



Modeling outcomes of soccer matches

Alkeos Tsokos¹ · Santhosh Narayanan² · Ioannis Kosmidis^{2,3} · Gianluca Baio¹ · Mihai Cucuringu^{3,4} · Gavin Whitaker¹ · Franz Király^{1,3}

Received: 10 July 2017 / Accepted: 25 June 2018 / Published online: 1 August 2018
© The Author(s) 2018

Abstract

We compare various extensions of the Bradley–Terry model and a hierarchical Poisson log-linear model in terms of their performance in predicting the outcome of soccer matches (win, draw, or loss). The parameters of the Bradley–Terry extensions are estimated by maximizing the log-likelihood, or an appropriately penalized version of it, while the posterior densities of the parameters of the hierarchical Poisson log-linear model are approximated using integrated nested Laplace approximations. The prediction performance of the various modeling approaches is assessed using a novel, context-specific framework for temporal validation that is found to deliver accurate estimates of the test error. The direct modeling of outcomes via the various Bradley–Terry extensions and the modeling of match scores using the hierarchical Poisson log-linear model demonstrate similar behavior in terms of predictive performance.

Keywords Bradley–Terry model · Poisson log-linear hierarchical model · Maximum penalized likelihood · Integrated nested laplace approximation · Temporal validation

1 Introduction

The current paper stems from our participation in the 2017 Machine Learning Journal (Springer) challenge on predicting outcomes of soccer matches from a range of leagues around the world (MLS challenge, in short). Details of the challenge and the data can be found in Berrar et al. (2017).

Editors: Philippe Lopes, Werner Dubitzky, Daniel Berrar, and Jesse Davis.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10994-018-5741-1>) contains supplementary material, which is available to authorized users.

✉ Alkeos Tsokos
alkeos.tsokos.10@ucl.ac.uk

¹ University College London, London, UK

² University of Warwick, Coventry, UK

³ The Alan Turing Institute, London, UK

⁴ University of Oxford, Oxford, UK

We consider two distinct modeling approaches for the task. The first approach focuses on modeling the probabilities of win, draw, or loss, using various extensions of Bradley–Terry models (Bradley and Terry 1952). The second approach focuses on directly modeling the number of goals scored by each team in each match using a hierarchical Poisson log-linear model, building on the modeling frameworks in Maher (1982), Dixon and Coles (1997), Karlis and Ntzoufras (2003) and Baio and Blangiardo (2010).

The performance of the various modeling approaches in predicting the outcomes of matches is assessed using a novel, context-specific framework for temporal validation that is found to deliver accurate estimates of the prediction error. The direct modeling of the outcomes using the various Bradley–Terry extensions and the modeling of match scores using the hierarchical Poisson log-linear model deliver similar performance in terms of predicting the outcome.

The paper is structured as follows: Sect. 2 briefly introduces the data, presents the necessary data-cleaning operations undertaken, and describes the various features that were extracted. Section 3 presents the various Bradley–Terry models and extensions we consider for the challenge and describes the associated estimation procedures. Section 4 focuses on the hierarchical Poisson log-linear model and the Integrated Nested Laplace Approximations (INLA; Rue et al. 2009) of the posterior densities for the model parameters. Section 5 introduces the validation framework and the models are compared in terms of their predictive performance in Sect. 6. Section 7 concludes with discussion and future directions.

2 Pre-processing and feature extraction

2.1 Data exploration

The data contain matches from 52 leagues, covering 35 countries, for a varying number of seasons for each league. Nearly all leagues have data since 2008, with a few having data extending as far back as 2000. There are no cross-country leagues (e.g. UEFA Champions League) or teams associated with different countries. The only way that teams move between leagues is within each country by either promotion or relegation.

Figure 1 shows the number of available matches for each country in the data set. England dominates the data in terms of matches recorded, with the available matches coming from 5 distinct leagues. The other highly-represented countries are Scotland with data from 4 distinct leagues, and European countries, such as Spain, Germany, Italy and France, most probably because they also have a high UEFA coefficient (Wikipedia 2018).

Figure 2 shows the number of matches per number of goals scored by the home (dark grey) and away (light grey) teams. Home teams appear to score more goals than away teams, with home teams having consistently higher frequencies for two or more goals and away teams having higher frequencies for no goal and one goal. Overall, home teams scored 304,918 goals over the whole data set, whereas away teams scored 228,293 goals. In Section 1 of the Supplementary Material, the trend shown in Fig. 2 is also found to be present within each country, pointing towards the existence of a home advantage.

2.2 Data cleaning

Upon closer inspection of the original sequence of matches for the MLS challenge, we found and corrected the following three anomalies in the data. The complete set of matches from

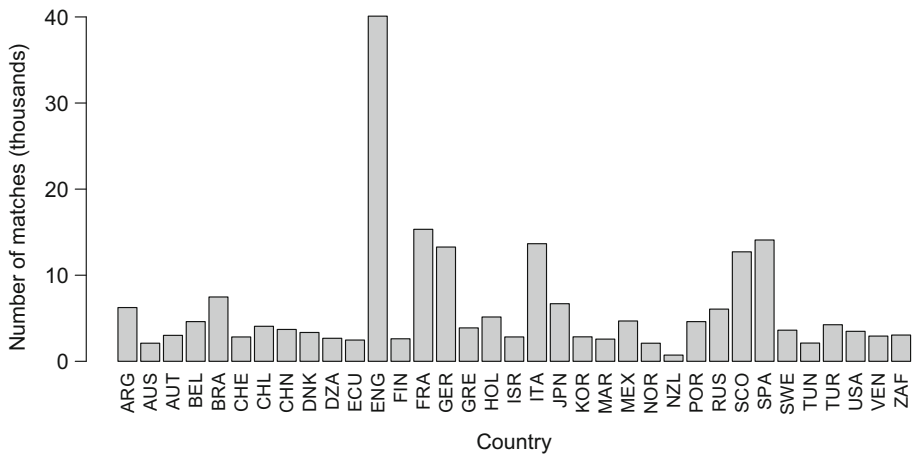


Fig. 1 Number of available matches per country in the data

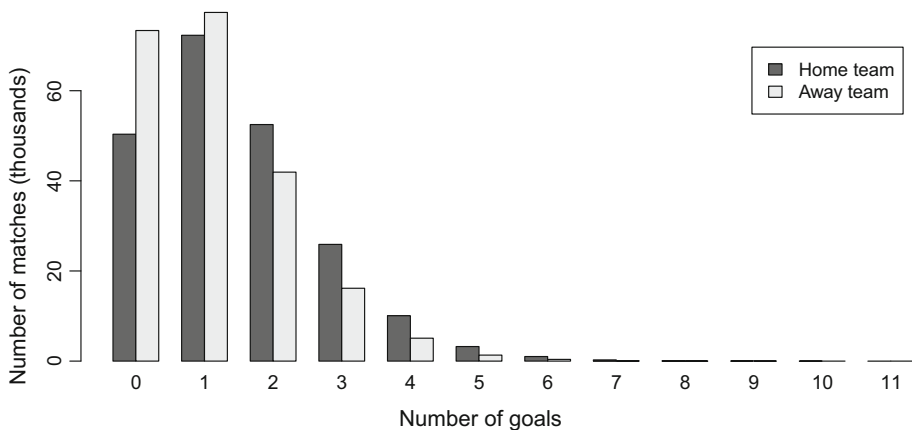


Fig. 2 The number of matches per number of goals scored by the home (dark grey) and away team (light grey)

the 2015–2016 season of the Venezuelan league was duplicated in the data. We kept only one instance of these matches. Furthermore, 26 matches from the 2013–2014 season of the Norwegian league were assigned the year 2014 in the date field instead of 2013. The dates for these matches were modified accordingly. Finally, one match in the 2013–2014 season of the Moroccan league (Raja Casablanca vs Maghrib de Fes) was assigned the month February in the date field instead of August. The date for this match was corrected, accordingly.

2.3 Feature extraction

The features that were extracted can be categorized into team-specific, match-specific and/or season-specific. Match-specific features were derived from the information available on each match. Season-specific features have the same value for all matches and teams in a season of a particular league, and differ only across seasons for the same league and across leagues.

Table 1 gives short names, descriptions, and ranges for the features that were extracted. Table 2 gives an example of what values the features take for an artificial data set with observations on the first 3 matches of a season for team A playing all matches at home. The team-specific features are listed only for Team A to allow for easy tracking of their evolution.

The features we extracted are proxies for a range of aspects of the game, and their choice was based on common sense and our understanding of what is important in soccer, and previous literature. Home (feature 1 in Table 2) can be used for including a home advantage in the models; newly promoted (feature 2 in Table 2) is used to account for the fact that a newly promoted team is typically weaker than the competition; days since previous match (feature 3 in Table 2) carries information regarding fatigue of the players and the team, overall; form (feature 4 in Table 2) is a proxy for whether a team is doing better or worse during a particular period in time compared to its general strength; matches played (feature 5 in Table 2) determines how far into the season a game occurs; points tally, goal difference, and points per match (features 6, 7 and 10 in Table 2) are measures of how well a team is doing so far in the season; goals scored per match and goals conceded per match (features 8 and 9 in Table 2) are measures of a team's attacking and defensive ability, respectively; previous season points tally and previous season goal difference (features 11 and 12 in Table 2) are measures of how well a team performed in the previous season, which can be a useful indicator of how well a team will perform in the early stages of a season when other features such as points tally do not carry much information; finally, team rankings (feature 13 in Table 2) refers to a variety of measures that rank teams based on their performance in previous matches, as detailed in Section 2 of the Supplementary Material.

In order to avoid missing data in the features we extracted, we made the following conventions. The value of form for the first match of the season for each team was drawn from a Uniform distribution in (0, 1). The form for the second and third match were a third of the points in the first match, and a sixth of the total points in the first two matches, respectively. Days since previous match was left unspecified for the very first match of the team in the data. If the team was playing its first season then we treated it as being newly promoted. The previous season points tally was set to 15 for newly promoted teams and to 65 for newly relegated teams, and the previous season goal difference was set to -35 for newly promoted teams and 35 for newly relegated teams. These values were set in an ad-hoc manner prior to estimation and validation, based on our sense and experience of what is a small or large value for the corresponding features. In principle, the choice of these values could be made more formally by minimizing a criterion of predictive quality, but we did not pursue this as it would complicate the estimation-prediction workflow described later in the paper and increase computational effort significantly without any guarantee of improving the predictive quality of the models.

3 Modeling outcomes

3.1 Bradley–Terry models and extensions

The Bradley–Terry model (Bradley and Terry 1952) is commonly used to model paired comparisons, which often arise in competitive sport. For a binary win/loss outcome, let

$$y_{ijt} = \begin{cases} 1, & \text{if team } i \text{ beats team } j \text{ at time } t \\ 0, & \text{if team } j \text{ beats team } i \text{ at time } t \end{cases} \quad (i, j = 1, \dots, n; \quad i \neq j; \quad t \in \mathbb{R}^+),$$

Table 1 Short names, descriptions, and ranges for the features that were extracted

Number	Short name	Description	Range
<i>Team-specific features</i>			
1	Home	1 if the team is playing at home, and 0 otherwise	{0, 1}
2	Newly promoted	1 if the team is newly promoted to the league for the current season, and 0 otherwise	{0, 1}
3	Days since previous match	Number of days elapsed since the previous match of the team	{1, 2, ...}
4	Form	A ninth of the total points gained in the last three matches in the current season	(0, 1)
5	Matches played	Number of matches played in the current season and before the current match	{1, 2, ...}
6	Points tally	The points accumulated during the current season and before the current match	{0, 1, ...}
7	Goal difference	The goals that a team has scored minus the goals that it has conceded over the current season and before the current match	{..., -1, 0, 1, ...}
8	Goals scored per match	Total goals scored per match played over the current season and before the current match	\mathbb{R}^+
9	Goals conceded per match	Total goals conceded per match over the current season and before the current match	\mathbb{R}^+
10	Points per match	Total points gained per match played over the current season and before the current match	[0, 3]
11	Previous season points tally	Total points accumulated by the team in the previous season of the same league	{0, 1, ...}
12	Previous season goal difference	Total goals scored minus total goals conceded for each team in the previous season of the same league	{..., -1, 0, 1, ...}
13	Team rankings	A variety of team rankings, based on historical observations; See Section 2 of the Supplementary Material	\mathbb{R}
<i>Season-specific features</i>			
14	Season	The league season in which each match is played	Labels
15	Season window	Time period in calendar months of the league season	labels
<i>Match-specific features</i>			
16	Quarter	Quarter of the calendar year based on the match date	labels

Table 2 Feature values for artificial data showing the first 3 matches of a season with team A playing all matches at home

	Match 1	Match 2	Match 3
<i>Match attributes and outcomes</i>			
League	Country1	Country1	Country1
Date	2033-08-18	2033-08-21	2033-08-26
Home team	Team A	Team A	Team A
Away team	Team B	Team C	Team D
Home score	2	2	0
Away score	0	1	0
<i>Team-specific features (Team A)</i>			
Newly promoted	0	0	0
Days since previous match	91	3	5
Form	0.5233	1	1
Matches played	0	1	2
Points tally	0	3	6
Goal difference	–	2	3
Goals scored per match	0	2	2
Goals conceded per match	0	0	0.5
Points per match	0	3	3
Previous season points tally	72	72	72
Previous season goal difference	45	45	45
<i>Season-specific features</i>			
Season	33–34	33–34	33–34
Season window	August–May	August–May	August–May
<i>Match-specific features</i>			
Quarter	3	3	3

where n is the number of teams present in the data. The Bradley–Terry model assumes that

$$p(y_{ijt} = 1) = \frac{\pi_i}{\pi_i + \pi_j},$$

where $\pi_i = \exp(\lambda_i)$, and λ_i is understood as the “strength” of team i . In the original Bradley–Terry formulation, λ_i does not vary with time.

For the purposes of the MLS challenge prediction task, we consider extensions of the original Bradley–Terry formulation where we allow λ_i to depend on a p -vector of time-dependent features \mathbf{x}_{it} for team i at time t as $\lambda_{it} = f(\mathbf{x}_{it})$ for some function $f(\cdot)$. Bradley–Terry models can also be equivalently written as linking the log-odds of a team winning to the difference in strength of the two teams competing. Some of the extensions below directly specify that difference.

3.1.1 BL: baseline

The simplest specification of all assumes that

$$\lambda_{it} = \beta h_{it}, \quad (1)$$

where $h_{it} = 1$ if team i is playing at home at time t , and $h_{it} = 0$ otherwise. The only parameter to estimate with this specification is β , which can be understood as the difference in strength when the team plays at home. We use this model to establish a baseline to improve upon for the prediction task.

3.1.2 CS: constant strengths

This specification corresponds to the standard Bradley–Terry model with a home-field advantage, under which

$$\lambda_{it} = \alpha_i + \beta h_{it}. \quad (2)$$

The above specification involves $n + 1$ parameters, where n is the number of teams. The parameter α_i represents the time-invariant strength of the i th team.

3.1.3 LF: linear with features

Suppose now that we are given a vector of features \mathbf{x}_{it} associated with team i at time t . A simple way to model the team strengths λ_{it} is to assume that they are a linear combination of the features. Hence, in this model we have

$$\lambda_{it} = \sum_{k=1}^p \beta_k x_{itk}, \quad (3)$$

where x_{itk} is the k th element of the feature vector \mathbf{x}_{it} .

Note that the coefficients in the linear combination are shared between all teams, and so the number of parameters to estimate is p , where p is the dimension of the feature vector. This specification is similar to the one implemented in the R package `BradleyTerry` (Firth 2005), but without the team specific random effects.

3.1.4 TVC: time-varying coefficients

Some of the features we consider, like points tally season (feature 6 in Table 1) vary during the season. Ignoring any special circumstances such as teams being punished, the points accumulated by a team is a non-decreasing function of the number of matches the team has played.

It is natural to assume that the contribution of points accumulated to the strength of a team is different at the beginning of the season than it is at the end. In order to account for such effects, the parameters for the corresponding features can be allowed to vary with the matches played. Specifically, the team strengths can be modeled as

$$\lambda_{it} = \sum_{k \in \mathcal{V}} \gamma_k(m_{it}) x_{itk} + \sum_{k \notin \mathcal{V}} \beta_k x_{itk}, \quad (4)$$

where m_{it} denotes the number of matches that team i has played within the current season at time t and \mathcal{V} denotes the set of coefficients that are allowed to vary with the matches played. The functions $\gamma_k(m_{it})$ can be modeled non-parametrically, but in the spirit of keeping the complexity low we instead set $\gamma_k(m_{it}) = \alpha_k + \beta_k m_{it}$. With this specification for $\gamma_k(m_{it})$, TVC is equivalent to LF with the inclusion of an extra set of features $\{m_{it} x_{itk}\}_{k \in \mathcal{V}}$.

3.1.5 AFD: additive feature differences with time interactions

For the LF specification, the log-odds of team i beating team j is

$$\lambda_{it} - \lambda_{jt} = \sum_{k=1}^p \beta_k (x_{itk} - x_{jtk}).$$

Hence, the LF specification assumes that the difference in strength between the two teams is a linear combination of differences between the features of the teams. We can relax the assumption of linearity, and include non-linear time interactions, by instead assuming that each difference in features contributes to the difference in strengths through an arbitrary bivariate smooth function g_k that depends on the feature difference and the number of matches played. We then arrive at the AFD specification, which can be written as

$$\lambda_{it} - \lambda_{jt} = \sum_{k \in \mathcal{V}} g_k(x_{itk} - x_{jtk}, m_{it}) + \sum_{k \notin \mathcal{V}} f_k(x_{itk} - x_{jtk}), \quad (5)$$

where for simplicity we take the number of matches played to be the number of matches played by the home team.

3.2 Handling draws

The extra outcome of a draw in a soccer match can be accommodated within the Bradley–Terry formulation in two ways.

The first is to treat win, loss and draw as multinomial ordered outcomes, in effect assuming that “win” \succ “draw” \succ “loss”, where \succ denotes strong transitive preference. Then, the ordered outcomes can be modeled using cumulative link models (Agresti 2015) with the various strength specifications. Specifically, let

$$y_{ijt} = \begin{cases} 2, & \text{if team } i \text{ beats team } j \text{ at time } t, \\ 1, & \text{if team } i \text{ and } j \text{ draw at time } t, \\ 0, & \text{if team } j \text{ beats team } i \text{ at time } t. \end{cases}$$

and assume that y_{ijt} has

$$p(y_{ijt} \leq y) = \frac{e^{\delta_y + \lambda_{it}}}{e^{\delta_y + \lambda_{it}} + e^{\lambda_{jt}}}, \quad (6)$$

where $-\infty < \delta_0 \leq \delta_1 < \delta_2 = \infty$, and δ_0, δ_1 are parameters to be estimated from the data. Cattelan et al. (2013) and Király and Qian (2017) use this approach for modeling soccer outcomes.

Another possibility for handling draws is to use the Davidson (1970) extension of the Bradley–Terry model, under which

$$\begin{aligned} p(y_{ijt} = 2 | y_{ijt} \neq 1) &= \frac{\pi_{it}}{\pi_{it} + \pi_{jt}}, \\ p(y_{ijt} = 1) &= \frac{\delta \sqrt{\pi_{it}\pi_{jt}}}{\pi_{it} + \pi_{jt} + \delta \sqrt{\pi_{it}\pi_{jt}}}, \\ p(y_{ijt} = 0 | y_{ijt} \neq 1) &= \frac{\pi_{jt}}{\pi_{it} + \pi_{jt}}. \end{aligned}$$

where δ is a parameter to be estimated from the data.

3.3 Estimation

3.3.1 Likelihood-based approaches

The parameters of the Bradley–Terry model extensions presented above can be estimated by maximizing the log-likelihood of the multinomial distribution.

The log-likelihood about the parameter vector θ is

$$\ell(\theta) = \sum_{\{i,j,t\} \in \mathcal{M}} \sum_y \mathbb{I}_{[y_{ijt}=y]} \log(p(y_{ijt} = y)),$$

where \mathbb{I}_A takes the value 1 if A holds and 0 otherwise, and \mathcal{M} is the set of triplets $\{i, j, t\}$ corresponding to the matches whose outcomes have been observed.

For estimating the functions involved in the AFD specification, we represent each f_k using thin plate splines (Wahba 1990, Section 2.4), and enforce smoothness constraints on the estimate of f_k by maximizing a penalized log-likelihood of the form

$$\ell^{\text{pen}}(\theta) = \ell(\theta) - \frac{1}{2} \sum_k d_k \theta^T P_k \theta$$

where P_k are penalty matrices and d_k are tuning parameters. For penalized estimation we only consider ordinal models through the R package `mgcv` (Wood 2006), and select d_k by optimizing the Generalized Cross Validation criterion (Golub et al. 1979). Details on the fitting procedure for specifications like AFD and the implementation of thin plate spline regression in `mgcv` can be found in Wood (2003).

The parameters of the Davidson extensions of the Bradley–Terry model are estimated by using the BFGS optimization algorithm (Byrd et al. 1995) to minimize $-\ell(\theta)$.

3.3.2 Identifiability

In the CS model, the team strengths are identifiable only up to an additive constant, because $\lambda_i - \lambda_j = (\lambda_i + d) - (\lambda_j + d)$ for any $d \in \mathbb{R}$. This unidentifiability can be dealt with by setting the strength of an arbitrarily chosen team to zero. The CS model was fitted league-by-league with one identifiability constraint per league.

The parameters δ_0 and δ_1 in (6) are identifiable only if the specification used for $\lambda_i - \lambda_j$ does not involve an intercept parameter. An alternative is to include an intercept parameter in $\lambda_i - \lambda_j$ and fix δ_0 at a value. The estimated probabilities are invariant to these alternatives, and we use the latter simply because this is the default in the `mgcv` package.

3.3.3 Other data-specific considerations

The parameters in the LF, TVC, and AFD specifications (which involve features) are shared across the leagues and matches in the data. For computational efficiency we restrict the fitting procedures to use the 20,000 most recent matches, or less if less is available, at the time of the first match that a prediction needs to be made. The CS specification requires estimating the strength parameters directly. For computational efficiency, we estimate the strength parameters independently for each league within each country, and only consider matches that took place in the past calendar year from the date of the first match that a prediction needs to be made.

4 Modeling scores

4.1 Model structure

Every league consists of a number of teams T , playing against each other twice in a season (once at home and once away). We indicate the number of goals scored by the home and the away team in the g th match of the season ($g = 1, \dots, G$) as y_{g1} and y_{g2} , respectively.

The observed goal counts y_{g1} and y_{g2} are assumed to be realizations of conditionally independent random variables Y_{g1} and Y_{g2} , respectively, with

$$Y_{gj} \mid \theta_{gj} \sim \text{Poisson}(\theta_{gj}) .$$

The parameters θ_{g1} and θ_{g2} represent the *scoring intensity* in the g th match for the home and away team, respectively.

We assume that θ_{g1} and θ_{g2} are specified through the regression structures

$$\begin{aligned} \eta_{g1} = \log(\theta_{g1}) &= \sum_{k=1}^p \beta_k z_{g1k} + \alpha_{h_g} + \xi_{a_g} + \gamma_{h_g, \text{Sea}_g} + \delta_{a_g, \text{Sea}_g}, \\ \eta_{g2} = \log(\theta_{g2}) &= \sum_{k=1}^p \beta_k z_{g2k} + \alpha_{a_g} + \xi_{h_g} + \gamma_{a_g, \text{Sea}_g} + \delta_{h_g, \text{Sea}_g}. \end{aligned} \quad (7)$$

The indices h_g and a_g determine the home and away team for match g respectively, with $h_g, a_g \in \{1, \dots, T\}$. The parameters β_1, \dots, β_p represent the effects corresponding to the observed match- and team-specific features z_{gj1}, \dots, z_{gjp} , respectively, collected in a $G \times 2p$ matrix \mathbf{Z} . The other effects in the linear predictor η_{gj} reflect assumptions of exchangeability across the teams involved in the matches. Specifically, α_t and ξ_t represent the latent attacking and defensive ability of team t and are assumed to be distributed as

$$\alpha_t \mid \sigma_\alpha \sim \text{Normal}(0, \sigma_\alpha^2) \quad \text{and} \quad \xi_t \mid \sigma_\xi \sim \text{Normal}(0, \sigma_\xi^2) .$$

We used vague log-Gamma priors on the precision parameters $\tau_\alpha = 1/\sigma_\alpha^2$ and $\tau_\xi = 1/\sigma_\xi^2$. In order to account for the time dynamics across the different seasons, we also include the latent interactions γ_{ts} and δ_{ts} between the team-specific attacking and defensive strengths and the season $s \in \{1, \dots, S\}$, which were modeled using autoregressive specifications with

$$\begin{aligned} \gamma_{t1} \mid \sigma_\epsilon, \rho_\gamma &\sim \text{Normal}\left(0, \sigma_\epsilon^2 \left(1 - \rho_\gamma^2\right)\right), \\ \gamma_{ts} &= \rho_\gamma \gamma_{t, s-1} + \epsilon_s, \\ \epsilon_s \mid \sigma_\epsilon &\sim \text{Normal}(0, \sigma_\epsilon^2) \quad (s = 2, \dots, S), \end{aligned}$$

and

$$\begin{aligned} \delta_{t1} \mid \sigma_\epsilon, \rho_\delta &\sim \text{Normal}\left(0, \sigma_\epsilon^2 (1 - \rho_\delta^2)\right), \\ \delta_{ts} &= \rho_\delta \delta_{t, s-1} + \epsilon_s, \\ \epsilon_s \mid \sigma_\epsilon &\sim \text{Normal}(0, \sigma_\epsilon^2) \quad (s = 2, \dots, S). \end{aligned}$$

For the specification of prior distributions for the hyperparameters $\rho_\gamma, \rho_\delta, \sigma_\epsilon$ we used the default settings of the R-INLA package (Lindgren and Rue 2015, version 17.6.20), which we also use to fit the model (see Sect. 4.2). Specifically, R-INLA sets vague Normal priors (centred at 0 with large variance) on suitable transformations (e.g. log) of the hyperparameters with unbounded range.

4.2 Estimation

The hierarchical Poisson log-linear model (HPL) of Sect. 4.1 was fitted using INLA (Rue et al. 2009). Specifically, INLA avoids time-consuming MCMC simulations by numerically approximating the posterior densities for the parameters of latent Gaussian models, which constitute a wide class of hierarchical models of the form

$$\begin{aligned} Y_i | \phi, \psi &\sim p(y_i | \phi, \psi), \\ \phi | \psi &\sim \text{Normal}(\mathbf{0}, \mathbf{Q}^{-1}(\psi)), \\ \psi &\sim p(\psi), \end{aligned}$$

where Y_i is the random variable corresponding to the observed response y_i , ϕ is a set of parameters (which may have a large dimension) and ψ is a set of hyperparameters.

The basic principle is to approximate the posterior densities for ψ and ϕ using a series of nested Normal approximations. The algorithm uses numerical optimization to find the mode of the posterior, while the marginal posterior distributions are computed using numerical integration over the hyperparameters. The posterior densities for the parameters of the HPL model are computed on the available data for each league.

To predict the outcome of a future match, we simulated 1000 samples from the joint approximated predictive distribution of the number of goals \tilde{Y}_1, \tilde{Y}_2 , scored in the future match by the home and away teams respectively, given features $\tilde{\mathbf{z}}_j = (\tilde{z}_{j1}, \dots, \tilde{z}_{j2})^\top$. Sampling was done using the `inla.posterior.sample` method of the R-INLA package. The predictive distribution has a probability mass function of the form

$$p(\tilde{y}_1, \tilde{y}_2 | y_1, y_2, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \mathbf{Z}) = \int p(\tilde{y}_1, \tilde{y}_2 | \mathbf{v}, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2) p(\mathbf{v} | y_1, y_2, \mathbf{Z}) d\mathbf{v},$$

where the vector \mathbf{v} collects all model parameters. We then compute the relative frequencies of the events $\tilde{Y}_1 > \tilde{Y}_2$, $\tilde{Y}_1 = \tilde{Y}_2$, and $\tilde{Y}_1 < \tilde{Y}_2$, which correspond to home win, draw, and loss respectively.

5 Validation framework

5.1 MLS challenge

The MLS challenge consists of predicting the outcomes (win, draw, loss) of 206 soccer matches from 52 leagues that take place between 31st March 2017 and 10th April 2017. The prediction performance of each submission was assessed in terms of the average ranked probability score (see Sect. 5.2) over those matches. To predict the outcomes of these matches, the challenge participants have access to over 200,000 matches up to and including the 21st March 2017, which can be used to train a classifier.

In order to guide the choice of the model that is best suited to make the final predictions, we designed a validation framework that emulates the requirements of the MLS Challenge. We evaluated the models in terms of the quality of future predictions, i.e. predictions about matches that happen after the matches used for training. In particular, we estimated the model parameters using data from the period before 1st April of each available calendar year in the data, and examined the quality of predictions in the period between 1st and 7th April of that year. For 2017, we estimated the model parameters using data from the period before 14th March 2017, and examined the quality of predictions in the period between 14th and 21st

Validation experiments

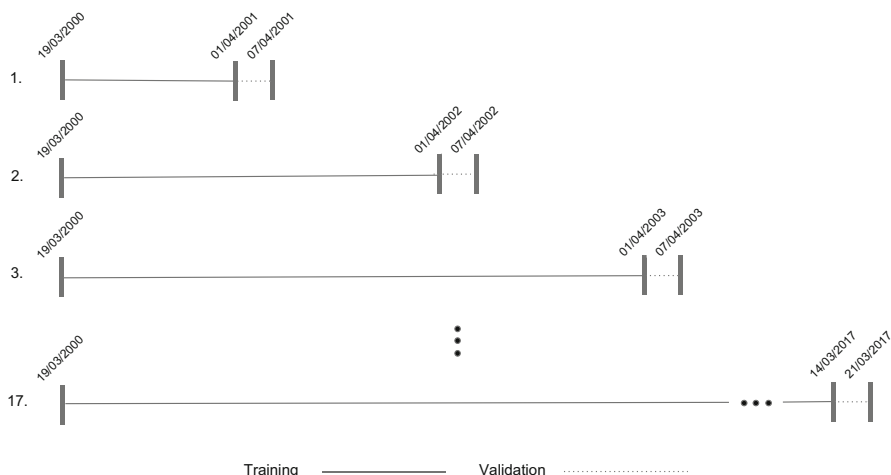


Fig. 3 The sequence of experiments that constitute the validation framework, visualizing their corresponding training and prediction periods

March 2017. Figure 3 is a pictorial representation of the validation framework, illustrating the sequence of experiments and the duration of their corresponding training and validation periods.

5.2 Validation criteria

The main predictive criterion we used in the validation framework is the ranked probability score, which is also the criterion that was used to determine the outcome of the challenge. Classification accuracy was also computed.

5.2.1 Ranked probability score

Let R be the number of possible outcomes (e.g. $R = 3$ in soccer) and \mathbf{p} be the R -vector of predicted probabilities with j -th component $p_j \in [0, 1]$ and $p_1 + \dots + p_R = 1$. Suppose that the observed outcomes are encoded in an R -vector \mathbf{a} with j th component $a_j \in \{0, 1\}$ and $a_1 + \dots + a_r = 1$. The ranked probability score is defined as

$$\text{RPS} = \frac{1}{r-1} \sum_{i=1}^{r-1} \left\{ \sum_{j=1}^i (p_j - a_j) \right\}^2. \quad (8)$$

The ranked probability score was introduced by Epstein (1969) (see also, Gneiting and Raftery 2007, for a general review of scoring rules) and is a strictly proper probabilistic scoring rule, in the sense that the true odds minimize its expected value (Murphy 1969).

Table 3 Illustration of the calculation of the ranked probability score and classification accuracy on artificial data

Observed outcome			Predicted probabilities			Predicted outcome			RPS	Accuracy
a_1	a_2	a_3	p_1	p_2	p_3	o_1	o_2	o_3		
1	0	0	1	0	0	1	0	0	0	1
1	0	0	0	1	0	0	1	0	0.5	0
1	0	0	0	0	1	0	0	1	1	0
1	0	0	0.8	0.2	0	1	0	0	0.02	1
0	1	0	0.33	0.33	0.34	0	0	1	0.11	0

5.2.2 Classification accuracy

Classification accuracy measures how often the classifier makes the correct prediction, i.e. how many times the outcome with the maximum estimated probability of occurrence actually occurs.

Table 3 illustrates the calculations leading to the ranked probability score and classification accuracy for several combinations of \mathbf{p} and \mathbf{a} . The left-most group of three columns gives the observed outcomes, the next group gives the predicted outcome probabilities, and the third gives the predicted outcomes using maximum probability allocation. The two columns in the right give the ranked probability scores and classification accuracies. As shown, a ranked probability score of zero indicates a perfect prediction (minimum error) and a ranked probability score of one indicates a completely wrong prediction (maximum error).

The ranked probability score and classification accuracy for a particular experiment in the validation framework are computed by averaging over their respective values over the matches in the prediction set. The uncertainty in the estimates from each experiment is quantified using leave-one-match out jackknife (Efron 1982), as detailed in step 9 of Algorithm 1.

5.3 Meta-analysis

The proposed validation framework consists of $K = 17$ experiments, one for each calendar year in the data. Each experiment results in pairs of observations $(s_i, \hat{\sigma}_i^2)$, where s_i is the ranked probability score or classification accuracy from the i th experiment, and $\hat{\sigma}_i^2$ is the associated jackknife estimate of its variance ($i = 1, \dots, K$).

We synthesized the results of the experiments using meta-analysis (DerSimonian and Laird 1986). Specifically, we make the working assumptions that the summary variances $\hat{\sigma}_i^2$ are estimated well-enough to be considered as known, and that s_1, \dots, s_K are realizations of random variables S_1, \dots, S_K , respectively, which are independent conditionally on independent random effects U_1, \dots, U_K , with

$$S_i \mid U_i \sim \text{Normal}(\alpha + U_i, \hat{\sigma}_i^2),$$

and

$$U_i \sim \text{Normal}(0, \tau^2).$$

The parameter α is understood here as the overall ranked probability score or classification accuracy, after accounting for the heterogeneity between the experiments.

The maximum likelihood estimate of the overall ranked probability or classification accuracy is then the weighted average

$$\hat{\alpha} = \frac{\sum w_i s_i}{\sum w_i},$$

where $w_i = (\hat{\sigma}_i^2 + \hat{\tau}^2)^{-1}$ and $\hat{\tau}^2$ is the maximum likelihood estimate of τ^2 . The estimated standard error for the estimator of the overall score $\hat{\alpha}$ can be computed using the square root of the inverse Fisher information about α , which ends up being $(\sum_{i=1}^K w_i)^{-1/2}$.

The assumptions of the random-effects meta-analysis model (independence, normality and fixed variances) are all subject to direct criticism for the validation framework depicted in Fig. 3 and the criteria we consider; for example, the training and validation sets defined in the sequence of experiments in Fig. 3 are overlapping and ordered in time, so the summaries resulting from the experiment are generally correlated. We proceed under the assumption that these departures are not severe enough to influence inference and conclusions about α .

5.4 Implementation

Algorithm 1 is an implementation of the validation framework in pseudo-code. Each model is expected to have a training method which trains the model on data, and a prediction method which returns predicted outcome probabilities for the prediction set. We refer to these methods as `train` and `predict` in the pseudo-code.

Algorithm 1 Pseudo-code for the validation framework

Input:

$\mathbf{x}_1, \dots, \mathbf{x}_G$	▷ feature vectors for all G matches in the data set
$d_1 \leq \dots \leq d_G$	▷ d_g is the match date of match $g \in \{1, \dots, G\}$
$\mathbf{o}_1, \dots, \mathbf{o}_G$	▷ match outcomes
train: $\{\mathbf{x}_g, \mathbf{o}_g : g \in A\} \rightarrow f(\cdot)$	▷ Training algorithm
predict: $\{\mathbf{x}_g : g \in B\}, f(\cdot) \rightarrow \{\bar{\mathbf{o}}_g : g \in B\}$	▷ Prediction algorithm
criterion: $\{\mathbf{o}_g, \bar{\mathbf{o}}_g : g \in B\} \rightarrow \{v_g : g \in B\}$	▷ observation-wise criterion values
D_1, \dots, D_T	▷ Cut-off dates for training for experiments
meta-analysis: $\{s_i, \hat{\sigma}_i^2 : i \in \{1, \dots, T\}\} \rightarrow \hat{\alpha}$	▷ Meta-analysis algorithm

Output: $\hat{\alpha}$

▷ Overall validation metric

```

1: for  $i \leftarrow 1$  to  $T$  do
2:    $A \leftarrow \{g : d_g \leq D_i\}$ 
3:    $B \leftarrow \{g : D_i < d_g \leq D_i + 10\text{days}\}$ 
4:    $n_B \leftarrow \dim(B)$ 
5:    $f(\cdot) \leftarrow \text{train}(\{\mathbf{x}_g, \mathbf{o}_g : g \in A\})$                                 ▷ fit the model
6:    $\{\bar{\mathbf{o}}_g : g \in B\} \leftarrow \text{predict}(\{\mathbf{x}_g : g \in B\}, f(\cdot))$           ▷ get predictions
7:    $\{v_g : g \in B\} \leftarrow \text{criterion}(\{\mathbf{o}_g, \bar{\mathbf{o}}_g : g \in B\})$ 
8:    $s_i \leftarrow \frac{1}{n_B} \sum_{g \in B} v_g$ 
9:    $\hat{\sigma}_i^2 \leftarrow \frac{n_B}{n_B - 1} \sum_{g \in B} \left( \frac{\sum_{h \in B/\{g\}} v_h}{n_B - 1} - s_i \right)^2$ 
10: end for
11:  $\hat{\alpha} \leftarrow \text{meta-analysis}(\{s_i, \hat{\sigma}_i^2 : i \in \{1, \dots, T\}\})$ 

```

Table 4 Description of each model in Section 3 and Section 4 in terms of features used, the handling of draws, the distribution whose parameters are modeled, and the estimation procedure that was used

Model	Draws	Features	Distribution	Estimation
BL (1)	Davidson	1	Multinomial	ML
BL (1)	Ordinal	1	Multinomial	ML
CS (2)	Davidson	1	Multinomial	ML
CS (2)	Ordinal	1	Multinomial	ML
LF (3)	Davidson	1, 6, 7, 12, 13	Multinomial	ML
LF (3)	Ordinal	1, 6, 7, 12, 13	Multinomial	ML
TVC (4)	Davidson	1, 6(<i>t</i>), 7(<i>t</i>), 12(<i>t</i>), 13	Multinomial	ML
TVC (4)	Ordinal	1, 6(<i>t</i>), 7(<i>t</i>), 12(<i>t</i>), 13	Multinomial	ML
AFD (5)	Davidson	1, 6(<i>t</i>), 7(<i>t</i>), 12(<i>t</i>), 13	Multinomial	MPL
HPL (7)		1, 2, 4, 6, 15, 16	Poisson	INLA
(†) TVC (4)	Ordinal	1, 2, 3, 4, 6(<i>t</i>), 7(<i>t</i>), 11(<i>t</i>)	Multinomial	ML

The suffix (*t*) indicates features with coefficients varying with matches played (feature 5 in Table 1)

The model indicated by † is the one we used to compute the probabilities for the submission to the MLS challenge

The acronyms are as follows: BL, Baseline (home advantage); CS, Bradley–Terry with constant strengths; LF, Bradley–Terry with linear features; TVC, Bradley–Terry with time-varying coefficients; AFD, Bradley–Terry with additive feature differences and time interactions; HPL, Hierarchical Poisson log-linear model

6 Results

In this section we compare the predictive performance of the various models we implemented as measured by the validation framework described in Sect. 5. Table 4 gives the details of each model in terms of features used, the handling of draws (ordinal and Davidson, as in Sect. 3.2), the distribution whose parameters are modeled, and the estimation procedure that has been used.

The sets of features that were used in the LF, TVC, AFD and HPL specifications in Table 4 resulted from ad-hoc experimentation with different combinations of features in the LF specification. All instances of feature 13 refer to the least squares ordinal rank (see Subsection 2.5 of the supplementary material). The features used in the HPL specification in (7) have been chosen prior to fitting to be home and newly promoted (features 1 and 2 in Table 1), the difference in form and points tally (features 4 and 6 in Table 1) between the two teams competing in match *g*, and season and quarter (features 15 and 16 in Table 1) for the season that match *g* takes place.

For each of the models in Tables 4 and 5 presents the ranked probability score and classification accuracy as estimated from the validation framework in Algorithm 1, and as calculated for the matches in the test set for the challenge.

The results in Table 5 are indicative of the good properties of the validation framework of Sect. 5 in accurately estimating the performance of the classifier on unseen data. Specifically, and excluding the baseline model, the sample correlation between overall ranked probability score and the average ranked probability score from the matches on the test set is 0.973. The classification accuracy seems to be underestimated by the validation framework.

The TVC model that is indicated by † in Table 5 is the model we used to compute the probabilities for our submission to the MLS challenge. Figure 4 shows the estimated time-varying coefficients for the TVC model. The remaining parameter estimates are 0.0410 for

Table 5 Ranked probability score and classification accuracy for the models in Table 4, as estimated from the validation framework of Section 5 (standard errors are in parentheses) and from the matches in the test set of the challenge

Model	Ranked probability score			Accuracy			Test
	Draws	Validation		Test	Validation		
BL	Davidson	0.2242	(0.0024)	0.2261	0.4472	(0.0067)	0.4515
BL	Ordinal	0.2242	(0.0024)	0.2261	0.4472	(0.0067)	0.4515
CS	Davidson	0.2112	(0.0028)	0.2128	0.4829	(0.0073)	0.5194
CS	Ordinal	0.2114	(0.0028)	0.2129	0.4779	(0.0074)	0.4951
LF	Davidson	0.2088	(0.0026)	0.2080	0.4849	(0.0068)	0.5049
LF	Ordinal	0.2088	(0.0026)	0.2084	0.4847	(0.0068)	0.5146
TVC	Davidson	0.2081	(0.0026)	0.2080	0.4898	(0.0068)	0.5049
TVC	Ordinal	0.2083	(0.0025)	0.2080	0.4860	(0.0068)	0.5097
AFD	Ordinal	0.2079	(0.0026)	0.2061	0.4837	(0.0068)	0.5194
*HPL		0.2073	(0.0025)	0.2047	0.4832	(0.0067)	0.5485
†TVC	Ordinal	0.2085	(0.0025)	0.2087	0.4865	(0.0068)	0.5388

The model indicated by † is the one we used to compute the probabilities for the submission to the MLS challenge, while the one indicated by * is the one that achieves the lowest estimated ranked probability score

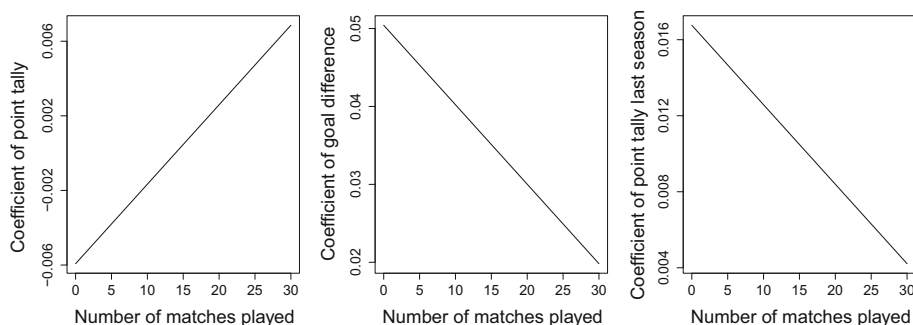


Fig. 4 Plots of the time-varying coefficients in the TVC model that is indicated by † in Table 5, which is the model we used to compute the probabilities for our submission to the MLS challenge

the coefficient of form, -0.0001 for the coefficient of days since previous match, and 0.0386 for the coefficient of newly promoted. Of all the features included, only goal difference and point tally last season had coefficients for which we found evidence of difference from zero when accounting for all other parameters in the model (the p values from individual Wald tests are both less than 0.001).

After the completion of the MLS challenge we explored the potential of new models and achieved even smaller ranked probability scores than the one obtained from the TVC model. In particular, the best performing model is the HPL model in Sect. 4.1 (starred in Table 5), followed by the AFD model which achieves a marginally worse ranked probability score. It should be noted here that the LF models are two simpler models that achieve performance that is close to that of HPL and AFD, without the inclusion of random effects, time-varying coefficients, or any non-parametric specifications.

The direct comparison between the ordinal and Davidson extensions of Bradley–Terry type models indicates that the differences tend to be small, with the Davidson extensions appearing to perform better.

We also tested the performance of HPL in terms of predicting actual scores of matches using the validation framework, comparing to a baseline method that always predicts the average goals scored by home and away teams respectively in the training data it receives. Using root mean square error as an evaluation metric, HPL achieved a score of 1.0011 with estimated standard error 0.0077 compared to the baseline which achieved a score of 1.0331 with estimated standard error 0.0083.

7 Conclusions and discussion

We compared the performance of various extensions of Bradley–Terry models and a hierarchical log-linear Poisson model for the prediction of outcomes of soccer matches. The best performing Bradley–Terry model and the hierarchical log-linear Poisson model delivered similar performance, with the hierarchical log-linear Poisson model doing marginally better.

Amongst the Bradley–Terry specifications, the best performing one is AFD, which models strength differences through a semi-parametric specification involving general smooth bivariate functions of features and season time. Similar but lower predictive performance was achieved by the Bradley–Terry specification that models team strength in terms of linear functions of season time. Overall, the inclusion of features delivered better predictive performance than the simpler Bradley–Terry specifications. In effect, information is gained by relaxing the assumption that each team has constant strength over the season and across feature values. The fact that the models with time varying components performed best within the Bradley–Terry class of models indicates that enriching models with time-varying specifications can deliver substantial improvements in the prediction of soccer outcomes.

All models considered in this paper have been evaluated using a novel, context-specific validation framework that accounts for the temporal dimension in the data and tests the methods under gradually increasing information for the training. The resulting experiments are then pooled together using meta-analysis in order to account for the differences in the uncertainty of the validation criterion values by weighing them accordingly.

The meta analysis model we employed operates under the working assumption of independence between the estimated validation criterion values from each experiment. This is at best a crude assumption in cases like the above where data for training may be shared between experiments. Furthermore, the validation framework was designed to explicitly estimate the performance of each method only for a pre-specified window of time in each league, which we have set close to the window where the MLS challenge submissions were being evaluated. As a result, the conclusions we present are not generalizable beyond the specific time window that was considered. Despite these shortcomings, the results in Table 5 show that the validation framework delivered accurate estimates of the actual predictive performance of each method, as the estimated average predictive performances and the actual performances on the test set (containing matches between 31st March and 10th April 2017) were very close.

The main focus of this paper is to provide a workflow for predicting soccer outcomes, and to propose various alternative models for the task. Additional feature engineering and selection, and alternative fitting strategies can potentially increase performance and are worth pursuing. For example, ensemble methods aimed at improving predictive accuracy like calibration,

boosting, bagging, or model averaging (for an overview, see Dietterich (2000)) could be utilized to boost the performance of the classifiers that were trained in this paper.

A challenging aspect of modeling soccer outcomes is devising ways to borrow information across different leagues. The two best performing models (HPL and AFD) are extremes in this respect; HPL is trained on each league separately while AFD is trained on all leagues simultaneously, ignoring the league that teams belong to. Further improvements in predictive quality can potentially be achieved by using a hierarchical model that takes into account which league teams belong to but also allows for sharing of information between leagues.

8 Supplementary material

The supplementary material document contains two sections. Section 1 provides plots of the number of matches per number of goals scored by the home and away teams, by country, for a variety of arbitrarily chosen countries. These plots provide evidence of a home advantage. Section 2 details approaches for obtaining team rankings (feature 13 in Table 2) based on the outcomes of the matches they played so far.

Acknowledgements This work was supported by The Alan Turing Institute under the EPSRC Grant EP/N510129/1.

Author contributions The authors are grateful to Petros Dellaportas, David Firth, István Papp, Ricardo Silva, and Zhaozhi Qian for helpful discussions during the challenge. Alkeos Tsokos, Santhosh Narayanan and Ioannis Kosmidis have defined the various Bradley–Terry specifications, and devised and implemented the corresponding estimation procedures and the validation framework. Gianluca Baio developed the hierarchical Poisson log-linear model and the associated posterior inference procedures. Mihai Cucuringu did extensive work on feature extraction using ranking algorithms. Gavin Whitaker carried out core data wrangling tasks and, along with Franz Király, worked on the initial data exploration and helped with the design of the estimation-prediction pipeline for the validation experiments. Franz Király also contributed to the organisation of the team meetings and communication during the challenge. All authors have discussed and provided feedback on all aspects of the challenge, manuscript preparation and relevant data analyses.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Agresti, A. (2015). *Foundations of linear and generalized linear models*. Hoboken: Wiley.
- Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2), 253–264.
- Berrar, D., Dubitzky, W., Davis, J., & Lopes, P. (2017). Machine learning for Soccer. Retrieved from osf.io/ftuwa.
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3–4), 502–537.
- Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, 16, 1190.
- Cattelan, M., Varin, C., & Firth, D. (2013). Dynamic BradleyTerry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1), 135–150.
- Davidson, R. R. (1970). On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329), 317.
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177.
- Dietterich, T. G. (2000). *Ensemble methods in machine learning* (pp. 1–15). Berlin: Springer.

- Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, 46(2), 265.
- Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. New Delhi: SIAM.
- Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6), 985–987.
- Firth, D. (2005). Bradley-Terry Models in R. *Journal of Statistical Software*, 12(1)
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.
- Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing good ridge parameter. *Technometrics*, 21(2), 215.
- Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society D*, 52, 381–393.
- Király, F.J., & Qian, Z. (2017). Modelling competitive sports: Bradley–Terry–Élo models for supervised and on-line learning of paired competition outcomes (pp. 1–53). [arXiv:1701.08055](https://arxiv.org/abs/1701.08055).
- Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software, Articles*, 63(19), 1–25.
- Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3), 109–118.
- Murphy, A. H. (1969). On the ranked probability score. *Journal of Applied Meteorology*, 8(6), 988–989.
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*, 71, 319–392.
- Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.
- Wikipedia (2018). UEFA coefficient — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=UEFA%20coefficient&oldid=819064849>. Accessed February 09 2018.
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65(1), 95.
- Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton: CRC Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.