Computational Statistics
https://doi.org/10.1007/s00180-024-01585-z

ORIGINAL PAPER

Alternative ranking measures to predict international football results

Roberto Macrì Demartino[1] · **Leonardo Egidi**[2] · **Nicola Torelli**[2]

Received: 30 April 2024 / Accepted: 25 November 2024

Abstract

Over the last few years, there has been a growing interest in the prediction and modelling of competitive sports outcomes, with particular emphasis placed on this area by the Bayesian statistics and machine learning communities. In this paper, we have carried out a comparative evaluation of statistical and machine learning models to assess their predictive performance for the 2022 FIFA World Cup and the 2023 CAF Africa Cup of Nations by evaluating alternative summaries of past performances related to the involved teams. More specifically, we consider the Bayesian Bradley-Terry-Davidson model, which is a widely used statistical framework for ranking items based on paired comparisons that have been applied successfully in various domains, including football. The analysis was performed including in some canonical goal-based models both the Bradley-Terry-Davidson derived ranking and the widely recognized Coca-Cola FIFA ranking commonly adopted by football fans and amateurs.

Keywords Bayesian statistics · Bradley-Terry-Davidson model · Prediction · World Cup

✉ Roberto Macrì Demartino
roberto.macridemartino@phd.unipd.it

Leonardo Egidi
legidi@units.it

Nicola Torelli
nicola.torelli@deams.units.it

[1] Department of Statistical Sciences, University of Padova, Via C. Battisti 241, 35121 Padova, Veneto, Italy

[2] Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Via Valerio 4/1, 34127 Trieste, Friuli-Venezia Giulia, Italy

Published online: 06 December 2024

1 Introduction

The application of statistical and machine learning models in forecasting international football competitions, such as the FIFA World Cup or UEFA Champions League, has always attracted the interest of several analysts.

From a statistical perspective, the outcome of a football match may be predicted using two different approaches. The result-based approach (Koning 2000; Carpita et al. 2019, among others), which directly predicts the match result, explicitly modelling the so-called three-way process - home win, draw, or away win - using a logistic or multinomial regression model. Alternatively, the goal-based approach models the goals scored and conceded in each match by modelling count variables, typically by using Poisson regressions, and then determines the exact match result by comparing these scores. It is worth noting that once a goal-based model has been estimated, it becomes possible to derive the three-way process by simply aggregating the estimated probabilities. However, for a more comprehensive comparison of goal-based and result-based statistical methods, see Egidi and Torelli (2021).

This paper investigates the potential improvement in predictive performance achieved by some statistical goal-based methods and machine learning result-based algorithms when an appropriate measure of the teams' relative strength is used as an additional predictor. To assess it, we analyze the outcomes of the 2022 FIFA World Cup in Qatar and the 2023 Africa Cup in Ivory Coast. We considered a Bayesian Bradley-Terry-Davidson derived ranking system, using the posterior median of the log-strength parameters as a novel predictor. Subsequently, we compare the predictive performances of this approach with those obtained using the well-established FIFA ranking – in terms of FIFA ranking points – to determine which provides more reliable forecasts.

In the goal-based approach, each game involves evaluating the goal counts for each team. Furthermore, the expected number of goals is determined by team attributes such as offensive and defensive abilities, and home advantage when applicable. Under some basic assumptions, the team-specific Poisson distributions are considered independent, resulting in a double Poisson model (Maher 1982; Baio and Blangiardo 2010; Groll and Abedieh 2013; Egidi et al. 2018, among others). However, these Poisson-based approaches can be generalized in different ways to capture the dependence between scores. For instance, Dixon and Coles (1997) extended the work of Maher (1982) by allowing (a slightly negative) correlation between scores and incorporating a dependence parameter within their model to account for it. Furthermore, the bivariate Poisson model, designed to account for positive goal dependencies, was developed by Karlis and Ntzoufras (2003) within a frequentist framework and by Ntzoufras (2011) from a Bayesian perspective. A key limitation of previous models lies in their assumption of invariant team-specific parameters, implying that team performance remains constant over time based on their offensive and defensive abilities. However, it is recognised that team performance is inherently dynamic, fluctuating over years and possibly within seasons. Rue and Salvesen (2000) proposed a dynamic extension

for the double Poisson model on continuous time, while Owen (2011) proposed a discrete time random walk approach for both offensive and defensive parameters. In addition, Koopman and Lit (2015) further extended the bivariate Poisson model into a state-space framework, allowing team abilities to vary according to a state vector.

A fundamentally different modelling approach has emerged due to the availability of large volumes of data, which has led to the development of a range of machine learning result-based techniques. These include artificial neural networks (ANNs), multivariate adaptive regression splines (MARS) (Friedman 1991), and ensemble learning methods such as random forests (Breiman 2001). Notably, the predictive performance of several random forests configurations has been examined and evaluated in the context of international football matches (Schauberger and Groll 2018; Groll et al. 2019, 2021, among others).

While the aforementioned methodologies provide robust frameworks for modelling football match outcomes, their predictive performance may be improved by incorporating additional historical information. Typically, one approach to potentially improve predictive accuracy is to integrate established rankings, such as the FIFA ranking, as additional model covariates. Specifically, these covariates are based on the quantitative measures of team strength from which the overall ranking is then derived. Notably, the algorithm underlying the FIFA ranking system has undergone a significant revision since 2018. The new algorithm takes into account not only the outcome of the single matches but also the strength level of teams before each match. For further details see Szczecinski and Roatis (2022). The resulting algorithm offers benefits in terms of both simplicity and relative transparency. However, an alternative and particularly interesting methodology for obtaining a ranking, based on pairwise comparisons between teams, is the Bradley-Terry model (Bradley and Terry 1952). The model assigns a strength parameter to each team, and the odds of winning a match are determined by the ratio of these parameters. The estimated strengths can be employed to construct a rating system that reflects the ranking of teams based on the outcome of matches and the competitive interactions between teams.

In order to encompass a range of limitations, a number of extensions and generalizations to the original Bradley-Terry model have been developed over the years. For instance, Rao and Kupper (1967) and Davidson (1970) extended the model applicability to scenarios in which a draw is a possible outcome by including a novel parameter which affects the probability of a tie in a match. Springall (1973) proposed a generalization of the Bradley-Terry model with team-dependent linear predictors. Davidson and Solomon (1973) proposed a Bayesian version of the Bradley-Terry model, using a family of conjugate prior distributions to compute the posterior distribution of the log-strength parameters. Moreover, Leonard (1977) suggested a more flexible Bayesian approach using non-conjugated multivariate Gaussian prior distributions for the log-strengths. Since these early works, there have been numerous contributions that have further extended the Bayesian framework for paired comparison models (Chen and Smith 1984; Caron and Doucet 2012; Whelan 2017; Osei and Davidov 2022; Wainer 2023, among others). The Bradley-Terry model can be further extended to account for situations where the order of comparisons can

influence the outcome. A classic example is the home-field advantage in football, where the team playing at home may have a psychological or logistical advantage compared to the visiting team. Specifically, Beaver and Gokhale (1975) and Davidson and Beaver (1977) introduced additive and multiplicative order effects, respectively. Additionally, several extensions have been developed to model the dynamic variation of strengths over time. (Fahrmeir and Tutz 1994; Glickman 1999, 2001; Cattelan et al. 2012; Tian et al. 2023, among others).

Our work focuses on extending the well-known goal-based and result-based protocols by introducing alternative ranking measures for international football matches that could serve as a valuable computational routine for practitioners. The rest of the paper is organized as follows. Section 2 presents the theoretical framework, introducing the standard Bradley-Terry model and its Davidson extension for handling draws, followed by a discussion of the Bayesian approach. Furthermore, Sect. 3 describes the statistical goal-based methods and the machine learning result-based algorithms used in this study. In Sect. 4, we evaluate the application of these methodologies on the data from both the last World Cup and Africa Cup of Nations. Finally, Sect. 5 provides concluding remarks, outlining the limitations, advantages, and potential future research directions.

2 The Bradley-Terry model

The Bradley-Terry model (Bradley and Terry 1952) is one of the most popular modelling techniques in a pairwise comparison context for ranking players or teams. The model assumes that each team T_k, with $k = 1, \dots, N_T$, is characterized by a latent parameter, $\alpha_k > 0$, representing its intrinsic strength. The outcome of any given comparison is modelled as an independent Bernoulli random variable, where the probability of each outcome is a function of the strengths of the teams involved. Specifically, for a match between team T_i and team T_j, with $i \neq j = 1, \dots, N_T$, the probability that T_i defeats T_j in the n-th match, with $n = 1, \dots, N$, is

$$p_{ij}^{W} = \mathbb{P}(T_i \text{ defeats } T_j) = \frac{\alpha_i}{\alpha_i + \alpha_j}, \tag{1}$$

where α_i and α_j are the strength parameters of the teams involved in the match. These parameters are invariant to a multiplicative constant. Therefore, parameter identifiability is obtained by imposing a constraint such as $\sum_{k=1}^{N_T} \alpha_k = 1$. Furthermore, the final ranking of the teams can be determined by sorting their respective strength parameters α_k.

The model in (1) is commonly reparameterized by the logarithm of the strength parameters

$$p_{ij}^{W} = \frac{\exp(\psi_i)}{\exp(\psi_i) + \exp(\psi_j)}, \tag{2}$$

where $\psi_i = \log(\alpha_i)$ and $\psi_j = \log(\alpha_j)$. Since the α values are invariant to multiplicative constants, the ψ values are invariant to additive constants. Consequently, the parameters are identifiable if $\sum_{k=1}^{N_T} \psi_k = 0$. This transformation offers several advantages. Notably, it enables the estimation of the log-strength parameter across an expanded parameter space, $\psi \in (-\infty, +\infty)$, providing greater flexibility for the application of a wide class of priors in the Bayesian setting. Furthermore, the logit transformation facilitates parameter estimation within the frequentist framework using generalized linear models (GLMs) (Cattelan 2012).

2.1 Dealing with draws

The standard Bradley-Terry model does not account for draws. Several alternatives have been proposed to address this limitation, including the assignment of draws as wins to both teams, the spreading of draws as half a win to each team, the non-consideration of draws as wins, and the random assignment of draws as wins to one of the two teams. However, none of these approaches directly incorporates the possibility of a tie into the model.

To address this, Rao and Kupper (1967) extended the Bradley-Terry model to accommodate draws by introducing an additional parameter η, and explicitly modelling its probability as follows

$$p_{ij}^W = \frac{\alpha_i}{\alpha_i + \eta\alpha_j},$$
$$p_{ij}^D = \mathbb{P}(T_i \text{ draw } T_j) = \frac{(\eta^2 - 1)\alpha_i\alpha_j}{(\alpha_i + \eta\alpha_j)(\eta\alpha_i + \alpha_j)}.$$

If $\eta = 1$ then the Rao-Kupper model reduces to the standard Bradley-Terry model. Furthermore, using the log-parametrization as in (2), the model is

$$p_{ij}^W = \frac{\exp(\psi_i)}{\exp(\psi_i) + \exp(\gamma + \psi_j)},$$
$$p_{ij}^D = \frac{(\exp(2\gamma) - 1)\exp(\psi_i + \psi_j)}{[\exp(\psi_i) + \exp(\gamma + \psi_j)][\exp(\gamma + \psi_i) + \exp(\psi_j)]},$$

where $\gamma = \log(\eta)$.

An alternative approach, which adheres to the ratio scale required by the so-called choice axiom (Luce 1959), was proposed by Davidson (1970). As in Rao and Kupper (1967), the Bradley-Terry-Davidson (BTD) model introduces an additional parameter that balances the probability of ties against the probability of not having ties and computes three different probabilities. The log-parametrization of the model is

$$
\begin{aligned}
p_{ij}^{W} &= \frac{\exp(\psi_i)}{\exp(\psi_i) + \exp(\psi_j) + \exp(\gamma + (\psi_i + \psi_j)/2)}, \\
p_{ij}^{D} &= \frac{\exp(\gamma + (\psi_i + \psi_j)/2)}{\exp(\psi_i) + \exp(\psi_j) + \exp(\gamma + (\psi_i + \psi_j)/2)}, \\
p_{ij}^{L} &= \frac{\exp(\psi_j)}{\exp(\psi_i) + \exp(\psi_j) + \exp(\gamma + (\psi_i + \psi_j)/2)}.
\end{aligned} \tag{3}
$$

Since the three-way process events are mutually exclusive, the following constraint is imposed $p_{ij}^{W} + p_{ij}^{D} + p_{ij}^{L} = 1$. It is worth noticing that if the draw parameter γ increases towards $+\infty$ then the probability of a tie p_{ij}^{D} approaches one. Conversely, if γ decreases towards $-\infty$ then p_{ij}^{D} approaches to zero. Finally, if γ is equal to zero then p_{ij}^{W}, p_{ij}^{D}, and p_{ij}^{L} depend solely on the strengths of the competing teams.

For the remainder of the paper, we will focus specifically on Davidson's proposal for dealing with draws.

2.2 The Bayesian approach

In the frequentist paradigm, teams are ranked using maximum likelihood estimates (MLE) of the strength parameters (Ford 1957; Hunter 2004). However, the Bayesian framework offers a different perspective by providing the posterior distribution of the strength parameters reflecting the inherent uncertainty in the ranking system. Here, we introduce the hierarchical Bayesian formulation of the Bradley-Terry model incorporating the Davidson extension for handling draws as in (3).

The Bayesian BTD model requires specifying prior distributions for both the team log-strength parameters and the draw parameter. Specifically, the prior placed on γ reflects our initial belief about the impact of teams' strengths on tie outcomes (Issa Mattos and Martins Silva Ramos 2022). When defining priors within this framework, Whelan (2017) proposed a set of desirable properties specifically suited for ranking systems. These properties aim to construct priors that avoid introducing unfair advantages or disadvantages for any particular team. Ideally, the prior should maintain invariance when teams are swapped, should not be affected by switching the outcome of the match for any given comparison, removing teams from the competition should not alter the prior distribution, and the prior should be proper. Specifically, employing a multivariate Gaussian for the log-strengths (Leonard 1977), or identical independent Gaussian distributions for each log-strength parameter, satisfies all the four conditions.

Based on this, let w_{ij} represent the binary outcome where team T_i defeats team T_j, and let d_{ij} indicate the binary outcome of a draw between teams T_i and T_j. Then, the hierarchical Bayesian BTD model is

$$\begin{aligned} w_{ij} \mid p_{ij}^{W} &\sim \text{Bernoulli}(p_{ij}^{W}), \\ d_{ij} \mid p_{ij}^{D} &\sim \text{Bernoulli}(p_{ij}^{D}), \\ \psi &\sim \text{N}(\mu_{\psi}, \sigma_{\psi}^{2}), \\ \gamma &\sim \text{N}(\mu_{\gamma}, \sigma_{\gamma}^{2}), \end{aligned} \tag{4}$$

where μ_{ψ} and μ_{γ} are the mean for the team log-strength and draw parameters, and σ_{ψ}^{2} and σ_{γ}^{2} denote the corresponding variances.

3 Statistical models and machine learning algorithms

This section describes the statistical models and machine learning algorithms employed to predict the outcomes of the considered competitions. Through a detailed analysis of goal-based models and result-based machine learning algorithms, this section aims to provide a comprehensive overview of the methodologies employed in the prediction of football matches, emphasizing their statistical foundations and practical implementations in sports analytics.

3.1 Goal-based models

Goal-based models assume that the number of goals scored in a match by each team follows a discrete distribution, typically two independent Poisson or a bivariate Poisson accounting for positive correlation. Thus, for each match, we need to consider the pair of counts (X_{in}, Y_{jn}), for $i \neq j = 1, \ldots, N_T$ and $n = 1, \ldots, N$. The first count X_{in} denotes the non-negative number of goals scored by the home team T_i and the second count Y_{jn} denotes the number of goals scored by the visiting team T_j, both in the n-th match. A simple double Poisson model is

$$\begin{aligned} X_{in} \mid \lambda_{1n} &\sim \text{Poisson}\left(\lambda_{1n}\right) \\ Y_{jn} \mid \lambda_{2n} &\sim \text{Poisson}\left(\lambda_{2n}\right) \\ \log\left(\lambda_{1n}\right) &= \theta + \text{att}_{h_n} + \text{def}_{a_n} + \frac{\phi}{2}\omega_n, \\ \log\left(\lambda_{2n}\right) &= \theta + \text{att}_{a_n} + \text{def}_{h_n} - \frac{\phi}{2}\omega_n, \end{aligned} \tag{5}$$

where λ_{1n} and λ_{2n} describe the expected number of goals for the home team and the away team, respectively. In particular, θ denotes a common baseline parameter, the parameters *att* and *def* represent the unknown attack and defense abilities for the home team h_n and the away team a_n in the n-th match. Furthermore, $\omega_n = (rank\_points_{h_n} - rank\_points_{a_n})$ captures the difference in FIFA ranking points (BTD relative log-strengths) between the home and away teams in the n-th match. Finally, the parameter ϕ tries to correct for the ranking points difference occurring

between two competing teams. A sum-to-zero constraint (Baio and Blangiardo 2010) is imposed on the attack and defense parameters to ensure model identifiability.

A key limitation of the double Poisson model lies in its assumption of conditional independence between the goals scored by competing teams. However, in interactive team sports like football, a degree of correlation between goal outcomes is likely due to on-field interactions. This correlation could reflect changes in playing style by one or both teams throughout the match. To address this limitation and account for the positive dependence between goal counts, a bivariate Poisson model (Karlis and Ntzoufras 2003) for each pair of counts can be considered

$$\begin{aligned}\left(X_{in}, Y_{jn} \mid \lambda_{1n}, \lambda_{2n}, \lambda_{3n}\right) &\sim \text{BivPoisson}\left(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}\right)\\ \log\left(\lambda_{3n}\right) &= \beta_0,\end{aligned} \tag{6}$$

where λ_{1n} and λ_{2n} are defined as in (5), whereas the coefficient λ_{3n} describes the dependence between the two random counts. Furthermore, all the other parameters have the same interpretation as in (5). Notably, when $\lambda_{3n} = 0$, the two components are independent, then the bivariate Poisson model reduces to a double Poisson model. We note that in (6) we let the covariance λ_{3n} to not depend on other predictors, thus we assume it is equal for each match n: However, one could assume an extended linear predictor with match-dependent covariates, as specified in Karlis and Ntzoufras (2003).

Poisson goal-based models may suffer from an underestimation of the number of draws, represented by the outcomes in the diagonal of the probability table. To address this issue, Karlis and Ntzoufras (2009) introduced a zero-inflated model for favoring the draw outcome. The diagonal-inflated bivariate Poisson model is defined as follows

$$\mathbb{P}(X_n = x_n, Y_n = y_n) = \begin{cases} (1-p)\,\text{BP}\left(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}\right) & \text{if } x_n \neq y_n \\ (1-p)\,\text{BP}\left(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}\right) + pD(x_n, \xi) & \text{if } x_n = y_n \end{cases}, \tag{7}$$

where $D(x_n, \xi)$ is a discrete distribution with parameter vector ξ.

Following Owen (2011) and Egidi et al. (2018), we introduce a dynamic assumption regarding team-specific effects for the models presented in equations (5), (6) and (7). A first-order autoregressive model is adopted by centering the effect of seasonal time τ on the lagged effect in $\tau - 1$, plus a fixed effect. This allows attack and defense parameters to vary across seasons, where a season corresponds to a year. Therefore, for each team i, where $i = 1, \ldots, N_T$, and each year τ, where $\tau = 2, \ldots, \mathcal{T}$, the prior distributions for the attack and defense abilities are defined as follows

$$\begin{aligned}\text{att}_{i,\tau} &\sim \text{N}\left(\text{att}_{i,\tau-1}, \sigma^2_{\text{att}}\right)\\ \text{def}_{i,\tau} &\sim \text{N}\left(\text{def}_{i,\tau-1}, \sigma^2_{\text{def}}\right).\end{aligned}$$

For the initial season $\tau = 1$, the prior distributions are initialized as

$$\begin{aligned}\text{att}_{i,1} &\sim \text{N}\left(\mu_{\text{att}}, \sigma^2_{\text{att}}\right)\\ \text{def}_{i,1} &\sim \text{N}\left(\mu_{\text{def}}, \sigma^2_{\text{def}}\right),\end{aligned}$$

where μ_{att} and μ_{def} are the mean for the initial attack and defense abilities, and σ^2_{att} and σ^2_{def} are their corresponding variances. As with the static models, the dynamic extension also imposes a sum-to-zero constraint on these random effects within each season for identifiability.

3.2 Result-based algorithms

Random forests (Breiman 2001) are ensemble learning algorithms that combine the predictions of a large number of decision trees. These methods are typically constructed from a large number of classification trees grown on bootstrap samples drawn from the original dataset. Notably, the aggregation of multiple trees offers several advantages. The resulting predictions inherit the unbiasedness of individual trees while exhibiting reduced variance. Additionally, the trees within a random forest are grown independently. This independence helps to reduce the overall variance of the ensemble compared to a single tree. To achieve this goal, random forests typically incorporate two key randomization steps during the tree building process. Furthermore, several studies have demonstrated the efficacy of random forests in predicting international football match outcomes. These contributions consistently report better performance compared to regression approaches (Schauberger and Groll 2018; Groll et al. 2019, 2021, among others).

Artificial neural networks (ANNs) represent a class of complex computational models inspired by the interconnected structure of neurons in the human brain. These models excel at processing information and learning from data through a layered architecture. Each layer comprises interconnected nodes that apply weights and biases to process inputs. The learning process involves adjusting these weights and biases to optimize the network's performance. Specifically, ANNs proved successful in predicting football match outcomes by considering historical data that include a wide range of information, such as team performance metrics, match results, and even individual player statistics (Huang and Chang 2010; Hucaljuk and Rakipovic 2011; Danisik et al. 2018, among others).

Multivariate Adaptive Regression Splines (MARS) was first proposed by Friedman (1991) as an algorithm to model non-linear relationships, particularly those that are nearly additive or involve low-order interactions between variables. Essentially, the algorithmic procedure involves a piecewise linear regression model. This allows the slope of the regression line to change from one interval to the other as the two knots are crossed. The selection of variables and knot locations is determined through a computationally efficient but intensive forward-backward search procedure. Notably, Abreu et al. (2013) applied a MARS algorithm to investigate the relationship between the number of goals scored and the final game statistics.

3.3 Computational procedure

In the statistical models and machine learning algorithms described in Sects. 3.1 and 3.2 an additional predictor, determined by the difference in FIFA ranking points (BTD relative log-strengths) between the home team and the away team,

is incorporated. Specifically, for the BTD relative log-strengths, this process involves initially fitting the Bayesian BTD model as described in Equation (4). Subsequently, the posterior median for each team's log-strength parameter is computed. Finally, the difference in the posterior medians of the competing teams is included as an additional predictor in both the goal-based models and result-based algorithms.

The computational steps for integrating the Bayesian BTD relative log-strengths into the statistical models and machine learning algorithms are summarized in Algorithm 1.

Algorithm 1 Bayesian BTD computational steps

1: Fit the Bayesian BTD model as described in (4).
2: For $k = 1, \ldots, N_T$ compute the posterior median for each team's log-strength parameter ψ_k.
3: For $n = 1, \ldots, N$ incorporate the difference in the posterior median of the competing teams' log-strengths, $\omega_n = \psi_{h_n} - \psi_{a_n}$, as an additional predictor in both the goal-based models and result-based algorithms for the n-th match of the competition.

4 Applications

The selection of the training and test sets is crucial and is likely to influence the predictions. We address this by employing an iterative training approach. Our goal-based statistical models and result-based machine learning algorithms are trained on a continuously updated dataset encompassing international matches from 2018 to 2023. The matches vary from the FIFA World Cups through the UEFA Euro Championships to normal friendly matches. The data excludes Olympic Games and matches in which at least one of the teams was the national B-team or a U-23 lineup. This approach allows for the incorporation of recent results, potentially improving predictive performance for two distinct scenarios: the group stage and the knockout stage of these two tournaments.

We evaluate the predictive performance of three dynamic goal-based Poisson models, implemented using the `footBayes` package (Egidi and Palaskas 2022), and three result-based machine learning techniques provided by the `caret` package (Kuhn 2022), both implemented in R, along the lines described in Sect. 3. Furthermore, we incorporate additional historical information from both the FIFA ranking – through the FIFA ranking points – and the Bayesian BTD derived ranking – through its relative log-strengths – using the `bpcs` R package (Issa Mattos and Martins Silva Ramos 2020).

To ensure a more comparable analysis, both FIFA ranking points and BTD relative log-strengths are normalized using the scaled median absolute deviation (MAD) normalization

$$x_{\text{MAD}} = \frac{x - \mathbb{M}(x)}{\mathbb{M}(|x - \mathbb{M}(x)|)},$$

where $\mathbb{M}(\cdot)$ is the median. Furthermore, to assess the predictive performance of the models described in Sect. 3, we employed the Brier score (Brier 1950) as recommended by Spiegelhalter and Ng (2009). It is essentially a mean squared error for forecasts where a lower score indicates greater model predictive accuracy. A common formulation is

$$b = \frac{1}{N} \sum_{n=1}^{N} \sum_{r=1}^{3} (p_{rn} - \delta_{rn})^2,$$

where p_{rn} represents the predicted probability of outcome r, with $r \in \{\text{win}, \text{draw}, \text{loss}\}$, for the n-th match. Here, δ_{rn} denotes the Kronecker delta, which equals 1 if the actual outcome of the n-th match corresponds to r. The lower bound of the Brier score is 0, which occurs when the predicted probabilities are perfectly accurate. In the case of three categories, the upper bound of the Brier score is 2. This happens when the worst prediction is made by assigning a probability of 1 to an incorrect category and 0 to the correct category. This results in a squared difference of 1 for both the correct and the selected incorrect category. Thus, if two such errors are made, the sum is 2.

4.1 2022 FIFA World Cup

The World Cup presents an interesting case study due to the diverse range of National teams with heterogeneous strengths. We specifically investigate the performance of teams in a structured environment, such as the group stage, and contrast it with the dynamic and high-pressure setting of the knockout stage, where single-elimination games can significantly impact teams' behaviour. This subsection describes how we employ both goal-based statistical models and machine learning algorithms to forecast match outcomes throughout the tournament, assessing their predictive performance by adding the historical information from the ranking systems outlined previously. Notably, we consider the FIFA ranking published just before the World Cup took place, available at https://inside.fifa.com/fifa-world-ranking/men?dateId=id13869.

Figure 1 presents scatterplot between the normalized Bayesian BTD relative log-strengths and the normalized FIFA points. The high value of the Pearson correlation coefficient ($\rho_p = 0.90$) suggests a positive linear relationship between these two variables. This is further corroborated by the Spearman correlation coefficient ($\rho_s = 0.88$). Additionally, the Kendall coefficient confirms a substantial positive association ($\tau = 0.69$), albeit slightly weaker than the Pearson and Spearman correlations.

The top panel of Fig. 2 presents scatterplots comparing the relative strengths of competing teams in both the group and knockout stages under the two ranking systems. Points above the dashed line represent matches where the "away" team had

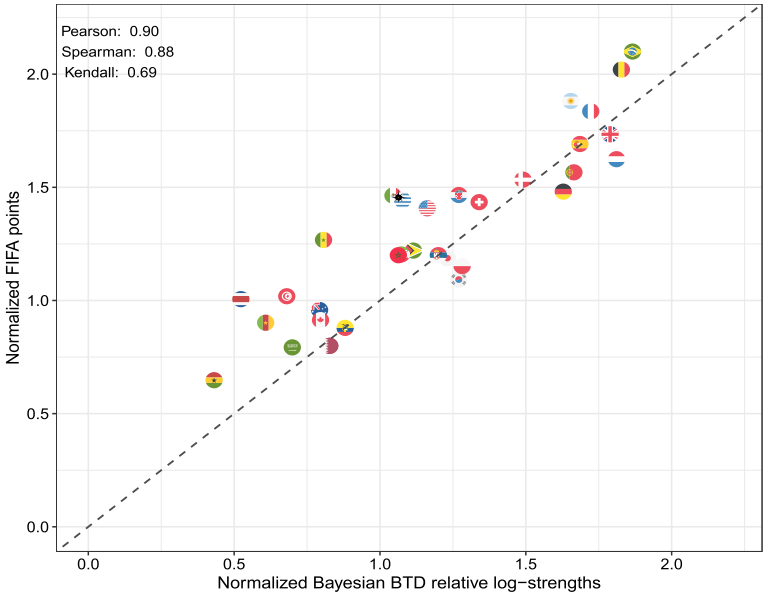

Fig. 1 2022 FIFA World Cup. Scatterplot comparing the FIFA ranking points and the BTD relative log-strengths of teams in the World Cup

higher relative strength than the “home" team, while points below the line indicate the opposite - we note that the terms ‘home’ and ‘away’ do not mean anything relevant in an international competition, where there are just one or two hosting teams. As expected, both the ranking systems reveal greater variability in team relative strengths during the group stage compared to the knockout stage. This is because teams in the group stage are typically more heterogeneous in terms of strength. As the tournament advances to the knockout stage, the remaining teams become increasingly similar in ability, leading to less variation in relative strength. This pattern is even more evident in the bottom panel of Fig. 2, which displays boxplots of relative strength differences between competing teams for each ranking system across the two stages. The boxplots further reveal that the Bayesian BTD ranking system exhibits more variability than the FIFA ranking during the group stage. In contrast, during the knockout stage, relative strength differences variability under both ranking systems are more concentrated, indicating a closer similarity in team abilities.

The main appeal of these models lies in their ability to predict the outcome of football matches. Table 1 illustrates the predictive accuracy of dynamic Poisson models and machine learning algorithms, measured by the Brier Score, across both group and knockout stages. The goal-based statistical models perform slightly better than the result-based methods during the group stage. In contrast, the machine learning approaches show better predictive accuracy in the knockout stage, where results may be

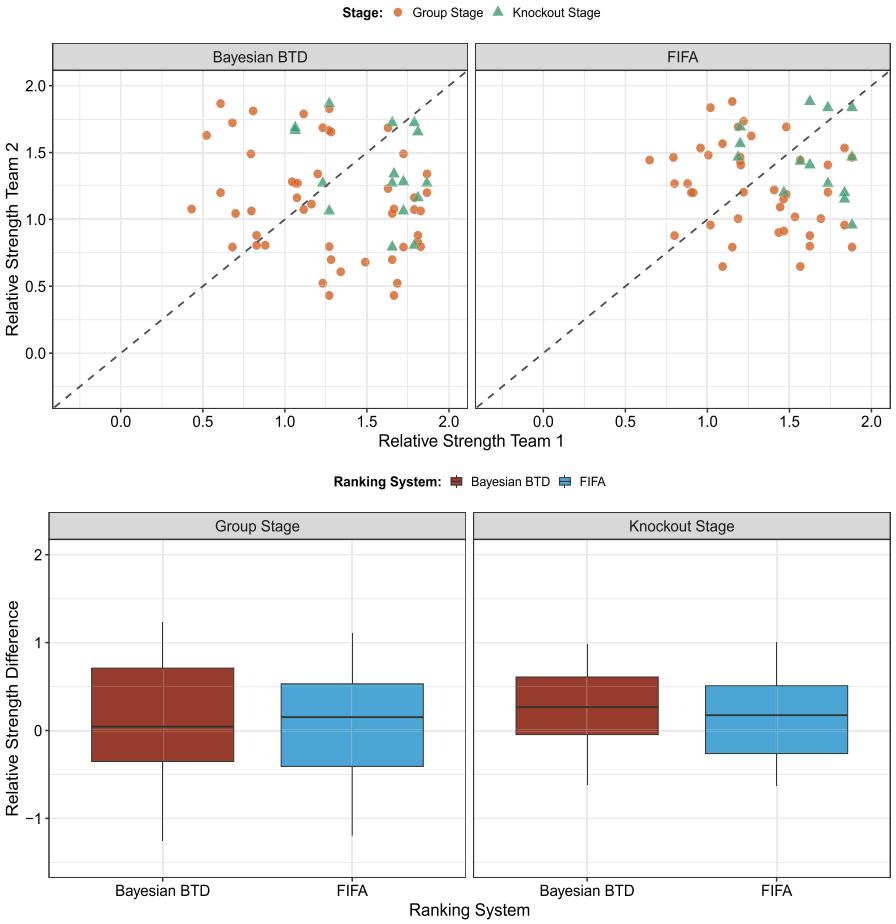

Fig. 2 2022 FIFA World Cup. The top panel displays scatterplots comparing the normalized relative strengths of teams in both the group stage (orange dots) and knockout stage (green dots) under the two ranking systems. The dashed grey line represents the bisector. The bottom panel presents boxplots of the normalized relative strength differences for the FIFA ranking (blue) and the Bayesian BTD derived ranking (red) across the two World Cup stages

Table 1 2022 FIFA World Cup. Brier score for the FIFA ranking and the Bayesian BTD derived ranking across the two World Cup stages

Model	Group stage		Knockout stage	
	FIFA	BTD	FIFA	BTD
Diag. Infl	0.620	0.629	0.530	0.510
Biv. Pois	0.617	0.618	0.546	0.535
Double Pois	0.622	0.623	0.543	0.527
MARS	0.640	0.660	0.486	0.503
ANN	0.627	0.660	0.465	0.471
Random Forest	0.713	0.745	0.493	0.461

less predictable. A similar trend is seen with ranking systems. While the FIFA ranking system shows marginally better predictive accuracy in the group stage, the Bayesian BTD ranking demonstrates overall better performance during the knockout stage. This suggests that the Bayesian BTD relative log-strengths are particularly apt at predicting outcomes when teams have comparable abilities, which is often the case in the latter stages of the competition.

4.2 2023 CAF Africa Cup of Nations

In this section, we fit the considered statistical models and machine learning algorithms to the data from the most recent CAF Africa Cup of Nations (AFCON) tournament held in Ivory Coast. Notably, the AFCON competition differs from the FIFA World Cup in that the participating teams tend to be more similar in terms of overall strength even during the group stage, making it a compelling case for analyzing the effectiveness of the Bayesian BTD ranking system. In order to conduct this analysis, we use as training set the data from matches played throughout 2018 to the end of 2023. Furthermore, the Bayesian BTD model was executed within this same period to generate team relative log-strengths. In addition, the FIFA ranking employed corresponds to those published on December 21st, available at https://inside.fifa.com/fifa-world-ranking/men?dateId=id14233.

Figure 3 presents a scatterplot for the 2023 CAF Africa Cup of Nations, showing the relationship between normalized Bayesian BTD relative log-strengths and normalized FIFA points. The results are consistent with those from the 2022 World Cup, showing a Pearson correlation coefficient of $\rho_p = 0.91$, a Spearman correlation coefficient of $\rho_s = 0.89$, and a Kendall coefficient of $\tau = 0.74$, all indicating strong positive associations.

The Bayesian BTD relative log-strengths exhibit less variability compared to the FIFA ranking points, in both the group and knockout stages, as illustrated in the top panel of Fig. 4. Furthermore, the boxplots displayed in the bottom panel of Fig. 4 show a significant reduction in variance for both ranking systems as we move from the group stage to the knockout stage. However, contrary to what was observed in the World Cup, the Bayesian BTD relative log-strengths exhibit lower variance compared to the FIFA ranking points in the group stage. This may indicate that the Bayesian BTD relative log-strengths more accurately reflect the inherent similarity in team strengths within this tournament stage.

As reported in Table 2, all the machine learning algorithms present similar prediction performance in the group stage of the AFCON. However, random forests show the weakest performance in the knockout stage. Furthermore, the other machine learning algorithms and statistical models exhibited similar levels of predictive accuracy in both stages, with statistical models performing slightly better in the group stage. In particular, the inclusion of this alternative ranking system generally improved the predictive performance of most of the statistical models evaluated in the knockout stage.

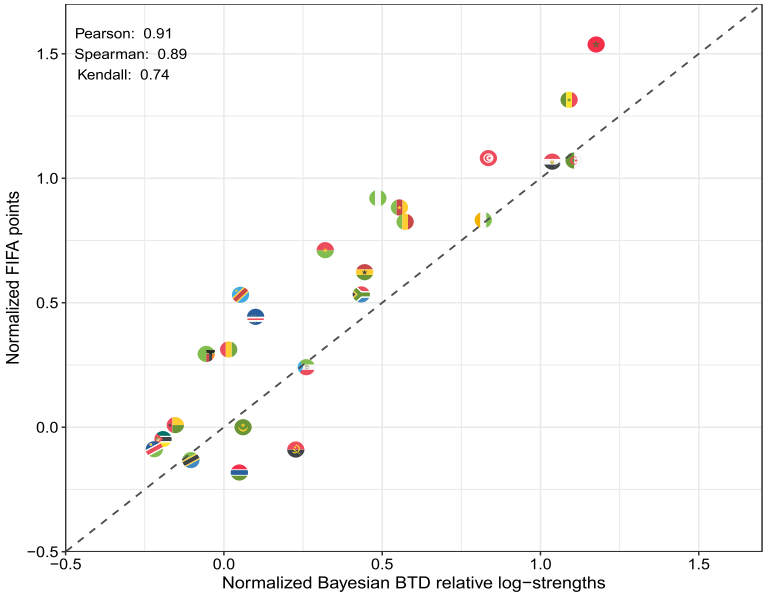

Fig. 3 2023 CAF Africa Cup of Nations. Scatterplot comparing FIFA ranking points and BTD relative log-strengths of teams in the Africa Cup of Nations

5 Discussion

This paper investigates the potential improvement in the predictive performance of statistical goal-based methods and machine learning result-based algorithms when a ranking system is incorporated as an additional predictor through its ranking points (relative strengths). We analyze data from the recent 2022 FIFA World Cup in Qatar and the 2023 CAF Africa Cup of Nations in Ivory Coast. Specifically, we explore the effectiveness of a Bayesian Bradley-Terry-Davidson derived ranking system in enhancing prediction accuracy compared to the well-established FIFA ranking system. We compare the performance of these two ranking systems across different tournament stages to identify their potential in predicting match outcomes.

While both the FIFA ranking points and the Bayesian BTD relative log-strengths provide valuable information for predicting outcomes, their effectiveness depends on the stage of the tournament. The FIFA ranking tends to be more accurate during the group stages of the 2022 World Cup, which features more heterogeneous team strengths. Conversely, the Bayesian BTD derived ranking is particularly effective in the group stage of the 2023 AFCON, and it also presents slightly better predictive performances in the knockout stages of both the World Cup and the AFCON. These stages typically exhibit smaller differences in team strengths. Consequently, this result suggests that the Bayesian BTD model effectively captures shifts in team strengths, making it especially valuable in tournaments where competing teams are similar, such as the AFCON, or in stages where the teams exhibit comparable strengths.

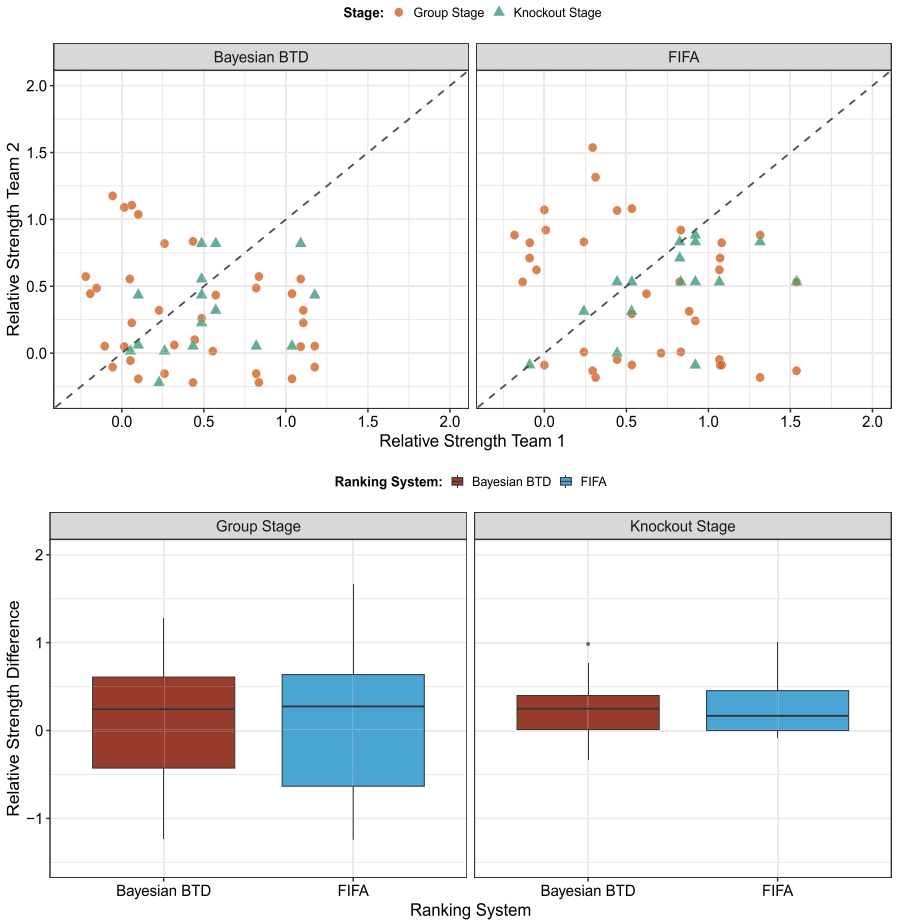

Fig. 4 2023 CAF Africa Cup of Nations. The top panel displays scatterplots comparing the normalized relative strengths of teams in both the group stage (orange dots) and knockout stage (green dots) under the two ranking systems. The dashed grey line represents the bisector. The bottom panel presents boxplots of the normalized relative strength differences for the FIFA ranking (blue) and the Bayesian BTD derived ranking (red) across the two Africa Cup of Nations stages

Table 2 2023 CAF Africa Cup of Nations. Brier score for the FIFA ranking and the Bayesian BTD derived ranking across the two Africa Cup of Nations stages

Model	Group stage		Knockout stage	
	FIFA	BTD	FIFA	BTD
Diag. Infl	0.679	0.682	0.681	0.677
Biv. Pois	0.673	0.682	0.658	0.660
Double Pois	0.670	0.677	0.670	0.656
MARS	0.679	0.690	0.645	0.650
ANN	0.703	0.702	0.666	0.661
Random Forest	0.736	0.687	0.834	0.884

It is important to note that while the Bayesian BTD derived ranking is a valuable alternative, it is more computationally intensive compared to the standard FIFA ranking. Specifically, this approach involves a two-step procedure. First, the Bayesian BTD model needs to be computed to derive the teams' relative log-strengths, and only then their difference can be used as additional predictor in the models being considered.

However, the potential to enhance the accuracy and predictive performances of these models remains significant. Further development could involve refining the Bayesian BTD model to include additional variables that impact match outcomes. These could include the overall market value of the players involved in a specific team, the number of Champions League players, an indicator of the hosting country, or the teams in its neighborhood. Even economic variables such as the GDP per capita or the national population size may be interesting. Furthermore, the implementation of a dynamic methodology, that enables the continuous adjustment for fluctuations in team strength over the course of a season or tournament, could potentially result in enhanced prediction accuracy.

The potential application of Bayesian Bradley-Terry derived rankings, as an alternative for the FIFA ranking or similar systems, represents a promising area for research. Further comparative studies across different sports or competition structures will be conducted to validate the effectiveness of the Bradley-Terry models. As broadly remarked, the interplay between the type of competition, the adopted ranking, and the chosen methodology represents a hot topic for football modellers and deserves a deep and further understanding, both in international matches and domestic leagues.

6 Software and data availability

All analyses were conducted in the R programming language version 4.2.3 (R Core Team 2023). The code to reproduce this manuscript is openly available at https://github.com/RoMaD-96/Bayesian_BTD. The data are available on Kaggle at https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017.

Acknowledgements This work has been supported by the project "SMARTsports: "Statistical Models and AlgoRiThms in sports. Applications in professional and amateur contexts, with able-bodied and disabled athletes", funded by the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) Bando 2022 - grant n. 2022R74PLE (CUP J53D23003860006).

References

Abreu PH, Silva DC, Mendes-Moreira J, Reis LP, Garganta J (2013) Using multivariate adaptive regression splines in the construction of simulated soccer team's behavior models. Int J Comput Intell Syst 6:893–910. https://doi.org/10.1080/18756891.2013.808426

Baio G, Blangiardo M (2010) Bayesian hierarchical model for the prediction of football results. J Appl Stat 37(2):253–264. https://doi.org/10.1080/02664760802684177

Beaver RJ, Gokhale DV (1975) A model to incorporat within-pair order effects in paired comparisons. Commun Stat 4(10):923–939. https://doi.org/10.1080/03610927308827302
Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324
Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthey Weather Rev 78(1):1–3
Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39(3/4):324–345
Cattelan M (2012) Models for paired comparison data: A review with emphasis on dependent data. Statistical Science 27(3):412–433. https://doi.org/10.1214/12-STS396
Carpita M, Ciavolino E, Pasca P (2019) Exploring and modelling team performances of the Kaggle European soccer database. Statistical Modelling 19(1):74–101. https://doi.org/10.1177/1471082X18810971
Caron F, Doucet A (2012) Efficient Bayesian inference for generalized Bradley-Terry models. J Comput Graph Stat 21(1):174–196. https://doi.org/10.1080/10618600.2012.638220
Chen C, Smith TM (1984) A Bayes-type estimator for the Bradley-Terry model for paired comparison. J Stat Plann Inf 10(1):9–14. https://doi.org/10.1016/0378-3758(84)90028-4
Cattelan M, Varin C, Firth D (2012) Dynamic Bradley-Terry modelling of sports tournaments. J Royal Stat Soc Series C: Appl Stat 62(1):135–150. https://doi.org/10.1111/j.1467-9876.2012.01046.x
Davidson RR (1970) On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. J Am Stat Association 65(329):317–328
Davidson RR, Beaver RJ (1977) On extending the Bradley-Terry model to incorporate within-pair order effects. Biometrics 33(4):693–702
Dixon MJ, Coles SG (1997) Modelling association football scores and inefficiencies in the football betting market. J Royal Stat Soc: Serie C (Appl Stat) 46(2):265–280. https://doi.org/10.1111/1467-9876.00065
Danisik N, Lacko P, Farkas M (2018) Football match prediction using players attributes, pp. 201–206. https://doi.org/10.1109/DISA.2018.8490613
Davidson RR, Solomon DL (1973) A Bayesian approach to paired comparison experimentation. Biometrika 60(3):477–487
Egidi L, Palaskas, V (2022) footBayes: Fitting Bayesian and MLE Football Models. R package version 0.2.0. https://github.com/leoegidi/footbayes
Egidi L, Pauli F, Torelli N (2018) Combining historical data and bookmakers' odds in modelling football scores. Stat Model 18(5–6):436–459. https://doi.org/10.1177/1471082X18798414
Egidi L, Torelli N (2021) Comparing goal-based and result-based approaches in modelling football outcomes. Soc Ind Res 156(2):801–813
Ford LR (1957) Solution of a ranking problem from binary comparisons. Am Math Monthly 64:28–33
Friedman JH (1991) Multivariate adaptive regression splines. Annals Stat 19(1):1–67. https://doi.org/10.1214/aos/1176347963
Fahrmeir L, Tutz G (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems. J Am Stat Association 89(428):1438–1449 (**Accessed 2024-04-20**)
Groll A, Abedieh J (2013) Spain retains its title and sets a new record - generalized linear mixed models on European football championships. J Quantitative Anal Sports 9(1):51–66. https://doi.org/10.1515/jqas-2012-0046
Groll A, Cristophe L, Hans VE, Gunther S (2019) A hybrid random forest to predict soccer matches in international tournaments. J Quantitative Anal Sports 15(4):271–287. https://doi.org/10.1515/jqas-2018-0060
Groll A, Hvattum LM, Ley C, Popp F, Schauberger G, Van Eetvelde H, Zeileis A (2021) Hybrid machine learning forecasts for the UEFA EURO 2020. arXiv preprint arXiv:2106.05799
Glickman ME (1999) Parameter estimation in large dynamic paired comparison experiments. J Royal Stat Soc SerieC (Appl Stat) 48(3):377–394
Glickman ME (2001) Dynamic paired comparison models with stochastic variances. J Appl Stat 28(6):673–689. https://doi.org/10.1080/02664760120059219
Huang K-Y, Chang W-L (2010) Neural network method for prediction of 2006 World Cup football game. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. https://doi.org/10.1109/IJCNN.2010.5596458
Hucaljuk J, Rakipovic A (2011) Predicting football scores using machine learning techniques. In: 2011 Proceedings of the 34th International Convention MIPRO, 1623–1627
Hunter DR (2004) MM algorithms for generalized Bradley-Terry models. Annals Stat 32(1):384–406

Issa Mattos D, Martins Silva Ramos E (2020) bpc: A package for Bayesian paired comparison analysis
Issa Mattos D, Martins Silva Ramos É (2022) Bayesian paired comparison with the bpcs package. Behav Res Meth 54(4):2025–2045. https://doi.org/10.3758/s13428-021-01714-2
Koopman SJ, Lit R (2015) A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. J Royal Stat Soc Serie A (Stat Soc) 178(1):167–186
Karlis D, Ntzoufras I (2003) Analysis of sports data by using bivariate Poisson models. J Royal Stat Soc: Serie D (Stat) 52(3):381–393. https://doi.org/10.1111/1467-9884.00366
Karlis D, Ntzoufras I (2009) Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. IMA J Manag Math 20(2):133–145. https://doi.org/10.1093/imaman/dpn026
Koning RH (2000) Balance in competition in Dutch soccer. J Royal Stat Soc: Serie D (Stat) 49(3):419–431. https://doi.org/10.1111/1467-9884.00244
Kuhn M (2022) Caret: Classification and Regression Training. R package version 6.0-93. https://CRAN.R-project.org/package=caret
Leonard T (1977) An alternative Bayesian approach to the Bradley-Terry model for paired comparisons. Biometrics 33(1):121–132
Luce RD (1959) Individual Choice Behavior: A Theoretical Analysis. Wiley, ???. https://books.google.it/books?id=c519AAAAMAAJ
Maher MJ (1982) Modelling association football scores. Stat Neerlandica 36(3):109–118. https://doi.org/10.1111/j.1467-9574.1982.tb00782.x
Ntzoufras I (2011) Bayesian Modeling Using WinBUGS vol. 698. John Wiley & Sons, Hoboken, New Jersey, USA
Osei PP, Davidov O (2022) Bayesian linear models for cardinal paired comparison data. Comput Stat Data Anal 172:107481. https://doi.org/10.1016/j.csda.2022.107481
Owen A (2011) Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. IMA J Manag Math 22(2):99–113. https://doi.org/10.1093/imaman/dpq018
R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2023). R Foundation for Statistical Computing. https://www.R-project.org/
Rao PV, Kupper LL (1967) Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. J Am Stat Association 62(317):194–204. https://doi.org/10.1080/01621459.1967.10482901
Rue H (2000) Salvesen: Prediction and retrospective analysis of soccer matches in a league. J Royal Stat Soc Serie D (Stat) 49(3):399–418
Schauberger G, Groll A (2018) Predicting matches in international football tournaments with random forests. Stat Model 18(5–6):460–482. https://doi.org/10.1177/1471082X18799934
Spiegelhalter D, Ng Y-L (2009) One match to go! Significance 6(4):151–153
Springall A (1973) Response surface fitting using a generalization of the Bradley-Terry paired comparison model. J Royal Stat Soc Serie C: Appl Stat 22(1):59–68. https://doi.org/10.2307/2346303
Szczecinski L, Roatis I-I (2022) FIFA ranking: Evaluation and path forward. J Sports Anal 8(4):231–250
Tian X-Y, Shi J, Shen X, Song K A spectral approach for the dynamic Bradley-Terry model. arXiv preprint arXiv:2307.16642 (2023)
Wainer J (2023) A Bayesian Bradley-Terry model to compare multiple ml algorithms on multiple data sets. J Mach Learn Res 24(341):1–34
Whelan JT (2017) Prior distributions for the bradley-terry model of paired comparisons. arXiv preprint arXiv:1712.05311

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com