INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Statistical analysis of sports data

Bichet Camille
Monédières Emmeline

5GMM - Département Génie Mathématiques et Modélisation

Under the supervision of
Sébastien Déjean
Philippe Saint Pierre
Javier López Sánchez

January 2020

Contents

Abstract

Keywords :

Acknowledgements

Introduction

- présentation du problème : est ce que les études stat peuvent s'appliquer au sport ? sur quels critères recruter un joueur ? Vaut il mieux quelques bons joueurs ou que des joueurs moyens ? un joueur donné a t il sa place dans une équipe particulière ?
- on a récupéré nos données sur whoscored
- plan

Can mathematics help your favorite football team to win ? What if, by using statistics, you were able to identify performance indicators for a team ? These are real questions and by using machine learning methods, it is possible to answer them.
The point of our project is to study what kind of actions can lead to a good ranking in a tournament. To do so, we created a data set based on data from the website *whoscored.com*.
We divided the study into three different parts. First we had an exploratory analysis to discover our data, find some interactions and have a first idea what the main variables are. Then we used variable selection methods to keep a reasonable number of variables and we applied regression to them. Finally we tried to implement machine learning algorithms such as Random Forest, Support Vector Machine or neural network methods.

1 Data presentation

1.1 The website : *WhoScored.com*

The website *WhoScored.com* is managed by a team of football analysts and software developers from London. It gathers football statistics, mainly from European divisions, with a lot of different variables. More than 500 tournaments, 15,000 teams and 250,000 players teams are represented. For some tournaments, detailed statistics are available.

When detailed statistics are available, we have around 80 variables for each tournament. These variables are the averages over a game.

1.2 The variables

Here we only study the team statistics (not player or referee statistics). These date are split into three main categories : `offensive`, `defensive` and `passing`.

1.2.1 `Offensive` statistics

`Offensive` data refer to actions from the players when they are in an offensive strategy. There are five `offensive` data : shots, goals, dribbles, possession loss and aerial duels. For the aerial duels and the dribbles, we only know if the action was successful or unsuccessful. Shots and goals data are more detailed. We have statistics on the zones of the football field where shots and goals have been tempted. We can only find the type of situation of these shots and goals, the accuracy of these actions and the players body parts with which they have been performed.

1.2.2 `Defensive` statistics

`Defensive data` refer to actions from the players when they are in a defensive strategy. They do not control the game anymore. There are eight `defensive` data : tackles, interception, fouls, cards, off-sides, clearances, blocks and saves. They are less detailed than `offensive` data.

1.2.3 `Passing` statistics

`Passing` data refer to all the passing actions between players during the game. They are divided into three types of passing : passes, key-passes and assists.
Assists are passes that always lead to a goal, just after the pass.
Key passes are important passes which can lead to assists and then to a goal.
The third type of passes, just called `passes` are all the other kind of passes during the game.

1.3 Our data

At first, we tried to work only with the data from the French tournament, the "*Ligue 1*". We only had 20 teams to study, which meant 20 individuals only. It was not enough to do statistics on it, in front of the large number of explanatory variables. Two options were available :

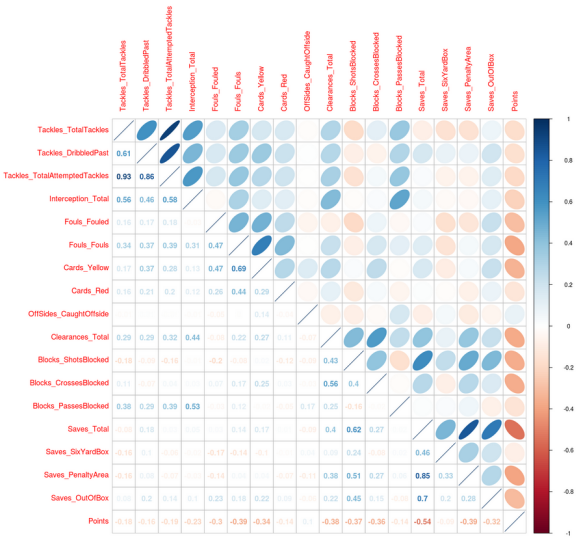
- Gather data from several last seasons, which was not a good idea because most of the teams keep a similar composition and a similar playing style over the years. Using data from several years would have made individuals dependant.
- Use data from different tournaments

We chose to use data from different tournaments, from six different countries : `France`, `Spain`, `Germany`, `Italy`, `England` and `Argentina`. Each tournament is from the 2018-2019 season. By this way, the data set is composed of 124 independent teams and 86 variables.

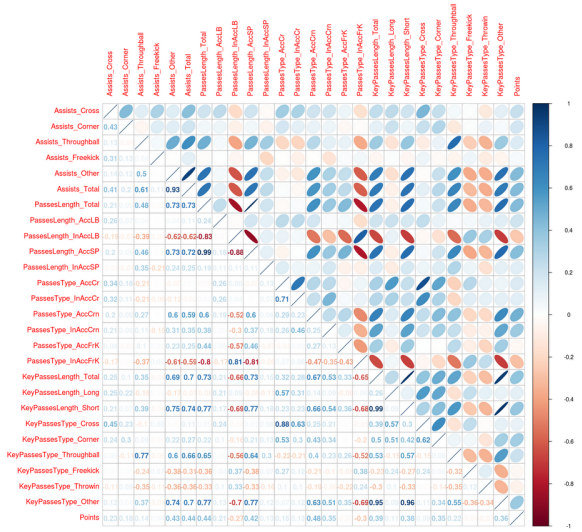
2 Exploratory analysis

2.1 Analysis of the correlations between variables

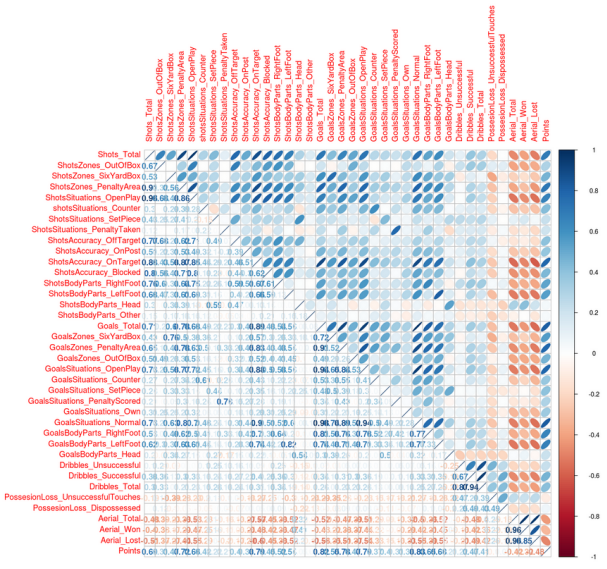
The aim of this section is to carry out a qualitative analysis of the data. We use the `corrplot` package in order to display different correlation matrices. Indices of the correlation matrices are computed on each data set (`offensive`, `defensive` and `passing`) for more readability.



(a) `Defensive` data



(b) `Passing` data



(c) `Offensive` data

Figure 1: Correlation matrices of `offensive`, `defensive` and `passing` data

We would like to identify some correlated variables and particularly which variables are correlated with the number of points at the end of a tournament.
On the Figure 1, we represent on a same graph the real correlation coefficient between each variable and an ellipse standing for this coefficient. A red ellipse means that two variables are negatively correlated, whereas a blue one means that they are positively correlated. The finer the ellipses, the more correlated the variables are.

On Figure 1a, we notice some highly correlated variables. This is the case for `TotalTackles` and `TotalAttemptedTackles`, because `TotalTackles` is included in `TotalAttemptedTackles`. It works as well for `SavesTotal` and `SavesPenaltyArea`, with the same explanation.
About `Rating`, our interest variable, we do not really have any very positively correlated variable, but `SavesTotal` and `ClearanceTotal` are negatively correlated.

Several variables are correlated negatively, as we can see on Figure 1b. Those variables are `PassesLength_AccSP` (Accurate Short Passes) and `PassesLength_InAccLB` (Inaccurate Long Balls), `PassesLength_InAccLB` (Inaccurate Long Balls) and `PassesLength_Total`, `PassesType_InAccFrK` (Inaccurate Freekicks) and `PassesLength_AccSP` (Accurate Short Passes). Each variable is the "opposite" of the other, which can explain why they are negatively correlated. For example, when a team spends the game doing short passes, it does not do long passes.
There are positively correlated variables as well. We can cite `KeyPasssesType_Other` with `KeyPassesLength_Total` and `KeyPassesLength_Short`, or `Assists_Other` with `Assists_Total`.

The variable `Rating` is positively correlated with `Assits_Other`, `Assits_Total`, `PassesLength_Total`, `KeyPassesLength_Total`, `KeyPassesLength_short` and `KeyPassesType_Other`. Every `Assist` leads to a goal, so it makes sense to be correlated. `Key Passes` are as well important to score, so they are positively correlated.

Finally, the last figure, Figure 1c represents the correlation matrix of `offensive` data. As for the other figures, some variables are positively correlated. `Aerial_Total`, `Aerial_Loss` and `Aerial_Win` are really correlated, they must carry the same information. The same phenomenon is observed for `Dribbles_Successful`, `Dribbles_Unsuccessful` and `Dribbles_Total`. Some other variables are positively correlated, they have skinny blue ellipses. There are only a few negative correlations, and they are low.
`Rating` is positively correlated with `Shots_Total`, `ShotsAccuracy_OnTarget`, `Goals_Total`, `GoalsZones_PenaltyArea`, `GoalsSituations_OpenPlay` and `GoalsSituations_Normal`. It makes sense, and it is quite trivial. Having a high average value in shots or goals, means that the team had the ball and had goals opportunities, which is helpful to win.

2.2 Principal Component Analysis

Another way to analyse our data is to do a Principal Component Analysis (PCA) to obtain a display of individuals and variables on a same plane. We chose this method because it provides a better representation of individuals and variables by maximizing the variance between variables.
This linear dimensionality reduction method converts correlated variables into a basis of linearly uncorrelated variables, called principal components.

Variables and individuals are projected on principal planes, as we can see on the Figure 2. Such graphs make it possible to see links between variables. For example, two variables which are close on a principal plane are strongly and positively correlated, in regard to the two main variables that carry this plan. In contrast, two variables which are radially opposed are strongly but negatively correlated.

On a factorial map, we can also see the links between individuals and variables by using a biplot display like the Figure 2. Individuals which are close to an arrow tip have a high value of the variable associated to this arrow.

On Figure 2, we have computed the PCA on the entire data set but `Rating` is a quantitative supplementary variable : it is displayed on the factorial map but it has not been taken into account to compute the PCA. That is why the arrow is a red dotted line. Even if `Rating` is a supplementary variable, it is highly correlated with the first dimension : it is almost horizontal. We can say that the first principal axis stands for the `Rating` and hence for the teams rank.

There are two types of variables that are opposite along the first axis. Variables related to aerial duels, fouls, tackles and ball losses (among others) are displayed on the left, whereas variables related to pass length, shots and goals are displayed on the right. We notice that these last variables, correlated to `Rating` according to the PCA, are logically influential on the `Rating` of a team. To have a good rating, a team needs to score goals and construct its game, especially thanks to the numerous and short passes. In contrast fouls, cards and inaccurate actions will lead to a low rating.

Then, teams have been plotted by color according to their countries. We can see that teams of each country, excepting Argentina, are spread all along the horizontal axis. On the right of the map, we find the best

team (or at least the second best team) of each tournament (Bayern Munich, Manchester City, Napoli, Paris Saint-Germain and Barcelona) excepting the Argentina. On the left of the graphic, the less performing teams, excepting for Argentina, are plotted. The center of the biplot gathers teams in the middle of the ranking.
We notice that all the Argentine teams are displayed on the left of this graph, maybe because they do not play the same football as in Europe. It can be explained by the fact that it is the only non European team in our data set.
At the sight of this graph, best teams of the tournaments can be considered as outliers. Eibar and Parma Calcio 1913 are also outliers from the "middle" teams.

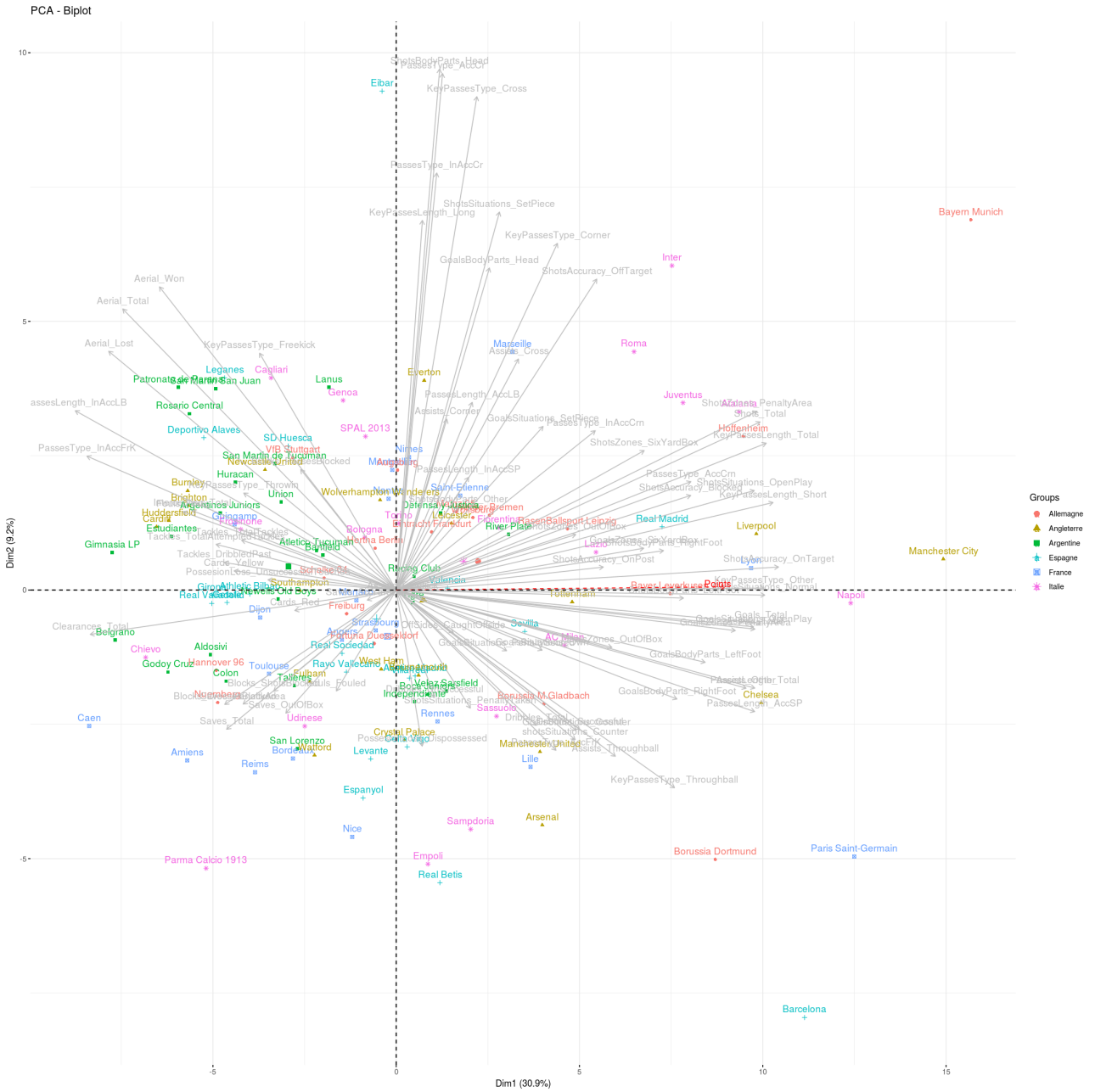


Figure 2: PCA biplot

2.3 Classification

In this part, we will try to find similarities between individuals or variables. This way, we can determine if some variables or individuals are close according to a specific measure, and then understand the similarities in a specific class. To do so, we use Ascending Hierarchical Classification (AHC). This method consists on gathering the close clusters in order to, at the end, have only one cluster containing all the objects. It is quite easy to interpret and very visual method.

2.3.1 Individuals classification

To create individual clusters, we started from the results of the PCA. Thanks to the PCA, the dimensions are already reduced, and then, it is easier to classify the football clubs. To create clusters, we need to know how many clusters we need. To do so, we use Figure 3, where we choose 5 as the optimal number of clusters.

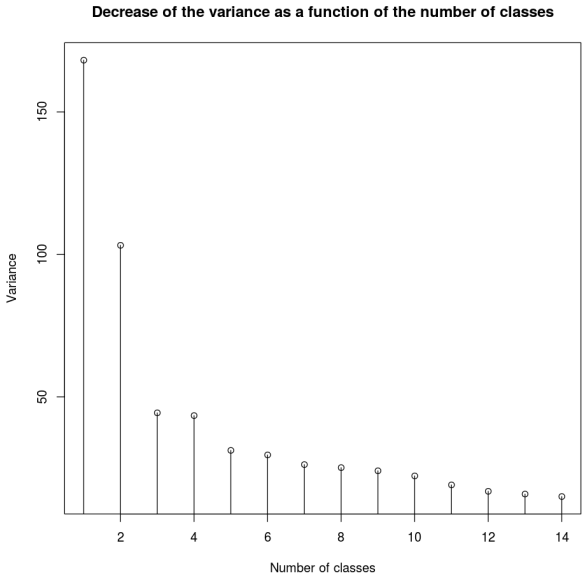


Figure 3: Decrease of the variance

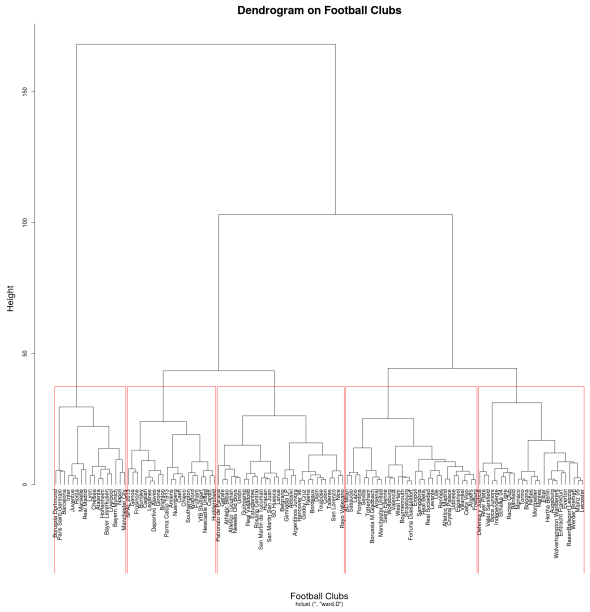


Figure 4: Dendrogram on individuals

Knowing the optimal number of clusters, we are able to divide individuals in 5 clusters, as it is shown in Figure 4. Football clubs seem to be ranked according to their performances. On the left of the dendrogram are the best clubs and on the right the weaker. To have a more precise idea on the specificity of each class, we plot the PCA biplot where we add the clusters. It is Figure 5.

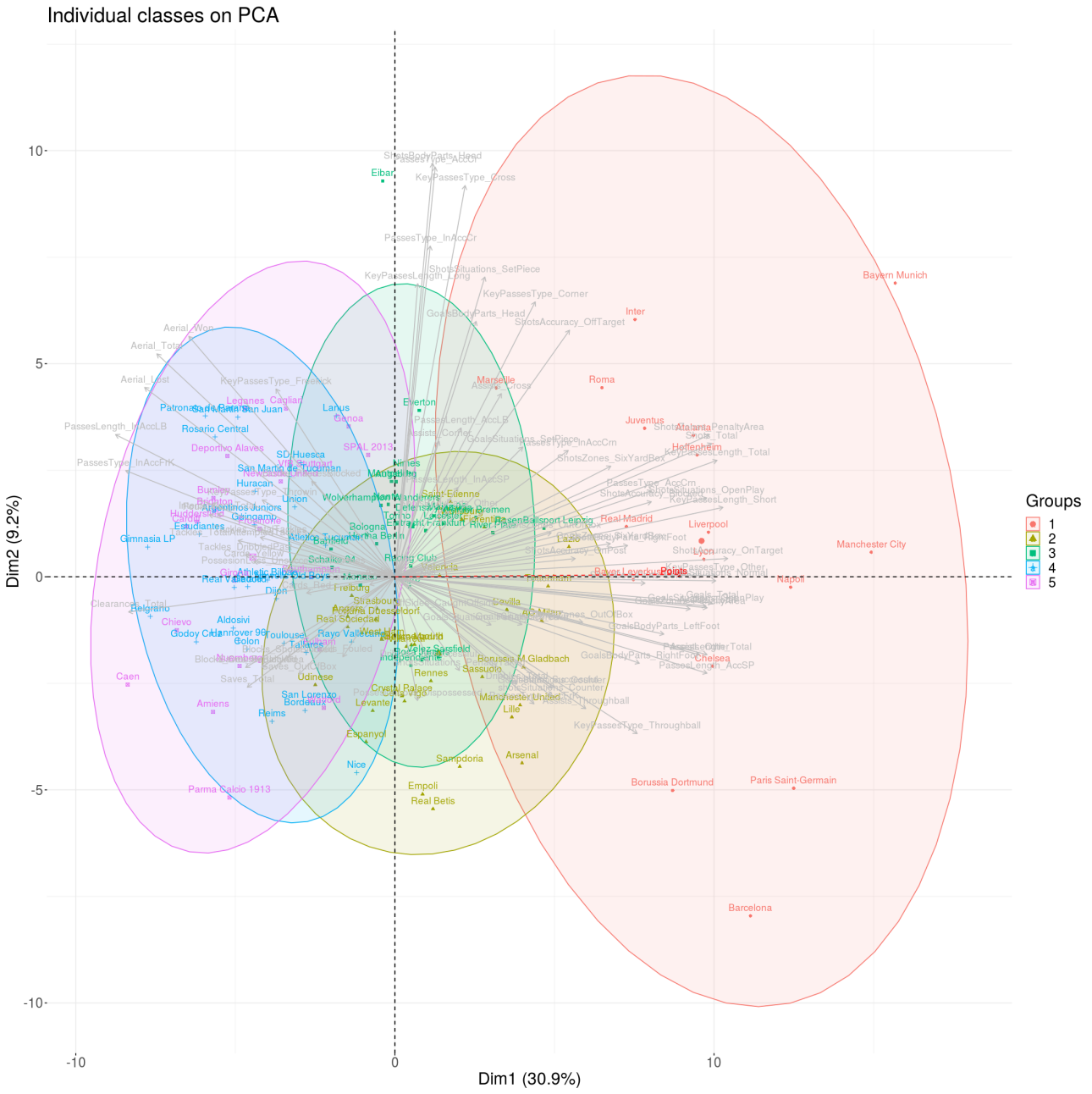


Figure 5: Individuals clusters PCA biplot

Since the first dimension represents the rating of a team, as mentioned in Part 2.2, we can clearly visualize the first cluster, in red, composed of the best teams. This cluster is the most spread out because each team seems to have its characteristics. The separation is obvious for the top of the ranking, as evidenced by the PCA biplot and the dendrogram (Figure 4) because this cluster is linked to the others only at the end. The other clusters are more or less mixed up. The divisions are fuzzy.

2.3.2 Variables classification

To create a dendrogram on variables, we compute the Spearman correlation coefficients between each variables. Then the dendrogram is build using the Ward distance method applied on the correlation coefficients.

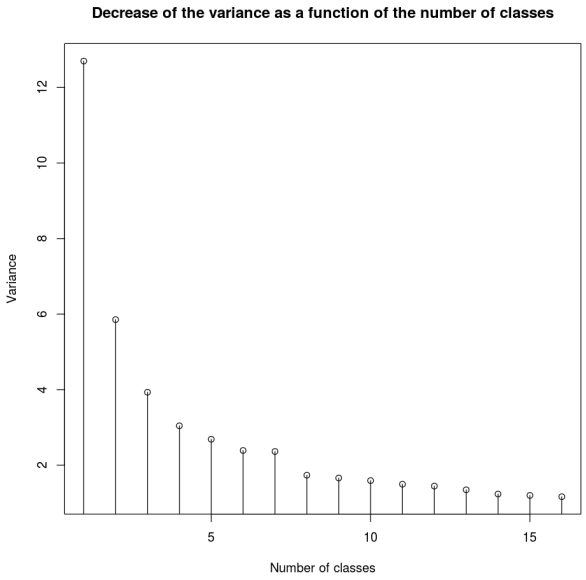


Figure 6: Decrease of the variance

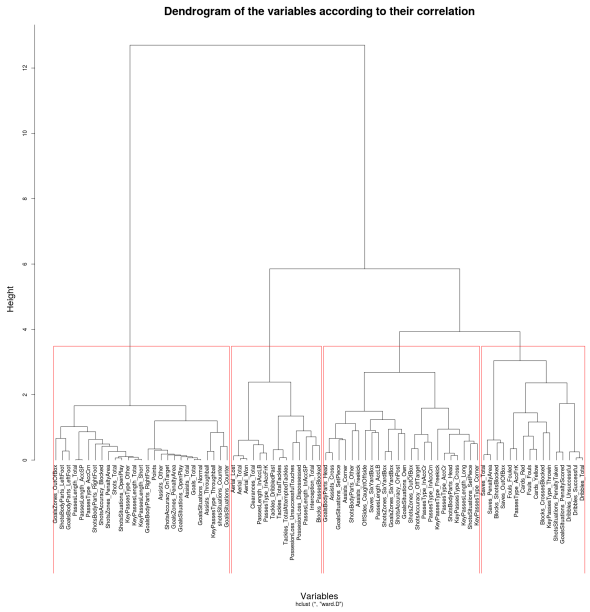


Figure 7: Dendrogram on variables

Regarding Figure 6, we can choose to split the variables into 3, 4 or 8 clusters. Since we want to obtain a precise gathering of variables, we choose to keep 8 clusters, as shown on the Figure 7.
We can see a clustering gathering aerial duels and long passes (clearances for example). Another clustering gathers actions happening after fouls like corners and free kicks. Another clustering is relatively different from the others because it is separated very early on the dendrogram. It is related to actions that are important to score goals and hence to goals themselves. This is the first rectangle on the left. The other clusters are less easy to analyse.

On the Figure 8 we see the 8 clusters of variables perfectly separated. This is therefore the PCA main interest, namely to separate the variables as much as possible in order to have the greatest possible variance, while reducing the size of individuals.

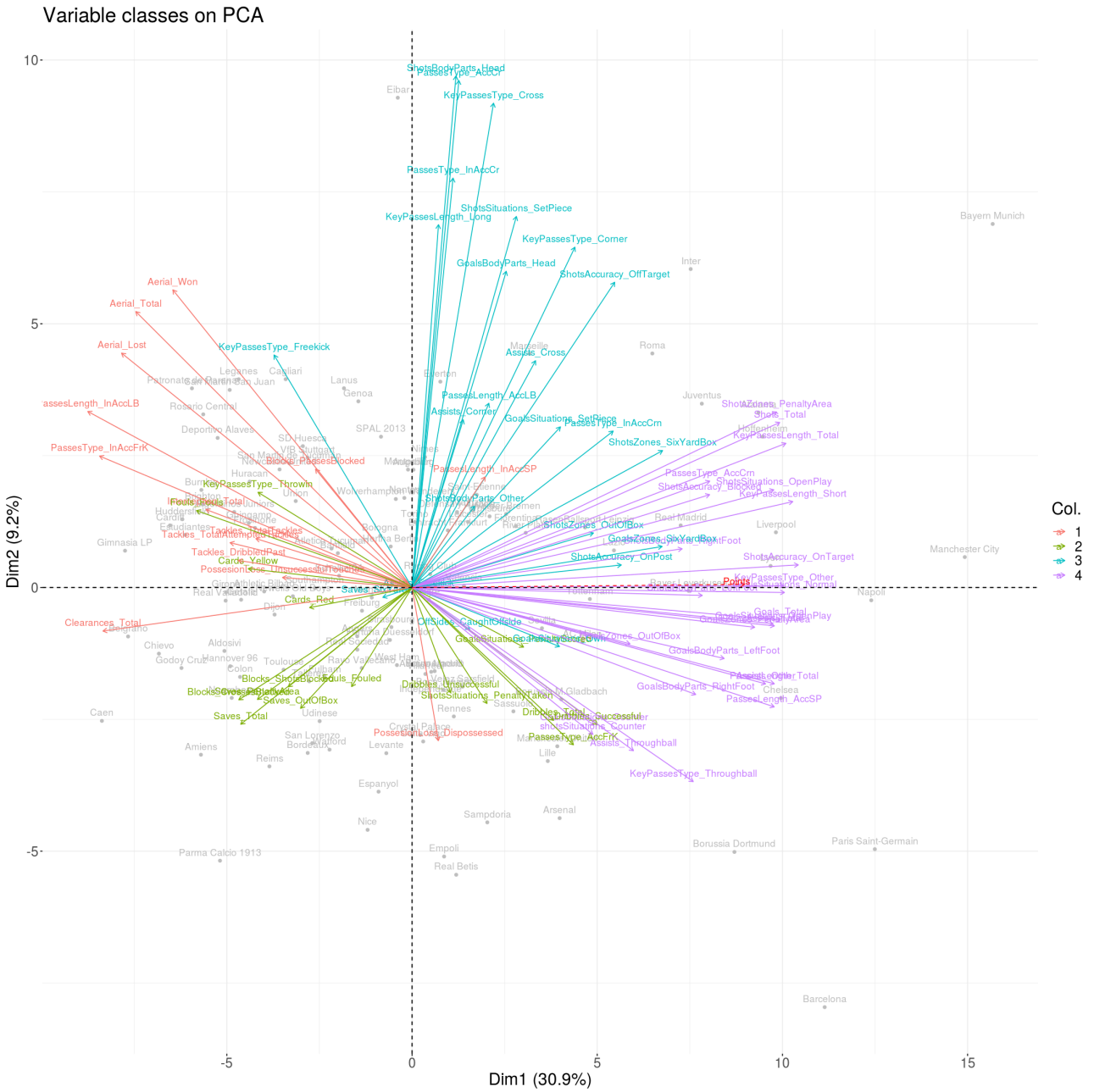


Figure 8: Variables clusters PCA biplot

2.3.3 Total classification

It is possible to visualize clusters on variables and individuals. To do so, we can use a `heatmap`. On Figure 9, each cell stands for the contribution of an individual to a variable. The whiter a cell, the more positively influenced by this variable the individual is. For instance, we an see a really white cell between `ShotsBodyParts_LeftFoot` and `Barcelona`. It makes sense because Messi plays in Barcelona, and he is left-footed.
On the bottom left corner are the whiter cells. It corresponds to good ranked clubs, hence, these variables lead to a good ranking. Next to those variables, there are dark blue cells. The corresponding variables are negatively influential on the ranking. There are aerial duels and inaccurate actions.

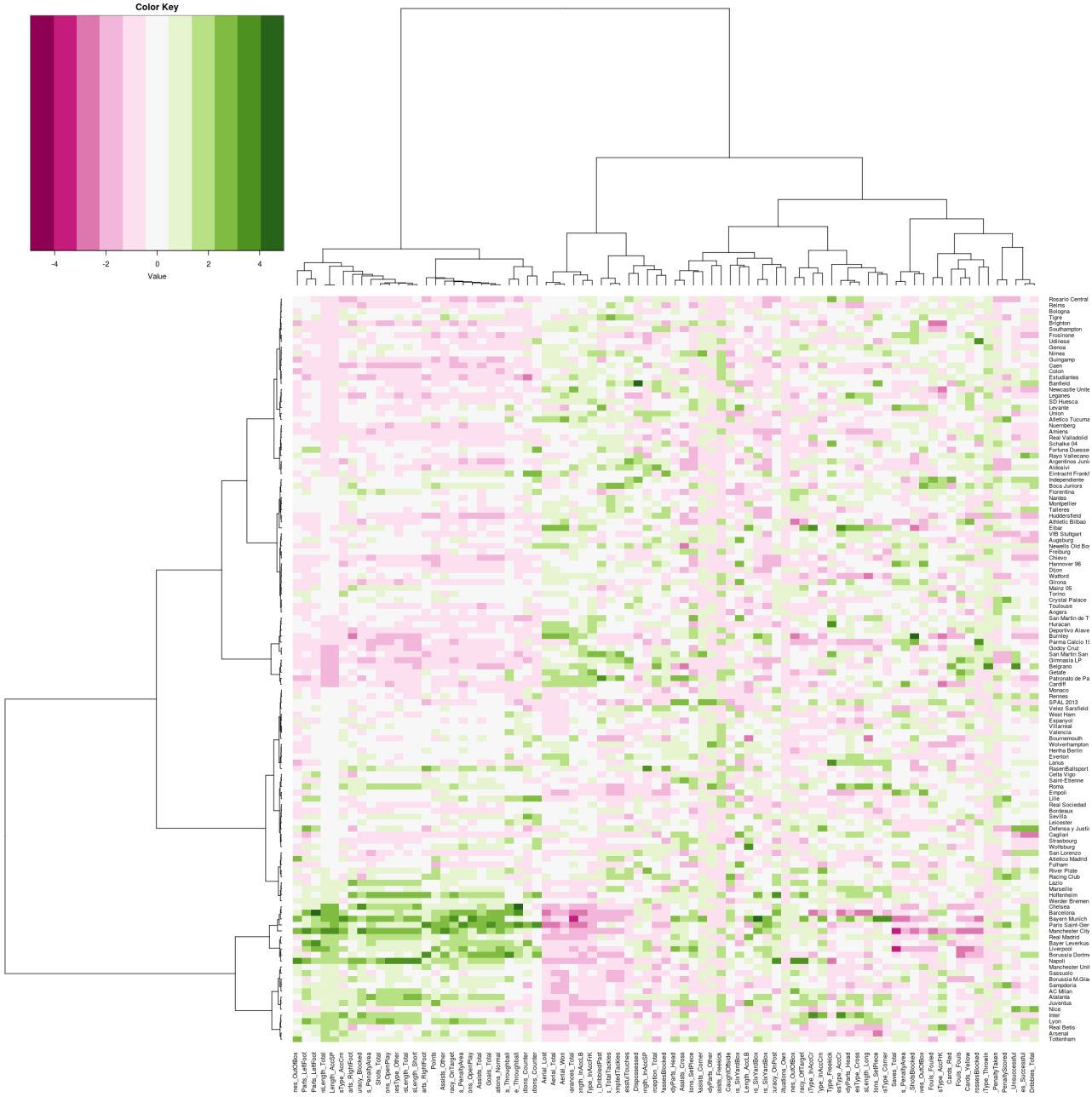


Figure 9: Heatmap on the two dendrograms

3 Variable selection and regression

3.1 Bayesian Information Criterion

The Bayesian Information Criterion, also named *BIC*, is a criterion used to select a model among a finite set of models. The finite set of models is composed of the complete model and every possible combination with different numbers of variables. A good model minimizes the criterion.
The BIC is defined as

$$BIC = ln(x)k - 2ln(\hat{L})$$

where

- $\hat{L}$ is the maximized value of the likelihood function;
- x is the observed data;
- n is the number of individuals;
- k is the number of parameters estimated by the model.

To select variables, we worked stepwise. At each iteration, the models tries to add a variable to the model and checks if one variable can be removed.

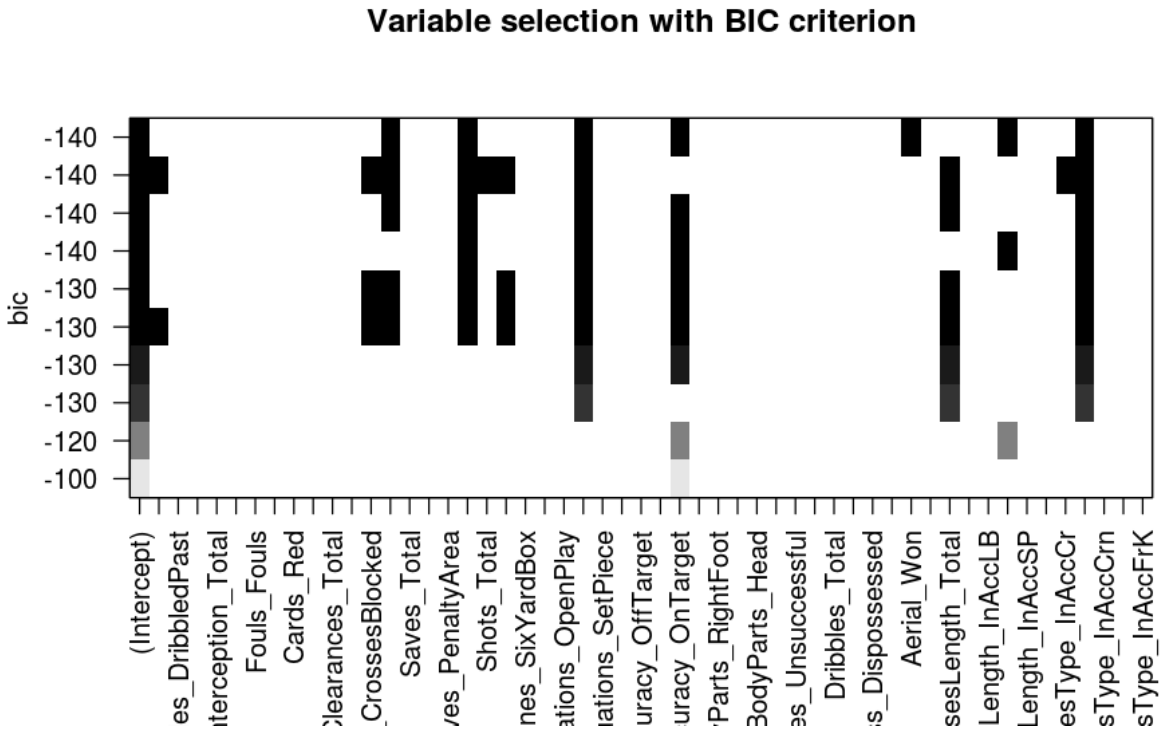


Figure 10: BIC Criterion

The variables selected are those allowing the lower BIC. The best model according to this criterion includes the black variables at the top. There are 7 variables left.

3.2 Lasso and ElasticNet

Our data set is composed of a number of explanatory variables (actions) really greater than the number of individuals (football teams). A classical regression will lead to estimators with high variance.
A solution can be to use regularization methods, by introducing penalization criterion in the optimisation problem. Let us consider a linear model with n observations :

$$Y_i = \theta_0 + \theta_1 X_i^1 + \theta_2 X_i^2 + ... + \theta_p X_i^p + \epsilon_i$$

where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$. We assume that the variables ϵ_i are independent and identically distributed, centered and with a variance σ^2.
We can rewrite this equation using matrix $\mathbf{X}$ with general term X_i^j (we set $X_i^0 = 1 \forall i$) and the vectors θ and $\mathbf{Y}$. We want to minimize $\|\mathbf{Y} - \mathbf{X}\theta\|^2$. The classical least squares estimator, in the case of the matrix $\mathbf{X}$ is of full rank, is :

$$\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

In our case, since the number of explanatory variables is greater than the number of individuals, the matrix $\mathbf{X}$ is not of full rank. That is why we introduce the LASSO regression, where the optimisation problem can be written as :

$$\min_{\theta} L(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2 + \lambda \sum_{i=1}^{p} |\theta_i| \tag{1}$$

With this method we obtain sparse solutions, which means that several coefficients θ_i are set to 0. In other words, some variables are not used to explain the target variable Y. That is why we can say that LASSO regression is a variable selection method.
The parameter λ in equation 1 has to be chosen by us. To help us find the better value of λ, we used a cross-validation method.

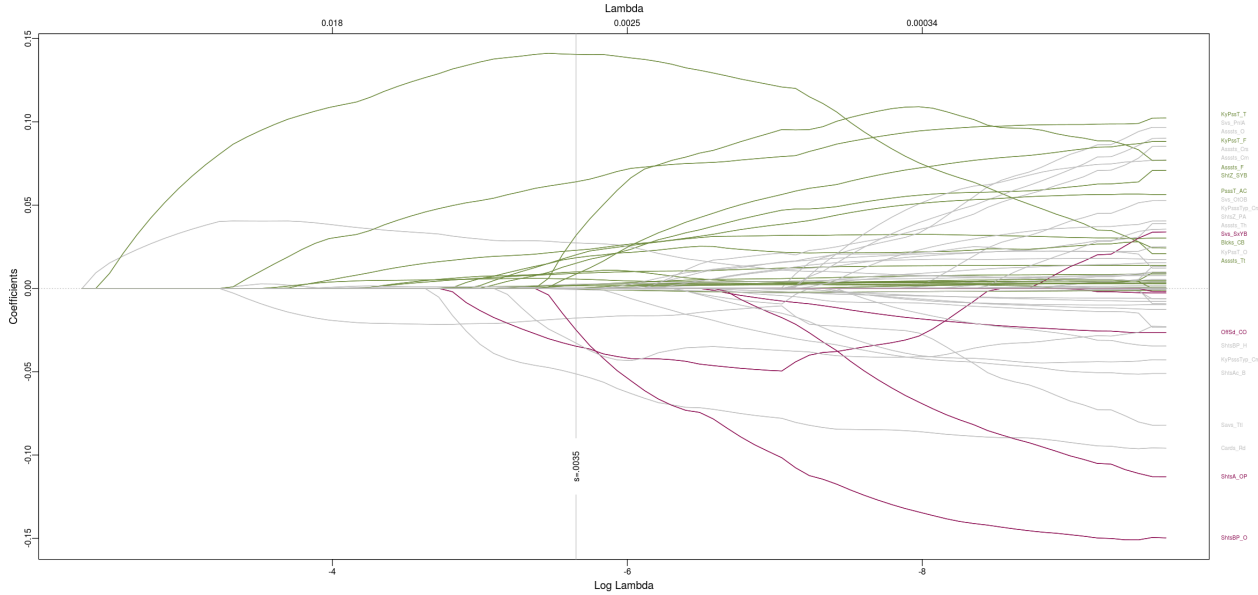


Figure 11: Lasso regression on the data set

Another regularization method in regression, called Ridge regression consists in minimizing :

$$\min_{\theta} L(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2 + \lambda \sum_{i=1}^{p} |\theta_i|^2 \tag{2}$$

This methods do not lead to sparse solutions but it prevents parameters θ_i from taking high values. This is not a variable selection method.

LASSO and Ridge regressions combined form the ElasticNet method, which can be written :

$$\min_{\theta} L(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2 + \lambda \left(\alpha \sum_{i=1}^{p} |\theta_i| + (1-\alpha) \sum_{i=1}^{p} |\theta_i|^2 \right) \tag{3}$$

ElasticNet method is a good compromise between sparse solutions and solutions with enough non-zero parameters. There is not as much null coefficients as in the LASSO regression, but we keep a reasonable number of explanatory variables.

3.3 Results comparison

Using different methods, we obtained different results. The best method seems to be ElasticNet. It keeps around 20 variables, which is a good number. If we want to reduce the dimension, we can remove the variables with a coefficient close to 0, but keeping all of them allows a larger study.

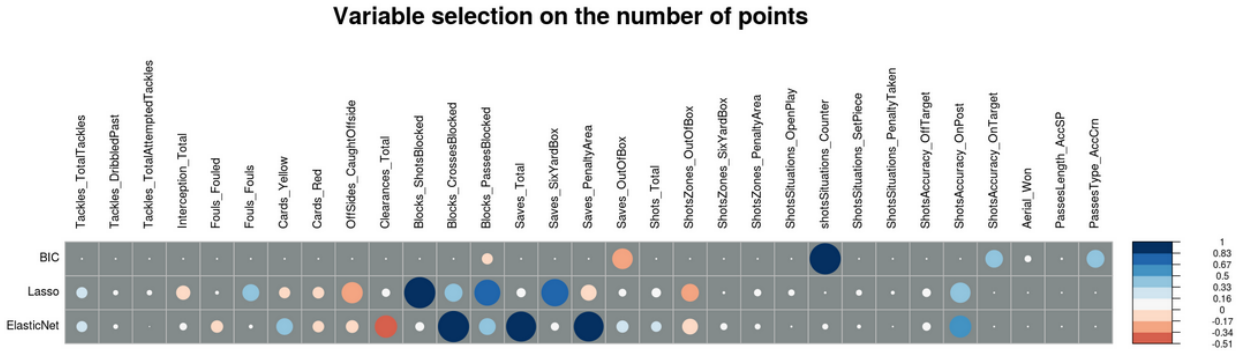


Figure 12: Variables kept according to the method

Based on ElasticNet, crosses blocked, saves and particularly saves in penalty area are highly positively correlated to the number of points at the end of the tournament. `ShotsAccuracy_OnPost` has also a high positive correlation. `Clearance_Total` is negatively correlated to the number of points. We have already seen it in part 2.1, and it is confirmed.

4 Random Forest

4.1 Classification and Regression Tree

A Classification and Regression Tree (CART) is a non parametric algorithm which aims in constructing a binary decision tree, easy to interpret.
Figure 13 is an example of a tree computed with the number of points as target variable.
For each node, the algorithm chooses the explanatory variable which minimize the sum of heterogeneities of the two son nodes. The heterogeneity of the node κ is defined by

$$D_\kappa = \sum_{i \in \kappa} (y_i - \overline{y_\kappa})^2$$

On this tree, we can see several nodes, standing for a variable. Each node split the individuals according to a threshold on the variable indicating by the node. For example, the root (first node on top of the tree) on Figure 13 stands for `ShotsAccuracy_OnTarget` with its threshold 165.

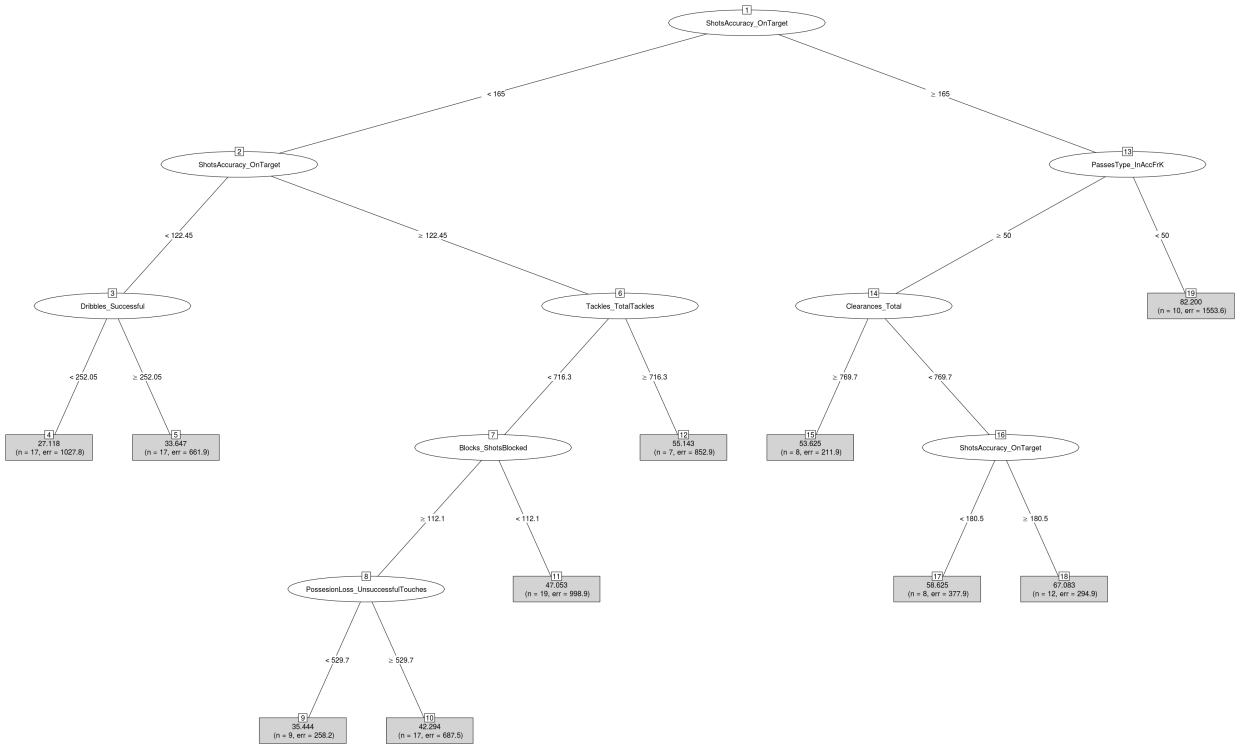


Figure 13: Example of a tree used in Random Forest algorithm

4.2 Feature selection using Random Forest

Random Forest is an aggregation method based on CART. In order to reduce variance of the model, for each node split we select randomly a set of explanatory variables among all of them. This procedure reduce correlation between trees.

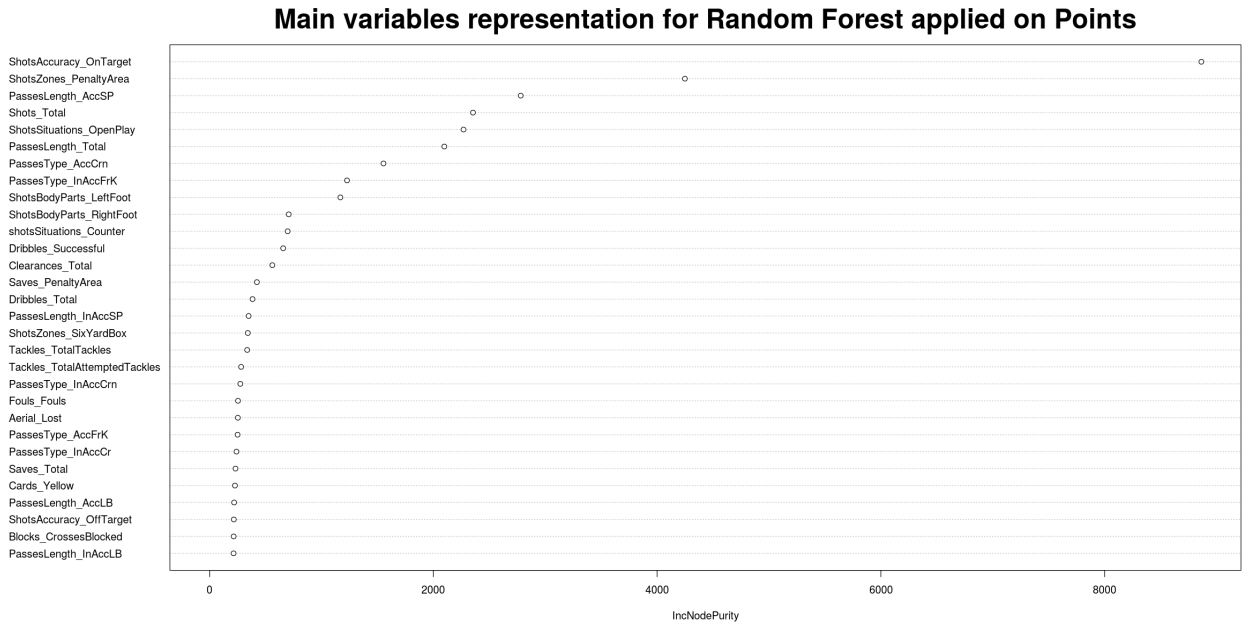


Figure 14: Variable importance based on Random Forest

Conclusion

References