate and predict goals for each match as in section 3.4.

For a systematic examination and evaluation of the models proposed above, we have fitted the Skellam model and the parametrized version Skellam2, along with the zero-inflated version for both of them. Additionally, we include the results from the double-Poisson model as a reference benchmark. All models were estimated in R using the function optim, except for the double-Poisson model, which was fitted using the R function glm. The results are displayed in Table 3.9 below.

It is important to note in advance that the model fitting process was somewhat challenging. We found that the parameter estimates were sensitive to the choice of initial values, which therefore had to be selected with care. For the ZI model under the Skellam2 parameterization, we used the estimated parameters from the Skellam2 model as starting values; nevertheless, the initial value of p still had a strong influence on the final estimates. Consequently, the starting value of p is manually chosen for each league to improve the

Bundesliga

Season	Model	Df	Loglik	Fit			Special Parameters	
				Home	Draw	Away	σ^2	p
2020–2021	Actual			130.00	81.00	95.00		
	Double Poisson	36	−874.54	137.11	68.33	100.56		
	Skellam2	36	−586.60	142.67	53.74	109.59	4.35	
	ZI SKellam2	37	−581.48	127.29	80.30	98.41	4.53	0.11
	Skellam	70	−545.53	133.91	80.00	92.09		
	ZI Skellam	71	−545.25	131.01	85.16	89.83		0.03
2021–2022	Actual			139.00	76.00	90.00		
	Double Poisson	36	−893.42	142.27	66.35	96.38		
	Skellam2	36	−593.31	139.66	60.49	104.85	3.21	
	ZI SKellam2	37	−591.46	135.79	73.17	96.04	3.34	0.06
	Skellam	70	−567.19	138.70	74.81	91.49		
	ZI Skellam	71	−567.17	138.55	74.96	91.49		0.00
2022–2023	Actual			147.00	73.00	86.00		
	Double Poisson	36	−926.40	151.83	65.54	88.64		
	Skellam2	36	−603.81	150.83	62.33	92.84	3.17	
	ZI SKellam2	37	−602.16	147.11	73.04	85.85	3.32	0.05
	Skellam	70	−583.75	152.98	69.09	83.94		
	ZI Skellam	71	−582.85	147.97	77.44	80.59		0.05
2023–2024	Actual			135.00	81.00	90.00		
	Double Poisson	36	−889.87	143.21	62.56	100.23		
	Skellam2	36	−580.60	147.87	55.24	102.88	3.60	
	ZI SKellam2	37	−576.42	132.78	80.89	92.33	3.64	0.10
	Skellam	70	−549.04	139.53	76.55	89.93		
	ZI Skellam	71	−549.05	139.40	77.23	89.37		0.00
2024–2025	Actual			119.00	75.00	112.00		
	Double Poisson	36	−915.98	126.27	68.66	111.07		
	Skellam2	36	−581.98	126.55	68.29	111.16	2.74	
	ZI SKellam2	37	−581.48	122.27	73.92	109.82	2.85	0.03
	Skellam	70	−562.20	121.23	77.07	107.70		
	ZI Skellam	71	−562.20	121.15	77.09	107.76		0.00

Table 3.9: Observed and model-implied match outcome frequencies, together with log-likelihood values and special parameters, for Poisson and Skellam models fitted to Bundesliga data across the 2020–2025 seasons.

fit, and the team-specific parameters are directly inherited from the Skellam2 model. On the other hand, the "regular" Skellam model was not as sensitive to initial values, and the fitting was smoother throughout all leagues and seasons.

The zero-inflated specification, however, in a few cases failed to converge to parameter estimates that fit the championship properly, as we can see from the results in Table 3.9 above and in Tables C.10, C.11, C.12, C.13 in the Appendix.

We suspect the issue may stem from the fact that football is a low-scoring sport

and, as a result, when fitting the model we can have σ^2 close to μ. However, the chosen parametrization imposes the constraint $\left|\sigma^2\right| \geq \mu_i \ \forall i$. This might be an issue for fitting the model despite the convenient interpretation of the model's parameters; better optimization algorithms or a Bayesian approach may be considered.

The first Skellam model steadily predicts more draws than the double Poisson model. There are cases, however, where it predicts way more draws than those that actually occurred, for example (in Table C.10), in the Premier League in the season 2020-2021, the actual were 84, but the model predicted 102. The parametrized version initially did not manage to fit each league, but with the incorporation of the (properly initialized value selected) zero-infated parameter p, it greatly improved its performance. Specifically, this Skellam version offers a very agile model that can fit many different leagues, as it requires a mean score difference estimation $\mu_{i,j}$ and a common σ^2 for its variance, along with the "correcting" parameter p. Overall, these models appear to perform okay, not great, despite some instances where they outperform the double Poisson model, particularly in leagues with an unusually high number of draws.

3.8.2 Bivariate Skellam

Based on the encouraging results achieved by fitting the Skellam distribution in Section 3.8.1 and its parametrization, Skellam2, we further pursue modelling the score differences. In particular, we apply the same model specification as described in the previous section 3.8.1 to each half separately and then model the dependence between the two using a copula. In the past, McHale and Scarf (2011) used copulas not only to forecast, but to explain match outcomes, by determining what characterises winning teams. They focused more on the explanatory aspect of modelling rather than forecasting, as discussed in the introduction of Chapter 3.

An important characteristic of the Skellam distribution is its additive property: the sum of Skellam-distributed random variables is itself Skellam-distributed. In our specific bivariate setting, this implies that, for each half of each match, the goal difference follows a Skellam distribution, as outlined in Karlis et al. (2024). In the following, we investigate whether scoring performance is stable between the first and second halves of a match and assess the potential of this type of modelling for the Bundesliga over the 2020–2025 period, in direct analogy to the univariate analysis, for comparison.

An immediate difference emerges in the score differences for the separate halves, as can be seen in Figure 3.13 below. In the first half, the goal difference appears to be centered at 0, whereas in the second half, it is more dispersed, indicating the different performance of the teams in each situation. Both histograms express a slight advantage for home teams, as they exhibit a small surplus of positive values relative to negative values. This behaviour of score differences can be observed in all leagues for which we have data (see figure B.9 of the appendix for reference).

In plain English, all of the above translate to games being more conservative in the first half, whereas the second halves seem more decisive and show a slight positive correlation (see Table 3.10 below). Although the score differences Z_1 and Z_2 are integer-valued and may take negative values, the Pearson correlation remains well-defined and provides a natural measure of linear dependence between the two halves. Below in Table 3.10, the correlation by each half throughout the 5 major European championships during

the 2020-2025 seasons is displayed.

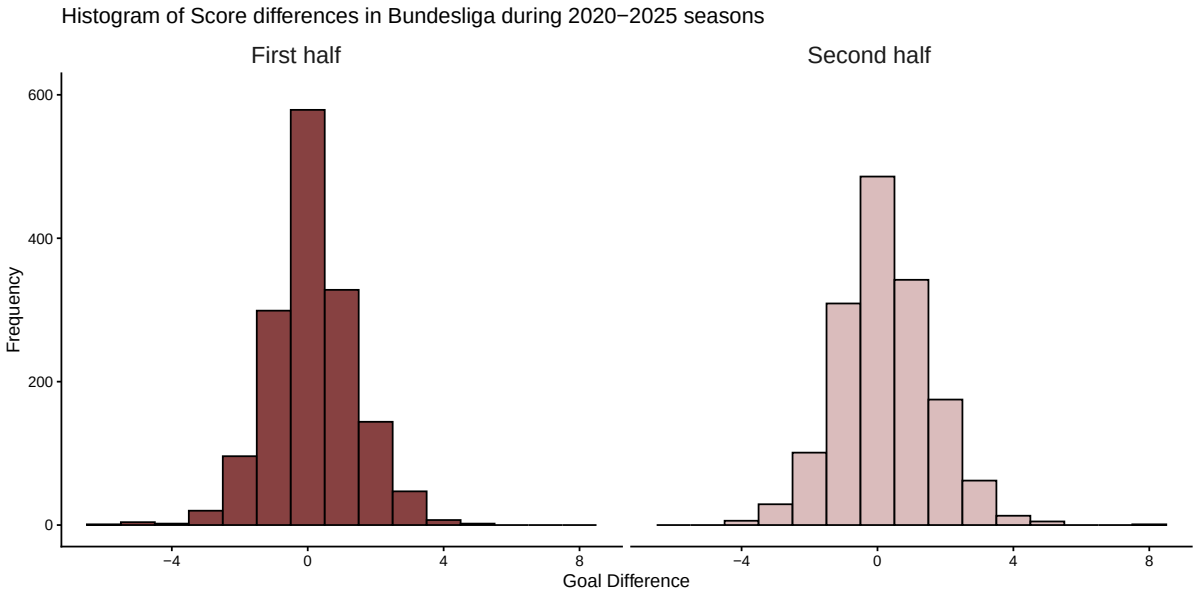


Figure 3.13: Empirical distributions of score differences by half for Bundesliga matches over the 2020–2025 seasons.

League	Pearson	Spearman	Kendall
Premier League	0.101	0.079	0.064
Serie A	0.046	0.023	0.018
La Liga	0.010	0.004	0.003
Bundesliga	0.096	0.083	0.067
Ligue 1	0.055	0.042	0.035

Table 3.10: Correlation measures between first-half and second-half score differences across major European leagues for the 2020-2025 seasons.

As in the univariate case, we consider the Skellam2 parametrisation to include covariates. As described in 2.10 and Karlis et al. (2024), we model two random variables Z_{i1} and Z_{i2} as the score difference for each half, where each of them marginally is Skellam2 reparametrised variable with mean μ_i for $i = 1, 2, \ldots, n$, where n is the number of matches and σ_j^2 where $j = 1, 2$ for the separate halves.

To clarify the model specifications we need to estimate,

$$\mu_{i1} = b_1^T Z_{1i}$$
$$\mu_{i2} = b_1^T Z_{2i}$$

where b corresponds to the team's coefficients for each match. We further model the dependence between the two variables using a Frank copula, allowing for positive or negative dependence with parameter θ. Additionally, we have considered fitting a zero-inflated version to improve model fit. Overall we need to estimate $\Theta = (\beta_1, \beta_2, \sigma_1^2, \sigma_2^2, p_1, p_2, \theta)$.

Bundesliga

Season	Model	Df	LogLik	Fit			Special Parameters					
				Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ	corr
2020–2021	Actual			130.00	81.00	95.00						0.05
	Double Poisson	36	-874.54	137.11	68.33	100.56						
	Skellam2	72	-967.96	136.35	54.62	115.01	1.40	2.84				
	ZI Skellam2	74	-958.71	130.42	64.92	110.64	1.40	2.84	0.13	0.10		
	FC Skellam2	73	-967.05	135.56	56.92	113.49	1.40	2.84			-0.66	
	ZI FC Skellam2	75	-958.03	129.90	66.54	109.54	1.40	2.84	0.13	0.10	-0.51	
2021–2022	Actual			139.00	76.00	90.00						0.08
	Double Poisson	36	-893.42	142.27	66.35	96.38						
	Skellam2	72	-982.30	145.72	57.07	102.20	1.40	2.26				
	ZI Skellam2	74	-981.93	144.59	58.63	101.77	1.40	2.26	0.00	0.03		
	FC Skellam2	73	-982.28	145.72	57.31	101.97	1.40	2.26			-0.08	
	ZI FC Skellam2	75	-981.91	144.59	58.86	101.54	1.40	2.26	0.00	0.03	-0.08	
2022–2023	Actual			147.00	73.00	86.00						0.05
	Double Poisson	36	-926.40	151.83	65.54	88.64						
	Skellam2	72	-996.06	154.48	57.66	93.85	2.09	1.62				
	ZI Skellam2	74	-988.36	149.79	63.92	92.28	2.09	1.62	0.14	0.00		
	FC Skellam2	73	-996.00	154.43	57.17	94.39	2.09	1.62			0.15	
	ZI FC Skellam2	75	-988.32	149.74	63.57	92.68	2.09	1.62	0.14	0.00	0.11	
2023–2024	Actual			135.00	81.00	90.00						0.07
	Double Poisson	36	-889.87	143.21	62.56	100.23						
	Skellam2	72	-969.83	152.19	52.27	101.53	1.83	2.12				
	ZI Skellam2	74	-963.82	147.24	59.67	99.08	1.83	2.12	0.10	0.08		
	FC Skellam2	73	-968.77	152.23	54.21	99.54	1.83	2.12			-0.71	
	ZI FC Skellam2	75	-962.55	147.40	61.46	97.13	1.83	2.12	0.10	0.08	-0.72	
2024–2025	Actual			119.00	75.00	112.00						0.08
	Double Poisson	36	-915.98	126.27	68.66	111.07						
	Skellam2	72	-953.31	130.01	61.95	114.04	1.55	1.75				
	ZI Skellam2	74	-953.04	129.10	63.56	113.35	1.55	1.75	0.02	0.03		
	FC Skellam2	73	-953.30	129.92	62.20	113.88	1.55	1.75			-0.07	
	ZI FC Skellam2	75	-953.03	129.02	63.78	113.20	1.55	1.75	0.02	0.03	-0.07	

Table 3.11: Model comparison for the Bundesliga over the 2020–2025 seasons. The table lists degrees of freedom (Df), model fit loglikelihood, empirical outcome frequencies (Home, Draw, Away), and key special parameters for each model.

Once more, we took the Bundesliga during the 2020–2025 seasons as our illustrative example. Similarly to the univariate case, we fitted a double Poisson model as a baseline, a Skellam2 model without copula dependencies, a zero-inflated version of that, a Skellam2 model with copula dependencies, and, finally, a zero-inflated version of that as well. We again fitted the models using Maximum Likelihood Estimation via the R function optim.

The first model is a double Poisson just for baseline reference. For each half, we have specified its own mean score difference μ, as well as its own variance σ^2, following the considerations outlined above. We estimated the model both under the assumption

that the two halves are independent—referred to as Skellam2, and ZI Skellam2 when accounting for zero inflation—and using a Frank copula to model dependence between the halves, which we denote by FC. The parameter p_1 corresponds to the first half zero inflation, while p_2 corresponds to the second half zero inflation. The results are presented in Table 3.12 below, as well as in the corresponding tables in the appendix for the other leagues.

As the plot 3.13 indicates, the observed frequency of the draws is higher in the first half than in the second. This pattern was repeated in most seasons as $p_1 > p_2$, indicating that draws occur more frequently in the first half than in the second. In the same essence, generally $\sigma_1^2 < \sigma_2^2$, which indicates a greater dispersion of score differences in the second half, since these scalar variables capture the overall variability of score differences for a given league in a particular season.

We observe in Table 3.11 that teams generally play more aggressively in the second half, as $p2$ is smaller than $p1$ in most cases. Additionally, variance was greatly reduced compared to the univariate setting (since splitting the match into two parts also decreased the underlying variability) and in general $\sigma_2^2 \geq \sigma_1^2$. This aligns with the patterns already suggested by Figure 3.13. Similar observations can be made for all the other European championships reported in the Appendix.

Despite the general slightly positive correlation that is calculated throughout all leagues and seasons, the copula's dependence parameter θ is estimated to be mostly negative (see Table 3.11, and the Tables C.14, C.16, C.18, C.20 of the Appendix). This is very similar to the overdispersion case that we discussed in section 3.5.3.

Specifically, here, the correlation measures the overall dependence of score differences between halves, including the general variation that logically and empirically exists between each different team's performance. After including covariates, essentially meaning teams, into the Skellam2 model estimation and subsequently introducing the dependence parameters, the resulting correlation is predominantly negative. Nevertheless, there are still a few instances of positive correlation, suggesting that the model fit is not truly stable and does not truly capture the actual nature of dependence.

This issue would likely be resolved if we added covariates to the copula parameter as well. This approach, however, would lead to substantial identifiability problems and result in an overparameterized model that would be very difficult to interpret.

Nonetheless, this model specification produces results comparable to those of the double Poisson model (see Table 3.11, and Tables C.14, C.16, C.18, C.20 in the Appendix), while offering the additional advantage of enabling the estimation of first- and second-half outcomes, as shown in Table 3.12 below. This, in turn, makes it possible to analyze the joint distribution of half-time and full-time results directly, rather than restricting attention to marginal outcome frequencies alone.

By presenting the empirical transition matrices alongside the corresponding model-based expectations, the table makes it possible to evaluate not only the overall goodness of fit, but also how well each model captures the dependence structure between the two halves of a match. In particular, differences between the independent and copula-based specifications become apparent, indicating precisely where the independence assumption across halves fails to match the transition patterns observed in the data, if it actually fails.

Bundesliga: Joint Half-Time / Full-Time Results

Season	Actual			Independent ZI Skellam2			Frank Copula ZI Skellam2		
	69	16	5	64.6	9.8	8.0	62.6	10.7	9.1
2020–2021	50	53	34	53.2	42.4	40.6	53.1	42.1	40.9
	11	12	56	12.6	12.8	62.1	14.2	13.7	59.6
	89	16	6	87.1	13.0	7.9	86.8	13.2	8.0
2021–2022	40	35	30	45.7	31.2	32.0	45.7	31.2	32.0
	10	25	54	11.8	14.4	61.9	12.0	14.5	61.5
	97	14	7	96.6	12.9	5.8	96.9	12.7	5.7
2022–2023	40	45	33	45.7	38.3	31.8	45.6	38.4	31.9
	10	14	46	7.5	12.7	54.6	7.3	12.4	55.1
	84	17	7	93.5	11.4	7.1	91.0	12.6	8.3
2023–2024	40	46	27	44.3	36.4	31.3	45.2	35.8	31.0
	11	18	56	9.4	11.9	60.7	11.1	13.1	57.8
	80	12	9	83.1	13.9	7.4	82.9	14.1	7.5
2024–2025	31	43	32	37.4	35.1	33.2	37.5	35.1	33.3
	8	20	71	8.5	14.5	72.7	8.7	14.7	72.5

Table 3.12: Observed and model-based joint half-time/full-time outcome matrices for Bundesliga seasons 2020–2025. Each cell contains a 3×3 matrix describing transitions from half-time (rows: Home, Draw, Away) to full-time (columns: Home, Draw, Away). The two model-based columns represent the Independent and Frank Copula Zero-Inflated Skellam2 specifications.

Generally, from table 3.11 and table 3.9, we conclude that the introduction of the copula does not substantially improve the Skellam model specification we already have. This can also be observed in the corresponding tables for the rest of the leagues, which are included in the appendix.

Additionally, the fitting of all models presented was once again difficult to achieve using the R function optim. We had to estimate each parameter separately to yield consistent results. To obtain a strongly accurate model, one could employ alternative estimation methods; we leave this for future research. This set of models illustrates the differences in performance across teams, a topic we will explore further in the following section, where we will be returning to Poisson-based models.

3.9 Dependence Structures with Poisson

Throughout this thesis, we have discussed various methods for modelling or describing the scoring patterns of football matches. The main idea, however, has been to estimate the overall defensive and attacking capabilities of each team, except for the previous Section 3.8 where we discussed various models based on the Skellam distribution and their slightly

different applications and interpretation .

Conceptually, this idea is like cutting the pitch in half and observing what happens on each goalpost. We abstract a team's attack and "place" it against its opponent's defence, and then we observe what the outcome will be without taking into consideration the interactions between the two teams. It is a rather simplified approach to how the game of football is actually played, but an effective approach as we have seen so far.

In the previous section, where we modelled the score differences, we incorporated the dependence that might emerge from each half of a game. There, we had the chance to explore how a team's performance varies in each game's half and the dependence structure between them.

In this section, we take a closer look at the possible overall dependence that may be present in a football match. We will construct a model for the interactions between the two opposing teams and, building on the earlier approach, examine the different forms of dependence between the two halves. To capture (or "couple," as the names suggest) the multivariate distribution of such a complex structure, will use copulas, which were discussed in section 2.10. The foundation will be a Poisson distribution, similar to what was discussed in section 3.4, upon which we will incorporate the dependencies we wish to explore.

3.9.1 Team Dependence

We will begin the exploration of the various dependencies in a football match, rather simply, and gradually we will proceed with more complex model structures. Before delving into mathematical equations and models, let's briefly note the overall dependence of scoring patterns that opposing teams exhibit in the 5 major European leagues for the 2020-2025 seasons. Table 3.13 below displays exactly that.

League	**Pearson**	**Spearman**	**Kendall**
Bundesliga	-0.1320	-0.1010	-0.0832
La Liga	0.0044	0.0141	0.0120
Ligue 1	-0.0868	-0.0697	-0.0580
Premier League	-0.1310	-0.1180	-0.0980
Serie A	-0.0853	-0.0657	-0.0547

Table 3.13: Correlation coefficients between Goals scored by opposing teams across major European leagues for the 2020-2025 seasons.

Across all major European leagues, the correlation between the team's and opponents' goals scored is slightly negative, with the only exception being the Spanish La Liga, which is practically zero. The different correlation metrics: Pearson, Spearman, and Kendall are consistent in sign and magnitude, which is reassuring. The small negative values suggest a weak inverse relationship, meaning that the more goals a team scores, the less its opponent does, and vice versa.

Analogously to section 3.4 for the double Poisson model, we assume that X and Y, representing the number of goals scored by the home and away teams respectively, are

random variables following Poisson distributions. On estimating the lambda parameters in each game, we use covariates as regressors based on the following equations

$$\log(\lambda_{h,i}) = \gamma + att_{h_{(i)}} + def_{a_{(i)}}$$
$$\log(\lambda_{\alpha,i}) = \gamma + att_{a_{(i)}} + def_{h_{(i)}}$$

where γ denotes the constant parameter, att represents the attacking strength of either the home or away team, and def represents the defensive strength of the corresponding home or away team for each match i.

Premier League

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	θ	corr
2020–2021	Actual			146.00	84.00	150.00		-0.05
	Double Poisson	40	-1074.59	145.14	92.15	142.71		
	Double Poisson 2	78	-1058.92	140.13	99.95	139.92		
	Frank Copula Poisson	79	-1058.57	143.17	94.11	142.72	0.30	
2021–2022	Actual			162.00	88.00	130.00		-0.15
	Double Poisson	40	-1059.38	159.67	85.34	134.99		
	Double Poisson 2	78	-1037.47	154.98	92.80	132.22		
	Frank Copula Poisson	79	-1037.45	159.88	83.97	136.15	-0.09	
2022–2023	Actual			180.00	90.00	110.00		-0.04
	Double Poisson	40	-1079.43	174.87	88.23	116.89		
	Double Poisson 2	78	-1059.76	168.66	96.06	115.28		
	Frank Copula Poisson	79	-1059.27	173.06	90.57	116.37	0.36	
2023–2024	Actual			174.00	83.00	123.00		-0.15
	Double Poisson	40	-1121.17	171.89	79.42	128.69		
	Double Poisson 2	78	-1108.95	168.82	86.77	124.42		
	Frank Copula Poisson	79	-1107.99	175.05	75.10	129.86	-0.54	
2024–2025	Actual			152.00	96.00	132.00		-0.12
	Double Poisson	40	-1081.79	151.00	86.58	142.42		
	Double Poisson 2	78	-1061.41	147.81	95.09	137.10		
	Frank Copula Poisson	79	-1060.95	153.67	83.71	142.62	-0.40	

Table 3.14: Model comparison for the Premier League over the 2020–2025 seasons. The table lists degrees of freedom (Df), model fit loglikelihood, observed match outcome frequencies (Home, Draw, Away), and dependence parameters for the Frank copula Poisson model.

In the attempt to capture the dependence between the scoring rates of the two opposing teams, we use a Frank copula (see Section 2.10 for details), which estimates a parameter, denoted by θ, representing their correlation. Based on the results of Table 3.13 above, the Premier League exhibits the greatest (negative) correlation, and thus its data are gonna be used as an illustrative example for the model specification described above. Table 3.14 above displays the results of such fitting, and in Table C.22, C.23, C.24, C.25 of the Appendix, one can find similar results for the rest of the leagues.

In detail, we estimated a double Poisson model (identical to the one in Section 3.4), referred to as Double Poisson, and a second specification, Double Poisson 2, which is the same model but without the common home-effect coefficient. The Frank Copula Poisson model is the extension introduced above that incorporates the dependence parameter between the two opposing teams. All models were fitted using the glm function in R, and the dependence parameter was obtained via the optim function, using the glm estimates as starting values.

We note that, although the correlation is consistently computed as slightly negative, the specified model still produces varying estimates for the parameter θ in both absolute value and sign. This is clear, especially when comparing the 2021-2022 and 2023-2024 seasons of the Bundesliga, where both have a calculated correlation of -0.11 (see table C.24 of the Appendix), but the dependence parameter θ is estimated to be -0.45 and 0.08, respectively.

The concerning result in Table 3.14 and the corresponding tables in the Appendix is that there is no clear pattern when estimating the dependence of the opposing team's scoring. Despite the calculated correlation throughout all leagues and seasons indicating a slight or even no existing negative dependence for competing teams, the estimated θ copula parameter does not have a similar behaviour.

These inconsistent results are probably due to two main contributing factors. First is the fact that the parameter θ captures the overall variation of the scoring patterns between two opposing teams in a league without accounting for their different abilities. Since we have not added covariates (for reasons discussed above), the parameter θ represents the dependence parameter common for all teams, resulting in the troubling estimates we have seen. This is a problem we have addressed repeatedly in this thesis across many of the cases we examined. The other reason is the fact that the correlation is very little to start with, and therefore, the region where we are trying to estimate is essentially flat, meaning optim "jumps" around positive and negative values.

3.9.2 Handling Absolute Chaos

Up to this point, we have studied several forms of dependence. In Section 3.8, we analysed and modelled the correlation between the two halves of football matches using score differences. Subsequently, in Subsection 3.9.1, we studied the dependence between the scoring patterns of the two opposing teams. All of these models were formulated using copulas.

In this section, we combine the ideas mentioned above. We model each team's abilities using a Poisson distribution, as in Section 3.4, but now separately for each half. Subsequently, we capture the dependencies arising from all possible interactions through a 4-variate copula, which is presented in Section A of the Appendix.

Specifically, we aim to investigate how, for example, the home team's scoring in the first half affects its scoring in the second half, as well as the away team's scoring

in either the first or second half, and vice versa. The same applies to all other possible combinations.

For clarity, let H_j denote the number of goals scored by the home team and A_j the number scored by the away team, where $j = 1, 2$ correspond to the first and the second half, respectively. Thus, for each team's performance in a given half, we estimate its λ parameter as, as

$$\log(\lambda_{h,i,j}) = \gamma + att_{h_{(i,j)}} + def_{a_{(i,j)}}$$
$$\log(\lambda_{\alpha,i,j}) = \gamma + att_{a_{(i,j)}} + def_{h_{(i.j)}}$$

just as we did in the previous sections. In addition to this specification, we wish to capture all potential interactions that could arise among these four variables.

The concept of bivariate modelling has already been introduced in the previous sections, as its foundation has been section 2.10, where copulas were briefly presented. We now extend this framework to the case of 4 variables using the Gaussian copula based on equation (2.15) (see section A of the appendix for further details). By analogy, this setting requires 6 different dependence parameters to capture all possible interactions between 4 variables.

To further illustrate this complexity, we computed the cross-correlation of the average scoring rates in the first and second halves for both home and away teams. Figure 3.15 below shows the correlation structure between first-half and second-half goals across the major European leagues for the 2020–2025 seasons.

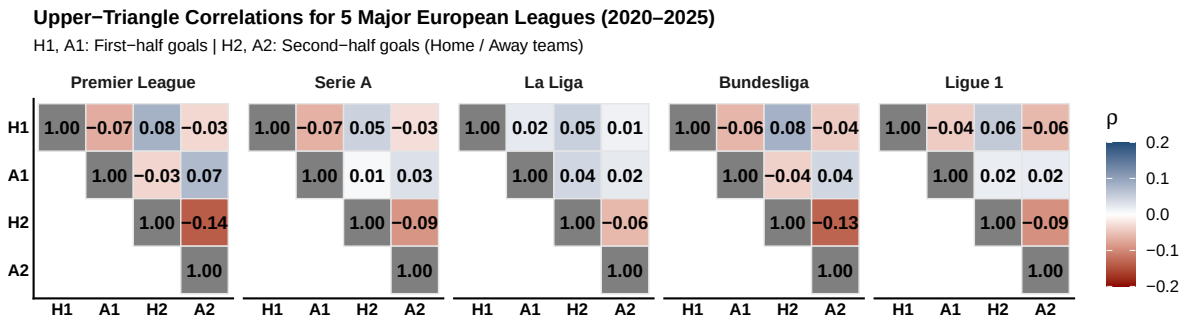


Figure 3.14: Correlations between first- and second-half goals across major European leagues for 2020-20215 seasons. H1/A1 and H2/A2 denote home/away goals in the first and second half, respectively.

The first observation that immediately stands out is that, in every league, the strongest relationship seems to be the average goals scored by the home and away teams in the second half, and this association is negative. Additionally, teams generally show a positive relationship in their performance between the first and second halves.

The overall pattern suggests that teams exhibit a positive correlation between their performances across the two halves, while still showing a negative correlation relative

to their opponents. To further examine these interactions, we also computed the same correlations separately for each league and each season. Figure 3.15 below presents these correlations for the Premier League for every season in the 2020–205 period.

The overall pattern remains the same each season, and this also holds true across the other leagues, as is shown in Figure B.10 of the appendix. These dependencies, however, appear to be minor in many instances and require further investigation, which can be precisely achieved through modelling.

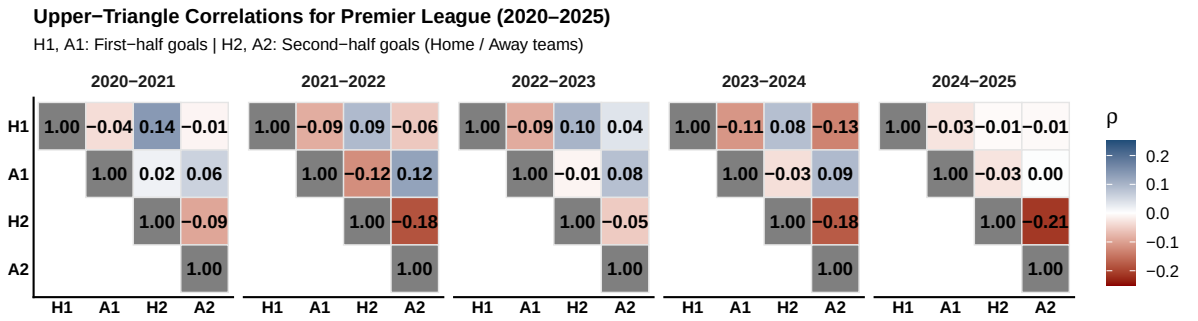


Figure 3.15: Correlations between first- and second-half goals in the Premier League by season (2020–2025). H1/A1 and H2/A2 denote home/away goals in the first and second half.

We rely on the model specification described above and apply equation (A.1) of the Appendix to estimate the model via the R function optim. As initial values, we use the parameter estimates obtained by fitting a double Poisson to each half separately via the glm function in R, which together yield a four-dimensional Poisson model overall. In Table 3.15 below, we include all the aforementioned estimations along with fitted probabilities and actual outcomes, as we have previously done in many different models. We have also included a double Poisson model to serve as a baseline reference. Similar tables for the remaining championships can be found in the Appendix.

Based on the results of 3.15, we infer that, despite the fancy modelling and complicated pmf (see Appendix, section A), the model offers no substantial improvement over the 4-variate Poisson model, or even over the standard Double Poisson model when applied to the Premier League data. The estimated copula parameters do not have a constant behaviour as they differ in value from season to season, which is an indication of somewhat problematic fitting. As we have already discussed multiple times, this might be due to using scalar variables to account for dependencies without accounting for the different team strengths and how these affect the interactions between them. Adding covariates to an already complicated model would make it overparameterized and very difficult to interpret the estimations.

One additional drawback is the heavy computations required with the copula setting. Based on equation A.1 of the appendix, for each coupled probability, we wish to extract; we have to calculate 16 different integrals. For that matter, both the estimation and fitting of the model required some time, making it a bit impractical, but not forbidden

to use. If the reader is observant enough, would notice that the probabilities (or fitted probabilities from the copula model) do not sum to 380, which is the total number of matches played in the season. The reason is that, for the maximum goals scored in a half, we assumed the value to be 4, which did not cover the full range of probabilities and therefore led to the observed deficiency.

Premier League

Season	Model	Df	LogLik	Fit			Dependence Parameters					
				Home	Draw	Away	r_{12}	r_{13}	r_{14}	r_{23}	r_{24}	r_{34}
2020–2021	Actual			146.00	84.00	150.00						
	DP	40	-1074.59	145.14	92.15	142.71						
	4P	156	-1504.86	144.22	91.53	144.25						
	4CP	162	-1500.39	140.32	94.31	141.97	0.03	0.10	-0.04	0.19	0.04	0.01
2021–2022	Actual			162.00	88.00	130.00						
	DP	40	-1059.38	159.67	85.34	134.99						
	4P	156	-1516.61	159.67	84.60	135.72						
	4CP	162	-1511.86	159.47	83.19	132.20	0.01	-0.19	0.03	0.06	-0.01	-0.17
2022–2023	Actual			180.00	90.00	110.00						
	DP	40	-1079.43	174.87	88.23	116.89						
	4P	156	-1533.26	174.02	87.72	118.24						
	4CP	162	-1529.72	169.90	90.09	114.84	-0.04	0.02	0.17	0.05	0.06	-0.02
2023–2024	Actual			174.00	83.00	123.00						
	DP	40	-1121.17	171.89	79.42	128.69						
	4P	156	-1633.03	174.20	78.41	127.37						
	4CP	162	-1629.56	170.54	74.74	127.72	-0.03	-0.01	-0.10	0.05	-0.03	-0.15
2024–2025	Actual			152.00	96.00	132.00						
	DP	40	-1081.79	151.00	86.58	142.42						
	4P	156	-1568.20	152.38	86.73	140.88						
	4CP	162	-1557.51	150.71	84.30	140.70	0.00	-0.10	0.08	0.04	-0.16	-0.28

Table 3.15: Model comparison results for the Premier League across the 2020–2025 seasons. The table reports degrees of freedom (Df), log-likelihood values, empirical and fitted match outcome frequencies (Home, Draw, Away), and pairwise dependence parameters for the 4-variate copula Poisson model.

Such a model setting, however, which is estimated through halves, offers a more detailed estimation of the turnover of the game. As we have already seen in section 3.8.2, we can approximate the first and second half results, which are displayed in Table 3.16, below, or the similar Tables of the Appendix for the rest of the leagues.

With this model, we conclude the chapter on modelling. We implemented, applied, and generally discussed various approaches on how to model football data. Their flaws, weaknesses, and limitations were also examined, and we tried to mitigate them with various methods, but many times unsuccessfully.

We presented in detail the idea of applying distributions to estimate outcomes, and we explained the intuition behind it. We also briefly discussed the general key elements behind modelling in the opening sections of this chapter.

Premier League: Joint Half-Time / Full-Time Results

Season	Actual			4-Variate Poisson			4-Variate Copula Poisson		
2020–2021	93	16	10	90.52	18.40	9.59	89.72	17.76	8.65
	43	57	58	47.07	59.04	51.05	43.41	61.79	53.06
	10	11	82	6.62	14.09	83.61	7.19	14.76	80.26
2021–2022	93	24	5	97.85	14.79	6.61	91.71	16.94	8.12
	60	49	46	54.03	54.87	45.20	57.73	49.58	46.28
	9	15	79	7.79	14.94	83.92	10.03	16.67	77.80
2022–2023	112	16	6	110.44	15.62	6.35	106.46	17.10	7.26
	53	55	41	55.22	56.55	41.41	54.59	56.58	38.09
	15	19	63	8.36	15.55	70.48	8.85	16.40	69.49
2023–2024	96	16	13	102.42	15.74	8.07	100.21	16.08	8.32
	63	48	43	59.33	45.13	42.59	56.97	40.68	45.30
	15	19	67	12.45	17.54	76.72	13.37	17.99	74.09
2024–2025	95	30	15	100.25	19.28	9.97	91.42	22.19	13.74
	45	47	47	45.71	53.21	49.91	50.13	45.15	52.92
	12	19	70	6.41	14.25	81.00	9.17	16.96	74.04

Table 3.16: Observed and model-implied joint half-time/full-time result matrices for the Premier League seasons 2020–2025. Each cell contains a 3×3 matrix of transitions from half-time (rows: Home, Draw, Away) to full-time outcomes (columns: Home, Draw, Away). The two model-based columns correspond to the Independent and Frank Copula Zero-Inflated Skellam2 specifications.

In the next chapter, we apply the knowledge gained, or more simply put, the models discussed in this chapter, to actually predict, or more delicately expressed, estimate the remaining of the current season (2025-2026) for the big 5 of European football. That is actually something not to miss out.

Chapter 4

Who will be first, who will be last?

And now we are talking. This chapter brings us to what people like to call the million-dollar question. This is likely the only chapter that may be read by anyone other than the very kind professors who will evaluate this thesis. Only if the findings presented in this chapter turn out to be accurate (we will find out in a few months after this is written), and someone else actually reads them, will they perhaps go back to the earlier chapters to explore the methods we used.

This is the chapter where we actually apply statistics for what it was made for: estimate final outcomes based on current circumstances, tackle uncertainty, tell the future. This is precisely the objective of the present chapter. Here, we report the estimations (or predictions) of selected models from chapter 3 for the final standings of the major 5 European championships.

We have collected data up until the first game week of January for each league: Premier League, Serie A, La Liga, Bundesliga, and Ligue 1. We report the fitting results from a double Poisson model (see section 3.4), a zero-inflated bivariate copula Skellam2 model (see section 3.8.2), and a 4-variate Poisson (see section 3.9.2). Each table presented below is essentially the same as those repeatedly encountered in Chapter 3. It reports, for each model, the number of parameters, the estimated log-likelihood, the predicted outcomes (total home wins, draws, and away wins) alongside the actual results, and, lastly, the estimated parameters for the ZI copula bivariate Skellam2. We make it clear that the data used for the estimation is up until the point of the first game week of January for each championship (total matches played differ a bit from league to league).

The two Poisson models were fitted with the R function glm, and the Skelam2 model was obtained using the R function optim, based on carefully chosen initial values. More precisely, each component was estimated separately: first, the team coefficients, then the inflation parameter (for both halves), and finally the Frank copula parameter.

Most importantly, we have included the predictions produced by the simulation for each remaining league, generated 10,000 times according to the model estimation described above. Below, we have divided everything into individual sections, each dedicated to a specific league. These include a "heatmap" of probabilities displaying the final estimated ranking produced by each model and the corresponding total goal prediction. At this point, we remind that Skellam2 models score difference, and to extract goals estimation, further assumptions are required to be made.

4.1 Premier League

Premier League (2025–2026)

Model	Df	LogLik	Fit			ZI-Copula Parameters				
			Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ
Actual			88.00	52.00	61.00					
Double Poisson	78	-532.53	88.36	48.48	64.16					
ZI fcop Skellam2	82	-1164.96	87.69	47.70	65.62	1.34	2.10	0.01	0.13	-1.15
4-Variate Poisson	156	-763.92	88.59	48.44	63.97					

Table 4.1: Model comparison results for the Premier League, season 2025–2026. Data are up to the first gameweek of January for that season. All models seem to yield very good results for the stage of the league. The models are the double Poisson discussed in section 3.4, the zero inflated frank copula skellam 3.8.2 and the 4-variate Poisson discussed in 3.9.2

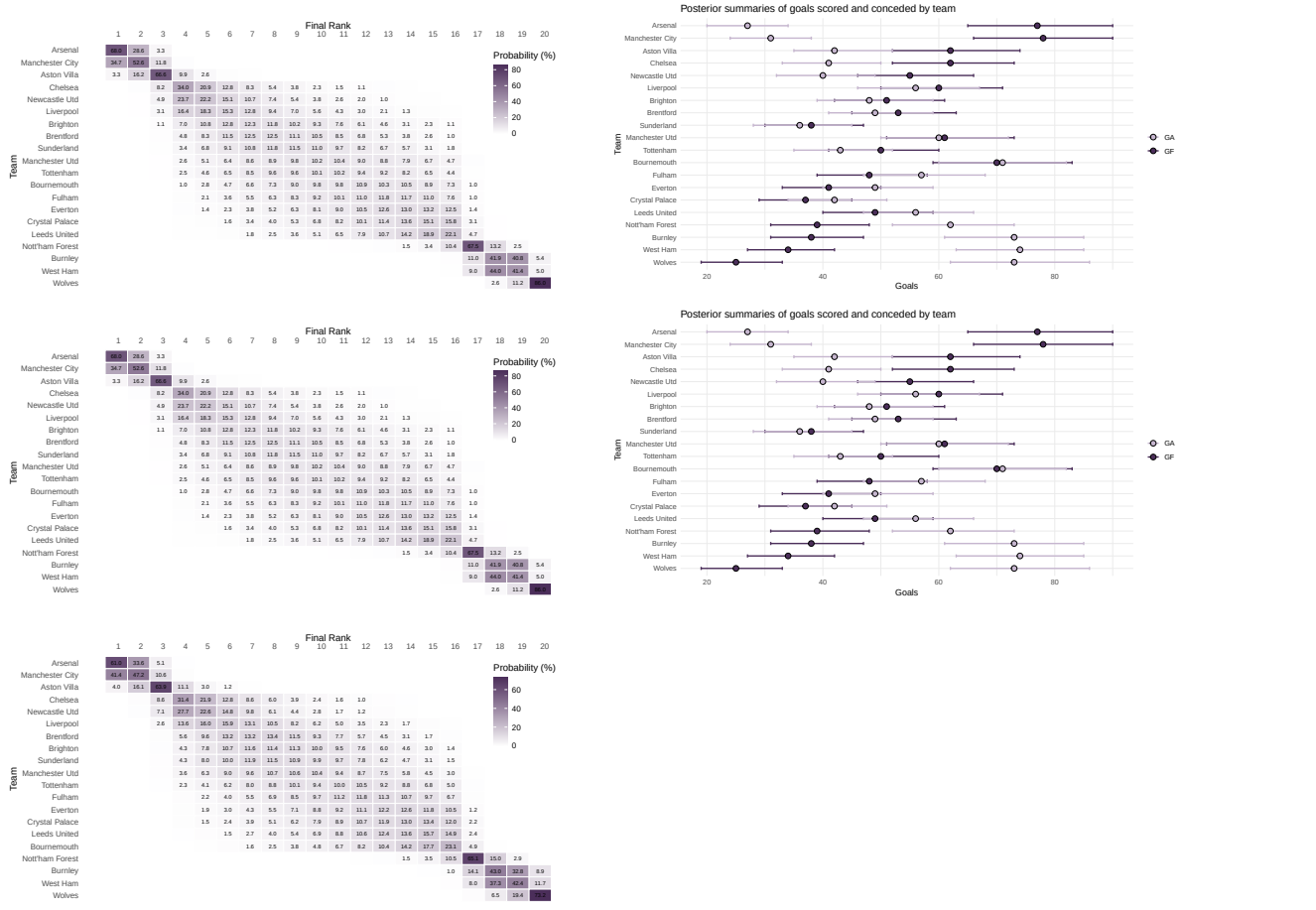


Figure 4.1: Simulated Premier League predictions under different models. Rows correspond to model classes (top to bottom: Double Poisson, 4-Variate Poisson, ZI Frank Copula Skellam2), while columns show final league rankings (left) and total goals distributions (right). The three models produce very similar results; now we just have to wait and see whether they will prove to be accurate as well.

4.2 Serie A

Serie A (2025–2026)

Model	Df	LogLik	Fit			ZI-Copula Parameters				
			Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ
Actual			73.00	61.00	62.00					
Double Poisson	78	-486.71	76.06	52.29	67.65					
ZI fcop Skellam2	82	-1074.68	77.16	50.58	68.26	0.97	1.84	0.04	0.11	-0.27
4-Variate Poisson	156	-689.76	76.03	52.40	67.56					

Table 4.2: Model comparison results for Serie A, season 2025–2026. Data are up to the first gameweek of January for that season. All models seem to yield good results for this stage of the league, even though a little draw inflation is observed. The models are the double Poisson discussed in section 3.4, the zero inflated frank copula skellam 3.8.2 and the 4-variate Poisson discussed in 3.9.2

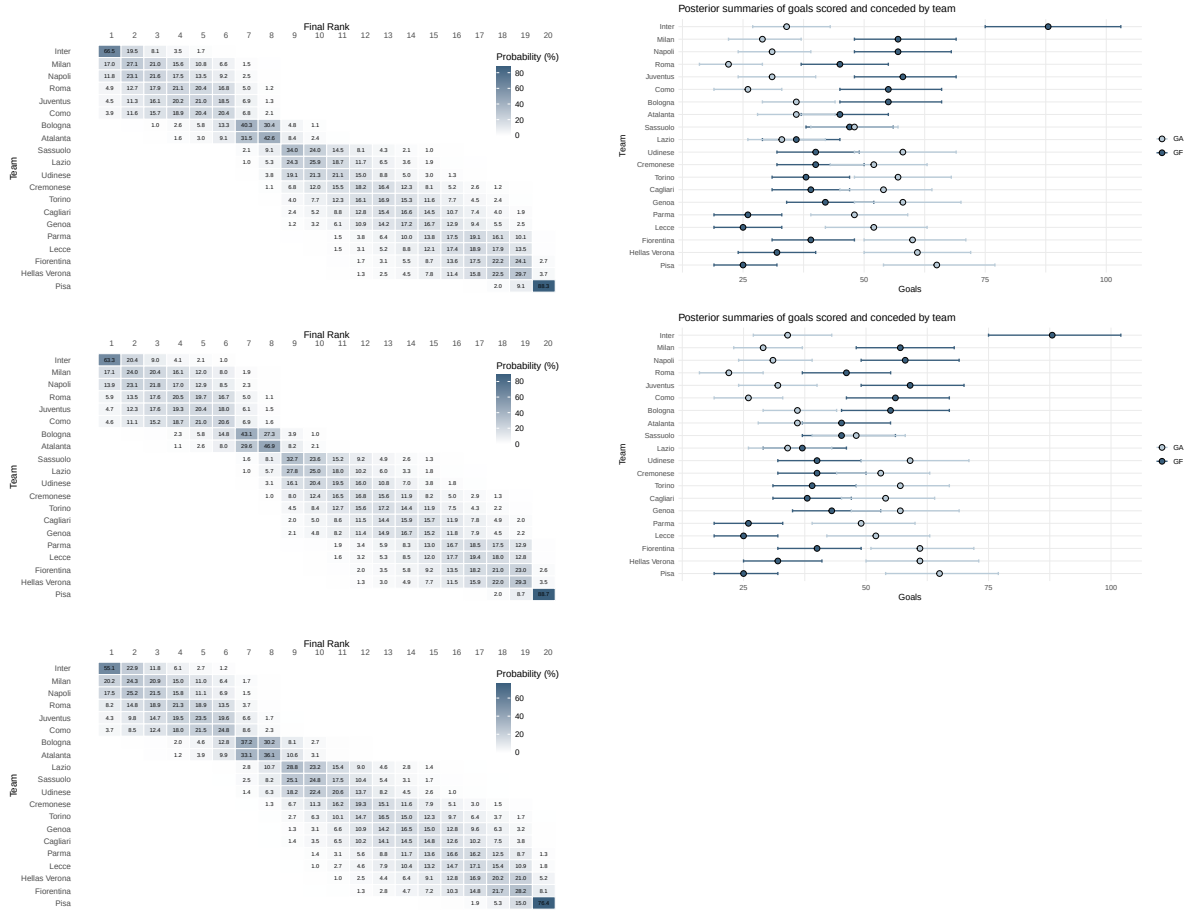


Figure 4.2: Simulated Serie A predictions under different models. Rows correspond to model classes (top to bottom: Double Poisson, 4-Variate Poisson, ZI Frank Copula Skellam2), while columns show final league rankings (left) and total goals distributions (right). The three models produce very similar results; now we just have to wait and see whether they will prove to be accurate as well.

4.3 La Liga

La Liga (2025–2026)

Model	Df	LogLik	Fit			ZI-Copula Parameters				
			Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ
Actual			88.00	52.00	49.00					
Double Poisson	78	-471.19	85.75	46.93	56.31					
ZI fcop Skellam2	82	-1100.72	86.67	43.24	59.09	1.59	2.21	0.11	0.08	-1.46
4-Variate Poisson	156	-678.24	85.58	46.97	56.45					

Table 4.3: Model comparison results for La Liga, season 2025–2026. Data are up to the first gameweek of January for that season. All models seem to yield very good results for this stage of the league. The models are the double Poisson discussed in section 3.4, the zero inflated frank copula skellam 3.8.2 and the 4-variate Poisson discussed in 3.9.2

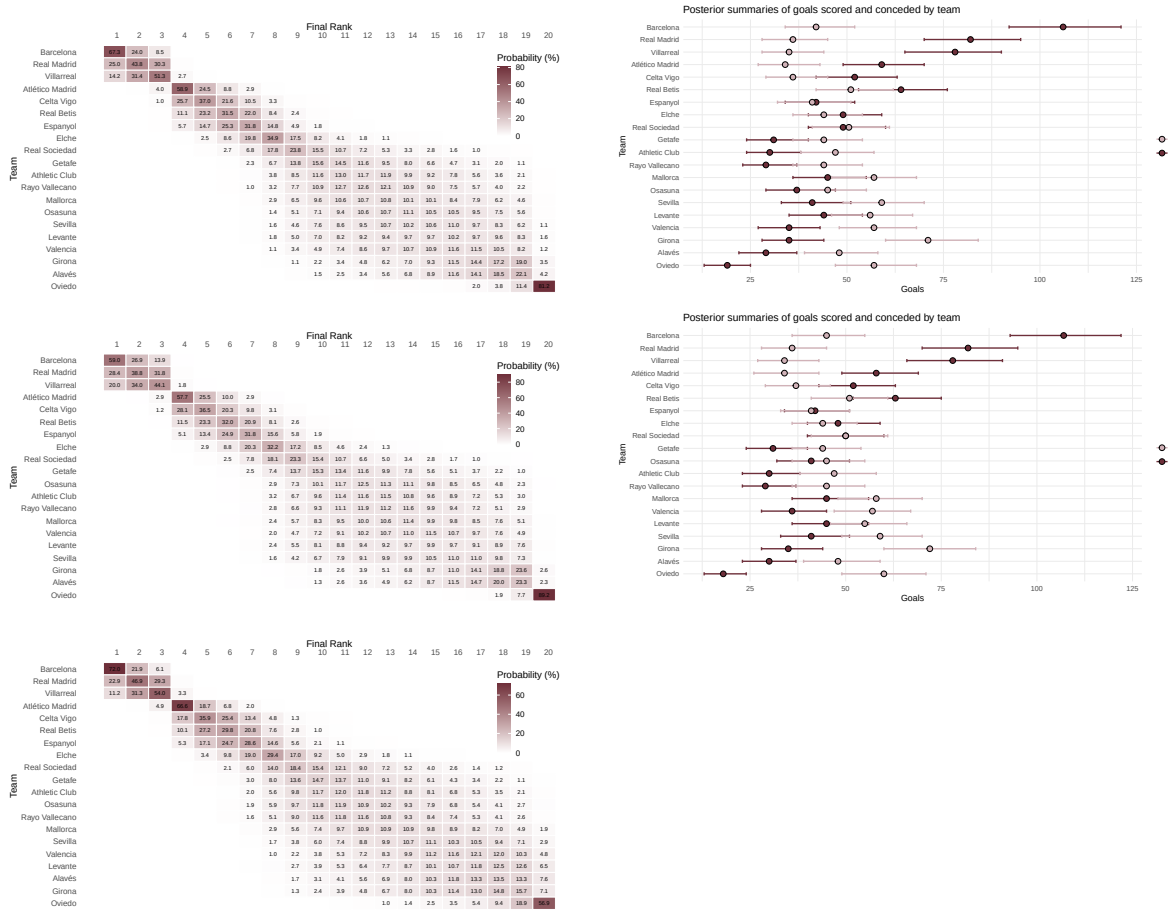


Figure 4.3: Simulated La Liga predictions under different models. Rows correspond to model classes (top to bottom: Double Poisson, 4-Variate Poisson, ZI Frank Copula Skellam2), while columns show final league rankings (left) and total goals distributions (right). The three models produce very similar results; now we just have to wait and see whether they will prove to be accurate as well.

4.4 Bundesliga

Bundesliga (2025–2026)

Model	Df	LogLik	Fit			ZI-Copula Parameters				
			Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ
Actual			64.00	34.00	44.00					
Double Poisson	70	-412.48	63.99	29.04	48.97					
ZI fcop Skellam2	74	-902.50	62.20	28.67	51.13	2.31	2.58	0.15	0.16	0.14
4-Variate Poisson	140	-587.14	63.96	29.06	48.96					

Table 4.4: Model comparison results for the Bundesliga, season 2025–2026. Data are up to the first gameweek of January for that season. All models seem to yield very good results for this stage of the league. The models are the double Poisson discussed in section 3.4, the zero inflated frank copula skellam 3.8.2 and the 4-variate Poisson discussed in 3.9.2

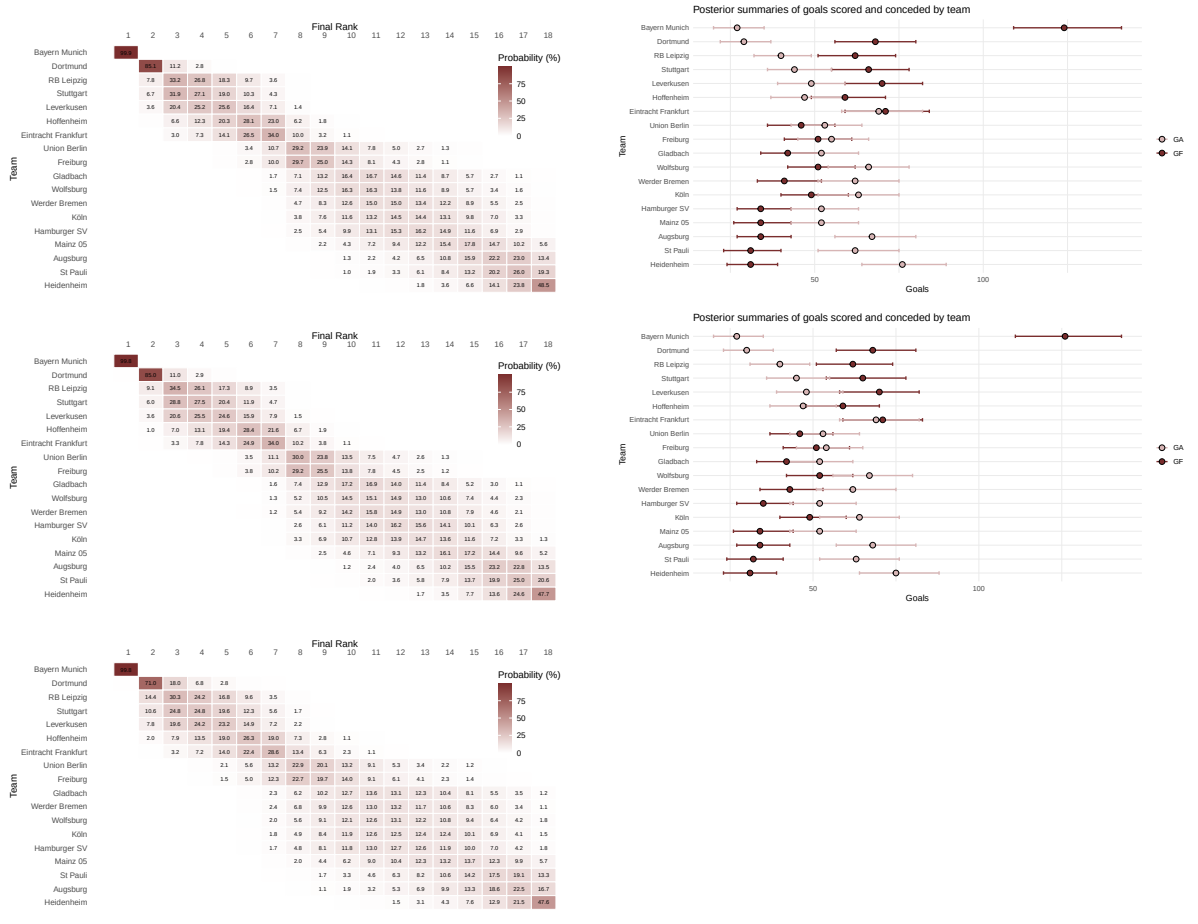


Figure 4.4: Simulated Bundesliga predictions under different models. Rows correspond to model classes (top to bottom: Double Poisson, 4-Variate Poisson, ZI Frank Copula Skellam2), while columns show final league rankings (left) and total goals distributions (right). The three models produce very similar results; now we just have to wait and see whether they will prove to be accurate as well.

4.5 Ligue 1

Ligue 1 (2025–2026)

Model	Df	LogLik	Fit			ZI-Copula Parameters				
			Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ
Actual			76.00	35.00	42.00					
Double Poisson	70	-418.65	74.16	34.52	44.32					
ZI fcop Skellam2	74	-883.84	73.50	30.47	49.02	1.27	1.99	0.04	0.00	0.22
4-Variate Poisson	140	-581.71	74.29	34.41	44.29					

Table 4.5: Model comparison results for Ligue 1, season 2025–2026. Data are up to the first gameweek of January for that season. All models seem to yield very good results for this stage of the league. The models are the double Poisson discussed in section 3.4, the zero inflated frank copula skellam 3.8.2 and the 4-variate Poisson discussed in 3.9.2

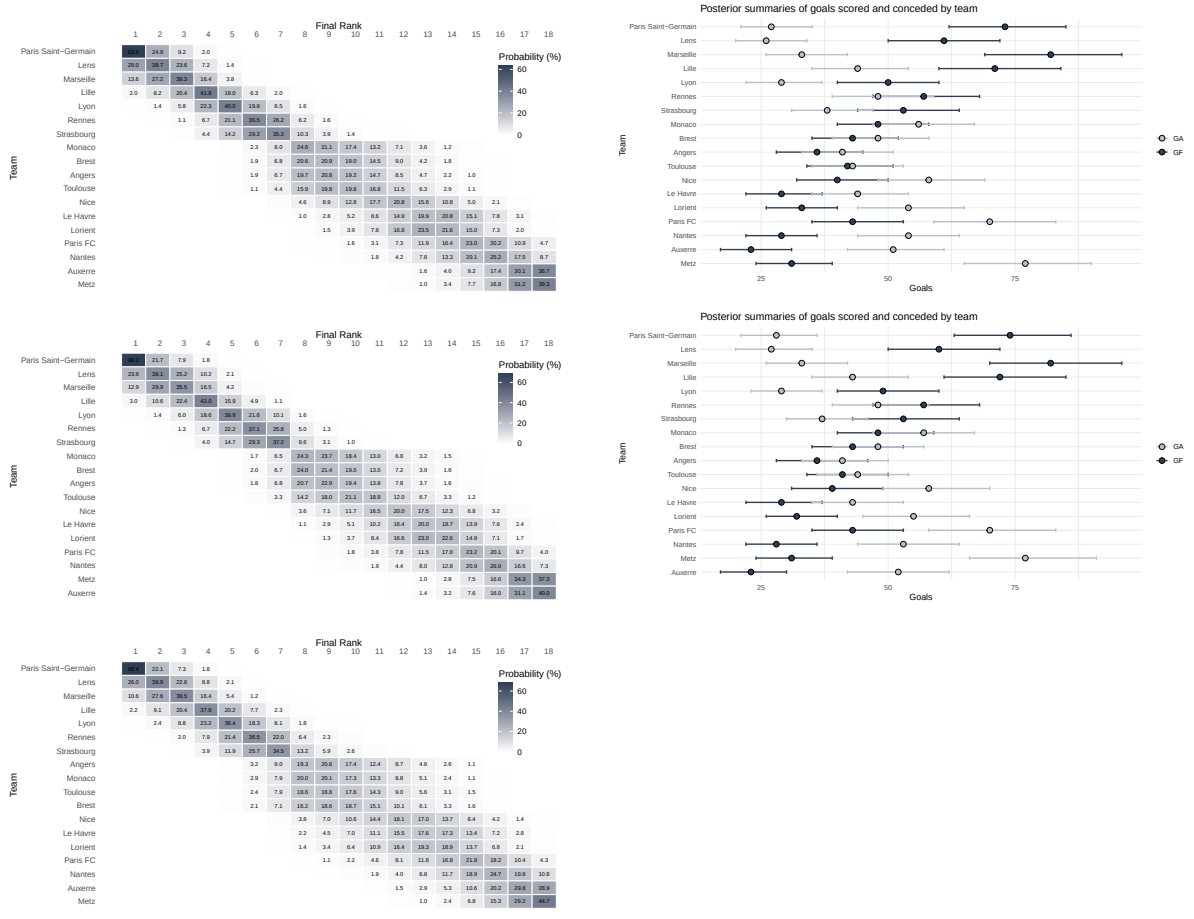


Figure 4.5: Simulated Ligue 1 predictions under different models. Rows correspond to model classes (top to bottom: Double Poisson, 4-Variate Poisson, ZI Frank Copula Skellam2), while columns show final league rankings (left) and total goals distributions (right). The three models produce very similar results; now we just have to wait and see whether they will prove to be accurate as well.

4.6 Conclusion

Surprisingly, or maybe not so surprisingly, but a rather welcoming ascertainment is that the three models yield almost identical predictions. This makes us confident that we are not entirely wrong, and actually, these predictions might turn out to be close to what we are going to see around May, barring, of course, any extreme unforeseen events.

After many years of suffering, Arsenal is expected to finally win the championship for the 2025-2026 season and put an end to the teasing for always finishing 2^{nd} or 4^{th}. On the bottom side of the table are Burnley, West Ham, and Wolves, with very few chances of escaping the relegation zone.

In Serie A, Inter seems to keep up the good pace that it has built in recent years and is predicted to win the championship this season. Napoli, on the other hand, won the championship the previous year, but is now estimated to finish 3^{rd}. Pisa appears almost certain to finish bottom of the table and has no chance of escaping relegation. The other two places that lead to Serie B are uncertain, and it is predicted that it will be a contest among many teams to avoid the unfortunate event.

The three models, despite predicting the same champion, Barcelona, differ noticeably in the estimated probability of achieving that (we, of course, are now examining the results of La Liga). They also differ (but not significantly) in the ranking of teams finishing in the lower middle of the table. All three predict Oviedo finishing in last place, but the Skellam model estimates a lower probability compared to the other two. Just like in Serie A, La Liga is also shaping up to have an interesting relegation battle, with many teams clustered near the bottom of the table.

In case Bayern Munich does not win the championship for the season 2025-2026, I dare to say that I should not be awarded the Statistics degree. All models agree on the same outcome and also agree on the astonishing probability of $99.8 - 99.9\%$. On the other side, the relegation zone seems to have great variation and uncertainty, as many teams are placed in that zone with close, each having very similar chances to one another.

The Ligue 1 results produce an interesting picture. All teams appear to be evenly distributed along the table's diagonal, indicating that the projected final rankings involve relatively low uncertainty. Paris Saint-Germain has very good chances of winning the championship, while Auxerre and Metz seem to be almost certainly relegated.

Similar conclusions can be made for every team in any league: given the objectives set at the beginning of the season, one can assess whether they are currently on course to meet those objectives or not. In other words, these predictions also provide a measure of evaluation for each team. Club analysts may run such simulations to determine whether the team is progressing as planned—implying that no major changes are needed—or whether it is underperforming relative to its targets, in which case adjustments must be made. Moreover, given the estimates produced by the models (discussed in detail in section 3.4), they can identify whether the team's main strengths or weaknesses lie in attack or in defence.

Additionally, each individual and fan of the game of football can observe their favourite team's final ranking. Helping to maintain the psychological stability so that anyone can be prepared for either celebration or the undesired events. Jokes aside, with such simulations, we can provide accurate predictions for the final ranking, as has been

reported many times in the literature. And we can also estimate outcomes (and goals scored) on short notice, for example, each next game week's matches for any league, given that we have all the previously played (or just the last 5 games played).

The plots we saw for each league also provide, in an illustrative and intuitive way, the dynamic distribution for each league. In the Premier League, for example, we can observe that the top 3 teams and bottom 3 teams are separated, while the rest are evenly clustered around the mid-table. In La Liga, similarly top 3 teams are close together, a few next fall a bit behind, and after mid-table, the uncertainty grows, meaning that teams are closely matched. In the Bundesliga, Bayern's dominance is quite apparent. On the other hand, in Serie A, there is no such clear separation of teams, indicating an overall competitive league. Finally, as mentioned earlier, in Ligue 1 it appears that each team has its own clear place in the standings, with little overlap compared to the other championships.

Lastly, we have to clarify and make explicit that the results above provide an estimation and an evaluation of what is likely to happen. To train each model, we have used all the games played so far this season, and to simulate the rest of the league, we have assumed that each team's performance will remain the same. One should keep in mind all these when interpreting the results, and should be extra careful if attempting sports betting.

Chapter 5

Discussion - The end

Everything good at some point comes to an end, so does this thesis. After some mediocre writing and a somewhat rough semester, we come to the point of concluding the entire work we have done so far. We have to painfully admit that, eventually, we did not succeed in bringing order to the chaos of football, but at least we are inclined towards this direction.

We examined in detail how the Poisson distribution can be applied to football data. First, we presented the simplest version, which is the double Poisson model. We analysed its limitations, which are the independence assumption, the inflation of draws, and the issue of overdispersion. For the last of these, overdispersion, we found that it does not occur in any of the seasons we studied. This is due to the fact that whenever we attempted to fit the double negative binomial model, it always collapsed to a double Poisson model. Considering the size of the sample we had, it would be safe to say that the incidence of overdispersion in football is no longer occurring, at least in the highest level that is played, which is the 5 Major European leagues, the data we had throughout this thesis.

We then discussed the well-established in the literature home effect, as is estimated through the double Poisson. Based on these estimates, we observed that the home effect is not fixed in value and that it varies depending on both league and season. The key insight from the data we had available is that home advantage was very small during the 2020–2021 pandemic season. It peaked in the 2022–2023 season across all leagues, then began to decline in each of them. It will be interesting to see what happens in the next seasons and whether home advantage will disappear entirely or bounce back.

We then explored the alternative of the compound Poisson model. This approach offers a more detailed evaluation for whatever metric is of interest. Shot and Shots on target were used as an example; their limitations were discussed, and we found out that the fitting of each league yielded almost the same results with the double Poisson model (which was used for goals), as was expected.

We then took a different approach and constructed a model to directly estimate the score difference instead of individual score rates. For that matter, we used the Skellam distribution, along with its zero-inflated counterpart and a reparameterized formulation. We found that, while this model yielded a better fit in some cases, it was less stable than the Poisson model and, overall, did not manage to outperform it. We further advanced the modelling process by adding a Frank copula to account for the dependence between the two halves, but it was not greatly improved either.

We then revisited the Poisson distribution and tried to extend the double Poisson we had previously discussed. We first examined the dependence between opposing teams and realised that the resulting estimates lacked stability and likely relied strongly on the covariates, meaning teams playing. We further advanced the model complexity by allowing all possible dependencies between the scoring rates in each half for both teams. However, we found that this extension did not offer much improvement over the independent 4-variate Poisson model.

Finally, based on the models, the double Poisson, the 4-variate Poisson (2 teams, 2 halves), and the zero-inflated parametrized Skellam with a Frank copula model, we made estimations for the current season (2025-2026) across the 5 Major European Leagues. Predictions essentially for the final standings table as of around May (with this being written in January), along with the goal-scoring estimates produced by the Poisson models.

And that's all it was. Essentially, we reproduced the well-established double Poisson model and tried to extend it by adding various dependency structures that may exist, but without noticeable improvement. Additionally, we explored the Skellam distribution, which did not manage to steadily surpass the standard approaches. It offered, however, a very agile and flexible model that's worth mentioning.

The diversity of leagues we studied and the multiple seasons we had as a sample allowed us to examine problems and limitations of various model fitting to different data. A reasonable conclusion is that a single model for all leagues is not optimised, and that developing separate models tailored to each league may be a better alternative. Each championship had its unique characteristic. All leagues exhibited, for example, a slight draw inflation expect Premier League when fitting the double Poisson model. The Skellam model seemed to be a better fit for Serie A when compared to the Poisson one. Removing the common home effect for teams significantly improved the double Poisson model fit for La Liga results. These are just a few of all the observations one can make. The main idea is that each league requires different handling in order to achieve optimal results.

There is one constant, however, a suitable model to fit all that steadily yields satisfactory results, and is rather straightforward to estimate: the double Poisson model. It has been the main focus throughout this thesis, and not with a bias, as we have seen that it managed to achieve very good results for all leagues and all seasons. Especially when incorporating different home effects for each team. The good news is that when we extended this to a 4-Variate model to estimate outcomes for each team for each half, yielded almost the same results while offering a more detailed insight into teams' performance (and more betting options, there it is, finally he said it). Therefore, the quote that we can conclude from this thesis is that "simplicity does not harm".

It might have gone unnoticed, but it is rather significant, the fact that throughout all fittings, estimations, and even predictions, we have assumed constant team parameters. Specifically, we have supposed that the teams' performance does not change over the duration of a season. This is a rather simplistic issue that can be resolved by adding time-varying coefficients to each model.

Another limitation we must acknowledge is that our analysis has been confined solely to the Maximum Likelihood Estimation method. While efficient, it might be, when the model parameters and complexity grew, it was not a straightforward process to yield accurate estimations. But rather careful handling was required.

Everything discussed in this thesis represents only a small snapshot of the knowledge and applications available in the literature and familiar to practitioners. As part of this work, we had the opportunity to explore some of the most fundamental applications related to forecasting football outcomes. There are, however, many different extensions and more sophisticated approaches applied and suggested. The outlook for upcoming methods or models appears both promising and intriguing.

Bibliography

Abramowitz, M. and Stegun, I. A. (1974). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, Washington, D.C.

Boyko, R. H., Boyko, A. R., and Boyko, M. G. (2007). Referee bias contributes to home advantage in english premiership football. *Journal of Sports Sciences*, 25(11):1185–1194. PMID: 17654230.

Clarke, S. R. and Norman, J. M. (1995). Home ground advantage of individual clubs in english soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44(4):509–521.

Dawson, P., Dobson, S., Goddard, J., and Wilson, J. (2007). Are football referees really biased and inconsistent? evidence on the incidence of disciplinary sanction in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):231–250.

Dixon, M. J. and Coles, S. G. (2002). 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.

Egidi, L., Karlis, D., and Ntzoufras, I. (2025). *Predictive Modelling for Football Analytics*. Chapman and Hall/CRC, 1st edition.

Garicano, L., Palacios-Huerta, I., and Prendergast, C. (2001). Favoritism under social pressure. Working Paper 8376, National Bureau of Economic Research, Cambridge.

Irwin, J. O. (1937). The frequency distribution of the difference between two independent variates following the same poisson distribution. *Journal of the Royal Statistical Society: Series A*, 100(3):415–416.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Karlis, D. and Mamode Khan, N. (2023). Models for integer data. *Annual Review of Statistics and Its Application*, 10:297–323.

Karlis, D., Michels, R., and Otting, M. (2024). Modelling handball outcomes using univariate and bivariate approaches. *arXiv preprint*.

Karlis, D. and Ntzoufras, I. (2000). On modelling soccer data. *Student*, 3:229–244.

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

Karlis, D. and Ntzoufras, I. (2005). Bivariate poisson and diagonal inflated bivariate poisson regression models in r. *Journal of Statistical Software*, 14(10):1–36.

Karlis, D. and Ntzoufras, I. (2006). Bayesian analysis of the differences of count data. *Statistics in Medicine*, 25(11):1885–1905.

Koopman, S. J., Lit, R., and Lucas, A. (2017). Intraday stochastic volatility in discrete price changes: The dynamic skellam model. *Journal of the American Statistical Association*, 112(520):1490–1503.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lee, A. J. (1997). Modeling scores in the premier league: Is manchester united really the best? *CHANCE*, 10(1):15–19.

Leite, W. and Pollard, R. (2018). International comparison of differences in home advantage between level 1 and level 2 of domestic football leagues. *German Journal of Exercise and Sport Research*, 48(2):271–277.

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

McHale, I. G. and Scarf, P. (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling*, 11:219–236.

Moroney, M. (1951). *Facts from Figures*. Mathematics and statistics. Penguin Books.

Nelsen, R. B. (2010). *An Introduction to Copulas*. Springer.

Nevill, A. M., Balmer, N. J., and Williams, A. M. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*, 3(4):261–272.

Pelechrinis, K. and Winston, W. (2020). A skellam regression model for quantifying positional value in soccer. *arXiv preprint*.

Pishro-Nik, H. (2014). *Introduction to Probability, Statistics, and Random Processes*. Kappa Research LLC.

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of Sports Sciences*, 4(3):237–248. PMID: 2884328.

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

Reep, C., Pollard, R., and Benjamin, B. (1971). Skill and chance in ball games. *Journal of the Royal Statistical Society: Series A*, 134(4):623–629.

Ross, S. M. (1998). *A First Course in Probability*. Prentice Hall, Upper Saddle River, NJ, 5 edition.

Skellam, J. G. (1946). The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society: Series A*, 109(3):296–296.

Sumpter, D. (2016). *Soccermatics: Mathematical Adventures in the Beautiful Game*. Bloomsbury Sigma Series. Bloomsbury Publishing.

Sutter, M. and Kocher, M. G. (2004). Favoritism of agents – the case of referees' home bias. *Journal of Economic Psychology*, 25(4):461–469.

Tomy, L. and Veena, G. (2022). A retrospective study on skellam and related distributions. *Austrian Journal of Statistics*, 51:102–111.

Venables, W. N. and Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York, 1 edition.

Κοντογιάννης, and Τουμπής, (2015). *Στοιχεία πιθανοτήτων*. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. Available online.

Appendix A

4-Variate copula PMF

Let $F_1(x_1)$, $F_2(x_2)$, $F_3(x_3)$ and $F_4(x_4)$ denote the marginal cumulative distribution functions, and let $C(u_1, u_2, u_3, u_4; \boldsymbol{\theta})$ be a four-dimensional copula, where $\boldsymbol{\theta}$ denotes the vector of dependence parameters. The joint probability mass function can then be obtained by applying finite differences to the copula distribution function as

$$
\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = \\
&C\big(F_1(x_1), F_2(x_2), F_3(x_3), F_4(x_4); \boldsymbol{\theta}\big) \\
&\quad - C\big(F_1(x_1-1), F_2(x_2), F_3(x_3), F_4(x_4); \boldsymbol{\theta}\big) \\
&\quad - C\big(F_1(x_1), F_2(x_2-1), F_3(x_3), F_4(x_4); \boldsymbol{\theta}\big) \\
&\quad - C\big(F_1(x_1), F_2(x_2), F_3(x_3-1), F_4(x_4); \boldsymbol{\theta}\big) \\
&\quad - C\big(F_1(x_1), F_2(x_2), F_3(x_3), F_4(x_4-1); \boldsymbol{\theta}\big) \\
&\quad + C\big(F_1(x_1-1), F_2(x_2-1), F_3(x_3), F_4(x_4); \boldsymbol{\theta}\big) \\
&\quad + C\big(F_1(x_1-1), F_2(x_2), F_3(x_3-1), F_4(x_4); \boldsymbol{\theta}\big) \\
&\quad + C\big(F_1(x_1-1), F_2(x_2), F_3(x_3), F_4(x_4-1); \boldsymbol{\theta}\big) \\
&\quad + C\big(F_1(x_1), F_2(x_2-1), F_3(x_3-1), F_4(x_4); \boldsymbol{\theta}\big) \\
&\quad + C\big(F_1(x_1), F_2(x_2-1), F_3(x_3), F_4(x_4-1); \boldsymbol{\theta}\big) \\
&\quad + C\big(F_1(x_1), F_2(x_2), F_3(x_3-1), F_4(x_4-1); \boldsymbol{\theta}\big) \\
&\quad - C\big(F_1(x_1-1), F_2(x_2-1), F_3(x_3-1), F_4(x_4); \boldsymbol{\theta}\big) \\
&\quad - C\big(F_1(x_1-1), F_2(x_2), F_3(x_3-1), F_4(x_4-1); \boldsymbol{\theta}\big) \\
&\quad - C\big(F_1(x_1), F_2(x_2-1), F_3(x_3-1), F_4(x_4-1); \boldsymbol{\theta}\big) \\
&\quad - C\big(F_1(x_1-1), F_2(x_2-1), F_3(x_3), F_4(x_4-1); \boldsymbol{\theta}\big) \\
&\quad + C\big(F_1(x_1-1), F_2(x_2-1), F_3(x_3-1), F_4(x_4-1); \boldsymbol{\theta}\big)\,.
\end{aligned}
\tag{A.1}
$$

where

$$\boldsymbol{\theta} = (\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}, \theta_{24}, \theta_{34})$$

denotes the vector of pairwise dependence parameters.

Appendix B

Figures

B.1 Poisson Assumptions

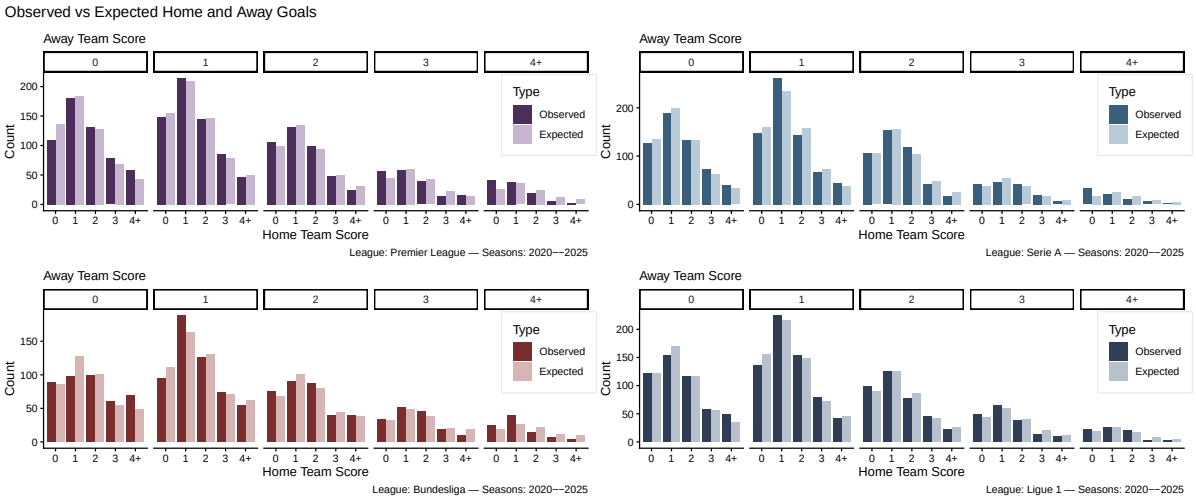


Figure B.1: Observed and model-implied joint distributions of home and away goals across major European leagues (2020–2025).

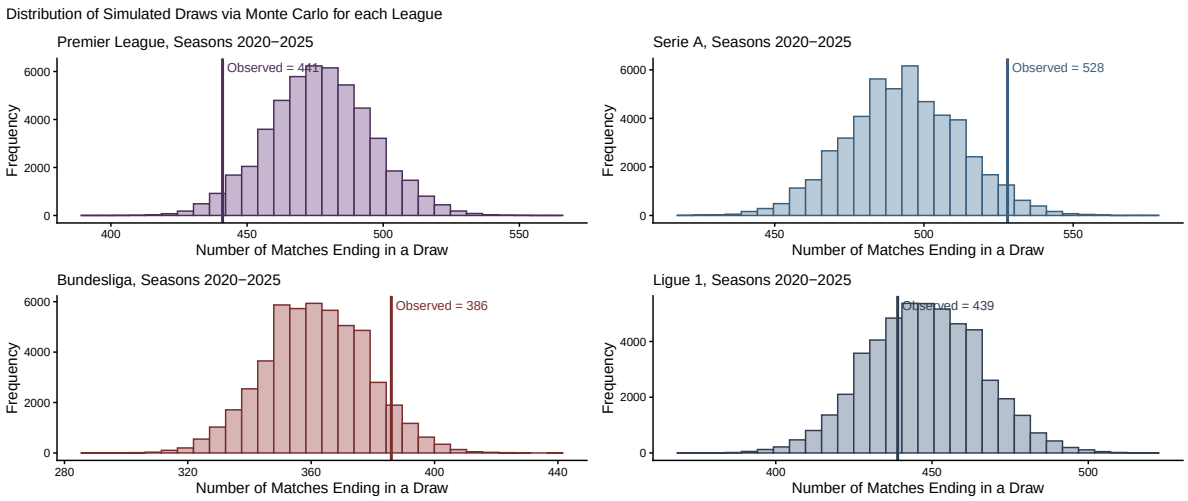


Figure B.2: Histograms of Monte Carlo–simulated draw counts for representative home–away matchups across major European leagues for the 2020-2025 seasons.

B.2 Overdispersion

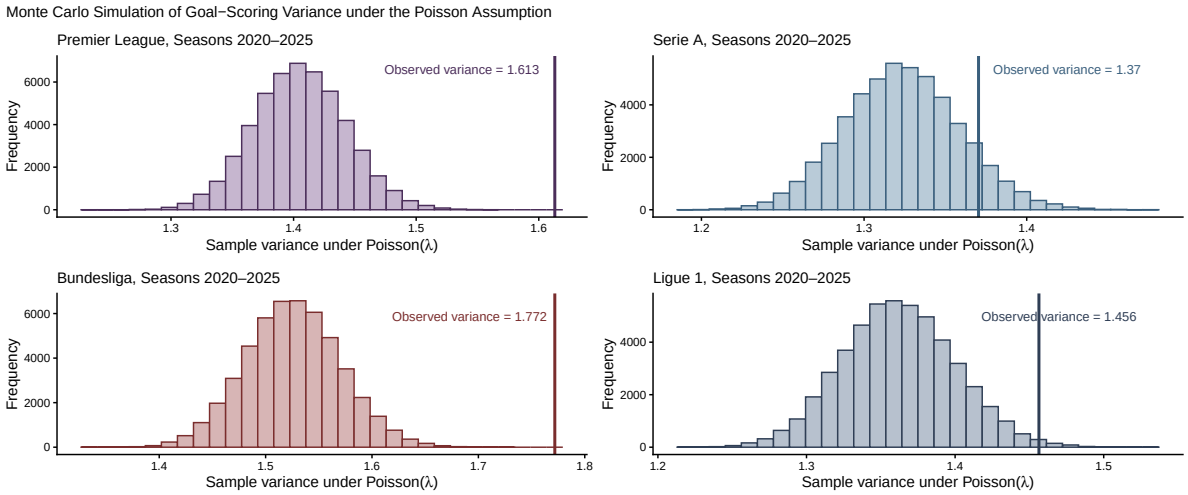


Figure B.3: Simulated distributions of team-level scoring variance for major European leagues over the 2020–2025 seasons.

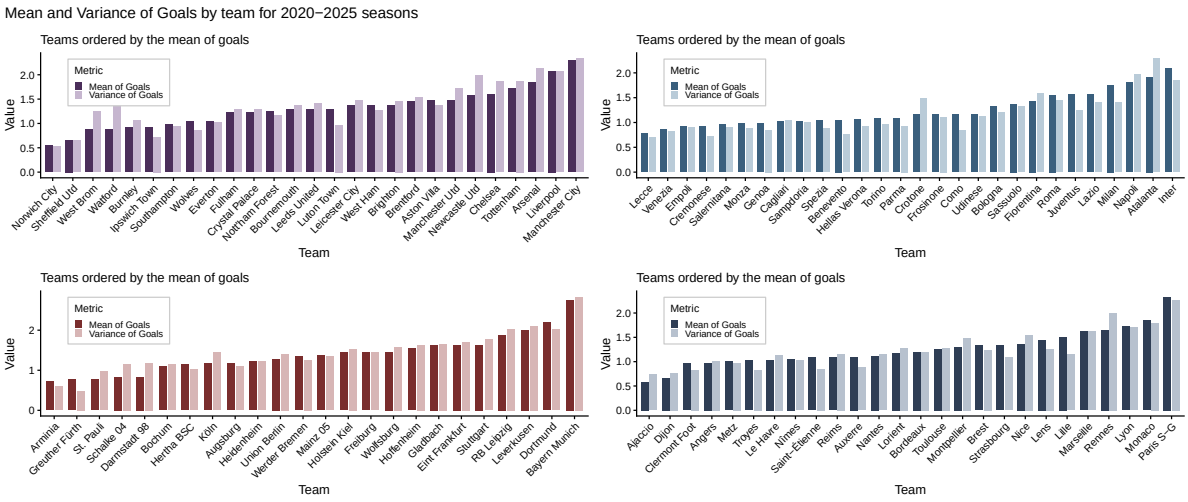


Figure B.4: Mean and Variance of goals scored by each team, in major European leagues during the 2020-2025 seasons.

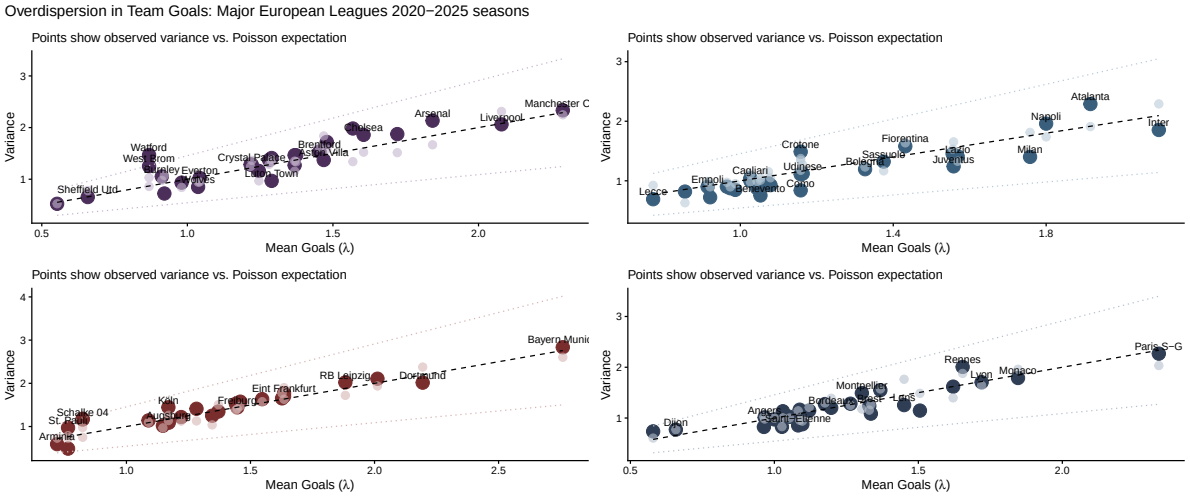


Figure B.5: Scatter plot of team-level mean and variance in goal scoring, with observed and simulated values, for major European leagues (2020–2025).

B.3 Shot and Shot On target Overdispersion

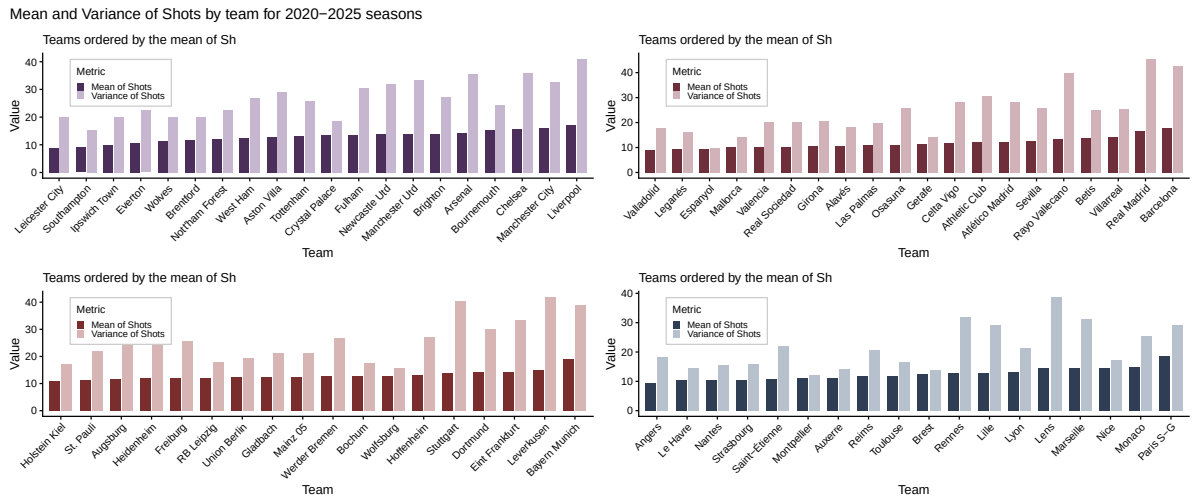


Figure B.6: Team-level mean–variance relationship for shots across major European leagues (2020–2025).

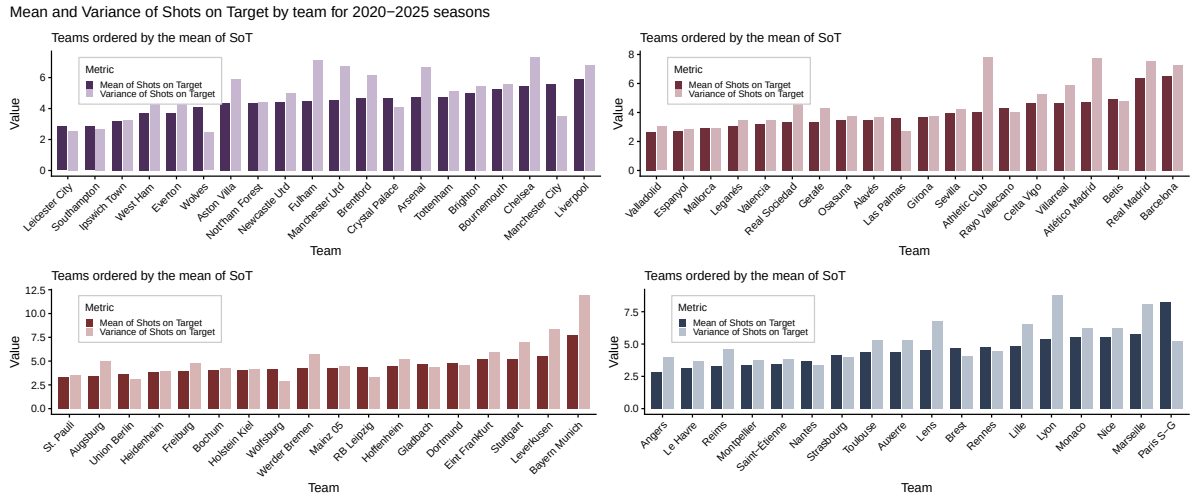
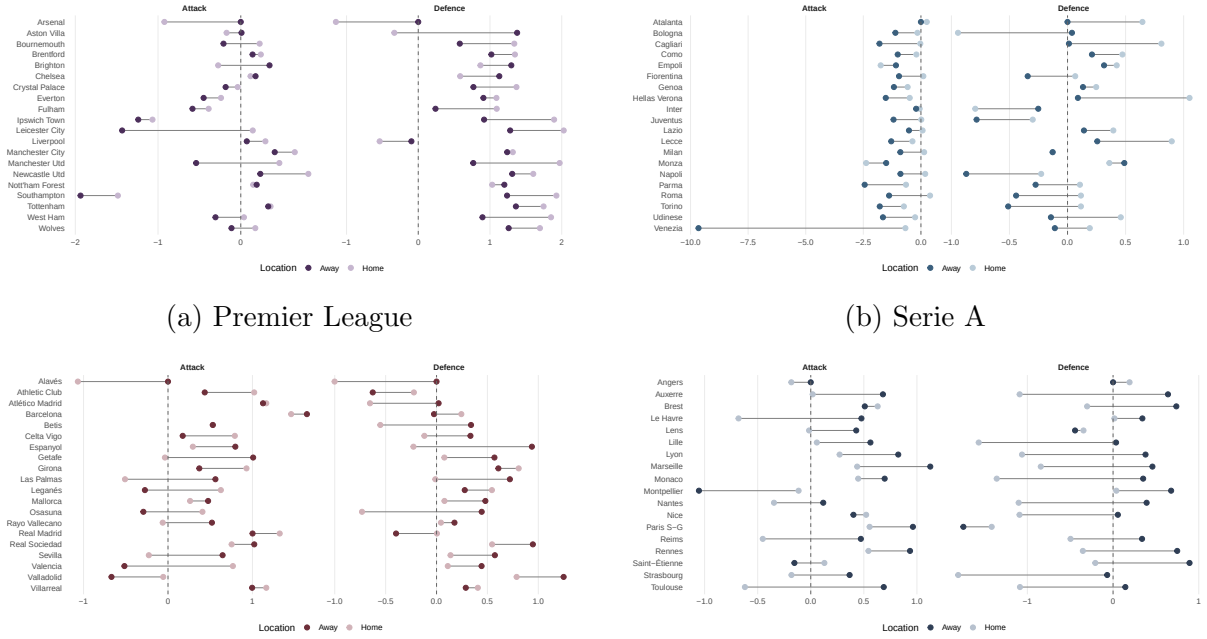


Figure B.7: Team-level mean–variance relationship for shots on target across major European leagues (2020–2025).

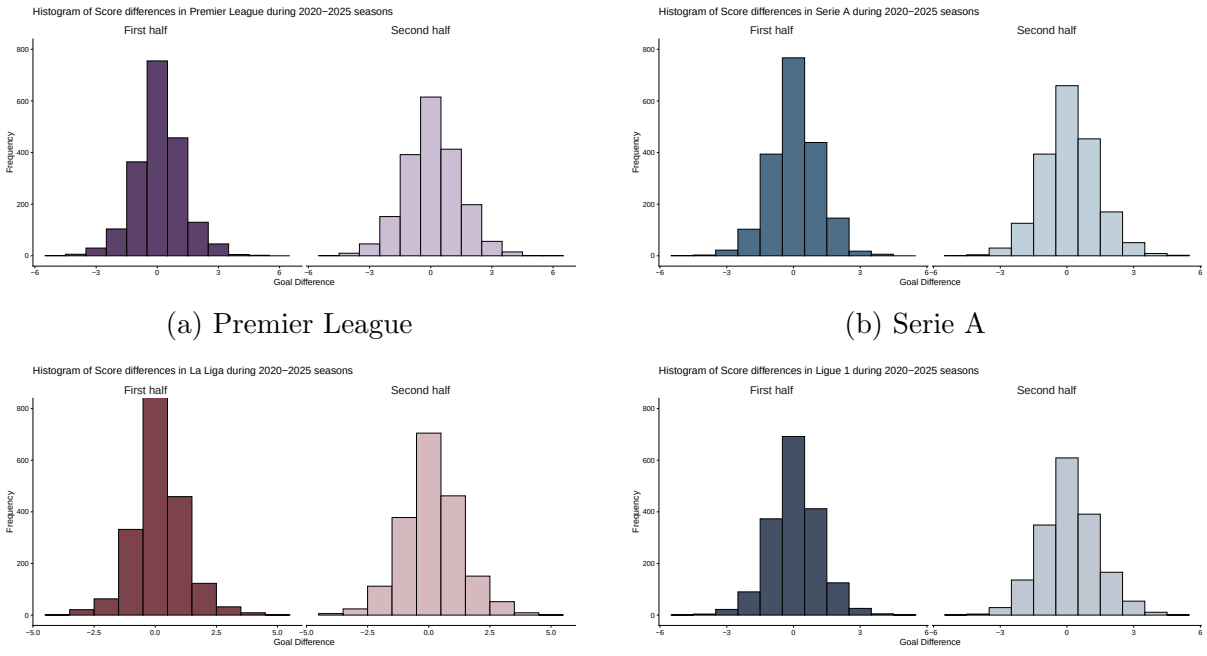
B.4 Half differences based on Skellam



(a) Premier League (b) Serie A

(c) La Liga (d) Ligue 1

Figure B.8: Estimated home and away attack–defence differences estimated by the Skellam model across major European leagues for the 2020-2025 seasons.

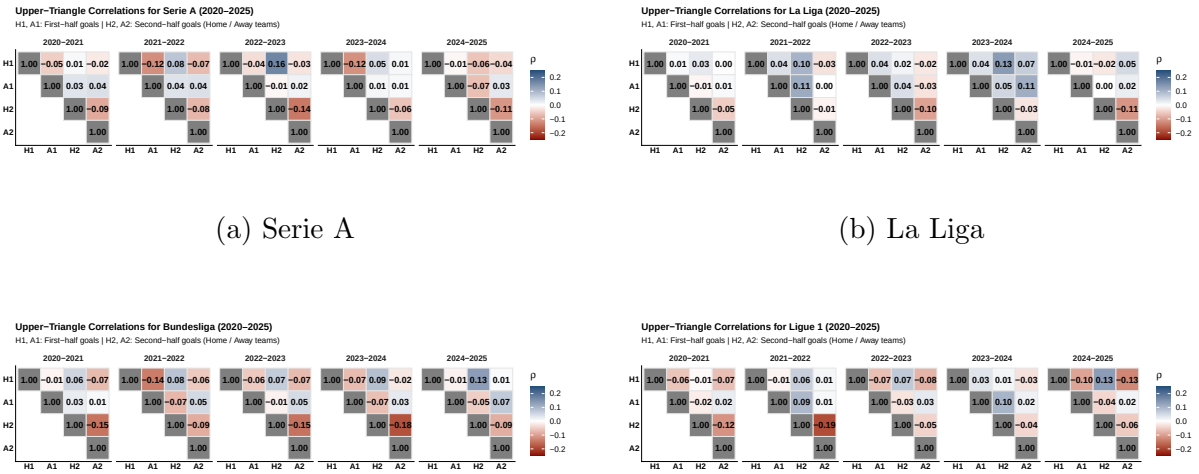


(a) Premier League (b) Serie A

(c) La Liga (d) Ligue 1

Figure B.9: Distribution of score differences by half across major European leagues for the 2020-2025 seasons.

B.5 Dependencies



(a) Serie A

(b) La Liga

(c) Bundesliga

(d) Ligue 1

Figure B.10: Upper-triangle correlation matrices by season for major European leagues. H1 and A1 denote first-half goals for the home and away teams, while H2 and A2 denote second-half goals.

Appendix C

Tables

C.1 Double Poisson and Negative Binomial

Premier League

Season	Model	Df	LogLik	Fit			Parameters	
				Home	Draw	Away	μ	r
	Actual			146.00	84.00	150.00		
2020–2021	Double Poisson	40	-1074.59	145.14	92.15	142.71	1.01	
	Negative Binomial	40	-1074.57	145.15	92.16	142.69	1.01	135.65
	Actual			162.00	88.00	130.00		
2021–2022	Double Poisson	40	-1059.38	159.67	85.34	134.99	1.14	
	Negative Binomial	40	-1059.39	159.52	85.45	135.03	1.14	15189.15
	Actual			180.00	90.00	110.00		
2022–2023	Double Poisson	40	-1079.43	174.87	88.23	116.89	1.33	
	Negative Binomial	40	-1079.43	174.77	88.20	117.03	1.33	10403.09
	Actual			174.00	83.00	123.00		
2023–2024	Double Poisson	40	-1121.17	171.89	79.42	128.69	1.22	
	Negative Binomial	40	-1121.17	172.01	79.44	128.55	1.22	21515.57
	Actual			152.00	96.00	132.00		
2024–2025	Double Poisson	40	-1081.79	151.00	86.58	142.42	1.04	
	Negative Binomial	40	-1081.79	151.02	86.45	142.53	1.04	20372.59

Table C.1: Double Poisson and Negative Binomial model fits for the Premier League (2020–2025).

Serie A

Season	Model	Df	LogLik	Fit			Parameters	
				Home	Draw	Away	μ	r
	Actual			150.00	103.00	126.00		
2020–2021	Double Poisson	40	-1092.90	161.78	83.27	133.95	1.14	
	Negative Binomial	40	-1092.90	161.72	83.26	134.02	1.14	23648.34
	Actual			147.00	99.00	134.00		
2021–2022	Double Poisson	40	-1059.97	156.62	86.83	136.55	1.10	
	Negative Binomial	40	-1059.98	156.51	86.84	136.65	1.10	19211.87
	Actual			159.00	98.00	123.00		
2022–2023	Double Poisson	40	-1016.58	162.63	94.08	123.29	1.22	
	Negative Binomial	40	-1016.58	162.73	93.93	123.34	1.22	17039.73
	Actual			157.00	114.00	109.00		
2023–2024	Double Poisson	40	-1010.01	162.23	93.13	124.64	1.22	
	Negative Binomial	40	-1010.02	162.21	93.12	124.67	1.22	19753.32
	Actual			148.00	114.00	118.00		
2024–2025	Double Poisson	40	-1011.26	152.35	94.26	133.39	1.10	
	Negative Binomial	40	-1011.27	152.43	94.12	133.45	1.10	18109.53

Table C.2: Double Poisson and Negative Binomial model fits for the Serie A (2020–2025).

Bundesliga

Season	Model	Df	LogLik	Fit			Parameters	
				Home	Draw	Away	μ	r
	Actual			130.00	81.00	95.00		
2020–2021	Double Poisson	36	-874.54	137.11	68.33	100.56	1.24	
	Negative Binomial	36	-874.55	137.08	68.26	100.66	1.24	19342.67
	Actual			139.00	76.00	90.00		
2021–2022	Double Poisson	36	-893.42	142.27	66.35	96.38	1.31	
	Negative Binomial	36	-893.42	142.34	66.27	96.39	1.31	15776.01
	Actual			147.00	73.00	86.00		
2022–2023	Double Poisson	36	-926.40	151.83	65.54	88.64	1.43	
	Negative Binomial	36	-926.40	151.72	65.59	88.69	1.43	6493.21
	Actual			135.00	81.00	90.00		
2023–2024	Double Poisson	36	-889.87	143.21	62.56	100.23	1.30	
	Negative Binomial	36	-889.88	143.24	62.65	100.11	1.30	20595.65
	Actual			119.00	75.00	112.00		
2024–2025	Double Poisson	36	-915.98	126.27	68.66	111.07	1.09	
	Negative Binomial	36	-915.98	126.37	68.52	111.11	1.09	15808.41

Table C.3: Double Poisson and Negative Binomial model fits for the Bundesliga (2020–2025).

Ligue 1

Season	Model	Df	LogLik	Fit			Parameters	
				Home	Draw	Away	μ	r
	Actual			142.00	94.00	144.00		
2020–2021	Double Poisson	40	-1059.75	147.25	89.91	142.84	1.02	
	Negative Binomial	40	-1059.75	147.22	89.86	142.92	1.02	18933.73
	Actual			165.00	100.00	115.00		
2021–2022	Double Poisson	40	-1062.88	170.50	89.62	119.89	1.28	
	Negative Binomial	40	-1062.88	170.54	89.60	119.86	1.28	19412.44
	Actual			164.00	94.00	122.00		
2022–2023	Double Poisson	40	-1058.05	159.16	88.78	132.07	1.14	
	Negative Binomial	40	-1058.05	159.13	88.78	132.09	1.14	17062.10
	Actual			118.00	83.00	105.00		
2023–2024	Double Poisson	36	-852.33	126.54	75.36	104.10	1.14	
	Negative Binomial	36	-852.33	126.56	75.36	104.07	1.14	17199.21
	Actual			140.00	67.00	98.00		
2024–2025	Double Poisson	36	-874.96	131.22	68.80	104.98	1.17	
	Negative Binomial	36	-874.96	131.40	68.70	104.91	1.17	18397.77

Table C.4: Double Poisson and Negative Binomial model fits for the Ligue 1 (2020–2025).

C.2 Home Effect

League	Season	$\hat{\mu}$	SE	p-value	LCI	UCI
Premier League	2020–2021	0.01	0.06	0.85	-0.11	0.14
Premier League	2021–2022	0.13	0.06	0.04	0.01	0.25
Premier League	2022–2023	0.29	0.06	0.00	0.17	0.41
Premier League	2023–2024	0.20	0.06	0.00	0.09	0.31
Premier League	2024–2025	0.04	0.06	0.50	-0.08	0.16
Serie A	2020–2021	0.13	0.06	0.03	0.01	0.25
Serie A	2021–2022	0.10	0.06	0.11	-0.02	0.22
Serie A	2022–2023	0.20	0.06	0.00	0.07	0.33
Serie A	2023–2024	0.20	0.06	0.00	0.07	0.32
Serie A	2024–2025	0.10	0.06	0.13	-0.03	0.23
La Liga	2020–2021	0.20	0.07	0.00	0.07	0.33
La Liga	2021–2022	0.27	0.07	0.00	0.14	0.40
La Liga	2022–2023	0.33	0.07	0.00	0.20	0.46
La Liga	2023–2024	0.23	0.06	0.00	0.11	0.36
La Liga	2024–2025	0.21	0.06	0.00	0.08	0.34
Bundesliga	2020–2021	0.22	0.07	0.00	0.08	0.35
Bundesliga	2021–2022	0.27	0.07	0.00	0.14	0.40
Bundesliga	2022–2023	0.35	0.07	0.00	0.22	0.48
Bundesliga	2023–2024	0.26	0.06	0.00	0.13	0.39
Bundesliga	2024–2025	0.09	0.07	0.19	-0.04	0.21
Ligue 1	2020–2021	0.02	0.06	0.73	-0.10	0.15
Ligue 1	2021–2022	0.25	0.06	0.00	0.12	0.37
Ligue 1	2022–2023	0.14	0.06	0.03	0.01	0.26
Ligue 1	2023–2024	0.14	0.07	0.06	0.00	0.27
Ligue 1	2024–2025	0.16	0.07	0.02	0.02	0.29

Table C.5: Estimated home effect under the double Poisson model with 95% confidence intervals for the five major European leagues across the 2020–2025 seasons.

C.3 Compound Poisson

Premier League

Season	Model	Df	LogLik	Fit			Parameters		
				Home	Draw	Away	μ	μ_n	μ_p
2020–2021	Actual			146.00	84.00	150.00			
	Double Poisson	40	-1074.59	145.14	92.15	142.71	1.01		
	Compound Poisson (Sh)	80	-10961.78	146.41	92.10	141.48		1.13	0.90
	Compound Poisson (SoT)	80	-389.36	144.60	92.31	143.09		1.10	0.88
2021–2022	Actual			162.00	88.00	130.00			
	Double Poisson	40	-1059.38	159.67	85.34	134.99	1.14		
	Compound Poisson (Sh)	80	-12263.29	160.49	85.14	134.37		1.18	0.97
	Compound Poisson (SoT)	80	-364.27	160.08	85.26	134.66		1.13	1.02
2022–2023	Actual			180.00	90.00	110.00			
	Double Poisson	40	-1079.43	174.87	88.23	116.89	1.33		
	Compound Poisson (Sh)	80	-11868.16	174.34	88.24	117.42		1.23	1.09
	Compound Poisson (SoT)	80	-388.00	174.64	88.00	117.36		1.25	1.09
2023–2024	Actual			174.00	83.00	123.00			
	Double Poisson	40	-1121.17	171.89	79.42	128.69	1.22		
	Compound Poisson (Sh)	80	-13798.42	172.58	79.27	128.14		1.27	0.97
	Compound Poisson (SoT)	80	-5.39	172.63	79.15	128.22		1.22	1.01
2024–2025	Actual			152.00	96.00	132.00			
	Double Poisson	40	-1081.79	151.00	86.58	142.42	1.04		
	Compound Poisson (Sh)	80	-12331.98	150.87	86.53	142.61		1.13	0.91
	Compound Poisson (SoT)	80	-325.39	151.67	86.56	141.78		1.14	0.88

Table C.6: Double Poisson and Compound Poisson model fits for the Premier League (2020–2025).

La Liga

Season	Model	Df	LogLik	Fit			Parameters		
				Home	Draw	Away	μ	μ_n	μ_p
2020–2021	Actual			155.00	117.00	108.00			
	Double Poisson	40	-1009.28	160.89	96.60	122.51	1.22		
	Compound Poisson (Sh)	80	-8466.08	159.93	96.98	123.08		1.15	1.06
	Compound Poisson (SoT)	80	-812.16	160.54	96.69	122.77		1.16	1.07
2021–2022	Actual			167.00	105.00	108.00			
	Double Poisson	40	-1041.37	167.89	97.41	114.71	1.31		
	Compound Poisson (Sh)	80	-10398.70	167.75	97.40	114.85		1.24	1.06
	Compound Poisson (SoT)	80	-619.31	167.55	97.43	115.02		1.23	1.09
2022–2023	Actual			180.00	94.00	106.00			
	Double Poisson	40	-1001.57	173.67	96.16	110.17	1.39		
	Compound Poisson (Sh)	80	-11443.44	173.40	95.99	110.61		1.30	1.07
	Compound Poisson (SoT)	80	-443.77	173.44	96.19	110.37		1.35	1.03
2023–2024	Actual			167.00	103.00	110.00			
	Double Poisson	40	-1037.39	165.52	94.57	119.90	1.26		
	Compound Poisson (Sh)	80	-11150.36	165.82	94.58	119.60		1.31	0.97
	Compound Poisson (SoT)	80	-378.66	165.60	94.42	119.97		1.33	0.93
2024–2025	Actual			163.00	102.00	115.00			
	Double Poisson	40	-1011.51	163.15	94.41	122.44	1.23		
	Compound Poisson (Sh)	80	-10651.28	162.50	94.19	123.31		1.25	0.98
	Compound Poisson (SoT)	80	-543.80	161.62	94.54	123.84		1.29	0.91

Table C.7: Double Poisson and Compound Poisson model fits for the La Liga (2020–2025).

Bundesliga

Season	Model	Df	LogLik	Fit			Parameters		
				Home	Draw	Away	μ	μ_n	μ_p
2020–2021	Actual			130.00	81.00	95.00			
	Double Poisson	36	-874.54	137.11	68.33	100.56	1.24		
	Compound Poisson (Sh)	72	-8969.27	137.69	68.07	100.24		1.11	1.14
	Compound Poisson (SoT)	72	-246.29	137.35	68.38	100.27		1.13	1.15
2021–2022	Actual			139.00	76.00	90.00			
	Double Poisson	36	-893.42	142.27	66.35	96.38	1.31		
	Compound Poisson (Sh)	72	-9743.79	143.85	66.02	95.13		1.19	1.13
	Compound Poisson (SoT)	72	-201.26	143.25	66.34	95.41		1.22	1.14
2022–2023	Actual			147.00	73.00	86.00			
	Double Poisson	36	-926.40	151.83	65.54	88.64	1.43		
	Compound Poisson (Sh)	72	-9416.72	152.17	65.33	88.50		1.25	1.17
	Compound Poisson (SoT)	72	-167.08	152.54	65.35	88.11		1.36	1.09
2023–2024	Actual			135.00	81.00	90.00			
	Double Poisson	36	-889.87	143.21	62.56	100.23	1.30		
	Compound Poisson (Sh)	72	-11182.40	143.62	62.24	100.14		1.28	1.02
	Compound Poisson (SoT)	72	-44.57	142.70	62.34	100.96		1.30	0.99
2024–2025	Actual			119.00	75.00	112.00			
	Double Poisson	36	-915.98	126.27	68.66	111.07	1.09		
	Compound Poisson (Sh)	72	-9867.63	127.25	68.66	110.09		1.23	0.88
	Compound Poisson (SoT)	72	-251.43	126.13	68.73	111.14		1.19	0.87

Table C.8: Double Poisson and Compound Poisson model fits for the Bundesliga (2020–2025).

Ligue 1

Season	Model	Df	LogLik	Fit			Parameters		
				Home	Draw	Away	μ	μ_n	μ_p
2020–2021	Actual			142.00	94.00	144.00			
	Double Poisson	40	-1059.75	147.25	89.91	142.84			
	Compound Poisson (Sh)	80	-10056.87	146.28	90.13	143.59		1.18	0.84
	Compound Poisson (SoT)	80	-612.44	146.17	90.09	143.74		1.15	0.82
2021–2022	Actual			165.00	100.00	115.00			
	Double Poisson	40	-1062.88	170.50	89.62	119.89			
	Compound Poisson (Sh)	80	-10551.91	170.88	89.72	119.39		1.23	1.05
	Compound Poisson (SoT)	80	-548.57	170.77	89.70	119.53		1.20	1.11
2022–2023	Actual			164.00	94.00	122.00			
	Double Poisson	40	-1058.05	159.16	88.78	132.07			
	Compound Poisson (Sh)	80	-11045.62	160.42	88.74	130.84		1.13	1.02
	Compound Poisson (SoT)	80	-261.09	159.23	88.46	132.31		1.13	1.01
2023–2024	Actual			118.00	83.00	105.00			
	Double Poisson	36	-852.33	126.54	75.36	104.10			
	Compound Poisson (Sh)	72	-9686.93	128.27	75.31	102.42		1.21	0.96
	Compound Poisson (SoT)	72	-175.15	128.65	75.03	102.31		1.19	0.98
2024–2025	Actual			140.00	67.00	98.00			
	Double Poisson	36	-874.96	131.22	68.80	104.98			
	Compound Poisson (Sh)	72	-9123.10	131.76	68.66	104.58		1.17	1.00
	Compound Poisson (SoT)	72	-132.64	131.62	68.78	104.60		1.16	1.02

Table C.9: Double Poisson and Compound Poisson model fits for Ligue 1 (2020–2025).

C.4 Univariate Skellam

Premier League

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	σ^2	p
2020–2021	Actual			146.00	84.00	150.00		
	Double Poisson	40	−1074.59	145.14	92.15	142.71		
	Skellam2	40	−704.74	146.07	86.33	147.60	2.68	
	ZI SKellam2	41	−704.74	146.07	86.33	147.60	2.68	0.00
	Skellam	78	−676.48	135.54	102.38	142.07		
	ZI Skellam	79	−676.46	135.57	102.34	142.10		0.00
2021–2022	Actual			162.00	88.00	130.00		
	Double Poisson	40	−1059.38	159.67	85.34	134.99		
	Skellam2	40	−709.40	167.67	71.65	140.69	3.40	
	ZI SKellam2	41	−708.85	161.17	87.65	131.18	3.46	0.05
	Skellam	78	−680.13	156.28	94.21	129.51		
	ZI Skellam	79	−680.06	156.51	94.19	129.30		0.00
2022–2023	Actual			180.00	90.00	110.00		
	Double Poisson	40	−1079.43	174.87	88.23	116.89		
	Skellam2	40	−711.60	183.74	72.15	124.10	3.63	
	ZI SKellam2	41	−710.36	174.22	89.10	116.67	3.69	0.06
	Skellam	78	−677.01	171.04	98.26	110.69		
	ZI Skellam	79	−677.16	170.72	98.92	110.35		0.00
2023–2024	Actual			174.00	83.00	123.00		
	Double Poisson	40	−1121.17	171.89	79.42	128.69		
	Skellam2	40	−732.79	176.40	71.96	131.63	3.52	
	ZI SKellam2	41	−731.98	173.06	82.15	124.79	3.48	0.03
	Skellam	78	−709.69	172.09	86.88	121.03		
	ZI Skellam	79	−709.51	172.18	87.17	120.65		0.00
2024–2025	Actual			152.00	96.00	132.00		
	Double Poisson	40	−1081.79	151.00	86.58	142.42		
	Skellam2	40	−701.47	159.07	81.04	139.88	2.97	
	ZI SKellam2	41	−699.73	140.92	95.16	143.91	2.85	0.04
	Skellam	78	−672.09	149.09	101.60	129.31		
	ZI Skellam	79	−672.07	149.00	101.52	129.48		0.00

Table C.10: Double Poisson and Skellam model fits for the Premier League (2020–2025).

Serie A

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	σ^2	p
2020–2021	Actual			150.00	103.00	126.00		
	Double Poisson	40	−1092.90	161.78	83.27	133.95		
	Skellam2	40	−688.30	167.53	74.63	136.84	3.32	
	ZI SKellam2	41	−685.41	156.40	102.55	120.05	3.36	0.09
	Skellam	78	−650.31	150.61	105.58	122.82		
	ZI Skellam	79	−650.31	150.69	105.53	122.79		0.00
2021–2022	Actual			147.00	99.00	134.00		
	Double Poisson	40	−1059.97	156.62	86.83	136.55		
	Skellam2	40	−677.70	158.81	85.81	135.38	2.57	
	ZI SKellam2	41	−677.65	148.36	98.98	132.66	2.62	0.05
	Skellam	78	−647.72	145.06	108.13	126.81		
	ZI Skellam	79	−647.72	144.98	108.21	126.81		0.00
2022–2023	Actual			159.00	98.00	123.00		
	Double Poisson	40	−1016.58	162.63	94.08	123.29		
	Skellam2	40	−669.72	166.05	89.81	124.14	2.50	
	ZI SKellam2	41	−669.43	162.81	98.04	119.15	2.50	0.03
	Skellam	78	−644.80	156.70	111.42	111.88		
	ZI Skellam	79	−644.84	156.57	111.68	111.75		0.00
2023–2024	Actual			157.00	114.00	109.00		
	Double Poisson	40	−1010.01	162.23	93.13	124.64		
	Skellam2	40	−661.74	168.31	82.47	129.22	2.86	
	ZI SKellam2	41	−656.65	155.51	113.71	110.78	2.90	0.11
	Skellam	78	−621.36	155.11	116.32	108.57		
	ZI Skellam	79	−621.38	155.04	116.64	108.32		0.00
2024–2025	Actual			148.00	114.00	118.00		
	Double Poisson	40	−1011.26	152.35	94.26	133.39		
	Skellam2	40	−664.36	153.61	86.14	140.25	2.72	
	ZI SKellam2	41	−657.47	150.47	103.77	125.76	2.59	0.05
	Skellam	78	−624.41	145.96	120.50	113.53		
	ZI Skellam	79	−624.41	145.91	120.53	113.56		0.00

Table C.11: Double Poisson and Skellam model fits for the Serie A (2020–2025).

La Liga

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	σ^2	p
2020–2021	Actual			155.00	117.00	108.00		
	Double Poisson	40	−1009.28	160.89	96.60	122.51		
	Skellam2	40	−654.40	164.60	94.94	120.47	2.30	
	ZI SKellam2	41	−651.85	152.18	116.36	111.47	2.32	0.08
	Skellam	78	−631.45	151.17	117.37	111.46		
	ZI Skellam	79	−631.45	151.17	117.42	111.41		0.00
2021–2022	Actual			167.00	105.00	108.00		
	Double Poisson	40	−1041.37	167.89	97.41	114.71		
	Skellam2	40	−676.43	166.73	86.28	126.98	2.82	
	ZI SKellam2	41	−673.87	162.34	105.80	111.86	2.83	0.07
	Skellam	78	−640.86	158.88	115.83	105.30		
	ZI Skellam	79	−640.87	158.80	115.75	105.45		0.00
2022–2023	Actual			180.00	94.00	106.00		
	Double Poisson	40	−1001.57	173.67	96.16	110.17		
	Skellam2	40	−659.91	169.73	90.82	119.44	2.56	
	ZI SKellam2	41	−660.34	167.89	93.06	119.06	2.57	0.01
	Skellam	78	−632.39	168.81	115.41	95.78		
	ZI Skellam	79	−632.38	168.79	115.40	95.81		0.00
2023–2024	Actual			167.00	103.00	110.00		
	Double Poisson	40	−1037.39	165.52	94.57	119.90		
	Skellam2	40	−675.27	170.76	77.65	131.60	3.31	
	ZI SKellam2	41	−671.00	158.66	103.27	118.08	3.31	0.09
	Skellam	78	−626.54	157.31	115.97	106.72		
	ZI Skellam	79	−626.59	157.05	116.20	106.75		0.00
2024–2025	Actual			163.00	102.00	115.00		
	Double Poisson	40	−1011.51	163.15	94.41	122.44		
	Skellam2	40	−674.73	170.08	76.75	133.17	3.36	
	ZI SKellam2	41	−671.86	157.81	102.15	120.03	3.39	0.09
	Skellam	78	−620.39	150.60	121.00	108.40		
	ZI Skellam	79	−620.39	150.60	120.95	108.45		0.00

Table C.12: Double Poisson and Skellam model fits for the La Liga (2020–2025).

Ligue 1

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	σ^2	p
2020–2021	Actual			142.00	94.00	144.00		
	Double Poisson	40	−1059.75	147.25	89.91	142.84		
	Skellam2	40	−690.55	148.63	86.01	145.35	2.69	
	ZI SKellam2	41	−690.37	143.90	94.31	141.79	2.71	0.03
	Skellam	78	−671.38	136.60	104.07	139.33		
	ZI Skellam	79	−671.39	136.58	104.11	139.31		0.00
2021–2022	Actual			165.00	100.00	115.00		
	Double Poisson	40	−1062.88	170.50	89.62	119.89		
	Skellam2	40	−691.57	168.54	80.22	131.24	3.08	
	ZI SKellam2	41	−690.04	163.58	99.45	116.97	3.16	0.07
	Skellam	78	−656.65	161.34	109.20	109.47		
	ZI Skellam	79	−656.61	161.25	109.24	109.52		0.00
2022–2023	Actual			164.00	94.00	122.00		
	Double Poisson	40	−1058.05	159.16	88.78	132.07		
	Skellam2	40	−691.00	159.10	81.68	139.22	2.91	
	ZI SKellam2	41	−692.78	160.06	94.71	125.23	3.04	0.05
	Skellam	78	−660.04	148.73	107.47	123.81		
	ZI Skellam	79	−660.04	148.70	107.48	123.82		0.00
2023–2024	Actual			118.00	83.00	105.00		
	Double Poisson	36	−852.33	126.54	75.36	104.10		
	Skellam2	36	−536.65	128.17	71.33	106.50	2.65	
	ZI SKellam2	37	−534.88	125.87	83.09	97.04	2.62	0.05
	Skellam	70	−516.78	121.45	90.68	93.87		
	ZI Skellam	71	−516.75	121.38	90.75	93.88		0.00
2024–2025	Actual			140.00	67.00	98.00		
	Double Poisson	36	−874.96	131.22	68.80	104.98		
	Skellam2	36	−569.58	136.08	59.40	109.52	3.33	
	ZI SKellam2	37	−568.93	134.50	67.18	103.32	3.30	0.03
	Skellam	70	−546.93	129.82	76.77	98.41		
	ZI Skellam	71	−546.93	129.78	76.80	98.41		0.00

Table C.13: Double Poisson and Skellam model fits for the Ligue 1 (2020–2025).

C.5 Bivariate Skellam

Premier League

Season	Model	Df	LogLik	Fit			Special Parameters					
				Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ	corr
2020–2021	Actual			146.00	84.00	150.00						0.07
	Double Poisson	40	-1074.59	145.14	92.15	142.71						
	Skellam2	80	-1158.81	143.89	76.26	159.84	1.40	2.05				
	ZI Skellam2	82	-1149.57	138.20	88.56	153.23	1.40	2.05	0.08	0.13		
	FC Skellam2	81	-1158.70	143.34	77.15	159.50	1.40	2.05			-0.19	
	ZI FC Skellam2	83	-1149.48	137.80	89.21	152.99	1.40	2.05	0.08	0.13	-0.15	
2021–2022	Actual			162.00	88.00	130.00						0.12
	Double Poisson	40	-1059.38	159.67	85.34	134.99						
	Skellam2	80	-1165.03	162.78	69.65	147.56	1.58	2.01				
	ZI Skellam2	82	-1160.01	159.31	76.05	144.63	1.58	2.01	0.11	0.03		
	FC Skellam2	81	-1164.45	162.26	71.22	146.51	1.58	2.01			-0.44	
	ZI FC Skellam2	83	-1159.39	158.87	77.43	143.69	1.58	2.01	0.11	0.03	-0.43	
2022–2023	Actual			180.00	90.00	110.00						0.05
	Double Poisson	40	-1079.43	174.87	88.23	116.89						
	Skellam2	80	-1165.34	177.03	74.96	128.01	1.49	1.90				
	ZI Skellam2	82	-1160.40	171.48	83.68	124.84	1.49	1.90	0.06	0.09		
	FC Skellam2	81	-1164.32	176.67	77.48	125.84	1.49	1.90			-0.59	
	ZI FC Skellam2	83	-1159.47	171.25	85.76	122.99	1.49	1.90	0.06	0.09	-0.53	
2023–2024	Actual			174.00	83.00	123.00						0.10
	Double Poisson	40	-1121.17	171.89	79.42	128.69						
	Skellam2	80	-1213.30	174.57	65.91	139.50	1.92	2.36				
	ZI Skellam2	82	-1203.64	171.91	70.61	137.47	1.92	2.36	0.15	0.00		
	FC Skellam2	81	-1213.28	174.53	66.26	139.20	1.92	2.36			-0.09	
	ZI FC Skellam2	83	-1203.58	171.83	71.16	136.99	1.92	2.36	0.15	0.00	-0.15	
2024–2025	Actual			152.00	96.00	132.00						-0.01
	Double Poisson	40	-1081.79	151.00	86.58	142.42						
	Skellam2	80	-1184.19	156.36	77.64	146.01	1.40	1.90				
	ZI Skellam2	82	-1184.17	156.30	77.71	145.98	1.40	1.90	0.00	0.00		
	FC Skellam2	81	-1180.27	154.53	82.53	142.94	1.40	1.90			-1.02	
	ZI FC Skellam2	83	-1180.25	154.48	82.60	142.92	1.40	1.90	0.00	0.00	-1.02	

Table C.14: Bivariate Skellam model comparison results for the Premier League (2020–2025).

Premier League: Joint Half-Time / Full-Time Results

Season	Actual			Independent ZI Skellam2			Frank Copula ZI Skellam2		
2020–2021	93	16	10	80.6	14.6	10.1	79.8	15.0	10.5
	43	57	58	47.7	58.7	51.0	47.7	58.6	51.1
	10	11	82	9.9	15.3	92.1	10.3	15.7	91.4
2021–2022	93	24	5	93.3	14.7	9.0	91.2	15.7	10.0
	60	49	46	56.2	46.5	51.3	56.6	45.9	51.5
	9	15	79	9.9	14.9	84.4	11.0	15.9	82.2
2022–2023	112	16	6	107.3	15.1	8.3	104.9	16.5	9.4
	53	55	41	53.8	53.1	40.7	54.6	52.6	40.5
	15	19	63	10.3	15.5	75.9	11.8	16.8	73.1
2023–2024	96	16	13	94.8	15.2	10.5	94.1	15.5	10.9
	63	48	43	63.5	39.2	50.7	63.6	39.1	50.6
	15	19	67	13.7	16.2	76.3	14.2	16.5	75.5
2024–2025	95	30	15	102.2	20.9	12.7	96.6	23.7	15.5
	45	47	47	45.3	41.2	50.1	46.7	40.7	49.3
	12	19	70	8.8	15.6	83.2	11.2	18.2	78.1

Table C.15: Observed and bivariate Skellam model-implied joint half-time/full-time result matrices for the Premier League (2020–2025).

Serie A

Season	Model	Df	LogLik	Fit			Special Parameters					
				Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ	corr
2020–2021	Actual			150.00	103.00	126.00						0.02
	Double Poisson	40	-1092.90	161.78	83.27	133.95						
	Skellam2	80	-1162.14	165.79	70.75	142.45	1.60	2.12				
	ZI Skellam2	82	-1157.77	161.61	78.08	139.31	1.60	2.12	0.08	0.07		
	FC Skellam2	81	-1156.13	163.69	77.45	137.86	1.60	2.12			-1.57	
	ZI FC Skellam2	83	-1152.58	159.89	83.65	135.45	1.60	2.12	0.08	0.07	-1.35	
2021–2022	Actual			147.00	99.00	134.00						0.03
	Double Poisson	40	-1059.97	156.62	86.83	136.55						
	Skellam2	80	-1146.39	157.09	77.69	145.22	1.57	1.58				
	ZI Skellam2	82	-1145.58	155.24	81.15	143.61	1.57	1.58	0.03	0.03		
	FC Skellam2	81	-1143.35	155.51	81.69	142.79	1.57	1.58			-0.97	
	ZI FC Skellam2	83	-1142.40	153.71	85.04	141.25	1.57	1.58	0.03	0.03	-0.97	
2022–2023	Actual			159.00	98.00	123.00						0.06
	Double Poisson	40	-1016.58	162.63	94.08	123.29						
	Skellam2	80	-1100.39	168.03	80.79	131.18	1.25	1.85				
	ZI Skellam2	82	-1100.21	167.49	81.67	130.84	1.25	1.85	0.02	0.00		
	FC Skellam2	81	-1099.71	167.50	83.15	129.34	1.25	1.85			-0.51	
	ZI FC Skellam2	83	-1099.52	166.98	84.01	129.02	1.25	1.85	0.02	0.00	-0.51	
2023–2024	Actual			157.00	114.00	109.00						0.00
	Double Poisson	40	-1010.01	162.23	93.13	124.64						
	Skellam2	80	-1137.42	170.84	72.01	137.15	1.83	1.96				
	ZI Skellam2	82	-1123.99	163.11	85.01	131.88	1.83	1.96	0.16	0.08		
	FC Skellam2	81	-1131.17	168.73	81.02	130.24	1.83	1.96			-1.84	
	ZI FC Skellam2	83	-1119.15	161.71	91.21	127.08	1.83	1.96	0.16	0.08	-1.37	
2024–2025	Actual			148.00	114.00	118.00						-0.02
	Double Poisson	40	-1011.26	152.35	94.26	133.39						
	Skellam2	80	-1125.54	163.09	74.71	142.20	1.84	1.76				
	ZI Skellam2	82	-1106.77	154.32	91.40	134.27	1.84	1.76	0.19	0.10		
	FC Skellam2	81	-1120.71	160.65	82.77	136.58	1.84	1.76			-1.59	
	ZI FC Skellam2	83	-1102.43	152.50	97.21	130.29	1.84	1.76	0.19	0.10	-1.25	

Table C.16: Bivariate Skellam model comparison results for Serie A (2020–2025).

Serie A: Joint Half-Time / Full-Time Results

Season	Actual			Independent			ZI Skellam2			Frank Copula			ZI Skellam2
2020–2021	92	23	9	96.3	14.9	9.5		89.9	18.1	12.7			
	47	59	40	53.2	46.7	44.5		53.9	46.0	44.6			
	11	21	77	12.1	16.4	85.3		16.2	19.6	78.1			
2021–2022	94	25	9	100.3	15.5	7.3		96.0	18.0	9.1			
	41	50	45	46.1	48.3	40.7		46.6	47.3	41.2			
	12	24	80	8.9	17.3	95.7		11.1	19.8	90.9			
2022–2023	93	26	9	96.9	17.3	9.3		94.2	18.8	10.5			
	57	51	46	59.3	46.8	47.7		59.9	46.4	47.5			
	9	21	68	11.3	17.6	73.9		12.9	18.9	71.0			
2023–2024	92	22	10	96.1	14.3	8.5		89.7	17.6	11.5			
	54	70	38	57.0	56.0	47.5		58.2	55.6	46.8			
	11	22	61	10.0	14.7	75.9		13.8	18.0	68.8			
2024–2025	77	19	9	86.8	12.7	6.6		81.4	15.7	9.0			
	60	68	41	58.1	63.7	46.9		58.4	63.3	46.9			
	11	27	68	9.5	15.0	80.8		12.7	18.2	74.4			

Table C.17: Observed and bivariate Skellam model-implied joint half-time/full-time result matrices for Serie A (2020–2025).

La Liga

Season	Model	Df	LogLik	Fit			Special Parameters					
				Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ	corr
2020–2021	Actual			155.00	117.00	108.00						0.02
	Double Poisson	40	-1009.28	160.89	96.60	122.51						
	Skellam2	80	-1079.51	168.29	85.53	126.18	1.28	1.56				
	ZI Skellam2	82	-1074.98	163.34	93.62	123.04	1.28	1.56	0.12	0.04		
	FC Skellam2	81	-1077.84	167.54	89.36	123.10	1.28	1.56			-0.77	
	ZI FC Skellam2	83	-1073.46	162.74	96.83	120.43	1.28	1.56	0.12	0.04	-0.69	
2021–2022	Actual			167.00	105.00	108.00						0.02
	Double Poisson	40	-1041.37	167.89	97.41	114.71						
	Skellam2	80	-1109.71	172.57	82.98	124.44	1.05	1.97				
	ZI Skellam2	82	-1102.61	164.84	95.41	119.74	1.05	1.97	0.04	0.12		
	FC Skellam2	81	-1108.74	172.12	85.66	122.22	1.05	1.97			-0.56	
	ZI FC Skellam2	83	-1101.79	164.52	97.56	117.92	1.05	1.97	0.04	0.12	-0.48	
2022–2023	Actual			180.00	94.00	106.00						-0.02
	Double Poisson	40	-1001.57	173.67	96.16	110.17						
	Skellam2	80	-1102.21	185.30	78.28	116.42	1.79	1.52				
	ZI Skellam2	82	-1080.80	172.84	95.72	111.44	1.79	1.52	0.22	0.07		
	FC Skellam2	81	-1100.55	185.28	82.58	112.14	1.79	1.52			-0.84	
	ZI FC Skellam2	83	-1079.27	172.89	98.88	108.22	1.79	1.52	0.22	0.07	-0.70	
2023–2024	Actual			167.00	103.00	110.00						0.06
	Double Poisson	40	-1037.39	165.52	94.57	119.90						
	Skellam2	80	-1133.25	170.49	70.58	138.92	1.78	2.28				
	ZI Skellam2	82	-1117.39	161.76	85.30	132.93	1.78	2.28	0.15	0.12		
	FC Skellam2	81	-1131.60	168.56	76.99	134.44	1.78	2.28			-1.23	
	ZI FC Skellam2	83	-1116.55	160.93	88.27	130.79	1.78	2.28	0.15	0.12	-0.65	
2024–2025	Actual			163.00	102.00	115.00						-0.06
	Double Poisson	40	-1011.51	163.15	94.41	122.44						
	Skellam2	80	-1156.55	174.43	69.81	135.75	1.97	2.10				
	ZI Skellam2	82	-1130.44	162.11	88.41	129.47	1.97	2.10	0.22	0.11		
	FC Skellam2	81	-1145.00	171.58	82.81	125.60	1.97	2.10			-2.47	
	ZI FC Skellam2	83	-1121.12	160.14	97.32	122.53	1.97	2.10	0.22	0.11	-1.83	

Table C.18: Bivariate Skellam model comparison results for La Liga (2020–2025).

La Liga: Joint Half-Time / Full-Time Results

Season	Actual			Independent			ZI Skellam2			Frank Copula			ZI Skellam2		
2020–2021	87	25	6	93.1	15.1	6.7	89.9	17.0	8.0						
	63	69	41	61.9	63.1	48.0	62.8	62.6	47.6						
	5	23	61	8.4	15.5	68.3	10.1	17.3	64.9						
2021–2022	102	14	8	93.0	14.5	8.2	90.6	15.8	9.2						
	57	69	49	61.9	66.5	46.7	62.5	66.1	46.4						
	8	22	51	10.0	14.4	64.9	11.3	15.6	62.3						
2022–2023	106	18	4	102.1	14.1	6.4	99.2	15.9	7.5						
	67	58	53	63.2	68.4	45.7	64.6	68.0	44.6						
	7	18	49	7.6	13.3	59.4	9.2	15.0	56.1						
2023–2024	95	26	4	95.2	14.6	9.9	91.9	16.2	11.6						
	59	66	36	55.7	56.6	48.2	56.3	56.4	47.8						
	13	11	70	10.9	14.1	74.8	12.8	15.7	71.3						
2024–2025	94	22	13	98.3	14.9	9.8	89.3	19.5	14.3						
	60	60	52	56.0	61.9	53.3	58.6	62.1	50.4						
	9	20	50	7.8	11.6	66.4	12.3	15.7	57.9						

Table C.19: Observed and bivariate Skellam model-implied joint half-time/full-time result matrices for the Premier League (2020–2025).

Ligue 1

Season	Model	Df	LogLik	Fit			Special Parameters					
				Home	Draw	Away	σ_1	σ_2	p_1	p_2	θ	corr
2020–2021	Actual			142.00	94.00	144.00						0.03
	Double Poisson	40	-1059.75	147.25	89.91	142.84						
	Skellam2	80	-1167.88	161.53	73.85	144.62	1.74	1.94				
	ZI Skellam2	82	-1161.91	156.61	82.61	140.78	1.74	1.94	0.07	0.09		
	FC Skellam2	81	-1166.37	160.13	77.49	142.39	1.74	1.94			-0.73	
	ZI FC Skellam2	83	-1160.47	155.44	85.69	138.87	1.74	1.94	0.07	0.09	-0.66	
2021–2022	Actual			165.00	100.00	115.00						-0.04
	Double Poisson	40	-1062.88	170.50	89.62	119.89						
	Skellam2	80	-1166.36	166.06	82.27	131.66	1.14	1.85				
	ZI Skellam2	82	-1165.67	164.04	85.34	130.62	1.14	1.85	0.00	0.04		
	FC Skellam2	81	-1162.16	164.56	87.25	128.19	1.14	1.85			-1.04	
	ZI FC Skellam2	83	-1161.54	162.61	90.17	127.21	1.14	1.85	0.00	0.04	-1.01	
2022–2023	Actual			164.00	94.00	122.00						0.05
	Double Poisson	40	-1058.05	159.16	88.78	132.07						
	Skellam2	80	-1143.22	156.81	76.79	146.40	1.72	1.60				
	ZI Skellam2	82	-1137.12	151.83	86.25	141.92	1.72	1.60	0.11	0.06		
	FC Skellam2	81	-1142.30	155.64	79.38	144.97	1.72	1.60			-0.57	
	ZI FC Skellam2	83	-1136.19	150.83	88.50	140.67	1.72	1.60	0.11	0.06	-0.53	
2023–2024	Actual			118.00	83.00	105.00						0.02
	Double Poisson	36	-852.33	126.54	75.36	104.10						
	Skellam2	72	-890.43	136.53	65.31	104.16	1.54	1.57				
	ZI Skellam2	74	-882.29	130.76	74.46	100.78	1.54	1.57	0.16	0.05		
	FC Skellam2	73	-889.58	136.05	67.94	102.01	1.54	1.57			-0.65	
	ZI FC Skellam2	75	-881.64	130.43	76.35	99.22	1.54	1.57	0.16	0.05	-0.51	
2024–2025	Actual			140.00	67.00	98.00						0.11
	Double Poisson	36	-874.96	131.22	68.80	104.98						
	Skellam2	72	-933.88	134.70	58.40	111.90	1.68	1.85				
	ZI Skellam2	74	-933.09	134.07	60.04	110.89	1.68	1.85	0.05	0.00		
	FC Skellam2	73	-933.87	134.67	58.53	111.80	1.68	1.85			-0.04	
	ZI FC Skellam2	75	-933.09	134.03	60.17	110.79	1.68	1.85	0.05	0.00	-0.04	

Table C.20: Bivariate Skellam model comparison results for Ligue 1 (2020–2025).

Ligue 1: Joint Half-Time / Full-Time Results

Season	Actual			Independent			ZI Skellam2			Frank Copula			ZI Skellam2		
2020–2021	91	21	8	97.8	15.2	8.6				94.5	16.9	10.1			
	38	55	47	48.1	50.7	41.4				48.3	50.3	41.5			
	13	18	89	10.8	16.7	90.8				12.6	18.5	87.3			
2021–2022	97	13	12	91.8	15.3	8.5				86.7	18.1	10.9			
	49	59	47	60.4	52.1	46.3				60.9	51.5	46.4			
	19	28	56	11.8	17.9	75.9				14.9	20.6	70.0			
2022–2023	97	19	8	96.0	15.4	7.8				93.5	16.8	8.8			
	53	59	38	47.9	55.5	45.3				48.2	55.1	45.5			
	14	16	76	8.0	15.3	88.8				9.1	16.7	86.4			
2023–2024	74	16	9	75.8	12.2	6.0				74.0	13.3	6.7			
	41	54	44	48.2	50.4	39.3				48.8	50.1	38.9			
	3	13	52	6.7	11.9	55.6				7.7	12.9	53.5			
2024–2025	86	16	2	80.7	12.1	6.3				80.5	12.2	6.4			
	41	33	34	43.1	32.3	32.3				43.1	32.3	32.3			
	13	18	62	10.3	15.6	72.3				10.4	15.6	72.1			

Table C.21: Observed and bivariate Skellam model-implied joint half-time/full-time result matrices for Ligue 1 (2020–2025).

C.6 Bivariate Poisson

Serie A

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	θ	corr
2020–2021	Actual			150.00	103.00	126.00		-0.04
	Double Poisson	40	-1092.90	161.78	83.27	133.95		
	Double Poisson 2	78	-1068.41	154.94	91.76	132.30		
	Frank Copula Poisson	79	-1066.55	157.44	89.33	132.23	0.77	
2021–2022	Actual			147.00	99.00	134.00		-0.06
	Double Poisson	40	-1059.97	156.62	86.83	136.55		
	Double Poisson 2	78	-1035.43	153.85	95.46	130.69		
	Frank Copula Poisson	79	-1035.25	157.56	89.59	132.85	0.32	
2022–2023	Actual			159.00	98.00	123.00		-0.11
	Double Poisson	40	-1016.58	162.63	94.08	123.29		
	Double Poisson 2	78	-996.87	156.18	102.76	121.06		
	Frank Copula Poisson	79	-996.83	161.92	92.97	125.11	-0.15	
2023–2024	Actual			157.00	114.00	109.00		-0.03
	Double Poisson	40	-1010.01	162.23	93.13	124.64		
	Double Poisson 2	78	-992.52	159.73	101.09	119.18		
	Frank Copula Poisson	79	-991.42	162.81	98.30	118.89	0.64	
2024–2025	Actual			148.00	114.00	118.00		-0.06
	Double Poisson	40	-1011.26	152.35	94.26	133.39		
	Double Poisson 2	78	-994.27	148.31	102.26	129.43		
	Frank Copula Poisson	79	-993.53	151.09	98.60	130.31	0.53	

Table C.22: Bivariate Poisson model comparison results for Serie A (2020–2025).

La Liga

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	θ	corr
2020–2021	Actual			155.00	117.00	108.00		-0.02
	Double Poisson	40	-1009.28	160.89	96.60	122.51		
	Double Poisson 2	78	-993.84	155.76	105.38	118.86		
	Frank Copula Poisson	79	-992.04	158.30	104.02	117.68	0.76	
2021–2022	Actual			167.00	105.00	108.00		0.05
	Double Poisson	40	-1041.37	167.89	97.41	114.71		
	Double Poisson 2	78	-1021.89	160.85	107.34	111.81		
	Frank Copula Poisson	79	-1019.21	163.81	107.08	109.11	0.84	
2022–2023	Actual			180.00	94.00	106.00		-0.02
	Double Poisson	40	-1001.57	173.67	96.16	110.17		
	Double Poisson 2	78	-988.86	169.30	104.30	106.40		
	Frank Copula Poisson	79	988.74	174.08	98.17	107.76	0.24	
2023–2024	Actual			167.00	103.00	110.00		0.05
	Double Poisson	40	-1037.39	165.52	94.57	119.90		
	Double Poisson 2	78	-1020.59	159.17	103.68	117.15		
	Frank Copula Poisson	79	-1016.36	160.07	105.78	114.15	1.14	
2024–2025	Actual			163.00	102.00	115.00		-0.01
	Double Poisson	40	-1011.51	163.15	94.41	122.44		
	Double Poisson 2	78	-994.79	155.54	103.55	120.90		
	Frank Copula Poisson	79	-994.16	159.15	99.88	120.97	0.49	

Table C.23: Bivariate Poisson model comparison results for La Liga (2020–2025).

Bundesliga

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	θ	corr
2020–2021	Actual			130.00	81.00	95.00		-0.06
	Double Poisson	36	-874.54	137.11	68.33	100.56		
	Double Poisson 2	70	-858.51	130.65	75.34	100.02		
	Frank Copula Poisson	71	858.49	134.97	68.09	102.93	-0.09	
2021–2022	Actual			139.00	76.00	90.00		-0.11
	Double Poisson	36	-893.42	142.27	66.35	96.38		
	Double Poisson 2	70	-876.70	139.51	72.25	93.24		
	Frank Copula Poisson	71	-876.13	144.36	63.25	97.39	-0.45	
2022–2023	Actual			147.00	73.00	86.00		-0.09
	Double Poisson	36	-926.40	151.83	65.54	88.64		
	Double Poisson 2	70	-910.83	147.20	71.85	86.95		
	Frank Copula Poisson	71	-910.43	152.54	63.09	90.37	-0.38	
2023–2024	Actual			135.00	81.00	90.00		-0.11
	Double Poisson	36	-889.87	143.21	62.56	100.23		
	Double Poisson 2	70	-872.95	141.20	68.48	96.32		
	Frank Copula Poisson	71	872.94	145.24	62.36	98.40	0.08	
2024–2025	Actual			119.00	75.00	112.00		-0.04
	Double Poisson	36	-915.98	126.27	68.66	111.07		
	Double Poisson 2	70	-899.23	121.52	74.81	109.67		
	Frank Copula Poisson	71	-899.14	124.41	69.69	111.90	0.24	

Table C.24: Bivariate Poisson model comparison results for the Bundesliga (2020–2025).

Ligue 1

Season	Model	Df	LogLik	Fit			Special Parameters	
				Home	Draw	Away	θ	corr
2020–2021	Actual			142.00	94.00	144.00		-0.09
	Double Poisson	40	-1059.75	147.25	89.91	142.84		
	Double Poisson 2	78	-1042.86	141.20	98.11	140.69		
	Frank Copula Poisson	79	-1042.86	145.35	89.86	144.79	0.01	
2021–2022	Actual			165.00	100.00	115.00		-0.02
	Double Poisson	40	-1062.88	170.50	89.62	119.89		
	Double Poisson 2	78	-1048.65	163.03	98.15	118.81		
	Frank Copula Poisson	79	-1047.97	166.62	93.86	119.52	0.45	
2022–2023	Actual			164.00	94.00	122.00		-0.07
	Double Poisson	40	-1058.05	159.16	88.78	132.07		
	Double Poisson 2	78	-1038.78	151.97	97.36	130.67		
	Frank Copula Poisson	79	-1038.59	155.69	91.33	132.98	0.27	
2023–2024	Actual			118.00	83.00	105.00		0.01
	Double Poisson	36	-852.33	126.54	75.36	104.10		
	Double Poisson 2	70	-833.22	124.17	81.84	99.99		
	Frank Copula Poisson	71	-831.27	125.10	81.48	99.42	0.85	
2024–2025	Actual			140.00	67.00	98.00		-0.10
	Double Poisson	36	-874.96	131.22	68.80	104.98		
	Double Poisson 2	70	-851.20	128.40	74.43	102.18		
	Frank Copula Poisson	71	-850.96	132.96	66.01	106.03	-0.31	

Table C.25: Bivariate Poisson model comparison results for Ligue 1 (2020–2025).

C.7 4-Variate Poisson

Serie A

Season	Model	Df	LogLik	Fit			Dependence Parameters					
				Home	Draw	Away	r_{12}	r_{13}	r_{14}	r_{23}	r_{24}	r_{34}
2020–2021	Actual			150.00	103.00	126.00						
	DP	40	-1092.90	161.78	83.27	133.95						
	4P	156	-1576.78	159.67	83.46	135.85						
	4CP	162	-1571.81	155.44	88.38	128.80	0.06	-0.14	0.07	0.17	0.01	-0.01
2021–2022	Actual			147.00	99.00	134.00						
	DP	40	-1059.97	156.62	86.83	136.55						
	4P	156	-1518.85	158.59	87.06	134.34						
	4CP	162	-1514.46	154.84	89.51	131.40	-0.09	-0.06	0.07	0.17	-0.10	-0.04
2022–2023	Actual			159.00	98.00	123.00						
	DP	40	-1016.58	162.63	94.08	123.29						
	4P	156	-1441.06	161.33	94.35	124.32						
	4CP	162	-1436.76	157.77	93.97	125.52	0.03	0.02	0.08	0.05	-0.13	-0.17
2023–2024	Actual			157.00	114.00	109.00						
	DP	40	-1010.01	162.23	93.13	124.64						
	4P	156	-1443.94	164.89	92.72	122.38						
	4CP	162	-1437.52	160.39	97.19	118.61	-0.14	-0.07	0.11	0.15	-0.17	0.06
2024–2025	Actual			148.00	114.00	118.00						
	DP	40	-1011.26	152.35	94.26	133.39						
	4P	156	-1446.98	153.08	93.90	133.02						
	4CP	162	-1442.96	152.41	95.68	128.81	0.08	-0.18	0.07	0.00	-0.08	-0.04

Table C.26: 4-variate Poisson model comparison results for Serie A (2020–2025).

Serie A: Joint Half-Time / Full-Time Results

Season	Actual			4-Variate Poisson			4-Variate Copula Poisson		
	92	23	9	94.76	16.22	7.78	87.73	18.10	8.91
2020–2021	47	59	40	54.24	50.60	47.66	55.10	51.87	47.37
	11	21	77	10.68	16.64	80.41	12.61	18.42	72.52
	94	25	9	96.36	14.43	6.14	93.43	16.77	7.84
2021–2022	41	50	45	53.92	56.11	44.25	50.49	53.24	44.21
	12	24	80	8.31	16.53	83.95	10.92	19.50	79.35
	93	26	9	98.91	15.71	6.40	92.54	17.79	8.42
2022–2023	57	51	46	54.49	63.01	46.57	55.91	58.86	50.48
	9	21	68	7.93	15.63	71.36	9.31	17.32	66.62
	92	22	10	98.02	15.43	6.94	96.02	18.14	8.67
2023–2024	54	70	38	58.98	62.41	46.44	54.12	60.82	43.29
	11	22	61	7.89	14.88	69.00	10.25	18.23	66.65
	77	19	9	86.54	13.91	5.65	79.87	15.80	6.92
2024–2025	60	68	41	57.85	63.00	47.48	62.79	61.82	48.12
	11	27	68	8.69	16.99	79.90	9.76	18.07	73.78

Table C.27: Joint half-time/full-time transition matrices for Serie A (2020–2025).

La Liga

Season	Model	Df	LogLik	Fit			Dependence Parameters					
				Home	Draw	Away	r_{12}	r_{13}	r_{14}	r_{23}	r_{24}	r_{34}
2020–2021	Actual			155.00	117.00	108.00						
	DP	40	-1009.28	160.89	96.60	122.51						
	4P	156	-1436.05	160.92	97.06	122.02						
	4CP	162	-1431.24	156.84	102.51	117.88	0.17	-0.05	0.00	0.18	-0.10	-0.07
2021–2022	Actual			167.00	105.00	108.00						
	DP	40	-1041.37	167.89	97.41	114.71						
	4P	156	-1449.98	166.29	99.03	114.68						
	4CP	162	-1446.19	160.84	108.32	107.56	0.14	0.00	0.04	0.27	-0.05	0.01
2022–2023	Actual			180.00	94.00	106.00						
	DP	40	-1001.57	173.67	96.16	110.17						
	4P	156	-1420.87	174.94	95.96	109.10						
	4CP	162	-1415.87	171.68	98.58	106.70	0.17	-0.03	0.03	0.11	-0.13	-0.13
2023–2024	Actual			167.00	103.00	110.00						
	DP	40	-1037.39	165.52	94.57	119.90						
	4P	156	-1449.14	164.07	95.49	120.43						
	4CP	162	-1443.73	156.77	104.07	114.49	0.08	0.10	0.12	0.19	0.11	0.08
2024–2025	Actual			163.00	102.00	115.00						
	DP	40	-1011.51	163.15	94.41	122.44						
	4P	156	-1464.42	160.58	95.25	124.15						
	4CP	162	-1451.38	160.32	98.96	116.01	-0.02	-0.34	0.15	0.07	-0.09	-0.06

Table C.28: 4-variate Poisson model comparison results for La Liga (2020–2025).

La Liga: Joint Half-Time / Full-Time Results

Season	Actual			4-Variate Poisson			4-Variate Copula Poisson		
	87	25	6	97.39	16.18	6.45	89.12	17.39	7.28
2020–2021	63	69	41	56.79	65.77	42.41	59.03	68.27	47.82
	5	23	61	6.73	15.11	73.16	8.68	16.86	62.77
	102	14	8	98.62	14.83	6.46	91.51	15.81	6.92
2021–2022	57	69	49	59.62	68.33	45.31	58.76	74.26	48.76
	8	22	51	8.05	15.87	62.90	10.57	18.24	51.88
	106	18	4	109.58	17.10	7.43	100.85	18.51	8.64
2022–2023	67	58	53	57.91	64.78	45.30	61.99	65.02	50.73
	7	18	49	7.45	14.09	56.37	8.83	15.06	47.33
	95	26	4	98.00	16.04	6.42	93.41	16.21	6.26
2023–2024	59	66	36	57.65	64.30	42.57	54.85	72.22	41.29
	13	11	70	8.43	15.15	71.45	8.51	15.65	66.94
	94	22	13	103.78	18.20	8.27	94.41	22.65	11.19
2024–2025	60	60	52	50.77	64.15	52.27	56.91	60.08	47.70
	9	20	50	6.03	12.90	63.62	9.00	16.22	57.12

Table C.29: Joint half-time/full-time transition matrices for La Liga (2020–2025).

Bundesliga

Season	Model	Df	LogLik	Fit			Dependence Parameters					
				Home	Draw	Away	r_{12}	r_{13}	r_{14}	r_{23}	r_{24}	r_{34}
2020–2021	Actual			155.00	117.00	108.00						
	DP	40	-1009.28	160.89	96.60	122.51						
	4P	156	-1436.05	160.92	97.06	122.02						
	4CP	162	-1431.24	156.84	102.51	117.88	0.17	-0.05	0.00	0.18	-0.10	-0.07
2021–2022	Actual			167.00	105.00	108.00						
	DP	40	-1041.37	167.89	97.41	114.71						
	4P	156	-1449.98	166.29	99.03	114.68						
	4CP	162	-1446.19	160.84	108.32	107.56	0.14	0.00	0.04	0.27	-0.05	0.01
2022–2023	Actual			180.00	94.00	106.00						
	DP	40	-1001.57	173.67	96.16	110.17						
	4P	156	-1420.87	174.94	95.96	109.10						
	4CP	162	-1415.87	171.68	98.58	106.70	0.17	-0.03	0.03	0.11	-0.13	-0.13
2023–2024	Actual			167.00	103.00	110.00						
	DP	40	-1037.39	165.52	94.57	119.90						
	4P	156	-1449.14	164.07	95.49	120.43						
	4CP	162	-1443.73	156.77	104.07	114.49	0.08	0.10	0.12	0.19	0.11	0.08
2024–2025	Actual			163.00	102.00	115.00						
	DP	40	-1011.51	163.15	94.41	122.44						
	4P	156	-1464.42	160.58	95.25	124.15						
	4CP	162	-1451.38	160.32	98.96	116.01	-0.02	-0.34	0.15	0.07	-0.09	-0.06

Table C.30: 4-variate Poisson model comparison results for the Bundesliga (2020–2025).

Bundesliga: Joint Half-Time / Full-Time Results

Season	Actual			4-Variate Poisson			4-Variate Copula Poisson		
	69	16	5	78.68	12.47	6.07	67.03	11.53	5.44
2020–2021	50	53	34	46.83	42.20	34.92	54.00	49.36	41.10
	11	12	56	9.21	13.96	61.62	7.31	12.36	52.11
	89	16	6	88.84	11.91	5.39	88.39	13.19	6.20
2021–2022	40	35	30	46.65	39.70	29.18	42.03	35.02	28.11
	10	25	54	8.48	13.94	60.89	9.95	15.66	60.80
	97	14	7	101.90	14.72	6.79	97.82	15.15	7.24
2022–2023	40	45	33	43.12	37.74	31.46	42.79	35.06	33.65
	10	14	46	7.08	12.62	50.55	7.81	12.94	48.33
	84	17	7	91.43	11.63	5.47	85.34	12.71	6.83
2023–2024	40	46	27	45.36	36.84	30.45	42.94	35.39	33.04
	11	18	56	8.61	13.39	62.81	8.49	14.16	60.08
	80	12	9	78.43	13.15	6.57	75.17	13.11	6.74
2024–2025	31	43	32	39.01	40.71	34.13	38.70	40.40	35.42
	8	20	71	7.82	14.18	72.00	7.55	14.08	69.96

Table C.31: Joint half-time/full-time transition matrices for the Bundesliga (2020–2025).

Ligue 1

Season	Model	Df	LogLik	Fit			Dependence Parameters					
				Home	Draw	Away	r_{12}	r_{13}	r_{14}	r_{23}	r_{24}	r_{34}
2020–2021	Actual			142.00	94.00	144.00						
	DP	40	-1059.75	147.25	89.91	142.84						
	4P	156	-1527.55	145.42	89.72	144.86						
	4CP	162	-1524.44	144.08	89.18	143.29	0.01	-0.12	-0.03	0.05	-0.08	-0.09
2021–2022	Actual			165.00	100.00	115.00						
	DP	40	-1062.88	170.50	89.62	119.89						
	4P	156	-1540.80	168.22	89.83	121.94						
	4CP	162	-1530.26	164.99	91.95	118.88	0.03	-0.11	0.07	0.21	-0.12	-0.22
2022–2023	Actual			164.00	94.00	122.00						
	DP	40	-1058.05	159.16	88.78	132.07						
	4P	156	-1520.50	156.59	89.03	134.38						
	4CP	162	-1517.29	152.67	90.71	132.32	0.00	-0.11	0.00	0.08	-0.15	-0.03
2023–2024	Actual			118.00	83.00	105.00						
	DP	36	-852.33	126.54	75.36	104.10						
	4P	140	-1197.98	128.26	75.08	102.65						
	4CP	146	-1191.64	124.60	82.41	95.76	0.25	-0.01	0.00	0.22	0.02	0.05
2024–2025	Actual			140.00	67.00	98.00						
	DP	36	-874.96	131.22	68.80	104.98						
	4P	140	-1245.80	132.32	67.74	104.93						
	4CP	146	-1244.11	127.55	67.58	104.81	-0.05	0.06	-0.03	0.02	-0.13	0.01

Table C.32: 4-variate Poisson model comparison results for Ligue 1 (2020–2025).

Ligue 1: Joint Half-Time / Full-Time Results

Season	Actual			4-Variate Poisson			4-Variate Copula Poisson		
	91	21	8	89.50	16.89	7.61	85.73	18.17	8.58
2020–2021	38	55	47	48.22	56.64	48.99	49.14	53.15	50.97
	13	18	89	7.70	16.19	88.26	9.21	17.86	83.74
	97	13	12	100.45	15.30	6.35	91.29	18.76	9.09
2021–2022	49	59	47	57.63	56.72	42.13	59.03	52.02	46.63
	19	28	56	10.14	17.81	73.46	14.67	21.18	63.16
	97	19	8	97.55	15.79	6.68	92.60	17.49	7.80
2022–2023	53	59	38	50.77	56.91	45.01	50.15	54.82	46.87
	14	16	76	8.27	16.33	82.69	9.92	18.39	77.66
	74	16	9	78.41	13.73	7.00	72.20	13.51	6.37
2023–2024	41	54	44	43.38	48.52	39.29	44.99	55.54	42.78
	3	13	52	6.48	12.84	56.36	7.41	13.36	46.61
	86	16	2	78.91	11.52	5.39	77.85	11.67	5.34
2024–2025	41	33	34	45.21	42.32	34.83	41.58	41.37	35.61
	13	18	62	8.20	13.91	64.71	8.12	14.55	63.85

Table C.33: Joint half-time/full-time transition matrices for Ligue 1 (2020–2025).