Norwegian School of Economics
Bergen, Spring 2022

Dissecting Poisson based prediction models in association football

A comprehensive look at methodology, assumptions, and accuracy using data from the main European Leagues (2011 – 2022)

Veton Kurtsmajlaj

Supervisor: Mario Guajardo

Master thesis, Economics and Business Administration

Major: Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Congratulations to A.C. Milan for the 2021 – 2022 Serie A title

Acknowledgments

Ahead I'd like to extend my reconnection to the staff at NHH. Despite the unusual circumstances surrounding the last two years, they ensured we had the opportunity to study and carry out our exams and obligations. Although ultimately, I didn't have the chance to join you in Norway, it was a great learning experience. I'd like to express my gratitude to my family and friends for their unconditional support despite my ongoing studies and this being a second Master's. A special thanks go to my team and colleagues at TheCityStats and Vigan Group, who took up the extra burden at work while I was busy writing this thesis. A special thanks also goes to my supervisor, Mario Guajardo, for his unreserved help and suggestions, although this work was submitted to him only in the last days before the deadline. Finally, I want to mention Leonardo Egidi of Universita degli Studi di Trieste for his courtesy in answering my questions. His research and the R – package '*footBayes*' are significant contributors to the present work.

Norwegian School of Economics

Bergen, 2022

Veton Kurtsmajlaj

Abstract

As the access to broader and better data increases, data analytics, statistical modeling, and data science generally find ever-growing interest in sports analytics, including association football. It is no secret that both clubs and even higher governing bodies in the sport implement data-driven strategies to give them insights and a competitive advantage in play. Recognizing the importance of the sport as a fan and from the point of view of an analyst, this work seeks to contribute to the current body of literature by offering a thorough investigation of one of the most elegant approaches to sports analytics in association football; The Poisson goal model. Based on the simple and intuitive idea that goals in football are rare discrete events that follow the Poisson distribution while conditional on team performance, the concept has been appealing to many researchers. At the same time, a simplistic idea at its core, its application to real-world data, has been met with much discussion regarding underlying assumptions and methodology. Much of the discussion in the last 40 years since the idea was formalized concerns addressing assumptions such as the applicability of the Poisson distribution, score interdependence, overdispersion, and parameter stability. In the present work, we take a step back and reexamine the idea, methodology, and assumptions in the light of the most recent data from Europe's major leagues. Furthermore, we examine sone novel concept such as considering xG (expected goals). Overall, some changing dynamics are revealed and some of the propositions made for the model do not hold given the recent developments in the sport.

Key words: Poisson distribution; Goal model; Football; Prediction; xG

Contents

List of Figures

List of Tables

Terms and Abbreviations

hfa	Home Ground/Field Advantage
hg	Home Team Goal
ag	Away (Visiting) Team Goal
DS	Dixon - Coles Model
RS	Ruse - Salvesen Model
CMP	Conway–Maxwell–Poisson Distribution
Skellam Distribution	Poisson Difference Distribution
DP	Double Poisson Distribution
BP	Bivariate Poisson Distribution
SD	Semi-Dynamic Model
TD	Time-Dynamic Model

1. Introduction

As one of the most popular and followed sports globally, football has gone well beyond just the competitive aspect for many people. Played in over 200 countries and with over 200 million active players, it has often revealed itself as an inseparable aspect of the very culture and background of the particular society where it is embedded. Individual and team play patterns may be distinctive to specific cultures. As an example, it is a long-held tradition of Brazilian players to have particular dribbling skills as a characteristic of the "way we play" thus showing itself as an aspect of Brazil's cultural identity (Araújo and Davids, 2016; Uehara et al., 2018)

Given the incredible popularity of the sport and its significant economic impact, an entire industry is built around it worldwide. Like other industries, modern data engineering and computerization play an ever-increasing role in the sports industry. In all aspects of the sports industry data-based, decision making is becoming a fundamental focus point (Mondello & Kamke, 2014). Such practices are encountered from the highest governing bodies of football, such as national federations, down to individual clubs. Clubs use statistical modeling to decide on ticket pricing, player rotation, injury prevention, and winning chances (Link, 2018). Even national federations and FIFA use modeling to reach important decisions (Monks & Husch, 2009). The gambling industry especially stands at the forefront of such developments, given the need to optimize odds and maintain their edge (Pearson, 2017). Various predictive models based on in-game team performance have shown promise in simulated settings. In the modern sport of football, monetization, professionalism, and optimization have shifted into primary focus and affected the game in many aspects. As the sport draws more high-paid and motivated professionals, the disparity between teams increases. These factors have made the game even more unbalanced and more predictable (Maimone and Yasseri, 2021)

Although a variety of analytical methods used in football target many indicators, this work at this time concerns modeling the final results of a single match between the home and vising side in terms of goals scored. Conceptually there are two ways of framing a final result prediction model in football. While one way is to model and forecast the potential number of goals that each team may score in 90 minutes, the other uses a classification algorithm to fit the likely outcome, such as a home win, irrespective of the goals scored. The classification models either fit a three-way result (home, draw, away) with ordered regressions (Koning, 2000) or make use of logistic regression and other classification algorithms to predict the likelihood of a particular single result such as home win or over 2.5 goals (Carpita et al., 2015, 2019; Prasetio and Harlili 2016). Although these models may be straightforward, have a simpler structure, and avoid interdependence, they are nested within the goal-based framework. It is clear that goals scored by the two sides determine the final match result, but from the point of view of the classification algorithms, results such as 1 – 0, 5 – 0, or 3 – 2 all result in home wins. Intuitively enough, one can conclude that it's possible to derive the outcome of the classification algorithms from that of the goal models, but the contrary isn't. (Egidi & Torelli, 2020).

It has been a long-established tradition since Maher (1982) to model the goal-scoring process of the home and away teams as Poisson random variables. The Poisson model is mainly adopted by linking team performance with distribution parameters and taking the notion that it reflects their average scoring rate in the length of a typical match. More specifically, the practice implies formulating the intensity parameter or expected goals as a function of team performance, home-field advantage, and other covariates while tuning for interdependence and parameter stability over time (Egidi & Torelli, 2020). This conceptualization becomes quite intuitive if one thinks of goals in football as rare discrete events, primarily dependent on the relative strength of the two sides facing each other. Although this simple approach offers an almost elegant alternative to modeling scores in football, several arguments rise regarding some fundamental aspects and assumptions. First, there is the argument of interdependence between the two teams. While some authors consider conditional independence (Maher, 1982), others modify the model to introduce parameter dependence between the scores (Dixon and Coles, 1997; Rue and Salvesen, 2000). Other functional structures that deal with score dependence include bivariate Poisson distribution (Karlis and Ntzoufras,

2003) or conditional independence within hierarchical Bayesian models (Baio and Blangiardo, 2010; Egidi et al.; 2018). The second major aspect discussed is parameter stability over time. While earlier models are static or use weighting to prioritize recent results (Dixon and Coles, 1997), more recent work introduces time dynamic models within either a frequentist or Bayesian approach.

Considering the vast interest, the sport generates and the quite elegant approach offered by the Poisson-based models in football prediction, the primary aim of this work is to provide a closer look and a review of the underlying assumptions these models build on and how current they are. As the concept spans the better part of the last 40 years many iterations have been proposed passed on the points laid out above and many concern data which is 10 to 20 years old. In the present work we reexamine much of the proposed models in terms of methodology and assumptions and test how they do in light of the latest data from Europe's leading football leagues. Furthermore, an attempt is made to introduce novel concepts such as xG to the models to improve predictability. The data used is the most recently available, spanning the last 12 seasons of the Premier Leagues, La Liga, Bundesliga, and Serie A. Results reconfirm some of the most basic concepts of the idea including the clear existence of a consistent hfa. In accordance with most of the literature the direct applicability of the Poisson distribution is brought into question, but the current findings are at odds with some of the solutions. More precisely there are discrepancies between some of the assumptions made in the literature and evidence from the latest data especially in terms of score dependence.

In section 2, the literature review provides an overview of the study and its evolution. Section 3 presents a formal framework built on theory focused on model specification and assumptions. Section 4 gives an overview of the data and estimation procedures. Section 5 presents the results from assumptions check and model accuracy, and finally, section 6 includes closing remarks and potential future research.

2. Literature Review

2.1. Background

Thomson (1975) proposed for the first time an approach similar to what would be used later for Poisson modeling in football. He uses maximum likelihood estimators to optimize team matching and quantifies other key team metrics. However, these attempts regard the 1973 season of NFL rather than European football. After Moroney (1956) indicated the applicability of the Poisson distribution to scoring in football, Maher (1982) introduced a similar approach linking the concept to team skills. MLE optimizes four parameters; home attack (α_i), home defense (β_i), away attack (α_j), and away defense (β_j) quantifying relative team strength in the context of the Poisson distribution. Another essential element that is clear to anyone who knows football is the so-called home-field advantage (γ). When looking at historical data, evidence from the early 80' and 90' points to hfa, linking primarily to crowd support, especially the size and density (Barnett, 1993). This approach has been widely used to predict the number of goals scored in football and served as a basis for several related works (Boshnakov et al., 2017; Koopman and Lit, 2015; Koopman and Lit, 2019; Owen, 2011).

The Poisson framework estimates the outcome probabilities after linking the scoring capabilities with the team's performance parameters and hfa. This connection is considered natural within the literature, as goals in football are regarded as discrete and rare events (Boshnakov et al., 2016). Based on goal interdependence, Egidi and Torelli (2020) broadly group Poisson based models into three main categories: Skellam or Poisson difference (Karlis & Ntzoufras 2009); compound Poisson distribution (Maher, 1982; Baio & Blangiardo, 2010; Egidi et al., 2018) and bivariate Poisson distribution (Dixon and Coles, 1997; Karlis and Ntzoufras, 2003). A secondary distinction is also made by considering parameter stability over time and diving the models into static or time dynamic (Egidi et al., 2018). With the advent of technology, much play data is recorded, and in-game analysis of every action is possible. From the analysis of every possible shot taken, the concept of xG (expected goals) arises as to the probability of a single shot resulting in a goal based on the characteristics and events leading up to it. However, expected goals are a more general performance metric as the final result alone seldom doesn't reflect all that happened on the pitch (Schmidt, 2020). These distinctions in assumptions are the primary target for evaluation in the present work. In the subsequent subsection a broader view of the development history is given alongside the main point each author makes and how we approach them.

2.2. History and Development

The application of goal-based modeling in sports analytics has a long history, and it's not only limited to football. Earlier work mainly favors adopting the negative binomial or "modified" Poisson as a better approximation for the distribution of scores, particularly by taking care of what sometimes appeared to be an overdispersion of results (Pollard, 1985). Such models apply to various ball games, including basketball and volleyball (Reep et al.; 1971). Maher (1982) is the first iteration where the Poisson distribution is used to model goals scored around a team's performance in football. The author examines the concept, assuming score independence and lacking it, resulting in good data approximation. Furthermore, Maher (1982) argues about the applicability of the Poisson distribution recognizing the fact that score dependence would affect the rates. Later this fact becomes a theme in many discussions, including the present work.

Dixon and Coles (1997) make a general review of the research on sports analytics and recognize the apparent dichotomy of long-run vs. short-run predictions. They take the position that, although many factors do indeed contribute to the evident noise we see in the results of a single match, it is still possible to model relative team strengths within the setting of a match. The authors look assumptions and applicability of the Poisson distribution and, except for minor corrections in the low scores, conclude that the home and away goals are generally mutually independent. Both the first and subsequent authors, such as Karlis and Ntzoufras (2003), consider positive correlations among sores which, according to them, inflate the probability of draws. Thus, a positive correlation is a prerequisite for their adjustment. Testing these assumptions is part of the aims of the present work. The introduction of weighting schemes and time-dependent variation to the models reflects each team's change in parameters. The corpus of this work remains relevant in score modeling and especially for optimizing betting strategies. Dixon and Coles (1997) propose an easy-to-implement modeling strategy that provides a semi-dynamic structure to parameter stability over time. It is one of the most popular reference papers in football modeling (Lindstrøm, 2014). Rue and Salvesen (2000) continue this approach by adding adjustments to the goals scoring intensities for the home and away teams. The authors add two main adjustments to their system, namely an adjustment for the intensity parameter for the home and away team capturing psychological effects and introducing a time-dynamic component to the model with Gaussian priors. These two elements are considered separately in this work, with the psychological extension added to the DS model as an additional covariate. Time-dynamic models are treated alongside the propositions of Egidi et al.; (2018).

One of the most critical underlying assumptions presented so far is the lack of interdependence between the scoring expectations of the home and away teams. Further modifications to the Poisson framework can address the assumption entirely. Karlis and Ntzoufras (2003) suggest using a bivariate Poisson distribution to manage the joint probability of the home and away team having a specific score, rather than considering them as separate processes. Misspecification introduced in the model as wrongly assumed independence introduces bias, especially for low score draws. Karlis and Ntzoufras (2008) revisit the issue of interdependence and propose a novel approach by using the Skellam distribution. Modeling the goal difference instead of the goals allows the removal of any residual correlation between the two teams. Estimation follows prior work with Bayesian analysis in count data (Karlis & Ntzoufras, 2006). This approach examples the correlation framework compared to Dixon and Coles (1997), making it more generalizable but to the assumes a positive correlation. On the other hand, a Bayesian approach is valuable for modeling outcomes since it incorporates information about a game using prior distributions. Such information in the model can be based on historical data and consider many factors. Finally, the Bayesian approach offers a unique advantage in the case of sports predictions with the posterior predictive distribution. The posterior distribution allows a probabilistic forecast of future games, making possible even the simulation of an entire tournament or season. A Bayesian inference to sports analytics and football appears in various research papers focusing on national and supranational competitions (Suzuki et al., 2010; Rue and Salvesen, 2000; Owen, 2011). In the present work, the Skellam and bivariate setups are tested under a Bayesian approach

Most recently, Egidi et al.; (2018) made an extensive overview of the proposed models and adjustments in goal modeling and Poisson distribution. They look into model assumptions by addressing goal interdependence and scoring stability over time. First, by following previous research, they propose a hierarchal Bayesian model to address the interdependence of scores. Baio and Blangiardo (2010) suggest hierarchical Bayesian models allow for correlation since the observed scores mix at an upper level. Empirical evidence points to a slight positive correlation or no correlation between any two teams in national leagues (McHale & Scarf 2011). The second important aspect addressed is the temporal stability of the parameters used to calculate the attack and defense values for the home and away teams, alongside the hfa. A dynamic model addresses time variability, and the authors adopt an autoregressive model with a team fixed effect. After addressing the usual issues of interdependence and parameter stability, introducing further covariates may improve the model's predictive power. Consequently, the widespread use of betting odds in sports modeling is not a surprise since they are the most accurate information on probability estimation for match results (Štrumbelj, 2014). However, accessing any potential information held within the bookies' odds may not be straightforward. First, one needs to derive scoring intensities expressed in the implicit probabilities contained in the inverse betting odds. Then, as betting companies offer unfair odds, normalization better reflects implicit probabilities.

Models considered in the literature compute relative team strength parameters such as attack, defense, and hfa using past results in their computational framework. With the advent of technology and better access to databases and computational power, much more data on football matches are easier to obtain and process. This enormous information can be accessed and modeled, resulting in better predictions. Within computational science and statistical machine learning framework, AI and neural networks may detect powerful patterns and estimate complex and otherwise difficult-to-detect relations. These methods are particularly suitable for this problem as available data is vast in numbers, quality, and multidimensionality. Examples of the most used and popular algorithms used are Artificial Neural Network (ANN) and Support Vector Machine (SVM). One of such metrics that has found great application in recent years is xG (expected goal) computation. This metric is similar, especially to the expected goal metric used in the Poisson Model but computed using machine learning from thousands of in-game metrics such as shots, passes, corners, goal attempts, penalties, and even individual player performances (Umami et al., 2021). In general terms, xG is the probability of a particular shot resulting in a goal based on a statistical model with parameters such as the distance and angle, type of shot or set-piece, and many more. There is much potential in the application of xG as it gives a complete overview of the team play and chances. One such rendering of the data is also known as a shot map which shows the locations of shots, passes, assists, and other vital actions about goal-scoring in the form of a heatmap. This procedure allows the xG approach to be an all-encompassing indicator for team performance (Schmidt, 2020). The usefulness of the xG approach becomes apparent when one considers that just the final score of a match may not always tell the whole story in terms of preparation and performance gap. In the present work, xG is used to model scoring intensities which then are used to calculate outcome probabilities based on the double Poisson approach.

3. Theoretical Framework

This section explores and explains the various formal models tackling goal-scoring models in football, starting from the basic formulation of the Poisson distribution. As discussed during the literature review, most models follow the general concept of Maher (1982) by framing four parameters; home attack (α_i), away defense (β_j), home defense (β_i), and away attack (α_j) in the context of expected goals and the Poisson distribution. Observed differences in literature stem from how various models address specifications and assumptions to maximize prediction accuracy.

3.1. The Poisson Distribution

Named after French mathematician Siméon Denis Poisson, who proposed it in 1837, the Poisson distribution is a class of discrete probability distributions in statistics that frames event frequency over a specified period (Hayes, 2021). Occurrences happen with a known average rate over the specified period, and each event is considered independent of the previous ones. By considering goal scoring in football as a rare and discrete event expected to happen a certain number of times within a specified time frame, say a match, the parallels with the Poisson distribution become apparent. Thus, a discrete random variable X (ex. Goals scored) is said to follow a Poisson distribution with parameter $\lambda > 1$ if it has a probability mass function:

$$f(k, \lambda) = Pr(X(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \qquad (1)$$
$$\lambda > 0$$

where

- k is the number of discrete occurrences (k = 0, 1, 2 … etc.)
- e is Euler's number
- t is the time interval within which the occurrences happen
- λ is the intensity parameter, which shows the expected number of occurrences per unit of time.

Within the contest of goal modeling, the general formulation above naturally simplifies. The models concern the expected number of goals in one match; thus, the time parameter always is $t = 1$. This fact reduces equation (1) above to the more familiar form: $Pr(X(1) = k) = \frac{e^{-\lambda}\lambda^k}{k!}$, giving the probability of any k goals scored within the length of one match by either side. Another defining aspect of the distribution is that the intensity parameter λ is not just the expected value of the random discrete variable X but also its variance:

$$\lambda = E(X) = Var(X) \qquad (2)$$

3.1.1. Poisson Regression

Using the Poisson distribution entails an additional advantage for estimation as it belongs to the family of exponential distributions. By rewriting (1) in exponential terms and using the log–link function, the concept of Poisson regression arises in the context of generalized linear models. In the case of the least squared regression, the parameter of interest is the average response μ_i for each observation i where the average response formulates as a linear combination of factors. Analogously, in the Poisson regression, the shift is moved from the average response μ_i to the intensity parameter λ_i, and it is modeled as a linear combination of different factors. Nonetheless, several issues arise when trying to fit the intensity parameter directly as a linear combination of factors, such as the one done for OLS. A setting like $\lambda_i = \beta_0 + \beta_1 x_i$ doesn't work for Poisson-generated data. This setting causes issues related to potential values of λ_i, which in this formulation may be negative for specific values of x_i. This fact violates both the Poisson parameter value constraint ($\lambda > 0$) and the equality of mean and variance (2). A log–link functional form instead may consider these issues (Roback and Legler, 2020):

$$log(\lambda_i) = \beta_0 + \beta_1 x_i \qquad (3)$$

By formulating $log(\lambda_i)$ as a linear function of covariates instead of λ_i solves the problem of negative values and potentially addresses the mean variance equality.

However, like with other models, to make estimations with Poisson regression, several assumptions are required:

a) The dependent variable Y_i is random and follows the Poisson distribution
b) Observations are independent of each other
c) Equality of mean and variance
d) Linearity of parameters, $log(\lambda_i)$ is a linear combination of the covariates.

3.2. The Poisson Goal Model

The premise presented in the goal-scoring model follows the formulation of the Poisson regression in (3). By assuming an independent and Poisson distributed process for goal–scoring in football, the log scoring rate, $log(\lambda_i)$ of each team is a linear combination of that team's and its opponent's relative strengths. Many authors describe the goal-scoring process as Poisson random variables based on empirical data.

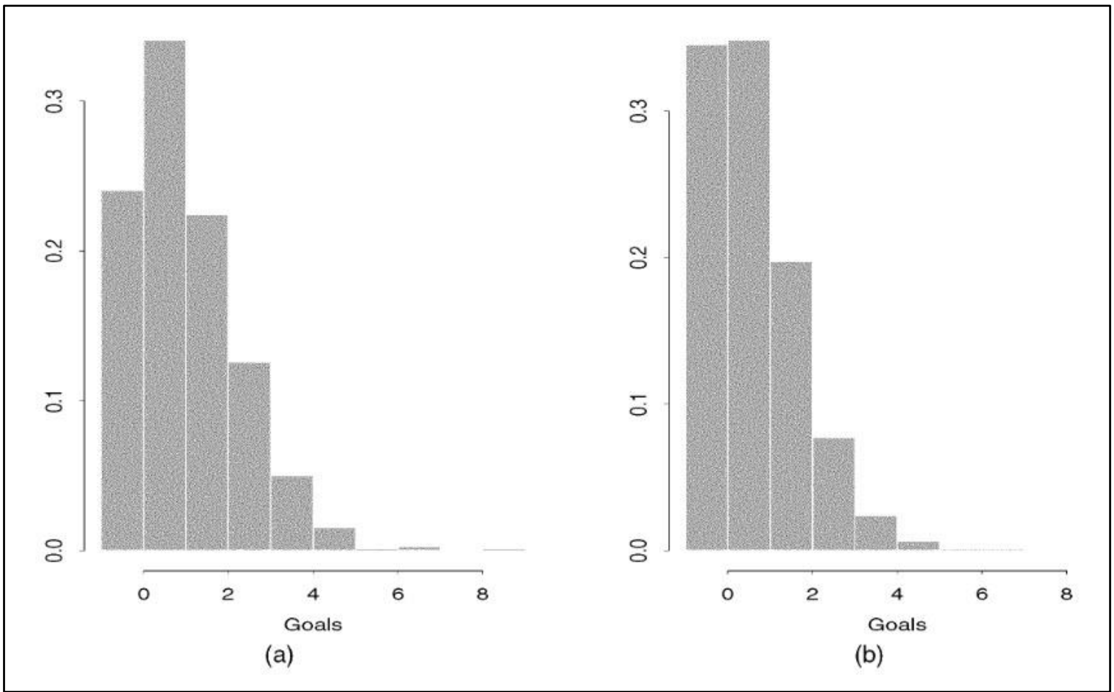

Figure 1. Histograms of (a) home and (b) away goals in 924 matches in the Premier League, 1993 – 1995. From Rue and Salvesen, 2000

The conceptualization of goal-scoring models based on the Poisson distribution encompasses a set of features required from the statistical model (Dixon and Coles, 1997):

a) The model should consider the differences between the two teams and link them to the outcome in a meaningful way.
b) Reasonably, the model should weigh the teams' recent performances.
c) The model should allow for a distinct advantage for the home team, as demonstrated by the empirical evidence from the game.
d) The model should quantify the team performance in numerous aspects, such as attack and defense strengths.
e) A team performance metric should consider the quality of recently faced teams.

Based on the points above, the basic model outlined in Maher (1982) formulates that the number of goals scored by each side are random Poisson variables, with means determined by each side's respective attack and defense parameters. More specifically, by making use of equation (1), if teams I and j face each other in a match, then the expected goals of each side x_{ij} and y_{ij} are:

$$\begin{gathered} x_{ij} \sim Poisson(\alpha_i \beta_j \gamma) \\ y_{ij} \sim Poisson(\alpha_j \beta_i) \\ \\ \alpha_i \beta_i > 0 \; \forall \; i \\ \alpha_j \beta_j > 0 \; \forall \; j \\ \gamma > 0 \end{gathered} \tag{4}$$

Where x_{ij} and y_{ij} are independent variables, and parameters α denote each team's attack metric and β each team's defense metric while γ indicates the home team advantage. Although simplified in terms of assumptions, the setting in (4) is a suitable generalization of the conceptual model, to which further modifications are applied. Expressing such a setting in terms of the Poisson regression shown in (3) gives:

$$\log(\lambda_i) = c_0' + \gamma + \alpha_i - \beta_j$$
$$\log(\lambda_j) = c_0'' + \alpha_j - \beta_i \qquad (5)$$

It follows from (4) and (5) that the probability of observing any combination of scores (x, y) from home and away teams I, j is:

$$P(x_{ij} = x, y_{ij} = y) = \frac{e^{-\lambda_i}{\lambda_i}^x}{x!}\frac{e^{-\lambda_j}{\lambda_j}^y}{y!}$$
$$\lambda_i = \alpha_i\beta_j\gamma \qquad (6)$$
$$\lambda_j = \alpha_j\beta_i$$

3.3. Modifications to the Model

The basic model conceptualization for goal-scoring in 4.2.1 and parameter estimation in 4.2.2 are good and easy-to-follow generalizations for the Poisson model in football. Nonetheless, there remain two issues regarding the model's assumptions and features.

First, the model needs to account for goal interdependence. According to the formulation in (4), the goal-scoring process needs to be completely independent for the two teams and solely dependent on their respective play parameters. As seen in the literature, several authors challenge this assumption and propose various modifications to address it. For example, some authors suggest departing from the assumption of independence and applying alternative Poisson model approximations such as the bivariate Poisson (Maher, 1982; Karlis and Ntzoufras, 2003; Boshnakov et al., 2016). Other authors favor relaxing this assumption by introducing corrections for lower-end scores (Dixon and Coles, 1997; Rue and Salvesen, 2000) or implementing conditional independence with Bayesian hierarchical models (Baio and Blangiardo, 2010; Egidi et al., 2018). At the same time, several extensions capture additional factors impacting the scoring capabilities. Secondly, there is the issue of parameter stability in time. The basic premise in (5) considers static models with unchanging parameters. However, this may not be the case as team performance may vary considerably within a season and even match to match. Various modifications adjust for this fact varying from weighting schemes emphasizing recent results (Dixon and Coles, 1997) to dynamic models with time components in more recent years.

3.3.1. Score Corrections

As discussed earlier, Dixon and Coles (1997) introduce an adjustment for low scores (at most 1), addressing interdependence and a weighting scheme to give more weight to recent performance. Therefore, the formulation in (6) is modified to allow for such corrections:

$$P\left(x_{ij} = x, y_{ij} = y\right) = \tau_{\lambda_i\lambda_j}(x, y)\frac{e^{-\lambda_i}{\lambda_i}^{x}}{x!}\frac{e^{-\lambda_j}{\lambda_j}^{y}}{y!}$$

$$\lambda_i = \alpha_i\beta_j\gamma$$
$$\lambda_j = \alpha_j\beta_i \qquad (7)$$

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

The adjustment $\tau_{\lambda_i\lambda_j}$ corrects the probability of encountering any given score combination based on ρ, the dependence parameter. The correction works in such a way as to allow complete independence for any score combination different from $x \leq 1$ and $y \leq 1$, in which cases probabilities alter. Furthermore, the authors propose a weighting function for parameter estimation in (5). The weighting scheme proposed is dynamic and allows down weighting of past matches based on internal optimization:

$$L(k \mid \alpha_e, \beta_e, \gamma) = \prod_{k=1}^{N}\left\{\tau_{\lambda_{ik}\lambda_{jk}}e^{-\lambda_{ik}}{\lambda_{ik}}^{xk}e^{-\lambda_{jk}}{\lambda_{jk}}^{yk}\right\}^{\varphi(t)} \qquad (8)$$

$$\varphi(t) = e^{-\xi t}$$

The weighting scheme works around the selection of the ξ parameter, which ensures an exponential decay for past match importance as the time horizon t moves forward. This case considers $\xi > 0$, with the static model (6) arising for $\xi = 0$. The parameter can be estimated internally to maximize prediction accuracy.

Rue and Salvesen (2000) introduce different modifications to account for additional effects and correct extreme scores. The first modification is introduced to the setting in (5) and serves as a modifier to the estimated intensities. It intends to capture any potential effects of the gap on paper between the two sides. Let $\Delta_{ij} = \frac{\alpha_i + \beta_i - \alpha_j - \beta_j}{2}$ be the average strength difference between the teams then:

$$\begin{aligned} \log(\lambda_i) &= c_0' + \alpha_i - \beta_j - \delta\Delta_{ij} \\ \log(\lambda_j) &= c_0'' + \alpha_j - \beta_i + \delta\Delta_{ij} \end{aligned} \tag{9}$$

In a given match, if team i is stronger than team j on paper, they may tend to underestimate them, $\delta > 0$. Although δ captures this type of misjudgment of stronger teams, it also can be negative if the weaker team is so perplexed as to develop an inferiority complex from the stronger team.

3.3.2. Poisson Bivariate Model

The models presented above still operate under a double Poisson setup, claiming independence of the scoring intensities. Therefore, modifications proposed to address potential deviations from independence mainly come as adjustments for certain score bands. Nonetheless, as shown by Karlis and Ntzoufras (2003), wrongly assuming independence will bias estimates complementarily to the correlation between the two teams. Moreover, empirical data from some authors claim a significant correlation between the scoring processes of the two sides (Lee, 1997; Karlis and Ntzoufras, 2003).

Figure 2 details the bias introduced in the probabilities of a draw for different levels of correlation λ_{ij} between the two sides as calculated by the authors in the reference paper. Calculations use a fixed rate for the home team $\lambda_i = 1$ and a varying rate for the away team $\lambda_j \coloneqq 0.1{:}\,2$. It is evident from the calculations that even marginal levels of correlation between 0.1 and 0.2 do affect the probability of a draw considerably (10% - 20%), especially for matches with low score expectations.

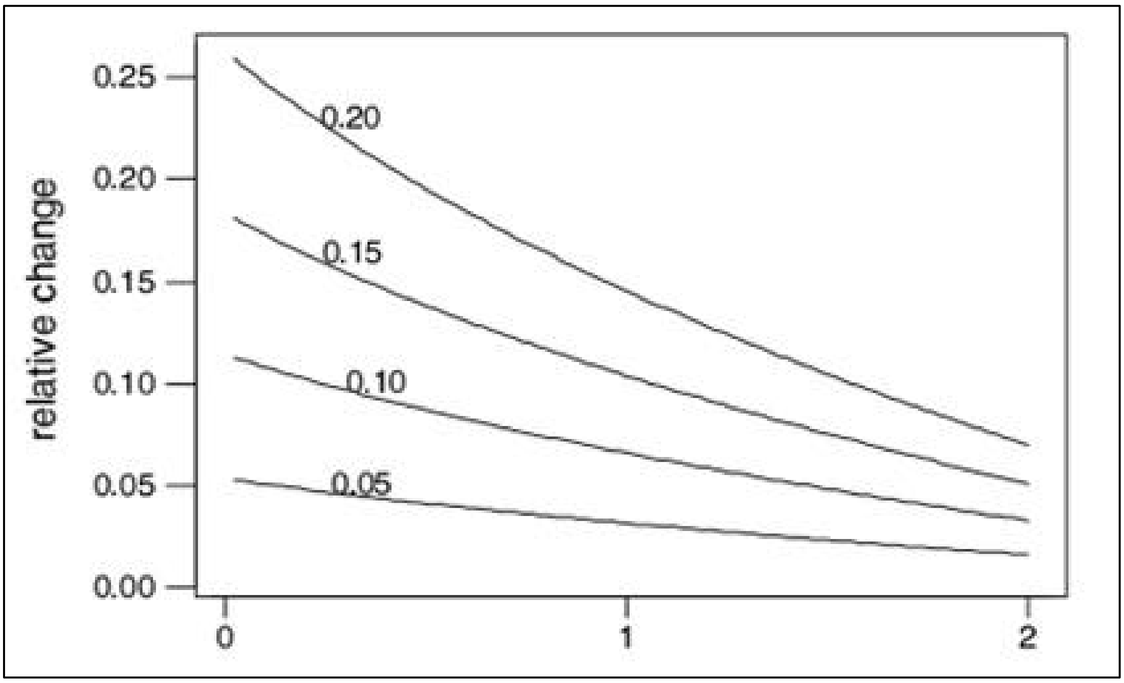

Figure 2. Relative change in the probability of a draw for different levels of correlation between the two scores, when $\lambda_i = 1$ and $\lambda_j \coloneqq 0.1{:}\,2$. From Karlis and Ntzoufras, 2003

Alternatively, considering a model where the goal–scoring process follows a bivariate instead of a double Poisson distribution may address the correlation problem (Kocherlakota, 1992). In this setup, the marginal distributions remain Poisson, but the two teams are allowed to have a level of correlation. Maher (1982) argued for using the bivariate Poisson distribution, but at the time, there were difficulties in estimating the results. If two random variables are such that $x_i = X_1 + X_2$ and $y_i = X_2 + X_3$, where X_k are independent random Poisson variables. Then x_i and y_i follow a bivariate Poisson distribution $PB(\lambda_i, \lambda_j, \lambda_{ij})$ with a joint probability function:

$$P(x,y) = e^{-(\lambda_i + \lambda_j + \lambda_{ij})} \frac{{\lambda_i}^x}{x!} \frac{{\lambda_j}^y}{y!} \sum_{k=0}^{\min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_{ij}}{\lambda_i \lambda_j} \right)^k \qquad (10)$$

The formulation above allows for dependence between the two variables. Marginally the two random variables continue to be Poisson, while $\lambda_{ij} = \mathrm{cov}(x, y)$ measures their dependence magnitude. It is evident that in case $\lambda_{ij} = 0$ the two variables are independent, and double Poisson applies. In terms of sports analytics parameters λ_i and λ_j reflect the net scoring ability of the two teams as we saw earlier, while λ_{ij} may represent other outside factors, which can modify the former two (Karlis and Ntzoufras, 2003). A third equation model added to the formulation in (5) estimates the covariance parameter λ_{ij}:

$$\log(\lambda_{ij}) = \beta_0 + \delta_1\beta_h + \delta_2\beta_a \qquad (11)$$

The formulation above allows for various linear estimations for the covariance parameter λ_{ij}. Parameters δ_1 and δ_2 are dummy variables that take values 0 or 1. If both are 0, the covariance parameter is constant and the same for all matches and teams. Alternatively, it can vary based on home or away team conditions based on weather δ_1 and δ_2 take values of 1. Consequently, λ_{ij} is interpreted as a random effect that impacts the marginal scoring averages by reflecting the game conditions.

3.3.3. Skellam Distribution

The Skellam distribution, also known as the Poisson difference distribution, is a discrete probability distribution reflecting the difference in count data. It derives from the difference between two independent Poisson random variables with expected values λ_2 and λ_1. The expected value for the resulting difference distribution is $K = \lambda_2 - \lambda_1$ (Skellam, 1946). Although the distribution usually applies to random variables from independent populations, Karlis and Ntzoufras (2003) show that it can also apply to dependent variables when the two have a common additive contribution, which is canalled by differencing. In a similar setting to (10), for any pair of variables $x_i = X_1 + X_2$ and $y_i = X_2 + X_3$, if X_1 and X_2 are Poisson distributed with intensities λ_i and λ_j then their difference $Z = x_i - y_i$ is a random variable with probability function:

$$Z(z|\, \lambda_i\lambda_j) = e^{-(\lambda_i + \lambda_j)} \left(\frac{\lambda_i}{\lambda_j}\right)^{\frac{z}{2}} I_{|x|}\left(2\sqrt{\lambda_i\lambda_j}\right)$$
$$E(Z) = \lambda_i - \lambda_j$$
$$Var(Z) = \lambda_i + \lambda_j \tag{12}$$

The Poisson difference can be applied to football to model the difference in scores between two teams. While the formulation in (12) addresses the expected goal difference, parameters λ_i and λ_j are again calculated as a linear combination of team metrics (5).

3.3.4. Score Dispersion

The Negative Binomial model offers a more flexible approach by allowing higher variance than the Poisson model. Although generally similar to the Poisson model in terms of scoring intensities, it can correct for overdispersion in the scores. Since the early days of the study, it has been one of the primary alternatives offered to model goal scoring in football (Moroney, 1956). The number of goals a team is expected to score given by the negative binomial distribution is:

$$P(x_{ij} = x) = \binom{r + x - 1}{x} p^r q^x \tag{13}$$

In this formulation, x is the number of goals scores with probability q before a failure occurs. Thus, the PDF of the negative binomial depends on $\mu = \frac{r(1-p)}{p}$ and $\sigma = \frac{\sqrt{r(1-p)}}{p}$. Pollard (1956) suggests a better fit of the negative binomial if the average score rate varies considerably, which he argues is also the case due to other factors that impact the game, such as the hfa. On the other hand, Conway–Maxwell–Poisson distribution (CMP) is an even more generalized approach to the Poisson distribution, correcting for both over and under dispersion, while the negative binomial addresses just the former (Shmueli, 2004).

3.3.5. Bookies' Odds as Covariates

Bookies' odds are considered one of the most accurate and publicly available sources of information on the probabilities of a match. Moreover, given many games, bet combinations, teams, and competitions, bookmakers try to be as efficient as possible in making the best prediction and keeping their edge (Štrumbelj, 2014). Therefore, it is possible within the setting of the Poisson regression to introduce information from the historic odds as covariates in the intensity regressions (5) in reverse form.

On the other hand, the odds need to be normalized before they can be used to accurately represent the probabilities of a specific outcome in a match. For instance, the posted odds of a particular bookie in the match Inter – Genoa played on 21/08/2021 for Serie A's 2021/2022 season were 1.33 for a home win, 5.25 for a draw, and 9 for an away win. However, as the odds represent the inverted perceived probabilities, they are 75.19%, 19.05%, and 11.11% for each result. Noticeably the sum of these probabilities is not 100% as expected, but 105.35%. This imbalance emerges because the odds posted by the bookies are 'unfair' as they seek to keep an edge on the punters, which in this case is 5.35%. Therefore, to normalize the probabilities and eliminate the bookie's edge, it's enough to divide each probability with the booksum. This procedure is known as simple normalization, and it's one of the most used methods precisely because of its simplicity (Štrumbelj, 2014).

3.3.6. Time Dynamic Models

Lastly, it is essential to consider the time stability of the estimated parameters. Rue and Salvesen (2000) propose a generalized linear Bayesian model for parameter estimation and stability. Each of the team's metrics at time t draws randomly from a normal distribution centered around values at time $t-1$. Time dynamic models depart from the ad – hoc modification in (8), with the parameters of attack and defense allowed to vary based on past matches for any team i. Variance is inflated gradually for further back matches to ensure a loss of information. Egidi et al. (2018) follow a similar approach but generalize the concept to an entire season rather than just the previous match. In this case, the parameters of attack and defense of each team in season t are allowed to vary around the values of season $t-1$, plus some constant:

$$\alpha_{it} \sim N(\mu_\alpha + \alpha_{it-1}, \sigma_{\alpha i}^2)$$
$$\mu_\alpha \sim N(0{,}10)$$
$$\sigma_{\alpha i}^2 \sim InvGamma(0.001, 0.001) \tag{14}$$

4. Data and Methodology

4.1. Data Overview

The football data used in this work primarily concerns final match results in terms of scored goals and includes information on metrics such as shots, xG, or odds. Data is retrieved online on sites like https://www.football-data.co.uk/, which offers free access to a wide range of game metrics from various major leagues spanning over 30 years. Most of the data come from the link above, except for xG scores retrieved from FBRef, through "*worldfootballR*[1]". In addition, match statistics used in modeling are identified and filtered during data preparation:

- Date when the game took place
- Name of the league and season the game pertains to
- Names of the home and away teams
- Number of goals scored by the home and away teams
- Number of shots attempted by the home and away teams
- xG of the home and away teams
- Average markets odds for each three-way result

The English Premier League is considered the most prestigious competition in the World, and hence most papers and articles refer to it for prediction modeling. Nonetheless, the concepts apply to other leagues, so we consider four major European competitions: The English Premier Leagues, Spanish La Liga, German Bundesliga, and Italian Serie A. Although modeled separately, it is valuable to compare them.

As the primary objective of this work is to optimize estimated parameters and compare prediction accuracy across models, available data is split into three sets based on seasons. The latest season, 2021 – 2022, is the target for prediction, thus making it the test sample. Most models are trained on a sample of 10 seasons from 2011 to 2020, while 2021 serves as a validation set. Validation is needed to optimize aspects of some models, such as weight allocation in Dixon and Coles (11). On the other hand, models

[1] https://github.com/JaseZiv/worldfootballR

estimated using xG have fewer observations as data is available only from 2018 onwards. In its entirety, the sample contains data on 17,338 matches across four leagues and twelve seasons. Table 1 below gives a comprehensive summary of available data and samples. The differences in observations are due to differences of league size.

League	*Sample*	*No. Seasons*	*No. Matches*
	Train	10	3,060
Bundesliga	Validation	1	306
	Test	1	306
	Train	10	3,800
La Liga	Validation	1	380
	Test	1	380
	Train	10	3,800
Premier League	Validation	1	380
	Test	1	366
	Train	10	3,800
Serie A	Validation	1	380
	Test	1	380
Total			**17,338**

Table 1. Dataset summary

4.2. Methodology

4.2.1. Model Estimation

The Poisson regression is part of a class of models that do not allow for a direct relationship between the dependent variable and other covariates in the model. As pointed out earlier, the functional form expresses $log(\lambda_i)$ and not λ_i as a linear combination of other covariates, and for this class of nonlinear models, estimators such as OLS don't apply; instead, a maximum likelihood estimation is necessary in such cases. In statistical estimations, MLE is a technique used to estimate a model's parameters based on available data by fitting parameter values so that the likelihood of encountering those observations maximizes. MLE is a case of the maximum a posteriori estimation (MAP), assuming a uniform prior distribution of the parameters. As an estimator, MLE is both efficient and consistent, and as data tends to approach the actual population, MLE will approach the actual parameters (Balaban, 2018).

Given a certain model specification θ, it is derived from (3) that the mean of the Poisson distribution dependent variable Y_i is $\lambda \coloneqq E(Y|x) = e^{\theta x}$. In this case, the model specification θ is the target for MLE estimation. Given a set of potential covariates x_i and a potential Poisson target variable Y_i, the probability of observing this data given θ is:

$$L(\theta|x,Y) = p(y_i|x_i) = \prod_{i=1}^{m} \frac{e^{y_i \theta x_i} e^{-\theta x_i}}{y_i!} \qquad for\ i \in 1 \dots m \qquad (15)$$

In this way, the likelihood function $L(\theta|x,Y)$ is set in terms of θ, for which the right-hand side of the equation gives the largest possible probability. It is practical to take the log of (4) and maximize the log-likelihood, but in practice, the optimization of θ is done by computerized means. Using MLE to estimate team parameters, in this case, is quite natural as the aim is to fit the best estimators, which may explain the scores we see. Parameter estimation for the attack and defense metrics is done at a team level, while usually, the home team advantage is the same for the entire league. For any given league, there is the same number of attack parameters $\{\alpha_1 \dots \alpha_n\}$ and defense parameters $\{\beta_1 \dots \beta_n\}$ along with a home advantage parameter γ to estimate.

In a league with n teams playing each other in k matches with corresponding scores $(x_k y_k)$, the likelihood function for the parameters of each team *m* are:

$$L(k \mid \alpha_m, \beta_m, \gamma) = \prod_{k=1}^{N} e^{-\lambda_{ik}} {\lambda_{ik}}^{xk} e^{-\lambda_{jk}} {\lambda_{jk}}^{yk}$$
$$\forall\, m \in Z \mp$$

$$\lambda_i = \alpha_{ik}\beta_{jk}\gamma$$
$$\lambda_j = \alpha_{jk}\beta_{ik} \qquad (16)$$

$$s.t.$$
$$\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \alpha_j = \sum_{i=1}^{m} \beta_i = \sum_{i=1}^{m} \beta_j = 0$$

Estimation methods vary in the literature depending on the nature of the model, being either static or dynamic. Static models usually use MLE estimation, while time dynamic approaches favor a Bayesian estimation with Monte Carlo simulations (Egidi et al., 2018). A Bayesian approach is also very useful as it can predict the results as part of posterior probabilities and includes information as a part of a priori distributions.

4.2.2. Strategy and Technical Implementation

The estimation strategy resolves around the typology of the considered models and builds a framework for comparison primarily regarding their prediction accuracy. The first distinction regards parameter stability between static and time dynamic models. Nonetheless, adding weights offers the opportunity for a semi-dynamic approach. The second dimension considers the different distribution types: double Poisson, Skellam, bivariate Poisson, and Negative Binomial. Finally, static models are further re-estimated using xG and adding bookie's odds as covariates. Ultimately, this strategy aims to check model assumptions as stated in the literature review, optimize model estimation and features, and see how they affect prediction accuracy. The table below gives a summary of models to be estimated and their features:

Category	***Distribution / Specification***	***Features***			
		No Covariates	***CMP***	***Odds***	***xG***
Semi-Dynamic (Weights)	Double Poisson				
	Bivariate Poisson / Dixon - Coles				
	Bivariate Poisson / Ruse - Salvesen				
Time-Dynamic	Double Poisson				
	Bivariate Poisson				
	Skellam				

Table 2. Models and features (grey areas denote possible combinations)

Alongside testing for model assumptions, we fit over 50 models across four leagues, two model time categories, five distribution types, and various features. Many of these models in the literature come as a solution to some underlying assumptions. Considering a wide range of results and combining them with what we see from the data allows us to conclude how well they address assumptions. Categorization into semi-dynamic and dynamic addresses parameter stability. In case there are significant differences between the two signifies that time variation is an important aspect. On the other hand, different distributions and specifications address the various forms of score dependence, such as complete independence for the double Poison to score corrections in the DC model. CMP addresses issues of dispersion, while additional covariates and xG are novel propositions to achieve better accuracy. All these models are estimated using the training sample as above. Weights for semi-dynamic models are determined in the validation sample, and model accuracy across the board is asserted with the test sample.

From the computational side, this work uses R – programming in all analytical procedures carried throughout, and model estimation primarily revolves around two packages specialized in football goal modeling; goalmodel[2] and footBayes[3]. The first package offers a wide range of specific models such as Dixon – Coles, Rue – Salvesen, and more generalized models such as negative binomial and CMP. These models are static, but there is the option of fitting weights, adding covariates, and estimating with xG. The second package is based mainly on the work of Egidi et al.; (2018) and focuses on the type of underlying distribution rather than on specific model formulations. The package estimates double Poisson, Skellam, and bivariate Poisson distributions through MLE in static form. In addition, it allows for Hamiltonian Monte Carlo (HMC) estimation for time dynamic models with various a priori distributions. The second point is particularly interesting to this work in estimating time-dynamic models. For HMC, we implement four chains with 300 replications in each case, and in addition, hypothesis testing related to the assumptions uses bootstrapped inference.

4.2.3. Evaluation

Ultimately, the model evaluation focuses on prediction accuracy for the matches played in each selected league's 2021 – 2022 season. As goal models return estimated probabilities for any goal combination, they potentially aggregate in estimates for a three-way result. Based on the literature, we implement two widespread measurements for prediction accuracy, pseudo-R^2 (Egidi and Torelli, 2020; Dobson et al., 2001) and the Brier score (Brier 1950; Egidi and Torelli, 2020; da Costa et al., 2021). These two

[2] *https://github.com/opisthokonta/goalmodel*
[3] *https://github.com/LeoEgidi/footBayes*

measurements are flexible and popular for measuring the accuracy of various prediction strategies in football.

The pseudo-R^2 is the geometric mean of all the probabilities assigned to the correct score by the model. For a set of m predictions with associated probabilities p_m assigned to the actual true outcome, the pseudo-R^2 is:

$$\text{pseudo} - R^2 = \left(\prod_1^m p_m\right)^{\frac{1}{m}}$$

The second measurement used is the Brier score, defined as a mean squared error of the forecast. The variant used in this case is the multi-category Brier score, which accounts for multiple potential outcomes in multiple predictions. In this case, final results predictions in multiple matches. For a set of n predictions with a set of k potential outcomes and associated probabilities p_{nk} assigned to any potential outcome, the Brier Score (BRS) is:

$$BRS = \frac{1}{n}\sum_1^n \sum_1^k (p_{nk} - o_{nk})^2$$

$$o_{nk} = \begin{cases} 1, if\ the\ outcome\ in\ event\ n, k\ happens \\ 0, Otherwise \end{cases}$$

Building a betting strategy that can postnatally beat the market is the focus of a wide range of works concerning football score modeling (Dixon and Coles, 1997; Rue and Salvesen, 2000; Štrumbelj, 2014; Egidi et al., 2018; da Costa et al., 2021). These strategies vary from value betting, based on the Kelly criterion (Kelly, 2011), to more complex systems involving Shin's method for odds normalization (Shin, 1993). Although building a betting strategy is not part of the aims of the current work, a simple setup can be valuable for model comparison. The proposed betting strategy revolves around bets for final results (home/draw/away) based on the estimated probabilities of each model. The evaluation considers a varying cut–off point for result categorization.

By assuming a pragmatic punter, a bet of 1$ is made in each case the probability of a result from a particular model is above the given threshold. In instances where two results are above the threshold, the most likely is selected. In this case, models may be compared on different sets of matches because different models may potentially result in different betting slips. For instance, one model may assign a 0.52 probability for a home win, while another 0.48, with a 0.5 threshold. This intends to capture the fact that a rational pointer would put in their betting slip whatever bets they feel safe with.

5. Results

5.1. Model Assumptions

As addressed in the literature review and theoretical framework, there is a series of assumptions that each formulation of the model considers the scoring process of the home and away team. Such assumptions regard important properties such as the hfa, parameter stability over time, score interdependence, and even the applicability of the Poisson distribution to the scores. These assumptions are checked against available data in the four leagues in this subsection.

5.1.1. HFA and parameter stability

An overview of the data given by descriptive statistics offers an insightful view into critical aspects of the analysis, especially when considering home and away goals. In addition, a comparison between the average home and away goals, along with a longitudinal view of the data, will offer insights into the hfa and parameter stability.

League	*Matches*	*Home Goals*		*Away Goals*		*Paired Samples t-test*		
		Mean	*StdDev*	*Mean*	*StdDev*	*Diff.*	*95% Low*	*95% High*
Bundesliga	3,060	1.65	1.36	1.30	1.22	0.35***	0.29	0.41
La Liga	3,800	1.59	1.37	1.13	1.16	0.46***	0.40	0.51
Premier League	3,800	1.55	1.30	1.19	1.17	0.36***	0.31	0.41
Serie A	3,799	1.52	1.26	1.19	1.14	0.33***	0.28	0.38

Table 3. Summary Statistics & Test (based on training samples)

First, descriptive statistics support the idea of goals being rare events in football (Arastey, 2019). In over 14,000 matches, there are just over 1.5 goals scored by the home team per game and just over 1.2 by the away team. Secondly, the data provide strong evidence for a significant and consistent hfa effect in all four leagues. In round-robin tournaments, each team plays with each other side once at home and away. This makes it possible to compare the goals a team scores in the first leg against an opponent with the score of the same team against the same opponent in the second leg. In addition, this allows the comparison of the home and away goals within a repeated sample setting by observing the same team at home and away. By taking advantage of this setup, a bootstrapped paired-samples t-test with is implemented (Table 3). Results indicate that when a team plays at home, they consistently score more on average than when they play away, which is in line with the entire corpus of literature.

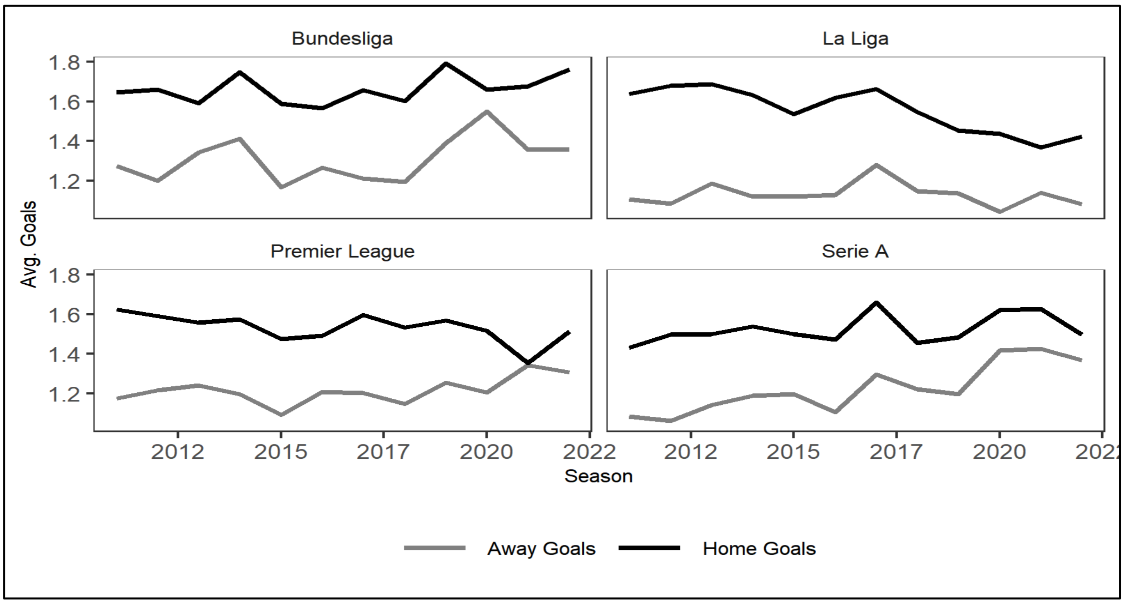

Figure 3. The average number of goals scored by the home and away teams; by league and season

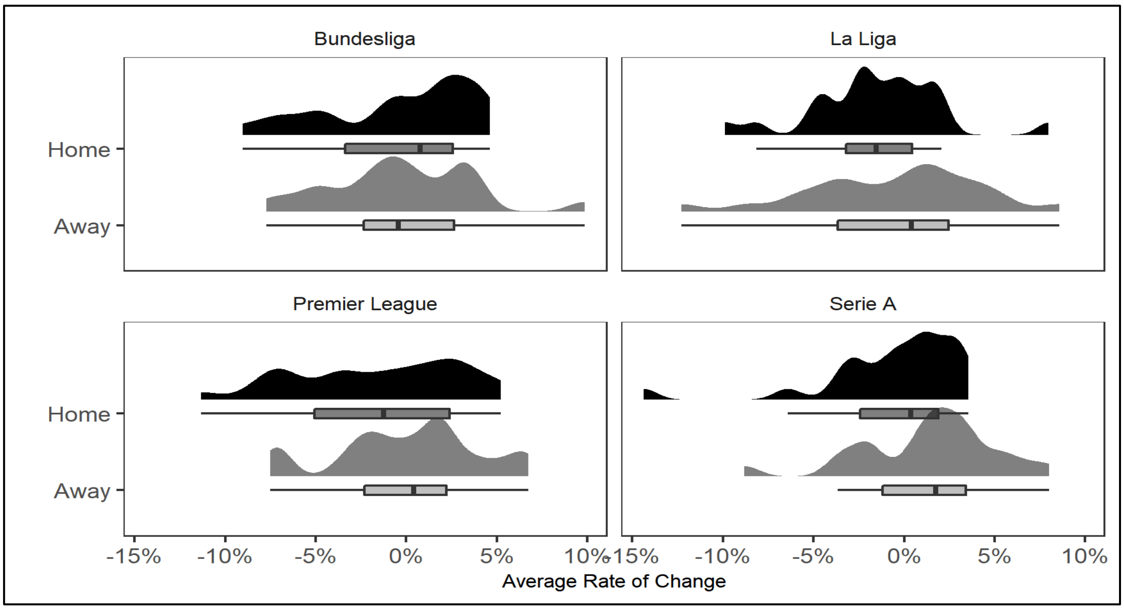

Figure 4. Distribution of average rates of change in scoring intensities across seasons at the team level[4]

[4] *Considers only teams that appear in at least 5 seasons*

The panel above, shown in Figures 3 and 4, displays a longitudinal view of the goals scored at home and away by the teams in the four leagues. Again, the average score rate at home is consistently higher than the away rate in all the leagues across all seasons. Additionally, both rates seem relatively stable over time and don't show much variation. This aspect is further explored at the team level, as displayed in figure 4. Concerning the home rate, out of 91 teams, 74 (81.32%) of them vary within ±5%, and 89 (97.80%) within ±10%. Regarding the away scoring rate, these proportions are 70 (76.92%) and 90 (98.9%), respectively. Nonetheless, results indicate a time competent impacting scoring intensities at the league and team level. These variations are in line with the corpus of literature arguing for time dynamic models.

5.1.2. Poisson Distribution Fit

First proposed by Moroney (1956), the Poisson distribution and negative binomial have been argued as a good fit for scores in football. Later the framework was expanded and formalized, as seen in the literature review. The R package fitdistrplus[5] provides functions to fit univariate distributions for both continuous and count data. Estimation methods use MLE with robust estimates based on resampling techniques. The package estimates distributions and tests their fit (Delignette-Muller and Dutang, 2015). Following Cullen and Frey (1999), a skewness and kurtosis graph may be an exploratory tool to examine likely distributions. Both Poisson and negative binomial distributions are considered based on preliminary fit and theory.

League	***Position***	***Poisson***		***Negative Binomial***		
		λ	χ^2	***Size***	μ	χ^2
Bundesliga	hg	1.65	16.22***	13.43	1.65	1.00
	ag	1.30	33.4***	8.55	1.30	3.16
La Liga	hg	1.59	53.52***	9.30	1.59	4.66
	ag	1.13	61.88***	6.69	1.13	9.95**
Premier League	hg	1.55	11.31**	18.14	1.55	1.02
	ag	1.19	39.73***	8.10	1.19	5.93
Serie A	hg	1.52	7.83*	33.87	1.52	4.45
	ag	1.19	13.46***	13.26	1.19	1.19

Table 4. Distribution and goodness–of–fit test

[5] https://cran.r-project.org/web/packages/fitdistrplus/index.html

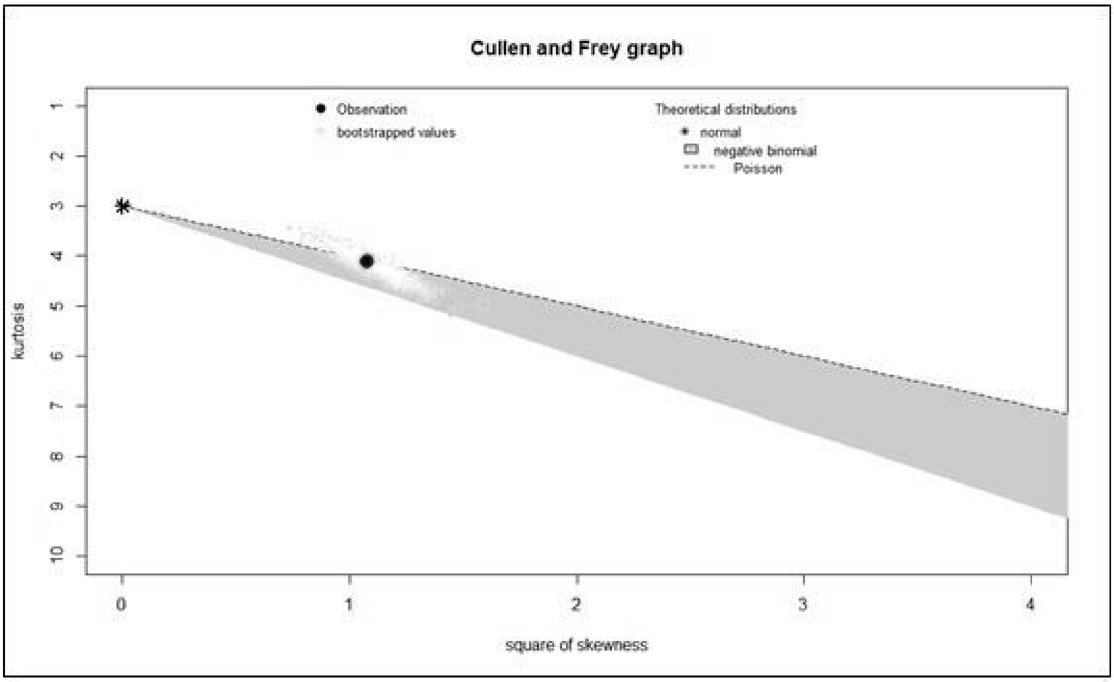

Figure 5. Skewness and Kurtosis graph of Bundesliga away goals (2011 – 2020)

Goodness–of–fit test (χ^2) suggests that the negative binomial distribution is a better fit for both the home and away goals in all leagues. These results may indicate that there is overdispersion present in the score distribution and that equality of mean and variance (2) is likely not to hold.

Preliminary results from the goodness-of-fit test seem to support the part of the literature arguing against the Poisson distribution (Pollard, 1985; Boshnakov, 1996). Nonetheless, these results only take into account the univariate distribution. Karlis and Ntzoufras (2005) argued that significant correlation present in the scores might lead to overdispersed data. Pollard (1985) argues against the Poisson distribution precisely because the hfa causes variation in the scoring rate. Maher (1982) recognizes the issue, while Dixon and Coles (1997) argue that after controlling for score dependencies, the general structure laid out in (4) still holds. Ultimately, the empirical deviation from the Poisson distribution may result from other factors not accounted for in the univariate analysis, such as score correlations.

5.1.3. Score Dependence

Score dependence is one of the main issues in the goal model, and it has been addressed empirically and methodologically in various instances (McHale and Scarf, 2011). For example, Dixon and Coles (1997) consider the empirical to marginal distribution ratio up to four goals. For any possible outcome (x_i, y_i) the ratio is given as $r = \frac{f(x_i,y_i)}{f(x_i)f(y_i)}$ with the denominator indicating the product of marginal empirical probabilities. This work follows McHale and Scarf (2011), considering unusual proportions that fall over/under two standard errors of 1, signifying an over/under-representation of certain results. In addition, a high number of significant departures from unity indicates a high score dependence. Robust estimates come from Monte Carlo simulations. The tables below show empirical ratios with bootstrapped standard errors in parenthesis and pairwise correlations.

The Bundesliga shows the highest number of significant deviations from r = 1, especially for results involving multiple goals by both sides. These results indicate a high degree of score correlation in the league. On the other hand, the Premier League shows an over-representation of once-sided results and under-representation of one-goal wins. La Liga likewise has an over-representation of deep one-sided matches but a lower number of significant deviations than the first two mentioned leagues. Lastly, Serie A has the least number of significant deviations, which mostly seem to amount to sampling error rather than anything else. Different results across leagues indicate different levels and dependence structures between scores. Results from the pairwise correlations in Table 5 serve as a general check for the following individual tables. As expected, no significant correlation is encountered in Serie A, while all other leagues show very significant negative correlations with the magnitude proportional to the number of unusual values in the following tables.

Observed correlations differ from observations made in the literature, but at the same time, different levels of correlations are stated even within the literature. Dixon and Coles (1997) detect significant correlation only in the lower scores and hence correct for them. McHale and Scarf (2011) find a significant negative correlation in scores of national teams while not finding any significant correlation in the Premier League during 2001 – 2005 in their previous work (McHale and Scarf, 2007). Karlis and Ntzoufras (2003) base their bivariate Poisson model on observations of weak but significant correlations in Serie A (1997 – 1998). Lastly, Egidi et al.; (2018) recognize the issue and argues for an implicit correction within hierarchical Bayesian models. Overall, in certain aspects, these results continue the trend found in literature and reveal new changing dynamics in score dependencies in different leagues at different points in time.

League	Pearson's r	Kendall's τ	Spearman's ρ
Bundesliga	-0.129***	-0.096***	-0.116***
La Liga	-0.076***	-0.051***	-0.06***
Premier League	-0.099***	-0.069***	-0.083***
Serie A	-0.024	-0.014	-0.016

Table 5. Bootstrapped Pairwise Correlations of home and away goals

Home Goals	Away Goals				
	0	1	2	3	4+
0	1.001 (0.055)	0.837 (0.049)	1.062 (0.07)	1.209 (0.111)	1.391 (0.158)
1	0.807 (0.039)	1.044 (0.038)	1.02 (0.053)	1.231 (0.084)	1.26 (0.116)
2	1.052 (0.049)	1.041 (0.045)	1.019 (0.064)	0.784 (0.09)	0.799 (0.12)
3	1.106 (0.073)	1.059 (0.062)	0.908 (0.084)	0.984 (0.129)	0.446 (0.132)
4+	1.346 (0.097)	1.026 (0.081)	0.884 (0.109)	0.344 (0.108)	0.62 (0.183)

Table 6. Bundesliga ratio of empirical bivariate probabilities to empirical independent bivariate probabilities (dark grey for significant overrepresentation; light gray for significant underrepresentation)

Home Goals	Away Goals				
	0	1	2	3	4+
0	0.966 (0.042)	0.945 (0.041)	0.972 (0.057)	1.177 (0.114)	1.593 (0.16)
1	0.986 (0.032)	0.984 (0.032)	1.098 (0.05)	0.941 (0.084)	0.877 (0.112)
2	0.914 (0.038)	1.065 (0.039)	1.028 (0.057)	1.049 (0.106)	0.956 (0.136)
3	1.137 (0.061)	1.003 (0.06)	0.882 (0.084)	0.823 (0.146)	0.689 (0.176)
4+	1.18 (0.073)	1.015 (0.071)	0.81 (0.095)	0.871 (0.176)	0.489 (0.174)

Table 7. La Liga ratio of empirical bivariate probabilities to empirical independent bivariate probabilities

Home Goals	Away Goals				
	0	1	2	3	4+
0	0.979 (0.042)	0.883 (0.043)	0.976 (0.058)	1.272 (0.098)	1.619 (0.17)
1	0.926 (0.033)	1.012 (0.034)	1.086 (0.048)	0.982 (0.071)	1.124 (0.123)
2	0.996 (0.04)	1.022 (0.04)	1.087 (0.057)	0.916 (0.086)	0.607 (0.122)
3	1.038 (0.061)	1.144 (0.061)	0.854 (0.08)	0.835 (0.12)	0.613 (0.17)
4+	1.312 (0.085)	0.978 (0.079)	0.68 (0.098)	0.831 (0.157)	0.595 (0.217)

Table 8. Premier League ratio of empirical bivariate probabilities to empirical independent bivariate probabilities

Home Goals	Away Goals				
	0	1	2	3	4+
0	1.056 (0.045)	0.941 (0.04)	0.995 (0.059)	0.977 (0.093)	1.166 (0.15)
1	0.932 (0.037)	0.996 (0.033)	1.036 (0.048)	1.163 (0.079)	1.036 (0.119)
2	1.001 (0.042)	1.072 (0.041)	0.995 (0.057)	0.737 (0.082)	0.955 (0.136)
3	1.063 (0.061)	0.936 (0.054)	0.957 (0.086)	1.155 (0.149)	0.944 (0.209)
4+	1.002 (0.087)	1.072 (0.083)	0.955 (0.114)	1.003 (0.196)	0.558 (0.206)

Table 9. Serie A ratio of empirical bivariate probabilities to empirical independent bivariate probabilities

5.2. Weighting Scheme

The weighting scheme introduces a semi-dynamic structure to the static models, such as Dixon and Coles (1997). Optimal weights are dependent on ξ, which is also the target for optimization. Previous work suggests testing values between 0 and 0.008 (Boshnakov et al., 2017). In the present work, the validation sample (season 2020 – 2021) is used to optimize the value of ξ. All model parameter estimations happen on the training set with different weights as determined by ξ, and selection criteria are based on validation sample predictability. More precisely, we seek to maximize the log-likelihood of the semi-dynamic model through the function $S(\xi)$, given k matches in the validation sample:

$$S(\xi) = \sum_{1}^{k} [h * \log(p_{hk}) + a * \log\,(p_{ak})]$$

Where p_{hk} and p_{ak} are the estimated probabilities for a home or away win from the training model, while h and a take the value one if the match resulted in a home win or away win, respectively, and zero otherwise. Essentially $S(\xi)$ is the sum of estimated (log) probabilities of the actual results. This work considers a weighting scheme for the double Poisson, Dixon-Coles (DS), and Rue-Salvesen models (RS). Usually, weighting is applied to the DC model, but the concept expands to the double Poisson and RS models, which is an extension of the DC.

Figure 6 below displays the relationship between ξ and $S(\xi)$ for each of the four leagues in the DC model. While this is a graphical display of the process, the complete picture is given in Table 9, which displays optimal values of ξ for all leagues and model combinations and the impact on model fit. As suggested by the literature, optimal values of ξ fall between 0 and 0.008, and in most cases, there is a clear improvement in model fit when prioritizing recent matches. La Liga is the only instance where the weighting scheme doesn't affect the model fit.

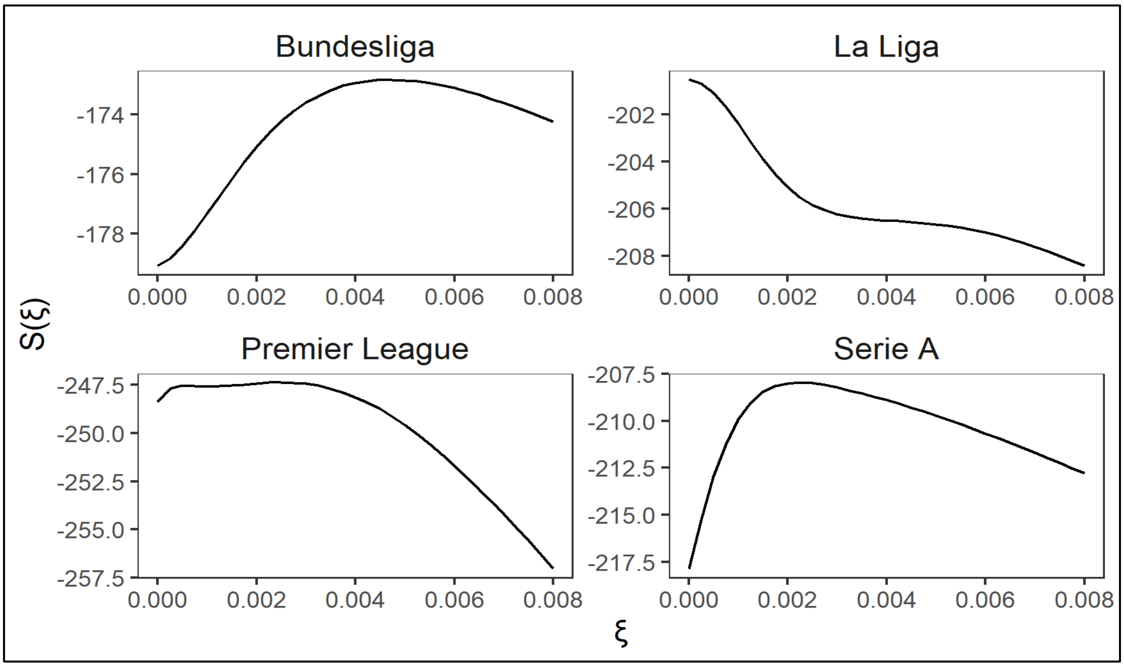

Figure 6. ξ against $S(\xi)$ - Dixon-Coles model

League	*Model*	ξ^*	$S(\xi^*)$	*ΔR²*	*ΔLog-Lik*	*ΔAIC*
	Dixon - Coles	0.0045	-172.82	4.15%	8,539.04	-17078.07
Bundesliga	Double Poisson	0.0035	-167.59	3.38%	8,385.49	-16770.98
	Rue - Salvesen	0.0050	-172.85	4.52%	8,595.21	-17190.41
	Dixon - Coles	-	-200.51	0.00%	-	-
La Liga	Double Poisson	0.0003	-199.17	-0.55%	3,756.62	-7513.23
	Rue - Salvesen	-	-200.53	0.00%	-	-
	Dixon - Coles	0.0023	-247.38	4.96%	9,703.43	-19406.86
Premier League	Double Poisson	0.0033	-240.98	6.03%	10,101.91	-20203.83
	Rue - Salvesen	0.0025	-247.37	5.25%	9,832.78	-19665.55
	Dixon - Coles	0.0023	-207.94	3.89%	9,570.16	-19140.33
Serie A	Double Poisson	0.0070	-199.95	11.55%	10,422.76	-20845.53
	Rue - Salvesen	0.0023	-207.93	3.89%	9,570.17	-19140.33

Table 10. Weight optimization for ξ^* and changes on model fit from $\xi = 0$

5.3. Model Performance

Accuracy is primarily addressed at the model level using the Brier score and Pseudo – R^2, and further comparisons consider how these indicators change across leagues. The same consideration applies when looking at the average income from the betting strategy above. Results indicate that differences across leagues are the main aspect to consider when modeling scores. Table 11 shows the average results in terms of accuracy across model types. When excluding league variation from the results, accuracy is very similar across all models, and differences are only marginal. These results are consistent with Egidi & Torelli (2020), finding minor differences across various models with similar magnitudes of BRS score and Pseudo – R^2. Subsequent panels give a broader picture of the differences in accuracy across leagues. Variation in model accuracy across leagues is due to the differences noted in league parameters during the assumptions check. Subsequent panels display differences across leagues, alongside the number of matches used to calculate accuracies. The way the package *goalmodel* works makes it impossible to predict games it has never encountered in previous years, reducing the number of matches used for testing. Nonetheless, we ensure that all models are tested on the same matches, and even after dropping the xG model, which has the least number of available predictions, the results do not change[6].

Type	***Model***	***Measure***	
		BRS	***Pseudo – R^2***
SD	SD_DC_Base	0.6053	0.3634
	SD_DC_CMP	0.6058	0.3631
	SD_DC_Odds	0.6057	0.3631
	SD_DP_Base	0.6075	0.3625
	SD_DP_CMP	0.6091	0.3616
	SD_DP_Odds	0.6081	0.3621
	SD_DP_xG	0.6029	0.3654
	SD_RS_Base	0.6061	0.3630
	SD_RS_CMP	0.6066	0.3626
	SD_RS_Odds	0.6064	0.3627
TD	TD_BP_Base	0.6045	0.3642
	TD_DP_Base	0.6026	0.3649
	TD_Skellam_Base	0.6046	0.3636

Table 11[7]. Average Brier Score and Pseudo – R^2 **across models**

[6] See figures 10, 11 in the appendix

[7] Model names are abbreviated. See Table 12 in the appendix for complete names

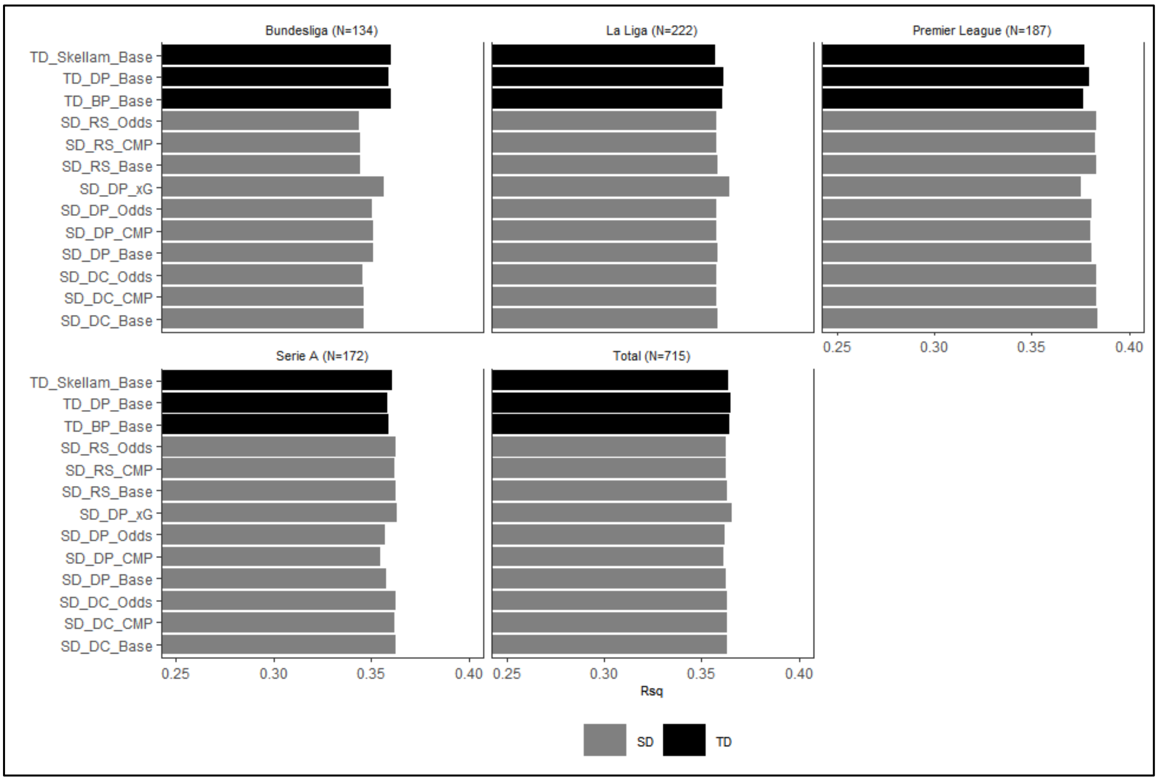

Figure 7. Pseudo – R^2 across leagues and models

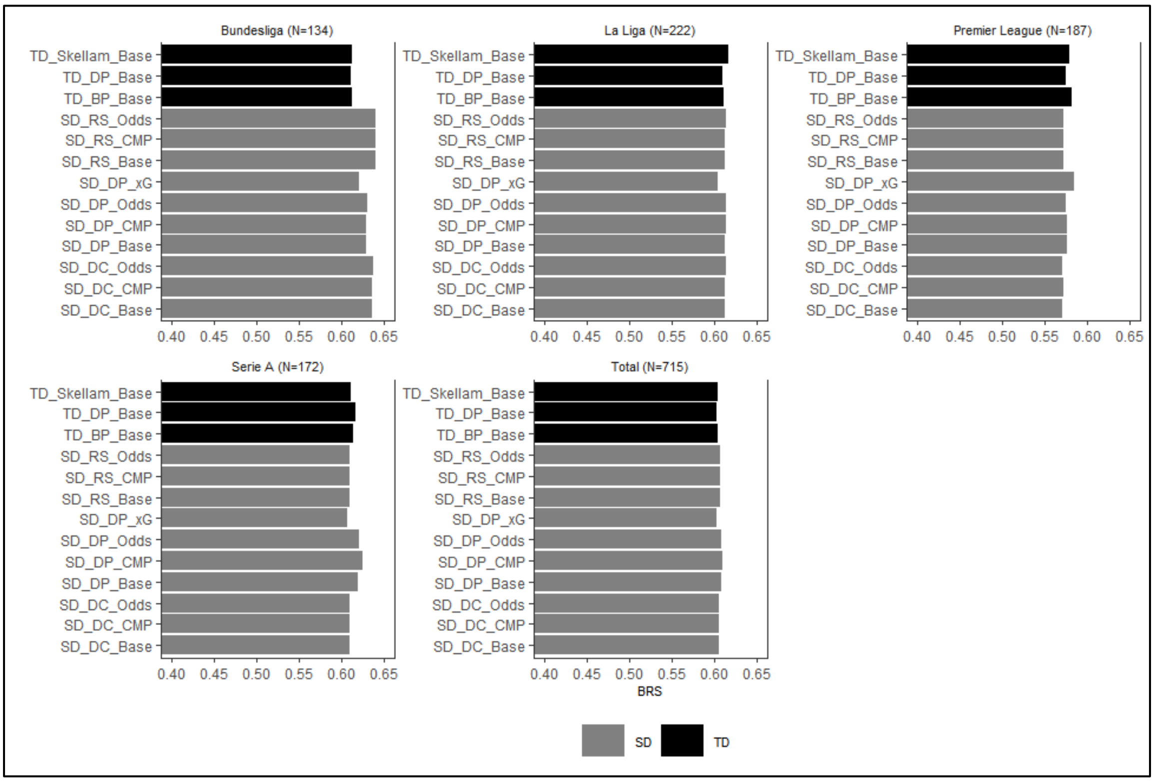

Figure 8. Brier Score across leagues and models

Da Costa et al. (2021) noted that accuracy variations, with specific models performing better in particular leagues. Accuracy results are consistent, with higher values of predictability (Pseudo – R^2) associated with lower values of average error (Brier score). Time dynamic models have the best performance within the Bundesliga, outperforming all other formulations, but with no differences across the three types of distributions (Skellam, Double Poisson, Bivariate Poisson). In La Liga, time-dynamic double Poisson and bivariate Poisson models performed the best, but with no significant difference compared to any other model. The English Premier League is the only instance where the semi-static models seem to outperform the time dynamic models, especially regarding the Dixon – Coles formulation. There is no distinction between dynamic and semi-dynamic models regarding Serie A, but there is an unusually low accuracy regarding double Poisson models. Except for a relatively good performance of the xG model in Bundesliga and La Liga, no other model containing additional covariates or accounting for dispersion (CMP) has any particular performance in any leagues. Lastly, as seen earlier, totals provide no new information as they simply average out differences present across leagues.

Figure 9 displays the results from the betting strategy elaborated at the end of section 4. Again, the betting strategy reflects accuracy results, and no method yields any profitability in the long run across leagues and considering different thresholds. In the Bundesliga, the only instances of some profitable bets involve the time dynamic, which shows the best accuracy in this league with the broadest margin observed overall. Even using a safer threshold at 70% doesn't ensure any consistent returns; instead, results point to the fact that profitability is connected to unique strategies applied to a specific number of teams or leagues. The same results are seen in Da Costa et al. (2021), where overall betting strategies are not profitable except for certain combinations of models and leagues. Serie A yields interesting results with a higher threshold associated with lower ROI. As a higher threshold means safer bets put on stronger sides, the negative correlation may suggest a higher number of unexpected results, such as the underdogs winning or drawing.

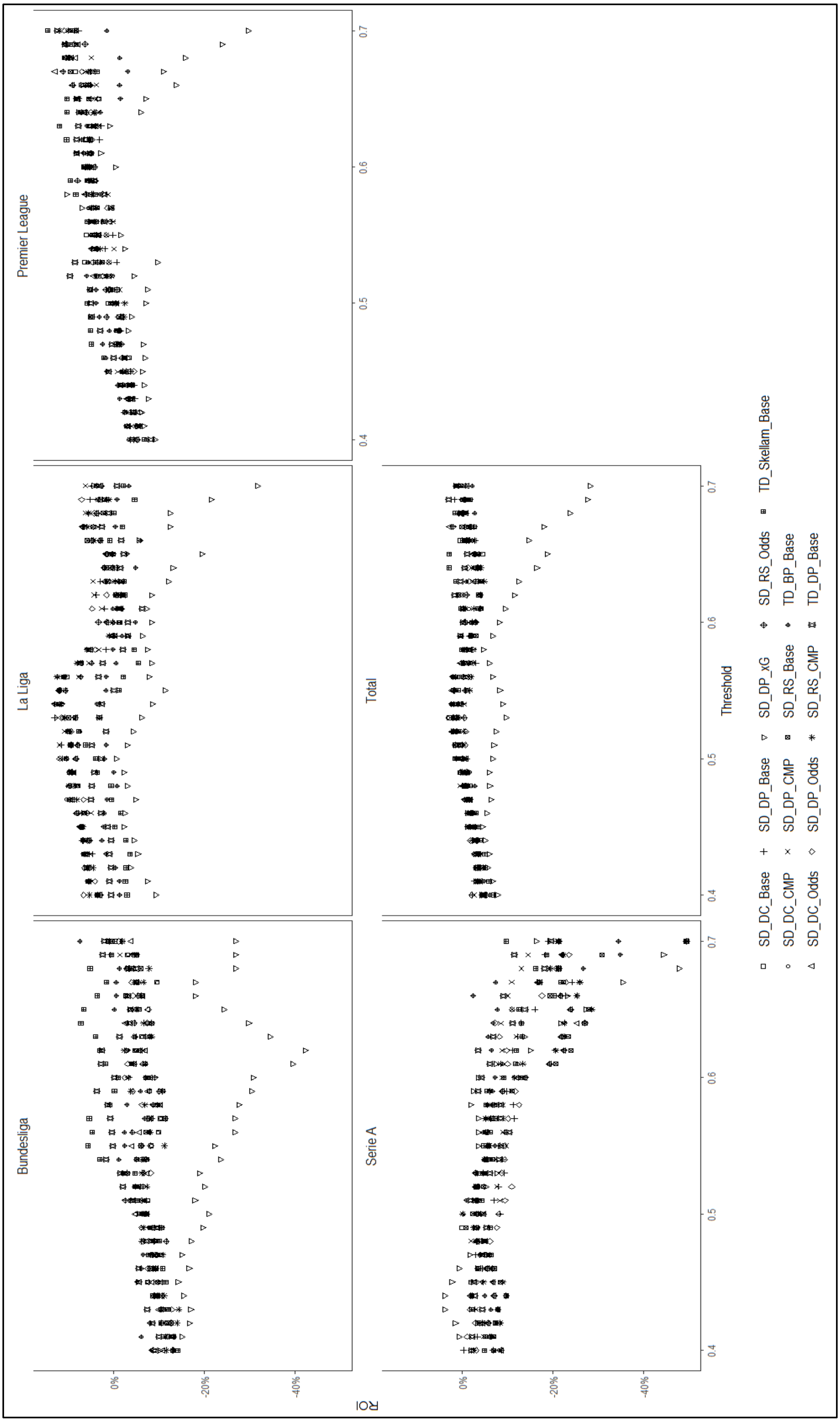

Figure 9. Model average ROI vs. decision threshold across leagues

6. Conclusions

This work offers a comprehensive look at Poisson-based prediction models as applied to association football. The intuitive and straightforward idea that the scoring process of the home and away team in football is associated with basic metrics such as team attack and defense come naturally to anyone who knows or follows the sport. Furthermore, what has come as natural to many researchers studying the sport has been linking this process to some underlying statistical distribution. Moroney (1956) first proposed the Poisson distribution for the goal-scoring process, and Maher (1985) formalized it in terms of relative team strength. Many more researchers suggested various ways to improve the model in light of new data and findings. This work contributes to this line of work by offering a thorough look at current and past developments. The review is not only in terms of model specification but by empirically testing and arguing the underlying assumptions and how they hold against the latest data from four of the major competitions in Europe and the World.

The direct connection between the Poisson distribution and the scoring process has been questioned because empirical evidence suggests a correlation between scores and overdispersion (Pollard, 1971; Maher, 1985; Dixon and Coles, 1997; Karlis and Ntzoufras, 2003). The current work agrees with these conclusions, and results show a consistent departure from the Poisson distribution in all considered leagues, at least when looking at the univariate goal distribution.

Arguably, many of the modifications and alternative propositions come from the departure from the Poisson distribution. Thus, understanding the structure of this departure has been vital in addressing it. In some aspects, the results from the current work disagree with the underlying assumption made in some of these modifications. For example, Dixon and Coles (1997) base their correction on the empirical fact that scores show a positive correlation, thus inflating the occurrences of draws. Karlis and Ntzoufras (2003) also share this conclusion. Our results are opposed and show a significant negative correlation in all the leagues but Serie A, which shows no score correlation. Negative score correlations indicate one-sided results and league disparity, which is consistent with the fact that an ever-smaller number of clubs dominates the sport (Maimone and Yasseri, 2021). For instance, Bayern München has won the last ten league titles practically unopposed in the German Bundesliga, where we also see the single largest negative score correlation. More general solutions are considered by Karlis and Ntzoufras (2003, 2008) by bivariate Poisson distribution and subsequently Skellam (Poisson difference) to account for positive correlation in general. Our results do not indicate any particular improvement in accuracy while considering these cases against the other models. In fact, not even the CMP model which supposedly accounts for both under and over dispersion shows any particular result.

The second most important aspect discussed in the literature is the question of adding a time component to parameter estimation. Results show variation is present both in the home and away scores in all leagues across seasons, although in moderate levels. While there is a clear improvement to model fit when moving from static to semi -dynamic models (Dixon and Coles, 1997), the advantage from semi-dynamic to time-dynamic models (Egidi et al., 2018; Rue & Salvesen 2000) is not quite clear. Time-dynamic models showed the best performance by far in the Bundesliga, but performed marginally worse compared to semi-dynamic models in the English Premier League. On the other hand, no major differences in performance were noted in La Liga and Serie A. Finally, beside addressing the various corrections and alternatives to the Poisson model alongside considering the time dimensions, this work attempts a limited novel approach by introducing bookies' odds and estimations via xG to the semi – dynamic models. The introduction of odds in the current form as covariates doesn't seem to contribute significantly in the current framework. In the same light considering estimates with xG, appears to yield conflicting results. While in terms of accuracy the xG model performs no differently to any other, its expected returns are significantly lower and more irregular.

In general results from this work agree with the conclusions of Egidi & Torelli (2020) who warn the reader in taking the results of any particular work in score modeling at face value especially concerning betting strategies. As we have shown the distinction and underlying assumptions made in many cases are not so straightforward when faced with the data, and some of the conclusions analyzed concern only one league and are seldom over 10 or 20 years old. On the other hand, in terms of future research there are some aspects that do need improvement. First, there was a technical issue that estimations from the *goalmodel* package could not predict some of the matches from the test sample, thus reducing it in size. This problem was further made worse by the xG data which was sparser compared to regular result data. Although this ultimately turned out not to be a major issue, certainly a better approach would improve estimations. Secondly, the approach of considering extra covariates and also the xG estimates can be improved by also including them into a time-dynamic perspective. Ultimately the single most significant unifier across leagues was the apparent negative distribution in the scores, thus a model addressing it would be a good starting point for future research.

References

Arastey, G. M. (2019). "What are expected goals (XG)? Sport Performance" Analysis. Retrieved May 25, 2022, from https://www.sportperformanceanalysis.com/article/what-are expected-goals-xg

Araújo, D., and Davids, K. (2016). "Team Synergies in Sport: Theory and Measures. Frontiers in Psychology", 7. doi:10.3389/fpsyg.2016.01449

Baio, G., & Blangiardo, M. (2010). "Bayesian hierarchical model for the prediction of football results." Journal of Applied Statistics, 37, 253 - 264.

Balaban, J. (2018). "A gentle introduction to maximum likelihood estimation. Medium." Retrieved May 8, 2022, from: https://towardsdatascience.com/a-gentle-introduction-to-maximum-likelihood-estimation-9fbff27ea12f

Barnett, M.; Kamke, C. (2014). "The introduction and application of sports analytics in professional sport organizations". Journal of Applied Sport Management, 6(2)

Benjamin, B., Pollard, R., Reep, C. (1971). "Skill and Chance in Ball Games". Journal of the Royal Statistical Society. Series A (General), 134(4), 623–629. doi:10.2307/2343657

Boshnakov, G., Kharrat, T., &Mchale, I. (2016.). "Are Goals Poisson Distributed?" Retrieved April 21, 2022, from https://www.sportstradingnetwork.com/wp-content/uploads/2016/06/GoalsDistribution-1.pdf

Boshnakov, G., Kharrat, T., McHale, I. G. (2017). "A bivariate Weibull count model for forecasting association football scores". International Journal of Forecasting, 33(2), 458–466.

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability". Monthey Weather Review, 78(1), 1–3.

Carpita, M., Ciavolino, E., Pasca, P., (2019). "Exploring and modelling team performances of the Kaggle European Soccer database. Statistical Modelling," doi:10.1177/1471082X18810971

Carpita, M; Sandri, M; Simonetto, A; Zuccolotto, P (2015). "Discovering the Drivers of Football Match Outcomes with Data Mining. Quality Technology & Quantitative Management", 12(4), 561–577. doi:10.1080/16843703.2015.11673436

Cullen AC, Frey HC (1999). "Probabilistic techniques in exposure assessment". Plenum Press, USA, pp. 81-155.

da Costa, I. B., Marinho, L. B., & Pires, C. E. (2021). Forecasting football results and exploiting betting markets: The case of "both teams to score." International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2021.06.008

Delignette-Muller, M. L., Dutang, C. (2015). "An R Package for Fitting Distributions. Journal of Statistical Software", 64(4), 1–34. https://doi.org/10.18637/jss.v064.i04

Dixon, M. J, Coles, S. G. (1997). "Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society. Series c (Applied Statistics)", 46(2), 265–280.

Dobson, S., Goddard, J. A., Dobson, S. (2001). "The economics of football". Cambridge: University Press Cambridge.

Egidi, L.; Pauli, F.; Torelli, N. (2018). "Combining historical data and bookmakers' odds in modeling football scores. Statistical Modelling", 18(5-6), 436–459. doi:10.1177/1471082X18798414

Egidi, L; Torelli, N. (2020). "Comparing Goal-Based and Result-Based Approaches in Modelling Football Outcomes. Social Indicators Research", (), –. doi:10.1007/s11205-020-02293-z

Galit. Sh., Thomas, P., Joseph, B., Sharad. B., Peter, B., (2005). "A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution". 54(1), 127–142. doi:10.1111/j.1467-9876.2005.00474.x

Havard Rue; Oyvind Salvesen (2000). "Prediction and Retrospective Analysis of Soccer Matches in a League"., 49(3), 399–418. doi:10.1111/1467-9884.00243

Hayes, A. (2022). "Poisson distribution. Investopedia". Retrieved April 24, 2022, from https://www.investopedia.com/terms/p/poisson-distribution.asp

Karlis, D., Ntzoufras, I. (2003). "Analysis of sports data by using bivariate Poisson models, 52(3), 381–393. doi:10.1111/1467-9884.00366

Karlis, D., Ntzoufras, I. (2006). "Bayesian analysis of the differences of count data" .25(11), 1885–1905. doi:10.1002/sim.2382

Karlis, D., Ntzoufras, I. (2008). "Bayesian modeling of football outcomes: using the Skellam's distribution for the goal difference." IMA Journal of Management Mathematics, 20(2), 133–145. doi:10.1093/imaman/dpn026

Kelly, J. L., Jr. (2011). "A new interpretation of information rate. In TheKelly capital growthinvestment criterion: theory and practice (pp.25–34). World Scientific.

Kocherlakota, S., Kocherlakota, K. (1992). "Bivariate Discrete Distributions (1st ed.)." CRC Press. https://doi.org/10.1201/9781315138480

Koning, R. H. (2000). Balance in Competition in Dutch Soccer. Journal of the Royal Statistical Society. Series D (The Statistician), 49(3), 419–431. http://www.jstor.org/stable/2681066

Koopman, S. J., & Lit, R. (2015). "A dynamic bivariate Poisson model for analyzing and forecasting match results in the English Premier League". Journal of the Royal Statistical Society: Series A (Statistics in Society), 178(1), 167–186.

Koopman, S. J., Lit, R. (2019). "Forecasting football match results in national league competitions using score-driven time series models". International Journal of Forecasting, 35(2), 797–809.

Lee, A. J. (1997). Modeling Scores in the Premier League: Is Manchester UnitedReallythe Best? CHANCE, 10(1), 15–19. https://doi.org/10.1080/09332480.1997.10554791

Lindstrøm, J. (2014, November 1). "The Dixon-Coles model for predicting football matches in R" (part 1) | opisthokonta.net. Opisthokonta .Net. Retrieved (2022, May 5) from https://opisthokonta.net/?p=890

Link, D. (2018). "Data analytics in professional soccer". Springer Vieweg, Wiesbaden
Maher, M. J. (1982). "Modeling association football scores." Retrieved (2022, March 13) from http://www.90minut.pl/misc/maher.pdf

Maimone, V., Yasseri, T. (2021). "Football is becoming more predictable; network analysis of 88 thousand matches in 11 major leagues". Royal Society Open Science. 8. 10.1098/rsos.210617.

Mark, Th. (1975). "On Any Given Sunday: Fair Competitor Orderings with Maximum Likelihood Methods". Journal of the American Statistical Association, 70(351), 536–541. doi:10.2307/2285929

McHale, I., Scarf, P. (2011). "Modelling the dependence of goals scored by opposing teams in international soccer matches. Statistical Modelling, 11(3), 219 236. doi:10.1177/1471082X1001100303

Moroney M. J., Facts from Figures, 3rd edition (Penguin, London) 1956.

Owen, A. (2011). "Dynamic Bayesian forecasting models of foot-ball match outcomes with estimation of the evolution variance parameter". IMA Journal of Management Mathematics, 22(2), 99–113.

Pearson, A., W. (2017). "How real-time analytics changes the face of the sports betting industry". In the Journal of Digital & Social Media Marketing, Volume 5, Issue 1

Pollard, R. (1985). "69.9 Goal-Scoring and the Negative Binomial Distribution. The Mathematical Gazette, 69(447), 45. https://doi.org/10.2307/3616453

Prasetio, D., Harlili, D., (2016). "IEEE 2016 International Conference On Advanced Informatics: Concepts, Theory And Application ICAICTA., Institute of Electrical and Electronics Engineers (IEEE) doi:10.1109/ICAICTA.2016.7803111

Roback, P., Legler, J. (2020). "Beyond Multiple Linear Regression: Applied Generalized Linear Models And Multilevel Models in R" (Chapman & Hall/CRC Texts in Statistical Science). Chapman and Hall/CRC.

Schmidt, I. G. (2020, April 21). "Una mirada al soccer analytics usando R-parte III. Medium." Retrieved May 6, 2022, from https://medium.com/datos-y-ciencia/una-mirada-al-soccer-analytics-usando-r-parte-iii-3bdff9cd3752

Shin, H. S. (1993). "Measuring the incidence of insider trading in a market for statecontingentclaims". The Economic Journal 103 (420), 1141–1153.

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.

Štrumbelj, E. (2014). "On determining probability forecasts from betting odds." International Journal of Forecasting, 30(4), 934–943. doi: 10.1016/j.ijforecast.2014.02.008

Suzuki. A, Salasar. L, Leite. J, Louzada-Neto, F. (2010). "A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup." The Journal of the Operational Research Society, 61(10), 1530–1539. doi:10.2307/40802329

Titman, A. C., Costain, D. A., Ridall, P. G., Gregory, K. (2015). "Joint modelling of goals and bookings in association football." Journal of the Royal Statistical Society. Series a (Statistics in Society), 178(3), 659–683. Retrieved: https://www.jstor.org/stable/43965756

Uehara, L; Button, Ch; Falcous, M; Davids, K(2014). "Contextualized skill acquisition research: a new framework to study the development of sport expertise. Physical Education and Sport Pedagogy." (), 1–16. doi:10.1080/17408989.2014.924495

Umami, I., Gautama, D., & Hatta, H. (2021). "Implementing the Expected Goal (xG) model to predict scores in soccer matches". International Journal of Informatics and Information Systems, 4(1), 38-54. doi:https://doi.org/10.47738/ijiis.v4i1.76

Appendix

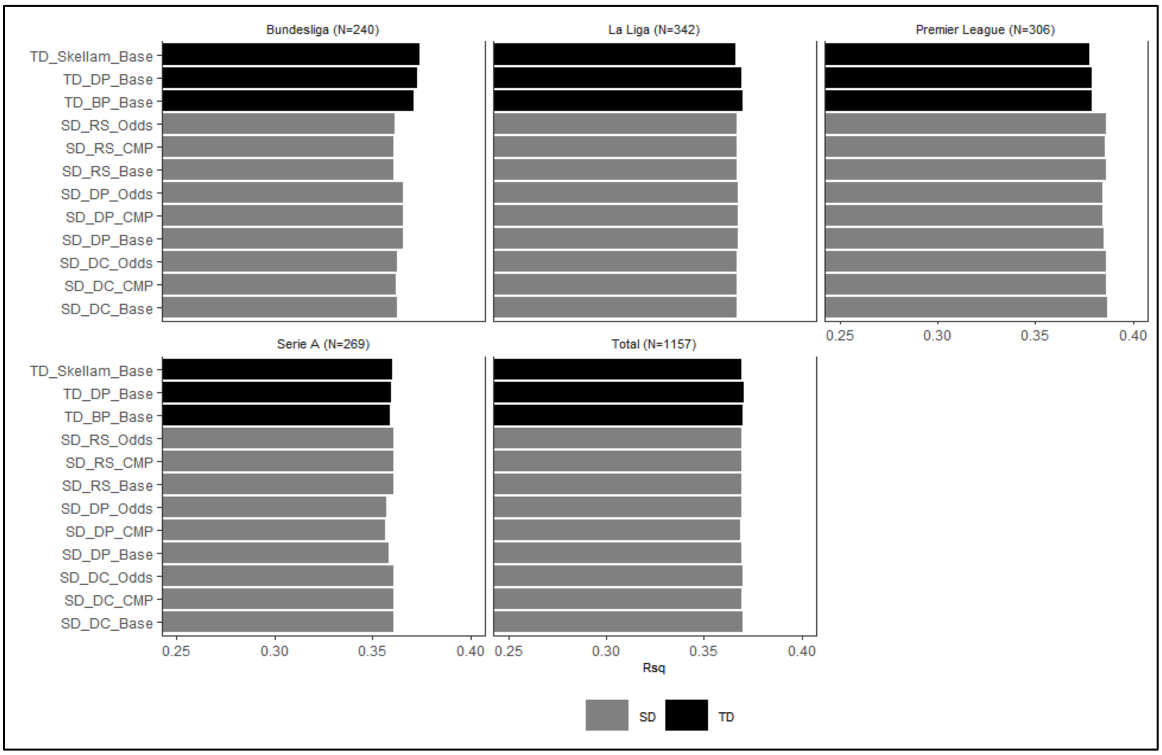

Figure 10. Pseudo – R^2 across leagues and models (xG dropped for larger N)

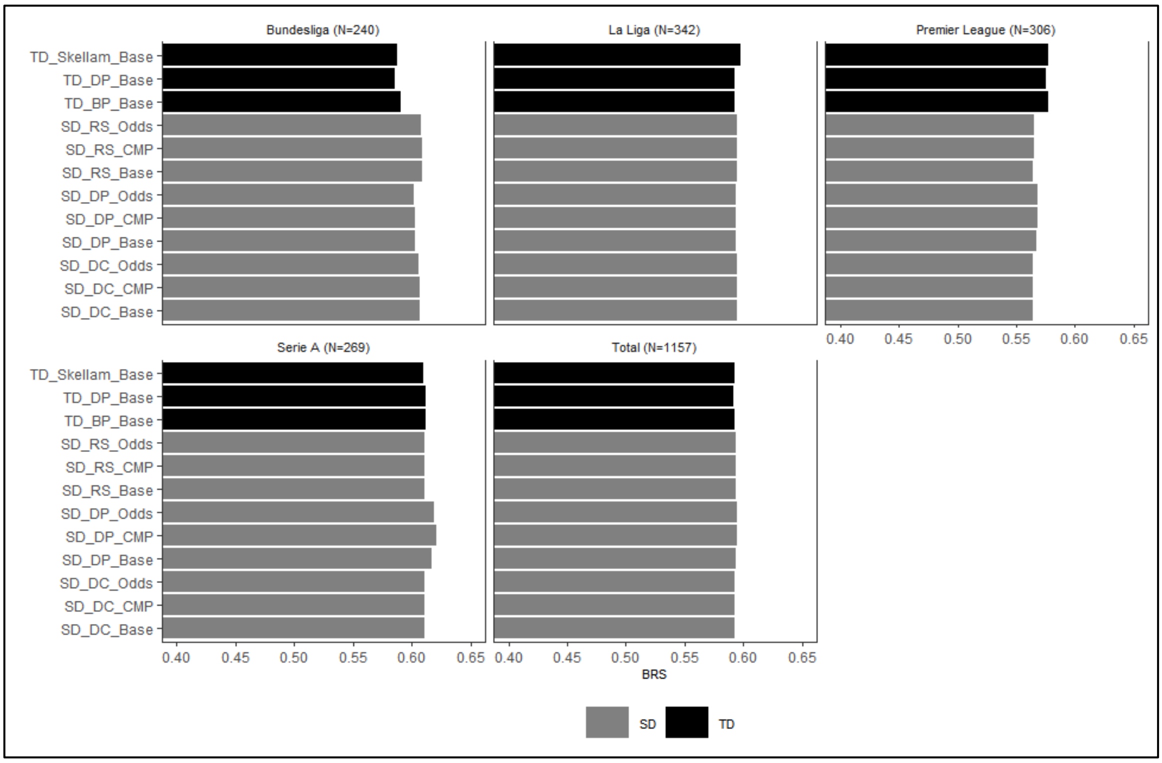

Figure 11. Brier score across leagues and models (xG dropped for larger N)

Abbreviation	*Long Name*
SD_DC_Base	Semi - Dynamic Dixon - Coles Model with no Covariates
SD_DC_CMP	Semi - Dynamic Dixon - Coles Model with CMP correction
SD_DC_Odds	Semi - Dynamic Dixon - Coles Model with Covariate Odds
SD_DP_Base	Semi - Dynamic Double Poisson Model with no Covariates
SD_DP_CMP	Semi - Dynamic Double Poisson Model with CMP correction
SD_DP_Odds	Semi - Dynamic Double Poisson Model with Covariate Odds
SD_DP_xG	Semi - Dynamic Double Poisson Model with xG
SD_RS_Base	Semi - Dynamic Rue - Salvesen Model with no Covariates
SD_RS_CMP	Semi - Dynamic Rue - Salvesen Model with CMP correction
SD_RS_Odds	Semi - Dynamic Rue - Salvesen Model with Covariate Odds
TD_BP_Base	Time - Dynamic Bivariate Poisson Model with no Covariates
TD_DP_Base	Time - Dynamic Double Poisson Model with no Covariates
TD_Skellam_Base	Time - Dynamic Skellam Model with no Covariates

Table 12. Model terminology