Author´s final version (how to cite):
Zambom-Ferraresi, F., Rios, V., & Lera-López, F. (2018). Determinants of sport performance in European football: What can we learn from the data?. *Decision Support Systems*, 114, 18-28. https://doi.org/10.1016/j.dss.2018.08.006

Determinants of sport performance in European football: What can we learn from the data?

Abstract

Nowadays game-related statistics in the sports industry are demanded by coaches, players, managers, journalists, supporters, fans, video games developers, betting markets and academics. However, the employment of game-related statistics to analyse performance in football (soccer) has inherent problems given it is a multifaced and complex phenomenon. This study analyses the importance of a large number of possible determinants of sport performance in the "Big Five" European football leagues during the period 2012/13–2014/15. To this end, Bayesian model averaging techniques and relative importance metrics are employed. The results obtained point to the existence of a set of robust determinants in sport performance. This set of drivers consists of (i) the assists, (ii) the shots conceded, (iii) the saves made by the goalkeeper, (iv) the number of precise passes with respect to the total number of passes, and (v) the shots on target. The results of the study support the idea that offensive actions are more relevant than defensive ones. In addition, we find the existence of some performance indicators that have usually been ignored by previous analyses such as the saves made by the goalkeeper and the assists. These findings could help the decision-making process of the coaching, scouting and managerial units of football clubs. Finally, the modelling techniques employed in this context can be generalized to gain knowledge in other fields of knowledge to extract factors affecting complex problems from large data set. This could be particular interesting when previous research has not yet obtained a well-defined and robust set of factors explaining these complex problems.

Keywords: performance analysis; sport management; soccer leagues; sport success; Bayesian model averaging; relative importance analysis.

1. Introduction

The worldwide sports events market defined as all ticketing, media and marketing revenues for major sports was 90.9 billion $ in 2017 (Statista, 2018). Football (soccer, in USA) is in the top spot in terms of relevance given that it accounts for 43% of the market share, above the rest of the sports. In Europe, football has a strong historical tradition and the leagues are highly competitive. The "Big Five" leagues (the English *Premier* League, the German *Bundesliga*, the Spanish *Liga*, *Serie A* Italian *Calcio*, and the French *Ligue 1*), are responsible for 54% of the revenues in the market of football (Deloitte, 2017). The main product of the football industry are the leagues or the championships, while the consumer are the fans who buy (i) stadium attendance, (ii) merchandising products and (iii) broadcasts. Fans and spectators are a key element of the success of sport contests (Mason, 1999). According to some studies, club identification and the win/lose phenomenon are the most considerable influences on the satisfaction of spectators (Byon, Zhang, & Baker, 2013; Scelles et al., 2017). Thus, increasing the understanding of the determinants of winning or losing is of major importance for football clubs.

A number of decision-making units within a football club are responsible for developing the conditions to win a contest. Coaching staff, scouting departments, and management are key entities of a professional organization, all of which must work together to build a successful team (Young II, 2010). Coaching staffs create playing philosophies and tactics using the talents of their players. The aim of the scouting department is to identify the skills that a player possesses, which later can be used to make a hiring decision. On the other hand, the responsibility of management is to ensure that each entity of a football club is working together towards a common goal. These units have to carry out collaborative efforts among themselves to determine successful global strategies.

As explained by Schumaker, Solieman, and Chen (2010) it is of paramount importance that the right decisions are made in order to maintain a competitive advantage. In this regard, among the biggest decisions a team faces, we find the selection of the players and the design of a strategy. Several considerations must be made by managers in order to make hiring decisions (i.e, skill, age,

media impact, complementarity with the squad, etc) or by the coach when developing and designing a strategy (i.e, direct play style, possession game, positional defense, etc).

The decisions of these organizational units can be influenced by the data, the challenge being to find ways to discover knowledge buried in the available data. The advances in information technologies have made it possible to collect, store and process massive and complex datasets. All this data holds valuable information such as trends and patterns, which can be used to improve decision making and increase chances of success (Cortez et al., 2009; MacHale & Relton, 2018). In fact, gaining a deeper understanding of the key determinants of positive sport results and the existing relationships between the countless actions during a match is not only a very important step towards a more predictive and prescriptive performance analysis but it is also a necessary condition to win matches, attract fans and increase incomes. However, despite the growing availability of football performance indicators, decision-making by the different units in a football club in most cases is still not supported by statistical models or a scientific process. Only recently, some empirical applications such as that of Schumaker Jarmoszko, and Labedz (2016) have shown that it could be possible to use data that is external to the playing field to forecast results.

The present study aims to show that computer driven mathematical and recent statistical modelling methods employing past statistics data can be used to provide European football teams' managers with additional information to improve their hiring decisions. Similarly, substantial knowledge gains can be obtained by the coaching staff to design and implement more effective strategies. Moreover, these techniques can be generalized to gain knowledge in other fields of knowledge where research has not yet obtained a well-defined and robust set of factors driving observable phenomena. Hence, the statistical methods employed here may allow researchers or decision-making units in many different contexts other than sports, to benefit from gaining a deeper understanding on the underlying drivers (and their relevance) of a given process. This knowledge could help to make suitable decisions in organizations, improving their decision support system.

An important issue to highlight at this point, is that in the case of football, previous sport performance research has not identified a clear set of indicators determining the main actions that

distinguish winners from losers. In this regard, the main criticisms of previous analyses on the determinants of sport performance, highlighted by Mackenzie and Cushion (2013), Carling et al. (2014), and Sarmento et al. (2014), are mainly methodological and refer to (i) the sample size, (ii) the set of variables considered and their definition, and (iii) the statistical methods employed to perform inference. Additionally, this strand of research usually fails to support decisions and to derive implications useful for football clubs.

The paper makes several novel contributions to the literature of football performance analysis and that of decision support systems in the management of sport teams. The proposed methodological approach can be important for the football industry given that (i) it is data driven and (ii) it can be integrated into a decision support system aiding the speed and quality of decision making. Furthermore, this paper solves the aforementioned methodological limitations in sport performance studies by analysing the relative importance of performance indicators in the final sports result through (i) the consideration of a greater set of determinants, (ii) a greater sample of observations, and (iii) the use of an innovative modelling methodology.

First, we analyse sport performance employing a set of 24 possible explanatory variables, which contrasts with the limited set of controls employed in the literature. Moreover, instead of restricting our study to a single regression model estimation, we perform inference based on a Bayesian Model Averaging (BMA) econometric analysis. In particular, we use the Monte Carlo Markov Chain Model Composition (MC^3) methodology for linear regression models developed by Madigan, York, and Allard. (1995). This analysis aims to compute the posterior inclusion probability (PIP) for the different variables in order to generate a probabilistic ranking of relevance for the various sport performance determinants.

Secondly, the sample used in this study includes a greater number of observations (i.e., teams) than most previous studies, which helps to obtain representative results of modern high-competition

football[1]. Therefore, to generalise our results to competitions with a high competitive level, we analyse the major national leagues of European football, the so-called "Big Five" during the period from 2012/13 to 2014/15. Notice that this implies our sample data cover 5,532 games. Also, we develop analyses for each league in order to detect differences among the five football leagues in terms of the key performance indicators.

Third, we complement the BMA analysis with a relative importance analysis. Assigning shares of relative importance to each or to a set of regressors is one of the key goals of researchers in applied studies and in sciences that work with observational data. Advances in computational capabilities have led to increased applications of computer-intensive methods like averaging over orderings that enable a reasonable decomposition of the model variance. Thus, in a second phase, relative importance metrics allowing for all possible causal patterns among the regressors are computed (Grömping, 2007). These metrics perform an R^2 decomposition enabling more detailed analysis of the relative contribution of each variable to sport performance differentials than previous decompositions.

After this introduction, the literature review is briefly presented in section 2. Section 3 describes the data used to analyse sports performance in the major European football leagues. Section 4 explains the modelling methodology. Section 5 discusses the main empirical findings of the paper, and Section 6 offers the main conclusions to be drawn from this work.

2. Literature Review

As highlighted in the introduction section, in spite of the increase in performance analysis research in recent decades, there is still no consensus about key performance indicators and how those impact the performance of football clubs.

The literature has focused on regular leagues such as the English *Premier* league (Carmichael, Thomas, & Ward, 2000; Oberstone, 2009; Vecer, 2014), the Spanish *Liga* (Lago-Ballesteros & Lago-

[1] The only exception is Collet (2013), who employed a data set covering 5,478 regular national league games, 395 UEFA Champions League games, and 205 Europe League games. However, an important drawback of this study is that to explain PIs such as the points, the goals, etc., only two regressors are employed (possession time and passing).

Peñas, 2010; Villa & Lozano, 2016), the Italian *Serie A* (Boscá et al., 2009), the French *Ligue One* (Collet, 2013), the German *Bundesliga* (Tiedemann, Francksen, & Latacz-Lohmann, 2011), and knockout competitions such as the UEFA Champions League and the UEFA Europe League (Collet, 2013; Barreira et al., 2014; Zambom-Ferraresi et al., 2017), besides the FIFA World Cup (Castellano Casamichana, & Lago, 2012; Delgado-Bordonau et al., 2013; Hughes & Franck, 2005; Moura, Martins, & Cunha, 2014).

Empirical research on the existing differences among styles of play across leagues is scarce. The findings of Boscá et al. (2009) indicated that to improve league ranking in Spain, the best-rewarded strategy is to improve offensive efficiency when playing at home, followed by increased offensive efficiency when playing away from home. In contrast, in order to obtain a better classification in the Italian league, it is more important to improve defensive, rather than offensive, efficiency. Considering these differences, after an overall analysis of the 'Big Five' main determinants of sports performance, we will analyse individual leagues.

The main samples, methodologies, and performance indicators employed by other studies related to our focus can be observed in table 1. A common drawback in many of these studies is the reduced sample size. Small sample size entails problems of generalisation and implies a low number of degrees of freedom, which could negatively affect the quality of statistical estimates. Examples of studies suffering from this problem included in table 1 are those of Barreira et al. (2014), Castellano et al. (2012), Delgado-Bordonau et al. (2013), Hughes and Franck (2005), and Moura et al. (2014), in which the coverage of games is limited.

INSERT TABLE (1) ABOUT HERE

Second, sport performance literature focusing on football has employed limited sets of variables, which is likely to create artificially narrow confidence intervals ignoring the uncertainty surrounding the true model or data generating process (DGP). Moreover, the omission of relevant explanatory variables that could affect sport performance patterns is of major importance from an econometric perspective, given that estimates may be inefficient and/or biased. The consequences of

biased and/or inefficient estimators include results with restricted reliability. This problem appears in a number of studies, such as those of Collet (2013) and Vecer (2014), in which the lack of controls is likely to create biased estimates.

Third, although the univariate tests and the ANOVA analyses in Lago-Ballesteros and Lago-Peñas (2010), Lago-Peñas et al. (2010), and Lago-Peñas and Lago-Ballesteros (2011) provide insights on the characteristics of different types of teams, they do not help to explain to what extent a variable is responsible for sport success in football. Similarly, the conventional regression analyses employed by Carmichael et al. (2000) and Vecer (2014), whenever regressors are correlated among themselves, as is likely to be the case, will fail to obtain precise estimates of importance. This is because in the case of correlated determinants, there is no obvious way to analyse how the fitted variability of the model can be decomposed across regressors (Grömping, 2007).

3. Data

Our sample of data is composed of three seasons ranging from 20012/13 to 2014/15 of the "Big Five", which implies data coverage of 5,532 games in total. The data source is the OPTAPro, a company with one of the largest sports databases of European football and whose data reliability has been previously tested by Liu et al. (2013).

To analyse sport performance, we take as our outcome variable the number of points of each of the teams in each league and season. However, problems may arise when comparing sport performance across teams and leagues given that different leagues have different numbers of teams, which implies the scores of leagues with more teams/games are likely to be higher than in the case of leagues with fewer teams. This is the case of the Bundesliga with 18 participants per season, while the other four leagues have 20 clubs playing the competition by season. To solve this problem, we apply a max–min normalisation to our raw data by scaling the total points between 0 and 1. This normalisation maintains the final ranking and the variability of the data, allowing us to homogenise the points of the different leagues and allowing us to perform comparisons across leagues. In particular, the normalised indicator of sport performance for each team *i* at period t, $I_{i,t}$, is calculated as:

$$I_{it} = \frac{e_{it}^{l} - e_{min,t}^{l}}{e_{max,t}^{l} - e_{min,t}^{l}} \quad (1)$$

where $e_{min,t}^{l}$ denotes the minimum score in points in league l during the season t, $e_{max,t}^{l}$ stands for the maximum score of any team in league l during season t, and e_{it}^{l} is the score of team i in league l during season t.

To explain differentials in football performance across clubs we have selected twenty-four variables based on the literature review. These variables have been grouped distinguishing between defense and attack. In turn, it is important to note that within the set of variables considered, some specific variables capture aspects of the game related to efficiency, while others capture aspects related to the total number of actions developed. Table 2 shows the descriptive statistics and operational definitions of these variables employed in the analysis. In addition, in Table 2 we include a column in which we provide information on the expected effect based on a review of the literature.

INSERT TABLE (2) ABOUT HERE

4. Empirical Methodology

To analyse the determinants of sport performance, we begin by considering a *linear regression model* given by Equation 2:

$$y = \alpha\iota_{nt} + X\beta + \varepsilon \quad (2)$$

where *y* denotes a *NT*x1 dimensional vector consisting of observations for the normalised sport performance index for each team *i* = 1,…, *N* and period *t* = 1,…, *T*, *X* is an *NT* x K matrix of exogenous aggregate covariates with associated response parameters β contained in a *K* x 1 vector. α reflects the constant term, and l_{nt} is an *NT* x 1 vector of ones. Finally, $\mathcal{E} = (\mathcal{E}_1, \ldots, \mathcal{E}_N)$ is a vector of i.i.d disturbances whose elements have zero mean and finite variance σ^2.

4.1. Bayesian Model Averaging (BMA)

A large literature on BMA in regression models already exists (for detailed reviews on the literature, see Fragoso & Neto, 2015; Hoeting et al., 1999; Moral-Benito, 2015). The key feature of this econometric procedure is that it eliminates the need to consider all possible models by

constructing a sampler that explores relevant parts of the large model space. Hence, contrary to previous studies on sport performance in which inference is based on single econometric model analysis, the BMA approach has the advantage of minimising the likelihood of producing (i) biased estimates and (ii) artificially low confidence intervals (Moral-Benito, 2015). To get an intuition behind the BMA approach, notice that for any set of possible explanatory variables of size *K*, there are a total of 2^K candidate models to be estimated, indexed by M_k, for $k = 1,..,2^K$ to explain the y data. Each model M_k depends upon parameters δ^k. This implies there are 2^K sub-structures of the model in Equation 2 given by subsets of coefficients $\delta^k = (\alpha, \beta^k)$ and combinations of regressors X_k. Hence, there are many different candidate models for estimating the effect of X_j on y with $j \epsilon K$. In this circumstance, one can either i) select a single model base and make inference using that selected model, ignoring the uncertainty surrounding the model selection process, or ii) estimate all candidate models and then compute a weighted average of all the estimates for the coefficient of X_j. In the second context, the researcher considers not only the uncertainty associated with the parameter estimate conditional on a given model, but also the uncertainty of the parameter estimated across different models. In particular, BMA inference on the parameters $\eta=(\delta,\sigma)$ is based on probabilistic weighted averages of parameter estimates of individual models:

$$p(\eta|y,X) = \sum_{k=1}^{2^K} p(\eta_k|M_k,y,X)p(M_k|y,X) \quad (3)$$

The weights and the posterior model probabilities (PMPs) are given by:

$$p(M_k|y,X) = \frac{p(y,X|M_k)p(M_k)}{\sum_{k=1}^{2^K} p(y,X|M_k)p(M_k)} \quad (4)$$

Model weights can be obtained using the marginal likelihood of each individual model after eliciting a prior over the model space. The marginal likelihood of model M_k is given by[2]:

[2] In particular, we employ a normal-gamma conjugate prior for $\delta = [\alpha, \beta]$ and σ :
$p(\delta) : N(c,\Sigma)$
$p\left(\frac{1}{\sigma^2}\right) : \Gamma(d,v)$

However, $p(\delta_k)$ is adjusted following the convention in BMA analysis by means of the g-prior hyper-parameter, which takes the value of $g_k = 1/max\{n, K^2\}$ such that: $p(\delta_k)(\delta_k \mid \sigma^2) : N\left[0, \sigma^2\left(g_k X_k^{'} X_k\right)^{-1}\right]$ The employment of the g-prior scales in the variance of the coefficients in δ_k reflects the strength of the prior.

$$p(y, X|M_k) = \int_0^{\infty} \int_{-\infty}^{\infty} p(y, X|\delta, \sigma, M_k) d\delta d\sigma \quad (5)$$

Inference on parameters of the model relies on the computation of the posterior mean (PM) and the posterior standard deviation (PSD).

$$E(\eta|y, X) = \sum_{k=1}^{2^K} E(\eta_k|M_k, y, X) p(M_k|y, X) \quad (6)$$

$$PSD = \sqrt{Var(\eta|y, X)} \quad (7)$$

where the $Var(\eta|y, X)$ is given by:

$$Var(\eta|y, X) = \sum_{k=1}^{2^K} Var(\eta_k|M_k, y, X) p(M_k|y, X) + \sum_{k=1}^{2^K} \left(E(\eta_k|M_k, y, X) - E(\eta|y, X)\right)^2 p(M_k|y, X) \quad (8)$$

where the first term reflects the variability of estimates across different regression models, and the second term captures the weighted variance across different models. Additionally, it is possible to compute the conditional posterior positivity of a parameter *h* as:

$$p(\eta_h \geq 0|y, X) = \sum_{k=1}^{2^K} p\left(\eta_{k,h}|M_k, y, X\right) p(M_k|y, X) \quad (9)$$

where values of conditional positivity close to 1 indicate that the parameter is positive in the vast majority of considered models. Conversely, values near 0 indicate a predominantly negative sign. Finally, with the aim of generating a probabilistic ranking of relevance for the various sport performance determinants, we compute the PIPs for a variable h as the sum of the PMPs including the variable h:

$$PIP = p(\eta_h \neq 0|y, X) = \sum_{k=1}^{2^K} p(\eta_k|M_k, y, X) p(M_k|\eta_k \neq 0, y, X) \quad (10)$$

In the BMA analysis, rather than estimating the 2^K possible models, we will work with a relevant sub-sample of the model space drawn by means of the MC^3 algorithm developed by Madigan et al. (1995). The algorithm to sample models relies on the following acceptance rule to explore the model space:

$$P = min\left[1, \frac{p(M'|y)}{p(M|y)}\right] \quad (11)$$

Lastly, we employ a binomial prior on the model space $p(M_k) = \phi^k (1-\phi)^{K-k}$, where each covariate k is included in the model with a probability of success ϕ. We set $\phi = 1/2$, which assigns equal probability $p(M_k) = 2^{-K}$ to all models under consideration.

where $p(M|y)$ denotes the probability of model M (i.e., the current model) and $p(M'|y)$ denotes the probability of an alternative model M'. Thus, if, $p(M'|y) > p(M|y)$ the sampler will move to model M'. The vector of log-marginal values for the current model M and the proposed alternative models M' are scaled and integrated to produce Equation (6).

4.2. Relative Importance Metrics

In order to complement the BMA analysis, we explore the relative importance of the various factors that could affect sport performance. To that end, we study the relative contribution of the various factors with the LMG method (Grömping, 2007; Lindeman, Merenda, & Gold, 1980), the Genizi, and the CAR scores (Genizi, 1993; Zuber & Strimmer, 2010, 2011). The decomposition procedures used in each of these metrics are detailed below.

Let the variance of the dependent variable Y be given by σ_y^2, the variance of the set of regressors contained in X be denoted by Σ, and the covariance of Y and the covariates by Σ_{yX}. Let P denote the correlations among regressors and P_{yX} marginal correlations between regressors and Y, such that:

$$\sum = V^{\frac{1}{2}} P V^{\frac{1}{2}} \tag{12}$$

and

$$\sum_{yX} V^{\frac{1}{2}} P_{yX} V^{\frac{1}{2}} \tag{13}$$

where $V = diag\big(Var(X_1), \dots, Var(X_{p)}\big)$. Defining the correlation between the model estimates and Y as $\Omega = corr\big(Y, \hat{Y}\big)$, then the squared multiple correlation coefficient is expressed as:

$$R^2 = \Omega^2 = P_{yX} P^{-1} P_{Xy} \tag{14}$$

Then, the unexplained variance can be written as $\sigma_Y^2(1 - \Omega)$ and the explained variance of a model with X_k regressors with indices in the set S as $evar_s = \left\lfloor \sigma_Y^2 \Omega \right\rfloor_{X_k, k \in S}$. Finally, the sequential added explained variance when adding the regressors with indices in M to a model that already contains the regressors with indices in S can be written as $svar = \left\lfloor \sigma_Y^2 \Omega \right\rfloor_{M \cup S} - \left\lfloor \sigma_Y^2 \Omega \right\rfloor_S$. This implies that the true coefficient of determination is given by:

$$R^2 = \Omega^2 = \frac{evar_{(s)}}{\sigma_y^2} \tag{15}$$

With these definitions in hand for any model with p regressors, the r-squared can be expressed as:

$$R^2 = \Omega^2 = \sum_{k=1}^{p} \phi^m X(k) \tag{16}$$

where m denotes the decomposition method. The LMG method assigns to each regressor X_k the following share:

$$\phi^{LMG} X(k) = \frac{1}{p} \sum_{i=0}^{p-1} \left(\sum_{S \subseteq k+1,\ldots,p, n(S)=i} \frac{svar(\{k\}|S)}{\binom{p-k}{i}} \right) \tag{17}$$

where $svar$ denotes the sequentially added explained variance as defined above. Thus, the share $\emptyset_k$ assigned to regressor k is the average over model sizes i of average improvements in explained variance when adding regressor k to a model of size i that did not contain k. Hence, the LMG metric performs a R^2 decomposition by averaging marginal contributions of independent variables over all orderings of variables and using sequential sums of squares from the linear model, the size of which depends on the order of the regressors in each particular model. Finally, to check the robustness of our results, we also compute two alternative metrics of relative importance: (i) the Genizi (1993) and the (ii) CAR scores. The weights associated to the Genizi (1993) and CAR measures are given by:

$$\theta^{GEN} Z(k) = \sum_{p=1}^{p} \left[\left(P^{\frac{1}{2}} \right)_{kp} \left(P^{\frac{-1}{2}} P_{Xy} \right)_p \right]^2 \tag{18}$$

and

$$\phi^{CAR} Z(k) = \omega_k^2 \tag{19}$$

with $\omega = P^{\frac{-1}{2}} P_{Xy}$.

5. Results and Discussion

5.1. Key Sport Performance Indicators for the Big Five Leagues

Table 3 reports the results obtained when implementing the MC^3 algorithm for the 5,000 top models out of the 8,149 generated by the sampler, where the number of draws to carry out the sampling exercise on the model space was 100,000. The concentration of the posterior density in this context was high, given that the top 1% of models concentrate 52% of the mass, whereas the top 5% concentrate 76%. We scale the PIPs of the different variables to classify evidence of robustness of

inequality regressors into three categories so that regressors with PIP $\in [0-20\%]$ are considered as weak determinants, with PIP $\in [20-80\%]$ of medium importance, and with PIP $\in [80-100\%]$ as very important.

Column 1 show the PIPs, while Columns 2 to 5 show the mean and the standard deviation of the posterior parameters' distributions, along with the lower and upper bounds, conditional on the variable being included in the model[3]. To complement these statistics, Column 6 reports the fraction of models where the t-stat of the corresponding variables is higher than 1.96 (which implies statistical significance at the 5% level), while Column 7 presents the results of the posterior sign certainty, which measures the posterior probability of a positive coefficient expected value, conditional on inclusion.

As observed in Column 1, there is a set of top variables that appears with high frequency in the group of very important determinants. The assists, the shots conceded, the saves made by the goalkeeper, and the passing accuracy appear to be the most relevant determinants and, in all cases, display PIPs of 99.9%. In the range of medium importance, we find the number of clearances blocks and interceptions (58%), the shots on target (23%), and the total number of fouls conceded (21%). On the other hand, the group of weak sport performance determinants with PIPs below 20% consists of a myriad of factors. Therefore, for the remainder of the paper, we will discuss only the results for the regressors with a PIP above 20%.

Column 6 shows that for the group of very important determinants, the variables appear to be significant at the 5% level in all of the regression models. On the other hand, the statistical significance of the regressors included in the group of medium relevance oscillates between 93% of the regression models in the case of the clearances, blocks, and interceptions and 34% of the models in the case of the fouls conceded. As shown in Column 7, the effects of the determinants of higher and

[3] The key difference with respect to unconditional posterior estimates of Equations 6 and 7 is that conditional posterior estimates for a particular variable are obtained as the weighted average over the models in which the variable is included. On the contrary, the unconditional posterior estimate is the averaged coefficient over all models, including those in which the variable does not appear, hence having a zero coefficient. Thus, the unconditional PM can be computed by multiplying the conditional mean in Column 3 times the PIP in Column 1.

medium levels of importance are robust across regression models and display the same sign of the PM in all cases. Among the top determinants of sport performance, only the number of shots conceded displays a negative effect, while the determinants that have a positive effect on sport performance are (i) the assists; (ii) the saves made; (iii) the passing accuracy; (iv) the number of blocks, clearances, and interceptions; (v) the shots on target; and (vi) the total fouls conceded.

INSERT TABLE (3) ABOUT HERE

One of the advantages of including regressors that capture both the efficiency in a type of behaviour or style of play (i.e., passing accuracy) and the intensity of this (i.e., total passes) is that by analysing the values of the PIPs we can see whether it is the brute force or the accuracy that matters. For the most remarkable determinants, the performance indicators related to efficiency/accuracy take a higher probability of inclusion than their absolute counterparts. In the context of passes, we find that passing accuracy (100%) displays a higher PIP than the total number of passes (11%). Similarly, regarding shots, we find that shots on target (23%) appear to be more relevant than the total number of shots attempted (10%). These results indicate that among the determinants with high PIP the accuracy/efficiency ratios are more important than the total actions performed.

Table 4 reports the results of the analysis of relative importance. For a proper interpretation of the R^2 decomposition performed, recall that in the context of a linear regression model the R^2 informs on the model's explained variability across observations. Thus, decompositions on the relative importance of a factor X^k tell us the percentage of explained disparities in sport performance across the observations by k. In the present context, the, $R^2 = 0.88$, while the unexplained variability is $\sigma_Y^2(1 - \Omega) = 0.1\,2$, which implies our decomposition explains most of the differences in sport performance across teams and seasons. Given that results produced by them were similar, we will discuss just the average share reported in the last column of Table 4.

The first salient feature of the relative importance decomposition is that the variability in sport performance that can be attributed to attack actions is 62%, while the sum of defense variables accounts for 38% of the differences in sport performance, which suggests that attack actions are more

relevant than defense ones. Among the set of attack factors, we find the most relevant factors are, in decreasing order, the assists (18%), the shots on target (10%), the passing accuracy (8%), the total passes (7%), and the total shots attempted (6%). However, the most relevant factor is a defensive one: the shots conceded. This factor explains by itself 21% of the sport performance of a football team. In a lower level of explanatory power of the sport performance, we find the saves made (6%) and the fouls conceded (3%). Notice that these results imply that relative importance metrics produce a similar group of factors to that suggested by the BMA analysis. The most remarkable differences can be seen in the fact that relative importance analysis attributes a relatively higher share to the total passes and to the total shots and a relatively lower importance to the total fouls conceded when compared to the BMA. Taken together, the two methodologies point to the existence of a set of key variables, such as (i) the number shots conceded, (ii) the assists, (iii) the passing accuracy, (iv) the saves made, and (v) the shots on target.

INSERT TABLE (4) ABOUT HERE

These results support previous analyses in the literature and provide new insights on the relevance for decision making of football clubs. Our findings regarding the positive and relevant effect of assists in the performance is in line with previous discriminant analyses (Lago-Ballesteros and Lago-Peñas, 2010; Lago-Peñas et al., 2010). Second, the relevance of the passing accuracy indicator supports Carmichael et al. (2000) and Oberstone (2009). Regarding defensive actions, two performance indicators appear to be key determinants in our empirical analysis: the shots conceded (Castellano et al., 2012), and the saves made. As far as we know, no previous empirical evidence analysing a set of determinants of sport performance have included the saves made in their modelling; a neglected indicator, which as we have highlighted is an important factor in the sporting success.

Furthermore, the results stemming from the group of medium importance such as the clearances, blocks and interceptions (Carmichael et al., 2000), shots on target (Delgado-Bordonau et al., 2013; Lago-Peñas et al., 2010, 2011; Moura et al., 2014), and total fouls conceded (Oberstone, 2009) are in agreement with the previous literature. Finally, the negative estimated effect of recovery

in the opposite half and the positive effect of all ball recoveries corroborate the results of Barreira et al. (2014), who found recovering directly ball possession in mid-defensive central zones increases attacking efficacy. However, some of the findings in our analysis contrast with those previously found in literature. This is the case in the set of regressors displaying relatively low PIPs. For instance, the indicator measuring the crosses attempts displays a PIP of 15% and has a positive effect on the success of the teams, which contrasts with the results of Vecer (2014).

The results obtained allow us to exemplify that by using statistical modeling techniques, football teams can obtain useful knowledge of what happened in the playing field to improve their performance at different levels. First, the results obtained, offer interesting implications from the point of view of strategic decisions to be developed by football clubs when selecting certain player profiles. Compared to the traditional importance given to the selection of good strikers and forward players, the results stemming from this empirical analysis highlights the importance of sports actions carried out by other players. In the first place, the relevance of shots conceded highlights the need to have defense players that are quick and agile to get ahead of the adversary and not allow him to shoot. Therefore, coaches and managers should select defensive players based on their physical characteristics instead of their technical skills.

Second, the relevance of other key determinants such as the variable assist and the passing accuracy emphasize the relevance of the midfielders, requiring eminently technical players who are safe in their game and make few mistakes (high degree of accuracy). Reinforcing the intuition that midfielders should have a marked technical profile and high precision, we find the fact that the through ball is more relevant than dribbles and runs, that do not appear to be a key determinant. Therefore, scouting units of football teams in which the offensive system is structured on an attacking midfielder (which connects the center of the field with the forwards) should prioritize players with high passing skills rather than dribbling.

Third, our results emphasize the importance of goalkeepers, through the variable "saves made". This result, which has traditionally been ignored in previous studies, coincides with the reality of the

main football teams in Europe, who attach great importance to the presence of a goalkeeper of high quality and safety among their players.

5.2. Robustness Check

The analysis carried out so far suggests the existence of a group of robust determinants of sport performance in the European football league. In the remainder of this section, the robustness of previous findings is investigated.

As a first robustness test, we examine to what extent the results may be sensitive to the choice of the measure used to quantify the sport results in the sample teams. To that end, we test an alternative measure of sport result based on a transformation of the final position in the league such that:

$$SP_{it} = ln\left(\frac{X+1-C_{it}}{C_{it}}\right) \quad (20)$$

where X is the number of teams in the league, and C_{it} denotes the classification of the team i in the league t.

Table 5 summarises the results of the BMA when using the alternative sport result metric. As is observed, (i) the number of shots conceded, (ii) the assists, (iii) the passing accuracy, (iv) the saves made, and (v) the shots on target also appear to be among the top determinants of SP.

INSERT TABLE (5) ABOUT HERE

5.3. Is there a unique play style? Main results and robustness checks by league

An additional issue to examine is to what extent previous findings are specific to the football league that was considered. This section analyses and compares the main results and robustness checks by league. We perform the BMA analysis and the relative importance decomposition for each individual league for the period between the 2012/13 and 2014/15 seasons. Tables 6 and 7 summarise the PIPs by factor and the average share of the R^2 attributed to each factor across metrics, respectively.

As observed in Table 6, the PIPs by factor in the different European football leagues are very similar, which implies that the primary determinants we identified previously using BMA analysis are robust across leagues. Assists, number of shots conceded and saves made all have a PIP of 100% in all leagues. However, passing accuracy does not display a high PIP value for each of the individual leagues. While the PIPs in the Bundesliga closely follow the overall aggregate European PIP values, without non-highlighted differences, there are some interesting differences in the actions related to success across the other leagues, which ultimately imply that the strategies for success and playing styles are different across countries.

INSERT TABLE (6) ABOUT HERE

The individual results obtained regarding the key determinants of sports performance in the Premier League reinforce the notion that the corners taken in the Premier League are more closely related with success than in other leagues. Nevertheless, the most remarkable performance indicator that differentiates the Premier League play style from the others is how consistently the English teams finished plays, through the shots. The shot on target, with almost 100% of PIP, and the related total shots attempts (37%) are highly important when compared with the results of the other four leagues.

A separate analysis of the Liga reveals that there are five performance indicators differentiating sport success and the play style of this competition. Clearances, blocks and interceptions, and total fouls conceded were determinants with medium importance in the overall analysis. When we analyse only the Liga, their importance increases from 58% to 93% and from 21 to 66% respectively. Other sports indicators such as recoveries in the opposite half (62%), dribbles and runs success rate (29%) and red cards (21%) also have more impact on sports performance than in the analysis of the 'Big Five'. Thus, four out of five highlighted determinants of the Liga are defensive, which is not in line with the colourful and attractive play style displayed by some Spanish teams in the last decade such as the FC Barcelona. This play style, characterized by high ball possession and a specific structure of passing sequence[4], contrasts with the other teams of the Liga. Our findings regarding the sports

[4] Gyarmati et al. (2014) have proposed a quantitative method to evaluate the styles of football teams through their passing structures. The analysis of the motifs in the pass networks allow them to compare and differentiate

performance in the Liga suggest that differences in the rankings for the majority of the teams can be explained by more efficient defensive tactics and a higher quality in the execution of the wide range of defensive actions. The high importance observed in the recoveries in the opposite half poses important implications for coaches who aim to achieve success in the Spanish competition. Specifically, this suggests that the implementation of an aggressive and risky defense, in which the lines of defense are placed in an advanced location is likely to increase sport performance. The reason is that this type of defense allows the execution of counter-attacks easily than other defense systems. However, this type of strategy is quite risky since it also generates vulnerabilities. Moreover, to be implemented correctly, this type of strategy requires sustained pressing over time, which is very physically demanding for players/both the forwards and the midfielders.

Regarding the results of the Serie A, two defensive variables appear to be more important in this league than in the rest of the leagues: tackle attempts (36%) and recoveries (26%). This result is consistent with Boscá et al. (2009) findings, where they find that to obtain a better classification in the Italian league, it is much more important to improve defensive efficiency rather than offensive efficiency. As a consequence, Italian teams might have developed a more direct attacking style, which is in agreement with the fact that two very specific attack indicators are more relevant in the Italian league than in the rest of leagues, i.e., through ball (61%) and offsides (24%). From the tactical point of view, our results suggest that coaches in Serie A should promote a style of game based on the concentration of the ball possession in midfield or even in defensive areas.

In the Ligue 1, fouls conceded in dangerous area (54%) and red cards (25%) are clearly above the other European leagues under study. This result might reflect a more physical and aggressive play style in the French league than in the other leagues. The probabilistic importance given to red cards and faults conceded in danger areas, suggests that, coaches in League 1, may consider developing highly cooperative and aid-based defenses in order to avoid high-risk situations in which defenders are forced to make faults that could leave the team in numerical inferiority.

---

the styles of different teams. Although most teams tend to apply homogenous style, surprisingly, a unique strategy of soccer is also viable—and quite successful, as we have seen in the recent years. Their results shed light on the unique philosophy of FC Barcelona quantitatively: does not consist of uncountable random passes but rather has a precise, finely constructed structure.

INSERT TABLE (7) ABOUT HERE

Finally, table 7 also aids in identifying the overall most important determinants of sports performance and the differences by league. The most notable differences identified in this analysis are: (i) the importance of shots on target in the Premier League, (ii) that shots conceded are less important across leagues than to the 'Big Five' and (iii) that fouls conceded in dangerous areas have the highest impact in Ligue 1.

6. Conclusions and general implications for decision making

This study analyses the relative importance of a large number of possible determinants of football performance during the period 2012/13–2014/15 for the most important European leagues. This paper makes two key contributions: methodological and empirical contributions. Firstly, we consider the effect of a great number of determinants employing two innovative methodologies in this context: the BMA technique and relative importance metrics analysis. These methods enable us to compute the PIPs for the different indicators to generate a probabilistic ranking of relevance for the various determinants driving success and decompose the R^2 of the model. These modelling techniques can be useful to gain knowledge in other decision support system contexts where there are complex problems and large datasets. Secondly, our empirical results reveal a set of robust determinants of sport performance in football. These performance indicators consist of (i) the assists, (ii) the shots conceded, (iii) the saves made by the goalkeeper, (iv) the passing accuracy, and (v) the shots on target. Moreover, we find strong support for the idea that offensive actions are more relevant than defensive ones. However, the indicator that exerts a stronger influence on the differentials of performance across football teams is a defensive indicator (shots conceded).

There are important implications of our findings that could be useful as inputs for the decision-making units in football teams. The first is related to tactics (for coaches) and techniques (for players). Based on our observation that assists and through balls are much more important than dribbles, runs, and crosses; hence, improving the technical and tactical execution in these plays is essential. However, we also observed that accuracy is more important than the amount of executions. The second main

implication is related to the clubs' management section. When hiring players, managers should consider hiring players with skills and abilities associated with those determinants that have an impact on the team success, in light of the possible combinations of players that the team already has and/or lacks. For example, considering the importance of shots conceded and saves made, if a team already has a relatively high value for shots conceded, a good strategy could be to increase the quality of the goalkeeper. Also, it should be highlighted that according to the results, the same style of plays and player characteristics are not adequate for all the leagues under study. It should therefore be mandatory for coaches and sports managers to know the peculiarities of each league in order to optimally select tactics and players.

More in general, the two modelling methodologies and statistical approaches employed in this research could be used to produce knowledge on the relevance of the determinants of other complex and multifaceted processes given that they allow the researcher to extract factors affecting complex problems for large datasets. This is a relevant contribution to the field of decision support systems as in many contexts and environments there is a substantial degree of uncertainty on the true determinants behind observable phenomena. Even if the methods are computationally intensive, we consider that this type of analysis could be integrated in the phase of data processing and knowledge creation of a wide spectrum of decision support systems and not only in the field of sport analytics.

References

Barreira, D., Garganta, J., Guimarães, P., Machado, J., & Anguera, M. T. (2014). Ball recovery patterns as a performance indicator in elite soccer. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, *228*(1), 61-72.

Boscá, J. E., Liern, V., Martínez, A., & Sala, R. (2009). Increasing offensive or defensive efficiency? An analysis of Italian and Spanish football. *Omega*, *37*(1), 63-78.

Byon, K.K., Zhang, J.J., & Baker, T.A. (2013). Impact of core and peripheral service quality on consumption behavior of professional team sport spectators as mediated by perceived value. *European Sport Management Quarterly, 13*(2), 232-263.

Carling, C., Wright, C., Nelson, L. J., & Bradley, P. S. (2014). Comment on 'Performance analysis in football: A critical review and implications for future research. *Journal of Sports Sciences*, *32*(1), 2-7.

Carmichael, F., Thomas, D., & Ward, R. (2000). Team performance: the case of English premiership football. *Managerial and Decision Economics*, 21(1), 31-45.

Castellano, J., Casamichana, D., & Lago, C. (2012). The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. *Journal of Human Kinetics*, 31(1), 137-147.

Collet, C. (2013). The possession game? A comparative analysis of ball retention and team success in European and international football, 2007-2010. *Journal of Sports Sciences*, 31(2), 123-136.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modelling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547-553.

Delgado-Bordonau, J. L., Domenech-Monforte, C., Guzmán, J. F., & Mendez-Villanueva, A. (2013). Offensive and defensive team performance: Relation to successful and unsuccessful participation in the 2010 Soccer World Cup. *Journal of Human Sport and Exercise*, 8(4), 894-904.

Deloitte (2017), "Annual Review of Football Finance 2017", available at: https://goo.gl/T58dsy (accessed November 6, 2017).

Fragoso, T. M., & Neto, F. L. (2015). Bayesian model averaging: A systematic review and conceptual classification. arXiv preprint arXiv:1509.08864.

Genizi, A. (1993). Decomposition of R2 in Multiple Regression with Correlated Regressors. *Statistica Sinica*, 3, 407-420.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2), 139-147.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382-401.

Hughes, M., & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences*, *23*(5), 509–514.

Lago-Ballesteros, J., & Lago-Peñas, C. (2010). Performance in Team Sports: Identifying the Keys to Success in Soccer. *Journal of Human Kinetics*, 25(1), 85-91.

Lago-Ballesteros, J., Lago-Peñas, C., & Rey, E. (2012). The effect of playing tactics and situational variables on achieving score-box possessions in a professional soccer team. *Journal of Sports Sciences*, 30(14), 1455-1461.

Lago-Peñas, C., & Lago-Ballesteros, J. (2011). Game location and team quality effects on performance profiles in professional soccer. *Journal of Sports Science and Medicine*, 10(3), 465-471.

Lago-Peñas, C., Lago-Ballesteros, J., Dellal, A., & Gómez, M. (2010). Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. *Journal of Sports Science and Medicine*, 9(2), 288-293.

Gyarmati, L., Kwak, H., & Rodriguez, P. (2014). Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308*.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). Introduction to bivariate and multivariate analysis (p. 119ff). Glenview, IL: Scott, Foresman.

Liu, H., Hopkins, W., Gómez, M. A., & Molinuevo, J. S. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. *International Journal of Performance Analysis in Sport*, 13(3), 803-821.

Mackenzie, R., & Cushion, C. (2013). Performance analysis in football: A critical review and implications for future research. *Journal of Sports Sciences*, 31(6), 639-676.

Madigan, D., York, J., & Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, 63(2), 215-232.

Mason, D. S. (1999) What is the sports product and who buys it? The marketing of professional sports leagues, *European Journal of Marketing*, 33 (3/4), 402-419

McHale, I. G., & Relton, S. D. (2018). Identifying key players in soccer teams using network analysis and pass difficulty. *European Journal of Operational Research*, *268*(1), 339-347.

Moral-Benito, E. (2015) Model averaging in economics: An overview, *Journal of Economic Surveys*, 29 (1), 46-75.

Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of Sports Sciences*, *32*(20), 1881-1887.

Oberstone, J. (2009). Differentiating the top English premier league football clubs from the rest of the pack: Identifying the keys to success. *Journal of Quantitative Analysis in Sports*, 5(3), 10.

OPTA (2012). Blog OptaPro's event definitions, and the importance of consistent data. Available in march 9th, 2017: (goo.gl/rXMXX4).

Sarmento, H., Marcelino, R., Anguera, M. T., Campaniço, J., Matos, N., & Leitão, J. C. (2014). Match analysis in football: a systematic review. *Journal of Sports Sciences*, 32(20), 1831-1843.

Scelles, N., Helleu, B., Durand, C., Bonnal, L., & Morrow, S. (2017). Explaining the number of social media fans for North American and European Professional sports clubs with determinants of their financial value. *International Journal of Financial Studies*, 5(4), art. No.: 25.

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). Sports knowledge management and data mining. *Annual Review of Information Science and Technology*, *44*(1), 115-157.

Schumaker, R.P., Jarmoszko, A.T., & Labedz Jr, C. S. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems*, *88*, 76-84.

Statista (2018). "Global sports market - total revenue from 2005 to 2017". Available in May 30th, 2018: (https://www.statista.com/statistics/370560/worldwide-sports-market-revenue/).

Tiedemann, T., Francksen, T., & Latacz-Lohmann, U. (2011). Assessing the performance of German Bundesliga football players: a non-parametric metafrontier approach. *Central European Journal of Operations Research*, 19(4), 571-587.

Vecer, J. (2014). Crossing in soccer has a strong negative impact on scoring: Evidence from the English Premier League, the German Bundesliga and the World Cup 2014 (September 30, 2014). Available at SSRN: https://ssrn.com/abstract=2225728 or http://dx.doi.org/10.2139/ssrn.2225728

Villa, G., & Lozano, S. (2016). Assessing the scoring efficiency of a football match. *European Journal of Operational Research*, *255*(2), 559-569.

Young II, W. A. (2010). A team-compatibility decision support system to model the NFL knapsack problem: An introduction to HEART. *Ohio University, ProQuest Dissertations Publishing*, 3413080.

Zambom-Ferraresi, F., García-Cebrián, L. I., Lera-López, F., & Iráizoz, B. (2017). Performance evaluation in the UEFA Champions League. *Journal of Sports Economics*, *18*(5), 448-470.

Zuber, V., & Strimmer, K. (2010). Variable importance and model selection by decorrelation. Preprint. http://arxiv. org/abs/1007.551

Zuber, V., & Strimmer, K. (2011). High-dimensional regression and variable selection using CAR scores. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 1-27.

Acknowledgements

Fernando Lera-López acknowledges the financial support from the Spanish Ministry of Education and Research (Project ECO2017-86305-C4-4-R). Fernando Lera-López and Fabiola Zambom-Ferraresi acknowledge the financial support from Foundation Caja Navarra,Foundation La Caixa and UNED Pamplona (Project 2018-19)

Table 1. Analysis of the Determinants of Football Performance

Study	Sample (N); period (T)	Methodology	Dependent variable
Barreira *et al*. (2014)	N= 4 (24 matches) FIFA WC; T= 1 cup (2010)	ANOVA (one and two way); Multinomial and logistic regression	BR with efficacy BR with no efficacy
Boscá *et al*. (2009)	N= IL; SL; T= 3 seasons (2000/01-2002/03)	DEA	Goals scored (attack output) goals conceded (defense output)
Carmichael *et al*. (2000)	N= PL (380 matches) T= 1 season (1997/98)	Panel data	Difference of goals (scored -conceded)
Castellano *et al*. (2012)	N= FIFA WC T= 3 cups (2002, 2006, 2010)	Discriminant analysis ANOVA; and multivariate analysis	Winning, drawing and losing teams
Collet (2013)	N= BF (5478 matches); N= UEFA CL (395 matches) T= 3 seasons (2007/08-2009/10); N= UEFA EL (205 matches), T= 1 season (2009/10)	2 stages - proportional odds models 1st aggregated team success 2nd individual match level	1st: points/match; goals/match; shots/match; FIFA points 2nd: home loss (-1), draw (0), win (1)
Delgado-Bordonau *et al*. (2013)	N= FIFA WC (54 matches) T= 1 cup (2010)	Student's independent t-test	Reach semi-finals (successful and unsuccessful teams)
Hughes and Franck (2005)	N= FIFA WC (52 and 64 matches) T= 2 cups (1990 - 1994)	Descriptive Ratios	(Un)Successful teams (shots; shots/goal; projected goals)
Lago-Peñas *et al*., (2010)	N= SL (380 matches) T= 1 season (2008/09)	Univariate (t-test) Multivariate discriminant analysis	Winning, drawing and losing teams
Lago-Peñas *et al*., (2011)	N= UEFA CL (288 matches of group stage) T= 3 seasons (2007/08- 2009/10)	One way ANOVA and discriminant analysis	Winning, drawing and losing teams
Lago-Ballesteros and Lago-Peñas (2010)	N= SL (380 matches) T= 1 season (2008/09)	One way ANOVA	3 groups: top 4, middle 12 clubs and bottom 4
Lago-Peñas and Lago-Ballesteros (2011)	N= SL (380 games) T= 1 season (2008/09)	Univariate (t-test and Mann-Whitney U) and multivariate (discriminant analysis)	4 groups (1-5, 6-10, 11-15, and 16-20 of final ranking)
Moura *et al*. (2014)	N= FIFA WC (Group stage); T= 1 cup (2006)	Principal component and cluster analysis	Winning, drawing and losing teams
Oberstone (2009)	N= PL (380 matches) T= 1 season (2007/08)	Multiple regression ANOVA (one-way)	Final league standings; 3 groups: top 4, middle 12 clubs and bottom 4
Villa and Lozano (2016)	N= SL (380 matches); T= 1 season (2013/14)	Network DEA	Goals
Vecer (2014)	N= PL (1780 games); T= (2008-2013)	Regression analysis	Goals

Note: WC= World Cup; CL= Champions League; EL= Europe League; BF= Big Five; BR= ball recovery.

Table 2. Definitions and Descriptive Statistics of the Explanatory Variables

Variable	Definition	Mean	STD	Expected Effect
Outcome Variable				
Sport Performance	Normalised total points archived by clubs at the end of a season	0.4	0.269	
A. Attack plays				
Total Shots Attempted	Shot: An attempt to score a goal, made with any part of the body, either on or off target. The outcomes of a shot could be: goal, shot on target, shot off target, blocked shot, post	367.80	60.83	+
Shots on Target	Total shots on target	164.76	36.42	+
Total Passes, (excl. Crosses, and Corners	Pass: An intentionally played ball from one player to another)	15902.81	2739.50	+
Passing Accuracy	Successful passes/total passes	0.78	0.05	+
(excl. Crosses and Corners)				
Assists	The final pass or cross leading to the recipient of the ball scoring a goal	34.13	12.66	+
Crosses Attempted	Any ball played into the opposition team's area from a wide position	603.79	131.86	+
Corners Taken (incl. Short Corners	A corner kick is a method of restarting play. It is awarded to the attacking team when the ball leaves the field of play crossing the goal line)	192.19	32.62	+
Dribbles and Runs Attempted	An attempt by a player to beat an opponent in possession of the ball. A successful dribble: the player beats the defender; unsuccessful: the dribbler is tackled	746.89	161.07	+
Dribble and Run Success Rate	Effective dribbles and runs with respect to the total number attempted	0.45	0.07	+
Long Pass Final Third	A pass over 32 metres on the final third of the field (attack of the reference team)	931.76	156.55	+
Through Ball	A pass playing a player through on goal, which could lead to a goal scoring opportunity. The pass needs to split the last line of defense and plays the teammate through on goal.	27.60	18.70	+
Offsides	Being caught in an offside position resulting in a free kick to the opposing team	88.71	19.14	?

Table 2. (Continued)

Variable	Definition	Mean	STD	Expected Effect
B. Defense plays				
Total Shots Conceded	Total shots attempted for the opposite team	164.76	29.78	-
Tackles Attempted	The act of gaining possession from an opposition player when he is in possession of the ball	754.60	79.47	-
Tackled Possession Retained (%)	A tackle won when a player makes a tackle and possession is retained by his team	0.23	0.03	+
Recoveries	The event given at the start of a team's recovery of possession from open play. The defending team must have full control of the ball and must start a new passage of play.	2071.00	297.60	+
Recoveries in Opp Half	A recovery on the opposite team's field (attack of reference team)	400.63	93.26	+
Clearances, Blocks, and Interceptions	Attempts to get the ball out of the danger zone when there is pressure. A defensive block, blocking a shot going on target. An interception is given when a player intercepts a pass with some movement	1750.02	246.97	?
Total Fouls Conceded	Any infringement that is penalised as foul play by a referee	517.86	73.38	-
Fouls Conceded in Danger Area	Infringement that is penalised as foul play by a referee in the lower 1/3	106.59	18.87	-
Yellow Cards	Indicates that a player has been officially cautioned/penalised due to infringement. A player receiving two yellow cards in a match is sent off.	75.59	20.28	-
Red Cards	A red card is shown by a referee to signify that a player has been sent off.	4.46	2.66	-
Saves Made	The goalkeeper prevents the ball from entering the goal with any part of his body.	112.22	20.79	+
Catches	The goalkeeper catching a cross or a ball played into the area when there is pressure from the rival	52.11	17.18	+

Notes: Own elaboration; Sources: Liu et al. (2013) and OPTA (2012)

Table 3. Main Results: Model Averaged Estimates

Variable	PIP	Lower 5%	Cond Post. Mean	Cond Post. Std	Upper 95%	T-Stat > 1.96	Sign Pos.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Assists	0.999	0.00868	0.00992	0.00110	0.01098	1.00	1.00
Shots Conceded	0.999	-0.01012	-0.00959	0.00065	-0.00921	1.00	0.00
Saves Made	0.999	0.00820	0.00882	0.00086	0.00945	1.00	1.00
Passing Accuracy	0.998	0.70022	0.90790	0.22871	1.22442	1.00	1.00
Clearances, blocks and interceptions	0.581	0.00005	0.00008	0.00003	0.00010	0.93	1.00
Shots on Target	0.229	0.00056	0.00076	0.00031	0.00117	0.40	1.00
Total Fouls Conceded	0.210	0.00006	0.00017	0.00009	0.00033	0.34	1.00
Recoveries in Opp. Half	0.183	-0.00030	-0.00016	0.00008	-0.00005	0.46	0.00
Recoveries	0.171	0.00002	0.00006	0.00003	0.00011	0.67	1.00
Crosses Attempted	0.154	0.00007	0.00011	0.00004	0.00017	0.63	1.00
Total Passes	0.110	-0.00001	0.00000	0.00000	0.00001	0.25	0.19
Total Shots Attempted	0.103	-0.00005	0.00024	0.00009	0.00053	0.17	0.91
Fouls Conceded (danger area)	0.100	-0.00156	-0.00088	0.00036	-0.00034	0.52	0.00
Tackles Attempted	0.087	-0.00023	-0.00016	0.00005	-0.00008	0.44	0.00
Red Cards	0.056	0.00144	0.00282	0.00082	0.00425	0.03	1.00
Dribble Run Success Rate	0.042	-0.08440	0.02976	0.02350	0.15284	0.01	0.64
Dribbles and Runs Attempted	0.042	-0.00006	-0.00002	0.00001	0.00001	0.00	0.12
Yellow Cards	0.040	-0.00012	0.00018	0.00008	0.00050	0.00	0.85
Through Ball	0.038	-0.00027	0.00014	0.00010	0.00062	0.01	0.69
Corners Taken	0.038	-0.00051	-0.00010	0.00006	0.00017	0.01	0.34
Catches	0.033	-0.00017	0.00004	0.00007	0.00028	0.00	0.56
Long Pass Final Third	0.033	-0.00011	-0.00001	0.00001	0.00005	0.16	0.63
Tackled Poss. Retained%	0.032	-0.00422	0.11059	0.04030	0.24587	0.00	0.94
Offsides	0.031	-0.00023	-0.00008	0.00006	0.00007	0.00	0.21

Notes: The dependent variable in all regressions is the normalized indicator of sport performance based on the points obtained during the season. All the results reported here correspond to the estimation of the top 5,000 models from the 16.77 million possible regressions including any combination of the 24 regressors. Prior mean model size is 12. Variables are ranked by Column (1), the posterior inclusion probability. Columns (2) to (5) reflect the lower 5% bound, the posterior mean, standard deviations and upper 95% bound for the linear marginal effect of the variable conditional on inclusion in the model, respectively. Column (6) is the fraction of regressions in which the coefficient has a classical t-test greater than 1.96, with all regressions having equal sampling probability. The last column denotes the sign certainty probability, a measure of our posterior confidence in the sign of the coefficient.

Table 4. Relative Importance Decomposition: Main Results

Variable	LMG Metric	CAR Scores	Genizi Decomposition	Average Importance
A. Attack	0.626	0.626	0.618	0.624
Assists	0.161	0.226	0.150	0.179
Shots on Target	0.103	0.101	0.090	0.098
Passing Accuracy	0.079	0.084	0.074	0.079
Total Passes	0.079	0.061	0.073	0.071
Total Shots Attempted	0.070	0.048	0.068	0.062
Through Ball	0.049	0.047	0.058	0.051
Long Pass Final Third	0.014	0.011	0.021	0.015
Offsides	0.013	0.012	0.018	0.014
Dribbles and Runs Attempted	0.010	0.005	0.013	0.009
Dribble and Run Success Rate	0.006	0.005	0.009	0.007
Crosses Attempted	0.005	0.004	0.007	0.005
B. Defense	0.374	0.374	0.382	0.376
Shots Conceded	0.182	0.294	0.168	0.215
Saves Made	0.084	0.006	0.085	0.058
Fouls Conceded in the Danger Area	0.030	0.032	0.037	0.033
Clearances, blocks and intercept.	0.017	0.010	0.022	0.017
Recoveries	0.012	0.016	0.016	0.015
Recoveries in Opp. Half	0.016	0.003	0.015	0.011
Total Fouls Conceded	0.011	0.001	0.010	0.007
Yellow Cards	0.006	0.003	0.007	0.005
Red Cards	0.005	0.004	0.007	0.005
Tackles Attempted	0.003	0.002	0.005	0.003
Tackled and Possession Retained %	0.004	0.001	0.005	0.003
Catches	0.003	0.001	0.004	0.003

Notes: The dependent variable in all regressions is the normalized indicator of sport performance based on the points obtained during the season. The decomposition applies to a model with $R^2 = 0.88$ while the unexplained variability is $\sigma_Y^2(1 - \Omega) = 0.12$

Table 5. Dependent Variable Robustness Check (I): BMA Ranking in League

Variable	PIP	Lower 5%	Cond Post. Mean	Cond Post. Std	Upper 95%	T-Stat > 1.96	Sign Pos.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Shots Conceded	1.000	-0.02929	-0.02660	0.00322	-0.02524	1.00	0.00
Saves Made	1.000	0.02436	0.02602	0.00416	0.02855	1.00	1.00
Assists	0.991	0.01535	0.01912	0.00393	0.02218	1.00	1.00
Catches	0.334	0.00268	0.00333	0.00163	0.00384	0.60	1.00
Clearances, blocks and intercept.	0.120	0.00011	0.00020	0.00008	0.00026	0.06	1.00
Recoveries	0.109	0.00008	0.00015	0.00006	0.00024	0.04	1.00
Corners Taken	0.109	0.00050	0.00148	0.00057	0.00377	0.12	1.00
Long Pass Final Third	0.087	-0.00047	-0.00033	0.00010	-0.00025	0.07	0.00
Red Cards	0.086	0.01380	0.01746	0.00603	0.02032	0.01	1.00
Dribble and Run Success Rate	0.069	-0.77155	-0.56249	0.19162	-0.31309	0.01	0.00
Passing Accuracy	0.066	0.48877	1.24473	0.35903	3.18904	0.11	0.99
Shots on Target	0.060	0.00050	0.00381	0.00083	0.00665	0.45	0.97
Total Shots Attempted	0.052	0.00065	0.00159	0.00033	0.00328	0.27	1.00
Crosses Attempted	0.052	-0.00006	0.00010	0.00006	0.00023	0.00	0.86
Offsides	0.050	-0.00150	-0.00098	0.00041	-0.00054	0.00	0.00
Recoveries in Opp. Half	0.039	-0.00058	-0.00013	0.00009	0.00009	0.00	0.22
Yellow Cards	0.039	-0.00014	0.00071	0.00035	0.00134	0.00	0.92
Through Ball	0.039	-0.00190	-0.00066	0.00041	0.00034	0.00	0.13
Fouls Conceded (danger area)	0.036	-0.00007	0.00096	0.00041	0.00186	0.00	0.94
Total Passes	0.036	-0.00004	0.00000	0.00000	0.00003	0.08	0.44
Tackled Possession Retained (%)	0.035	-0.35535	0.04169	0.18434	0.41473	0.00	0.57
Total Fouls Conceded	0.035	0.00008	0.00033	0.00010	0.00051	0.00	0.97
Tackles Attempted	0.035	-0.00013	0.00005	0.00007	0.00021	0.00	0.69
Dribbles and Runs Attempted	0.035	0.00006	0.00014	0.00004	0.00023	0.01	1.00

Notes: The sport performance dependent variable is the transformed of the ranking in the league such that $y_i = \log(X + 1 - C_i)/C_i)$ where X is the number of teams in the league and C denotes their classification. The results reported here correspond to the estimation of the top 5.000 models from the 16.77 million possible regressions including any combination of the 24 regressors. Prior mean model size is 12. Variables are ranked by Column (1), the posterior inclusion probability. Columns (2) to (5) reflect the lower 5% bound, the posterior mean, standard deviations and upper 95% bound for the linear marginal effect of the variable conditional on inclusion in the model, respectively. Column (6) is the fraction of regressions in which the coefficient has a classical t-test greater than 1.96, with all regressions having equal sampling probability. The last column denotes the sign certainty probability, a measure of our posterior confidence in the sign of the coefficient.

Table 6. Robustness Check (II): Posterior Inclusion Probabilities by League

Variable	*Big Five*	*Premier*	*La Liga*	*Serie A*	*Bundesliga*	*Ligue 1*
Total Shots Attempted	0.103	0.3744	0.0460	0.0960	0.1260	0.0465
Shots on Target	0.229	0.9903	0.0437	0.1073	0.0880	0.0336
Total Passes	0.110	0.0362	0.1418	0.0385	0.0521	0.0370
Passing Accuracy	0.998	0.0364	0.0569	0.0513	0.0530	0.0624
Assists	1.000	0.9986	1.0000	1.0000	0.9997	1.0000
Crosses Attempted	0.154	0.0423	0.1463	0.0418	0.0342	0.0426
Corners Taken	0.038	0.2561	0.0392	0.0347	0.1361	0.0344
Dribbles and Runs Attempted	0.042	0.0534	0.0489	0.0570	0.0380	0.1069
Dribble and Run Success Rate	0.042	0.0353	0.2908	0.0567	0.0962	0.0813
Long Pass Final Third	0.033	0.0426	0.0453	0.0740	0.0564	0.0476
Through Ball	0.038	0.1115	0.0401	0.6101	0.0329	0.0400
Offsides	0.031	0.0469	0.0327	0.2361	0.0328	0.0387
Shots Conceded	1.000	1.0000	1.0000	1.0000	1.0000	1.0000
Tackles Attempted	0.087	0.0341	0.0483	0.3646	0.0380	0.0563
Tackled and Possession Retained %	0.032	0.0437	0.0323	0.0499	0.0342	0.0320
Recoveries	0.171	0.1979	0.1404	0.2573	0.0340	0.0453
Recoveries in Opp. Half	0.183	0.0779	0.6225	0.0674	0.0346	0.0386
Clearances, block and intercept.	0.581	0.0369	0.9315	0.1989	0.2826	0.0396
Total Fouls Conceded	0.210	0.0475	0.6616	0.0478	0.0559	0.0445
Fouls Conceded (Danger Area)	0.100	0.0341	0.1334	0.0521	0.0354	0.5401
Yellow Cards	0.040	0.0345	0.0647	0.0279	0.0435	0.1119
Red Cards	0.056	0.0982	0.2114	0.1459	0.0742	0.2487
Saves Made	1.000	1.0000	0.9953	0.9991	0.9998	1.0000
Catches	0.033	0.0453	0.0550	0.0378	0.0782	0.0346

Notes: The dependent variable in all regressions is the normalized indicator of sport performance based on the points obtained during the season. All the results reported here correspond to the estimation of the top 5,000 models.

Table 7. Robustness Check (III): Relative Importance Decomposition by League

Variable	*Big Five*	*Premier*	*La Liga*	*Serie A*	*Bundesliga*	*Ligue 1*
A. Attack						
Total Shots Attempted	0.062	0.078	0.066	0.066	0.066	0.057
Shots on Target	0.098	0.156	0.085	0.085	0.092	0.067
Total Passes	0.071	0.064	0.058	0.058	0.099	0.076
Passing Accuracy	0.079	0.067	0.068	0.068	0.083	0.105
Assists	0.179	0.193	0.202	0.202	0.158	0.131
Crosses Attempted	0.005	0.002	0.004	0.004	0.004	0.012
Corners Taken	0.033	0.052	0.032	0.032	0.032	0.023
Dribbles and Runs Attempted	0.009	0.030	0.015	0.015	0.011	0.013
Dribble and Run Success Rate	0.007	0.005	0.010	0.010	0.013	0.005
Long Pass Final Third	0.015	0.031	0.005	0.005	0.005	0.025
Through Ball	0.051	0.040	0.055	0.055	0.052	0.047
Offsides	0.014	0.003	0.054	0.054	0.006	0.013
B. Defense						
Shots Conceded	0.215	0.135	0.148	0.148	0.157	0.189
Tackles Attempted	0.003	0.008	0.011	0.011	0.024	0.004
Tackled Possession Retained %	0.003	0.005	0.007	0.007	0.003	0.008
Recoveries	0.015	0.008	0.022	0.022	0.006	0.012
Recoveries in Opp. Half	0.011	0.027	0.020	0.020	0.008	0.012
Clearances, Blocks and Intercept	0.017	0.020	0.013	0.013	0.016	0.022
Total Fouls Conceded	0.007	0.007	0.016	0.016	0.022	0.009
Fouls Conceded (Danger Area)	0.033	0.023	0.028	0.028	0.041	0.104
Yellow Cards	0.005	0.002	0.027	0.027	0.036	0.003
Red Cards	0.005	0.003	0.006	0.006	0.016	0.003
Saves Made	0.058	0.040	0.046	0.046	0.024	0.049
Catches	0.003	0.005	0.004	0.004	0.024	0.009

Notes: The dependent variable in all regressions is the normalized indicator of sport performance based on the points obtained during the season. The decomposition applies to a model with $R^2 = 0.88$ while the unexplained variability is $\sigma_Y^2(1-\Omega) = 0.12$.