Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Predicting Football Matches with PARX-Copula Models

Leonardo Ribeiro Damiani Júnior

25 de fevereiro de 2024

Leonardo Ribeiro Damiani Júnior

Predicting Football Matches with PARX-Copula Models

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador(a): Prof. Dr. Flávio A. Ziegelmann

Porto Alegre
Fevereiro de 2024

Leonardo Ribeiro Damiani Júnior

Predicting Football Matches with PARX-Copula Models

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador(a) e pela Banca Examinadora.

Orientador(a):_____
Prof. Dr. Flávio A. Ziegelmann,
Doutor pela University of Kent at Canterbury, UKC, Grã-Bretanha.

Banca Examinadora:

Prof. Dr. Valdério Anselmo Reisen,
Doutor pela University of Manchester Institute of Science And Technology, UMIST, Inglaterra

Porto Alegre
Fevereiro de 2024

"The measure of greatness in a scientific idea,
is the extent to which it stimulates thought
and opens up new lines of research."
(Paul Dirac, 1968)

Agradecimentos

Primeiramente, agradeço a minha família por todo o apoio, amor e paciência.

Agradeço a minha mãe, Lisiane, por todo o amor dedicado a mim, por ser o meu exemplo na sala de aula e na vida, tanto com as pessoas ao meu redor como com a dedicação ao estudo.

Agradeço ao meu pai, Leonardo, pelo exemplo de profissionalismo, responsabilidade e companherismo, além de todos os momentos de amor e carinho que me distraíram ao longo deste tempo.

Agradeço a minha namorada, Maria Eduarda, por ser meu porto seguro e minha parceira ao longo destes anos de estudo, por ter paciência, atenção e por dividir cada vitória e cada derrota ao meu lado.

Agradeço ao meu irmão, João Paulo, por ser mais do que um irmão: um amigo. Sobrevivemos a uma pandemia no mesmo quarto!

Agradeço a minha sogra, Noeli, pelos momentos de paciência comigo e por cada vez que cedeu um quarto em sua casa para eu ficar.

Em especial, agradeço ao Flávio, meu orientador de iniciação científica e orientador deste trabalho, que me guiou durante todo o processo sempre com paciência e conselhos, onde pude aprender e me desenvolver como estudante.

Também agradeço ao Jean, professor e amigo da família, que me aconselhou nesses anos e foi importante pelas minhas escolhas tanto pela graduação de estatística quanto por outras já dentro do curso.

Além destes, agradeço a todos meus professores durante a graduação, cada um tem um pedaço de responsabilidade sobre este aluno que chegou ao final desta caminhada.

Agradeço também aos meus outros familiares: tios, primos, avós, dindos e amigos. Todos vocês foram importantes durante este processo de alguma forma. Seja nas confraternizações ou apenas nos momentos de conversa, obrigado por cada palavra, cada jogo e cada momento.

Novamente, agradeço a minha família: Lisiane, Leonardo, Maria e João.

Sem vocês isto não seria possível!

Resumo

O futebol é um esporte altamente lucrativo, que movimenta cada vez mais capital ao redor do mundo. Assim, nos tempos recentes, há um grande interesse em prever os resultados das partidas deste esporte. A adaptação de um Modelo de Poisson para os gols marcados por uma equipe em uma competição tornou-se uma ferramenta básica para essas análises. Nesse contexto, o Modelo de Poisson Autoregressivo com Covariáveis Exógenas é uma opção atraente, uma vez que tanto o número de gols quanto outras covariáveis relevantes podem ser incluídas no modelo a fim de fornecer informações preditivas adicionais. Além disso, por meio do uso de cópulas, uma possível dependência não linear entre ambos os gols marcados em uma partida podem aprimorar as previsões. Portanto, neste trabalho, integramos o processo Poisson Autoregressivo com Covariáveis Exógenas e Copula (PARX-Copula) para a previsão dos resultados das partidas de futebol da temporada 2022/23 da Premier League da Inglaterra. Avaliamos e comparamos as previsões obtidas com diferentes configurações de dependência para os gols marcados pelas equipes mandantes e visitantes, além de diferentes funções de ligação em nossos modelos. Ainda, testamos o uso de covariáveis para explicar as fraquezas dos oponentes para os nossos modelos. Finalmente, avaliamos como o uso de cópulas afeta as previsões dos resultados das partidas, uma vez que a suposição de independência entre o número de gols marcados em casa e fora é comum nesse contexto. Em nossos resultados, por meio do uso de métricas de performance, observamos o desempenho preditivo dos modelos no conjunto de testes. Em seguida, identificamos o melhor modelo preditivo da análise, $PARX_M Copula$, que considera o melhor ajuste para cada modelo marginal. No final, este modelo é aplicado a uma estratégia de apostas.

Palavras-Chave: Resultados de Futebol, Poisson, Copulas, PARX, Previsão, Premier League.

Abstract

Football, year after year, becomes a sport that increasingly moves billions of dollars around the globe. Thus, in recent times, there has been a great interest in predicting the matches results. Fitting a Poisson Model to the goals scored by a team in a competition has become a basic tool for these analyses and other approaches have also been used. In this context, Poisson Autoregressive with Exogenous Covariates is an attractive option, since both past number of goals and other relevant covariates can be included in the model to bring additional predictive information. Furthermore, taking into account, via copulas, a possible nonlinear dependence between the number of goals pro and against can improve the predictions. In this study we integrate Poisson Autoregressive with Exogenous Covariates and Copula (PARX-Copula) models for predicting the results of the 2022/23 football Premier League season matches in England. We evaluate and compare the forecasts obtained with distinct dependence settings for scored goals by home and away teams and different link function in our models. In addition, we test the use of covariates to explain the opponents weaknesses to our models. Finally, we assess how the use of copulas affects the predictions of the matches results, since the assumption of independence between home and away number of scored goals is common in this context. In our results, we see the predictive performance of all models in our analysis sample through performance metrics. Then the best predictive model $PARX_M Copula$ is presented, which considers the best fit for each marginal model. In the end, the best model is applied to a betting strategy.

Keywords: Football Results, Poisson, Copulas, PARX, Forecast, Premier League.

Sumário

1 Introdução

O futebol, ano após ano, se torna um esporte que cada vez mais movimenta milhões de pessoas e bilhões de dólares (Mathur, 2023) ao redor do globo. Assim, influenciado pelo advento das chamadas casas de apostas, tanto no Brasil (MKTEsportivo, 2023), quanto no mundo, o interesse pela previsão das partidas deste jogo se tornou não somente algo para lazer mas como o trabalho de muitos pesquisadores.

Os resultados de uma partida de futebol são o resumo dos processos que ocorrem durante os tempos regulamentares das partidas. Os times são definidos pelo que eles produzem em campo em um único número, por isso o gol é a principal finalidade. O futebol por essa razão se torna um desafio para a previsão, pois inúmeras variáveis influenciam no placar de uma única partida. No contexto de uma liga de futebol temos uma proposta mais atrativa para os pesquisadores, pois, geralmente, esta contempla um número limitado de partidas entre seus clubes por temporada anual. Por isso, diversos trabalhos tentam agregar informações teóricas a esse tipo de análise em busca de melhores previsões das próximas partidas.

Dessa forma, nos últimos tempos, tem havido um grande interesse em prever os resultados de partidas de futebol. Contribuindo com essa questão, o entendimento do comportamento das distribuições dos placares ao longo de um campeonato se tornou um atrativo para este fim. O ajuste de um Modelo de Poisson para os gols marcados por um time em uma competição se tornou uma ferramenta base dessas análises, onde Maher (1982) assume que o número de gols marcados por cada equipe em uma partida de futebol seguem processos de Poisson independentes. Dixon e Coles (1997) continuam os resultados de Maher e propõem um modelo bivariado capaz de perimitir depedência entre os gols marcados. Assim, outras abordagens também começaram a ser utilizadas, como métodos de machine learning (Santana et al., 2020), particularmente redes neurais (Bunker e Thabtah, 2019; Guan e Wang, 2022), métodos bayesianos (Baio e Blangiardo, 2010) e outros modelos multivariados (Koopman e Lit, 2015).

Na literatura também encontramos modelos construídos para variáveis de contagem que parecem ser interessantes no contexto futebolístico. Fokianos et al. (2009) apresenta modelos autorregressivos para séries temporais de contagem e discute a estimativa tanto para o modelo de Poisson quanto considerando a Binomial Negativa. Agosto et al. (2016) considera variáveis exógenas nesses modelos autorregressivos e Angelini e De Angelis (2017) o utiliza para a previsão de partidas de futebol considerando independência entre as séries de mandantes e visitantes. McShane et al. (2008) desenvolve um modelo de contagem baseado na distribuição Weibull que pode lidar com dados subdispersos e superdispersos. Com base nessa ideia, Kharrat et al.

(2019) estende essa abordagem para criar uma família diversificada e flexível de distribuições de contagem de renovação, que amplia muito a caixa de ferramentas de distribuições disponíveis para a modelagem de dados de contagem.

Estendendo esses modelos, a utilização de cópulas se torna um artífício recente para capturar a dependência entre os gols marcados em uma partida e permitir a modelagem bivariada não independente. Ou seja, Mchale e Scarf (2011), optam por permitir qualquer dependência potencial entre os gols marcados pelas duas equipes através da utilização de uma cópula para combinar as duas distribuições marginais de gols marcados. Nessa lógica, Halliday e Boshnakov (2018) apresentam uma abordagem da utilização dos modelos autorregressivos de Poisson integregados pela cópula de Frank, enquanto Boshnakov et al. (2017) utiliza os processos de Weibull renováveis com a mesma cópula. Por fim, vemos que copulas são utilizadas em diversas áreas Tootoonchi et al. (2022), Sabino da Silva et al. (2023) e Silva Filho et al. (2012).

Além disso, percebemos que estas modelagens e diferentes métodos de predição dos resultados, estão sendo aplicados contra as casas de apostas no intuito de um ganho financeiro em cima das "odds" (razão de chances) estimadas por essas plataformas para cada desfecho. Angelini e De Angelis (2017) utilizam estratégias para apostas de vitória do mandante, empate e vitória do visitante. Da Costa et al. (2022) focam no caso "ambos os times marcam"(BTTS). Shah et al. (2021) fazem uso do critério de Kelly para maximizar os ganhos e, utilizando o mesmo critério Boshnakov et al. (2017), considera ainda os casos de mais ou menos 2.5 gols por partida.

Sendo assim, os modelos discutidos e apresentados neste estudo concentram-se na dependência autorregressiva e na dependência cruzada entre os gols marcados por equipes de futebol em um campeonato. Consideraremos a natureza das equipes como mandantes e visitantes dentro de um campeonato nacional por meio de modelos bivariados utilizando cópulas com a possibilidade de inclusão de variáveis exógenas. Também serão utilizados modelos mais simples que consideram independência entre os gols marcados pelos adversários.

Portanto, este trabalho tem como objetivo avaliar e comparar as previsões obtidas para o número de gols de ambos os times em cada partida analisada. O foco da previsão será nos jogos da temporada 2022/23 da Premier League, a liga profissional de futebol da Inglaterra. A expectativa é que o uso de cópulas apresente um impacto positivo nos resultados devido a dependência entre os gols dos oponentes. Por fim, o segundo objetivo é obter retornos financeiros por meio de estratégias elaboradas contra uma casa de apostas a partir de um modelo de melhor perfomance.

Este trabalho será apresentado no formato de artigo que se encontra estruturado da seguinte forma: **Seção 2** apresentará os métodos usados nas análises; **Seção 3** demonstrará a fonte de dados e sua exploração; **Seção 4** conterá os resultados das estimativas realizadas juntamente com os desempenhos dos modelos e a estratégia de apostas adotada; **Seção 5** concluirá o artigo, apresentando as principais descobertas.

2 Artigo

Autores:

Leonardo Ribeiro Damiani Júnior, Flávio A. Ziegelmann

Título:

Predicting Football Matches with
PARX-Copula Models

Ano: 2024

Predicting Football Matches with PARX-Copula Models

Damiani, Leonardo[a,1], Ziegelmann, Flávio A.[a,2]

[a]*Department of Statistics, Federal University of Rio Grande do Sul (UFRGS, Brazil), Agronomia - 91509-900 Porto Alegre - RS, Av. Bento Gonçalves, 9500 - Prédio 43-111, Brazil*

Abstract

Football, year after year, becomes a sport that increasingly moves billions of dollars around the globe. Thus, in recent times, there has been a great interest in predicting the matches results. Fitting a Poisson Model to the goals scored by a team in a competition has become a basic tool for these analyses and other approaches have also been used. In this context, Poisson Autoregressive with Exogenous Covariates is an attractive option, since both past number of goals and other relevant covariates can be included in the model to bring additional predictive information. Furthermore, taking into account, via copulas, a possible nonlinear dependence between the number of goals pro and against can improve the predictions. In this study we integrate Poisson Autoregressive with Exogenous Covariates and Copula (PARX-Copula) models for predicting the results of the 2022/23 football Premier League season matches in England. We evaluate and compare the forecasts obtained with distinct dependence settings for scored goals by home and away teams and different link function in our models. In addition, we test the use of covariates to explain the opponents weaknesses to our models. Finally, we assess how the use of copulas affects the predictions of the matches results, since the assumption of independence between home and away number of scored goals is common in this context. In our results, we see the predictive performance of all models in our analysis sample through performance metrics. Then the best predictive model $PARX_M Copula$ is presented, which considers the best fit for each marginal model. In the end, the best model is applied to a betting strategy.

Keywords: Football Results, Poisson, Copulas, PARX, Forecast, Premier League

1. Introduction

Football continues to evolve into a sport that mobilizes billions of dollars (Mathur, 2023) worldwide each year. Thus, influenced by the advent of so-called betting houses, both in Brazil and globally (MKTEsportivo, 2023), the interest in predicting the outcomes of these matches has evolved from mere leisure to the occupation of numerous researchers.

The results of a football match summarize in a specific and very relevant way the processes occurring during the standard playing time. Teams are defined by what they produce on the field, represented by a single number, hence making scoring goals the primary objective. This complexity

Email addresses: leojunior.damiani@gmail.com (Damiani, Leonardo), flavioz@ufrgs.br (Ziegelmann, Flávio A.)

[1]Student at UFRGS.

[2]Professor at UFRGS.

renders football a challenge for prediction, given the countless variables influencing the score in a single match. In the context of a football league, typically encompassing a limited number of matches among clubs per season, annually, it poses an enticing challenge for researchers within this field, due to the limited sample size. Consequently, various studies attempt to integrate theoretical insights into this analysis for improved predictions of upcoming matches.

Hence, there has been much interest in predicting football match outcomes lately. Contributing to this quest, understanding the distribution patterns of scores throughout a championship has become a focal point. Employing a Poisson Model to analyze goals scored by a team in a competition has become a foundational tool in these analyses. For instance, Maher (1982) assumes that the number of goals scored by each team in a football match follows independent Poisson processes. Dixon and Coles (1997) extend Maher's findings and propose a bivariate model capable of allowing dependence between scored goals. Other approaches have emerged, such as machine learning methods (Santana et al., 2020), particularly neural networks (Bunker and Thabtah, 2019; Guan and Wang, 2022), Bayesian methods (Baio and Blangiardo, 2010) and other multivariate models (Koopman and Lit, 2015).

Literature also showcases models developed for count variables that seem useful in the football context. Fokianos et al. (2009) presents autoregressive models for count time series and discusses estimation for both Poisson and Negative Binomial models. Fokianos and Tjøstheim (2011) increases the process with log as a link function and Agosto et al. (2016) incorporates exogenous variables into these autoregressive models, while Angelini and De Angelis (2017) employ them for predicting football matches, considering independence between home and away series. McShane et al. (2008) develops a count model based on the Weibull distribution capable of handling underdispersed and overdispersed data. Building upon this idea, Kharrat et al. (2019) extends this approach to create a diverse and flexible family of renewal count distributions, significantly broadening the toolbox available for count data modeling.

Expanding on these models, copula modeling has emerged as a recent technique to capture dependence between the number of goals of each team scored in a match and enable non-independent bivariate modeling. For instance, Mchale and Scarf (2011) opt to allow potential dependence between goals scored by both teams by employing a copula to combine the two marginal distributions of scored goals. In line with this logic, Halliday and Boshnakov (2018) introduce an approach using autoregressive Poisson models integrated by the Frank copula, while Boshnakov et al. (2017) uses renewable Weibull processes with the same copula. Ultimately, copulas are applied across multiple domains, as seen in Tootoonchi et al. (2022), Sabino da Silva et al. (2023) and Silva Filho et al. (2012).

Furthermore, these models and various prediction methods are being applied against betting houses with the aim of financial gain based on the odds estimated by these platforms for each outcome. Angelini and De Angelis (2017) employ strategies for betting on home win, draw, and away win. Da Costa et al. (2022) focus on betting cases where both teams score goals, the famous "both teams to score" (BTTS). Shah et al. (2021) utilizes the Kelly criterion to maximize gains and Boshnakov et al. (2017), with the same criterion, considers another famous bet: over or under 2.5 goals per match.

Thus, the models discussed and presented in this study focus on autoregressive dependence and cross-dependence between goals scored by football teams in a championship. We will consider the nature of teams as home and away within a national championship through bivariate models using copulas, along with the possibility of adding exogenous variables. Simpler models that consider

independence between goals scored by opponents will also be used.

Therefore, the objective of this work is to evaluate and compare predictions obtained for the number of goals scored by both teams in each analyzed match. The forecast will focus on the matches of the 2022/23 season of the Premier League, the top professional football league in England. The expectation is that the use of copulas will have a positive impact on the results due to the dependence between the goals of the opponents. Finally, the second aim is to achieve financial returns through devised strategies against a betting house based on a model of better performance.

This article is structured as follows: **Section 2** will present the methods used in the analyses; **Section 3** will demonstrate the data source and its exploration; **Section 4** will contain the results of the estimations performed along with model performances and the betting strategy; **Section 5** will conclude the article, presenting the main findings.

2. Methodology

2.1. Poisson Autoregressive with Exogenus Covariates (PARX)

Following Agosto et al. (2016); Angelini and De Angelis (2017), the PARX model is now described.

Let $Y_t \in \{0, 1, 2, \dots\}$, $t = 1, \dots, T$, Y_t be an observed count time series. Let F_t be the information set available at time t, i.e., $F_t = \{y_{t-m}, x_{t-m} : m \geq 0\}$. It is stated that Y_t is a PARX process, with intensity parameter λ_t, denoted by $Y_t \sim \text{PARX}(p, q)$, if it can be written as follows:

$$Y_t | F_{t-1} \sim \text{Poisson}(\lambda_t), \tag{1}$$

$$\lambda_t = \omega + \sum_{l=1}^{p} \beta_l Y_{t-l} + \sum_{l=1}^{q} \alpha_l \lambda_{t-l} + \eta \mathbf{x_{t-1}}, \tag{2}$$

where $\omega > 0$, $\beta_1, \dots, \beta_q$ and $\alpha_1, \dots, \alpha_p$ are non-negative coefficients, $\eta > 0$ is a vector of coefficients for exogenous covariates and $\mathbf{x_{t-1}} \in \mathbb{R}^r$ is a vector of covariates considered in the model. Therefore, the conditional intensity, λ_t, depends on p past values of Y_t, q of its past values and covariates given by the vector $\mathbf{x_t}$.

The parameters ω, $\beta_1, \dots, \beta_q$, $\alpha_1, \dots, \alpha_p$, and η are time-invariant and ensure that the distribution of $Y_t | F_{t-1}$ is a non-degenerate Poisson ($\lambda_t \neq 0$) with a positive intensity ($\lambda_t > 0$). Specifically, when $\eta = 0$, the PARX model reduces to the one in Fokianos et al. (2009) which proposed an autoregressive model similar to GARCH models (Engle, 1982; Bollerslev, 1986).

Futhermore, the necessary e sufficient condition for (1) and (2) defining a unique strictly statiotionary process Y_t is $\sum_{j=1}^{\max(p,q)} \beta_j + \alpha_j < 1$ with x_t also strictly stationary.

A distinctive feature of the model is that, for a single covariate x_{t-1}, the expected value of the series Y_t is given by

$$E(Y_t) = E(\lambda_t) = \frac{\omega + E(x_{t-1})}{1 - \sum_{j=1}^{\max(p,q)} (\beta_j + \alpha_j)}. \tag{3}$$

Moreover, by incorporating past values of the response and covariates in the conditional intensity evolution, PARX models can capture overdispersion in the marginal distribution, i.e., $Var(Y_t|F_{t-1}) = E(Y_t|F_{t-1})$.

Further details on these and other properties can be found in Section 3 of Agosto et al. (2016).

2.2. Log Linear Model

Considering as shown in Fokianos and Tjøstheim (2011) and Liboschik et al. (2017), we can obtain the log-linear model with $\nu_t = \log(\lambda_t)$ and $t = 1, \ldots, T$, as follows:

$$Y_t|F_{t-1} \sim \text{Poisson}(\lambda_t), \tag{4}$$

$$\nu_t = \omega + \sum_{l=1}^{p} \beta_l \log(Y_{t-l} + 1) + \sum_{l=1}^{q} \alpha_l \nu_{t-l} + \eta \mathbf{x_{t-1}}, \tag{5}$$

where F_{t-1} is the information set available, ω is the intercept, $\beta_1, \ldots, \beta_q$ and $\alpha_1, \ldots, \alpha_p$ are the autoregressive coefficients, η is a vector of coefficients for exogenous covariates and $\mathbf{x_{t-1}} \in \mathbb{R}^r$ as a vector of additional covariates considered in the model.

The log-linear model allows the estimated parameters to be negative. However, note that the effect of the summations in the linear predictor on the conditional mean is multiplicative, and therefore, the parameters play a different role compared to the previous model.

The R package called **tscount**, proposed in Liboschik et al. (2017), is used for modeling these models. The package proposes likelihood-based estimation methods for analysis and modeling of count time series based on generalized linear models.

2.3. Copulas

The theory of Copulas was developed based on Sklar's theorem (Sklar, 1959), which states that any multivariate distribution can be represented as a function of its marginals.

Theorem. ***(Sklar, 1959)*** *Let F be a joint distribution with marginals $F_1, \ldots, F_d$. Then there exists a copula $C : [0,1]^d \longrightarrow [0,1]$ such that, for all $(y_1, \ldots, y_d)' \in \mathbb{R}^d$,*

$$F(y_1, \ldots, y_d) = C\Big(F_1(y_1), \ldots, F_d(y_d)\Big). \tag{6}$$

Moreover, the procedures for estimating the marginals and the dependence structure can be performed separately, as discussed in Joe and Xu (1996). The suggested approach is a two-stage procedure where the marginals are first estimated independently before the copula is adjusted. This is known as the "inference from the margins" (IFM) method. The resulting IFM estimator is asymptotically normal and consistent, supporting our choice of this method.

In this work we will use IFM method with empirical distributions of the residuals from the marginals, i.e., a more direct approach. Firstly, the empirical distributions are computed. The

residuals represent the differences between observed values and the values predicted by the estimated marginal distributions. Next, the residuals are utilized to construct a joint dependence structure. The rationale is that if there are correlations or patterns in the residuals of the marginal distributions, it could indicate some form of dependence between the variables.

Hence, a well-implemented open-source package for copulas is proposed in the **copula** package in Yan (2007). This package simplifies usage, enabling more individuals to benefit from copula properties, becoming an essential tool in recent years.

2.4. PARX-Copula Model

Let $D(\lambda_1, \lambda_2; \rho)$ be a bivariate distribution based on a copula with dependency parameter ρ, and marginals Poisson(λ_1) and Poisson(λ_2), based on Halliday and Boshnakov (2018). Also, let $\mathbf{Y_t} = (Y_{1,t}, Y_{2,t})'$ be a bivariate time series of counts where $Y_{1,t}$ and $Y_{2,t}$ are univariate PARX processes with intensities $\lambda_{j,t}$ and associated exogenous covariates $\boldsymbol{x_{j,t-1}}$ for $j = 1, 2$. Denoting $\boldsymbol{\lambda_t} = (\lambda_{1,t}, \lambda_{2,t})'$ where $t = 1, \ldots, T$, we denote by F_{t-1}, formed by past observations and exogenous covariates, the information set at time $t - 1$:

$$\boldsymbol{F_{t-1}} = \left\{ \boldsymbol{Y_{1-p}}, \ldots, \boldsymbol{Y_{t-1}}, \boldsymbol{\lambda_{1-q}}, \ldots, \boldsymbol{\lambda_{t-1}}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_{t-1}} \right\} \tag{7}$$

The process $\boldsymbol{Y_t}$ is a PARX-Copula($\rho, \lambda_1, \lambda_2$) if the bivariate distribution is

$$\boldsymbol{Y_t} | \boldsymbol{F_{t-1}} \sim D(\lambda_{1,t}, \lambda_{2,t}; \rho) = C_\rho \Big(F_1(y_1; \lambda_1), F_2(y_2; \lambda_2) \Big) \tag{8}$$

is such that the conditionals of $Y_{1,t}, Y_{2,t}$

$$Y_{j,t} | \boldsymbol{F_{t-1}} \sim \text{Poisson}(\lambda_{j,t}), \qquad j = 1, 2; \tag{9}$$

$$\lambda_{j,t} = \psi_j + \sum_{l=1}^{p} \alpha_{j,l} Y_{j,t-l} + \sum_{l=1}^{q} \beta_{j,l} \lambda_{j,t-l} + \eta_j \boldsymbol{x_{j,t-1}}, \qquad j = 1, 2; \tag{10}$$

where C_p is the chosen copula function, F_1, F_2 are the poisson distribution functions, $\alpha_{j,l}, \beta_{j,l}$, denote coefficients for past observations and intensities, ψ_j as an intercept term and η_j represents the non-negative vector of coefficients for exogenous covariates.

For the previously defined log-linear model, the development is analogous.

2.4.1. Akaike Information Criterion (AIC)

For the analysis of the marginal distribution models, we will consider the well-recognized Akaike criterion (Akaike, 1974) as the metric for selecting the best fit. The AIC is described as

$$AIC = 2K - 2\ln(\mathcal{L}) \tag{11}$$

AIC determines the relative information value of the model using the $\mathcal{L}$ log-likelihood estimate and the number K of parameters (independent variables) in the model.

2.5. Performance

To assess the performance of our models, we will employ the following methods.

2.5.1. Log Loss

Logarithmic Loss, commonly known as log loss or cross-entropy loss, is a performance metric used in the evaluation of classification models. It measures the performance of a classification model by quantifying the difference between the predicted probabilities and the actual class labels. The Log Loss for predictions of multiple classes (Bishop, 2006) is defined as:

$$\text{Log Loss} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} o_{ij}\log(p_{ij}) \tag{12}$$

where n is the number of instances, k is number of class, $o_{ij} = 1$ if the current class $y_i = j$, or $o_{ij} = 0$ if the class $y_i \neq j$, log is the natural logarithm and p_{ij} is the predicted probability that belongs to class.

2.5.2. Brier Score

Introduced by Brier (1950), the Brier Score is used to measure the accuracy of probabilistic forecasts. When applied to predictions of multiple classes, the Brier Score is defined as:

$$BS = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}(p_{ij} - o_{ij})^2 \tag{13}$$

where n is the number of instances, k is the number of classes, p_{ij} is the predicted probability, and $o_{ij} = 1$ if the current class $y_i = j$, or $o_{ij} = 0$ if the class $y_i \neq j$. This measure ranges from 0 to 2.

In the case with $k = 3$, we can consider a baseline value for the score. If each class receives an equal probability of 1/3 for an event to occur or not, the calculation will always result in

$$BS = \left(\frac{1}{3} - 1\right)^2 + \left(\frac{1}{3} - 0\right)^2 + \left(\frac{1}{3} - 0\right)^2 = \frac{2}{3}.$$

Therefore, the model is expected to have a Brier Score lower than the threshold = 2/3 to be considered better than the simple assignment of equiprobable probabilities.

2.5.3. Match Results Score

Finally, we will use a metric to assess the accuracy of predicting the actual match outcomes. We will consider the outcome with the highest estimated probability as the predicted outcome. The score will be based on the number of correct predictions, as outlined below.

$$MRS = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{14}$$

where n is the number of instances and $d = 1$ if the highest probability class for the instance matches the actual class; otherwise, $d = 0$.

3. Data Set and Analysis

The analyses were conducted using the R Statistical Software (R Core Team, 2023, v4.2.3).

3.1. Data Source

The data set utilized in this study are sourced from the Football Reference website (FBref). The focus lies on matches from the last thirteen seasons of the Premier League from 2010 to 2023, where the relevant matches are extracted via the worldfootballR R package (Zivkovic, 2022).

3.2. Data Set

Our dataset consists of 4940 match results from the Premier League, England's top professional football division, containing information about the round and date of these matches spanning the seasons from 2010/11 to 2022/23. Figure 1 display the distribution of number of matches at home for the 39 teams that participated in the league during this period.

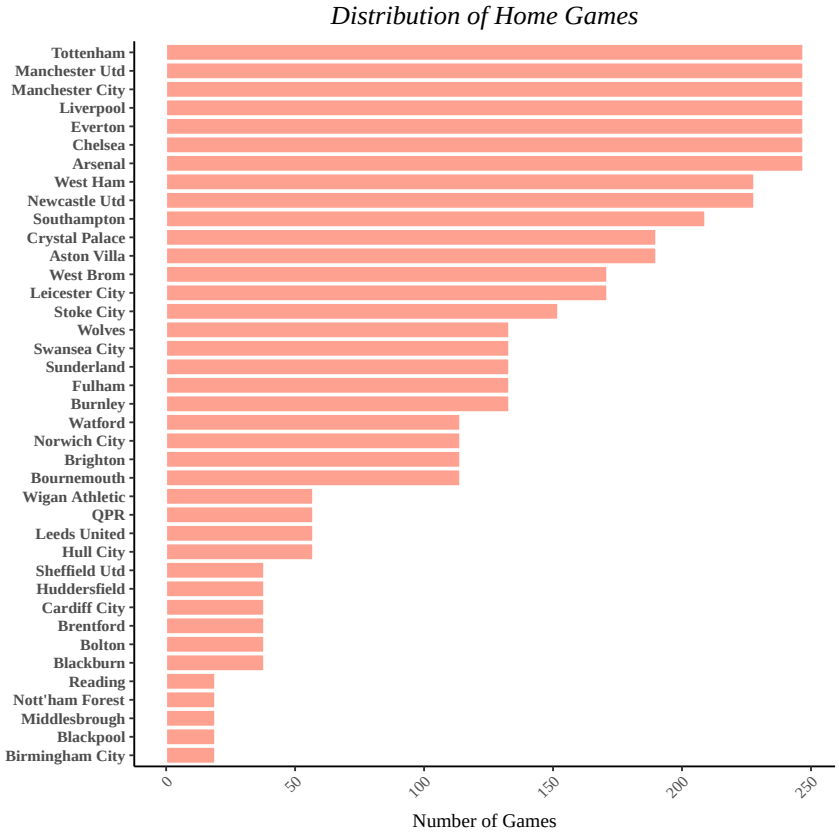

Figure 1: Matches from the seasons 2010/11 to 2022/23.

Note that the distribution will be the same for the number of away matches and the teams with the highest number of matches are those that played every league match throughout these years. On the other hand, this means that the other teams played in the lower divisions of the country at some point. This becomes a limitation of our study, because these teams that do not play in the league "escape" from our modeling sample and have fewer matches to be modeled.

Additionally, in our dataset, we will use one variable for both home and away models: weighted average of goals conceded by the opponent as a covariate. We are basing this metric on the measure used in Angelini and De Angelis (2017) and Santana et al. (2020). The metric serves as exogenous information to the model regarding the opponents' defensive capability. It is calculated from 2009/10 season matches, which are not included in our dataset, and aims to assign a linear weight to prioritize the most recently conceded goals.

3.3. Exploratory Analysis

Our emphasis will be on predicting matches from the last weeks of the 2022/23 season. The models rely on the assumption that the goals scored by the teams follows Poisson distributions. Therefore, Figure 2 illustrate the goals scored distributions that will be modeled for the 34th round of the top English football clubs according to UEFA. The distributions are similar to Poisson distributions and the others teams in the league have distributions with the same format.

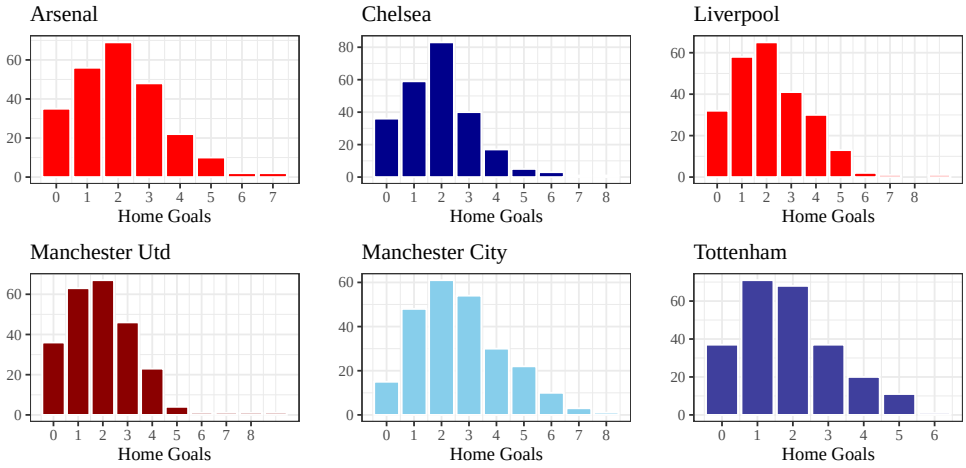

Figure 2: Goals as the home team during the period: 2010/11 to the 34th round of the 2022/23 season.

Thus, goals scored can be viewed as time series for modeling these processes, i.e., times series to model the rate parameter λ from the Poisson distribution. Consequently, each team will have its time series of goals scored modeled, wherein we differentiate between the team playing at home or away. In other words, each team will have one series for home matches and another one for away matches, as showed in Figure 3.

Then, in a match, we have the Poisson parameters λ_1 and λ_2 for home and away which leads us to an interpretation of the average rate of goals expected by the clubs. The distributions concentrate the observations in the lower values and then the averages are expected to be around this range. In Table 1 we present the goals scored average from the distrbutions in Figure 2.

Table 1: Average Goals Scored

	Arsenal	Chelsea	Liverpool	Man Utd	Man City	Tottenham
Average	2.06	1.92	2.16	1.96	2.67	1.87

The analysis for the away goals is analagous because the distributions are similar.

Therefore in Figure 3 we have the ilustration for the times series that will be modeled for the 34th round. Note that we present the times series only for two teams: Liverpool's series and Brighton's series. In this case, Liverpool played all the league matches during the period, but Brighton only started playing in the league in 2017 (due to the "escape" of teams from the sample).

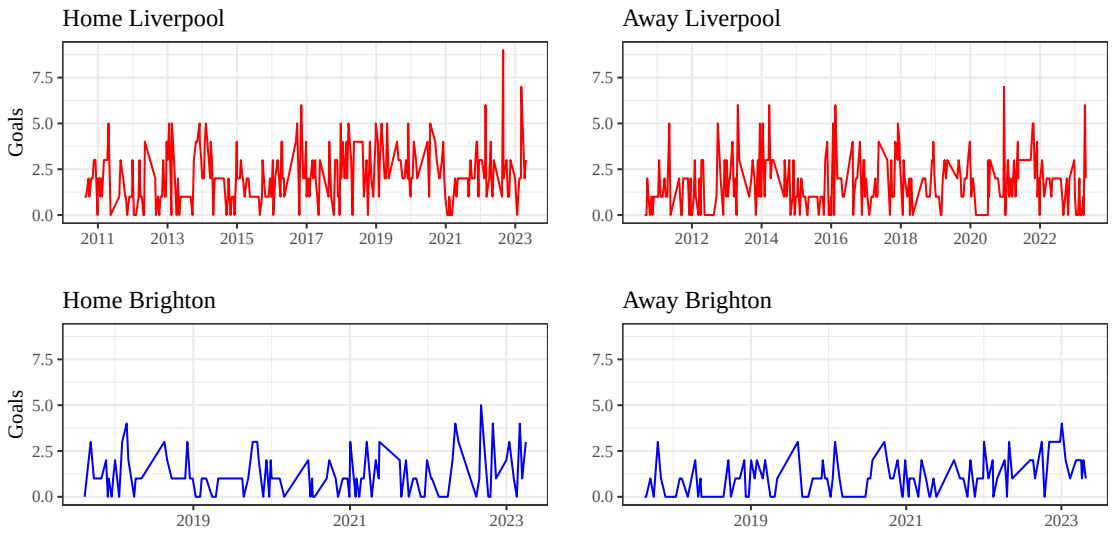

Figure 3: Time series of goals during the period: 2010/11 to the 34th round of the 2022/23 season.

Using the time series approach for modeling the rates, we will be interested in observing how past values and past rates influence these parameters. This influence will be obtained through the PARX models. Figure 4 illustrate the autocorrelation of the series presented in Figure 3.

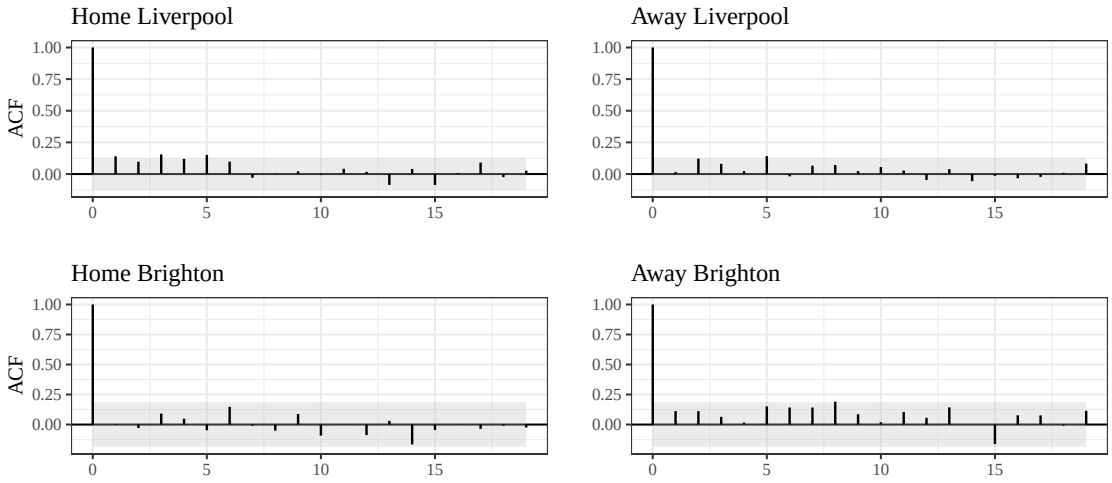

Figure 4: Autocorrelations of goals from the last 19 matches until the 34th round of the 2022/23 season.

The Figure 4 shows us a small autocorrelation in the series, but with some correlations overtanking the confidence limit in the Liverpool's series. This behavior is repeated in the other teams.

3.4. Models

The analysis will consider four different approaches to modeling the goals series. Hence, each team will have eight models, with four for the home series and four for the away series.

The models are detailed in Table 2.

Table 2: Marginals Models

	$\text{PARX}^*_I(p,q)$	$\text{PARX}_I(p,q)$	$\text{PARX}^*_L(p,q)$	$\text{PARX}_L(p,q)$
Covariate	×	✓	×	✓
Link Function	Identity	Identity	Log	Log

Each of these models will be selected from a range of models with varying number of parameters p and q using the Akaike criterion. For the construction of the bivariate distribution, we will consider two scenarios: assuming independence between the marginal distributions and employing the copula framework.

First, we are only considering bivariate models composed of the same marginal models. This approach aims to identify the impact of marginals and copulas on predicting the bivariate distribution. Thus, the possibilities explored are outlined in Figure 5 for a single match.

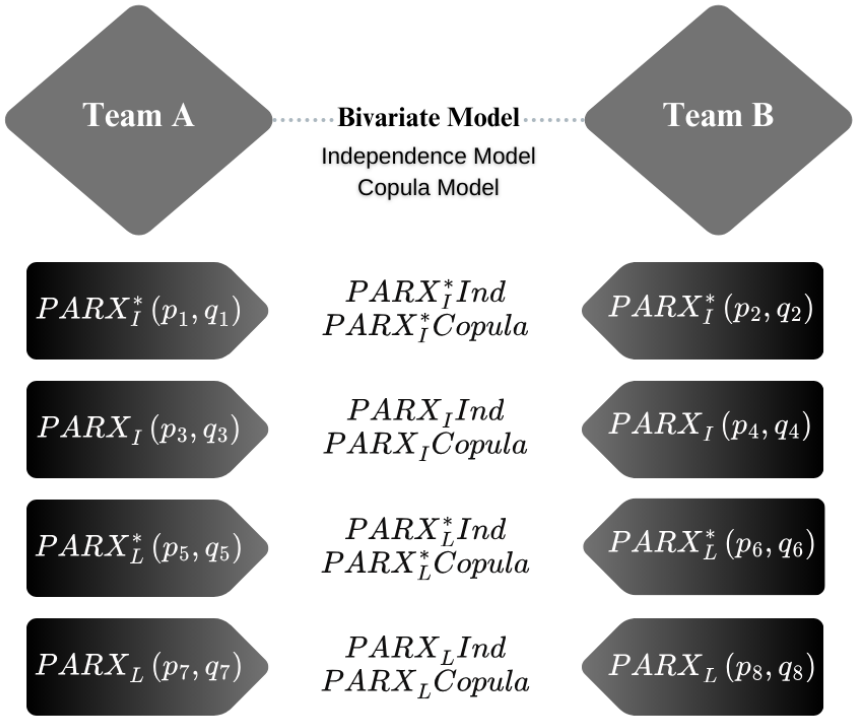

Figure 5: Modeling diagram for two teams

Therefore, a strategy that considers for each team only the best model among the four possible options is implemented. This strategy is denoted as $PARX_M$, with the bivariate models $PARX_M Ind$ and $PARX_M Copula$.

4. Results

For our forecasting window, we are considering 145 matches. We have divided these matches based on the rounds they occurred, from round 24 to round 38. Each round has the estimated models and the one-step-ahead forecast, where, in the end, we compute the metrics of interest to evaluate our strategies.

Hence, in our study, as 39 teams participated in the league during the study period, we estimated around 78 models for each marginal model. In other words, since we used four different model strategies for the marginals, we have a total of 312 models per round. Note that, clearly, in a single round, we will have only 20 teams competing, but for copula estimation, we are considering the entire history of Premier League matchups.

In Appendix C we acess the evaluation from our models with the Probability Integral Transform (PIT) histogram and a marginal calibration plot for assessing the fit.

4.1. Parameters

We considered the following number of parameter variations: $0 \leq p \leq 3$ and $0 \leq q \leq 3$. Through the Akaike criterion, the best models were selected for each of the individual fitting of home and away series. Figure 6 shows the identified models orders.

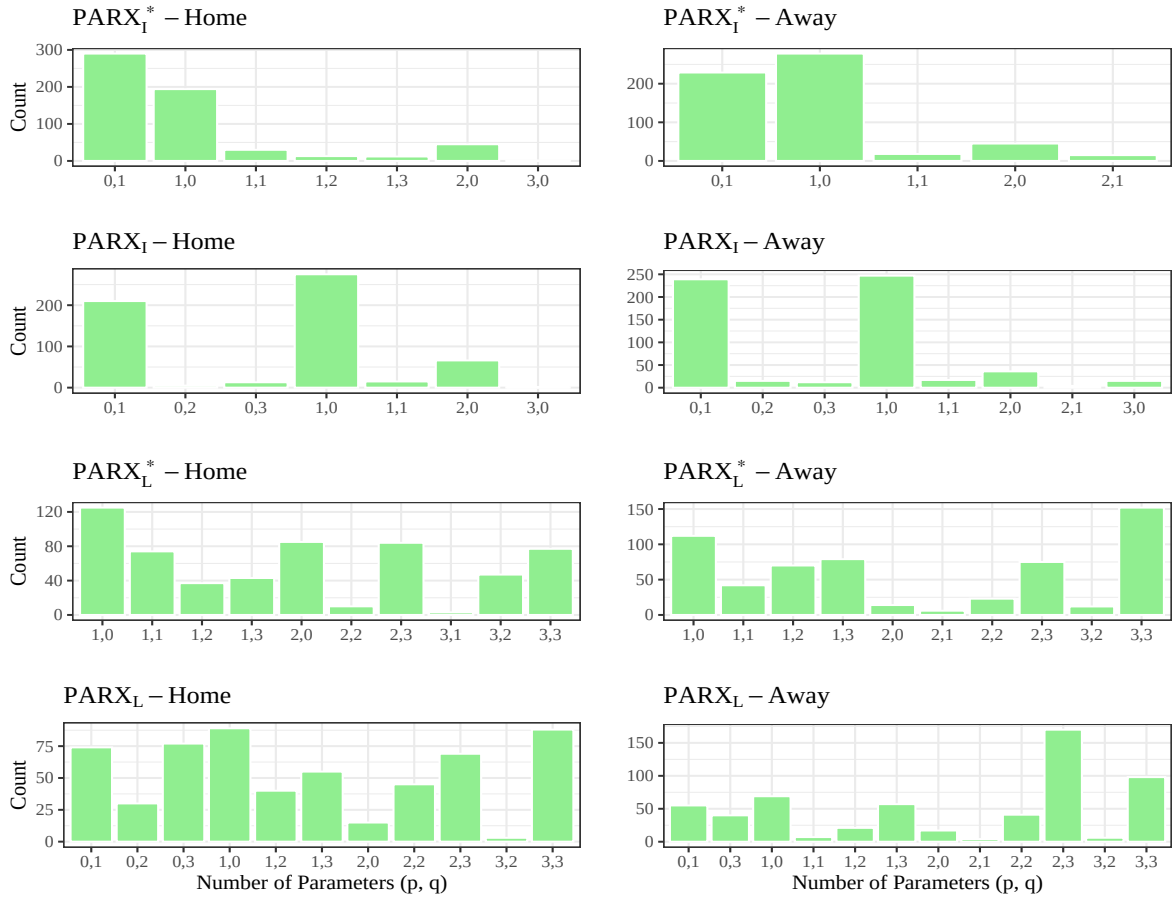

Figure 6: The distributions of the number of parameters for each marginal model.

We note that models estimated with the identity link function presented a small number of lags, with the main concentration on a single autoregressive parameter, either in past observations or past rates. This corroborates the result in Figure 3, i.e., we have a large number of series that show a small dependence on past observations and others that show a greater dependence.

On the other hand, the log-linear models presented a more distributed number of parameters, with the appearance of models with up to 2 or 3 lags for both past terms.

Furthermore, all models considered the intercept effect, and as described earlier, some models consider the effect of the variable used to explain the opponent's defensive strength.

4.2. Joint Distribution

These estimated models form the pair for multivariate modeling. Considering the assumption of independence, our estimated probability is constructed by taking into account the adjusted models with the estimated parameters $\hat{\lambda} = (\hat{\lambda_1}, \hat{\lambda_2})'$ and F_t the information set at time t.

$$F(y; \hat{\lambda}) = \hat{P}\Big(Y_{1,t+1} = y_1, Y_{2,t+1} = y_2 | F_t\Big) = \hat{P}\Big(Y_{1,t+1} = y_1 | F_t\Big) \times \hat{P}\Big(Y_{2,t+1} = y_2 | F_t\Big) \tag{15}$$

In the copula approach, Gaussian copula, Frank's copula and Clayton's copula (Appendix B) were tested, with the latter showing the best fits for each round, according to Akaike.

Thus, considering $\hat{\lambda} = (\hat{\lambda_1}, \hat{\lambda_2})'$, C the Clayton copula and ρ the copula parameter,

$$F(y; \hat{\lambda}, \rho) = C\Big(F_1(y_1; \hat{\lambda_1}), F_2(y_2; \hat{\lambda_2})\Big). \tag{16}$$

Thus, the probabilities of the home team's victory, draw, and away team's victory results are computed considering the possible outcomes of the joint distribution. Table 3 is an example to understand these probabilities with one match from the 2022/23 season.

Table 3: Arsenal x Chelsea Probabilities from the 34th round

	0	1	2	3	4	5	6
0	0.02754	0.04388	0.03496	0.01857	0.00740	0.00236	0.00063
1	0.05504	0.08770	0.06988	0.03712	0.01479	0.00471	0.00125
2	0.05500	0.08765	0.06983	0.03709	0.01478	0.00471	0.00125
3	0.03664	0.05839	0.04653	0.02471	0.00985	0.00314	0.00083
4	0.01831	0.02918	0.02325	0.01235	0.00492	0.00157	0.00042
5	0.00732	0.01166	0.00929	0.00494	0.00197	0.00063	0.00017
6	0.00244	0.00389	0.00310	0.00164	0.00066	0.00021	0.00006

We observe that, from the Table 3, we obtain the probability of Arsenal's victory by summing the red cells, the draw probability with the gray cells, and Chelsea's victory considering the blue cells. $P(Arsenal) \approx 0,47$, $P(Draw) \approx 0,22$ e $P(Chelsea) \approx 0,31$.

4.3. Performances

We replicated the analysis for each model in each round, where we computed the estimated probabilities of the home team's victory, draw, and the away team's victory. With these estimates, we can analyze the performances of our strategies with different model configurations. Table 4 contains the considered strategies along with the Log Loss, BS, and MRS metrics.

Table 4: Models with Log Loss, BS, and MRS

Models	**LogLoss**	**BS**	**MRS**
$PARX_I^*$ Ind	1.0115	0.6071	50.34%
$PARX_I^*$ Copula	1.0113	0.6070	50.34%
$PARX_I$ Ind	1.0019	0.6002	50.34%
$PARX_I$ Copula	1.0020	0.6002	50.34%
$PARX_L^*$ Ind	0.9982	0.5934	**55.17**%
$PARX_L^*$ Copula	**0.9980**	**0.5932**	**55.17**%
$PARX_L$ Ind	1.0279	0.6136	50.34%
$PARX_L$ Copula	1.0287	0.6139	50.34%

According to the metrics, all models performed better than the threshold calculated by the BS. Thus, we can observe that the model that stands out in performance according to the Log Loss and Brier Score was the $PARX_L^*Copula$ model, i.e., using copula modelling with marginals without considering covariate and with a logarithmic link function. This model also achieved an accuracy rate of 55.17% among the studied matches.

Finally, we implement the strategy $PARX_MInd$ and $PARX_MCopula$. First, we would like to use the best model among the four models, but we should not consider the $PARX_L$ in this strategy. We notice that the marginal model $PARX_L$ presented, in general, the best Akaike criterion for each team but a loss of predictive power.

Then, this strategy is composed of the $PARX_I^*$, $PARX_I$ and $PARX_L^*$ models to select the best marginal model. In this selection, the marginal models chosen using the Akaike criterion come from the models $PARX_I$ and $PARX_L^*$. The performance results are presented in Table 5.

Table 5: Best Marginal Model Strategy

Models	**LogLoss**	**BS**	**MRS**
$PARX_MInd$	0.9762	0.5784	**55.86**%
$PARX_MCopula$	**0.9758**	**0.5780**	**55.86**%

Therefore, among the fit strategies used, the $PARX_MCopula$ showed the best performance across the three metrics. We emphasize that the use of copulas slightly improves the results, but we stress that future investigation is needed to determine whether this difference is not significant.

4.4. Betting Aplication

Having identified the model with the best performance among the others, we can now use it as a reference for the implementation of a betting strategy aplication. We will adopt a strategy used in Angelini and De Angelis (2017). They relied on the strategies adopted by Dixon and Coles (1997) and Koopman and Lit (2015), who explored predictions obtained by their models.

Firstly, this strategy is based on analyzing profitable bets by relating the market odds and the probabilities estimated by the model. Each bet has odds, which represent the payout for the bet. For example, if a bet has odds of £4.50 and a stake of £1 is placed, the bettor will receive £4.50, making a profit of £3.50 (4.50 - 1 = £3.50).

Then, we will use bets with a value of £2 each and the odds considered will be an average of the odds proposed by the differents bookmakers considered in our study.

4.4.1. Betting Strategy

To illustrate the strategy used, let us consider an example from our sample: the match in the last round of the 2022/23 season between Manchester United and Fulham. The authors developed a betting strategy for the most popular bets offered in the market, for the results 1, X, and 2, corresponding to the home team's victory, a draw, and the away team's victory.

Let P_1, P_X, P_2 be the probabilities of interest and O_1, O_X, O_2 be the odds associated with the results 1, X, and 2, respectively. Then, the betting strategy proposed by the authors is based on:

1. The first step is to select the outcome with the highest probability.
   In the case of Manchester United vs. Fulham is the home team's victory ($P_1 = 0.7081$).

2. The second step is to decide if betting on this outcome is profitable.
   Let P_a^o be the implied probability defined by the inverse of the odds associated with result a, for a = 1, X, 2. In the example above, $P_1^o = 1.50^{-1} = 0.6667$. Therefore, according to the bookmakers, the probability of Manchester United winning is approximately 67%, against the predicted 70.8%. The proposed payout by the bookmaker is higher than expected.

Therefore, the expected value of the bet for outcome 1 is then given by

$$E[A_1] = \frac{P_1}{P_1^o} - 1 \tag{17}$$

The authors would only bet on the home team's victory if $E[A_1] > 0$, i.e., only if the estimated probability is higher than the implied probability proposed by the market ($P_1 > P_1^o$).

4.4.2. Threshold

The authors also consider an alternative strategy. In particular, they consider selecting only matches where $E[A_a] > \tau$, i.e., only if $P_a > P_a^o(1 + \tau)$, where $\tau > 0$, and $a = 1, X, 2$. For this reason, the authors only bet on matches with outcomes where the probabilities are higher than a specific threshold τ.

In the example considered, if $\tau = 0$, the case becomes the result mentioned above $E[A_1] > 0$. Adopting the alternative strategy, bets would only be placed if $0 < \tau < E[A_1] = \frac{P_1}{P_1^o} - 1 = 0.0621$. Hence, it is still convenient to bet on Manchester United's victory in this match as long as a threshold $\tau < 0.0621$ is selected.

Considering the $PARX_M Copula$ models that showed the best metrics for prediction, we present the results of comparisons with the odds in Table 6, considering different threshold values.

Table 6: Models Perfomances vs Odds

Threshold	**Nº Matches**	**Recommended for betting**	**Profit (£)**	**Return**
Threshold = 0	145	77	54.37	35.30%
Threshold = 0.25	145	45	39.71	44.12%
Threshold = 0.5	145	29	46.41	80.01%
Threshold = 1	145	13	27.352	105.20%

The Table 6 shows that it was possible to achieve the desired profit through the adopted strategy. Even with different values for the threshold parameter, the strategy functions as intended.

The chosen threshold plays an important role in profit and return obtained; higher values of this parameter select bets where the probabilities from our models are more distant from the implied probabilities, which explains the small number of indicated matches to bet on.

Another impactful factor in our result is the associated odds themselves since different bookmakers use different odds. That is, if this simulation were applied to a bookmaker considered inaccurate, the performance could be considered superior, just as a bookmaker with implied probabilities very close to those estimated by the models would indicate a lower number of matches for betting.

5. Conclusion

All bivariate models had a satisfactory performance in our analysis, each presenting an accuracy rate (MRS) exceeding 50%. Based on the performance metrics and Akaikes criterion, the marginal models that stand out are the $PARX_I$ and $PARX_L^*$, indicating significant results when considering log-linear models without covariates. Another crucial aspect in our results is the use of copulas, which generally improve prediction performances compared to models that assume independence.

Moreover, a bivariate model was constructed from the best marginal models for each team, achieving the most accurate prediction results with an accuracy rate (MRS) of approximately 55.86%. Through this model, an application in football betting was demonstrated, resulting in a positive financial return within the study's forecast window.

Therefore, we emphasize that the models studied in this article can be applied to other sports, as well as different contexts dealing with bivariate count data. For future studies, we recommend exploring others covariates in marginal models, mainly in log-linear models, and extending applications to other professional football leagues beyond the English Premier League. Finally, we conclude that the aims of this study were achieved.

Appendix A. PARX Estimation

The estimation for the PARX models are presented following the process in Agosto et al. (2016) and the summarized version in Angelini and De Angelis (2017). The conditional log-likelihood of the PARX model for the parameter vector $\boldsymbol{\theta} = (\omega, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta)'$ is given by

$$\ell_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} l_t(\boldsymbol{\theta}), \qquad l_t(\boldsymbol{\theta}) := y_t \log \lambda_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta}). \tag{A.1}$$

The maximum likelihood estimator (MLE) is then computed as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_T(\boldsymbol{\theta}). \tag{A.2}$$

The restrictions $\omega > 0, \beta_1, \ldots, \beta_q, \alpha_1, \ldots, \alpha_q, \eta \geq 0$ are required to guarantee that $\lambda_t > 0$ and $\sum_{j=1}^{\max(p,q)} \beta_j + \alpha_j < 1$ to ensure the stability of the process. These conditions imply that the PARX model admits a stationary and weakly dependent solution.

Then $\hat{\boldsymbol{\theta}}$ is obtained as the solution of $S_T(\boldsymbol{\theta}) = 0$, where the score $S_T(\boldsymbol{\theta})$ is defined as

$$S_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \left(\frac{y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\lambda)}{\partial \lambda}. \tag{A.3}$$

Furthermore, Theorem 2 in Agosto et al. (2016) shows that

$$\sqrt{T}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) \xrightarrow{d} N\big(0, H^{-1}(\boldsymbol{\theta}_0)\big), \qquad H(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right], \tag{A.4}$$

where the Hessian matrix $H(\boldsymbol{\theta})$ can be consistently estimated by

$$H_T(\boldsymbol{\theta}) = -\frac{1}{T} \sum_{t=1}^{T} \frac{1}{\lambda_t(\boldsymbol{\theta})} \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)'. \tag{A.5}$$

Therefore, the restrictions above mimic the ones used in GARCHX(p, q) models (Han and Kristensen, 2014) and are discussed more in detail in Agosto et al. (2016) along with other important properties of this estimation process.

Appendix B. Copulas

Gaussian Copula

Following Duque et al. (2021), we use the very popular Gaussian copula.

For our bivariate case (Ali et al., 2020) the Gaussian copula can be constructed based on Sklar (1959) using a standard normal distribution function $\Phi(\cdot)$ for the normal marginals with ρ as the dependece parameter ($\rho \in [-1, 1]$).

$$C\left(u, v\right) = P\Big(\Phi\left(u\right), \Phi\left(v\right)\Big) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{2\rho uv - u^2 - v^2}{2(1-\rho)^2}\right\} dudv. \quad \text{(B.1)}$$

Archimedean Copulas

Archimedean Copulas represent an important class of copulas with a specific property. This class uses Laplace transformations and mixtures of powers of univariate densities to create the multivariate distribution. Thus, the copula can be easily constructed from a generator function $\phi(\cdot)$ and a pseudo-inverse $\phi^{[-1]}(\cdot)$.

Definition. ***(Pseudo-Inversa)***. *Let ϕ be a continuous strictly decreasing function from $\boldsymbol{I} = [0, 1]$ to $[0, \infty]$ such that $\phi(1) = 0$. The pseudo-inverse of ϕ is given below:*

$$\phi^{[-1](t)} = \begin{cases} \phi^{-1}(t), & 0 \leq t \leq \phi(0) \\ 0, & \phi(0) \leq t \leq \infty \end{cases}$$

The pseudo-inverse will be a continuous and non-increasing function on $[0, \infty]$ and a strictly decreasing function on $[0, \phi(0)]$. If the generator $\phi(0) = \infty$, then $\phi^{[-1]}(t) = \phi^{-1}(t)$. Following Joe (1997), the generator function can be used to construct an Archimedean copula as follows:

$$C(u_1, \ldots, u_K) = \phi^{[-1]}\Big(\sum_{i=1}^{K} \phi(u_i)\Big) \quad \text{(B.2)}$$

Necessary and sufficient conditions on the generator functions $\phi(\cdot)$ and $\phi^{[-1]}(\cdot)$ in order to induce a valid Archimedean copula are discussed in Section 2 of McNeil and Nešlehová (2009).

In this work, using ρ as the dependence parameter, we consider Frank's copula ($p \in (-\infty, \infty)$) and Clayton's copula ($p \in (-1, \infty) \backslash \{0\}$) following Nelsen (2006),

$$C_{Frank}(u, v) = -\frac{1}{\rho} \ln\left(1 + \frac{(e^{-\rho u} - 1)(e^{-\rho v} - 1)}{e^{-\rho} - 1}\right), \quad \text{(B.3)}$$

$$C_{Clayton}(u, v) = \Big[\max\Big(u^{-\rho} + v^{-\rho} - 1, 0\Big)\Big]^{-1/\theta}. \quad \text{(B.4)}$$

Appendix C. PIT and Marginal Calibration

For model evaluation, the tscount package provides several tools. Thus, the package includes a graphical tool for checking probabilistic calibration with the Probability Integral Transform (PIT) histogram and a marginal calibration plot for assessing the fit.

As we have many estimated models, we will present an illustration of the evaluation process for one team in one of the rounds in the Figure C.7, as the models showed acceptable behaviors from the graphs.

Considering only the Everton team in round 34, we observe that the PIT histograms seem to be close to uniformity, and the marginal calibration plots are satisfactory, as they do not show outlier values.

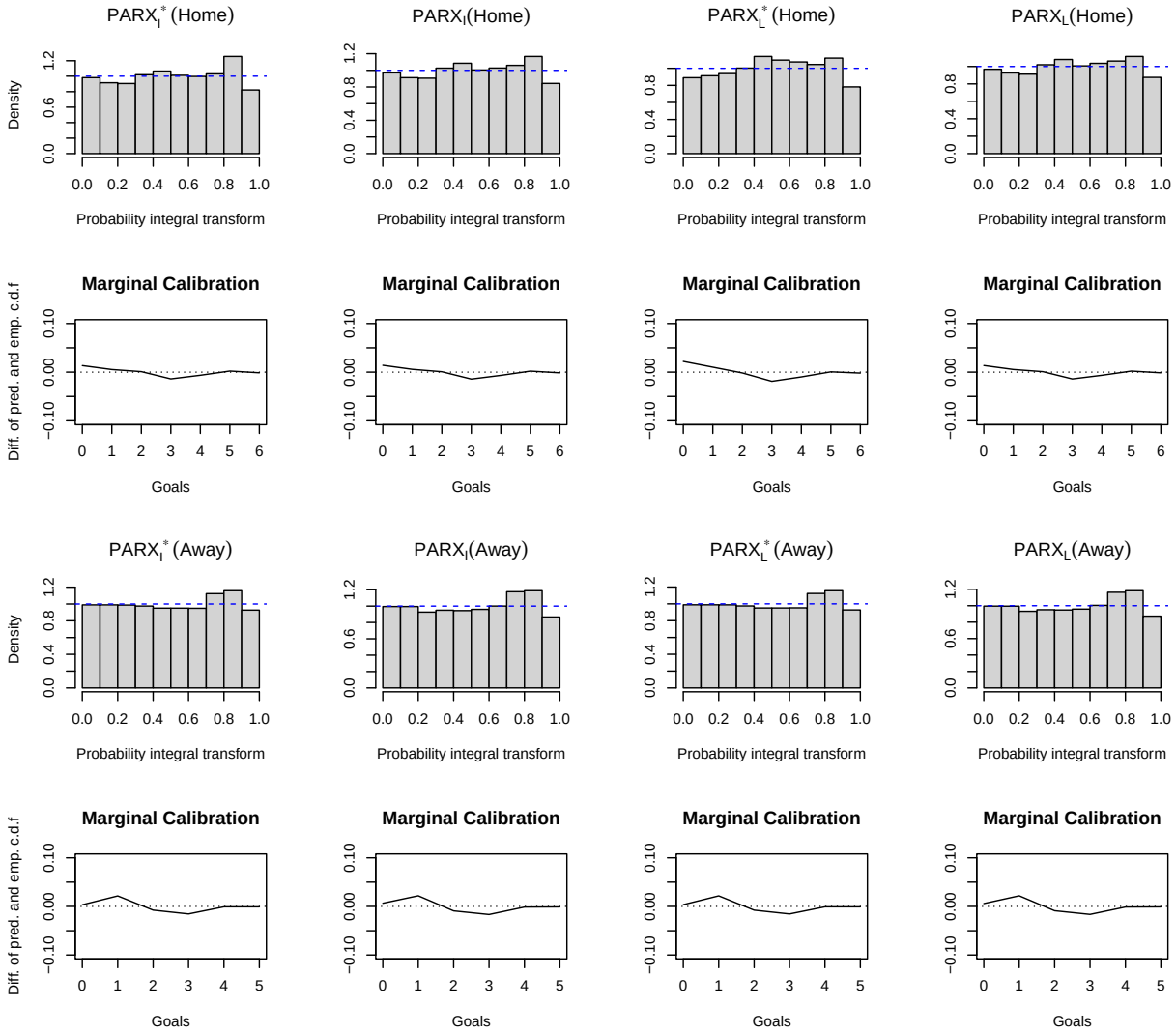

Figure C.7: Calibration of Everton models for round 34.

References

Agosto, A., Cavaliere, G., Kristensen, D., Rahbek, A., 2016. Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX). Journal of Empirical Finance 38, 640–663. URL: https://ideas.repec.org/a/eee/empfin/v38y2016ipbp640-663.html, doi:10.1016/j.jempfin.2016.02.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716–723. doi:10.1109/TAC.1974.1100705.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2020. Chapter 3 - monthly rainfall forecasting with markov chain monte carlo simulations integrated with statistical bivariate copulas, in: Samui, P., Tien Bui, D., Chakraborty, S., Deo, R.C. (Eds.), Handbook of Probabilistic Models. Butterworth-Heinemann, pp. 89–105. URL: https://www.sciencedirect.com/science/article/pii/B9780128165140000035, doi:https://doi.org/10.1016/B978-0-12-816514-0.00003-5.

Angelini, G., De Angelis, L., 2017. Parx model for football match predictions. Journal of Forecasting 36, 795–807. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2471, doi:https://doi.org/10.1002/for.2471, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2471.

Baio, G., Blangiardo, M., 2010. Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics 37, 253–264. URL: https://doi.org/10.1080/02664760802684177, doi:10.1080/02664760802684177, arXiv:https://doi.org/10.1080/02664760802684177.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Number 4 in Information science and statistics, Springer. URL: http://www.library.wisc.edu/selectedtocs/bg0137.pdf, doi:10.1117/1.2819119, arXiv:0-387-31073-8.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327. URL: https://www.sciencedirect.com/science/article/pii/0304407686900631, doi:https://doi.org/10.1016/0304-4076(86)90063-1.

Boshnakov, G., Kharrat, T., McHale, I.G., 2017. A bivariate weibull count model for forecasting association football scores. International Journal of Forecasting 33, 458–466. URL: https://www.sciencedirect.com/science/article/pii/S0169207017300018, doi:https://doi.org/10.1016/j.ijforecast.2016.11.006.

Brier, G.W., 1950. Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review 78, 1. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bunker, R.P., Thabtah, F., 2019. A machine learning framework for sport result prediction. Applied Computing and Informatics 15, 27–33. URL: https://www.sciencedirect.com/science/article/pii/S2210832717301485, doi:https://doi.org/10.1016/j.aci.2017.09.005.

Da Costa, I.B., Marinho, L.B., Pires, C.E.S., 2022. Forecasting football results and exploiting betting markets: The case of "both teams to score". International Journal of Forecasting 38, 895–909. URL: https://www.sciencedirect.com/science/article/pii/S0169207021001084, doi:https://doi.org/10.1016/j.ijforecast.2021.06.008.

Dixon, M., Coles, S., 1997. Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics) 46, 265–280. doi:10.1111/1467-9876.00065.

Duque, E.M.S., Vergara, P.P., Nguyen, P.H., van der Molen, A., Slootweg, J.G., 2021. Conditional multivariate elliptical copulas to model residential load profiles from smart meter data. IEEE Transactions on Smart Grid 12, 4280–4294. doi:10.1109/TSG.2021.3078394.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50, 987–1007. URL: http://www.jstor.org/stable/1912773.

FBref, . Football reference. https://fbref.com/pt.

Fokianos, K., Rahbek, A., Tjøstheim, D., 2009. Poisson autoregression. Journal of the American Statistical Association 104, 1430–1439. doi:10.1198/jasa.2009.tm08270.

Fokianos, K., Tjøstheim, D., 2011. Log-linear poisson autoregression. Journal of Multivariate Analysis , 563–578doi:10.1016/j.jmva.2010.11.002.

Guan, S., Wang, X., 2022. Optimization analysis of football match prediction model based on neural network. Neural Computing and Applications 34, 2525 – 2541. doi:10.1007/s00521-021-05930-x. cited by: 15.

Halliday, J., Boshnakov, G.N., 2018. Poarx modelling for multivariate count time series. arXiv: Methodology , 1–22URL: https://api.semanticscholar.org/CorpusID:54997080.

Han, H., Kristensen, D., 2014. Asymptotic theory for the qmle in garch-x models with stationary and nonstationary covariates. Journal of Business & Economic Statistics 32, 416–429. URL: https://doi.org/10.1080/07350015.2014.897954, doi:10.1080/07350015.2014.897954, arXiv:https://doi.org/10.1080/07350015.2014.897954.

Joe, H., 1997. Multivariate Models and Multivariate Dependence Concepts. Chapman and Hall/CRC eBooks. doi:10.1201/b13150.

Joe, H., Xu, J.J., 1996. The estimation method of inference functions for margins for multivariate models. Faculty Research and Publications URL: https://open.library.ubc.ca/collections/facultyresearchandpublications/52383/items/1.0225985, doi:http://dx.doi.org/10.14288/1.0225985.

Kharrat, T., Boshnakov, G.N., McHale, I., Baker, R., 2019. Flexible regression models for count data based on renewal processes: The countr package. Journal of Statistical Software 90, 1–35. URL: https://www.jstatsoft.org/index.php/jss/article/view/v090i13, doi:10.18637/jss.v090.i13.

Koopman, S., Lit, R., 2015. A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. Journal of the Royal Statistical Society. Series A. Statistics in Society 178, 167–186. doi:10.1111/rssa.12042.

Liboschik, T., Fokianos, K., Fried, R., 2017. tscount: An r package for analysis of count time series following generalized linear models. Journal of Statistical Software 82, 1–51. URL: https://www.jstatsoft.org/index.php/jss/article/view/v082i05, doi:10.18637/jss.v082.i05.

Maher, M.J., 1982. Modelling association football scores. Statistica Neerlandica 36, 109–118. URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9574.1982.tb00782.x, doi:https://doi.org/10.1111/j.1467-9574.1982.tb00782.x.

Mathur, N., 2023. Top 10 richest sports in the world in 2023. URL: https://thesportslite.com/net-worth/richest-sports-in-the-world/.

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

McNeil, A.J., Nešlehová, J., 2009. Multivariate archimedean copulas, d-monotone functions and l1-norm symmetric distributions. The Annals of Statistics 37. URL: http://dx.doi.org/10.1214/07-AOS556, doi:10.1214/07-aos556.

McShane, B., Adrian, M., Bradlow, E.T., Fader, P.S., 2008. Count models based on weibull interarrival times. Journal of Business & Economic Statistics 26, 369–378. URL: https://doi.org/10.1198/073500107000000278, doi:10.1198/073500107000000278, arXiv:https://doi.org/10.1198/073500107000000278.

MKTEsportivo, 2023. Estudo detalha crescimento de 360% no setor de apostas no brasil. URL: https://www.mktesportivo.com/2023/05/estudo-detalha-crescimento-de-360-no-setor-de-apostas-no-brasil.

Nelsen, R.B., 2006. An Introduction to Copulas. 2 ed., Springer, New York.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Sabino da Silva, F.A., Ziegelmann, F.A., Caldeira, J.F., 2023. A pairs trading strategy based on mixed copulas. The Quarterly Review of Economics and Finance 87, 16–34. URL: https://www.sciencedirect.com/science/article/pii/S1062976922001223, doi:https://doi.org/10.1016/j.qref.2022.10.007.

Santana, H., Ferreira da Silva, P.H., Ara, A., Louzada, F., Suzuki, A., 2020. Modelagem estatística e aprendizado de máquina: Previsão do campeonato brasileiro série a 2017. Matemática e Estatística em Foco 7.

Shah, K., Hyman, J., Samangy, D., 2021. A poisson betting model with a kelly criterion element for european soccer. MIT SLOAN Sports Analytics Conference URL: https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/607a445eee46ee3ac33595d3_KushalShah-PoissonBetting-RPpaper.pdf.

Silva Filho, O.C.d., Ziegelmann, F.A., Dueker, M.J., 2012. Modeling dependence dynamics through copulas with regime switching. Insurance: Mathematics and Economics 50, 346 – 356. doi:10.1016/j.insmatheco.2012.01.001. cited by: 75.

Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. l'Institut de statistique de l'Université de Paris 8, 229–231.

Tootoonchi, F., Sadegh, M., Haerter, J.O., Räty, O., Grabs, T., Teutschbein, C., 2022. Copulas for hydroclimatic analysis: A practice-oriented overview. Wiley Interdisciplinary Reviews: Water 9. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123889618&doi=10.1002%2fwat2.1579&partnerID=40&md5=7a5bc97108b760197dedd66d654a02ba, doi:10.1002/wat2.1579. cited by: 33; All Open Access, Green Open Access, Hybrid Gold Open Access.

UEFA, . Ten-year club coefficients. URL: https://www.uefa.com/nationalassociations/uefarankings/tenyears/.

Yan, J., 2007. Enjoy the joy of copulas: With a package copula. Journal of Statistical Software 21, 1–21. URL: https://www.jstatsoft.org/index.php/jss/article/view/v021i04, doi:10.18637/jss.v021.i04.

Zivkovic, J., 2022. worldfootballR: Extract and Clean World Football (Soccer) Data. URL: https://CRAN.R-project.org/package=worldfootballR. r package version 0.6.2.

Referências Bibliográficas

Agosto, A., Cavaliere, G., Kristensen, D., e Rahbek, A. (2016). Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX). *Journal of Empirical Finance*, 38(PB):640–663.

Angelini, G. e De Angelis, L. (2017). Parx model for football match predictions. *Journal of Forecasting*, 36(7):795–807.

Baio, G. e Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264.

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

Bunker, R. P. e Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1):27–33.

Da Costa, I. B., Marinho, L. B., e Pires, C. E. S. (2022). Forecasting football results and exploiting betting markets: The case of "both teams to score". *International Journal of Forecasting*, 38(3):895–909.

Dixon, M. e Coles, S. (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.

Fokianos, K., Rahbek, A., e Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439.

Guan, S. e Wang, X. (2022). Optimization analysis of football match prediction model based on neural network. *Neural Computing and Applications*, 34(4):2525 – 2541. Cited by: 15.

Halliday, J. e Boshnakov, G. N. (2018). Poarx modelling for multivariate count time series. *arXiv: Methodology*, pages 1–22.

Kharrat, T., Boshnakov, G. N., McHale, I., e Baker, R. (2019). Flexible regression models for count data based on renewal processes: The countr package. *Journal of Statistical Software*, 90(13):1–35.

Koopman, S. e Lit, R. (2015). A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, 178(1):167–186.

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

Mathur, N. (2023). Top 10 richest sports in the world in 2023.

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

McShane, B., Adrian, M., Bradlow, E. T., e Fader, P. S. (2008). Count models based on weibull interarrival times. *Journal of Business & Economic Statistics*, 26(3):369–378.

MKTEsportivo (2023). Estudo detalha crescimento de 360% no setor de apostas no brasil.

Sabino da Silva, F. A., Ziegelmann, F. A., e Caldeira, J. F. (2023). A pairs trading strategy based on mixed copulas. *The Quarterly Review of Economics and Finance*, 87:16–34.

Santana, H., Ferreira da Silva, P. H., Ara, A., Louzada, F., e Suzuki, A. (2020). Modelagem estatística e aprendizado de máquina: Previsão do campeonato brasileiro série a 2017. *Matemática e Estatística em Foco*, 7.

Shah, K., Hyman, J., e Samangy, D. (2021). A poisson betting model with a kelly criterion element for european soccer. *MIT SLOAN Sports Analytics Conference*.

Silva Filho, O. C. d., Ziegelmann, F. A., e Dueker, M. J. (2012). Modeling dependence dynamics through copulas with regime switching. *Insurance: Mathematics and Economics*, 50(3):346 – 356. Cited by: 75.

Tootoonchi, F., Sadegh, M., Haerter, J. O., Räty, O., Grabs, T., e Teutschbein, C. (2022). Copulas for hydroclimatic analysis: A practice-oriented overview. *Wiley Interdisciplinary Reviews: Water*, 9(2). Cited by: 33; All Open Access, Green Open Access, Hybrid Gold Open Access.