Statistical Modelling 2019; **19(1)**: 74–101

Exploring and modelling team performances of the Kaggle European Soccer database

Maurizio Carpita[1], Enrico Ciavolino[2] and Paola Pasca[2]
[1]Department of Economics and Management, University of Brescia, Brescia, Italy.
[2]Department of History, Society and Human Studies, University of Salento, Lecce, Italy.

Abstract: This study explores a big and open database of soccer leagues in 10 European countries. Data related to players, teams and matches covering seven seasons (from 2009/2010 to 2015/2016) were retrieved from Kaggle, an online platform in which big data are available for predictive modelling and analytics competition among data scientists. Based on both preliminary data analysis, experts' evaluation and players' position on the football pitch, role-based indicators of teams' performance have been built and used to estimate the win probability of the home team with the binomial logistic regression (BLR) model that has been extended including the *ELO* rating predictor and two random effects due to the hierarchical structure of the dataset.

The predictive power of the BLR model and its extensions has been compared with the one of other statistical modelling approaches (Random Forest, Neural Network, k-NN, Naïve Bayes). Results showed that role-based indicators substantially improved the performance of all the models used in both this work and in previous works available on Kaggle. The base BLR model increased prediction accuracy by 10 percentage points, and showed the importance of defence performances, especially in the last seasons. Inclusion of both *ELO* rating predictor and the random effects did not substantially improve prediction, as the simpler BLR model performed equally good. With respect to the other models, only Naïve Bayes showed more balanced results in predicting both win and no-win of the home team.

Key words: Kaggle European Soccer (KES) database, binomial logistic regression (BLR) model, role-based player performance indicators, prediction of match results, comparison of classification models, statistical learning models

1 Introduction

In recent years, a substantial growth in the use of statistical models in the sport field occurred. Bookmakers got more and more interested in fine-tuning the forecast of match outcomes (Odachowski and Grekow, 2013), teams want to gather an in-depth view of player's characteristics and performance in order to both enhance it and to potentially buy new players (Ahuja et al., 2017; Zelenkov and Solntsev, 2017; Carling et al., 2005). Researchers committed to the insightful collection of the most

Address for correspondence: Maurizio Carpita, University of Brescia, Piazza del Mercato, 15, 25121 Brescia, Italy.
E-mail: maurizio.carpita@unibs.it.

 10.1177/1471082X18810971

significant statistical applications to a wide range of sports (Albert et al., 2005, 2017): Forefront data analysis techniques have been employed in order to discover what really determines a good performance of players (Šlaton, 2012) or would be helpful in constructing players' perfomance indexes (McHale et al., 2012; McHale and Szczepański, 2014), and in finding the best way to interpret and predict match outcomes (Carpita et al., 2014; Leung and Joseph, 2014; Liti et al., 2017).

There are two distinct strands of empirical literature on modelling and forecasting football match results (McHale and Scarf, 2007; Dobson and Goddard, 2011), depending on what their focus is and on the statistical methodology used: a direct approach, in which regression-based models are usually employed to predict an ordered outcome variable such as win–draw–loss (Koning, 2000), and an indirect approach which rather focuses on the distribution of goals scored by teams through models based on probability distributions, such as Poisson's; this last approach allows to draw inferences about the most likely outcome in terms of result or exact score (Karlis and Ntzoufras, 2003) and to efficiently handle some aspects concerning the distribution of sports data, as the excess of draws (Karlis and Ntzoufras, 2009). With respect to forecasting capabilities, results-based models might be expected to outperform the goals-based models on the grounds that the model selection and specification issues are more straightforward (Dobson and Goddard, 2011, p. 80), but in some applications none of them seems to be better (Goddard, 2005). The present study makes use of the results-based modelling approach, by enriching its simple nature with information about players' performance and their role on the soccer pitch.

In parallel to the progress of methodologies in sports, online platforms for predictive modelling and analytics competitions such as Kaggle (*www.kaggle.com*) emerged, representing a meeting place for data scientists. This online platform offers a wide variety of freely shared big data, so that statisticians and data miners can take part in competitions aimed at producing the best descriptive or predictive model. Our exploration and statistical modelling focuses on a European soccer (football) dataset, due to both its worldwide fame, to the fact that this sport lends itself to many different statistical techniques (Stern, 2005) and to the lack of team's perfomance indicators for the prediction of match outcomes. In this context, the Kaggle European Soccer (KES) database cointains data about 28 000 players and about 21 000 matches of the championship leagues of 10 countries and 7 seasons from 2009/2010 to 2015/2016. To the best of our knowledge, the KES database is the biggest open database devoted to the soccer leagues of European countries.

The main purpose of this study is to test the power of *experts*' evaluation of players' performance by role in predicting home team winning the match. In doing so, an original three-step approach has been adopted:

1. based on 33 player-related variables, 7 players' performance indicators have been built using experts' classification of the EA Sports FIFA videogame;
2. players' performance indicators have been further combined in order to reflect coach's decisions before each match takes place, in terms of

players' position or role on the pitch (*forward*, *midfielder*, *defender* and *goalkeeper*);
3. differences in role-based players' performance indicators between home and away team are used as predictors in a statistical model to assess how much they affect the probability of the home team winning the match.

In particular, role-based indicators of teams' performance can be useful from the interpretative point of view, as they are related to the strategic choices of the coach of each team. With respect to 3, a classical way to model soccer data focuses on the measure of teams' strength viewed as a latent variable, so that the observed result of a match is determined by this latent variable. In statistics, models based on this approach are known as paired comparison models and the most famous one is the Bradley–Terry (BT) model (Tutz and Schauberger, 2015). In its original specification, the probability that one team beats the opponent in a match only depends on the difference between the strength parameters of each of the two teams; the BT model can be extended in order to include both the possible results (win, draw and loss) and home team's advantage via *home effect* parameter. The original BT model allows for teams' strength estimation and ranking as well as for clustering teams; however, it does not explain why some teams are better than others. A standard way to explain the variation in performance is to include the difference in covariates between the two teams in the model (Tutz and Schauberger, 2015); in more complex models, different parameters for both the covariates of teams and matches can be specified (Cattelan et al., 2013; Schauberger et al., 2016). As the focus of our study is to assess whether team's performance indicators are able to predict the win of home team in a generic match, rather than estimating the strength of each team, a simple binomial logistic regression (BLR) model with only teams' difference in covariates has been adopted, assuming therefore that these predictors would capture the main effects on the result of interest (home team win). The BLR model, which largely used in previous studies (Alves et al., 2011; Magel and Melnykov, 2014) and not only in the soccer context (Lisi and Zanella, 2017), has been extended by adding the *ELO* rating as predictor and two random effects that take into account the hierarchical structure of data (matches within seasons, seasons within countries).

The article is organized as follows: In Section 2, we present the KES database and explain the building process of the role-based indicators of team's performance; Section 3 gives an introduction to the BLR model for estimating the win probability of the home team, with subsections devoted to model fit and model prediction; in Section 4, estimation results of the BLR model applied to KES database are presented and discussed for the whole and the by-country datasets; in Section 5, a comparison between the two extensions of the base BLR model and other competing models (Random Forest, Neural Network, k-NN, Naïve Bayes) for prediction of match results is carried out; in the last section, conclusions are reported. All computations have been performed using the R statistical computing software, ver. 3.4.3 (R Core Team, 2014).

2 The KES database and the role-based indicators of teams' performance

The full original structure of the KES database is described in the Appendix and is based on two main tables. The *Match_table* contains the date, the positions (X and Y coordinates) on the pitch for the 22 players of the two (home and away) teams and the final result (home and away team goals) of each match. The *Player_Attributes_table* contains other 33 variables, with periodic player's performance on a 0–100 scale with respect to different abilities. In this study, data belong to 10 countries/leagues (Belgium, England, France, Germany, Italy, Netherlands, Portugal, Scotland, Spain and Switzerland) and to 7 seasons (from 2009–2010 to 2015–2016). See Tables A1 and A2 in the Appendix for details and statistics.

The left-plot of Figure 1 shows the average number of observations available for each player within the seven seasons (grey points represent the considered countries/seasons). As it can be noted, the average number of observations available per player is just below 2 only in the first three seasons. This happens because, until January 2013, players' performance indicators are available only twice a year (February and August), while they have a monthly availability in the remaining seasons (precisely, after January 2013). However, since in the latter seasons the average number of observations per player is lower than 6, each player's performance update cannot still be considered monthly.

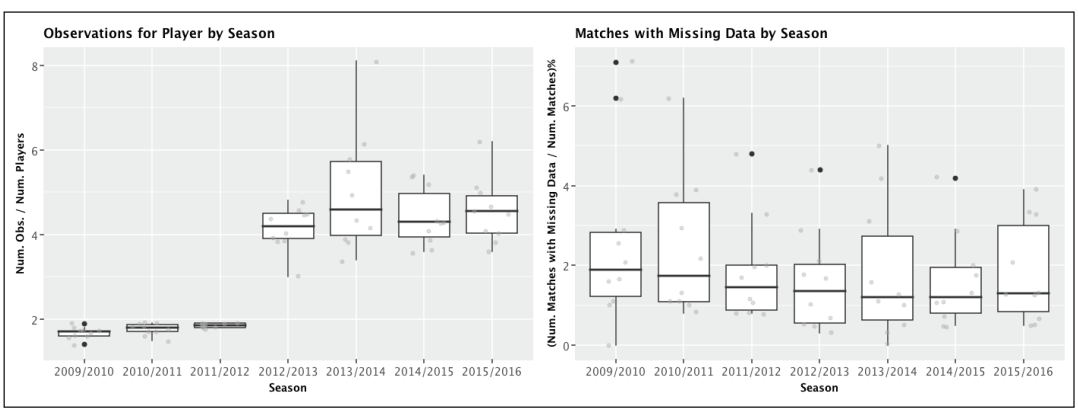


Figure 1 Mean observations for player by season (left) and matches with missing data by season (right)
Note: Grey points represent countries used to draw the box plots.

Plot on the right side of Figure 1 shows the average number of matches with missing data for the seven seasons (grey points represent the considered countries/seasons). The percentage of matches with missing data are rather low, on average lower that 2%.

In order to reduce the number of the predictors in the model, we adopted the *sofifa.com* classification. In this famous website, largely used by soccer fans (2.37 million of links in Google on February 2018), 33 players' performance variables

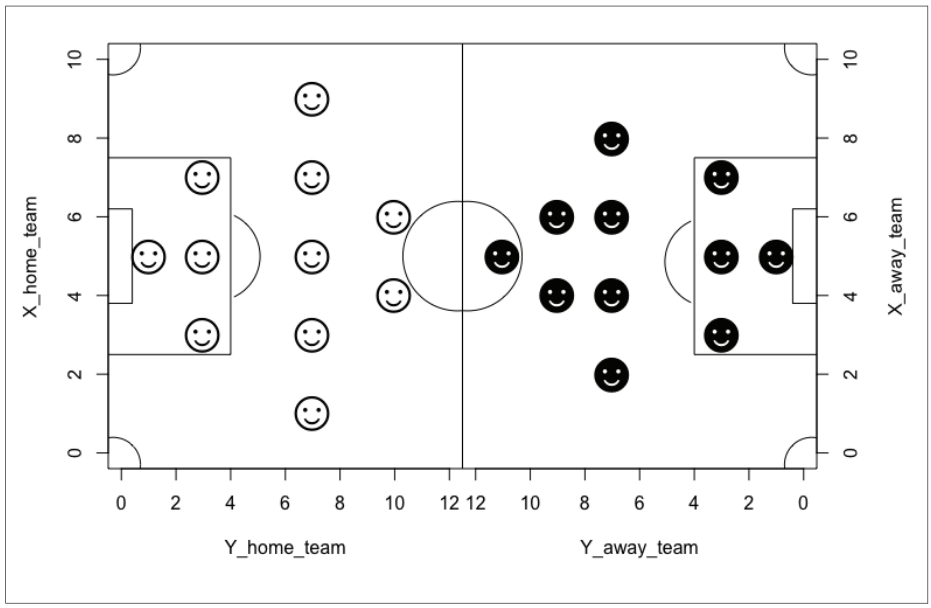


Figure 2 Example of team line-up with coordinates X and Y of the match Juventus (home team) versus Rome (away team), Turin (Italy), 24 January 2016

in the *Player_Attributes_table* of the KES database have been grouped by experts into 7 dimensions: *power*, *mentality*, *skill*, *movement*, *attacking*, *defending* and *goalkeeping*. For each dimension, the simple average of its variables has been computed to obtain the corresponding performance indicator.

Player's role in each match has been identified through X and Y coordinates of players' position on the soccer pitch in the *Match_table* of the KES database. For illustration purposes, Figure 2 shows the line-up of Juventus (home team) and Rome (away team) of the match held in Turin (Italy) on 24 January 2016. Coordinates X (1–9) and Y (1, 3, 5–11) position the players in the pitch according to their role. Clearly, the role of a player is defined by the Y coordinate: Thus, the role has been assigned to each player according to the following rules:

- if Y = 1, then player's role is *goalkeeper*;
- if Y = 3, then player's role is *defender*;
- if $5 \leq Y \leq 9$, then player's role is *midfielder*;
- if $10 \leq Y \leq 11$, then player's role is *forward*.

A correlation network between the seven players' performance indicators (first three letters of their name in capital) and four players' roles (in capital) is reported in Figure 3: The thicker and darker the line is, the stronger and more significant is the positive correlation.

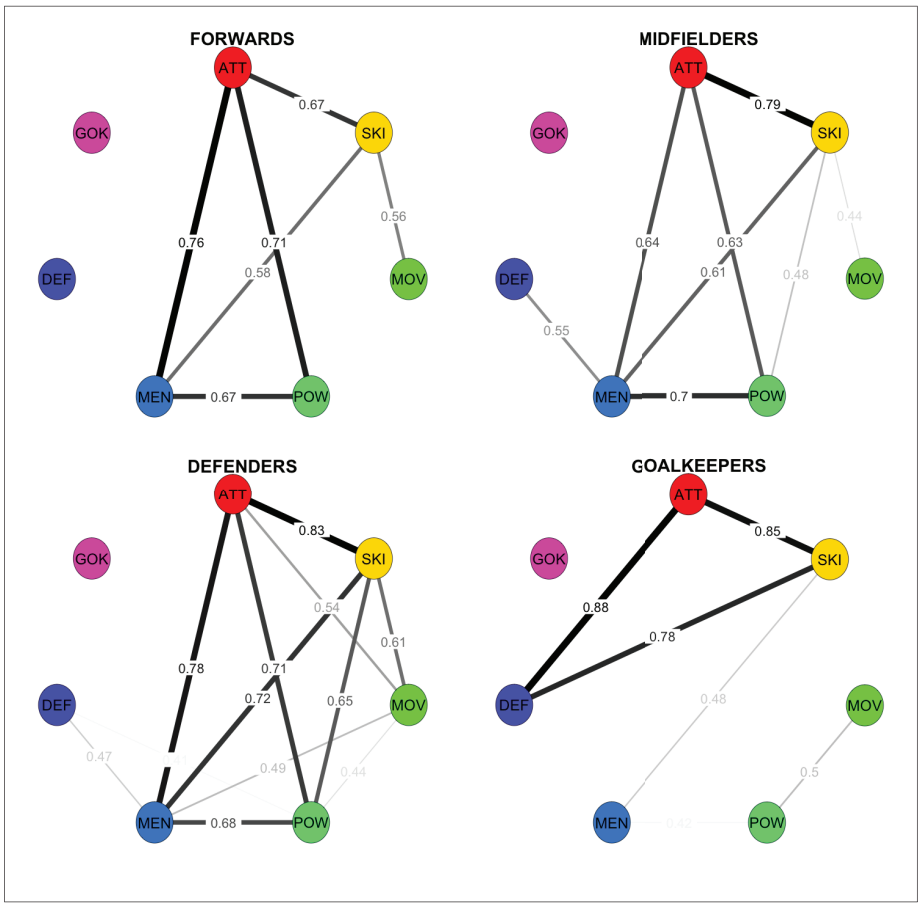


Figure 3 Correlation networks between the seven role-based indicators of players' performance. Lines are correlations greater than +0.4 (thicker lines are higher correlations)

It is interesting to note how strong correlations vary according to players' role. Except for *goalkeepers*, *skill*, *power* and *mentality* are highly correlated to the *attacking* ability: The latter shows a stronger correlation with *mentality* and *power* for the *forwards*, while a stronger correlation with the *skill* dimension can be observed as players' position shifts back towards the *goalkeeper* position (0.79 for *midfielders* and 0.83 for *defenders*); these high correlations can create *multicollinearity* problems for the model, which is an aspect we considered in Section 4.1. In general, *defenders*' role shows the highest number of correlations: Positive values greater than 0.4 can be observed between all but the *goalkeeping* role. The latter is characterized by three dimensions highly correlated with each other: *attacking*, *skill* and *defending*; The reason of such high positive correlations among the four players' performance indicators is that they are average variables with very low range scores, thus not

being coherent with players' role (see var.long.name in Table A2 of the Appendix). For this reason, we did not consider some performance indicators as predictors in the *goalkeepers* position in model specification of Section 3.

The link between performance indicators of *Player_Attributes table* and players' role of the home and away teams in the *Match table* was performed by using *player identification code* and *reference date*. With respect to this second key, the *rolling join* procedure of the R package `data table` was used to link the performance indicators of the immediately previous date *before* the match.

Finally, the average of players' performance indicators grouped by players' roles has been computed for each match for both home and away teams. These role-based indicators of players' performance have been used as predictors for the match result in the model presented in the next section. In this study, the match result is the difference between home and away team goals and is classified into three categories: win (if the difference is positive), loss (if negative) and draw (if null) of the home team. The relative frequencies for win range from 40% to 51%, while they range from 26% to 34% and 21% to 29% for loss and draw, respectively (see Table A1 in the Appendix).

Our preliminary attempts to predict the three match results revealed a substantially different performance of the model, with the draw results being particularly difficult to predict with respect to win and loss. This evidence has already been observed in previous attempts to model the KES database in Kaggle (Airback, 2017; Hodge, 2017; O'Brien, 2017; Rambier, 2018) and has been extensively documented in previous studies (Carpita et al., 2014, 2015). It could be due to both the fact that a draw is an outcome intrinsically characterized by a higher degree of uncertainty and to the problem of class imbalance (Menardi and Torelli, 2014), as draw has lower frequencies. In view of the earlier discussion, we first fitted an imbalanced model to data (Carpita et al., 2014), obtaining similar results: a slightly better result for draw at the expense of win outcome; such result cannot be considered profitable when the prediction of win is largely more crucial.

As stated in Carpita et al. (2014), even if a model that satisfactorily predicts all the three results should be preferred, from a practical point of view a coach might be more interested in the key factors that would help his team win the match. In this case, one would prefer the model with the best predictive power for the win result rather than the draw or loss. For this reason, we reduced the dependent variable of the model into two categories: WIN and NOWIN (which incorporates loss or draw results).

3 The binomial logistic regression model

A wide variety of models is suitable for prediction of a binomial outcome on large datasets, ranging from simple to more sophisticated ones. However, as suggested by James et al. (2013) and Kosinski et al. (2016), the best starting point is a simple model that can be easily built and quickly computed. This choice allows for a thorough examination of the contribution of each predictor to the model, thus helping researchers in developing and comparing more complex approaches (Carpita et al.,

2014; Lisi and Zanella, 2017). In view of the above, we start with a BLR model presented in this section, which will be further extended and compared with other ones in Section 5.

The BLR model is a generalized linear model used to estimate the probability linked to a binomial dependent variable Y using a set of covariates or predictors X (Hosmer Jr et al., 2013). As explained at the end of the previous section, the response variable Y consists of two categories, referred to two possible match results of the home team: WIN and NOWIN. Along with the binary response variable, distributed as a Bernoulli with probability π, the BLR model includes a vector $\mathbf{x} = (x_1, ..., x_k)'$ of k covariates or predictors. In our study, predictors $\mathbf{x}$ are the difference of home and away teams' performance indicators by players' role. This model relies on the *log-odds* or *logit transformation*:

$$logit[\pi(\mathbf{x})] = log\text{-}odds[\pi(\mathbf{x})] = log\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \mathbf{x}'\boldsymbol{\beta} \tag{1}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ being the vector of parameters estimated by maximum likelihood method and $\mathbf{x}'\boldsymbol{\beta}$ being the linear predictor.

Assuming $\pi < 1$, in this study the *odds*:

$$odds[\pi(\mathbf{x})] = \frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = exp(\beta_0 + \mathbf{x}'\boldsymbol{\beta}) \tag{2}$$

is the ratio of the probabilities of WIN to NOWIN the match for the home team. As an example, if the probability of success of the home team is 0.75, *odds* = $0.75/0.25 = 3$ or 3-to-1. Whereas probabilities are bounded between 0 and 1, odds can be any non-negative number, and can be easily used to assess the effect of performance indicator x_i on the probability that home team WIN the match. If $\mathbf{x}_{-i}$ is $\mathbf{x}$ without predictor x_i:

$$odds[\pi(\mathbf{x_i} + \mathbf{1}; \mathbf{x_{-i}})] = exp(\beta_i) \cdot odds[\pi(\mathbf{x})] \tag{3}$$

so that in our case $exp(\beta_i)$ is the increase of the odds of WIN-to-NOWIN for the home team due to a one unit increase in the role-based indicator of team's performance x_i.

The BLR model is implemented through the `glm` command in R. Moreover, to assess it's effectiveness we used the R packages `caret` and `pscl` to compute some goodness-of-fit (GoF) statistics and model prediction indexes, as explained in the next two sections.

3.1 Model goodness-of-fit statistics

For inferential purposes, McFadden (1973) proposed the *likelihood-ratio* as overall GoF statistic of the BLR model. It compares two *extreme* models:

$$G^2 = 2 \cdot (l_M - l_0), \tag{4}$$

where l_M is the maximized log-likelihood for the model with all the covariates, and l_0 is the maximized log-likelihood for the model without covariates, that is, with only the intercept β_0. The test statistic G^2 is distributed as a chi-square random variable with k degrees of freedom under the null hypothesis that the model without covariates is true against the alternative hypothesis that the current model is true.

As in linear regression, also in logistic regression there is a need of statistics possibly ranging from zero to one that summarize the overall GoF of the model, with zero indicating the worst fit and one indicating the perfect fit. It is well known that in linear regression models, the coefficient of determination R^2 provides this kind of information. The attempt to develop a similar statistic for the logistic regression led to the pseudo-R^2 statistics (Smith and McKenna, 2013). As Hu et al. (2006) summarize in their article, pseudo-R^2 is usually based on a latent model structure, in which the binary outcome is related to the predictors through a linear model. In this context, McFadden (1973) proposed the following:

$$R^2_{McF} = 1 - l_M / l_0 = -G^2/(2 \cdot l_0). \tag{5}$$

Other adjustments made pseudo-R^2 applicable to any model estimated by the maximum likelihood method with a correction based on the general principle formulated by Cragg and Uhler (1970). This popular and widely used statistic can be expressed as:

$$R^2 = \frac{1 - (L_0/L_M)^{\frac{2}{n}}}{1 - L_0^{\frac{2}{n}}} = \frac{1 - exp(-G^2/n)}{1 - exp(-G_0^2/n)}, \tag{6}$$

where L_M and L_0 are the maximized likelihood previously defined, G_0^2 is the likelihood ratio in the case of perfect fit of the model and $[1 - exp(-G_0^2/2)]$ is the norming term.

From an interpretative point of view, it is important to underline that pseudo-R^2 statistics for binomial regression assumes lower values with respect to the R^2 used in linear regression model: A satisfactory $R^2 = 0.5$ for linear regression in a micro-data context would correspond to $R^2_{McF} = 0.25$ and to $R^2 = 0.4$ in binomial regression (Veall and Zimmermann, 1996, Table 1).

3.2 Model prediction indexes

Among the classification tasks, the binary one is certainly the most popular: Based on the simple majority rule, the binary task aims at classifying cases into one and only one of two non-overlapping classes (Sokolova and Lapalme, 2009). Correctness of model classification can be evaluated by computing the number of correctly predicted matches referred to WIN of the home team (*true positives*), the number of correctly predicted matches referred to NOWIN of the home team (*true negative*) and matches incorrectly assigned to WIN (*false positives*) or NOWIN (*false negatives*). For the case of a binary classification, these four counts constitute the model *confusion table*:

		Predicted	
		WIN	NOWIN
Actual	WIN	true win (t_W)	false no win (f_{NW})
	NOWIN	false win (f_W)	true no win (t_{NW})

Starting from the confusion table, it is possible to compute some model prediction indexes, such as:

- *Accuracy*, which expresses how effectively the model predicts match results:

$$Acc = \frac{t_W + t_{NW}}{t_W + f_{NW} + f_W + t_{NW}} \tag{7}$$

- *Sensitivity*, which expresses how effectively the model predicts WIN matches:

$$Sen = \frac{t_W}{t_W + f_{NW}} = \frac{t_W}{a_W} \tag{8}$$

- *Specificity*, which expresses how effectively the model predicts NOWIN matches:

$$Spe = \frac{t_{NW}}{f_W + t_{NW}} = \frac{t_{NW}}{a_{NW}}, \tag{9}$$

where a_W and a_{NW} are the actual match results of WIN and NOWIN, respectively.

Finally, note that:

$$Acc = \frac{a_W \cdot Sen + a_{NW} \cdot Spe}{a_W + a_{NW}} \tag{10}$$

so that accuracy is the weighted mean of sensitivity and specificity indexes. The accuracy index of the model in (10) can be compared with the accuracy index obtained without the model: This is the *null accuracy*, and it is simply equal to the highest observed frequency of the two possibile results of the match.

4 Modelling match results using team performances

To estimate parameters and assess performance of the BLR model (1) described in Section 3, we used the following *bootstrap approach* with a nested *statistical learning-testing procedure*. For each of 1 000 bootstrap replication, the dataset containing teams' performance indicators and matches' results has been randomly split into two parts: a *learning set*, consisting of the 70% of observations, and a *testing set* made up of the remaining 30%. The *learning sets* have been used for parameter estimation and assessment of model through the GoF statistics mentioned

in Section 3.1, while the *testing sets* have been used to compute the prediction indexes introduced in Section 3.2.

The modelling step has been carried out taking into account the two different periods (seasons 2009–2011 and 2012–2015) on both: the whole KES dataset (without any distinction among countries), and on the dataset grouped by countries. In the following sections, results for the whole and the by-country KES datasets are presented.

4.1 Results for the whole KES dataset

Table 1 shows GoF statistics (Section 3.1) and prediction indexes (Section 3.2) for the BLR model (1) in Section 3. For both periods, the *learning set* values of McFadden (1973) likelihood ratio statistics G^2 described in (4) are very high (the null hypothesis that the model without covariates is true against the alternative that our model is true has to be rejected at the usual significance levels), whereas the *testing set* values of the Cragg and Uhler (1970) pseudo-R^2 displayed in (6) are rather low (21% and 23 % for the first and second period, respectively) but sufficient (see interpretation at the end of Section 3.2). To verify potential *multicollinearity* issues, we computed the usual *variance inflation factor* or VIF (James et al., 2013) index: in 1 000 bootstrap replications, the VIF ranged from $min = 1.1$ to $max = 5.1$, with $mean = 2.5$, so that multicollinearity is not a problem for our model.

For both periods, the *AccMod* index of the *testing set* (i.e., the prediction accuracy in (6) for the BLR model) is about 64%, that is, 10 percentage points greater of the *AccNull* index (the accuracy without model specification). Bootstrap accuracy intervals for the BLR model (lower bound is *L_AccMod* and upper bound is *U_AccMod*) range from 62% to 66%, therefore not containing *AccNull*. The prediction ability of the BLR model for each result of the match (home team WIN or NOWIN) can be evaluated by examining the *sensitivity* and *specificity* indexes described in (8, 9) and computed on the *testing set*: *Sen* indicates that the BLR model correctly predicts 56% of WIN for seasons 2009–2011 and 53% of WIN for seasons 2012–2015; *Spe* indicates that the BLR model correctly predicts 70% of NOWIN for the first period and 74% of NOWIN for the second one.

Table 1 GoF statistics and prediction indexes for the BLR model by seasons

Seasons	n.match	G^2	R^2	AccNull	AccMod	L_AccMod	U_AccMod	Sen	Spe
2009–2011	8 990	1 160	0.230	0.533	0.635	0.617	0.654	0.563	0.699
2012–2015	11 983	1 418	0.211	0.550	0.643	0.627	0.659	0.525	0.741

Parameter estimates and bootstrap *t*-statistics for the BLR model are reported in Table 2; for comparison purposes, team's performance indicators have been standardized. For seasons 2009–2011, 9 *t*-statistics (out of the 22) allow to reject the null hypothesis of coefficient being equal to 0 at the significance level of 10%; moreover, all but one of them (FOR.*mentality*, with *t-stat* = −2.7)

Table 2 Parameter estimates and *t*-statistics for the BLR model by seasons

Perf. indicator	Seasons 2009–2011		Seasons 2012–2015		Perf. indicator	Seasons 2009–2011		Seasons 2012–2015	
	$\hat{\beta}$	*t-stat*	$\hat{\beta}$	*t-stat*		$\hat{\beta}$	*t-stat*	$\hat{\beta}$	*t-stat*
FOR.*attacking*	0.063	1.969	0.059	2.360	DEF.*attacking*	0.098	3.062	0.104	3.467
FOR.*skill*	−0.003	−0.120	0.052	2.476	DEF.*skill*	0.037	1.233	0.061	2.103
FOR.*movement*	0.172	8.600	0.022	1.375	DEF.*movement*	0.015	0.714	0.060	3.529
FOR.*power*	0.050	2.000	0.016	0.842	DEF.*power*	0.025	0.962	−0.028	−1.273
FOR.*mentality*	−0.069	−2.654	0.051	2.217	DEF.*mentality*	0.032	1.185	−0.020	−0.714
FOR.*defending*	0.018	1.059	0.015	1.000	DEF.*defending*	0.233	9.320	0.221	11.632
MID.*attacking*	0.046	1.314	0.118	4.069	GOK.*movement*	0.029	1.611	0.040	2.500
MID.*skill*	0.096	3.200	0.079	3.038	GOK.*power*	0.019	1.056	0.008	0.533
MID.*movement*	0.163	7.762	0.117	6.882	GOK.*mentality*	−0.028	−1.647	0.010	0.714
MID.*power*	−0.045	−1.731	−0.041	−2.050	GOK.*goalkeeping*	−0.023	−1.211	0.114	6.706
MID.*mentality*	0.145	5.000	0.117	4.034					
MID.*defending*	−0.001	−0.048	0.008	0.400					

determine an expected increase in the probability of home team WIN. For the first period, DEF.*defending* (*t-stat* = 9.3), FOR.*movement* (*t-stat* = 8.6), MID.*movement* (*t-stat* = 7.8) and MID.*mentality* (*t-stat* = 5.0) significantly and positively affect the probability of home team WIN. In the first period, no statistically significant estimates emerged for the GOK player position.

For seasons 2012–2015, 14 *t*-statistics (out of the 22) allow to reject the null hypothesis of coefficient being equal to 0 at the significance level of 5%; moreover, all but one of them (MID.*power*, with *t-stat* = −2.1) determine an expected change, in terms of increase in the probability of home team WIN. The second period confirmed the significant and positive effects of DEF.*defending* (*t-stat* = 11.6), MID.*movement* (*t-stat* = 6.9) and MID.*mentality* (*t-stat* = 4.0) on the probability of home team WIN; this is not the case for FOR.*movement* (*t-stat* = 1.4), while the GOK.*goalkeeping* parameter results to be positive and statistically significant (*t-stat* = 6.7).

These results suggest that in the second period the European home teams may increase their WIN chances by changing their playing strategies, that is, decreasing the dimension *power* in both *midfielders* and *defenders*, and increasing *defenders*'and *goalkeepers*' performance (in particular the dimensions *defending* and *goalkeeping*). Considering that the attribute *power* refers to the amount of power a player can put in a shot while still keeping it precise, these results are not surprising and suggest that reinforcing more tactical skills (such as the ability to provide conditions for an *attacking* advantage, along with agility and general efficiency) rather than the brute force can be a better strategy for players having a less aggressive role in the field.

In summary, 16 out of 22 teams' performance indicators used as predictors in the BLR model show statistical significance and all but two of them increase the probability that home team WIN the match.

Figure 4 gives an insight on the *effect sizes*: The 95% confidence intervals, along with the bootstrap standard errors related to the increase in odds of WIN-to-NOWIN of the home team due to a one-unit increase in each of the 16 teams' performance

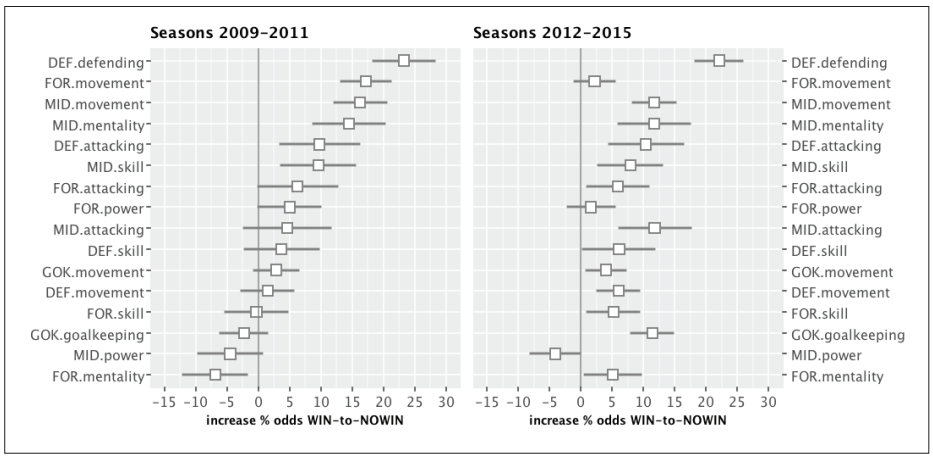


Figure 4 Confidence intervals for increase (%) of odds WIN-to-NOWIN for the home team due to an unit increase in the 16 significative role-based indicators of teams' performance with the BLR model by seasons

indicators (see (3) at the end of Section 3) are displayed. Comparing the two periods, the difference between the effect sizes of teams' performance indicators emerges. Hovewer, in both periods DEF.*defending* increases the odds WIN-to-NOWIN of more than 20%; MID.*movement* and MID.*mentality* increase by 15% in the first period and by 10% in the second one, while FOR.*movement* increases by 15% only in the first period.

4.2 Results for the by-country KES dataset

In this section, results obtained applying the BLR model (1) to the KES dataset split by country are presented. Table 3 shows GoF statistics and prediction indexes for the BLR model applied to the by-country dataset.

For both periods, values of likelihood ratio statistic G^2 in (4) are high and significant at the usual levels, while values of the pseudo-R^2 in (6) are generally rather low but still satisfactory in some cases (end of Section 3.2): This occurs in Portugal ($R^2 = 45\%$), Scotland ($R^2 = 44\%$) and Belgium ($R^2 = 37\%$) in the first period, but only in Portugal in the second period ($R^2 = 38\%$).

For the first period, the *AccMod* index for the BLR model is about 70% for Scotland, Netherlands and Portugal, that is, 10% points greater than the *AccNull* index. Accuracy null, 95% confidence intervals along with bootstrap standard errors of the BLR model accuracy index by country and periods are shown in Figure 5 (vertical lines represent null and model accuracy for the whole KES dataset): Model accuracy is always significantly greater than null accuracy, except for France and Switzerland in the first period, and for Scotland and Switzerland in the second period.

In seasons 2009–2011, the highest *sensitivity* (correct predictions of home team WIN: *Sen* in Table 3) is achieved by Netherlands (69%), followed by Spain (65%)

Table 3 GoF statistics and prediction indexes of the BLR model by country and seasons

	Seasons 2009–2011						
Country	n.matches	G^2	R^2	AccNull	AccMod	Sen	Spe
Belgium	690	144	0.365	0.516	0.590	0.547	0.643
England	1 140	140	0.216	0.522	0.651	0.606	0.683
France	1 140	114	0.183	0.551	0.582	0.445	0.694
Germany	918	101	0.198	0.564	0.616	0.474	0.715
Italy	1 118	134	0.213	0.532	0.611	0.574	0.651
Netherlands	918	180	0.331	0.523	0.678	0.687	0.660
Portugal	720	196	0.453	0.563	0.673	0.561	0.751
Scotland	684	178	0.437	0.589	0.691	0.542	0.781
Spain	1 140	163	0.250	0.519	0.614	0.645	0.591
Switzerland	522	67	0.229	0.539	0.577	0.476	0.657
	Seasons 2012/2013–2015/2016						
Country	n.matches	G^2	R^2	AccNull	AccMod	Sen	Spe
Belgium	732	108	0.262	0.546	0.630	0.536	0.705
England	1 520	172	0.201	0.558	0.645	0.532	0.728
France	1 520	174	0.205	0.559	0.628	0.524	0.716
Germany	1 224	120	0.176	0.536	0.622	0.524	0.685
Italy	1 519	187	0.218	0.548	0.629	0.544	0.716
Netherlands	1 224	165	0.236	0.548	0.630	0.553	0.714
Portugal	1 092	248	0.384	0.557	0.664	0.522	0.781
Scotland	912	131	0.257	0.582	0.611	0.415	0.752
Spain	1 520	246	0.278	0.517	0.658	0.624	0.676
Switzerland	720	111	0.273	0.557	0.596	0.472	0.708

and England (61%), while the highest *specificity* (correct predictions of home team NOWIN: *Spe* in Table 3) is higher for Scotland (78%), Portugal (75%) and Germany (72%). In seasons 2012–2015, the highest sensitivity values can be observed for Spain (62%), Netherlands (55%) and Italy (54%), while the highest specificity is achieved by Portugal (78%), followed by Scotland (75%) and England (73%).

The 95% confidence intervals of *sensitivity* and *specificity* indexes for the BLR model, along with bootstrap standard errors, divided by period and country are displayed in Figures 6 and 7, respectively (vertical lines represent index values for the whole KES dataset): With respect to the average of the whole KES dataset, the BLR model of the dataset by-country shows better results *sensitivity* for Netherlands, Spain and England, and better results in *specificity* for Scotland and Portugal. Conversely, in the first period *sensitivity* is worse for Germany and France, while *specificity* is worse for Spain.

In general, it is known that a better *sensitivity* is obtained at the expense of *specificity* and vice versa: As showed by these last two figures, whole and by-country *sensitivity* indexes decrease of about 5% shifting from the first to the second period, while *specificity* shows roughly the same increase. However, while country order remains the same from a *specificity* point of view (except for Germany), this does not occur for *sensitivity*.

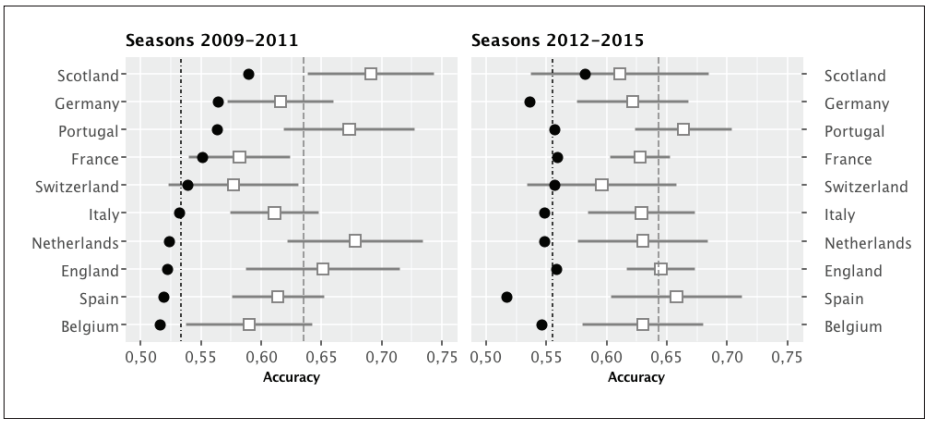


Figure 5 Accuracy null (black dots) and confidence intervals for BLR model *accuracy*

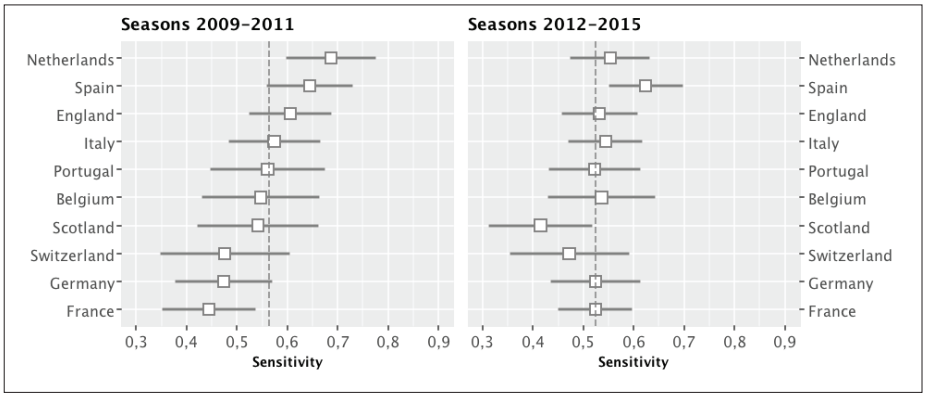


Figure 6 Confidence intervals for BLR model *sensitivity*

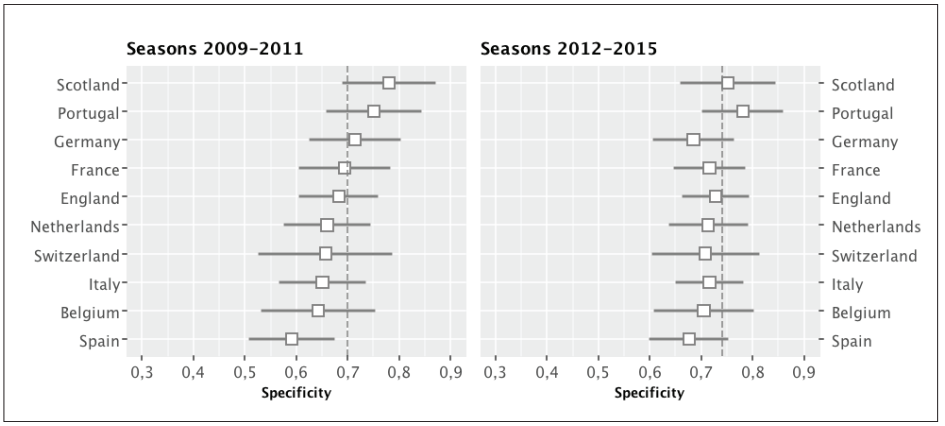


Figure 7 Confidence intervals for BLR model *specificity*

5 Extensions and predictions of the BLR model

For the sake of a more detailed evaluation of our BLR model (1), in this section two extensions have been taken into account. First, the possibility to rely on an overall performance index referred to the whole team before the actual match and independent from the performance achieved by the players (grouped by roles) chosen by the coach to play the match; This aspect has been evaluated by adding the *ELO* rating as additional regressor to our model. A second important aspect that deserves attention is related to the hierarchical structure of the dataset: The *match–season–country* structure can be modelled as a three-level model using two *random effects*. Both extensions and results are presented in Section 5.1.

Lastly, predictive power of our BLR model with and without the two extensions has been compared to other statistical models: Random Forest, Neural Networks, k-Nearest Neighbours and Naïve Bayes. For this purpose, the `caret` package, that is, a set of R functions that streamlines the creation of predictive models, has been entirely used. As a preliminary step, appropriate treatment of some missing data left was necessary (516 for the *forwards* role, 42 for the *midfielders* role, 45 for the *defenders* and 107 for the *goalkeepers*): Imputation based on the mean of *k-Nearest Neighbors* found in the data allowed for simple, accurate and efficient missing data handling (Kuhn et al., 2008). Results of this comparison are showed in Section 5.2.

5.1 Extensions of the BLR model

In order to incorporate the overall perfomance achieved by each team before each match into our BLR model (1), the *ELO* rating has been used. *ELO* is a well-known probability-based rating system proposed by Elo (1978) for chess and today also used in soccer: Hvattum and Arntzen (2010) used the *ELO* rating for match predictions and concluded that this rating appears to be useful in encoding the overall information of past results for team strength measurement. For each team and after each match, the *ELO* rating is computed using the following recursive formula:

$$ELO_{new} = ELO_{old} + K \cdot G \cdot (W - We), \tag{11}$$

where the K-factor is a positive constant chosen by the rater, G is the goal difference index, that is, a match importance index corresponding to the difference between goals of the home and away team in absolute value; W is the final result of the match (W = 1 for win, 0.5 for draw e 0 for loss), and We is the expected match result, based on the difference in ratings of the two teams and that considers the advantage for the team playing at home. In this study, the *ELO* rating was computed using the `elo` R package: Given that the goals difference has been computed in a binary form (0 or 1) G is set to 1, and the ELO index is updated for each team over each match.

For this analysis, a starting value of the ELO_{old} rating of 1 500 and match results of the first six months of season 2009–2010 were used to obtain different ELO_{new} rating for each team of each country; the K-factor of 30 and the advantage

of 100 for the home team were used, as in the World Football Elo Ratings (`www.eloratings.net/about`).

The three-level hierarchical structure of the data (matches within seasons, seasons within countries) is added to the intercept of the BLR model (1) as a double random effect:

$$logit[\pi(\mathbf{x})] = \beta_{0sc} + \mathbf{x}'\boldsymbol{\beta} \tag{12}$$

$$\beta_{0sc} = \beta_0 + v_c + u_{sc},$$

with $v_c \sim \mathcal{N}(0, \sigma_v^2)$ and $u_{sc} \sim \mathcal{N}(0, \sigma_u^2)$ being the two random effects, one for country (3rd level: 10 groups) and one for season within country (2nd level: $7 \times 10 = 70$ groups), and $\mathbf{x}'\boldsymbol{\beta}$ being the linear predictor (with or without the *ELO* rating). The BLR model (12) is named *multilevel/hierarchical generalized linear* model (Gelman and Hill, 2007) or *non-linear mixed-effects* model (Pinheiro and Bates, 2010), and its parameters $\boldsymbol{\gamma} = (\beta_0, \beta_1, ..., \beta_k, \sigma_v, \sigma_u)'$ are estimated through maximum likelihood method, using the Laplace approximation of the integral over the random effects space. Note that, instead of 10 + 70 = 80 fixed group-level coefficients, only the two random group-level coefficients (σ_v, σ_u) are estimated: This choice is parsimonious and *always* advisable in presence of a multilevel structure (Gelman and Hill, 2007, p. 246).

Table 4 displays estimates and *t*-statistics for the following four BLR models, estimated with all countries and seasons (except the first 6 months of season 2009–2010, to initialize the *ELO* rating):

- The BLR model (1), with role-based indicators of teams' performance (BLR1);
- The BLR1 model, with the *ELO* rating added as predictor (BLR2);
- The BLR model (12), with the predictors of BLR1 and random effects (BLR3);
- The BLR model (12), with the predictors of BLR2 and random effects (BLR4).

As shown by the parameter estimates of the BLR1 and BLR2 models in Table 4, the *ELO* rating provided a significant contribution to model improvement, slightly reducing the statistical effects of the other predictors. In order to assess the presence of multicollinearity of the *ELO* rating with our role-based indicators of teams' performance, the VIF index as in Section 4.1 has been computed: No regressor showed values of this index greater than 5, thus confirming the absence of such problem. The β parameter estimates of the BLR3 and BLR4 models in Table 4 are essentially the same of BLR1 and BLR2 models, respectively, and estimates of the season/country random effect parameters are very low (about 0.04 and 0.06), considering that the residual standard deviation of the logistic distribution is 1.7. In other words, the hierarchical structure of this data doesn't seem worth being taken into account. Note that standard errors for the two random effects are not reported, as estimators of variances do not tend to have a symmetric distribution (Bates, 2009).

In Table 5, a comparison of the GoF statistics of the four BLR models is showed. The usual information criteria indexes (AIC and BIC) are only slightly lower for the

Table 4 Parameter estimates for the four BLR models

	BLR1		BLR2		BLR3		BLR4	
Perf. indicator	$\hat{\beta}$	*t-stat*	$\hat{\beta}$	*t-stat*	$\hat{\beta}$	*t-stat*	$\hat{\beta}$	*t-stat*
FOR.*attacking*	0.014	0.400	−0.025	−0.700	0.015	0.416	−0.024	−0.683
FOR.*skill*	0.051	1.665	0.037	1.207	0.050	1.655	0.036	1.196
FOR.*movement*	0.076	3.393	0.061	2.728	0.076	3.385	0.061	2.722
FOR.*power*	0.063	2.299	0.063	2.292	0.063	2.281	0.062	2.277
FOR.*mentality*	0.018	0.567	0.011	0.345	0.018	0.578	0.011	0.359
FOR.*defending*	0.019	0.937	0.023	1.129	0.019	0.910	0.022	1.097
MID.*attacking*	0.065	1.584	0.019	0.447	0.064	1.557	0.017	0.420
MID.*skill*	0.090	2.483	0.066	1.805	0.090	2.493	0.066	1.815
MID.*movement*	0.133	5.617	0.101	4.235	0.133	5.610	0.101	4.232
MID.*power*	−0.041	−1.413	−0.032	−1.076	−0.041	−1.414	−0.032	−1.075
MID.*mentality*	0.149	4.056	0.127	3.430	0.149	4.060	0.127	3.432
MID.*defending*	−0.016	−0.611	−0.023	−0.881	−0.016	−0.609	−0.023	−0.879
DEF.*attacking*	0.133	3.255	0.110	2.679	0.133	3.250	0.110	2.677
DEF.*skill*	0.019	0.511	−0.004	−0.100	0.020	0.522	−0.003	−0.090
DEF.*movement*	0.036	1.465	0.014	0.561	0.035	1.442	0.013	0.537
DEF.*power*	0.002	0.059	−0.005	−0.167	0.002	0.071	−0.005	−0.153
DEF.*mentality*	0.003	0.089	0.002	0.047	0.003	0.075	0.001	0.031
DEF.*defending*	0.228	8.000	0.152	5.186	0.229	8.033	0.153	5.217
GOK.*movement*	0.015	0.711	0.004	0.196	0.015	0.711	0.004	0.199
GOK.*power*	0.030	1.394	0.025	1.184	0.030	1.397	0.025	1.188
GOK.*mentality*	−0.023	−1.134	−0.011	−0.540	−0.023	−1.128	−0.011	−0.536
GOK.*goalkeeping*	0.079	3.321	0.029	1.215	0.080	3.343	0.030	1.241
ELO.rating			0.379	12.323			0.379	12.312
Season effect (σ_u)					0.042		0.048	
Country effect (σ_v)					0.063		0.062	

mixed-effects models BLR3 and BLR4; also, the two likelihood ratio test statistics G^2 as in (4), comparing the maximized log-likelihood of these last models with that of BLR1 and BLR2 models, respectively, are rather low (about 4.5). Moreover, the p-values (about 0.1) of a Chi-squared distribution with 2 degrees of freedom do not allow to draw conclusions about the null hypothesis of no random effects, considering that the likelihood ratio test for variance parameters tends to be conservative in mixed-effects models (Pinheiro and Bates, 2010, pp. 84, 323).

Lastly, this GoF analysis is not conclusive about the choice of the 'best' model between BLR1–4. In the next section, a comparison between these and other models has been carried out in order to test their predictive ability, thus providing further evaluation elements.

5.2 Predictions of the BLR model and the other models

Within the Kaggle competition context, some experts in statistics and data science attempted to develop well-performing models for football outcome prediction (Hodge, 2017; O'Brien, 2017; Rambier, 2018). All of them considered only part

Table 5 GoF statistics for the four BLR models

Model	df	log-lik	AIC	BIC	G^2
BLR1	23	−8 547	17 139	17 312	
BLR2	24	−8 469	16 986	17 166	
BLR3	25	−8 544	17 139	17 327	4.550
BLR4	26	−8 467	16 985	17 181	4.344

of the KES database, such as a particular season or a particular country, along with bookmakers' predictions or players' performance indicators. However, none of them considered players' role on the pitch as a relevant aspect to be integrated into players' performance indicators. Moreover, difficulty in predicting draws is evident in all models that barely achieved 50% in accuracy. Airback (2017) presents an interesting application with the KES database, considering dimension reduction techniques and comparing the performance of various models, showing the superiority of Gaussian Naïve Bayes with respect to the others (BLR model included); when players' performance indicators are included in his model, *sensitivity* and *specificity* indexes are 43% and 55%, respectively. Indeed, in our study, building new role-based indicators of teams' performance and dichotomizing the response variable considerably increased the predictive power of the BLR model, which performed better than the one built by Airback (2017).

Following the model competition approach of Airback (2017), in this section the BLR model (1) and its extensions (12) have been compared with other well-known classification models (James et al., 2013; Perl and Weber, 2004; Carpita et al., 2014), in the spirit of the *two cultures* (Breiman, 2001): *Random Forest* (RF), a tree-based classification model that reduces heterogeneity in the bagged trees, approximating very complex relationships and enhancing reliability of results; *Neural Network* (NN), a learning model based on the estimation of non-linear combinations of the covariates to model a response variable composed of simple computational units (neurons), structured on different levels and interconnected through an architecture of linkages; *k-Nearest Neighbor* (k-NN), a learning model that assigns to each covariate profile the most common category of the response variable among its *k nearest neighbors*; *Naïve Bayes* (NB), a classification model based on the probabilistic approach with the strong (naïve) assumption that covariate values are independent from each other given the category of the response variable. For NN and k-NN, the default of the `caret` package allowed for optimal solutions, consisting in size 1 and decay 0 in the first case, and in k = 9 in the second case.

Prediction assessment of the models is similar to the one described in Section 4: In this case, 500 bootstrap replications have been used, each randomly splitting the dataset into 70% (*learning set*) and 30% (*testing set*). Predictions indexes (Acc, Sen, Spe) are reported in Table 6 (for all models, their standard errors range between 0.001 and 0.003).

All models show similar *accuracy* (about 64%), with the exception of the k-NN model (about 61%). The four BLR models have very similar performances and better than in previous attempts to model the KES database (Airback, 2017), confirming

Table 6 Models' prediction indexes

Classifier	Acc	Sen	Spe
BLR1	0.642	0.544	0.725
BLR2	0.643	0.540	0.730
BLR3	0.642	0.544	0.724
BLR4	0.643	0.545	0.725
RF	0.646	0.519	0.752
NN	0.643	0.544	0.726
k-NN	0.611	0.536	0.674
NB	0.635	0.623	0.645

that the role-based indicators of teams' performance are useful as predictors, and that the *ELO* rating and random effects do not improve the prediction. Except for NB, the other models show lower predictive for WIN (about 54%, with 52% for RF) with respect of NOWIN (about 73%, with 67% for k-NN). Thus, results confirm that NOWIN prediction is more precise than the WIN prediction: This happens because NOWIN actually consists of two distinct categories (draw and loss) but, as explained in the conclusions, further insights are necessary about this issue. Although *accuracy* of NB model is slightly lower (63.5%) than the one of other models, it exhibits *sensitivity* and *specificity* values closer to each other (62.3% and 64.5%, respectively), that is, it is more balanced in predicting both WINs and NOWINs.

6 Conclusions

In this article, original results of exploratory data analysis and modelling performed on the European soccer data of KES database (10 countries/leagues and 7 seasons, from 2009/2010 to 2015/2016) are presented.

First, 33 players' performance indicators have been aggregated (averaged) using experts' evaluation and linked to players' role assigned by the coach before each match. We think that considering these indicators can be useful from the interpretative point of view, as they are related to the strategic choices of the coach of each team. Then, the role-based indicators of team performance have been used to estimate their effects on the home team winning probability using the simple BLR model. Results highlight the importance of *defenders*: a one unit increase in the home-away team difference for the defending indicator increases odds of win for the home team by more than 20%. Performance of *midfielders* is important as well: A one unit increase in the home-away team difference of the *movement* and *mentality* indicators increases the WIN odds by about 15% in the period 2009–2011 and by about 10% in the period 2012–2015. Instead, *forwards*' performance is not so important from a match results point of view: In fact, *forwards*' movement indicator increases by 15% the odds of WIN for the home team only in the first period.

In the light of these first evidences, it seems interesting to consider the possibility to aggregate a more *neutral* outcome, such as *draw*, to the *win* outcome. However, this may be possible when the distribution of such outcomes is fair enough to ensure balanced results. In our dataset, combining *draw* along with the *win* outcome would have resulted in too much imbalance in the data, thus undermining the analyses.

Second, the bootstrap approach with a statistical learning–testing procedure has been used to assess model GoF and prediction accuracy in both the whole and the by-country dataset for seasons 2009–2011 and 2012–2015. For both periods, the BLR model shows rather good statistical properties, its classification *accuracy* is 10% greater than the classification accuracy without the model (about 64% with respect to 54%), and it correctly predicts about 54% of the home team WINs (*sensitivity*) and about 73% of the home team NOWINs (*specificity*). *Sensitivity* of the BLR model is significantly higher in Spain and Netherlands leagues (in the second period only) and lower for France (in the first period only), while *specificity* is significantly higher for Scotland (in the first period only) and lower for Spain.

Third, the BLR model has been extended towards two directions: (a) the well-known *ELO* rating for the overall performance of each team before each match has been added as regressor; (b) the hierarchical structure of the dataset (matches within seasons, seasons within countries) has been modelled as random effect. Even if the BLR model confirmed the significance of the *ELO* rating effect on the home team winning probability, both these extensions do not substantially improve the performance of our simpler model in terms of *accuracy*, *sensitivity* and *specificity*.

Fourth, the BLR model has been further compared with other ones reflecting a wide variety of statistical approaches (RF, NN, k-NN and NB). Except for the k-NN model, which seemed to perform slightly worse than the others, the remaining ones performed good, but they all showed a lower ability to predict WINs with respect of NOWINs. The NB model seems to provide more balanced classification ability, in terms of WINs and NOWINs (as it emerged in another study with the KES database [Airback, 2017]). Of course, the choice of one over the other depends on the goal to be achieved. Where the WIN prediction becomes of crucial importance, choosing the last one would result in better performance.

This study can be extended in various directions. First, it can be interesting to verify if there are other ways to aggregate the 33 players' performance indicators, in order to improve the model GoF. Teams' performance indicators by player's role might be defined and validated on a statistical basis rather than using the simple average on experts' classification: In particular, a factor analytic approach would clarify correlation and structure, allowing to obtain more robust composite indicators. Second, it can be useful to try to improve the predictive performances of the models, in particular for the DRAW result. In addition to find other or more accurate predictors for the models with three categories of the dependent variable (WIN, LOSS and DRAW), it would be interesting to investigate the effect of the classification rule applied to transform the estimated probabilities into the expected result for the match.

Appendix: The KES database

The original version of KES database refers to 11 different countries (Belgium, England, France, Germany, Italy, Netherlands, Portugal, Scotland, Spain, Switzerland and Poland) and 8 football seasons (from 2008–2009 to 2015–2016): It is made up of data coming from different sources (Mathien, 2016). Two main tables of the KES database are considered in this study:

- *Player_Attributes table*: The original dimension of this data table is 183 978 rows (players) and 42 columns (variables). The first 3 variables are players' identification codes, followed by 6 variables: reference date (yyyy-mm-dd), overall and potential rating (on a 0 to 100 scale), preferred foot, attacking and defensive work rates (ordinal scales). The last 33 variables represent player's performance (0–100 scale) with respect to different abilities of the soccer player (var.long.name in Table A2) from the EA Sports FIFA videogame.
- *Match table*: Originally, this data table counts 25 979 rows (matches) and 115 columns (variables). The first 8 variables are matches' identification codes, country, league, season, stage, home and away team, followed by reference date (yyyy-mm-dd) and number of goals of the two teams. The following 44 variables

Table A1 Statistics of players and matches by country, league, seasons of KES database

		Seasons 2009/2010–2011/2012						
Country	League	n.players	n.obs	n.matches	n.miss	f.win	f.loos	f.draw
Belgium	Belgium Jupiler League	1 053	1 790	690	36	0.481	0.261	0.258
England	England Premier League	1 463	2 701	1 140	6	0.476	0.261	0.263
France	France Ligue 1	1 404	2 318	1 140	20	0.452	0.254	0.294
Germany	Germany 1. Bundesliga	1 271	2 348	918	18	0.441	0.310	0.248
Italy	Italy Serie A	1 462	2 695	1 118	16	0.474	0.258	0.268
Netherlands	Netherlands Eredivisie	1 192	2 235	918	13	0.510	0.277	0.214
Portugal	Portugal Liga ZON Sagres	1 144	2 000	720	30	0.444	0.296	0.260
Scotland	Scotland Premier League	902	1 520	684	26	0.406	0.346	0.247
Spain	Spain LIGA BBVA	1 447	2 603	1 140	11	0.508	0.257	0.235
Switzerland	Switzerland Super League	717	1 174	522	9	0.462	0.308	0.230
		Seasons 2012/2013–2015/2016						
Country	League	n.players	n.obs	n.matches	n.miss	f.win	f.lose	f.draw
Belgium	Belgium Jupiler League	1 256	5 484	732	20	0.449	0.305	0.246
England	England Premier League	1 907	9 289	1 520	8	0.443	0.303	0.254
France	France Ligue 1	1 973	7 299	1 520	23	0.447	0.282	0.271
Germany	Germany 1. Bundesliga	1 691	8 814	1 224	10	0.453	0.306	0.241
Italy	Italy Serie A	2 044	12 514	1 519	11	0.451	0.285	0.264
Netherlands	Netherlands Eredivisie	1 604	6 278	1 224	13	0.454	0.293	0.252
Portugal	Portugal Liga ZON Sagres	1 692	8 145	1 092	40	0.440	0.301	0.258
Scotland	Scotland Premier League	1 190	4 573	912	24	0.423	0.337	0.240
Spain	Spain LIGA BBVA	1 957	8 191	1 520	17	0.475	0.293	0.232
Switzerland	Switzerland Super League	987	4 152	720	21	0.447	0.301	0.251

represent, respectively, the team line-up with X and Y coordinates for each of the 22 players of the match (22 variables) and the identification codes of the corresponding 22 players (22 variables). Eight variables describe some actions of the match and the last 30 variables are the bets of results (win, loss and draw) of 10 betting companies.

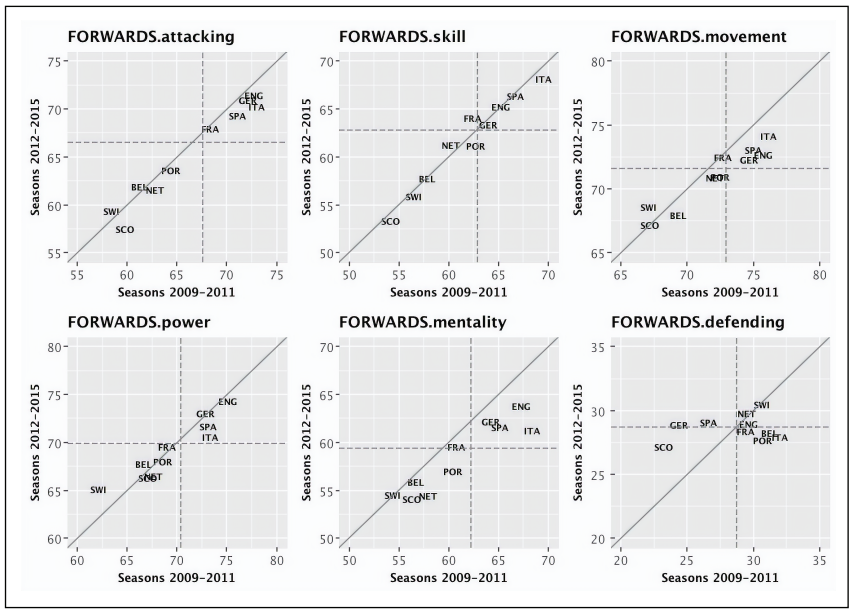
Support tables in the KES database contain country and league names, players' name, birthday, height and weight, and teams' name. After a preliminary inspection, we excluded season 2008/2009 from our analysis, as there was a 50% to 80% of missing data in five countries; we excluded the Poland League as well, as its seasons 2008/2009 and 2012/2013 contained 50% to 80% of missing data too.

Table A1 shows some statistics of players and matches by country, league and period from the two main tables of the KES database. The first period (three seasons, from 2009/2010 to 2011/2012) includes a total of 12 055 players and 21 384 observations, 8 990 matches and a 2.1% missing data. The second period (four seasons, from 2012/2013 to 2015/2016) involves a total of 16 301 players with 74 739 observations, 11 983 matches and 1.6% of missing data. During the first period, the number of players per country and matches per country range, respectively, from 717 to 1 463 and from 522 to 1 140. In the second period, they range, respectively, from 987 to 2 044 and from 720 to 1 520.

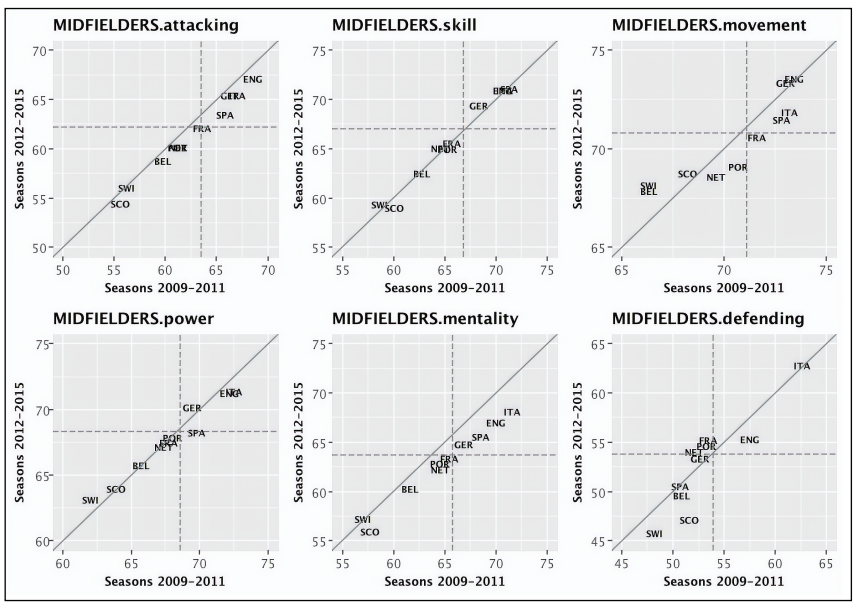
Results of each match are the difference between home and away team goals, and can be classified into three categories: win, loss and draw of the home team. The last three columns in Table A1 show relative frequencies for each result: Win ranges from 40% to 51%, whereas for loss and draw ranges are 26%–34% and 21%–29%, respectively.

Table A2 Variables and experts' dimensions (EA Sports FIFA) of the KES database

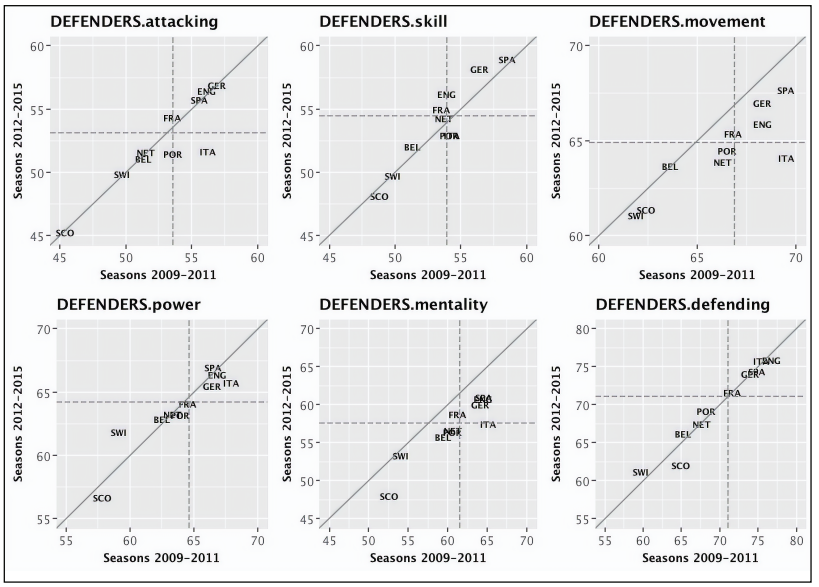
dimension	var.short.name	var.long.name	dimension	var.short.name	var.long.name
power	pow1	shot_power	*movement*	mov1	acceleration
power	pow2	jumping	*movement*	mov2	sprint_speed
power	pow3	stamina	*movement*	mov3	agility
power	pow4	strength	*movement*	mov4	reactions
power	pow5	long_shots	*movement*	mov5	balance
mentality	men1	aggression	*attacking*	att1	crossing
mentality	men2	interceptions	*attacking*	att2	finishing
mentality	men3	positioning	*attacking*	att3	heading_accuracy
mentality	men4	vision	*attacking*	att4	short_passing
mentality	men5	penalties	*attacking*	att5	volleys
skill	ski1	dribbling	*defending*	def1	marking
skill	ski2	curve	*defending*	def2	standing_tackle
skill	ski3	free_kick_accuracy	*defending*	def3	sliding_tackle
skill	ski4	long_passing	*goalkeeping*	gok1	gk_diving
skill	ski5	ball_control	*goalkeeping*	gok2	gk_handling
			goalkeeping	gok3	gk_kicking
			goalkeeping	gok4	gk_positioning
			goalkeeping	gok5	gk_reflexes



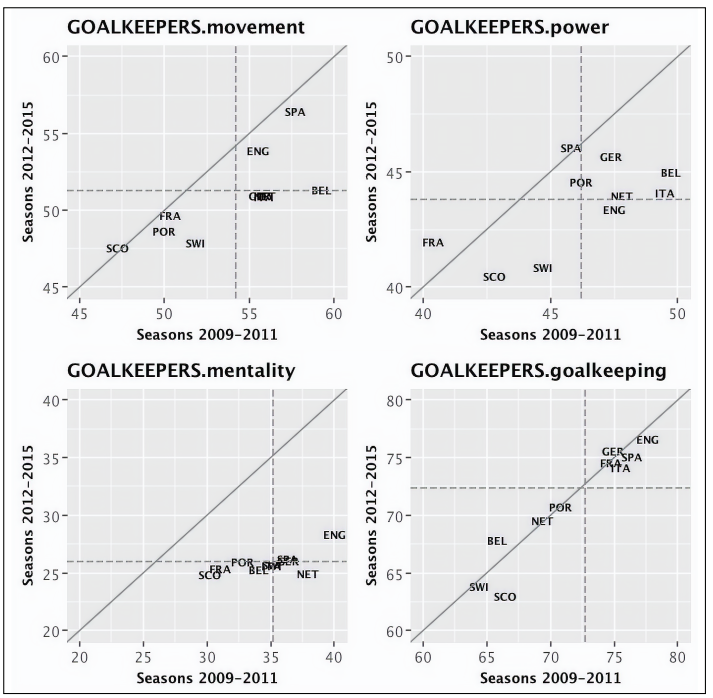
(a) *Forwards*



(b) *Midelders*



(c) *Defenders*



(d) *Goalkeepers*

Figure A1 Scatterplots of average players' role performances indicators by teams' country in Seasons 2009–2011 and 2012–2015

After the exploratory data analysis of the KES database, using the expert's classification of *sofifa.com* (see Section 2), the 33 players' performance variables of the *Player_Attributes table* have been grouped in 7 dimensions (Table A2): Due to missing data, the average score for each players has been preferred over the total score, so obtaining 7 players' performance indicators (*power*, *mentality*, *skill*, *movement*, *attacking*, *defending* and *goalkeeping*).

The link between performance indicators of *Player_Attributes table* and players of the home and away teams in the *Match table* was performed by using *player identification code* and *reference date*. With respect to this second key, the *rolling join* procedure of the R package `data table` was used to link the performance indicators of the immediately previous date *before* the match.

Finally, the average of players' performance indicators grouped by players' roles have been computed for both home and away teams and for each match. Figure A1 shows scatterplots of average performance indicators by country and players' role in seasons 2009–2011 and 2012–2015: Within each player's performance indicator considered, labels are the position of countries in both periods, while vertical and horizontal dotted lines represent the European average and the diagonal continuous line is the bisector line.

Among players' roles, country indicators are above the European average for England, France, Germany, Italy and Spain. Some performance indicators show a stable country average within the two periods (FORWARDS.*attacking* and *skill*; MIDFIELDERS.*attacking*, *skill* and *power*; DEFENDERS.*skill*, *power* and *defending*; GOALKEEPERS.*goalkeeping*), whereas others have a very low range in the second period with respect to the first one (FOR.*defending* and GOK.*mentality*).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

References

Ahuja K, Dey K, Nagar S and Vaculin R (2017) *Determining player performance statistics using gaze data*. US Patent App, 15/184 229. URL `https://patents.google.com/patent/US20170361157A1/en`(last accessed 25 October 2018).

Airback (2017) Match outcome prediction in football. URL `www.kaggle.com/airback/matchoutcome-prediction-in-football?scriptVersionId=796746` (accessed 8 October 2018).

Albert J, Bennett J and Cochran J (2005) *ASA-SIAM Series on Statistics and Applied Probability: Anthology of Statistics in Sports*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Albert J, Glickman ME, Swartz TB and Koning RH (2017) *Handbook of Statistical Methods and Analyses in Sports*. Boca Raton, FL: CRC Press.

Alves AM, de Mello JCCBS, Ramos TG and Sant'Anna AP (2011) Logit models for the probability of winning football games. *Pesquisa Operacional*, **31**, 459–65.

Bates DM (2009) *Assessing the precision of estimates of variance components*. URL `lme4.r-forge.r-project.org/slides/2009-07-21-Seewiesen/4Precision-4a4.pdf` (last accessed 25 October 2018).

Breiman L (2001) Statistical modeling: The two cultures. *Statistical Science*, **16**, 199–231.

Carling C, Williams AM and Reilly T (2005) *Handbook of soccer match analysis: A systematic approach to improving performance*. New York: Routledge.

Carpita M, Sandri M, Simonetto A and Zuccolotto P (2014) Football mining with R. In *Data Mining Applications with R*, edited by Y Zhao and Y Cen, pages 398–433. Waltham, MA: Academic Press.

——— (2015) Discovering the drivers of football match outcomes with data mining. *Quality Technology & Quantitative Management*, **12**, 561–77.

Cattelan M, Varin C and Firth D (2013) Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**, 135–50.

Cragg JG and Uhler RS (1970) The demand for automobiles. *The Canadian Journal of Economics/Revue canadienne d'Economique*, **3**, 386–406.

Dobson S and Goddard J (2011) *The Economics of Football, 2nd edition*. Cambridge, UK: Cambridge University Press.

Elo AE (1978) *The Rating of Chessplayers: Past & Present*. London: Battsford.

Gelman A and Hill J (2007) *Analytical Methods for Social Research: Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Goddard J (2005) Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, **21**, 331–40.

Hodge P (2017) *Predicting EPL scores for fun*. URL `www.kaggle.com/petehodge/predictingepl-scores-for-fun` (accessed 8 October 2018).

Hosmer Jr DW, Lemeshow S and Sturdivant RX (2013) *Applied Logistic Regression*. Vol. 398. Hoboken, NJ: John Wiley & Sons.

Hu B, Shao J and Palta M (2006) Pseudo-R2 in logistic regression model. *Statistica Sinica*, **16**, 847–60.

Hvattum LM and Arntzen H (2010) Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, **26**, 460–70.

James G, Witten D, Hastie T and Tibshirani R (2013) *An Introduction to Statistical Learning*. Vol. 112. New York: Springer.

Karlis D and Ntzoufras I (2003) Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**, 381–93.

——— (2009) Bayesian modelling of football outcomes: Using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, **20**, 133–45.

Koning RH (2000) Balance in competition in Dutch soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **49**, 419–31.

Kosinski M, Wang Y, Lakkaraju H and Leskovec J (2016) Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, **21**, 493.

Kuhn M (2008) Building predictive models in R using the caret package. *Journal of Statistical Software*, **28**, 1–26.

Leung CK and Joseph KW (2014) Sports data mining: Predicting results for the college football games. *Procedia Computer Science*, **35**, 710–19.

Lisi F and Zanella G (2017) Tennis betting: Can statistics beat bookmakers? *Electronic Journal of Applied Statistical Analysis*, **10**, 790–808.

Liti C, Piccialli V and Sciandrone M (2017) Predicting soccer match outcome using machine learning algorithms. In *Proceedings of MathSport International 2017 Conference*, edited by C De Francesco, L De Giovanni, M Ferrante, G Fonseca, F Lisi and S Pontarollo, page 229. Padova: Padova University Press.

Magel R and Melnykov Y (2014) Examining influential factors and predicting outcomes in European soccer games. *International Journal of Sports Science*, **4**, 91–6.

Mathien H (2016) *European Soccer Database*. URL `www.kaggle.com/hugomathien/soccer` (last accessed 25 October 2018).

McFadden D (1973) *Conditional Logit Analysis of Qualitative Choice Behavior*, pages 105–42. New York, NY: Academic Press.

McHale IG and Scarf P (2007) Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, **61**, 432–45.

McHale IG, Scarf PA and Folker DE (2012) On the development of a soccer player performance rating system for the English Premier League. *Interfaces*, **42**, 339–51.

McHale IG and Szczepanski L (2014) A mixed effects model for identifying goal scoring ability of footballers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **177**, 397–417.

Menardi G and Torelli N (2014) Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, **28**, 92–122.

O'Brien C (2017) *Logistic regression for betting*. URL `www.kaggle.com/colinobrienbi/logistic-regression-for-betting?scriptVersionId=655052` (last accessed 25 October 2018).

Odachowski K and Grekow J (2013) Using bookmaker odds to predict the final result of football matches. In *Knowledge Engineering, Machine Learning and Lattice Computing with Applications*, edited by M Graña, C Toro, RJ Howlett and LC Jain, pages 196–205. Berlin and Heidelberg: Springer Berlin Heidelberg.

Perl J and Weber K (2004) A neural network approach to pattern learning in sport. *International Journal of Computer Science in Sport*, **3**, 67–70.

Pinheiro JC and Bates DM (2010) *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rambier E (2018) *Match outcome prediction*. URL `www.kaggle.com/rambieres telle/matchoutcomepredictions` (last accessed 8 October 2018).

Schauberger G, Groll A and Tutz G (2016) *Modeling football results in the German Bundesliga using match-specific covariates*. URL `epub.ub.uni-muenchen.de/29390/` (last accessed 25 October 2018).

Slaton Z (2012) A beautiful numbers game. URL `www.abeautifulnumbersgame.com` (last accessed 8 October 2018).

Smith TJ and McKenna CM (2013) A comparison of logistic regression pseudo R2 indices. *Multiple Linear Regression Viewpoints*, **39**, 17–26.

Sokolova M and Lapalme G (2009) A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, **45**, 427–37.

Stern H (2005) Introduction to the football articles. In ASA-SIAM Series on Statistics and Applied Probability: Anthology of Statistics in Sports, edited by A Jim, B Jay and JC James, Ch. 3, pages 13–15. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Tutz G and Schauberger G (2015) Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Advances in Statistical Analysis*, **99**, 209–27.

Veall MR and Zimmermann KF (1996) Pseudo-R2 measures for some common limited dependent variable models. *Journal of Economic Surveys*, **10**, 241–59.

Zelenkov Y and Solntsev I (2017) Measuring the efficiency of Russian football premier league clubs. *Electronic Journal of Applied Statistical Analysis*, **10**, 773–89.