



The role of passing network indicators in modeling football outcomes: an application using Bayesian hierarchical models

Riccardo Ievoli¹ · Aldo Gardini² · Lucio Palazzo³

Received: 28 December 2020 / Accepted: 22 June 2021
© The Author(s) 2021

Abstract

Passes are undoubtedly the more frequent events in football and other team sports. Passing networks and their structural features can be useful to evaluate the style of play in terms of passing behavior, analyzing and quantifying interactions among players. The present paper aims to show how information retrieved from passing networks can have a relevant impact on predicting the match outcome. In particular, we focus on modeling both the scored goals by two competing teams and the goal difference between them. With this purpose, we fit these outcomes using Bayesian hierarchical models, including both in-match and network-based covariates to cover many aspects of the offensive actions on the pitch. Furthermore, we review and compare different approaches to include covariates in modeling football outcomes. The presented methodology is applied to a real dataset containing information on 125 matches of the 2016–2017 UEFA Champions League, involving 32 among the best European teams. From our results, shots on target, corners, and such passing network indicators are the main determinants of the considered football outcomes.

Keywords Network intensity · Poisson regression · Regularized horseshoe prior · Skellam distribution · UEFA Champions League

✉ Riccardo Ievoli
riccardo.ievoli@unife.it

Aldo Gardini
aldo.gardini2@unibo.it

Lucio Palazzo
lucio.palazzo@unina.it

¹ Department of Economics and Management, University of Ferrara, Ferrara, Italy

² Department of Statistical Sciences “P. Fortunati”, University of Bologna, Bologna, Italy

³ Department of Political Sciences, University of Naples Federico II, Napoli, Italy

1 Introduction

New technologies, such as wearable devices, multiple-camera player trackers, and drone-based analysis of training sessions, are increasing the ways to retrieve data and provide new opportunities in team sports analysis. The statistical analysis of collected data recently spread in real applications concerning team sports and football especially (Albert et al. 2005; Memmert 2019).

In this applied field, most statistical methods focus on modeling football outcomes, such as the team winning, exact match result, scored goals, and ball possession. In fact, statistical models are useful to assess the main determinants that explain and/or predict football outcomes. In this sense, in-match indicators, political-economic factors, or socio-geographic features are often used as explanatory variables.

According to Egidi and Torelli (2020), two main types of statistical models can be distinguished in this context: *result-based* and *goal-based*. The first type is based on a multinomial outcome, typically constituted by the following categories: home win, draw, and home loss (labeled as 1, X, 2). The second one considers the number of goals scored by each competing team. For clarity, we propose to distinguish a further family of models characterized by the goal difference as the response variable. We decide to denote them as *difference-based models*, even if, sometimes, they are included among the goal-based models. In this paper, the difference-based and the goal-based approaches are considered.

Goal-based models are usually defined assuming a probability distribution suitable for counts to model the response. From the seminal work by Maher (1982), the use of conditionally independent Poisson distributions represents the default choice in modeling the number of scored goals by each team in a match. Some notable extensions are the works by Dixon and Coles (1997) and Rue and Salvesen (2000) that take into account changes in team conditions usually occurring along the season. Karlis and Ntzoufras (2003) proposed to include the possibility that scored goals are positively correlated within matches, specifying a bivariate Poisson distribution. On the other hand, the work by Karlis and Ntzoufras (2009) can be denoted as the first example of a difference-based model: the authors assumed a Skellam distribution to fit the *goals-difference* between two teams. Many of these fundamental contributions belong to the Bayesian framework, the inferential approach considered in this paper. Among the others, it is worth citing Baio and Blangiardo (2010), Egidi and Torelli (2020), and Manderston et al. (2018) as empirical applications based on Bayesian inference.

Besides individual skills of players, tactics and team strategies are key elements for succeeding in football, and appropriate methodology to deal with these elements is still under debate. Furthermore, network metrics are recently becoming more popular in football, as highlighted in Pena and Touchette (2012) and Clemente et al. (2015). In particular, network analysis is applied to football passing distributions: a relevant contribution is Grund (2012), and other interesting applications can be found in Gonçalves et al. (2017), Mclean et al. (2018), Clemente et al. (2020), and Ichinose et al. (2021). This approach is also exploited

in other team sports (see, e.g., Braham and Small 2018, for an application to Australian football). Passing networks present some advantages: they help to detect patterns or strong/weak ties among players and their positions in the lineup. They also provide valuable evidence of players' skills, tactics, and connections between positions. Unfortunately, the data at the individual level, required to generate passing networks, are open access only for the major international competitions.

The first main contribution of the paper is to show the potential of Bayesian hierarchical models in either managing covariates and finding the determinants of football outcome. In fact, the flexibility of Bayesian models and the usage of appealing computational tools allow us to review and discuss the practical meaning of different model specifications, considering diverse dependence relationships between the response and the covariates. Several probabilistic programming languages can be employed in a Bayesian framework to draw samples from the posterior distributions of the parameters. The models presented in this work are estimated using Stan (Stan Development Team 2020), and the code is available as supplementary material. We aim to carefully compare several expressions for the linear predictors, assessing their potential drawbacks on real data, in order to advise applied researchers and experts. Regularized horseshoe priors (Piiironen and Vehtari 2017) are assumed for the regression coefficients to control their posterior variance, avoid multicollinearity, and limit the occurrence of over-fitting issues that might lead to poor out-of-sample performances. Regularized estimates of the coefficients were considered in recent works (Groll et al. 2015; Schauberger et al. 2018) from a pure frequentist perspective.

A further aim of this work is to exploit performance indicators derived from passing networks in the analysis football outcomes, including network indices as explanatory variables. We decide to take advantage of this information with other in-match variables to explore their interplay in determining the game outcome. In fact, passes represent more than 80% of the events in football (Cintia et al. 2015), and they convey crucial information on the strength of a team. To summarize, we include network indices with team-level control variables (usually free available match statistics) to suggest to football insiders which factors are the most important in determining the outcome on the pitch.

To the best of our knowledge, none of the previously cited works takes into account the strength of relationships among players, neither the type of interactions among them. Exceptions can be found in Grund (2012), which does not use in-match covariates, and, more recently, in Diquigiovanni and Scarpa (2019), which exploited network-based clusters to model the scored goals. Furthermore, Ievoli et al. (2021) used network indicators to model the probability of winning the game for a team. From a slightly different perspective, Carpita et al. (2019) included the a priori evaluation of players' abilities (involving passing skills) in predicting the win, without using passing network information.

The rest of the paper is organized as follows: in Sect. 2, we present the main variables of our analysis, dividing them into in-match variables and network summary measures. In Sect. 3, we introduce goal based and difference based Bayesian hierarchical models, using four different specifications of the linear predictor. Section 4

contains the main results of proposed models applied to a real dataset regarding the 2016–2017 UEFA Champions League (UCL). They are followed by a brief discussion regarding the meaning of our results and their practical implications in football (Sect. 5). Finally, concluding remarks are summarized in Sect. 6.

2 Variable measurement

In this section, we define the set of variables that will be used in the proposed statistical models. The football outcomes and the in-match covariates are firstly introduced, then we skip to the definition of the network-based summary measures. It is worth stressing that, for each generic match g , we record observations both for the home team (H) and the away team (A).

2.1 Football outcomes

Throughout the paper, the football outcome is defined in the following ways:

- (a) y_g^H is the number of goals scored by the home team
- (b) y_g^A is the number of goals scored by the away team
- (c) $z_g = y_g^H - y_g^A$ is the difference between goals scored by two competing teams (or margin of victory).

Definitions (a) and (b) characterize the goal-based modeling approach and can be found, for instance, in Maher (1982), Dixon and Coles (1997), Rue and Salvesen (2000), and Karlis and Ntzoufras (2003). On the other hand, definition (c) refers to the difference-based approach and appears in Karlis and Ntzoufras (2009) or Manderson et al. (2018), among others.

2.2 In-match variables

Regarding the in-match covariates, we collect indices for the two competing teams of each match. The focus is mainly on conventional quantities often used in applied works (Castellano et al. 2012; Carpita et al. 2015; Schauburger et al. 2018; Lepschy et al. 2020). These variables refer to actual events in match g , and their observed values are stored in vectors \mathbf{x}_g^H and \mathbf{x}_g^A for the home team and the away team, together with the network-based variables. The set of in-match variables is described in the following.

Shots on target: it is the number of attempted shots on goal per team in a match. It measures the ability to produce concrete opportunities.

Corners: it is the count of obtained corner kicks, which can be another relevant output of the offensive actions.

Fouls suffered: it is the raw number of suffered fouls (including hands), which can interrupt the offensive actions.

Ball possession: it is the ratio between the time in which a team plays the ball and the total match time.

Distance: it is the sum of meters covered by all the players of a team, representing a synthetic measure of athletic skills.

2.3 Network-based variables

Network analysis deals with relational data, emphasizing the investigation of the structure generated between units, driven by quantity and quality of ties occurring among them. Network theory includes all possible methods to analyze data presenting interactions between a set of units (agents) to investigate patterns and community structures (Wasserman and Faust 1994).

A network is defined as the ordered triple $(\mathcal{V}, \mathcal{A}, \mathcal{W})$ consisting of a set of vertices $v \in \mathcal{V}$, a set of arcs $a \in \mathcal{A} \subseteq \mathcal{V} \times \mathcal{V}$, and a set of weights $\omega \in \mathcal{W}$. Both sets of vertices \mathcal{V} and arcs \mathcal{A} are assumed to be finite. An ordered pair of vertices denotes an arc through a function $\psi: \psi(v_i, v_j) = a_{ij} \in \mathcal{A}$, mapping the one-directional tie from vertex v_i to v_j , with $i, j = 1, \dots, n$. The mapping $\omega: \mathcal{A} \rightarrow \mathbb{R}$ defines a weight related to each arc. A network can be expressed in the form of an *adjacency matrix* $P = (p_{ij})$, with $p_{ij} = \omega(a_{ij})$ if $\exists a_{ij} \in \mathcal{A}$ and $p_{ij} = 0$ otherwise and $i \neq j$. We also assume that “loops”, i.e., arcs connecting a vertex to itself, as are not allowed. Therefore, $p_{ij} = 0$ when $i = j$.

In football, a specific adjacency matrix can be obtained for each match of a team. Considering the starting line-up, eleven players ($n = 11$) are depicted in rows and columns of the matrix, and cells contains the number of completed passes between players of a given team. In practice, this matrix can be read in two different ways, i.e., row-wise and column-wise. In the former, a generic cell p_{ij} contains the number of passes that i -th player gives to j -th player, while, in the latter, the cell expresses the number of passes that j -th player receives from i -th player. To summarize, this matrix is composed by 110 cells, since we assume that a player cannot pass the ball to himself, and it is generally not symmetric.

In Table 1, an example of adjacency matrix is reported. It concerns the team passing network distribution of Arsenal observed for the match against PSG of 09/13/2016. Figure 1 represents the directed and weighted graph obtained through the adjacency matrix of Table 1. Players p_i are connected by arrows (representing the arcs), and the strength of the relationships (i.e., the number of passes) is expressed through the arrows' width. The graph is obtained using the package `igraph` (Csardi and Nepusz 2006) in R software (R Core Team 2020). This representation primarily allows to make comparisons at the individual (micro) level involving players, but some information can also be extracted at the team (macro) level to evaluate the overall performance.

After introducing the passing networks, it is necessary to find techniques able to synthesize the information contained in them. Several network indices can be

Table 1 Example of an adjacency matrix in football

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}
Ospina	p_1		1	1	1	1	4	1	11	2	1
Koscielny	p_2	1		3	2	0	13	11	12	1	1
Alexis Sanchez	p_3	0	3		9	4	4	0	4	0	4
Ozil	p_4	0	2	8		2	8	3	6	2	3
Oxlade-Chamberlain	p_5	0	2	2	2		0	0	4	0	2
Iwobi	p_6	0	4	6	8	0		7	3	4	4
Monreal	p_7	3	3	8	2	3	7		5	1	0
Santi Cazorla	p_8	0	10	11	13	5	7	5		8	7
Mustafi	p_9	6	13	1	1	2	2	0	15		13
Bellerin	p_{10}	1	0	2	1	4	5	0	2	11	
Coquelin	p_{11}	0	3	5	3	1	4	3	9	2	3

It represents the team passing network of Arsenal observed in a Group Stage match of the 2016–2017 UCL

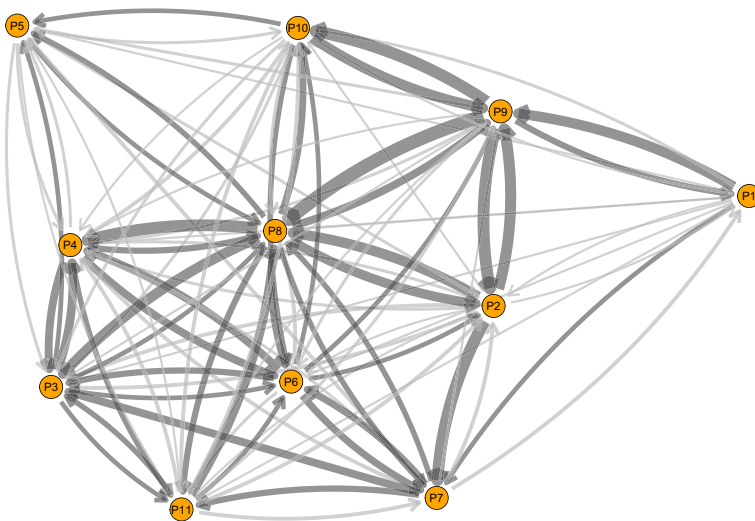


Fig. 1 Graph of the team passing network of Arsenal (vs PSG, 09/13/2016), generated from the adjacency matrix illustrated in Table 1. Dark lines emphasize links presenting weights higher than the median of adjacency matrix

found in the literature (see, Wasserman 1994; Carrington et al. 2005; De Nooy et al. 2018, among others). They can be computed from team passing networks and interpreted as performance indicators (see, e.g., Clemente et al. 2016). We focus on some network summary measures that are able to capture the complexity of network topology with a meaningful interpretation for football. These indices are described in the following.

Pass accuracy: at the team level, it is defined as the sum of ratios between completed (P_i^c) and attempted (P_i^a) passes for each i -th player. It is proposed as a measure of the overall technical skills of a football team since it may have a key role in creating offensive actions and, in some cases, can avoid to receive counterattacks from the opponent. To summarize, this index represents a “bridge” between in-match covariates and network summary measures.

Network intensity: it is defined as $\text{Time}_e^{-1} \sum_i \sum_j p_{ij}$, where Time_e is the actual ball possession time (in minutes), and p_{ij} is the generic element of the adjacency matrix of a team in a match (e.g., see Table 1). Since it is crucial to take into account the real minutes of ball possession (i.e., excluding when the ball is out of play), we propose the usage of effective time of ball possession (Time_e). This is a modification of the index introduced by Grund (2012), which used the overall time of ball possession instead of Time_e . From a football perspective, this index quantify the passing speed of a team in a match, i.e., the aptitude to circulate the ball quickly among teammates.

Network diameter: it is the *geodesic* distance between most distant vertices of a graph, without taking into account the link weights. Given a set of vertices of a network \mathcal{V} and the *geodesic* distance $d(u, v)$, measured between two vertices $u, v \in \mathcal{V}$, the diameter can be expressed as $\max_{u,v \in \mathcal{V}} d(u, v)$. High network diameter values express the ability to generate as many direct connections as possible in terms of passes, even considering that the theoretical maximum in our setting corresponds to the number of players (11). It can be also viewed as a measure of tactical variety (e.g., the ability to make cross-field passes).

Reciprocity: it is computed as the proportion of mutual connections in a directed graph, i.e., the frequency of opposite counterpart of a directed arc also included in the graph. Given $L^\leftrightarrow = \{a \mid a \text{ is a bidirectional arc} \} \subseteq \mathcal{A}$, reciprocity corresponds to $|L^\leftrightarrow|/|\mathcal{A}|$. In football, it measures the ability of two players to have mutual connections with each other. Moreover, it also evaluates the balance of a team in terms of passing directions. For example, high values of reciprocity can be related to the propensity of certain type of relationships, such as the so-called “give and go” or “wall passes”.

Median of average nearest neighbors (MANN): it is computed as the median of the average nearest neighbors. For each player p_i , the average degrees of partners for the i -th vertex can be computed as (Clemente et al. 2016):

$$\text{ANN}_i = \frac{\sum_j (p_{ij} + p_{ji})(p_j + p_{\cdot j})}{2(p_i + p_{\cdot i})}$$

where p_i and $p_{\cdot i}$ are, respectively, the row and column marginal sums of the adjacency matrix. This is an individual index expressing the correlation levels between pairs of players. The overall index (MANN), at the team level, measures the cohesion in terms of passing behaviors: the presence of one or few key players on the pitch, in terms of completed and received passes, leads to higher values of this index.

Third quartile of hub (Q_3 -Hub): it is computed as the 75-th percentile of the individual hub indices. The algorithm to compute hubs can be found in Kleinberg

Table 2 Input variables setting scheme

	Description	Dimension	Domain	Expression	Ref.
In-match	Shots on target	Effective scoring chances	\mathbb{N}	–	[1]
	Corners	Inactive scoring chances	\mathbb{N}	–	[1]
	Fouls suffered	Interruption of play	\mathbb{N}	–	[1]
	Ball possession	Control of the ball	[0, 1]	–	[1]
	Distance	Health and physical strength	\mathbb{R}^+	–	[1]
Net-work	Accuracy	Passing skills	[0, 1]	$\sum_i \frac{P_i^c}{P_i^a}$	[2]
	Network intensity	Passing speed	\mathbb{R}^+	$\text{Time}_e^{-1} \sum_i \sum_j P_{ij}$	[3]
	Network diameter	Quality of the play	{0, ..., 11}	$\max_{u,v \in V} d(u, v)$	[4]
	Reciprocity	Balance in terms of passing directions	[0, 1]	$\frac{ L^{\leftrightarrow} }{ A }$	[4]
	MANN	Cohesion in terms of passing behaviors	\mathbb{R}^+	$\text{Med}_i \left(\frac{\sum_j (p_{ij} + p_{ji})(p_{j\cdot} + p_{\cdot j})}{2(p_{i\cdot} + p_{\cdot i})} \right)$	[4]
	Q ₃ -Hub	Level of playmaking	[0, 1]	HITS algorithm	[5]

Legend: [1], see, e.g., Castellano et al. (2012); Carpita et al. (2015); Schaubberger et al. (2018); Lepschy et al. (2020) among others. [2], Schaubberger et al. (2018). [3], Grund (2012) [4], Clemente et al. (2016). [5], Ievoli et al. (2021)

et al. (2011). High values of such index are associated to players with good ability in passing the ball to other players. Hub players can be seen as play-makers of a team. Q₃-Hub can be considered a synthetic measure expressing the level of “play-making” of a team.

Table 2 has the role of summarizing the variables of this section, depicting their description, the covered dimension of the game, the domain, the mathematical expression (when required), and, lastly, the associated references. A detailed explanation on how passing network matrices can be processed to obtain graphs and compute the previously illustrated summary measures is provided as supplementary material, including code and data example.

3 Bayesian modeling of football outcomes

In this section, statistical models that will be estimated on UCL data are illustrated. Recalling the football outcomes defined in Sect. 2, we can distinguish two families of statistical models. Goal-based models are specified if the couple of scored goals

(y_g^H, y_g^A) is considered as the response for match $g = 1, \dots, G$. Alternatively, difference-based models are fitted if the goal difference between the competing teams z_g is employed as the response. Starting from goal-based models, the first setting we discuss relies on Maher (1982) model. He proposed to specify two conditionally independent Poisson distributions for the scored goals (labeled with IP in following):

$$y_g^H | \lambda_g^H \sim \text{Poi}(\lambda_g^H), \quad y_g^A | \lambda_g^A \sim \text{Poi}(\lambda_g^A); \quad g = 1, \dots, G. \tag{1}$$

The Poisson parameters λ_g^H and λ_g^A are modeled specifying the following linear models on their logarithmic transformations:

$$\begin{aligned} \log(\lambda_g^H) &= \mu + h + \mathbf{u}_g^T \boldsymbol{\beta}_H + \alpha_{H_g} + \delta_{A_g}; \\ \log(\lambda_g^A) &= \mu + \mathbf{v}_g^T \boldsymbol{\beta}_A + \alpha_{A_g} + \delta_{H_g}. \end{aligned} \tag{2}$$

In football, such linear predictors are characterized by the following fixed components: a common location parameter μ , a home effect parameter h , accounting for the possible favorable conditions that the team hosting the game may have, and a linear combination of p covariate values ($\mathbf{u}_g \in \mathbb{R}^p$ for the home team and $\mathbf{v}_g \in \mathbb{R}^p$ for the away team) with the associated regression coefficients $\boldsymbol{\beta}_H$ and $\boldsymbol{\beta}_A$. Teams-crossed random effects are included too: α_t conveys the attacking ability of team $t = 1, \dots, T$, whereas δ_t concerns its defense ability. The subscripts of the effects (H_g and A_g) represent the indices of teams involved in match g , remembering that the attacking effect of a given team and the defense effect of the opponent team concur in explaining the number of scored goals.

Karlis and Ntzoufras (2003) noted that setting a model by means of conditionally independent Poisson distributions might neglect the positive correlation that is commonly observed between the number of goals scored by the competing teams. To overcome this issue, they proposed to model the couples of scored goals through a bivariate Poisson (BP) distribution:

$$(y_g^H, y_g^A) | \lambda_g^H, \lambda_g^A, \lambda_g^C \sim \text{Biv-Poi}(\lambda_g^H, \lambda_g^A, \lambda_g^C); \quad g = 1, \dots, G. \tag{3}$$

Under this model, the following moment's expressions hold: $\mathbb{E}[y_g^H] = \lambda_g^H + \lambda_g^C$, $\mathbb{E}[y_g^A] = \lambda_g^A + \lambda_g^C$, and the covariance is $\mathbb{C}[y_g^H, y_g^A] = \lambda_g^C$. For parameters λ_g^H and λ_g^A , the same predictors of (2) are assumed, whereas for the correlation parameter, in line with Karlis and Ntzoufras (2003):

$$\log(\lambda_g^C) = \mu_C + \rho_{H_g} + \rho_{A_g}, \tag{4}$$

where μ_C is the baseline correlation level and ρ_t is a team-specific random effect.

In parallel, moving to the framework of difference-based modeling, following Karlis and Ntzoufras (2009), we model the margin of victory z_g assuming a Skellam distribution (Sk):

$$z_g | \lambda_g^H, \lambda_g^A \sim Sk\left(\lambda_g^H, \lambda_g^A\right); \quad g = 1, \dots, G. \quad (5)$$

The Skellam distribution is defined as the difference between two Poisson distributions having intensity parameters λ_g^H and λ_g^A , which are defined as in (2). We stress that, unlike goal-based models, the use of the z_g leads to the loss of the match outcome magnitude, and the two Poisson intensity parameters do not directly pertain to the number of scored goals by a given team. On the other hand, assuming a Skellam distribution for the difference implies marginal distributions for the scored goals that are more flexible than the Poisson (even in the bivariate case). As a matter of fact, the BP model accounts for the correlation of the couple (y_g^A, y_g^H) through a Poisson distribution (that has intensity λ_g^C in our notation), whereas, under the Sk model, the correlation is implicitly modeled by means of any discrete random variable. For further details of this aspect see Karlis and Ntzoufras (2009).

3.1 Prior distributions

Since we are in the Bayesian inferential framework, we need to specify a prior distribution for each parameter included in the model. Starting from the random effect vectors α , δ and ρ , we assume them as a priori independent, following zero-mean Gaussian distributions:

$$\alpha | \sigma_\alpha^2 \sim \mathcal{N}_T(\mathbf{0}, \sigma_\alpha^2 \mathbf{I}_T); \quad \delta | \sigma_\delta^2 \sim \mathcal{N}_T(\mathbf{0}, \sigma_\delta^2 \mathbf{I}_T); \quad \rho | \sigma_\rho^2 \sim \mathcal{N}_T(\mathbf{0}, \sigma_\rho^2 \mathbf{I}_T),$$

and the classical sum-to-zero constraints are imposed: $\sum_{t=1}^T \alpha_t = 0$, $\sum_{t=1}^T \delta_t = 0$, and $\sum_{t=1}^T \rho_t = 0$. In doing so, the random effects capture the deviations due to the attacking and defensive abilities of the specific team from the fixed part of the linear predictor, i.e., the overall mean, the possible home effect, and the linear combination of coefficients with the observed covariates (Karlis and Ntzoufras 2003). To sample from the constrained posterior distribution, see the manual of the Stan software (Stan Development Team 2020).

The regression coefficients (β_H and β_A) included in (2) might be estimated through the introduction of a penalty term. In fact, several variables could affect the response, all being included in the linear predictor. For this reason, regularization methods are used for variable selection since they shrunk to zero coefficient estimates related to negligible covariates, reducing the parameters' variance. Among the others, Groll et al. (2015) and Tutz and Schauburger (2015) considered the LASSO framework, whereas the problem has not been tackled yet from the Bayesian perspective. A plethora of shrinkage priors for the regression coefficients are available (Bhadra et al. 2019), here we decide to adopt the regularized horseshoe prior by Piironen and Vehtari (2017): it easily allows to incorporate prior information about sparseness and can be interpreted as the continuous version of the popular spike-and-slab priors. The prior setting for the generic regression coefficient is defined as follows:

$$\beta_{k,j} | \tau_k^2, \lambda_{k,j}^2 \sim \mathcal{N}\left(0, \tau_k^2 \tilde{\lambda}_{k,j}^2\right), \quad k \in \{H, A\}, \quad j = 1, \dots, p;$$

$$\tilde{\lambda}_{k,j}^2 = \frac{c^2 \lambda_{k,j}^2}{c_k^2 + \tau_k^2 \lambda_{k,j}^2},$$

where τ_k represents a global scale, $\lambda_{k,j}$ a local scale, and c is a further scale parameter that controls the prior assumed on the coefficients not shrunk toward 0, i.e., the *slab* part of the prior.

The hierarchy is completed assuming:

$$\lambda_{k,j} \sim \mathcal{C}^+(0, 1),$$

$$\tau_k \sim \mathcal{C}^+(0, \tau_0),$$

$$c_k^2 \sim \text{Inv-Gamma}\left(\frac{\nu_{\text{slab}}}{2}, \frac{\nu_{\text{slab}}}{2} s_{\text{slab}}^2\right),$$

where ν_{slab} and s_{slab}^2 can be interpreted as the degrees of freedom and scale of the Student's t prior on the slab part, since the prior on the regression coefficient tends to a $\mathcal{N}(0, c_k^2)$ in absence of shrinking. The prior scale of τ_k is fixed as $\tau_0^2 = p_0 \tilde{\sigma} / ((p - p_0) \sqrt{G})$: p_0 is interpreted as the prior number of expected non-null effects and $\tilde{\sigma}^2$ is the pseudo-variance of generalized linear models that in the Poisson case with logarithmic link can be fixed as the reciprocal of the sample mean (Pii-ronen and Vehtari 2017).

Eventually, in line with Gelman et al. (2006), half-Student's t priors are set for the scale hyperparameters σ_α , σ_δ and σ_ρ , choosing 3 degrees of freedom. A non-informative Gaussian prior centered in zero having large variance is chosen for the parameters μ , h , and μ_C .

3.2 Model specifications: linear predictors

One of this paper aims is to explore and compare possible relationships that can be assumed between the response variable and the covariates included in the model. As discussed in Sect. 2, variables listed in Table 2 are used as auxiliary information, and, referring to game g , they are contained in vector \mathbf{x}_g^H for the home team and \mathbf{x}_g^A for the away team.

Firstly, baseline models without any auxiliary information are considered. Throughout the paper, we will label this formulation of the linear predictors as M_0 , where M will be replaced by the specified model (i.e., IP, BP, or Sk). Then, following Groll et al. (2015) and Groll et al. (2018), the differences between the covariates observed for the two teams in match g are used: we refer to this specification with M_1 . With M_2 , we indicate the most flexible model specification considered: we link each linear predictor to the covariates observed on the specific team. Finally, the latter framework is simplified in M_3 by assuming a common vector of regression coefficients for both the competing teams. Table 3 has the role of summarizing the considered specifications in terms of algebraic relationships and restrictions imposed on equations terms in (2).

Table 3 Assumptions of linear predictors under the different models

Specification	Assumptions
M_0	$\mathbf{u}_g^T \boldsymbol{\beta}_H = \mathbf{v}_g^T \boldsymbol{\beta}_A = 0$
M_1	$\mathbf{u}_g = \mathbf{x}_g^H - \mathbf{x}_g^A, \mathbf{v}_g = \mathbf{x}_g^A - \mathbf{x}_g^H$ $\boldsymbol{\beta}_H = \boldsymbol{\beta}_A = \boldsymbol{\beta}$
M_2	$\mathbf{u}_g = \mathbf{x}_g^H, \mathbf{v}_g = \mathbf{x}_g^A$ $\boldsymbol{\beta}_H \neq \boldsymbol{\beta}_A$
M_3	$\mathbf{u}_g = \mathbf{x}_g^H, \mathbf{v}_g = \mathbf{x}_g^A$ $\boldsymbol{\beta}_H = \boldsymbol{\beta}_A = \boldsymbol{\beta}$

3.3 Model checking and model selection

To draw samples from the posterior distribution of the parameters that characterize the models described in this section, the Stan software is used through its R interface rstan. In this way, Markov Chain Monte Carlo (MCMC) samples are obtained from each parameters posterior distribution and can be used to carry out inference. Any posterior distribution can be synthesized by computing Monte Carlo estimates of the usual summary statistics such as the mean, the standard deviation, and the quantiles, that are used to produce the credible intervals.

The natural way in which prediction is carried out represents an appealing feature of the Bayesian inferential framework. In fact, the posterior predictive distribution allows to perform the prediction of a potentially unknown future observation of the outcome, here labeled as $(\tilde{y}^H, \tilde{y}^A)$ and \tilde{z} . For example, considering the difference response variable z , and indicating with $\boldsymbol{\theta}$ the vector containing all the model parameters, the posterior predictive distribution of \tilde{z} is defined by the following integral:

$$p(\tilde{z}|z) = \int_{\boldsymbol{\theta}} p(\tilde{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|z)d\boldsymbol{\theta},$$

where $p(\tilde{z}|\boldsymbol{\theta})$ is the likelihood function of the predicted observation and $p(\boldsymbol{\theta}|z)$ is the posterior distribution of the model parameters.

The posterior predictive distribution is obtained integrating out the model parameters, and therefore the predictions include all the uncertainty due to the estimation procedure. Moreover, since the MCMC samples from the posterior of $\boldsymbol{\theta}$ are available, sampling from $p(\tilde{z}|z)$ is trivial. This distribution can be exploited both for forecasting purposes and for checking the model goodness-of-fit through the posterior predictive checks (Rubin 1984).

The samples generated from the posterior predictive distribution of the scored goals can be further combined to obtain the prediction of the multinomial game outcome $O_g \in \{1, X, 2\}$. More in detail, the posterior probabilities of the outcomes of the g -th match (p_{1g}, p_{Xg}, p_{2g}) are computed, then the predicted result by the model is fixed as $\hat{O}_g = \max_{i=\{1,X,2\}} p_{ig}$.

As synthetic measures of the models' ability in capturing the final result of a match, we used the correct classification rate (CC) both computed for the G in-sample units (CC_{in}) and for the G^{test} matches belonging to an out-of-sample set (CC_{out}):

$$CC_{in} = \frac{\sum_{g=1}^G \mathbf{1}_{\{O_g = \hat{O}_g\}}}{G}, \quad CC_{out} = \frac{\sum_{g=1}^{G^{test}} \mathbf{1}_{\{O_g = \hat{O}_g\}}}{G^{test}}$$

and the Brier (1950) score (BS), a popular goodness-of-fit measure for categorical outcomes:

$$BS = \frac{\sum_{g=1}^G \sum_{i=\{1,X,2\}} \left(p_{ig} - \mathbf{1}_{\{O_g=i\}} \right)^2}{G}.$$

Note that $\mathbf{1}_{\{E\}}$ is an indicator function that assumes value 1 when the event E occurs and 0 otherwise.

Lastly, to compare models with the same response variable, information criteria aimed at estimating the point-wise out-of-sample model prediction accuracy are widely used in Bayesian inference. According to Vehtari et al. (2017), an efficient way to estimate this quantity is through the approximate leave-one-out cross-validation information criterion using Pareto-smoothed importance sampling (LOOIC-PSIS). Its computation is implemented in the R package `loo` (Vehtari et al. 2017), and the best model is the one with the smallest LOOIC value.

4 Analysis of UEFA champions league data

The models described in Sect. 3 have been applied to real data regarding the 2016-2017 UCL, including variables described in Sect. 2 as covariates. Data were collected using freely available press kits from the official UEFA website¹. They include 125 matches and 250 passing networks for the $T = 32$ most competitive European teams. We point out that UCL is constituted by two phases:

Group stage: it consists of 8 groups composed of 4 teams (6 matches per team, 12 matches per group) for a total of 96 matches;

Knockout phase: it is composed of Round of 16, Round of 8, Semi-Finals, and Final, for a total of 29 matches. Note that the final is the only one-off match.

The empirical relationships among the selected variables are firstly investigated. Figure 2 summarizes the correlations between couples of quantitative variables measured through the Spearman’s coefficient. From this figure, regarding the in-match covariates, we can notice that *shots on target*, *ball possession*, and *corners* show a positive correlation between them (all values exceed 0.4). On the contrary, *distance* and *fouls suffered* seem not to show any monotonic relationship with other in-match variables. Considering network summary measures, *network intensity* expresses a high positive correlation with both *pass accuracy* (0.8) and *ball possession* (0.6), while *reciprocity* is also positively correlated with all network summary

¹ www.uefa.com

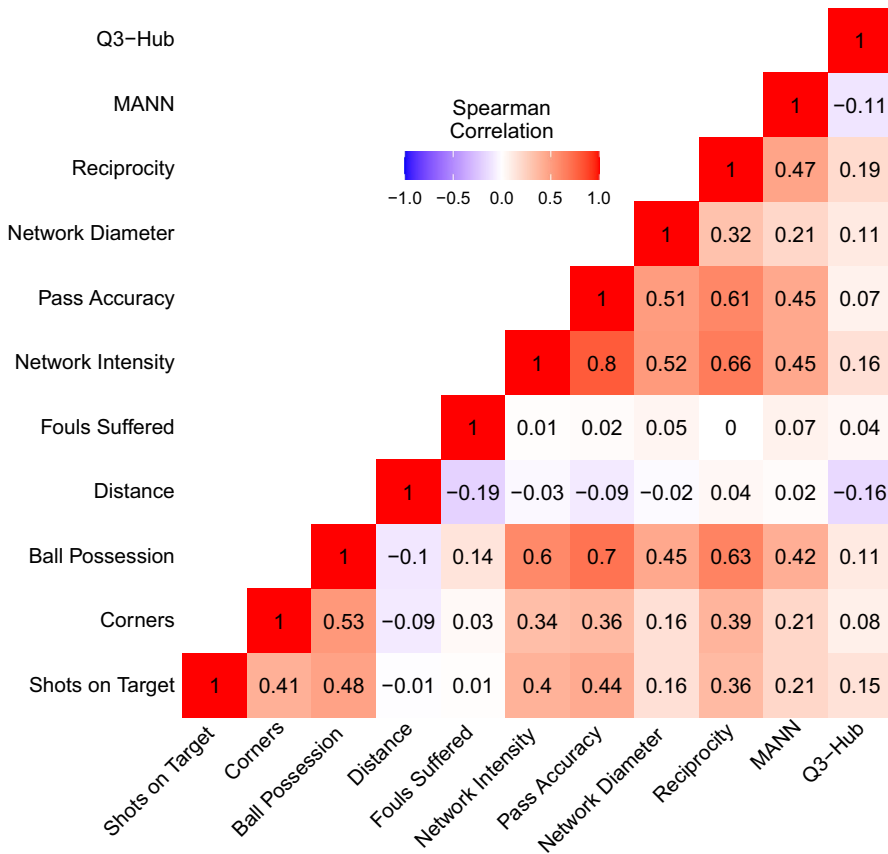


Fig. 2 Correlation Matrix of the explanatory variables based on Spearman's correlation coefficient

measures, presenting high correlation values (greater than 0.6) with *ball possession*, *network intensity*, and *pass accuracy*. The *diameter* also shows remarkable positive correlations with *ball possession*, *network intensity*, and *pass accuracy*, whereas *fouls suffered* and *distance* are generally uncorrelated to overall network summary measures. Thus, we can conclude that quantities regarding physical activities and contacts between players are not mutually related to the precision of offensive actions and the level of passing interactions. Q₃-Hub also presents weak correlations with all variables.

Bayesian models described in Sect. 3 are fitted to the considered data using the two types of football outcomes, i.e., the *scored goals* (y_g^H, y_g^A) and the *difference in goals* (z_g) . As mentioned, we use IP and BP models on the scored goals and the Sk model for the difference in goals, considering four different specifications of the linear predictors for each model previously presented in Subsection 3.2 and summarized in Table 3. The specification of the horseshoe prior introduced in Sect. 3.1 is completed choosing $v_{\text{slab}} = 7$ and $s_{\text{slab}} = 2.5$. The prior number of relevant effects p_0

Table 4 Goodness-of-fit measures for the three Bayesian hierarchical models according to four different covariates specifications

	IP				BP				Sk			
	M_0	M_1	M_2	M_3	M_0	M_1	M_2	M_3	M_0	M_1	M_2	M_3
LOOIC	583	543	522	517	583	543	522	518	397	338	343	337
(S.E.)	(21)	(24)	(19)	(19)	(21)	(23)	(19)	(19)	(21)	(16)	(15)	(14)
CC_{in}	0.56	0.58	0.59	0.60	0.56	0.57	0.59	0.59	0.57	0.64	0.69	0.67
CC_{out}	0.59	0.79	0.79	0.76	0.59	0.59	0.79	0.76	0.55	0.69	0.62	0.69
BS	0.53	0.50	0.50	0.50	0.53	0.51	0.50	0.50	0.48	0.47	0.46	0.47

Best values for each model are depicted in bold

was fixed equal to 3, using the information from a preliminary analysis using the Bayesian LASSO (Park and Casella 2008). Posterior inference is carried out on 12000 MCMC replicates that are obtained from 4 parallel chains using 6000 iterations for each. The first 3000 iterations of each chain are discarded as a warm-up period.

The convergence of the MCMC algorithm is carefully checked by visual inspection, monitoring the posterior effective sample sizes, and computing the Gelman-Rubin statistic. We remark that a tutorial concerning the estimation of considered models on our data is provided as supplementary material.

To assess the reliability of our analysis, we check the fitted models' performance both inside and outside the sample: matches coming from the *group stage* ($G = 96$) are used as the training set, the remaining $G^{test} = 29$ (*knockout phase*) ones constitute the test set. Although prediction is not our primary aim, this splitting procedure is useful to understand the consistency of the statistical relationships between the outcome and the set of covariates captured by the model.

Table 4 shows the performances of three different models according to the four goodness-of-fit indicators (LOOIC, CC_{in} , CC_{out} , and BS) presented in Subsection 3.3. We remark that a proper comparison between all the three models can not be carried out through the LOOIC value, since Sk is specified on a different response variable with respect to IP and BP.

As expected, the addition of covariates improves all the performance indicators for each model, confirming the usefulness of the available variables set. This can be immediately checked by observing the decreasing of LOOIC values from M_0 to other linear predictors specifications. Considering also the corresponding standard error (S.E.), the IP and BP models have remarkably lower LOOIC values with formulations M_2 and M_3 . On the other hand, Sk_1 and Sk_3 appears preferable with respect to Sk_0 and Sk_2 . Comparing the three regression models in terms of predictive abilities of the match outcome, the best CC_{in} can be found for models Sk_2 and Sk_3 , noting that the Sk model with auxiliary information dominates IP and BP models in the CC_{in} values. On the other hand, the CC_{out} indicator is higher in models with Poisson likelihoods. The mismatch between CC_{in} and CC_{out} values for some models can be due to UCL-specific features. In fact, a *class imbalance* between training and

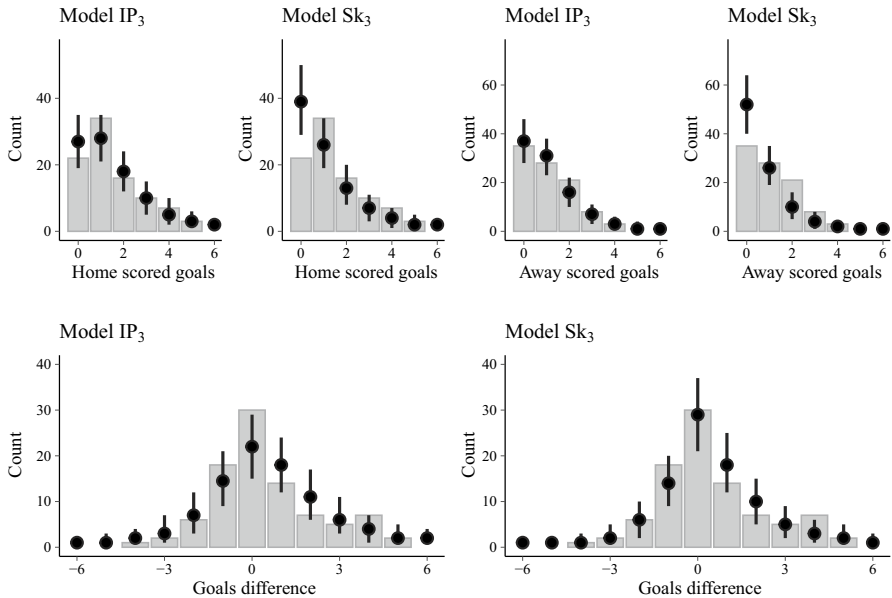


Fig. 3 Comparison of empirical data (histograms) with the posterior predictive distributions. The bold lines represent the 90% uncertainty intervals of counts around the medians

test set is observed: the draw occurrence in the knockout phase is largely lower than in the group stage (10.3% vs. 31.3%).

This comparison helps to understand which specification of the linear predictor is preferable. We decided to select one model for each football outcome, in order to clarify the implications of proposed statistical analysis on football. Since the addition of a correlation term in BP models seems not to improve their performances, IP_3 is chosen to model the team scored goals, preferring its parsimony compared to IP_2 . On the other hand, Sk_3 is selected as model for the goals difference outcome, since it performs similarly to Sk_1 , but the interpretation of the coefficients is more intuitive.

After selecting the final models, we focus on their ability to fit the data exploiting the replicated datasets generated from the posterior predictive distributions and summarizing the results with the tools provided by the `bayesplot` package (Gabry and Mahr 2021). In the plots of Fig. 3, histograms and the 90% uncertainty intervals, centered on the median related to the simulated datasets under models Sk_3 and IP_3 , are reported. The top line refers to the distributions of goals scored by home and away teams, whereas the bottom line concerns the goals difference sample distribution. We highlight that in the case of the Skellam model, the goals distributions are generated assuming independent Poisson distributions.

The skewness of the goals difference distribution emerges from Fig. 3: the winning of home teams is usually larger compared to the away teams winning. Considering this outcome, the Sk_3 model shows its good performance in predicting the draw (i.e., a difference in goals equal to zero), while the IP_3 model tends to underestimate the probability that draws occur. This is a known feature of such

Table 5 Posterior means and 90% credible intervals (C.I.) for models parameters

Parameter/Variable	Sk ₃		IP ₃	
	Post. mean	90% C.I.	Post. mean	90% C.I.
σ_α	0.161	[0.016; 0.372]	0.111	[0.011; 0.259]
σ_δ	0.160	[0.012; 0.412]	0.081	[0.007; 0.201]
μ	-0.466	[-0.839; -0.109]	0.046	[-0.124; 0.213]
h	0.276	[0.015; 0.548]	0.186	[-0.028; 0.400]
Shots on target	0.384	[0.292; 0.477]	0.304	[0.239; 0.369]
Corners	-0.109	[-0.233; 0.000]	-0.050	[-0.133; 0.004]
Ball possession	0.099	[-0.006; 0.256]	0.011	[-0.028; 0.074]
Distance	-0.042	[-0.264; 0.068]	-0.016	[-0.120; 0.042]
Fouls suffered	-0.023	[-0.113; 0.033]	-0.023	[-0.090; 0.015]
Network intensity	0.036	[-0.033; 0.160]	0.058	[-0.007; 0.161]
Pass accuracy	0.014	[-0.081; 0.136]	0.023	[-0.026; 0.114]
Network diameter	-0.001	[-0.068; 0.064]	-0.002	[-0.049; 0.040]
Reciprocity	-0.010	[-0.110; 0.058]	0.009	[-0.035; 0.072]
MANN	-0.069	[-0.194; 0.009]	-0.013	[-0.074; 0.022]
Q ₃ -Hub	0.029	[-0.020; 0.110]	0.011	[-0.019; 0.059]

models, and a possible way to face this issue is to introduce a further correlation structure in the goal-based models to inflate the draw probability (Dixon and Coles 1997; Rue and Salvesen 2000). On the other hand, observing plots in the top line, we deduce that models based on the Skellam distribution achieve the results overestimating the event of a zero goal scored by a team.

Table 5 summarizes the results of the two selected models in terms of posterior means and 90% credible intervals (C.I.) for their parameters and the regression coefficients, remarking that standardized covariates are included in the model. Focusing on the determinants of football outcome, we consider the posterior probability that a coefficient is higher or lower than 0 as a measure of importance for the covariate. In particular, *shots on target* is the most relevant variable for both models, showing a large positive effect on both scored goals and goals difference. Surprisingly, *corners* are also relevant, but show a negative impact on the two outcomes ($\mathbb{P}[\beta_2 < 0|y] = 0.89$ for IP₃ and $\mathbb{P}[\beta_2 < 0|z] = 0.95$ for Sk₃). The third variable appearing relevant in IP₃ is the *network intensity* ($\mathbb{P}[\beta_6 > 0|y] = 0.87$). Then, focusing on Sk₃, the *ball possession* has a positive impact on the outcome ($\mathbb{P}[\beta_3 > 0|z] = 0.90$), whereas the corresponding coefficient is shrunk toward 0 for IP₃. Lastly, in Sk₃, a negative coefficient can be found for the *MANN* indicator ($\mathbb{P}[\beta_{10} < 0|z] = 0.86$). Observing the remaining estimated coefficients, the other covariates share a lower relevance in explaining the football outcomes.

To conclude, we highlight that IP₃ is also characterized by lower values for the random effects scales (both for attack and defense) and home effect parameter h .

5 Discussion

In this section we propose a football oriented interpretation of the previous results. In particular, Subsection 5.1 aims at investigating the meaning of the fitted models, whereas Subsection 5.2 extrapolates useful indications starting from the estimated model coefficients.

5.1 Lessons from empirical models

In the previous section, we reported the general results of the two selected models. Here, we point out some features emerging from the comparison between different fitted models, discussing how covariates and model specifications impact in this applied field.

Let us consider the features of the analyzed linear predictors. Some useful empirical indications about model formulations can be deduced from the results contained in Table 4. From a predictive point of view, the outcomes of the model with covariates are comparable, but LOOIC suggests that the best trade-off between accuracy and parsimony is M_3 . Therefore, when the outcome (y_g^H, y_g^A) is analyzed, the usage of specific team covariates is preferable. On the other hand, possible differences on the pitch between the home and the away teams are conveyed through the covariate information, hence the parsimony induced by the use of the same vector β is favored. In parallel, the correlation term in equation (4), introduced with the bivariate Poisson distribution, does not apparently improve the model performance in fitting our data.

Our results suggest that the Skellam model represents an interesting trade-off between the result-based models and the pure goal-based ones. It takes into account the gap in terms of scored goals, avoiding to discard possible information available on the pitch, as happens in multinomial models for the final match result. Additionally, the fact that the draw constitutes a value of the response variable leads to higher performance in prediction. Despite the good in-sample behavior of Sk models, they are sensitive to the changes in the results distribution occurring between the group stage and the knock-out phase (see, e.g., the out-of-sample indicators).

Regarding the role of covariates, interesting hints can be deduced from Fig. 4, in which the posterior means of the random effects under models IP_0 and IP_3 are compared. We note that the auxiliary information is able to explain a large portion of data variability, that in the case of model IP_0 is absorbed by the random effects α and δ ($\sigma_\alpha = 0.402$ and $\sigma_\delta = 0.424$ under IP_0 ; $\sigma_\alpha = 0.111$ and $\sigma_\delta = 0.081$ under IP_3). This feature is encouraging for several reasons. Firstly, the fitted model appears to be correctly specified and the fixed effects well capture the scored goals behavior. Secondly, the included explanatory variables indirectly incorporate also the defensive ability of the opponent team. In addition, also the home effect h tends to be less important in models with covariates: its posterior mean under IP_0 is 0.391 (90% C.I.: [0.191; 0.593]), compared to 0.186 for IP_3 (90% C.I.: [-0.028; 0.400]). The fact that *shots on target* result the most relevant covariate can be analyzed recalling that they

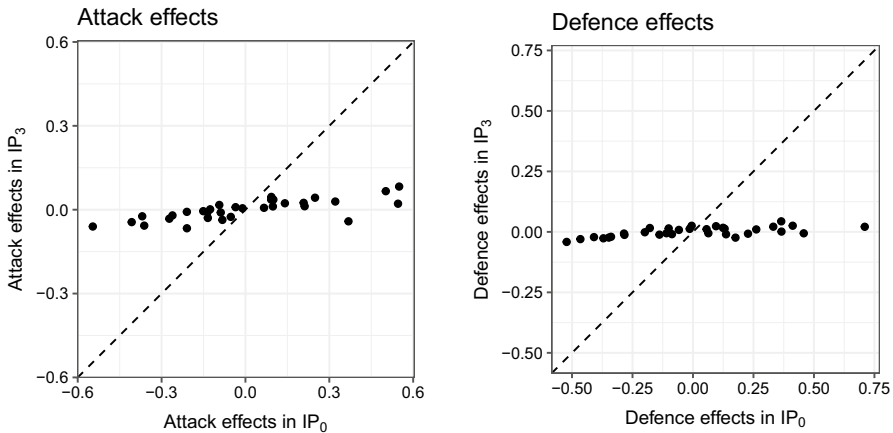


Fig. 4 Comparisons of the attack random effects α and defense random effects δ under models IP_0 and IP_3

represent necessary events in order to score a goal. For this reason, it might be interesting to develop a modeling framework in which this information is included as an offset in a Poisson regression, to study how the remaining in-match covariates and network-based measures influence the *relative risk* of scoring a goal.

5.2 Implications in football

Undoubtedly results from statistical models can be affected by the peculiarities and rules of the analyzed football tournament. Throughout UCL, away goals are very relevant because of the “double value” rule, possibly helping a team to progress beyond the *group stage* or move forward in the *knockout phase*. As evidence of this, 40% of goals are scored by the away teams along the whole competition. In the group stage, draws are frequent (31%) and the away team wins are not rare (29%). In the knockout phase, a change in the distribution of the results is observed: home team wins is the dominant outcome (64.3%), and fewer draws occur (10.7%).

The precision of offensive actions is crucial to score goals, or score more goals than the opponent. A competitive European team has the burden of finding and buying players able to become very accurate in shooting, i.e., with a high propensity to score goals. According to our data, the winning home teams present a median of 7 *shots on target*, while this number decreases to 5 if we consider the away teams able to win. Therefore, home and away teams share the same median in terms of *shots on target* (equal to 3) when they lose the match. This finding is crucial considering that in UCL data the ratio of scored goals over shots on target is equal to 33.4%, where the home teams are slightly more precise than the away ones (36% vs. 31%).

More surprisingly, our results show that collecting corner kicks may have a negative impact on the football outcome. Although it is not easy to give a unique interpretation for this finding, the corner kicks can be viewed as a result of the offensive actions and are positively correlated with the *shots on target* (see, e.g., Fig. 2).

Moreover, a corner kick can be caused both by an error in shot of a forward or by an effective defensive behavior of the opposing team (e.g., a good save by the goalkeeper or a defensive stop), avoiding the goal. In addition, following this second point of view, the collection of too many corners can be harmful and may negatively affect the effectiveness of the offensive actions. Not all teams have the characteristics to take advantage from this particular game situation and, consequently, to earn additional shots on goal. To give an example, in the four matches with the highest difference in corners between home and away teams, the home one lost by a narrow margin or draw. The number of corners is not correlated (in terms of Spearman's coefficient) with the number of goals.

The importance of *network intensity* suggests that a competitive team should include, in its roster, players capable of making many passes per minute, even using more “first-time” passes. Hyballa and Te Poel (2015), in their book “*Passing Drills*”, emphasize the need of “*fast, precise, variable, and creative passing*”. This indicator is especially relevant to win the away matches: we observe that the 36 winning away teams present a median of network intensity equal to 9.2, outperforming their opponents by two passes per minute (median equal to 7.2). Conversely, in the 33 draws, the median of *network intensity* is slightly higher for the home teams (8.0 vs 7.6), while the winning home teams express a network intensity of 8.3 (the median of losing away teams is equal to 7.4). According to the results in Fig. 2, a team able to increase its passing speed usually expresses an high passes *accuracy* and dominates in terms of *ball possession*. The positive effect of *network intensity* is also stressed by an observed positive correlation with the number of goals (0.32).

Concerning the Sk_3 model, *ball possession* has an impact on the goals difference, adding value to the positive correlation (0.38) we found between these two quantities. In addition, if the *ball possession* of the home team is almost 20% higher than the opponent's one, the home team scores two goals more than the opponent, in median. Conversely, if a difference of almost 20% in *ball possession* is in favor of the away team, they score one goal more than the home team, in median.

6 Conclusion

In this work, we firstly explored how Bayesian hierarchical models are suitable in finding the determinants of football outcomes, using auxiliary information only available from the pitch. Exploiting the widespread computational tools that are nowadays available to fit Bayesian models, alternative model specifications are proposed and compared in order to provide useful indications regarding both in-match and network-based variables. In particular, we take advantage from the flexibility of Bayesian modeling to include auxiliary information.

To summarize the main results, the effectiveness of an offensive action is crucial to determine the football outcome, but variables such as the passing speed (number of passes in the temporal unit) can improve the propensity of scoring goals or more goals than the opponent. Potential over-fitting issues are mitigated by using the regularized horseshoe prior for the regression coefficients. This feature leads us to models with good fitting results also for observations outside the training set. In fact, the

relationships between covariates and the football outcomes appear robust in terms of portability of the results, and might represent also an important tool for forecasting. To investigate this aspect, the application of these models to larger competitions, such as national leagues, might be appealing, keeping in mind the peculiarities of the competition on the fore. In these cases, a large volume of past information is available and it can be exploited to produce summary measures for covariates (such as means or medians) that can be used to forecast the result of a football match, without using covariates from the game itself. The inclusion of network summary measures will be a key point in further models, even in a predictive perspective.

In this paper, we also discuss how the structural passing network features can be informative for football teams' staff, managers, and match analysts. Passing speed, balance in terms of passing directions, team cohesion can be reasonable determinants for the football outcome.

Possible improvements on the network measures should consider spatial information on the pitch, such as the positions of players and other events as the shots on target. For example, qualitative attributes of the connection, such as long versus short passes, measures related to the evolution of the passing structure during the match, and interactions between players of two opposite teams could also provide more information about the overall match-level performances.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10182-021-00411-x>.

Funding Open access funding provided by Università degli Studi di Ferrara within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Albert, J., Bennett, Y., Cochran, J.J.: *Anthology of Statistics in Sports*. SIAM, Philadelphia (2005)
- Baio, G., Blangiardo, M.: Bayesian hierarchical model for the prediction of football results. *J. Appl. Stat.* **37**(2), 253–264 (2010)
- Bhadra, A., Datta, J., Polson, N.G., Willard, B.: LASSO meets horseshoe: a survey. *Stat. Sci.* **34**(3), 405–427 (2019)
- Braham, C., Small, M.: Complex networks untangle competitive advantage in australian football. *Chaos* **28**(5), 053105 (2018)
- Brier, G.W.: Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**(1), 1–3 (1950)
- Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P.: Discovering the drivers of football match outcomes with data mining. *Qual. Technol. Quant. M.* **12**(4), 561–577 (2015)

- Carpita, M., Ciavolino, E., Pasca, P.: Exploring and modelling team performances of the kaggle European soccer database. *Stat. Model.* **19**(1), 74–101 (2019)
- Carrington, P.J., Scott, J., Wasserman, S.: *Models and methods in social network analysis*, vol. 28. Cambridge University Press, Cambridge (2005)
- Castellano, J., Casamichana, D., Lago, C.: The use of match statistics that discriminate between successful and unsuccessful soccer teams. *J. Hum. Kinet.* **31**(1), 137–147 (2012)
- Cintia, P., Rinzivillo, S., Pappalardo, L.: A network-based approach to evaluate the performance of football teams. In: *Machine Learning and Data Mining for Sports Analytics Workshop*. Porto, Portugal (2015)
- Clemente, F.M., Couceiro, M.S., Martins, F.M.L., Mendes, R.S.: Using network metrics in soccer: a macro-analysis. *J. Hum. Kinet.* **45**(1), 123–134 (2015)
- Clemente, F.M., Martins, F.M.L., Mendes, R.S., et al.: *Social Network Analysis Applied to Team Sports Analysis*. Springer, New York (2016)
- Clemente, F.M., Sarmiento, H., Aquino, R.: Player position relationships with centrality in the passing network of world cup soccer teams: Win/loss match comparisons. *Chaos Soliton. Fract.* **133**, 109625 (2020)
- Csardi, G., Nepusz, T.: The igraph software package for complex network research. *InterJ. Complex Syst.* **1695**, 1–9 (2006)
- De Nooy, W., Mrvar, A., Batagelj, V.: *Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software*, vol. 46. Cambridge University Press, Cambridge (2018)
- Diquigiovanni, J., Scarpa, B.: Analysis of association football playing styles: an innovative method to cluster networks. *Stat. Model.* **19**(1), 28–54 (2019)
- Dixon, M.J., Coles, S.G.: Modelling association football scores and inefficiencies in the football betting market. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **46**(2), 265–280 (1997)
- Egidi, L., Torelli, N.: Comparing goal-based and result-based approaches in modelling football outcomes. *Soc. Indic. Res.* 1–13 (2020)
- Gabry, J., Mahr, T.: bayesplot: plotting for Bayesian models. R package version 1.8.0. (2021). <https://mc-stan.org/bayesplot/>
- Gelman, A., et al.: Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal.* **1**(3), 515–534 (2006)
- Gonçalves, B., Coutinho, D., Santos, S., Lago-Penas, C., Jiménez, S., Sampaio, J.: Exploring team passing networks and player movement dynamics in youth association football. *PLoS ONE* **12**(1), e0171156 (2017)
- Groll, A., Schaubberger, G., Tutz, G.: Prediction of major international soccer tournaments based on team-specific regularized poisson regression: an application to the fifa world cup 2014. *J. Quant. Anal. Sport.* **11**(2), 97–115 (2015)
- Groll, A., Kneib, T., Mayr, A., Schaubberger, G.: On the dependency of soccer scores—a sparse bivariate poisson model for the uefa european football championship 2016. *J. Quant. Anal. Sport.* **14**(2), 65–79 (2018)
- Grund, T.U.: Network structure and team performance: The case of english premier league soccer teams. *Soc. Netw.* **34**(4), 682–690 (2012)
- Hyballa, P., Te Poel, H.D.: *German soccer passing drills: more than 100 drills from the Pros*. Meyer & Meyer Verlag (2015)
- Ichinose, G., Tsuchiya, T., Watanabe, S.: Robustness of football passing networks against continuous node and link removals. *Chaos Soliton. Fract.* **147**, 110973 (2021)
- Ilevoli, R., Palazzo, L., Ragozini, G.: On the use of passing network indicators to predict football outcomes. *Knowl. Based Syst.* **222**, 106997 (2021)
- Karlis, D., Ntzoufras, I.: Analysis of sports data by using bivariate poisson models. *J. R. Stat. Soc. Ser. D. Stat.* **52**(3), 381–393 (2003)
- Karlis, D., Ntzoufras, I.: Bayesian modelling of football outcomes: using the Skellam’s distribution for the goal difference. *IMA J. Manag. Math.* **20**(2), 133–145 (2009)
- Kleinberg, J.M., Newman, M., Barabási, A.L., Watts, D.J.: *Authoritative sources in a hyperlinked environment*. Princeton University Press, Princeton (2011)
- Lepschy, H., Wäsche, H., Woll, A.: Success factors in football: an analysis of the german bundesliga. *Int. J. Perf. Anal. Spor.* **20**(2), 150–164 (2020)
- Maher, M.J.: Modelling association football scores. *Stat. Neerl.* **36**(3), 109–118 (1982)
- Manderson, A., Murray, K., Turlach, B.: Dynamic Bayesian forecasting of afl match results using the Skellam distribution. *Aust. N. Z. J. Stat.* **60**(2), 174–187 (2018)

- McLean, S., Salmon, P.M., Gorman, A.D., Stevens, N.J., Solomon, C.: A social network analysis of the goal scoring passing networks of the 2016 European football championships. *Hum. Mov. Sci.* **57**, 400–408 (2018)
- Memmert, D.: Data analytics in football: positional data collection, modeling, and analysis. *J. Sport Manag.* **33**, 574 (2019)
- Park, T., Casella, G.: The Bayesian LASSO. *J. Am. Stat. Assoc.* **103**(482), 681–686 (2008)
- Pena, J.L., Touchette, H.: A network theory analysis of football strategies. (2012).arXiv preprint arXiv:1206.6904
- Piironen, J., Vehtari, A.: Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron. J. Stat.* **11**(2), 5018–5051 (2017)
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (2020). URL <http://www.R-project.org/>
- Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.*, 1151–1172 (1984)
- Rue, H., Salvesen, O.: Prediction and retrospective analysis of soccer matches in a league. *J. R. Stat. Soc. Ser. D. Stat.* **49**(3), 399–418 (2000)
- Schauberger, G., Groll, A., Tutz, G.: Analysis of the importance of on-field covariates in the German Bundesliga. *J. Appl. Stat.* **45**(9), 1561–1578 (2018)
- Stan Development Team. Stan modeling language users guide and reference manual, 2.25 (2020). URL <https://mc-stan.org>
- Tutz, G., Schauburger, G.: Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Adv. Stat. Anal.* **99**(2), 209–227 (2015)
- Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Stat. Comput.* **27**(5), 1413–1432 (2017)
- Wasserman, S.: *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*. Sage, Thousand Oaks (1994)
- Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge (1994)

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