Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v15n1p249

Exploring essential variables for successful and unsuccessful football teams in the "Big Five" with multivariate supervised techniques

By Malagón-Selma, Debón, Ferrer

Published: 20 May 2022

Electronic Journal of Applied Statistical Analysis
Vol. 15, Issue 01, May 2022, 249-276
DOI: 10.1285/i20705948v15n1p249

Exploring essential variables for successful and unsuccessful football teams in the "Big Five" with multivariate supervised techniques

P. Malagón-Selma*[a], A. Debón[a], and A. Ferrer[b]

[a] *Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia*
[b] *Grupo de Ingeniería Estadística Multivariante. Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia*

Published: 20 May 2022

This research proposes multivariate techniques for discovering the game actions that contribute to the final ranking of football teams. This study uses data from the "Big Five" teams that competed in the Bundesliga First Division, Premier League, LaLiga, Ligue 1, and Serie A in the 2018-2019 season. The principal component analysis is used for outlier detection and for providing an overall preliminary insight. The statistically significant game actions of the top and bottom teams were studied using three supervised multivariate techniques, namely the partial least squares discriminant analysis, random forest and logistic regression. The partial least squares discriminant analysis model best identifies the variables with the most statistically significant contribution to a team's success or failure. The results were compared with those obtained using two-sample univariate tests (such as the Student's t-test or the Mann–Whitney test), demonstrating the advantages of multivariate approaches over univariate approaches. The results indicate that the top teams have both offensive and defensive power, and emphasise the high number of attacking actions; in contrast, the bottom teams have weak defences and few offensive actions.

*Corresponding author: pimasel@doctor.upv.es

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

keywords: multivariate methods, two-sample tests, partial least squares discriminant analysis (PLS-DA), random forest (RF), logistic regression (RL), game actions.

1 Introduction

Sports, especially football, have been entrenched as essential elements in society, both culturally and economically. Deloitte recently performed a study on the "Big Five's" incomes in 2017/2018. The research indicated that the Premier League was at the top of earnings with €5.440 million, followed by the Bundesliga (€3.168 million), LaLiga (€3.073 million), Serie A (€2.217 million), and Ligue 1 (€1.692 million) (Barnard et al., 2019).

However, not all teams receive the same profit. A team's profit varies greatly depending on the team's ranking, so a team's position at the end of the season is of great importance. In Spanish football, 90% of television revenues go to the First Division (LaLiga), and only 10% is allocated to teams that play in the Second Division (LaLiga2). In this context, it is not surprising that data analysis has fostered extensive literature that investigates which game actions promote a team's success.

Discriminant analysis has been widely used to identify game actions that allow discrimination between losing, drawing, and winning teams (Lago-Ballesteros and Lago-Peñas, 2010; Lago-Peñas et al., 2011; Castellano et al., 2012; Liu et al., 2016). These analyses combine performance variables with categorical variables such as the match location (home or away) and the opposition quality.

The players' position and performance have been studied in other fields (Carpita and Golia, 2021; Carpita et al., 2021); however, the researchers' results are difficult to compare because no standard criteria exist to determine the number of positions in the field. The investigators classified the positions according to their roles or according to their location in the match. In general, three classification groups are common for all studies: defender, midfielder, and forward. Through video analysis, researchers have analysed how players' physical activity varies (e.g., distance covered during high-speed running and low-speed running) depending on the player's position. These studies have been performed in the country leagues: LaLiga (Di Salvo et al., 2007), Premier League (Gregson et al., 2010), Serie A (Vigne et al., 2010), and Ligue 1 (Carling, 2011); and in the Champions League (Rampinini et al., 2007; Bradley et al., 2010).

The essential variables for team classification at the end of a season have also been an object of analysis (Oberstone, 2009; Lago-Peñas et al., 2010; Brito de Souza et al., 2019). The authors compared the position of the teams and identified the variables that discriminate between them. These investigators carried out the research using one-way analysis of variance (ANOVA) (Oberstone, 2009; Lago-Ballesteros and Lago-Peñas, 2010) and the Student's t-test (Brito de Souza et al., 2019). These univariate statistical techniques might overlook valuable information. When analysing variables independently, univariate methods do not take into account the relationship between variables. Therefore, it is impossible to know the correlations between the statistically significant variables.

This is a crucial disadvantage of these methods because the information they provide is incomplete, making it difficult to achieve a global vision of football teams' behaviour in the field of play.

Oberstone (2009) also used points earned during the 2007-2008 English Premier League season to perform a multiple linear regression to identify statistically significant game actions for a team's success at the end of a season. However, he pointed out the possible existence of multicollinearity issues, so he was forced to refine the analysis using backwards elimination in a stepwise regression. A disadvantage of using this method is that correlated variables, important for discrimination between groups, could be eliminated, making for an incomplete interpretation of the vital game actions.

To overcome the limitations of these previous studies, in this paper, we will analyse the game actions that have a greater contribution to the success or unsuccess of a team using different multivariate analysis methods more useful than the univariate techniques used thus far by some authors. As in previous analyses which focus on the difference between successful and unsuccessful teams (Liu et al., 2015; Brito de Souza et al., 2019), the main objective of this research is to determine which game actions contribute significantly to reaching Champions League or Europa League positions (top teams) or avoiding being relegated to the second division (bottom teams) and, therefore, which variables would indicate a high probability of success or failure in future seasons.

This paper has five sections. The Section 2 is devoted to describing the database and explaining the data processing. The Section 3 presents the methodology and summarises how the multivariate methods have been applied. The Section 4 introduces the game actions with a higher contribution to a team's success to discriminate between each analysed group depending on the method used. In addition, results from multivariate methods are compared with those from univariate techniques. Finally, the Section 5 presents the discussion and conclusions of the paper.

2 Database

The database comprises 35 observations (top and bottom teams) corresponding to the clubs that competed in LaLiga, Premier League, Bundesliga, Serie A, and Ligue 1 in the 2018-2019 season, and has 53 variables. The two first variables are the team's name (categorical) and the league ranking's final position, either top or bottom (ordinal). This ordinal variable was determined based on previous investigations (Oberstone, 2009; Lago-Peñas et al., 2010; Brito de Souza et al., 2019), with the top clubs classified as the Champions League or Europe League (20 teams), and the bottom clubs are those that descended to the second division (15 teams). The 51 remaining variables are quantitative and collect information about game actions. These variables were collected by accessing several internet sources, such as WhoScored, FBref and Fichajes.net

The variables used stored information accumulated about the game actions carried out throughout the season. These variables were mean-centred and scaled to unit variance. In this way, variables measured at different scales have the same influence, a priori, in the models and can be analysed together. Table 1 shows all the variables classified into

four groups according to the nature of the game's actions.

Table 1: Variables classified by type of game actions

Type of variables	**Game actions and abbreviations**
Variables related to defensive actions	Shots conceded blocked (SCB), Recoveries (R), Clean sheets (CS), Penalties conceded (PC), Interceptions (I), Shots conceded on target inside the box (SCTI), Shots conceded on the target outside the box (SCTO), Tackles won (TW), Tackles lost (TL), Yellow cards (YC), Clearances (Cl), Fouls conceded (FC), Tackles accuracy (TA), Aerial duel won (ADW), Aerial duel lost (ADL), Aerial duel accuracy (ADA), Goal conceded inside the box (GCIB), and Goal conceded outside the box (GCOB)
Variables related to offensive actions	Key passes (KP), Penalties took (PT), Shots off target (SOT), Shots blocked (SB), Corners won (CW), Crosses unsuccessful (CU), Shots accuracy (SA), Assists (A), Shots on target (ST), Successful crosses (SC), Dribbles successful (DS), Fouls won (FW), Dribbles unsuccessful (DU), Crossing accuracy (CrA), Dribbles accuracy (DrA), Duels Accuracy (Dl_A), Duels lost (DlL) and Duels won (DlW)
Variables related to the goal	Goals accuracy (GA), Goals inside the box (GIB), Goals outside the box (GOB) and Direct free kick goals (DFKG)
Variables related to passes and possession	Passing accuracy (PA), Average possession (AP), Passes unsuccessful opponents half (PUOpp), Passes successful opponents half (PSOpp), Successful longpasses (SLP), Unsuccessful shortpasses (PUS), Successful shortpasses (PSS), Unsuccessful longpasses (ULP), Passes per 90 mins (P_90), Passing accuracy in opponents half (PAOppH), Passing accuracy in own half (PAOwnH) and Longpasses success (LPS)

Note that, before standardising the variables, to check to what extent the distribution of some of the variables may differ between the different leagues, a partial least squares discriminant analysis was performed using the leagues as dependent variables for both bottom and top teams. The results (not shown and available under request) indicated that neither variable was statistically significant (p-values lower than 0.05) in discriminating between leagues in both cases. Appendix shows comparative boxplots of game actions with standardised values (Figures A.1. and A.2.) and comparative tables with the mean and standard deviation (Tables A.1. and A.2.) of game actions for each league and team level. After these analyses, it was concluded that the behaviour between leagues was pretty similar. Thus, the standardisation process will not lose im-

portant information about a particular league or playing style, and we can study all the leagues together. Also, like us, some other authors study all the teams (Collet, 2013; Malagón-Selma et al., 2022) or players (Decroos et al., 2019) of the "Big Five".

3 Statistical methods

The following section proposes different statistical methods. Firstly, an exploratory analysis was carried out using the principal component analysis (PCA) (Wold et al., 1987). We propose to use PCA as a multivariate exploratory tool capable of detecting outliers that can compromise any subsequent statistical analysis, visualizing complex network relationships between the data, and providing some insight into the variables with different behaviours between the top and bottom teams. Secondly, a confirmatory analysis was performed to determine which game actions had a statistically significant contribution to a team's success at the end of the season. The multivariate models selected for these analyses were the following: random forest (RF) (Breiman, 2001), a machine learning algorithm that has been used to identify important game actions within several sports (Carpita et al., 2015; Migliorati, 2020; Smithies et al., 2021; Whitehead et al., 2021); partial least squares discriminant analysis (PLS-DA) (Wold et al., 1993), a method with excellent results in most of the fields where it has been used, especially indicated when there are many correlated variables (Gottfries et al., 1995; Worley and Powers, 2013; Noçairi et al., 2016); and logistic regression (LR) (Nelder and Wedderburn, 1972), a classical statistical model used as a benchmark. Previous studies have used regression models to find the match statistics that best explain the number of points obtained at the end of the season (Oberstone, 2009; Brito Souza et al., 2019), the effect of ball possession in team success (Lago-Peñas et al., 2010; Collet, 2013) and the final result in a match (Lago, 2009; Liu et al., 2016).

The free software R was used to analyse the database (R Core Team, 2019). This computer language uses high-quality statistical and graphical methods. RStudio (RStudio Team, 2020), an integrated development environment (IDE), was used to program in R.

3.1 Exploratory analysis

The exploratory analysis was carried out by using PCA, an unsupervised learning technique that uses the covariances between variables to create principal components (PCs). The PCA model provides detailed multivariate information about each observation's value for each extracted principal component, allowing the study and visualization of the relationship between observations (through the scores scatterplot) and also the relationships between variables (through the loadings scatterplot). PCA also provides two valuables' statistics for outlier detection (Ferrer, 2007) such as SPE and Hotelling's T^2.

SPE is defined as:

$$SPE_n = e'_n e_n = (x'_n - \hat{x}'_{n,A})(x_n - \hat{x}_{n,A}) = \sum (x_{n,k} - \hat{x}_{n,k,A})^2$$

where x'_n is the row vector of actual values of each *n-th* observation ($n = 1, \ldots, N$) in the K variables, and $\hat{x}'_{n,A}$ is the row vector of predicted values of that observation in the K variables obtained from the PCA model with A PCs. SPE measures the squared Euclidean distance of an observation from the subspace of the principal components. The SPE statistic can be modelled by the (noncentral) χ^2 distribution:

$$SPE \sim g\chi^2_h$$

An observation with a high SPE is an anomalous observation due to the breakage of the correlation structure modelled by the PCA model.
On the other hand, Hotelling's T^2 statistic is defined as:

$$T^2_n = \sum_{a=1}^{A} \left(\frac{t_{n,a}}{s_a} \right)^2,$$

where $t_{n,a}$ is the score value of the *n-th* observation in the *a-th* dimension ($a = 1, \ldots, A$) and s_a is the variance of the *a-th* principal component. Hotelling's T^2 measures the estimated squared Mahalanobis distance of the projection of observations onto the principal components' subspace to the centre of this subspace. Hotelling's T^2 statistic can be modelled as a Snedecor F-distribution:

$$T^2 \sim \frac{A(N^2-1)}{N(N-A)} F_{A,(N-A)}.$$

An observation with a high Hotelling's T^2 is an extreme observation due to the extreme values (high or low) in certain variables.

The 95% and 99% control limits are calculated for both statistics. Therefore, it is expected that 5% or 1% of the observations (depending on the control limit, respectively) will have values for those statistics slightly higher than these control limits. This is called the false alarm rate. On the other hand, if an observation highly exceeds any of these control limits (e.g. more than two or three times its corresponding control limit), it will be considered an outlier and will be analysed in detail.

3.2 Confirmatory analysis

The second step was to carry out a confirmatory analysis of the results obtained in the previous step. Three multivariate supervised models were selected to determine which variables are related to the team position. These models help us to interpret the results regarding which variables have the most statistically significant information to discriminate between top versus bottom teams.

3.2.1 Partial least squares discriminant analysis

Partial least squares discriminant analysis (PLS-DA) (Barker and Rayens, 2003) is a PLS variant for classification models. The Y matrix is built with dummy variables (as many as classes to be considered, in our case: two classes, Top and Bottom). PLS models

find latent variables in the X space (predictors) and Y space (responses) with maximum covariance. In our case, the dimension of X space is K, and the dimension of Y space is 2. As a result, a few latent variables explain the sources of variability in the X space that are related to the Y space. The variable influence on projection (VIP) measures the importance of the variables given by the PLS-DA model. The VIP score provides information about the contribution of each variable to the model (Eriksson et al., 2013) by means of the following expression:

$$VIP_k = \sqrt{K \times \left(\frac{\sum_{a=1}^{A} w_{ka}^2 \times ssy_{total}}{ssy_{total}} \right)}$$

where w_{ka}^2 is the squared PLS weight of a particular $k-th$ variable for each PLS $a-th$ dimension, SSY_a is the sum of squares explained by that PLS dimension, SSY_{total} is the total sum of squares explained by the PLS model and K is the number of variables. Thus, the result is a weighted sum of the squared correlations between the PLS-DA components and the original $k-th$ variable. The weights correspond to the percentage variation explained by each PLS-DA component in the model.

Jackknife confidence intervals are calculated to select the statistically significant variables. The jackknife method (Quenouille, 1949) is a resampling technique that estimates the variability of the PLS regression coefficients. The procedure calculates the model's coefficients with N-1 observations (i.e. deleting one observation at a time) as many times as observations, N, the database has. Finally, the 95% jackknife confidence intervals are obtained from the N regression coefficients (one for each iteration). The statistically significant variables will be those whose 95% jackknife confidence intervals do not contain the zero value.

3.2.2 Random forest

The RF algorithm is an ensemble method typically used in machine learning that uses the bagging technique to combine trees at random and improve predictability. The process begins with random sampling (with replacement) of L observations (approximately two-thirds of the data) from the original database that constitutes the training dataset to build the tree, and the remainders are out of bag observations (OOB). Unsampled observations (approximately one-third of the data) make up the test set and will be used to calculate the prediction error. A decision tree, independent from the rest, is created from each dataset, with a subset of variables randomly selected at each split. In addition, each tree grows deep and without pruning.

One of the advantages of the RF algorithm is that it allows us to know the importance of variables in the classification model. Out of the four ways proposed by Breiman (2001) to calculate the importance of variables, (Liaw and Wiener, 2002) integrated only two, mean decrease accuracy (MDA) and mean decrease Gini (MDG), in their R-package ramdonForest. The MDA is obtained by calculating the increase in the prediction error in OOB when one variable's values are permuted in the training dataset while the others remain unchanged. Thus, the greater the decrease in precision, the greater the

importance of the variables. On the other hand, MDG follows the logic of the Gini criterion. Every time a split on a node is on a particular variable, the impurity of the descendant nodes decreases. The impurity reaches its minimum (0) when all the split node observations belong to the same group. Thus, the greater the decrease in the Gini coefficient, the greater its importance. The sum of all the Gini reductions for each variable, normalised by the number of trees, will result in the MDG.

Since the R-package developed by Liaw and Wiener (2002) does not include the p-value calculation that provides the statistical significance, two tests will be used in this paper to determine the statistical significance of the variables in the RF model. First, the p-value is calculated using the one-sided binomial test (Paluszyńska, 2017), where the null hypothesis is that the selection of a variable occurs by chance. Let W be the number of nodes that allocate the variable X_k in a forest if the selection is random. W can be modelled as a binomial statistical distribution $B(F, p)$, where F is the total number of nodes in the forest, and $p = 1/K$. Therefore, a variable X_k will be statistically significant if the observed number of nodes where this variable appears in the forest (r) is greater than the 95% percentile of the binomial distribution, and the p-value will be calculated as $P(W \geq r)$. Second, we propose carrying out a permutation test to calculate the p-value. Permutation tests are non-parametric procedures for determining statistical significance based on rearrangements of the labels of a dataset (Edgington and Onghena, 2007). As in all statistical hypothesis tests, the statistical significance is represented by its p-value, where the null hypothesis is defined as the labels assigning samples to classes are interchangeable (Edgington and Onghena, 2007). Thus, a p-value lower than 0.05 indicate that the labels are not interchangeable, and that the original classification of each observation is relevant (Knijnenburg et al., 2009). In the same way as Tusher et al. (2001) and Subramanian et al. (2005), who use permutation tests to compute statistical significance through the randomly rearranged class labels, we also randomly rearranged the teams' classification to obtain the permutation values. We fitted 1000 RF with and without the permuted response variable. P-values were calculated as the proportion of times the MDA and MDG values obtained from permuted data are equal or greater than the original MDA and MDG, where the response variable remained without permutation. Both for one-sided binomial and permutation test, the statistically significant variables are those corresponding to p-values lower than 0.05.

3.2.3 Logistic regression

LR is a statistical model used to model binary response variable (1 or 0). The logistic model is a well-known model (Cox, 1958) that gives the probability that an individual belongs to a class; this probability is obtained from the explanatory variables' relationship. An increase or decrease in the value of an explanatory variable can change the corresponding probability. However, before carrying out the logistic model, it was necessary to avoid multicollinearity between the explanatory variables. If the variables are too correlated, multiple models with different statistically significant explanatory variables and regression coefficients but similar classification performance will be obtained, making model interpretation difficult. The variance inflation factor (VIF) quantifies the

multicollinearity between those variables.

The VIF is defined as:

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the multiple R^2 for the regression of X_k on the remaining covariates.

A high VIF value indicates high multicollinearity. Thus, it was necessary to remove the variables with the highest VIF values. Once multicollinearity was controlled, it was possible to apply the LR to the remaining explanatory variables. Finally, the stepwise method was used to select the most relevant variables.

3.2.4 Resampling method for comparison of the model's performance

The cross-validation technique (G-fold) was used to evaluate the model's performance. This method consists of randomly dividing the database into G subgroups (in this case, G=5). Once the division is done, the training set with subgroups G-1 is created, and the test set (to validate the model) is created with the remaining subgroup. This process is repeated G times, using independent datasets to train and validate the model, concluding when all the individuals have been in the validation subgroup once (Stone, 1974). Note, that it was taken into account that the subgroups were balanced. Therefore, each subgroup consisted of 7 individuals (4 top and 3 bottom).

To evaluate and compare the performance of each test set, receiver operating characteristic (ROC) curves were used (Fawcett, 2006), and summarised as the area under the ROC curve (AUC). AUC is a single scalar value representing a portion of the unit square area underneath the ROC curve. This value will always be between 0 and 1, so the greater the AUC is, the better the classification method. Finally, a two-way ANOVA, with the model as the main factor and the test set as a block factor, was used to test the statistically significant differences between the models.

Note that in addition to assessing the prediction error, cross-validation helps us to select models with results that can be generalized to other data sets (Refaeilzadeh et al., 2009).

3.2.5 Univariate approach: two-sample tests

Although in this paper we want to show the advantages of using multivariate techniques for analysing the most significant variables for discriminating between successful and unsuccessful teams, since there are papers (Oberstone, 2009; Lago-Peñas et al., 2010; Brito de Souza et al., 2019) that use univariate techniques, we are going to compare the results of both approaches.

The Student's t-test was used to statistically contrast each univariate variable's effect on the final ranking (top vs. bottom). There was a check that the variables complied with the conditions of normality and homoscedasticity using the Shapiro–Wilk test (Shapiro and Wilk, 1965) and Levene test (Levene, 1961), respectively. If normality or homoscedasticity assumptions were rejected, the nonparametric Mann–Whitney test

(Wilcoxon, 1945) or the Welch t-test (Welch, 1947) were performed, respectively. Again a p-value lower than 0.05 was considered the threshold for statistical significance.

4 Results

4.1 Exploratory analysis

Principal component analysis (PCA) was used as an exploratory multivariate analysis method to provide a global vision of the relationships between the variables. Note that this is not possible using univariate techniques. PCA was obtained by the pca function from the mixOmics R-package (Rohart et al., 2017). This library offers the possibility of using a wide range of multivariate methods for data exploration (PCA, PLS, PLS-DA). First, it is necessary to determine the number of components to obtain lower-dimensional data while preserving as much of the data variation as possible. The evolution of the cumulative explained variance ratio vs. the number of principal components in the PCA model showed that the first component explains 47% of the total variance, the second: 9%, the third: 9%, the fourth: 7%, the fifth: 4%, and so on in decreasing order. In order to explain at least 80% of the total variance, we decided to keep seven components in the PCA model. Next, SPE and Hotelling's T^2 with 95% and 99% control limits were used to check for anomalous or extreme observations, respectively. It was concluded that there were no anomalous or extreme observations.

Focusing on the first two PCs, that jointly explain 56% of the total variability, the scores scatterplot (Figure 1) allows us to visualise the relationship between the teams and the leagues, whereas the loadings scatterplot (Figure 2) shows the relationship between the variables. Figure 1 shows the scores scatterplot of the first two PCs; the top teams are coloured in blue and the bottom teams are coloured in red. Figure 1 reveals that the teams appear clustered along the first principal component, which splits the top (right) and bottom (left) teams. These two components were selected after performing all score plots for each pair of PCs and verifying that the first PC discriminates the most between the groups. From Figure 1 it is clear that the teams of the five leagues studied are not grouped within each top and bottom cluster. This matches the results shown in Aappendix and justifies that all the leagues can be analysed together.

Figure 2 shows the loadings scatterplot of the first two PCs: the further a variable is from the centre of the plot and the closer to a particular PC, the greater its relationship to that PC, and vice versa. On both extremes along the x-axis, there are the most correlated variables with PC1, the most negatively correlated, in blue, and the most positively correlated, in red. Figure 2 shows two clusters of variables: on the left, those related to defensive actions (SCB, SCTO, SCTI, GCOB, GCIB, ADW, ADL, CI, and I) and passes unsuccessful (ULP); and on the right, those related to offensive actions (A, SA and ST), goal (GA and GIB), and passes and possession (AP, PA, P90, PAOwnH, LPS, PAOppH, PSS, and PSOpp). Variables from both clusters will be positively correlated inside a cluster and negatively correlated between clusters.

By comparing Figures 1 and 2, one can expect that variables on the plot's right will take higher values in top teams (shown in red) than in bottom teams (shown in blue),

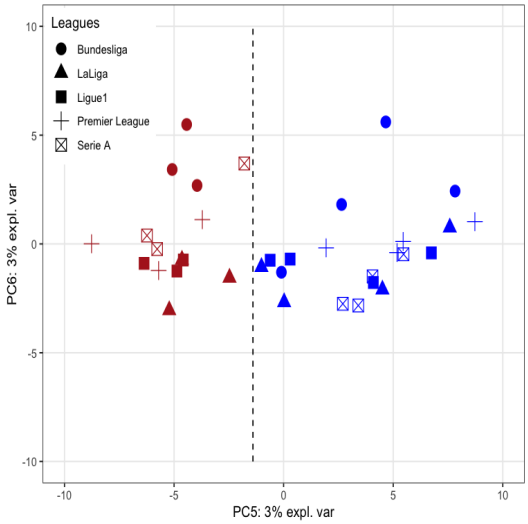


Figure 1: PCA scores scatterplot of the teams and leagues projected in the PC1/PC2 space: top teams in blue and bottom teams in red

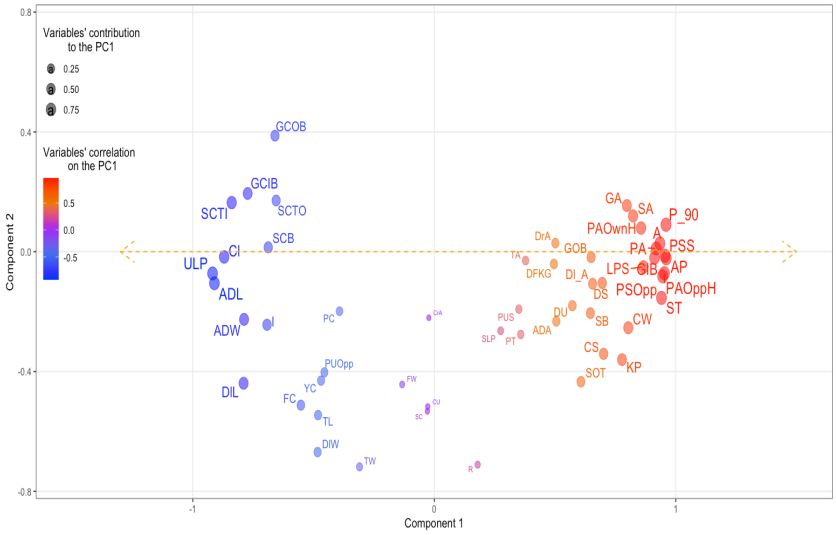


Figure 2: PCA loadings scatterplot of the variables in the PC1/PC2 space sized by a variable's correlation strength to PC1.The colour of the dots indicates the negative (blue) or positive (red) correlation of the variables with PC1. Orange dotted arrow indicates the direction of the most discriminating PC

while variables on the plot's left will take higher values in the bottom teams than in the top teams. This can provide some preliminary clues on the variables with discriminant behaviour between the top and bottom teams.

It is important to note that the fact that the first component is the one that most discriminates is a coincidence that might not occur in other analyses. In any case, by exploring the different PCs, one may find any combination of PCs with discriminant behaviour.

4.2 Confirmatory analysis

The confirmatory data analysis was carried out using PLS-DA, RF, and LR. As we mentioned before, these methods were selected because they are supervised models that can provide information about which variables have the most statistically significant information for classifying top versus bottom teams.

4.2.1 PLS-DA

In PLS-DA, the first step was to calculate the optimal number of latent variables using the mixOmics R-package's perf function, which estimates the mean squared error of prediction (MSEP) and indicates the number of PLS components necessary to obtain the best predictive model. In this case, the model chosen had 2 components and a goodness of fit (i.e., explained percentage of total variance) of 55%. The second step was to analyse the SPE and Hotelling's T^2 statistics to verify that there were no anomalous or extreme data, respectively. As in the case of the PCA, no anomalous or extreme observations were detected.

Figure 3 shows the scores scatterplot of the first two PLS-DA components coloured by the classification label (top teams in blue and bottom teams in red). The first PLS-DA component clearly discriminates between top (right) and bottom (left) teams.

Figure 4 shows the weightings scatterplot, revealing the relationships between the explanatory variables and the top and bottom categories. The furthest away from the centre, and the closest a variable is to the centroids of the top and bottom categories, the greater its contribution to classifying those classes (with higher values in the closest category than the opposite -furthest- one).

Figure 4 shows that bottom teams are close in shots and goals conceded inside and outside the box (SCTI, SCTO, GCIB, and GCOB). In contrast, top teams are close in the number of matches that the team finished without receiving a goal (clean sheets, CS), goals scored (GA), and percentage of duels, one-to-one, where a player wins the ball (Dl_A). Thus, the higher the number of shots and goals conceded, the higher the probability of belonging to a bottom team. Similarly, the higher the number of goals scored and the number of matches that the team finished without conceding a goal, the higher the probability of belonging to a top team.

Figure 5 shows the regression coefficients with 95% jackknife confidence intervals of the PLS-DA model for the top class, ordered from most positive to most negative values, and calculated from the mdatools R-package (Kucheryavskiy, 2020). The statistically significant variables are those whose jackknife confidence intervals do not contain the zero value. These are summarised in Table 2. Variables with positive regression coefficients (such as CS, A, ST, GIB, GA, PSOpp, P90, AP and PSS) will take, on average, higher

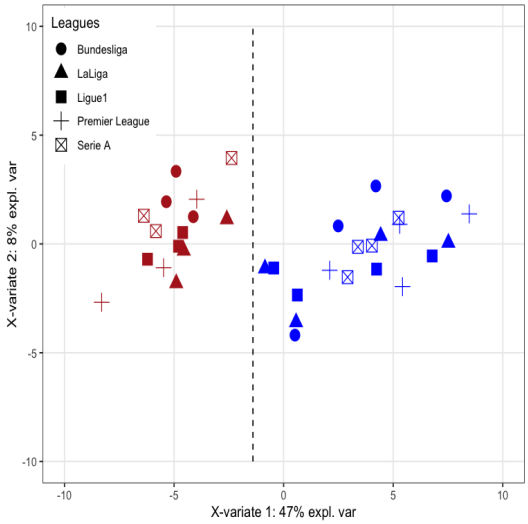


Figure 3: PLS-DA scores scatterplot of the distribution of the teams and leagues projected in the PLS-DA1/PLS-DA2 space: top teams in blue and bottom teams in red

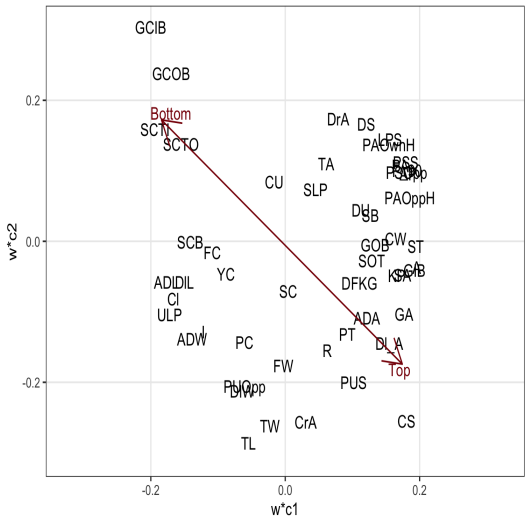


Figure 4: PLS-DA weightings scatterplot showing the relationship between the explanatory variables and the response variables in the PLS1/PLS2 space

mean values in top teams than in bottom teams. In contrast, variables with negative regression coefficients (such as SCTI, SCTO, GCIB, GCOB, ADL, Cl, DIL and ULP) will take, on average, higher mean values in the bottom teams than in top teams. The PLS-DA regression coefficients to predict the bottom class are not shown, as they are a

specular image of the model to predict the top class.

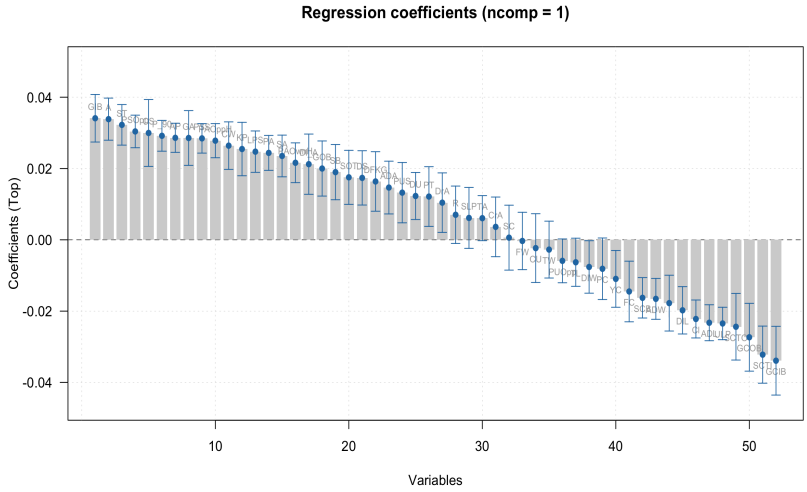


Figure 5: PLS-DA regression coefficients with 95% jackknife confidence intervals for the variables to predict the top teams

Therefore, according to the results of the weighting plot (Figure 4), and jackknife confidence intervals (Figure 5), it is possible to confirm that the variables that contribute the most to discriminate between top vs. bottom teams match those explained by the first principal component (Figure 2).

4.2.2 Random forest

The randomForest R-package was implemented in R by Liaw and Wiener (2002). First, the optimisation of hyperparameters mtry, the random number of variables selected in each tree and the nodesize, the minimum size of the terminal nodes, were calculated. Thus, the algorithm was carried out using these obtained optime values mtry=7 and nodesize=9. The output gave the MDA and MDG for each variable.

Figure 6 shows the variance importance scatterplot with the MDA values (x-axis) and the MDG values (y-axis). Variables are labelled in red, green or blue depending on the p-value calculated through the one-sided binomial test obtained using the randomForestExplainer R-package (Paluszynska et al., 2020). Figure 6 highlights the variables with a p-value lower than 0.01 (in red), which match the essential variables for the two measures MDA and MDG: number of shots on target (ST), shots conceded on target inside the box (SCTI), effectivity (GA), assists (A), goals conceded on target inside the box (GCIB), number of goals inside the box (GIB), key passes (KP), average possession (AP), number of passes per 90 mins (P90), passes successful opponents half (PSOpp), and number of matches that the team finished without receiving a goal (CS).

P-values for MDA and MDG using the permutation tests were also used. Table 2 shows

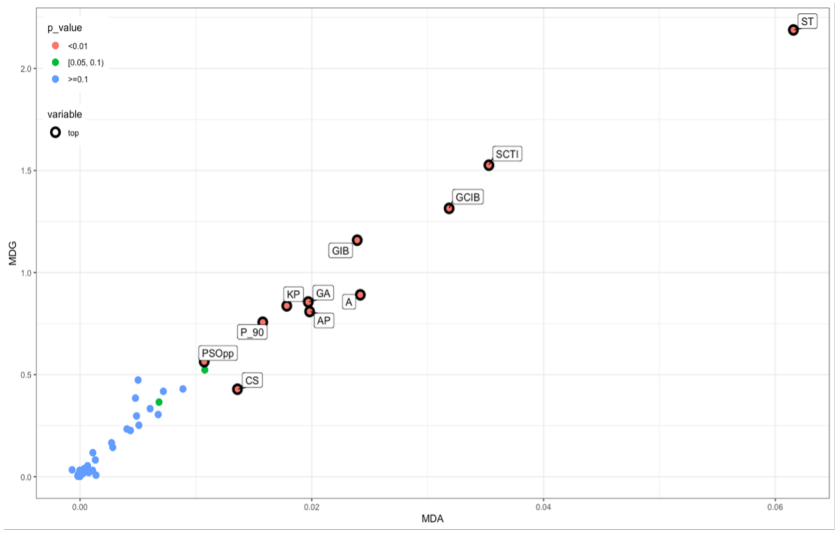


Figure 6: Multiway importance plot with mean decrease accuracy (MDA) and mean decrease Gini (MDG)

that the one-sided binomial test and the MDG permutation tests yield similar results. However, the MDA permutation tests selects, in addition to the variables selected by the other two approaches, more variables already selected from PLS-DA. These differences could be due to both the MDG and the one-sided binomial test using the Gini criterion. Several papers have indicated that the Gini index may be less reliable because it benefits variables that offer many categorical cut-off points or have missing values (Kim and Loh, 2001; Strobl et al., 2007).

4.2.3 Logistic regression

The last proposed model, LR, was carried out in 2 steps. First, `vif_function` was used to alleviate the multicollinearity (Beck, 2013) by eliminating correlated variables. The function calculates the VIF value of all explanatory variables, and removes the variable with the highest VIF. Next, the VIF is recalculated, and the variable with the highest VIF value is eliminated again. This process is repeated until all variables have a VIF lower than a threshold. The function output is the name of these variables. The threshold values used for the VIF were 2.5, 5, and 10, since these values are commonly used in different contexts (Kutner et al., 2005; Sheather, 2009; Johnston et al., 2018). Finally, the LR was fitted using the stats R-package. Later, the model with the most relevant variables is selected according to the Akaike information criterion (AIC) (Akaike, 1974) by means of the stepAIC function. Table 2 shows the statistically significant variables for the LR models with the three thresholds considered, and the above models PLS-DA and RF.

Table 2 shows that the LR is the model that selects the least number of statistically

Table 2: Comparison of the statistically significant variables in the PLS-DA, RF and LR (thresholds 2.5, 5 and 10) models

Methods	**Variables related to defensive actions**	**Variables related to offensive actions**	**Variables related to the goal**	**Variables related to passes and possession**
PLS-DA: Jackknife intervals	SCTI[6], GCIB[4], CS[3], GCOB[3], SCTO[3], ADL[3], Cl[3], YC[2], ADA[2], SCB[1], I[1], FC[1], and ADW[1]	A[4], ST[4], KP[4], Dl_A[4], DlL[2], CW[2], SA[2], PT[1], SOT[2], SB[1], DrA[2], DS[1], DlW[1], and DU[1]	GIB[4], GA[4], DFKG[3], and GOB[1]	AP[4] P_90[4], PSOpp[4], PUS[4], PAOwnH[3], PSS[3], LPS[2], ULP[2], PA[2], and PAOppH[2]
RF: Binomial test	SCTI, GCIB, CS, and GCOB	A, ST, KP, and Dl_A	GIB and GA	AP, P_90, PSOpp, and PSS
RF: MDA	SCTI, GCIB, CS, GCOB, SCTO, ADL, and Cl	A, ST, KP, Dl_A, DlL, CW, and SA	GIB and GA	AP, P_90, PSOpp, PAOwnH, PSS, LPS, ULP, PA, and PAOppH
RF: MDG	SCTI, GCIB, and ADL	A, ST, and KP	GIB and GA	AP, P_90, and PSOpp
LR:10	SCTI, SCTO, ADA, Cl, TW[1], and TA[1]	Dl_A, SOT, and DrA	-	PUS, PUOpp[1]
LR:5	SCTI	CrA[1]	DFKG	PUS and PAOwnH
LR:2.5	YC	SOT and FW[1]	DFKG	PUS

The first time a variable appears, it is accompanied by a number that indicates the number of models for which it was statistically significant.

significant variables. The reason is that, unlike PLS-DA and RF, LR penalizes the inclusion of collinear regressors through the VIF factor. Therefore, the LR model tries to keep only variables that are not, or slightly correlated, between them, regardless of whether the variables not selected have a relationship with the response variable. For this reason, LR suffers from interpretation problems with collinear regressors.

4.2.4 Comparison of models performance

Using cross-validation for each test set, the AUC statistic was obtained by means of the ROCR R-package (Sing et al., 2005). Later, a two-way ANOVA was used to test

statistical differences between average model performance using AUC. The model factor was found to be statistically significant (p-value=0.0028). Therefore, to determine which models differed statistically from each other, we performed Fisher's post hoc 95% LSD interval test implemented in the agricolae R-package (de Mendiburu, 2021). Note that Fisher's post hoc 95% LSD interval test does not take into account the AUC constraint (values between 0 and 1), so the LSD intervals have been trimmed.

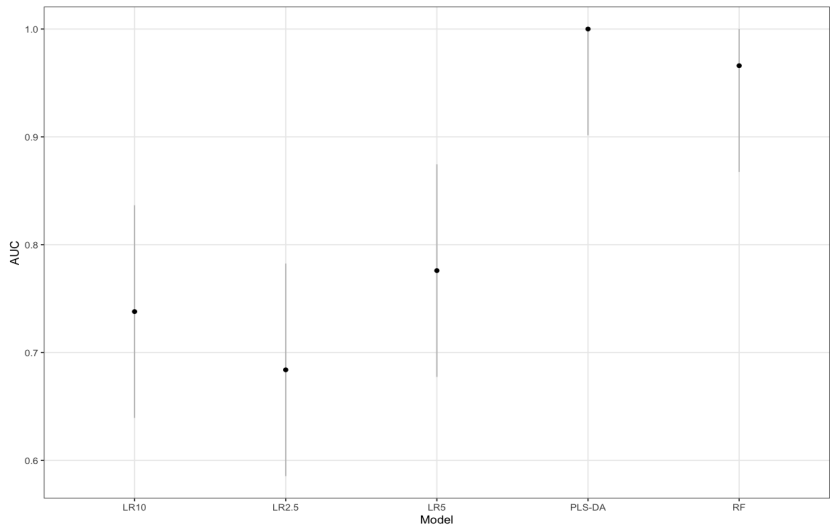


Figure 7: Multiple comparisons of the models (X-axis) vs. the AUC (Y-axis). The black points indicate the mean AUC for each model, and the intervals are based on 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences

From Figure 7, it is possible to conclude that the average AUC is statistically higher in the PLS-DA and RF models than in the three LR models. However, there were no statistically significant differences in the average AUC between RF and PLS-DA. Nevertheless, Table 2 shows that in this study, PLS-DA is the method that selects more essential variables for successful and unsuccessful soccer teams.

4.2.5 Univariate approach: two-sample tests

This section carries out univariate statistical two-sample tests using the stats (R Core Team, 2019) and car (Fox and Weisberg, 2019) R-packages. Previously, normality and homoscedasticity were tested through the Shapiro–Wilk test and the Levene test, respectively, from the library stats. The Mann–Whitney test was used if normality could not be accepted, and the Welch approximation was used in cases of heteroscedasticity. Table 3 shows the statistically significant variables for the univariate analysis for a p-value less than 0.05.

The only group in which all the game actions were statistically significant were vari-

Table 3: Statistically significant variables (p-values<0.05) for the two-sample test (top vs. bottom teams)

Type of variables	**Game actions**
Variables related to defensive actions	$SCTI^{B}$, $GCIB^{B}$, CS^{T}, $GCOB^{B}$, $SCTO^{B}$, ADL^{B}, Cl^{B}, YC^{B}, ADA^{T}, SCB^{B}, I^{B}, FC^{B} and ADW^{B}
Variables related to offensive actions	A^{T}, ST^{T}, KP^{T}, Dl_A^{T}, DlL^{B}, CW^{T}, SA^{T}, PT^{T}, SOT^{T},SB^{T}, DS^{T}, and DU^{T}
Variables related to the goal	GIB^{T}, GA^{T}, $DFKG^{T}$, and GOB^{T}
Variables related to passes and possession	AP^{T}, P_90^{T}, $PSOpp^{T}$, PUS^{T} $PAOwnH^{T}$, PSS^{T}, LPS^{T}, ULP^{B}, PA^{T}, and $PAOppH^{T}$

^{T} indicates the variables that take higher mean values in the top teams than the bottom and ^{B} vice versa.

ables related to the goal. The univariate statistical test results shown in Table 3 basically agree with the results preliminary highlighted from PCA and finally confirmed from PLS-DA.

5 Discussion and conclusion

Although the comparison between successful and unsuccessful teams has been previously explored (Oberstone, 2009; Lago-Peñas et al., 2010; Brito de Souza et al., 2019), the current analysis is a novel proposal because it includes the game actions of the teams that competed in the first division of the Bundesliga, Premier League, LaLiga, Ligue 1 and Serie A throughout a season. To the best of our knowledge, although previous researchers have used data from European competitions (Collet, 2013; Decroos et al., 2019), this study constitutes the first analysis performed with match statistics of the five best leagues of the world on this topic.

Another novelty of this paper is that the comparison between the top and bottom teams was carried out using multivariate statistical techniques that have not been used previously in the literature. Additionally, the multivariate statistical methods have also been compared with the results of the classical two-sample univariate tests commonly used in the literature. In this sense, this paper shows the benefits of using PCA as a very effective technique to carry out a preliminary exploratory data analysis. Instead of looking at one variable at a time, as done with univariate exploratory analysis techniques, PCA allows us to look at all variables at a time and interpret the multivariate data information jointly. In particular, PCA is highly efficient for detecting outliers, finding patterns in the observations and visualising the relationship between the vari-

ables. All this information provides a preliminary global insight into the game actions that differentiate between top and bottom teams that were confirmed using supervised multivariate techniques. Out of these supervised techniques studied, PLS-DA selects the highest number of most relevant variables with statistically significant power to discriminate between the top and bottom teams (confirming the preliminary results obtained from PCA) and provide one of the models (jointly with random forest) with statistically better classification performance.

In previous analyses, only the goal average conceded per match (Oberstone, 2009), shots conceded, recoveries (R), yellow cards (YC), and fouls conceded (FC) (Brito de Souza et al., 2019) were statistically significant defensive variables. However, the PLS-DA model detected a high number of variables related to defensive actions (SCTI, GCIB, CS, GCOB, SCTO, ADL, Cl, ADA, I and ADW) in addition to the variables found in previous studies. Regarding the offensive variables, researchers differ in their results depending on the leagues, variables, and the number of seasons analysed (Oberstone, 2009; Lago-Peñas et al., 2010; Brito de Souza et al., 2019). However, the PLS-DA model found statistically significant variables studied in the previous analysis, and highlighted new variables not detected previously (DlL, CW, SA, PT, SOT, SB, DS, DrA, DlW and DU). In the variables related to goals, while previous studies only found the effectivity (GA) (Oberstone, 2009; Lago-Peñas et al., 2010; Brito de Souza et al., 2019), goals outside the box (GOB) (Oberstone, 2009) and free-kick goals (DFKG) (Brito de Souza et al., 2019) to be statistically significant, our PLS-DA model detected all the variables studied (GA, GOB, GIB and DFKG). In the case of the variables related to passes and possession, the PLS-DA detected game actions statistically significant for previous studies (Oberstone, 2009; Lago-Ballesteros and Lago-Peñas, 2010; Lago-Peñas et al., 2010; Casal et al., 2019; Brito de Souza et al., 2019). Additionally, the PLS-DA also found that the number of unsuccessful shortpasses (PUS), longpass success (LPS), unsuccessful longpasses (ULP), and passing accuracy in its own half (PAOwnH) were variables with high discriminant power.

Regarding the method commonly used for data analysis in soccer, the two-sample Student's t-test (Rampinini et al., 2009; Lago-Peñas et al., 2010; Brito de Souza et al., 2019), this univariate analysis is not very efficient. First, it is necessary to perform as many statistical tests as there are variables in the database, which also can lead to multiple comparison issues due to the high number of hypotheses tests performed. Second, since the variables (i.e. game actions) are analysed independently (i.e. one at a time), the tests do not provide information about the game actions relationships, making it difficult to achieve a global vision of football teams' behaviour in the field of play. Additionally, most of the variables that the Student's t-tests identified as statistically significant (see Table 3) were also detected by the multivariate models, in particular by the PLS-DA, which also selected additional game actions such as DrA and DlW (see Table 2). Anyway, note that the univariate approach may be useful only for testing the statistical significance of a single predictor but cannot be used if the study aims to predict the top/bottom teams.

This analysis represents a significant advance in sports analytics by proposing powerful multivariate techniques such as PCA and PLS-DA to incorporate into the analysis

toolkit. Regarding the game actions selected with discriminant power between top and bottom teams, this study highlights that not everything depends exclusively on the number of goals scored but that defensive and offensive strategies are necessary. It is important to note that although the variables used in the study measure the accumulated information of the game actions, and that at the end of the season all the teams know their final classification (so it is useless to use any predictive model to know it), this information can be helpful for coaches and data analysts to planning future seasons.

Acknowledgement

The authors want to express their gratitude to the Universitat Politècnica de València for the financial support through the FPI-UPV grant (PAID-01-19).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(3):166–173.

Barnard, M., Boor, S., Winn, C., Wood, C., and Wray, I. (2019). *World in motion: annual review of football finance 2019*. Deloitte.

Beck, M. (2013). Collinearity and stepwise vif selection. *Retrieved from http://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/*.

Bradley, P. S., Di Mascio, M., Peart, D., Olsen, P., and Sheldon, B. (2010). High-intensity activity profiles of elite soccer players at different performance levels. *The Journal of Strength & Conditioning Research*, 24(9):2343–2351.

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

Brito de Souza, D., Campo, L.-D., Blanco-Pita, H., Resta, R., Del Coso, J., et al. (2019). An extensive comparative analysis of successful and unsuccessful football teams in laliga. *Frontiers in Psychology*, page 2566.

Brito Souza, D., López-Del Campo, R., Blanco-Pita, H., Resta, R., and Del Coso, J. (2019). A new paradigm to understand success in professional football: analysis of match statistics in LaLiga for 8 complete seasons. *International Journal of Performance Analysis in Sport*, 19(4):543–555.

Carling, C. (2011). Influence of opposition team formation on physical and skill-related performance in a professional soccer team. *European Journal of Sport Science*, 11(3):155–164.

Carpita, M., Ciavolino, E., and Pasca, P. (2021). Players' role-based performance composite indicators of soccer teams: A statistical perspective. *Social Indicators Research*, 156(2):815–830.

Carpita, M. and Golia, S. (2021). Discovering associations between players' performance

indicators and matches' results in the European Soccer Leagues. *Journal of Applied Statistics*, 48(9):1696–1711.

Carpita, M., Sandri, M., Simonetto, A., and Zuccolotto, P. (2015). Discovering the drivers of football match outcomes with data mining. *Quality Technology & Quantitative Management*, 12(4):561–577.

Casal, C. A., Anguera, M. T., Maneiro, R., and Losada, J. L. (2019). Possession in football: more than a quantitative aspect–a mixed method study. *Frontiers in Psychology*, 10:501.

Castellano, J., Casamichana, D., and Lago, C. (2012). The use of match statistics that discriminate between successful and unsuccessful soccer teams. *Journal of Human Kinetics*, 31(2012):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.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

de Mendiburu, F. (2021). *agricolae: Statistical Procedures for Agricultural Research*. R package version 1.3-5.

Decroos, T., Bransen, L., Van Haaren, J., and Davis, J. (2019). Actions speak louder than goals: valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1851–1861.

Di Salvo, V., Baron, R., Tschan, H., Montero, F. C., Bachl, N., and Pigozzi, F. (2007). Performance characteristics according to playing position in elite soccer. *International Journal of Sports Medicine*, 28(3):222–227.

Edgington, E. and Onghena, P. (2007). *Randomization tests*. Chapman and Hall/CRC.

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., and Vikström, C. (2013). *Multi- and megavariate data analysis basic principles and applications*, volume 1. Umetrics Academy.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Ferrer, A. (2007). Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process. *Quality Engineering*, 19(4):311–325.

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage publications, third edition.

Gottfries, J., Blennow, K., Wallin, A., and Gottfries, C. (1995). Diagnosis of dementias using partial least squares discriminant analysis. *Dementia and Geriatric Cognitive Disorders*, 6(2):83–88.

Gregson, W., Drust, B., Atkinson, G., and Salvo, V. (2010). Match-to-match variability of high-speed activities in premier league soccer. *International Journal of Sports Medicine*, 31(4):237–242.

Johnston, R., Jones, K., and Manley, D. (2018). Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of british voting behaviour. *Quality & Quantity*, 52(4):1957–1976.

Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96(454):589–604.

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J., and Shmulevich, I. (2009). Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):i161–i168.

Kucheryavskiy, S. (2020). mdatools — R package for chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 198.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw Hill Irwin, New York. NY.

Lago, C. (2009). The influence of match location, quality of opposition, and match status on possession strategies in professional association football. *Journal of Sports Sciences*, 27(13):1463–1469.

Lago-Ballesteros, J. and Lago-Peñas, C. (2010). Performance in team sports: identifying the keys to success in soccer. *Journal of Human Kinetics*, 25(2010):85–91.

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

Lago-Peñas, C., Lago-Ballesteros, J., and Rey, E. (2011). Differences in performance indicators between winning and losing teams in the UEFA Champions League. *Journal of Human Kinetics*, 27(1):135–146.

Levene, H. (1961). Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pages 279–292.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3):18–22.

Liu, H., Gómez, M.-A., Gonçalves, B., and Sampaio, J. (2016). Technical performance and match-to-match variation in elite football teams. *Journal of Sports Sciences*, 34(6):509–518.

Liu, H., Gomez, M.-Á., Lago-Peñas, C., and Sampaio, J. (2015). Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup. *Journal of Sports Sciences*, 33(12):1205–1213.

Malagón-Selma, P., Debón, A., and Ferrer, A. (2022). Modelos de machine learning y estadística multivariante para predecir la posición de los equipos de primera división. *Journal of Sports Economics & Management*, 12(1):3–22.

Migliorati, M. (2020). Detecting drivers of basketball successful games: an exploratory study with machine learning algorithms. *Electronic Journal of Applied Statistical Analysis*, 13(2):454–473.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Noçairi, H., Gomes, C., Thomas, M., and Saporta, G. (2016). Improving stacking

methodology for combining classifiers; applications to cosmetic industry. *Electronic Journal of Applied Statistical Analysis*, 9(2):340–361.

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).

Paluszyńska, A. (2017). Understanding random forests with randomforestexplainer. *The Comprehensive R Archive Network*.

Paluszynska, A., Biecek, P., and Jiang, Y. (2020). *randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance*. R package version 0.10.1.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 11(1):68–84.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rampinini, E., Coutts, A. J., Castagna, C., Sassi, R., and Impellizzeri, F. (2007). Variation in top level soccer match performance. *International Journal of Sports Medicine*, 28(12):1018–1024.

Rampinini, E., Impellizzeri, F. M., Castagna, C., Coutts, A. J., and Wisløff, U. (2009). Technical performance during soccer matches of the Italian serie a league: effect of fatigue and competitive level. *Journal of Science and Medicine in Sport*, 12(1):227–233.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. In Liu, L. and Özsu, M. T., editors, *Encyclopedia of Database Systems*, pages 532–538. Springer US.

Rohart, F., Gautier, B., Singh, A., and Le Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11):e1005752.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Sheather, S. (2009). *A modern approach to regression with R*. Springer Science & Business Media.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):7881.

Smithies, T. D., Campbell, M. J., Ramsbottom, N., and Toth, A. J. (2021). A random forest approach to identify metrics that best predict match outcome and player ranking in the esport rocket league. *Scientific Reports*, 11(1):1–12.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133.

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

matics, 8(1):1–21.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

Vigne, G., Gaudino, C., Rogowski, I., Alloatti, G., and Hautier, C. (2010). Activity profile in elite Italian soccer team. *International Journal of Sports Medicine*, 31(05):304–310.

Welch, B. L. (1947). The generalization of "Students's" problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35.

Whitehead, S., Till, K., Jones, B., Beggs, C., Dalton-Barron, N., and Weaving, D. (2021). The use of technical-tactical and physical performance indicators to classify between levels of match-play in elite rugby league. *Science and Medicine in Football*, 5(2):121–127.

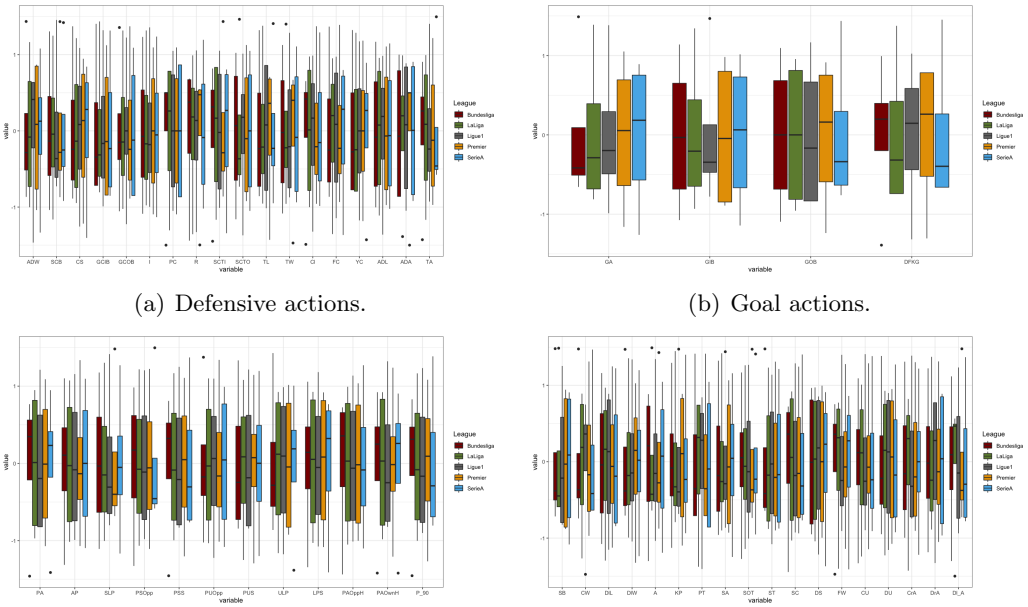
Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1(6):80–83.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Wold, S., Johansson, E., Cocchi, M., et al. (1993). PLS: partial least squares projections to latent structures. In *From 3D QSAR in Drug Design: Theory, Methods and Applications*, pages 523–550. Kubinyi H (eds.). ESCOM Science Publishers.

Worley, B. and Powers, R. (2013). Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1):92–107.

Appendix



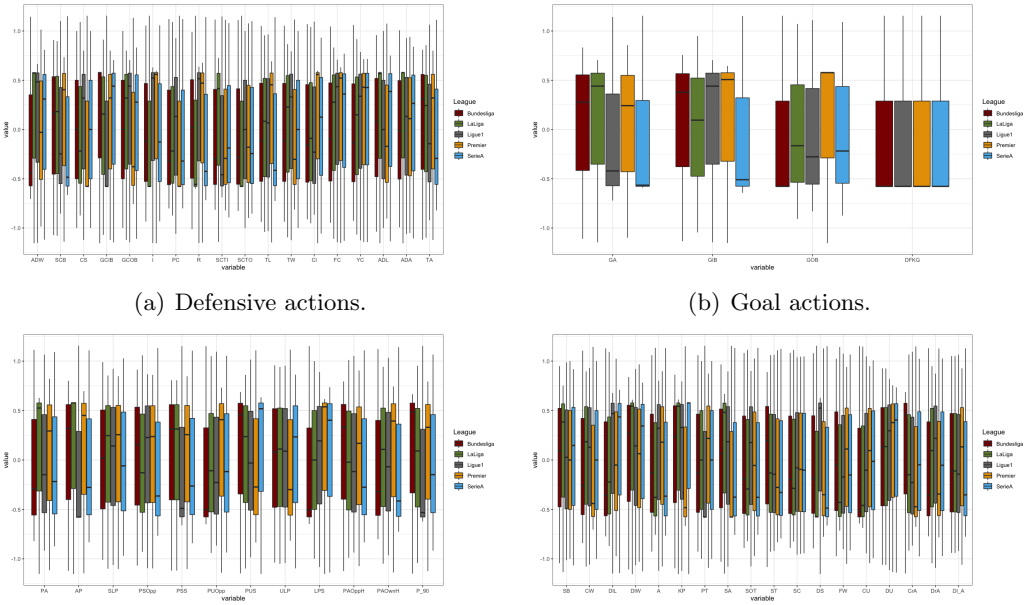
(a) Defensive actions.

(b) Goal actions.

(c) Passing and possession actions.

(d) Ofensive actions.

Figure A.1.: Boxplot with standardised values for the Top teams in each league.



(a) Defensive actions.

(b) Goal actions.

(c) Passing and possession actions.

(d) Ofensive actions.

Figure A.2.: Boxplot with standardised values for the Bottom teams in each league.

Table A.1.: Mean and standard deviation of the variables for the Top teams in the Big Five

	Premier		**Ligue 1**		**Bundesliga**		**Serie A**		**LaLiga**	
Variables	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
SCB	81.5	21.3	108.0	19.2	76.5	14.9	108.8	15.0	98.3	17.4
R	2333.8	75.3	2253.0	29.4	2166.3	111.9	2304.5	136.8	2273.0	38.3
CS	17.5	3.7	13.8	3.0	11.8	3.1	14.0	3.6	15.5	3.7
PC	3.0	1.8	7.5	2.4	5.8	0.5	4.5	0.6	5.8	2.9
I	338.0	21.9	380.5	41.6	373.5	34.5	381.8	61.6	410.8	39.6
SCTI	83.3	27.0	98.3	12.1	88.5	23.8	91.0	11.2	98.8	1.5
SCTO	34.3	12.2	45.8	4.2	33.5	7.3	39.5	2.4	44.0	9.6
TW	360.0	22.4	401.0	24.1	344.5	46.6	379.8	26.4	403.3	59.8
TL	232.5	29.0	259.5	6.4	206.0	40.8	219.5	15.3	233.8	45.8
YC	46.5	8.0	61.5	7.2	47.8	10.0	69.0	5.6	83.0	10.1
Cl	636.3	97.4	628.3	106.0	569.3	151.2	600.5	22.8	660.0	157.8
FC	339.5	26.2	462.0	17.5	359.0	77.5	448.5	28.2	448.8	28.6
TA	60.9	1.8	60.7	0.9	62.7	1.6	63.4	2.2	63.5	1.2
ADW	574.5	40.9	567.5	101.9	583.8	125.5	536.8	96.4	567.8	107.5
ADL	565.3	48.5	538.5	137.7	535.8	81.4	533.0	72.3	522.8	127.6
ADA	50.4	0.6	51.7	3.1	52.0	2.8	50.0	3.6	52.3	2.4
GCIB	27.0	7.1	35.0	6.1	32.8	9.4	31.5	6.5	33.8	7.1
GCOB	3.8	3.1	4.0	0.8	6.5	3.3	4.8	2.1	2.8	1.7
KP	400.8	49.9	364.5	51.9	341.0	54.1	463.3	33.4	356.5	48.7
PT	5.0	1.4	10.0	3.6	5.0	1.4	6.3	2.6	7.3	2.4
SOT	213.0	10.9	214.8	32.5	200.5	38.1	269.8	20.7	205.8	20.8
SB	169.0	31.8	129.0	27.9	119.5	10.5	163.3	8.5	114.5	21.2
CW	240.5	46.9	213.8	33.7	209.5	53.2	261.5	27.6	211.5	10.7
CU	395.0	41.9	344.3	61.6	311.3	69.4	498.8	85.5	335.3	109.8
SA	50.5	2.6	49.8	0.6	50.6	3.2	44.3	1.1	48.8	5.0
A	56.5	10.8	53.5	13.0	53.5	9.6	48.3	10.3	43.5	11.7
ST	218.5	31.7	213.5	33.8	204.8	35.4	214.5	17.7	198.8	44.0
SC	93.8	14.7	90.3	20.8	85.0	19.4	144.0	37.8	86.3	31.3
DS	403.0	46.2	398.5	72.0	375.5	66.8	340.3	40.4	404.0	57.6
FW	359.0	50.9	469.8	29.5	375.8	22.9	468.0	38.3	515.5	44.6
DU	278.0	24.2	280.0	23.4	257.3	28.7	254.8	30.0	263.5	46.9
CrA	19.3	3.9	20.8	3.4	21.4	1.4	22.2	1.6	20.4	1.7
DrA	59.1	3.8	58.5	2.5	59.2	2.7	57.2	1.9	60.6	2.5
Dl_A	50.7	1.0	50.9	1.5	51.7	1.4	51.0	1.4	52.4	1.0
DlL	1862.0	56.9	2002.3	104.4	1751.0	192.1	1846.5	147.5	1914.0	122.8
DlW	1918.8	92.8	2075.8	76.4	1868.8	161.3	1925.3	170.9	2106.5	127.4
GA	18.1	2.3	17.7	3.9	18.8	3.5	14.4	1.6	15.9	2.9
GIB	67.8	16.5	67.3	19.6	66.3	7.7	60.0	7.9	55.8	15.8
GOB	10.8	4.6	8.3	1.5	9.0	3.7	9.5	5.9	9.0	5.2
DFKG	1.3	1.0	2.3	1.7	1.8	1.3	1.3	1.9	2.8	2.4
PA	86.1	2.7	84.3	3.8	83.3	5.5	86.0	1.4	84.0	4.5
AP	63.0	3.7	57.0	6.1	58.3	7.0	58.0	1.8	54.8	8.6
PUOpp	3037.3	249.9	2757.8	202.6	2826.3	366.6	2833.0	155.4	2789.8	261.2
PSOpp	11654.5	2326.9	8591.3	2429.7	8671.8	1391.9	10027.5	964.1	9531.8	2649.3
SLP	1116.0	33.8	996.3	149.0	1070.5	188.3	1263.5	135.0	1258.3	212.5
PUS	2414.5	155.0	2086.5	86.0	2291.8	274.7	2117.8	179.1	2182.3	144.2
PSS	19705.3	2574.2	15597.0	3609.1	15419.0	3341.0	16941.3	1371.6	15913.5	4387.3
ULP	909.5	247.8	905.8	345.9	858.3	154.6	845.3	62.4	923.5	248.4
P_90	635.4	59.2	515.4	85.8	577.7	90.7	557.1	35.1	533.7	113.7
PAOppH	81.8	4.3	77.9	6.0	77.9	5.6	81.9	1.8	79.3	5.9
PAOwnH	91.7	1.6	91.6	2.1	90.3	4.0	91.3	1.2	90.4	2.5
LPS	55.7	6.4	53.7	8.5	55.4	8.2	59.8	3.5	57.9	10.1

Table A.2.: Mean and standard deviation of the variables for the Bottom teams in the Big Five

	Premier		**Ligue 1**		**Bundesliga**		**Serie A**		**LaLiga**	
Variables	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
SCB	140.3	9.1	123.7	14.8	120.3	4.0	138.3	11.0	102.7	23.8
R	2319.3	48.2	2256.0	33.0	1969.3	74.5	2188.7	27.4	2278.3	24.0
CS	6.7	2.9	8.3	2.1	5.0	1.0	7.0	2.0	7.3	1.5
PC	7.3	0.6	5.7	2.5	4.7	2.1	5.7	2.1	9.3	1.5
I	497.7	56.0	463.7	37.0	403.3	52.8	393.7	68.4	434.0	1.7
SCTI	163.0	17.1	123.0	13.1	144.7	15.5	152.3	7.1	139.7	3.2
SCTO	48.0	5.6	62.0	1.0	54.7	12.9	71.0	8.2	52.0	13.9
TW	387.7	35.2	425.0	15.1	351.0	8.5	352.0	23.0	412.3	42.3
TL	265.7	29.4	264.3	10.0	218.3	20.6	229.7	16.1	245.7	16.0
YC	63.0	7.0	68.0	8.9	59.3	2.5	91.7	17.2	94.7	15.6
Cl	869.0	182.8	788.0	91.3	798.3	39.9	835.7	90.5	863.7	18.6
FC	405.7	27.5	539.3	47.5	396.3	23.6	535.3	68.4	561.0	21.6
TA	59.4	1.7	61.7	1.2	61.7	2.6	60.5	1.1	62.6	2.6
ADW	864.3	115.0	779.7	44.1	673.7	60.6	609.7	146.4	700.7	127.9
ADL	846.7	126.9	858.0	28.0	713.0	25.1	685.7	109.8	734.7	106.2
ADA	50.5	0.6	47.6	0.8	48.5	2.5	46.8	2.3	48.7	1.0
GCIB	64.3	8.3	49.3	4.0	60.7	2.3	58.3	3.8	55.3	10.6
GCOB	11.0	2.6	11.3	3.8	9.0	1.0	13.0	3.6	7.3	2.1
KP	290.7	32.5	265.7	55.8	256.7	27.1	289.7	14.4	293.0	29.5
PT	2.7	1.5	4.7	1.2	3.3	2.5	6.0	1.0	6.0	1.0
SOT	179.7	12.0	176.7	30.4	171.3	17.2	190.0	21.2	192.0	17.1
SB	114.0	2.0	106.7	12.5	99.7	3.5	112.3	18.1	101.7	24.4
CW	164.7	3.8	162.7	33.7	154.7	13.9	161.0	28.0	164.3	25.3
CU	367.0	94.8	392.7	64.3	303.3	83.0	433.7	48.0	450.0	8.7
SA	41.9	1.1	40.1	1.3	39.9	2.1	41.1	5.2	43.7	2.1
A	19.0	5.6	19.3	2.1	20.7	5.0	24.7	10.0	25.0	2.6
ST	130.0	14.4	118.3	23.2	113.3	8.6	132.3	19.3	148.3	2.5
SC	90.3	14.0	92.0	12.5	83.0	36.0	109.7	35.6	118.7	9.3
DS	283.3	63.5	323.7	48.3	256.7	19.8	249.0	73.9	388.7	126.7
FW	367.3	24.1	469.0	28.8	400.0	19.5	484.3	48.4	474.7	45.8
DU	215.0	2.6	246.7	24.8	208.7	47.8	200.7	30.4	270.7	46.8
CrA	20.0	2.0	19.1	2.2	21.0	3.5	19.9	3.5	20.8	1.6
DrA	56.4	5.0	56.6	6.0	55.6	4.1	54.9	5.0	58.3	4.9
Dl_A	49.7	0.8	48.9	1.2	49.6	2.4	48.1	0.5	49.1	0.6
DlL	2181.0	58.6	2328.3	48.2	1916.7	178.2	2050.7	168.4	2278.0	54.0
DlW	2155.3	121.7	2224.3	60.5	1879.7	68.4	1897.7	123.1	2195.7	15.3
GA	9.7	1.8	10.1	1.7	10.4	0.4	10.8	4.1	11.8	0.4
GIB	25.3	7.2	25.3	3.8	25.0	2.6	31.7	15.0	34.7	3.5
GOB	4.7	0.6	4.0	3.6	4.7	1.2	3.3	1.5	5.7	4.0
DFKG	0.3	0.6	0.3	0.6	0.3	0.6	0.3	0.6	1.0	0.0
PA	73.3	8.5	76.6	1.6	76.2	0.6	78.7	2.6	77.1	2.9
AP	43.3	8.1	45.3	2.3	44.3	2.1	44.0	3.6	46.7	2.3
PUOpp	3273.3	261.8	3043.3	167.8	2799.0	201.5	2834.3	113.6	2955.0	239.1
PSOpp	5805.3	1427.8	5373.7	599.2	4517.7	433.9	5744.3	1187.1	5287.7	160.0
SLP	1054.7	165.6	1184.0	127.0	980.0	49.5	1079.7	112.1	1103.0	60.9
PUS	2221.0	197.6	2060.3	140.6	1962.7	87.8	1999.3	68.8	1967.0	126.7
PSS	9876.0	3824.4	9784.3	880.2	9187.7	330.4	10829.7	1548.3	9980.3	1259.5
ULP	1456.0	166.7	1274.7	48.6	1214.7	54.6	1197.7	142.9	1306.3	78.4
P_90	384.4	102.6	376.4	23.5	392.5	13.2	397.5	39.3	377.8	30.6
PAOppH	66.3	7.7	67.4	1.7	64.9	1.7	70.5	4.8	68.4	3.0
PAOwnH	83.7	6.0	87.4	1.5	87.6	0.6	87.3	1.0	86.6	1.6
LPS	42.0	6.3	48.1	1.7	44.7	2.1	47.5	4.5	45.8	1.0