Quality Technology & Quantitative Management

ISSN: (Print) 1684-3703 (Online) Journal homepage: http://www.tandfonline.com/loi/ttqm20

Discovering the Drivers of Football Match Outcomes with Data Mining

Maurizio Carpita, Marco Sandri, Anna Simonetto & Paola Zuccolotto

Quality Technology &
Quantitative Management
Vol. 12, No. 4, pp. 561-577, 2015

QTQM
© ICAQM 2015

Discovering the Drivers of Football Match Outcomes with Data Mining

Maurizio Carpita[*], Marco Sandri, Anna Simonetto and Paola Zuccolotto

Department of Quantitative Methods, University of Brescia, Brescia, Italy

(*Received* January 2014*, accepted* May 2014)

Abstract: In this paper the relationship between the outcome of a football match (win, lose or draw) and a set of variables describing the game actions is investigated across time, by analyzing data from 4 consecutive yearly championships. The aim of the study is to discover the factors leading to win the match. More precisely, the goal is to select, from hundreds of covariates, those that most strongly affect the probability of winning a match, to recognize regularities across time by identifying the variables whose importance is confirmed in different analyses, and finally to construct a small number of composite indicators to be interpreted as drivers of match outcome. These tasks are carried out using the Random Forest machine learning algorithm, in order to select the most important variables, and Principal Component Analysis, in order to summarize them into a small number of drivers. Variable selection is performed using the novel approach developed by Sandri and Zuccolotto [33-34].

Keywords: Classification, football data, machine learning algorithms, variable selection.

1. Introduction

In recent years a growing interest has been devoted to the use of statistical methods with the aim of identifying factors determining sporting events outcomes [17]. The use of statistical thinking in sports is documented since the early papers from the 1970s, followed by many scientific articles regularly published on the statistical analysis of sports data. In 2005 and 2008 two insightful collections of statistical analyses applied to a wide range of sports, including American football, baseball, basketball, and ice hockey [2-3]. In the literature, several data analysis techniques have been applied to football [11, 30] (known as soccer in the United States) with the aim of studying player performance [22, 36], predicting match outcomes [10, 27, 32], or identifying optimal game strategies [37].

To forecasting purpose, a fairly common approach to model the distribution of the match results is based on the double Poisson distribution. Starting from the work of Maher [25], Karlis and Ntzoufras [20] introduced the idea of using a dependence parameter to improve prediction of the number of draws. Koopman and Lit [21] highlighted the importance of transforming these models in dynamic models, as the team's attack and defense strengths vary over time. Their analysis is based on nine seasons of the English Premier League.

Some studies have focused on arbitrage. In 2010, Buraimo *et al.* [8] studied the effect of arbitration decisions on the outcome of matches, trying to determine the possible bias introduced by the assignment of sanctions (yellow or red cards) in the English Premier League and the German Bundesliga. Other researchers addressed the same issue, focusing

[*] Corresponding author. E-mail: maurizio.carpita@unibs.it

on different leagues. For the Italian league we can cite the works of Lucey and Power [24] and Scoppa [35]; for the Primera División in Spain the contribution of Garicano *et al.* [13]; for the German Bundesliga the studies of Dohmen [12] and Sutter and Kocher [39]; for the English Premier League Rickman and Witt [31].

Another line of research addressed the analysis on individual players, focusing on the goal scoring ability. McHale and Szczepański [26] proposed a two mixed effects models, separating the scoring process into the generation of shots and the conversion of shots to goals. They validated their model on data from two seasons of the English Premier League.

Finally, it is interesting to mention the contribution of van Dijkhuizen, the author of the Soccernomics 2012 Report [40], who tried to study the relationship between sporting results and economic performance in a given region. The aim of his work is to highlight the link between the "perception of the crisis, the economic maneuvers carried out by the member states of the European Union and the attitude of the football teams".

From a methodological point of view, the huge amount of raw data available on the actions and strategies – combined with the absence of a sound theory explaining the relationships between the many variables affecting match outcome – make football an interesting subject for exploratory data analyses using data mining [14, 15].

In this paper we follow and extend the analysis presented in Carpita *et al.* [10], where the authors applied a data mining approach to examine a dataset containing information about the matches played in the Italian Football League "Serie A" (2010-2011 season), with the aim of identifying the key factors that cause the outcome of the football match (win, lose or draw). Here we focus on the first part of the analysis, concerned with the following tasks: (i) the selection of informative variables using the Gini variable importance measure computed by Random Forests [6] and corrected with the novel approach proposed by Sandri and Zuccolotto [33-34]; (ii) the use of the selected variables to construct composite indicators to be interpreted as drivers of match outcome by Principal Component Analysis. Finally, we check the extent to which the obtained drivers influence the match outcome, using a Multinomial Logistic Regression model, which had revealed effective in the former study.

The specific aim of this paper is to extend the analysis to data of the 4 seasons, from 2008-2009 to 2011-2012, checking for regularities across different championships, discovering which variables are confirmed to be important, and building more stable drivers of the match outcome. Our interest is to investigate which factors are most crucial for the match result, so we focus on the overall game of each team (and not on individual players) and we do not consider the numbers of goals. The strengths of this study are the data mining approach, which allows to integrate different models and analytical techniques, and the use of measures of variable importance to select the most influential factors on the final result.

The paper is organized as follows: in Section 2 we briefly describe the main features of the analyzed data; the variable selection with Random Forest, the construction of the composite indices with PCA, and the inspection of their ability as drivers of the match outcome are treated in Sections 3, 4 and 5, respectively. Concluding remark can be found in Section 6.

2. The Dataset

The top Italian professional football league "Serie A" has 20 teams, selected based on their performance during the previous season (the best 17 from "Serie A" plus the top 3 teams from the next highest ranking Italian league, “Serie B”). A true round-robin format is used: each of the 20 teams plays twice (as home and away team) against each of the other 19 teams, for a total of 38 matches. The season is divided into two parts: the “going round” (“andata”) and the “return round” (“ritorno”). Teams are awarded three points for a win, one point for a draw, and no points for a loss. The top four teams in the final ranking qualify for the UEFA Champions League and the fifth and sixth teams qualify for the UEFA Europa League Tournament. The season lasts from August to June.

In this paper we analyze data of the 4 seasons from 2008-2009 to 2011-2012, available from the Panini Digital football database. Panini Digital is a leader in the collection of statistical data on football, providing data services to football clubs and the media (www.paninidigital.com). The football database contains detailed information about plays made during each match (e.g. free kicks and shots, action type, fouls, crosses, recovered balls, goal assists, average time of ball possession, saves, goals on free kicks, etc.). The data collection technique is based on the proprietary software *DigitalScout*: a trained operator observes the match and records each action in real time, entering data using a point-and-click device and voice recorder. For each match, approximately 1,300 events are processed, summarized, and recorded in a dataset with 482 variables, that we use in the statistical analysis described below. The statistical unit of interest for our analysis is the single match.

We analyze 4 databases, one for each season, composed of $n = 380$ observations (10 matches for each of the 38 rounds) and 482 variables: one categorical target variable Y describing the outcome of the match (three categories: win, lose, or draw; henceforth *W*, *L*, *D*) and $p = 481$ covariates $\boldsymbol{X}$, that describe the actions during the match.

During the observation period, 14 teams participated in all four championships, while there are 12 teams that moved from or to the lower division.

Looking at the list of players, on average each team is made up of 28.7 players (sd 2.8). We can observe that in the 2008-2009 season, the percentage of Italians players is equal to 61%. This percentage is considerably reduced over time by switching to 55.4% in 2009-2010 to 49.8% in 2010-2011, and it has stabilized at 49.5% in 2011-2012. This seems to be due primarily to the fact that teams of average/low ranking made purchases outside of Italy, with the intention of buying players at low cost, make them "explode" in our league and then resell them to exorbitant prices to the big European football. It is also interesting to note the trend of the mean age of the players, which increased from 27.2 in 2008-2009 to 28 years in 2011-2012. It is a rather high value compared to other leagues. A plausible explanation is that the Italian league has always been considered among the most difficult at the European level, so the teams tend to acquire successful and experienced players, who have already played in a league similar to ours.

3. Variable Selection

The first step of the analysis consists of selecting, among the p covariates, those variables having the strongest influence on the outcome variable Y. Following Carpita *et al.* [10], we propose to carry out variable selection using the Random Forest (RF) algorithm [6], because it is well suited to treat the case: $n < p$. This methodology allows us to approximate the relationships between the p covariates in $\boldsymbol{X}$ and the target variable Y, even those that are very complex, thus identifying which variables play a major role in the mechanism generating the outcome. At the same time, we can identify any redundant or non-influential covariates [23].

The RF with randomly selected inputs consists of a sequence of classification or regression trees (CART, [4]) grown by selecting randomly from $\boldsymbol{X}$, at each node, a small group of h covariates on which to split the node. As is customary when using the CART methodology, the splitting criterion is the maximum heterogeneity reduction in the target variable Y, which has three categories in this case study: $(Y_1, Y_2, Y_3) = (W, L, D)$. This procedure is used together with *Bagging* [5], which is the random selection of a subsample from the original training set at each tree. This simple and effective idea is based on a theoretical framework described by Breiman [6] in his seminal work. The RF prediction is an average of the tree predictions for Y, computed by passing down each tree only the observations that did not contribute to its construction (*out-of-bag* predictions).

The RF algorithm returns two main variable importance measures (VIMs) that are used to identify the most important predictors [7] in $\boldsymbol{X}$. The first VIM is the *Mean Decrease in Accuracy* (MDA). For each tree in the RF, the out-of-bag prediction accuracy is recorded and compared to the accuracy of the same tree when the values for the j-th covariate are randomly permuted. The MDA VIM of the j-th covariate is then obtained by averaging over all trees the decrease in accuracy due to this permutation. The second VIM is the *Total Decrease in Node Impurities* (TDNI). At each split of the RF in each tree, the heterogeneity reduction in the target variable Y is defined as the importance measure attributed to the splitting variable in that particular split. The TDNI VIM is then obtained by summing up these measures, separately for each covariate, for all of the trees in the RF.

In general, the heterogeneity reduction in the target variable Y due to the split of node *w* into the two daughter nodes *wl* and *wr*, is measured as

$$d_w = \frac{n_w}{n} \left\{ \hat{H}_{Y|w} - \left(\frac{n_{wl}}{n_w} \hat{H}_{Y|wl} + \frac{n_{wr}}{n_w} \hat{H}_{Y|wr} \right) \right\},$$

where $\hat{H}_{Y|w}$, $\hat{H}_{Y|wl}$, and $\hat{H}_{Y|wr}$ are the estimated heterogeneities, of *Y* in node *w* and in the left and right splits, respectively. Similarly, n_w, n_{wl}, and n_{wr} are the sample sizes in node *w* and in the left and right splits.

The heterogeneity measure $\hat{H}$ we adopt in our analysis is the Gini index

$$\hat{H}_{Y|w} = \hat{G}_{Y|w} = 1 - \sum_{k=1}^{3} f_{kw}^2,$$

where f_{kw} is the observed relative frequency of k-th category in node w. The corresponding TDNI measure is the Gini VIM. In the variable selection analysis below, we rely on the Gini VIM, as recommended by Calle and Urrea [9]. These authors showed that variable ranking according to the Gini VIM is generally more stable than the ranking obtained by the MDA VIM. Nonetheless, VIMs should be used cautiously for variable selection because they may lead to spurious results under some circumstances [28]. In addition, as observed by Strobl *et al.* [38] and Sandri and Zuccolotto [33], the Gini VIM is biased in favor of variables that have either more distinct numerical values (or categories) or fewer missing values. In this case study, we deal with bias in the Gini VIM using the heuristic correction strategy proposed by Sandri and Zuccolotto [33-34]. The idea behind this correction algorithm is to estimate bias by means of p "pseudo-covariates" $\boldsymbol{Z}$, independent on Y (thus, uninformative) and having the same marginal distributions of $\boldsymbol{X}$. A set of S observations for $\boldsymbol{Z}$ is generated from $\boldsymbol{X}$ by a heuristic permutation procedure, and the estimated VIMs of the pseudo-covariates are then shown to approximate the unknown bias [33-34].

The main steps of this algorithm are:

1) Generate a matrix $\boldsymbol{Z}$ of pseudo-covariates by randomly permuting the rows of $\boldsymbol{X}$.

2) Build an RF for the prediction of Y using the covariates in $\boldsymbol{X}$ and the pseudo-covariates in $\boldsymbol{Z}$ as explanatory variables.

3) Compute the Gini VIMs of covariates and pseudo-covariates.

4) For each $j = 1, 2, \ldots, p$, calculate the difference between the estimated VIM of variable X_j and the corresponding pseudo-covariate Z_j.

5) Repeat steps 1 to 4 a total of S times (in this case study we set $S = 100$).

6) Calculate the mean values of these differences over the S replications.

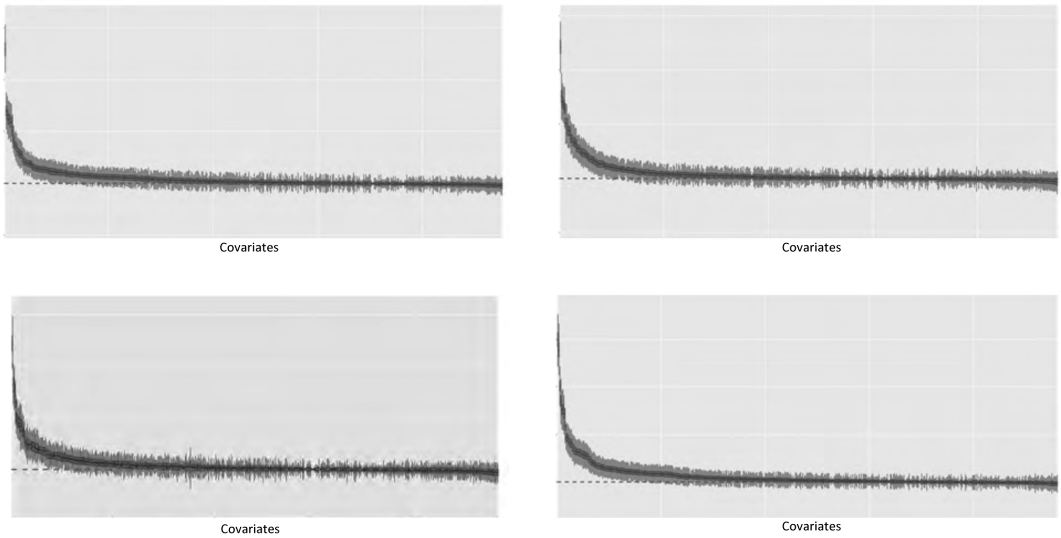

Figure 1. Box-plots of bias-corrected Gini VIMs for 4 datasets (*top left*: season 2008-2009; *top right*: season 2009-2010; *bottom left*: season 2010-2011; *bottom right*: season 2011-2012).

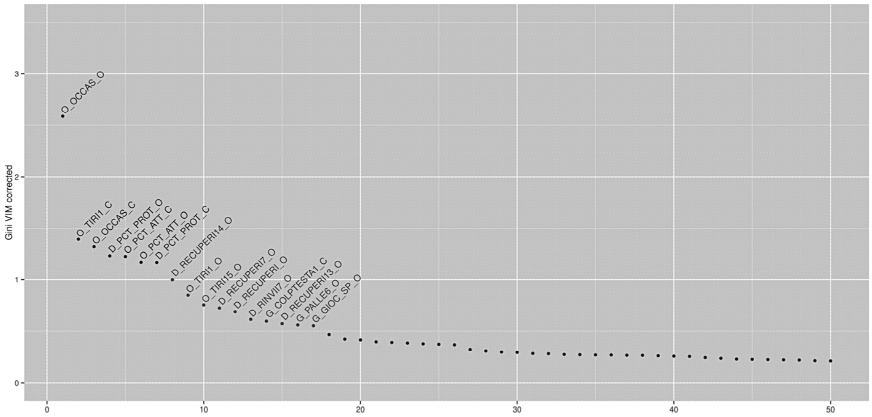

Figure 2. The 50 most important covariates according to the mean bias-corrected Gini VIM (season 2008-2009).

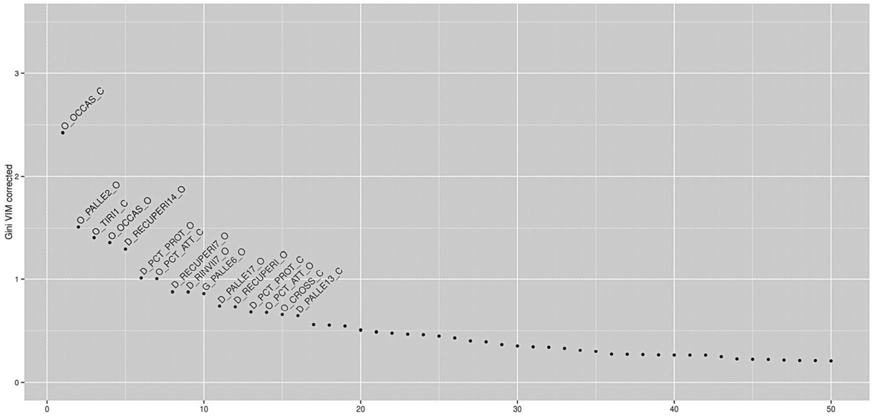

Figure 3. The 50 most important covariates according to the mean bias-corrected Gini VIM (season 2009-2010).

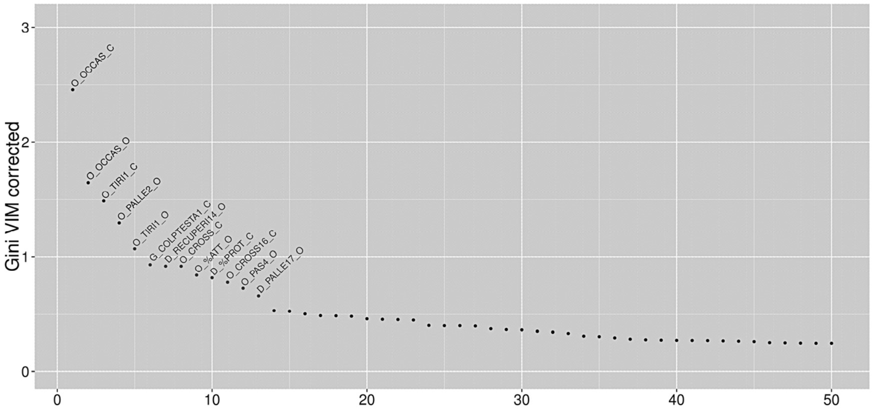

Figure 4. The 50 most important covariates according to the mean bias-corrected Gini VIM (season 2010-2011)

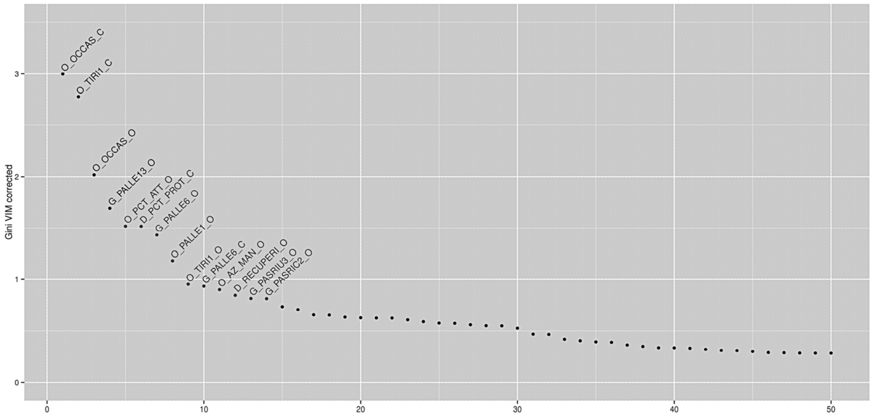

Figure 5. The 50 most important covariates according to the mean bias-corrected Gini VIM (season 2011-2012).

R codes for implementing this bias-correction procedure can be found in the supplementary material of the paper by Sandri and Zuccolotto [33], while in Carpita *et al.* [10] the codes are implemented with parallel computing features, which appreciably fastens computations, and their usage is exhaustively described within the text. For each analysed season, the box-plots of the bias-corrected Gini VIMs for the 481 covariates obtained with $S = 100$ replications are displayed in Figure 1. The 4 graphs show that, for each season, only a small part of the 481 covariates is relevant for the prediction of the match outcome, the remaining covariates are non-influential and potentially noisy.

Figures 2-5 show the mean unbiased Gini VIMs of the 50 most important covariates, for the 4 seasons. Each graph highlights the presence of three groups of covariates: some "highly" informative covariates (blue dots), some covariates with "medium" importance (red dots), and a residual group (green dots) of covariates having low VIMs, that are approximately constant. We observe that there is a fair regularity from one season to another, at least for the top ranked covariates. This finding is very important, as it confirms that the proposed procedure is quite stable in detecting the covariates which typically affect the match outcome, regardless of the peculiarities of the single championships. Figure 6 shows the Venn chart summarizing the variable selection in the 4 datasets, with superimposed pie charts indicating the frequency distribution of the position of the corresponding covariates among the blue dots (high importance) or the red dots (medium importance) in the 4 cases (see Figures 2-5).

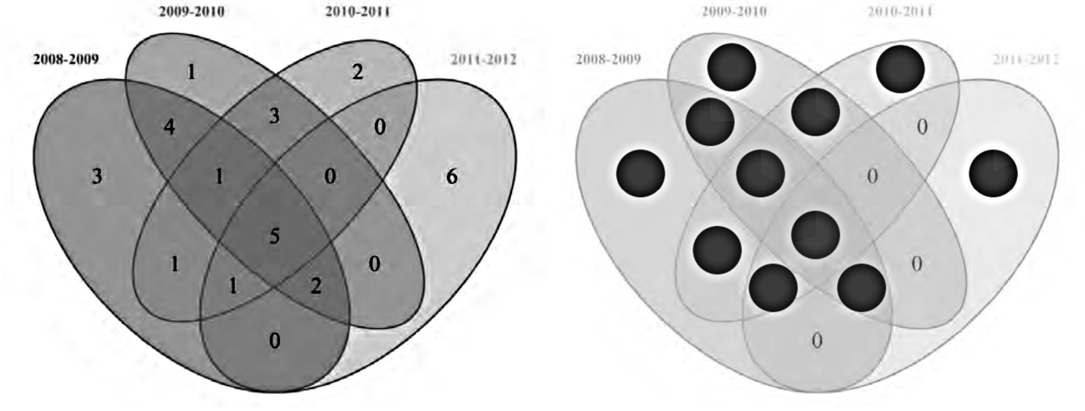

Figure 6. Venn diagram of the covariates selected in the 4 cases.

We observe that 5 covariates are selected in each of the 4 analyses, while 4 and 8 covariates are selected in 3 and 2 analyses, respectively. The pie charts in the right panel show that the covariates repetitively selected with different dataset, also tend to be the same that belong to the "highly informative" groups in the single analyses (blue dots in Figures 2-5).

On the whole, we decide to consider informative the $q = 17$ covariates selected at least two times (Table 1). It is worth noting that 10 of the 13 covariates identified as important in the work of Carpita *et al.* [10], are confirmed informative in this new analysis dealing with an appreciably larger amount of data.

Table 1. Covariates selected at least two times.

2008-2009 2009-2010 2010-2011 2011-2012	2008-2009 2009-2010 2010-2011	2008-2009 2009-2010 2011-2012	2008-2009 2010-2011 2011-2012	2008-2009 2009-2010	2008-2009 2010-2011	2009-2010 2010-2011
O_OCCAS_O O_TIRI1_C O_OCCAS_C O_PCT_ATT_O D_PCT_PROT_C	D_RECUPERI14_O	D_RECUPERI_O G_PALLE6_O	O_TIRI1_O	D_PCT_PROT_O O_PCT_ATT_C D_RECUPERI7_O D_RINVII7_O	G_COLPTESTA1_C	O_PALLE2_O D_PALLE17_O O_CROSS_C
O_OCCAS_O: NUMBER OF SCORING OPPORTUNITIES CREATED BY AWAY TEAM						
O_TIRI1_C: NUMBER OF SHOTS ON GOAL FOR HOME TEAM						
O_OCCAS_C: NUMBER OF SCORING OPPORTUNITIES CREATED BY HOME TEAM						
O_PCT_ATT_O: GOAL ATTACK PERCENTAGE BY AWAY TEAM						
D_PCT_PROT_C: PERCENTAGE OF PENALTY AREA'S DEFENSE BY HOME TEAM						
D_RECUPERI14_O: NUMBER OF DEFENSIVE HEADINGS IN THE PENALTY AREA BY AWAY TEAM						
D_RECUPERI_O: NUMBER OF DEFENSIVE ACTIONS BY AWAY TEAM						
G_PALLE6_O: NUMBER OF BALL KICKS IN THE MIDFIELD BY AWAY TEAM						
O_TIRI1_O: NUMBER OF SHOTS ON GOAL FOR AWAY TEAM						
D_PCT_PROT_O: PERCENTAGE OF PENALTY AREA'S DEFENSE BY AWAY TEAM						
O_PCT_ATT_C: GOAL ATTACK PERCENTAGE BY HOME TEAM						
D_RECUPERI7_O: NUMBER OF DEFENSIVE ACTIONS IN THE PENALTY AREA BY AWAY TEAM						
D_RINVII7_O: NUMBER OF LONG BALL KICKS BY AWAY TEAM						
G_COLPTESTA1_C: NUMBER OF HEADING SHOTS BY THE HOME TEAM IN THE AWAY TEAM'S PENALTY AREA						
O_PALLE2_O: NUMBER OF BALL KICKS TO BYPASS THE MIDFIELD BY AWAY TEAM						
D_PALLE17_O: NUMBER OF BALLS LOST DURING FORWARD ACTIONS BY HOME TEAM						
O_CROSS_C: NUMBER OF CROSSES BY HOME TEAM						

4. Building Drivers of the Match Outcome

After selecting the $q=17$ variables that most influence match results, in the second step of the analysis we combine those variables into a smaller number of easily interpretable key indicators, using Principal Component Analysis (PCA). The main function of PCA is to reduce the dimensionality of a data set of q correlated variables while retaining as much of the variation within the data set as possible. The earliest descriptions of this technique were published by Pearson [29] and Hotelling [18].

Briefly, given the $(n \times q)$ data matrix $\boldsymbol{X}_s$, containing the observations of the q numerical variables X, selected in the first step, PCA results in a linear transformation to a new coordinate system such that the new $(n \times q)$ data matrix $\boldsymbol{W}$ contains the q uncorrelated the Principal Components (PCs) The orthogonal projections of data onto the one-dimensional spaces spanned by $\boldsymbol{W}$ have, respectively, the greatest variance, the second greatest variance, and so on. This reduction in complexity is achieved by selecting components containing a significant portion of the total observed variance.

PCA is based on the singular value decomposition of the data matrix $\boldsymbol{X}_s = \boldsymbol{UDV}'$ [16], where $\boldsymbol{U}$ is an $n \times q$ orthogonal matrix of left singular vectors, $\boldsymbol{V}$ is a $q \times q$ orthogonal matrix whose column vectors $\mathbf{v}_j$ are right singular vectors, and $\boldsymbol{D}$ is a $q \times q$ diagonal matrix of the singular values d_j of $\boldsymbol{X}_s$. The number of PCs is equal to q. The PCs are in the form $\boldsymbol{W} = \boldsymbol{X}_s\boldsymbol{V}$, so that $\boldsymbol{w}_1 = \boldsymbol{X}_s\boldsymbol{v}_1$ has the largest sample variance, $Var(\boldsymbol{w}_1) = Var(\boldsymbol{X}_s\boldsymbol{v}_1) = d_1 / q$, among all of the normalized linear combinations of the original variables. In order to reduce the dimensionality of our data, we keep the first b PCs; the amount of variance they account for is referred to as the Cumulative Variance Accounted For (CVAF):

$$\text{CVAF} = \sum_{j=1}^{b} Var\left(\boldsymbol{w}_j\right) = \sum_{j=1}^{b} \frac{d_j}{q}.$$

Further details about this technique and how results should be interpreted can be found in Jolliffe [19]. To facilitate interpretation of the results, we choose to conduct two separate PCAs, one for the 6 variables directly related to the home team (O_TIRI1_C, O_OCCAS_C, D_PCT_PROT_C, O_PCT_ATT_C, G_COLPTESTA1_C, O_CROSS_C) and one for the 11 variables directly related to the away team (O_OCCAS_O, O_PCT_ATT_O, D_RECUPERI14_O, D_RECUPERI_O, G_PALLE6_O, O_TIRI1_O, D_PCT_PROT_O, D_RECUPERI7_O, D_RINVII7_O, O_PALLE2_O, D_PALLE17_O). Firstly, we carry out separate analyses in the 4 datasets.

3.1. PCA for the Home Team

The results of the PCAs carried out on the data from the 4 championships are displayed in Tables 2 and 3.

Table 2. Variances of the extracted PCs and CVAF (4 seasons separately).

	2008-2009		**2009-2010**		**2010-2011**		**2011-2012**	
PC	***d*** d_j	**CVAF(%)**	d_j	**CVAF(%)**	d_j	**CVAF(%)**	d_j	**CVAF(%)**
1	2.10	35.0	2.01	33.6	1.87	31.1	1.95	32.5
2	1.46	59.4	1.44	57.5	1.54	56.9	1.47	57.0
3	0.99	75.8	1.06	75.1	0.98	73.2	1.04	74.3
4	0.65	86.6	0.70	86.8	0.75	85.7	0.74	86.7
5	0.49	98.9	0.46	94.5	0.52	94.3	0.51	95.2
6	0.31	100.0	0.33	100.0	0.34	100.0	0.29	100.0

Table 3. Varimax rotated loadings of the first 3 PCs (4 seasons separately).

	2008-2009			**2009-2010**			**2010-2011**			**2011-2012**		
VARIABLES	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
O_TIRI1_C	**0.608**	0.054	0.081	**0.630**	0.050	0.011	**0.637**	-0.014	0.009	**0.625**	0.032	0.023
O_OCCAS_C	**0.620**	0.038	0.064	**0.650**	-0.029	-0.001	**0.651**	0.041	0.011	**0.647**	0.023	-0.001
D_PCT_PROT_C	0.023	-0.028	**0.959**	0.223	-0.035	**0.843**	0.229	-0.043	**0.829**	0.216	-0.081	**0.842**
O_PCT_ATT_C	**0.495**	-0.122	**-0.225**	**0.361**	-0.047	**-0.538**	**0.343**	-0.06	**-0.554**	**0.375**	-0.117	**-0.532**
G_COLPTESTA1_C	0.004	**0.706**	-0.100	0.009	**0.701**	-0.010	0.004	**0.680**	0.057	-0.016	**0.702**	-0.053
O_CROSS_C	0.003	**0.694**	0.091	0.009	**0.708**	0.014	0.013	**0.728**	-0.051	0.054	**0.696**	0.061

We observe that the results are pretty stable across the different championships, both from the point of view of the CVAF and the loadings of the first 3 PCs. So, we are allowed to collect all the data together and carry out a single PCA for all the 4 championships (Tables 4 and 5).

Table 4. Variances of the extracted PCs and CVAF (all data together).

	ALL SEASONS	
PC	d_j	**CVAF(%)**
1	1.97	32.8
2	1.49	57.5
3	1.03	74.8
4	0.71	86.5
5	0.51	95.0
6	0.30	100.0

Table 5. Varimax rotated loadings of the first 3 PCs (all data together).

	ALL SEASONS		
VARIABLES	**PC1** **SHOT.ATTACK.HOME**	**PC2** **AERIAL.ATTACK.HOME**	**PC3** **DEFENSE.HOME**
O_TIRI1_C	**0.631**	0.030	0.022
O_OCCAS_C	**0.646**	0.008	0.006
D_PCT_PROT_C	0.196	-0.055	**0.869**
O_PCT_ATT_C	**0.381**	-0.084	**-0.494**
G_COLPTESTA1_C	0.000	**0.703**	-0.016
O_CROSS_C	0.026	**0.703**	0.023

Recalling that the commonly used rule for retaining a PC is that its variance be greater than 1, we choose to retain $b = 3$ PCs, with the a CVAF of 74.8%.

For interpretation of the rotated solution, we refer to the loadings of Table 5.

The first (rotated) PC represents the ability of the home team to create opportunities to make shots on goal, so we name it "*shot.attack.home*". The second PC represents aerial abilities (crosses and heading) when the home team is on the attack, so we name it "*aerial.attack.home*". The third PC represents the capability of the home team defense, so we name it "*defense.home*".

3.2. PCA for the Away Team

The results of the PCAs carried out on the data from the 4 championships are displayed in Tables 6 and 7.

Table 6. Variances of the extracted PCs and CVAF (4 seasons separately).

	2008-2009		2009-2010		2010-2011		2011-2012	
PC	d_j	**CVAF(%)**	d_j	**CVAF(%)**	d_j	**CVAF(%)**	d_j	**CVAF(%)**
1	4.94	44.9	5.06	46.0	5.02	45.6	5.08	46.2
2	1.84	61.6	1.78	62.2	1.84	62.3	1.94	63.8
3	0.97	70.5	1.11	72.2	1.12	72.5	1.03	73.2
4	0.74	77.2	0.76	79.1	0.72	79.0	0.70	79.6
5	0.73	83.8	0.62	84.7	0.64	84.8	0.57	84.8
6	0.57	89.0	0.48	89.1	0.49	89.3	0.50	89.3
7	0.43	92.9	0.39	92.6	0.37	92.7	0.41	93.0
8	0.31	95.7	0.35	95.7	0.33	95.7	0.30	95.8
9	0.27	98.2	0.28	98.3	0.26	98.1	0.27	98.2
10	0.15	99.5	0.14	99.6	0.16	99.6	0.15	99.6
11	0.05	100.0	0.03	100.0	0.05	100.0	0.05	100.0

Table 7. Varimax rotated loadings of the first 3 PCs (4 seasons, separately).

	2008-2009			2009-2010			2010-2011			2011-2012		
VARIABLES	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**	**PC1**	**PC2**	**PC3**
O_OCCAS_O	-0.001	**0.650**	0.040	-0.005	**0.654**	-0.018	-0.001	**0.651**	0.003	-0.026	**0.588**	0.144
O_PCT_ATT_O	0.051	**0.399**	**-0.449**	0.076	**0.332**	**-0.551**	0.042	**0.342**	**-0.509**	0.046	**0.529**	**-0.219**
D_RECUPERI14_O	**0.336**	0.089	**0.140**	**0.317**	0.105	**0.211**	**0.301**	0.102	**0.273**	**0.263**	0.071	**0.361**
D_RECUPERI_O	**0.430**	-0.029	0.036	**0.422**	-0.006	0.062	**0.427**	0.015	0.047	**0.423**	-0.012	0.079
G_PALLE6_O	**0.384**	0.012	-0.112	**0.396**	0.039	-0.162	**0.367**	0.050	-0.114	**0.359**	0.070	-0.002
O_TIRI1_O	-0.032	**0.619**	0.036	-0.039	**0.629**	0.047	-0.039	**0.636**	0.040	-0.040	**0.598**	0.055
D_PCT_PROT_O	-0.012	0.142	**0.822**	-0.036	0.219	**0.729**	-0.045	0.175	**0.788**	-0.127	-0.031	**0.885**
D_RECUPERI7_O	**0.449**	0.007	-0.152	**0.439**	0.012	-0.122	**0.431**	0.035	-0.109	**0.430**	0.017	-0.030
D_RINVII7_O	**0.276**	0.029	**0.211**	**0.282**	-0.013	**0.260**	**0.333**	-0.053	**0.119**	**0.340**	-0.009	0.072
O_PALLE2_O	**0.281**	-0.022	**0.134**	**0.310**	-0.063	**0.076**	**0.314**	-0.081	0.051	**0.336**	-0.064	-0.039
D_PALLE17_O	**0.441**	-0.059	-0.030	**0.434**	-0.041	-0.007	**0.437**	-0.021	-0.024	**0.435**	-0.025	0.006

Also in this case, results are rather stable from one season to another. Again, we collect all data together and carry out an overall analysis (Tables 8 and 9).

Table 8. Variances of the extracted PCs and CVAF (all data together).

	ALL SEASONS	
PC	**d_i**	**CVAF(%)**
1	5.03	45.7
2	1.90	63.0
3	1.02	72.2
4	0.71	78.7
5	0.62	84.3
6	0.49	88.8
7	0.44	92.8
8	0.31	95.6
9	0.28	98.1
10	0.15	99.6
11	0.05	100.0

Table 9. Varimax rotated of the first 3 PCs (all data together).

	ALL SEASONS		
VARIABLES	**PC1 MIDFIELD-DEFENSE. COUNTERATTACK.AWAY**	**PC2 SHOT.ATTACK. AWAY**	**PC3 AREA-DEFENSE. AWAY**
O_OCCAS_O	-0.008	**0.642**	0.039
O_PCT_ATT_O	0.044	**0.405**	**-0.462**
D_RECUPERI14_O	**0.307**	0.101	**0.255**
D_RECUPERI_O	**0.428**	-0.006	0.043
G_PALLE6_O	**0.374**	0.050	-0.100
O_TIRI1_O	-0.038	**0.620**	0.049
D_PCT_PROT_O	-0.034	0.144	**0.822**
D_RECUPERI7_O	**0.436**	0.020	-0.111
D_RINVII7_O	**0.314**	-0.016	0.122
O_PALLE2_O	**0.312**	-0.064	0.047
D_PALLE17_O	**0.439**	-0.035	-0.030

For the away team we use similar criteria as for the home team (see above) to determine the number of PCs to retain. As a result, also in this case we choose to retain $b = 3$ PCs, with a CVAF of 72.2%. For interpretation of the rotated solution, we refer to loadings shown in Table 9. The first (rotated) PC represents general defense abilities, long-range kicks and sudden counterattacks of the away team, with specific reference to actions in the midfield, so we name it "*midfield-defense.counterattack.away*". The second PC represents the ability of the away team to create opportunities to make shots on goal, so we name it "*shot.attack.away*". The third PC represents the attitude of the away team to condense defense in the crucial penalty area, also at the expense of the possibility to create effective counterattack actions, so we name it "*area-defense.away*".

In the following, we consider the 6 extracted PCs as possible drivers of the match outcome. It is worth noting that the 6 drivers built with this analysis are roughly the same that were obtained in the work presented by Carpita *et al.* [10], based only on data from the season 2010-2011.

In the next section we investigate the influence of the proposed drivers on the outcome variable Y by means of Multinomial Logistic Regression, a statistical classification model specifically suited to the case of polytomous dependent variable.

5. Inspecting the Influence of Drivers on the Match Outcome

In this section we use the obtained composite indices as predictors in a classification model for the variable Y, in order to inspect their actual ability as drivers of the match outcome.

In Carpita *et al.* [10], several different statistical models had been applied in this step, namely two machine learning algorithms (Random Forest and Neural Network), a nonparametric model (K-Nearest Neighbor), a probabilistic model (Naïve Bayes), and a model that belongs to the class of generalized linear models (Multinomial Logistic Regression). After evaluating the predictive performance of classifiers by means of a number of different indices (Accuracy, Cohen's Kappa, Sensitivity and Specificity for each of the three categories *W*, *L*, *D*) computed over 1,000 crossvalidated replications of each classification model, the Multinomial Logistic Regression (MLogit) classifier had been considered the model with the best balance between complexity and accuracy of predictions. As a matter of fact, despite the simplicity and the intrinsically linear nature of MLogit, this model behaves favorably compared to more complex methods, and this suggests that the relationships under investigation could be characterized by an essentially linear structure. Hence, in the present study we carry out the analysis by directly resorting to MLogit, without any further comparison to other methods.

The MLogit is a generalized linear model used to estimate the probabilities for the m categories of a qualitative dependent variable *Y*, using a set of explanatory variables $\boldsymbol{X}$:

$$\Pr\left(Y_{ik}\right)=\Pr\left(Y_i=k\middle|\boldsymbol{x}_i;\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_m\right)=\frac{\exp\left(\boldsymbol{\beta}_{0k}+x_i\boldsymbol{\beta}_k'\right)}{\sum_{j=1}^{m}\exp\left(\boldsymbol{\beta}_{0j}+x_i\boldsymbol{\beta}_j'\right)}\text{ with }k=1,2,\ldots,m,$$

where $\boldsymbol{\beta}_k$ is the row vector of regression coefficients of $\boldsymbol{X}$ for the *k*-th category of Y. For further details see Agresti [1].

Due to the substantial stability of results across the different championships observed in the previous sections, we compute MLogit estimates on the dataset obtained by gathering all the data from the 4 seasons.

Tables 10 and 11 report the confusion matrix and some main goodness-of-fit statistics, respectively, while the relative risks ratios are displayed in Table 12.

Table 10. Confusion matrix (MLogit model).

		REFERENCE		
		W	L	D
PREDICTION	W	612	64	181
	L	43	253	107
	D	75	68	117

Table 11. Multinomial logistic model: Accuracy, Cohen's Kappa, Sensitivity and Specificity for each of the three categories W, L, D.

ACCURACY COHEN'S KAPPA	0.64 (95% CI = 0.62 - 0.67) 0.43 (95% CI = 0.41 - 0.44)		
	W	L	D
SENSITIVITY SPECIFICITY	0.84 0.69	0.66 0.87	0.29 0.87

Accuracy and Cohen's Kappa confirm a good predictive ability of the model. However, looking at Sensitivity and Specificity, we notice a substantially different performance of the model for the three categories of Y, with the *D* (draw) result being particularly difficult to predict.

This result had been already observed in the former study. This could be due, in part, to the fact that a draw is an outcome intrinsically characterized by a major degree of uncertainty, and, in part, to the problem of class imbalance. About the second issue, we have fitted to data the same imbalanced MLogit model proposed in Carpita *et al.* [10], obtaining similar results: a slightly higher Sensitivity for *D* at the expenses of the Sensitivity of *W*, which is not to be considered a profitable result as the prediction of *W* is largely more crucial in this context.

Table 12. Multinomial logistic model: relative risks ratios (with 95% confidence intervals).

VARIABLES	L	D
SHOT.ATTACK.HOME	0.16 (0.12 – 0.21)	0.30 (0.24 – 0.38)
AIR.ATTACK.HOME	1.92 (1.50 – 2.46)	1.71 (1.39 – 2.11)
DEFENSE.HOME	1.58 (0.95 – 2.62)	1.44 (0.94 – 2.20)
MIDFIELD-DEFENSE.COUNTATTACK.AWAY	2.20 (1.70 – 2.83)	1.56 (1.26 – 1.93)
SHOT.ATTACK.AWAY	4.78 (3.69 – 6.20)	1.86 (1.49 – 2.32)
AREA-DEFENSE.AWAY	0.51 (0.31 – 0.85)	0.63 (0.41 – 0.96)

The relative risks ratios give insights about the impact of the different factors on the probability that the home team wins the match.

On the one hand, from the point of view of the game strategies put into practice by the home team, we notice that increases in the factor "*shot.attack.home*" reduce the probability of both *L* and *D* with respect to *W*. Hence, creating opportunities and making shots on goal turns out to be the most important game strategy in terms of the probability of winning the match. The factor "*aerial.attack.home*" tends to slightly favor the probability of *L* and *D*, meaning that this game strategy tends to be very poorly effective or even counteractive. The same happens for the factor related to the defense ("*defense.home*"), meaning that the home team should maximally exploit the advantage of playing at home by putting in practice a game strategy as much aggressive as possible.

On the other hand, from the away team's perspective, the most effective game strategy is confirmed to be described by the driver "*shot.attack.away*", which is the exact counterpart of "*shot.attack.home*" for the home team. From this point of view there is no difference between playing at home or away. In fact, increments in "*shot.attack.away*" largely increase the probability of *L* (which means that the away team wins the match). For the away team, a more protective game strategy could be recommendable. In fact, the defense is divided into two different factors, "*midfield-defense.counterattack.away*" and "*area-defense.away*", the former being related to defense played in the midfield, which often can be converted into counterattacks, the latter describing defense condensed in the penalty area, which, beyond being a more dangerous game attitude, only occasionally can be turned into counterattacks, due to the rearward position of players. As expected, "*midfield-defense.counterattack.away*" and "*area-defense.away*" have, respectively, a positive and negative impact on the probability of *L* with respect to *W*, which means that the former style of defense is advantageous for the away team, while the latter is counteractive.

Again, all these results largely confirm the main part of the findings described in the former work by Carpita *et al.* [10].

6. Concluding Remarks

In this paper, we show the results of a data mining analysis developed to identify the key factors driving the result of a football match. Following the approach presented in a previous paper [10], we carried out the analysis on 4 large datasets, composed of 482 variables for each match of the Italian football championship league "Serie A" during the 4 seasons from 2008-2009 to 2011-2012. The two main tasks performed in this paper are: (i) the selection of informative variables, (ii) the construction of composite indicators to be interpreted as drivers of the match outcome. The main results can be briefly summarized as follows.

In the variable selection step, 17 covariates were selected out of the original set of 481, and considered important for determining the match outcome. This selection was made by choosing those covariates that have revealed stably relevant across the 4 analysed championships. The 17 selected variables were then summarized, using Principal Component Analysis, in six composite indicators: three related to the home team and three related to the away team. At this stage, as characteristic element, we successfully applied the Gini variable importance measure with heuristic correction strategy.

The extent to which one statistical indicator is particularly influential during a given match suggests the use of a specific game strategy by the corresponding team. In terms of football strategy, this data analysis highlights that the ability to create opportunities to make shots on goal increases the probability of scoring and, ultimately, winning the match, both for the home and the away team. On the other hand, we found that there is a difference between playing at home or away. More specifically, the home team should put into practice an aggressive game style, while the away team can take advantage of a more defensive strategy, provided that it is prudently played far from the penalty area. In addition, a strategy based on aerial abilities is positively associated with the probability of a draw or defeat; therefore, teams with a high degree of aerial play are less likely to win.

The results of our study may therefore be particularly interesting for coaches and team managers, who can use the ideas presented in the study when they are making decisions related to game strategies and skill acquisitions for the team. In addition, knowledge of well-controlled statistical analyses can help practitioners better understand what they need to do. In fact, the data analysis of the most influential factors for the final match outcome, considered in conjunction with the average value of each factor for the opposing team, can be useful to define a strategy for attack or defense during a direct match.

Finally, we can remark that results are fairly stable across the 4 analysed championships. This important finding confirms that, in spite of the obvious differences from one year to another (different teams and also different players), there are some fundamental regularities due to the main characteristics of the football game. Of course, this can be of maximum utility for decision-making in football.

Acknowledgements

The authors are grateful to dott. Aldo Maccagni of Panini Digital for his interest in this work and for making data available.

References

1. Agresti, A. (2003). Logit models for multinomial responses. In: *Categorical Data Analysis*, 2nd edition. John Wiley & Sons, Hoboken, NJ.

2. Albert, J., Bennet, J. and Cochran, J. J. (2005). *Anthology of Statistics in Sport, ASA-SIAM Series in Statistics and Probablity*, SIAM, Philadelphia, ASA, Alexandria.

3. Albert, J., Koning, R. H. (2008). *Statistical Thinking in Sports*. Chapman & Hall, Boca Raton.

4. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall, New York.

5. Breiman, L. (1996). Bagging predictors, *Machine Learning*, 24, 123-140.

6. Breiman, L. (2001). Random forests, *Machine Learning*, 45(1), 5-32.

7. Breiman, L. (2002). *Manual on Setting up, Using, and Understanding Random Forests v3.1.* Technical report. http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf.

8. Buraimo, B., Forrest, D. and Simmons, R. (2010). The 12th man?: refereeing bias in English and German soccer. *Journal of the Royal Statistical Society Series A*, Royal Statistical Society, 173(2), 431-449.

9. Calle, M. L. and Urrea, V. (2010). Letter to the editor: stability of random forest importance measures. *Briefings in Bioinformatics*, 12(1), 86-89.

10. Carpita, M., Sandri, M., Simonetto, A. and Zuccolotto, P. (2013). Football mining with R. in: *Data Mining Applications with R* (Edited by Y. Zhao, Y. Cen), Chapter 14. Elsevier.

11. Carroll, B., Palmer, P. and Thorn, J. (1988). *The Hidden Game of Football*. Warner Books, New York.

12. Dohmen, T. (2008). The influence of social forces: evidence from the behavior of football referees. *Economic Inquiry*, 46, 411–424.

13. Garicano, L., Palacios-Huerta, I. and Prendergast, C. (2005). Favoritism under social pressure. *Review of Economics and Statistics*, 87, 208–216.

14. Han, J., Kamber, M. and Pei, J. (2011). *Data Mining: Concepts and Techniques*, 3rd edition. The Morgan Kaufmann Publishers, San Francisco.

15. Hand, D. J., Mannila, H. and Smyth, P. (2001). *Principles of Data Mining: Adaptive Computation and Machine Learning.* MIT Press, Cambridge.

16. Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

17. Hopkins, W. G. (2012). The impact-factor Olympics for journals in sport and exercise science and medicine. *Sportscience*, 16, 17-19, http://sportsci.org/2012/wghif.htm.

18. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417-441, 498-520.

19. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Verlag, New York.

20. Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Statistician*, 52, 381–393.

21. Koopman, S. J. and Lit, R. (2014). A dynamic bivariate Poisson model for analysing and forecasting match results in the English premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, early view: DOI: 10.1111/rssa.12042.

22. Kuper, S. (2011). A football revolution. *Financial Times Magazine*, June 17, 2011, http://gilesrevell.com/files/championsleague.pdf.

23. Liaw, A. and Wiener, M. (2002). Classification and regression by random forest. *R News*, 2 (3), 18-22.

24. Lucey, B. and Power, D. (2004). *Do soccer referees display home team favouritism?* Mimeo. Trinity College Dubli, Dublin.

25. Maher, M. J. (1982) Modelling association football scores. *Statistics Netherlands*, 36, 109-118.

26. McHale, I. G. and Szczepański, Ł. (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.

27. Min, B., Kim, J., Choe, C., Eom, H. and McKay, R. I. (2008). A compound framework for sports results prediction: a football case study. *Knowledge-Based Systems*, 21, 551-562.

28. Nicodemus, K. K. (2011). Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics*, 12(4), 369-373.

29. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, Series 6, 2(11), 559-572.

30. Pollard, R. and Reep, C. (1997). Measuring the effectiveness of playing strategies at soccer. *The Statistician*, 46(4), 541-550.

31. Rickman, N. and Witt, R. (2008). Favouritism and financial incentives: a natural experiment. *Economica*, 75, 296–309.

32. Rue, H. and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *The Statistician*, 49(3), 399-418.

33. Sandri, M. and Zuccolotto, P. (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics,* 17(3), 611-628.

34. Sandri, M. and Zuccolotto, P. (2010). Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms. *Statistics and Computing*, 20, 393-407.

35. Scoppa, V. (2008). Are subjective evaluations biased by social factors or connections? An econometric analysis of soccer referee decisions. *Empirical Economics*, 35, 123-140.

36. Slaton, Z. (2012). *A Beautiful Numbers Game* - Statistically informed soccer writing. http://www.abeautifulnumbersgame.com.

37. Stern, H. (2005). Introduction to the football articles. In *Anthology of Statistics in Sport* (Edited by Albert, J., Bennet, J., Cochran, J. J.), ASA-SIAM Series in Statistics and Probability, SIAM, Philadelphia, ASA, Alexandria.

38. Strobl, C., Boulesteix, A. L., Zeileis, A. and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8-25.

39. Sutter, M. and Kocher, M. (2004). Favouritism of agents - the case of referees' home bias. *Journal of Economic Psychology*, 25, 461-469.

40. Van Dijkhuizen, A. (2012). *Soccernomics 2012 - Euro Football Poland/Ukraine*, http://www.abnamromarkets.be/fileadmin/user_upload/TA/2012/120529_-_Soccernomics_2012_ENG.pdf.

Authors' Biographies:

Maurizio Carpita is Full Professor of Statistics at the Department of Economics and Management, Scientific Director of the DMS StatLab (Data Methods and System Statistical Laboratory) at the University of Brescia - Italy and Scientific Coordinator of the Statistical Area at the EURICSE (European Research Institute of Cooperatives and Social Enterprises) - Italy. His main research interests relate on collection and organization of database, statistical methods, models, and algorithms for the measurement of perceptions, the applications of data analysis in economics and social sciences. His studies are about the measurement of subjective work quality and the socio-economic impact of cooperatives, nonprofit organizations and social enterprises.

Marco Sandri is Statistical Consultant and member of the DMS StatLab (Data Methods and Systems Statistical Laboratory) at the University of Brescia. He has authored/co-authored over 70 academic papers and articles in international journals and international conferences in the field of statistics and applications of statistics to life sciences. His main research areas are data mining, computational statistics and biostatistics.

Anna Simonetto is Research Fellow of Statistics at the Department of Economics and Management and member of the DMS StatLab (Data Methods and System Statistical Laboratory) at the University of Brescia - Italy. Her main research interests focus on multivariate data analysis, structural equation modeling, statistics for social science data analysis and big data analysis.

Paola Zuccolotto is Associate Professor of Statistics at the Department of Economics and Management and member of the Scientific Committee of the DMS StatLab (Data Methods and System Statistical Laboratory) at the University of Brescia - Italy. Her research topics cover data analysis, data mining and statistical modelling, with specific interest to prediction, feature selection, classification, dimensionality reduction, latent variables measurement, and applications in several different contexts (marketing, finance, social sciences, sensory analysis, sport, medicine, genetics).