ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ
ΔΙΟΙΚΗΣΗΣ
ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF
BUSINESS

MSc IN
BUSINESS ANALYTICS

Department of Management Science & Technology

MSc in Business Analytics

«CLUSTERING ANALYSIS FOR THE FIFA BALLON D'OR PARTIAL RANKING DATASETS OF THE PERIOD 2010 - 2015»

By

Dimitrios – Taxiarchis Gkoumas

Student ID Number: f2821812

Name of Supervisor: Prof. Dimitrios Karlis

August 2020

Athens, Greece

DEDICATION

To Elmina…

AKNOWLEDGEMENTS

Firstly, I would like to thank my Thesis supervisor Prof. Karlis Dimitris of Department of Statistics at Athens University of Economics and Business. He was always available whenever I had a question or trouble about my research, by steering me in the right direction.

Moreover, I would like to thank my family for their abiding support in my whole life.

Last but not least, I would like to thank all the people that supported me throughout this year.

ABSTRACT

The rapid increase in the volume of the data, in recent years, makes the notion of clustering and the extraction of useful information through it all the more important. A special kind of data, which is applied to many fields such as politics, elections, psychology, sports, market research, etc., is the ranking data. In particular, we are going to work with partial ranking data, which is a very interesting and challenging type of ranking data.

The main purpose of this Thesis, is the clustering of the voters of the FIFA Ballon d'Or partial ranking datasets for the period 2010 – 2015. Our goal is to separate them in different groups according to their preferences, for each one of the years in the period under study. Moreover, we are going to attempt to distinguish possible voting behavioral patterns through further analysis of the clustering results, and extrinsic factors that could have affected the final preference of a voter. Also, we are going to present the fundamental notions in the context of ranking data and provide ways for visualizing and modeling partial ranking data.

TABLE OF CONTENTS

Page

APPENDICES

Chapter 1

Introduction

As we move forward in time, the data driven view has been introduced and adopted from many industries throughout the planet. It is also observed, that this turn to data and analytics, is making determined steps in the sports industry. The combination of statistical methods with the usage of technology and big data tools, accompanied with the knowledge of the sport, that is under research, can create insights of great importance that are very helpful for the sports organizations. The last years more and more teams and organizations are hiring data scientists in order to improve performance and decision making that are about concerns around the matches (which player to put in the game, which one should be substituted during the match, etc.) but also financial concerns (players that are overpaid according to their performance, price of tickets, segmentation of the fans in order to attract them in a better way, etc.). Examples of leading sports associations and leagues, that are in the process of adopting the data driven view, are the FIFA, NBA, American Association of Professional Baseball, Premier League, Bundesliga, etc.

1.1 ***Fifa Ballon d'Or Voting System***

One of the highest honor for a football player, is to be awarded with the FIFA Ballon d'Or award. The Ballon d'Or is an annual football award presented by French news magazine 'France Football'. It is one of the oldest since it has been awarded back in 1956 and is considered as the most prestigious individual award for football players [75]. The name of the award has been changed to FIFA Ballon d'Or , due to the agreement that was made with FIFA for the merge of the Ballon d'Or with the 'FIFA World Player of the Year'. The agreement was made in 2010 and ended in 2016, when the award reverted to its first name [75]. Since the period that we are going to examine is 2010 – 2015, we refer to the award as FIFA Ballon d'Or.

The award is based on a voting procedure, from which the winner is declared. At first, out of the professional football players that exist, FIFA selects 23 players, who thinks that are the best 23 players in the world for this specific year that the competition takes place. After that, this list of players is presented to the eligible voters and is made public. The people who are entitled to vote are divided into three categories [18]:

1) the coaches of the national teams,

2) the captain of each registered country and

3) a group of journalists, limited to one per country.

Each of the three categories has the same electoral weight, notwithstanding the actual sizes of the classes the voters represent [18]. There is not any restriction on whom an eligible voter is able to vote, except the one that states that if a candidate is also eligible to vote, he is prohibited to vote for himself. The eligible voters must choose the top 3 players out of the list in order of their preference. For each vote, the corresponding points are given to the preferred player. The distribution of those points is the following:

5 points for the first player, 3 points for the second player and 1 point for the third player. Thus, the smaller the rank the more points that a player receives. Finally, the player who receives the most points is awarded with the Ballon d'or. In case of draw, the player with the most first – place votes gets the award. As someone can observe, we start mention the word 'rank' and this is the kind of the data that we are going to work with.

1.2 ***Ranking Data***

Ranking data commonly arise from situations where it is desired to rank a set of individuals or objects in accordance with some criterion [45]. This kind of data can be observed directly or as a result from a ranking of a set of assigned or as a transformation of continuous/discrete data. Examples of ranking data, in the literature, can be found in politics, voting and elections, psychology, market research, medical treatments, house reviews, horse racing, etc.

A definition of ranking of n objects, can be the following :

A *ranking* or *permutation* of n distinct objects is a vector of length n, with each component corresponding to an object, and the value of each component being the rank of that object, namely the quantity 1+ the number of other objects that are considered superior, in either a qualitative or quantitative sense [66]. We use

$$\pi = [\pi(1), \ldots, \pi(n)]$$

to denote this ranking or permutation. In terms of preference, an object that receives the lowest rank is the most preferred among the others. The inverse situation of a ranking is the ordering [66].

It is very important to make clear the difference between these two notions, because it is essential to have data in the correct notation, ranking or ordering, that each modeling algorithm needs as input. A definition of ordering could be the following :

An *ordering* or *inverse permutation* of n objects, labeled 1 to n, is a vector of length n, with each component i giving the label of the object that has rank $\mathrm{i}, \mathrm{i} = 1, \ldots, \mathrm{n}$ [66]. The *ordering* or *inverse permutation* associated with π is specified by the mapping $\pi^{-1}(\mathrm{j}) = \mathrm{i}$ if $\pi(\mathrm{i}) = \mathrm{j}$, $\mathrm{i} = 1, \ldots, \mathrm{n}$, $\mathrm{j} = 1, \ldots, \mathrm{n}$ [66]. In other words, it is a permutation which, given an array of size n of integers in range from 1 to n, is obtained by inserting position of an element at the position specified by the element value in the array.

The ranking data can be partitioned to complete and incomplete rankings.

A complete ranking is a permutation, in which all the n objects of the set are ranked by the judges. In the case of FIFA Ballon d'Or data, in case of a complete ranking, all the 23 players of the list should be assigned with a rank from the voters.

In some cases, though, incomplete ranking data are observed, especially when the evaluation of an object is time consuming or takes much effort. In that case, instead of ranking all objects of the set, each individual may be asked to rank the top q objects only for $q \leq t$, called top q partial rankings [45]. This is exactly the case that we have to face in analyzing the FIFA Ballon d'Or datasets. Since the voters are asked to rank only the top 3 out of 23 players of the list, the ranking datasets that we are going to work with, are typical top 3 partial rankings.

The presence of partial ranking data makes the analysis more challenging, in comparison with the case of complete ranking data. First of all, the literature and the software regarding incomplete ranking data is limited, in comparison with the corresponding research and software for complete ranking data. This makes the detection and implementation of algorithms, that can suit to this kind of data, more difficult. It has to be pointed out that the missing positions in an incomplete permutation are not missing data. In terms of interpretation purposes, they can be viewed as an expression of preference, since they represent the non – preference of a judge to a specific object. Thus, they have to be handled in such a way when the ranking is trying to be interpreted. But, in what concerns clustering purposes or different kind of analysis, the researcher has to detect appropriate methods for the estimation of the missing positions in order to conduct inference.

Such algorithms, which have been constructed and presented in the literature, are in great use on cases where incomplete rankings exist. These algorithms sometimes may be complex and difficult to use, in compare with the classic non parametric approaches for clustering complete ranking datasets. Also, some of such algorithms have a computational burden, in terms of the needed time for the simulation and estimation of the missing positions, which adds an extra challenge in the context of modeling partial ranking datasets. Furthermore, the presence of partially ranked data is a very challenging issue, in terms of visualizing them. This happens due to the fact that the traditional methods for visualizing ranking data can note be used. Thus, we have to find alternative ways and methods in order to achieve the graphical representation of such rankings. Besides that, in analysis (e.g clustering) methods that require the use of distance, the presence of such data is an obstacle since many distance metrics are not able to be used because of the scaling. Thus, alternative paths of imputation of the missing positions or different distance metrics have to be used.

1.3 *The notion of Clustering*

After presenting the type of data that we are going to work with, it is time to get a brief overview of clustering, since this is the main purpose of the analysis that will take place. There are plenty of definitions that try to explain what clustering is about:

We could define clustering as a technique which goal is, given a set of data points, to find groups of observations which they ‘look similar’ within the cluster and are ‘different’ from observations of different groups [36].

Clustering has an enormous amount of uses in a variety of industries. Some common applications for clustering include marketing segmentation, social network analysis, anomaly detection, image segmentation, search result grouping, etc. In the context of ranking data, cluster analysis is performed usually in consumer questionnaires, voting forms or other inquiries of preferences [41],[45]. Its main goal is to identify typical groups of rank choices. After that, the further exploration, for trying to find relationships that are based on the common characteristics that objects of the same group have, lies on the researcher’s hand.

1.4 ***The Aims of Thesis***

This Thesis contains eight chapters. In the second chapter, the fundamental classes of models for ranking data are going to be described. Clustering models and models regarding partial rankings, will be emphasized. In the third chapter, the transformation of the raw ‘FIFA Ballon D’or’ datasets to partial rankings, are going to be presented. Moreover, the fundamental descriptive statistics that are applied in such data are going to be presented and implemented, in order to get a better understanding of the data. In Chapter 4, we are going to present visualizations methods for ranking data. We start by describing fundamental approaches for complete ranking data (e.g permutation polytope) and consequently we present visualizations methods for partial ranking data such as metric multidimensional scaling, multidimensional unfolding technique, multidimensional preference analysis and we implement the non – metric multidimensional scaling technique in order to graphically represent our data. The fundamental objective of this Thesis is the clustering of the FIFA Ballon d’Or voters, for each specific year of the period 2010 – 2015. Thus, in Chapters 5,6,7 we present different clustering algorithms that can be applied to the data. In each of these chapters there are two main sections. The first section contains the theoretical framework of the method that is going to be used and in the second section, we apply the method on the datasets for each year of the period under study. The goal of the applications is to detect how these algorithms work out and to create groups of voters that have common characteristics, in terms of preference. Through the clusters that are going to be created, we will try to identify trends of preference for players, possible voting behavioral patterns which can characterize the way that a player is voted and possible ascendance of certain players that can be recognized in a cluster.

Furthermore, we will try to detect if external factors (the word ‘external’ is used due to the fact that these factors are not included in the model) affect the vote decisions of persons. The examined factors are the continent where a voter comes from and the job of the voter. Through this extra analysis we will attempt to understand if, for example, the fact that a voter and a player come from the same continent, plays a vital role in the voter’s final preference. It would be very interesting to reveal such a pattern because this would be a strong indication of the existence of such a relationship, which many football fans assume. In the closing chapter we will present the work and the main discoveries that the research will have advanced, with the hope that the results are going to inspire researchers to delve deeper into this subject.

Chapter 2

Models for Ranking Data

In this chapter we are going to present the fundamental classes of models for ranking data. In the analysis of a ranking dataset, in order to make inference on the preferences of the voters, modeling of the data is needed. In the next sections, there will be presented general models for ranking data but, also, models which concern the clustering of such datasets. It has to be pointed out that most of the following classes of models are used for clustering purposes. Here, we present the general terminology and methodology of these methods. The more specialized theoretical framework of each method, that is used for the clustering implementation on the FIFA datasets, is presented as a separate section before the application part of the method. In the following lines of this chapter we are going to snapshot the general picture of ranking data models and features of those models that make them more preferable than the others, based on the purposes of the analysis.

2.1 ***Probability models for Ranking Data***

The probability modelling for ranking data can be described as an efficient way to understand people's perception and preference on different objects [45]. We are going to present the probability models that have been developed, through the four categories that Critchlow et.al classified them, in 1991 [17]. These four groups of probability models are [45]:

a) Order statistics models,

b) Paired comparison models,

c) Distance – based models and

d) Multistage models.

In the section 2.5, we are going to present the Finite Mixtures model which is also a fundamental class of probability models. It is very interesting to describe these models and their properties, since the Insertion Sorting Rank algorithm and the Bayesian approach of the Plakett – Luce model, that are going to be implemented in the next chapters, are model based clustering algorithms that are founded on probability models. Thus, it is important to describe the fundamental concepts and the properties of such models.

2.1.1 Order Statistics Models

The sense of order statistics models has been introduced from Louis Leon Thurstone, in 1927 [68]. The American psychologist, who was instrumental in the development of psychometrics and statistical techniques for the analysis of performance on psychological tests, published a paper in which the ranking of two objects was considered. The fundamental idea of his proposal was that the final ranking of a judge, on a set of objects, is determined by the ordering of random variables that represent the tastes of a judge. Since these tastes can fluctuate, according to the understanding of the judge to each object, they cannot be predicted. Thus, they are random variables. The probability of observing a ranking π_j under the class of such a model can be described from the following formula :

$$P(\pi_j) = P\big(y_{[1]j}j > \ y_{[2]j}j > \cdots > y_{[t]j}j\big), \pi_n \in S \ ,$$

where π_j is the ranking of t objects, the set $([1]_j, [2]_j, \dots, [t]_j j$ is the ordering of objects corresponding to the ranking π_j such that the judge j assigns rank i to object $[i]_j$ [45]. The set $y_{1j},\ y_{2j}, \dots, y_{tj}$ represents the random utilities from which the ranking is dependent. The term S in which every ranking belongs to, is the set of all $t!$ possible rankings. In order to make the model simpler, some probabilistic structures on the random utilities are assumed. Also, Critchlow et al. (1991) [17], observed that if these utilities are allowed to have arbitrary dependencies, any probability distribution can be expressed as in the upon formula [45]. Such type of models that can be presented through this formula are referred to as Thurstone order statistics models (Yellot 1977 [77], Critchlow et al. 1991 [17]). The two most famous Thurstone models, that have been further developed in the following years, are the Thurstone model (Thurstone 1927 [68], Daniels 1950 [19], Mosteller 1951 [9]) and the Luce model (Bradley and Terry 1952 [5], Luce 1959 [40]).

The ranking probability in the Luce model can be expressed as a function of top – choice probabilities only. Also, the model satisfies the Independence of Irrelevant Alternatives (IIA) axiom, which introduced by the mathematical psychologist Tversky in 1972 [35]. The axiom states that the choice of a judge between two objects, depends on the preferences between these two objects only and is irrelevant to another object. The axiom is also being satisfied from extensions of the Luce model, such as the Rank – Order Logit models, which include judge – specific covariates, object – specific covariates and their interactions. It has to be mentioned at this point that the Plackett – Luce model, thus the Bayesian approach of the model that is going to be used as a clustering method later, is based on this Luce's axiom of choice.

The main drawback of this axiom is that is impractical because the correlation among the errors is not included in the models and this can lead to unrealistic patterns in many real life ranking problems. Thus, some order statistics models that do not satisfy the IIA property have been developed. An example of such a model is the Multivariate (Generalized) Extreme Value model (GEV), which was introduced by McFadden (1978) [47]. The model assumes that the error terms of the simplified probabilistic structures, that have been mentioned previously, follow a generalized extreme value distribution with the following cumulative distribution function :

$$F(\varepsilon_1, \dots, \varepsilon_t) = \exp\,[-H(e^{-\varepsilon_1}, \dots, e^{-\varepsilon_t})],$$

where H is t – dimensional and all the univariate marginal distributions are Gumbel distributed [20]. The key pros of the GEV model is that it is able to fit many different types of ranking data, as Joe (2001) [32] stated.

2.1.2 ***Paired Comparison Models***

The notion of ‘Paired Comparison models’ was first introduced from Babington Smith (1950) [4], whose proposal was about a family of probability models for ranking data based on the paired comparisons idea. By assuming mutual independence of these paired comparisons, the probability of observing a ranking π_j, under the Smith model, is given by the formula :

$$P(\pi_j) = C \prod_{\{(a,b):\pi_j(a)<\pi_j(b)\}} p_{ab} ,$$

where the constant C is appropriately selected in order to make the probabilities sum to 1 and p_{ab} is the probability of object a being preferred to object b.

Based on this model, Mallows (1957) [42] proposed the addition of constraints on the $\{p_{ab}\}$ term. His idea lead to two subclasses of the Smith model : the Mallows – Bradley – Terry model and the Mallows model. The Mallows model (1957) [42] came after the work of Bradley and Terry (1952) [5], and it was a try for simplification of their model. The Mallows work is also crucial, as it was introduced in the ranking literature the notion of modal ranking.

Definition: A probability model is said to be strongly unimodal with modal ranking π_0 , if its ranking probability has the unique maximum at $\pi = \pi_0$ [45] .

Also, based on [30] , the modal ranking rule is supremely robust to noise, in the sense of being correct in the face of any ‘reasonable’ type of noise. At this point, before presenting the general formula of the Mallows model, we have to define the Kendall and Spearman distances. For any permutation π,σ the Kendall distance is defined as :

$$D_K(\pi, \sigma) = \sum_{i<j} I \,\{[\pi(i) - \pi(j)][\sigma(i) - \sigma(j)] < 0\},$$

where $I\{\cdot\}$ is the indicator function taking values 1 or 0 depending on whether the statement in brackets holds or not [59].

Moreover, the Spearman distance is defined as :

$$D_S(\pi, \sigma) = \frac{1}{2} \sum_{i=1}^{t} [\pi(i) - \sigma(i)]^2 \text{ [59]}$$

Based on the above, the formula of the Mallows model is the following :

$$P(\pi_j) = c(\theta, \varphi)\theta^{d_S(\pi,\pi_0)}\varphi^{d_K(\pi,\pi_0)} ,$$

where $c(\theta, \varphi)$ is selected in order to make the probabilities sum to 1, $d_S(\pi, \pi_0)$ is the Spearman distance and $d_K(\pi, \pi_0)$ is the Kendall distance between π and π_0 [30]. The model states that as the distance from π to the modal ranking increases, the ranking probability decreases geometrically according to this increase.

It has to be pointed out, that the Paired Comparison models satisfy many properties of the ranking models. In specific, Marley (1968) [44] showed that the class of these models satisfy the reversibility property, which states that the reversing of a ranking π has no effect on the probability models, based on the reverse function $\gamma(\pi) = t + 1 - \pi$. Moreover, it satisfies the property of L – decomposability which states that the ranking of t objects can be decomposed into t – 1 stages. There are more properties that are fulfilled from this type of models, but with some conditions that have to be activated.

2.1.3 Distance – Based Models

A class of distance – based models was developed by Diaconis (1988) [21]. His work rests on the idea that a distance function can measure the discrepancy between two rankings and the fact that is reasonable to assume most of the judges to have rankings close to the modal ranking π_0. The general form of the distance – based model can be described as

$$P(\pi|\lambda,\pi_0) = \frac{e^{-\lambda d(\pi,\sigma)}}{C(\lambda)},$$

where $\lambda \geq 0$ is the dispersion parameter and $d(\pi,\sigma)$ is an arbitrary right - invariant distance. It has to be mentioned at this point that it is required for the selected distance to satisfy the property of right invariance. This is a property that is explained in further sections and ensures that a possible relabeling of the ranking objects does not affect the distance. Examples of such distances are the Spearman Footrule, the Spearman distance and the Kendall distance. In that case where the Kendall distance is used as the distance function in the formula, the model is called the Mallow's φ – model and is a proved relationship between the distance – based models and the paired – comparison models (Critchlow et.al., 1991). The Mallow's φ – model is also a special case of the φ – components models class, which has been introduced from Flinger and Verducci (1986) [22]. Their work was mainly based on an extension of the distance – based models, by decomposing the distance metric $d(\pi,\sigma)$ into $t - 1$ objects, where t is the number of the ranked objects. Regarding the general formula that has been presented above, the ranking probability in a distance – based model holds the largest value at the modal ranking π_0 and at the same time it declines when it is away from π_0 . The decline rate of the ranking probability is dependent on the dispersion parameter λ. Thus, if the value of λ is small, the distribution of rankings will be more concentrated around modal ranking and vice versa. To get a better understanding of this fact, we have to analyze the general formula of the distance – based models. The term $\frac{e^{-\lambda d(\pi,\sigma)}}{C(\lambda)}$ can be converted to $\frac{\frac{1}{e^{\lambda d(\pi,\sigma)}}}{C(\lambda)}$ which is equal to $\frac{1}{C(\lambda)e^{\lambda d(\pi,\sigma)}}$. Based on the last term, it can be noticed that if the value of the dispersion parameter λ is large, it leads the denominator of the term to increase, which causes the value of the fraction to getting smaller. On the other hand, if the value of λ is small, the value of the term is getting larger. Thus, we can observe that when the dispersion parameter is large the ranking probability is small, which means that is away from the modal ranking and vice versa. The maximum likelihood estimator (MLE) $\hat{\lambda}$, can be found by solving different equation if the modal ranking is a known value and if it is an unknown value.

We have to point out that the distance – based models satisfy both the reversibility and the label invariance properties, that have been mentioned previously. Besides that, this class of model is able to handle partial ranking data, with the implementation of some modifications on the used distance measures. The estimation of the model parameters using the EM algorithm, was introduced by Beckett (1993) [7]. Contrarily, an approach which does not include the EM algorithm was proposed by Adkins and Flinger (1998) [45], who showed a non – iterative maximum likelihood estimation procedure for the Mallow's φ – model. Moreover, the mixture models have also be considered for the class of distance – based models. Specifically, Murphy and Martin (2003) [51] extended their use, in order to describe the presence of heterogeneity among the judges of a dataset. Their work was proved very helpful, because they smoothed the assumption of the homogeneous population in the distance – based models.

2.1.3.1 Weighted Distance – Based Models

Despite the fact of the great usefulness of the φ – component models that have been mentioned above, some distance properties are not satisfied in specific cases of such class of models. Thus, Lee and Yu (2012) [39], provided the notion of weighted distance measures which are able to retain all the required properties of a distance and also allow different weights for different ranks, which enhance the model flexibility.

Thus, many distances that are used in the distance – based models (like Kendall, Spearman, Spearman Footrule), were introduced in a weighted format. For example, the Spearman weighted distance formula is

$$d_S(\pi, \sigma; w) = \sum_{i=1}^{t} w_{\pi_0(i)} [\pi(i) - \sigma(i)]^2 \ .$$

In general, the probability of observing a ranking π under the weighted distance – based ranking model is [45]

$$P(\pi | w, \pi_0) = \frac{e^{-d(\pi, \pi_0; w)}}{C(w)} \ .$$

Based on this formula, the value of weight can be interpreted in three ways. At first, in the case of a large value of w_i, is a supporting factor to the assumption that the ranking of the object *i* is close in π_0 . Secondly, if the value of w_i is close to zero, then a change in the rank of the object ranked *i* will not have a serious impact on the distance. Finally, if the value of weight is zero, then the model is uniform. Besides these, Lee and Yu (2012) [39] motivated from the work of Murphy and Martin (2003) [51], and took into account the finite mixtures to the weighted distance – based models. In order to estimate the model parameters they applied the EM algorithm, by computing for each observation the probability of belonging to every subpopulation and maximizing the conditional expected log – likelihood, given the estimates in the first step. In order to derive the EM algorithm, they defined a latent variable, which indicated if an observation belonged to the specific subpopulation.

2.1.4 Multistage Models

In 1988, Flinger and Verducci [23] introduced the class of multistage models. Multistage ranking models, including the popular Plackett-Luce distribution (PL), rely on the assumption that the ranking process is performed sequentially, by assigning the positions from the top to the bottom one (forward order) [13]. The general idea of this class of models was to decompose the ranking process into a sequence of independent stages.

For example, if *t* objects are about to be ranked, the ranking process can be decomposed into *t* – 1 stages, where at stage *i*, the *i*th object is chosen. In specific, the most preferred item is selected at the first stage, the best of the remaining items at the second stage and this procedure keeps going until the least preferred object is selected. Flinger and

Verducci (1988) [23] proposed a general multistage model with $\frac{t(t-1)}{2}$ parameters and three more specialized models. These models are named as the free model, the strongly unimodal model and the exponential factor model. The main difference between these three models is the number of constraints that each of the model has.

The free model, as its name indicates, has the least constraints while in the exponential factor model some conventions are required in order to run. Besides these models, that Flinger and Verducci proposed, another multistage model was also proposed, in Hu (2000). He showed the decomposition of the ranking process can be also done for $(t-1)^2$ parameters c_{ij}, where both i and $j = 1,2,\dots,t-1$. The parameters c_{ij} are used in order to determine which object will be selected in each stage. Furthermore, finite mixtures of multistage models have been introduced in the literature. These mixtures can provide interesting adequacy power, for the assessment of the modeling. On the other hand, if we compare them with the mixtures of distance – based models, the adequacy power of the distance – based may not provide a powerful assessment but they have more meaningful parameters and, also, are easier to be implemented.

2.1.4.1 Connection with Plackett – Luce Ranking Model

It has to be pointed out that under the decomposition process that was described above, the Luce models and the φ – component models can also belong to the class of multistage models. This is very important, if we consider that one of the three clustering approaches that are implemented in this Thesis is the Bayesian finite mixture of Plackett – Luce model. As it is going to be described in the corresponding section, the Plackett – Luce model is a powerful stagewise model for analyzing partial ranking data. Based on the decomposition process of ranking data, we consider a set of items, and a set of choice probabilities that satisfy the Luce's axiom. Next, we consider a pick of one object at a time out of the set, according to the choice probabilities. Such samples give a total ordering of objects, which can be considered as a sample from a distribution over all possible rankings [33]. The form of such a distribution was first considered by Plackett (1975) in order to model probabilities in a K – horse race [33]. The most important aspect of the Plackett – Luce model, is the fact that it is applicable either each observation is provided by a complete ranking of all items, or a partial ranking of some items. Because the data we are analyzing are partial rankings, a multistage model such as the Plackett – Luce, is a very appropriate choice for clustering our data.

2.1.5 Finite Mixture Models for Ranking Data

The finite mixture models were first introduced by Newcomb (1886) [52] who used them in order to model outliers. Since then this class of models has definitely gained ground in the literature, as it has been researched and developed on a great scale. A basic interpretation of a finite mixture model is that it provides a natural representation of heterogeneity in a finite number of latent classes. The heterogeneity concerns the effects on different groups of observations [54]. The general formula of a finite mixture model density with parameter vector $\theta = (\pi', \theta'_1, \theta'_2, \dots, \theta'_K)'$ is the following :

$$f(x;\theta) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k),$$

where π_k represents the kth mixing proportion or the probability that the observation x_i belongs to the kth subpopulation with corresponding density $f_k(x)$ [48]. The term K represents all the number of components with $\pi = (\pi_1, \pi_2, \ldots, \pi_K)'$,where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The most general form of a mixture is to suppose that f_k's are a parametric form which is known e.g $f_k(x) = f_k(x; \theta_k)$, where in that case only θ has to be estimated. But when K is also not provided, the number of components in the mixture have to be additionally estimated.

In general, the clustering approaches that depend on probability models are more and more being developed and used in the literature. There are many times that the data from such models are considered to come from finite mixture probability distributions. Moreover, in the finite mixture model framework, each group is assumed to have its own distribution and corresponding probability of representation [48]. It is observed, also, that the finite mixture models and the context of model – based clustering is preferred over distance – based approaches or more heuristic approaches (e.g hierarchical clustering), due to the more statistical oriented view of the clustering problem, the various types of ranking data that these models can analyze, the more robust answers in questions such as the number of clusters etc.

In respect of the clustering goals, after the mixture model has already been fitted, is being used in order to identify any grouping that is probably exist in the data. For example, if a four component model has been fitted, we would want to detect any pattern that can be identified between these groups. Thus, the overall target of the usage of mixture models in the clustering context is to separate the data into a number of groups – components, in which the objects in each group would have common characteristics but the objects of different groups would differ. When the finite mixture models are used for clustering purposes, there is one initiative that has been introduced in this approach. This is about the assignments of data points to the different clusters, which is the fundamental information that concerns a clustering problem, and is missing from the observed sample.

In order to deal with this issue, in the finite mixture models clustering approach, a random variable which can be noted as z_{jp} is introduced. This random variable can take the following values : $z_{jp} = 1$, if a data point y_j belongs to a population p and $z_{jp} = 0$, otherwise. These variables can also be referred as latent variables, because they are not directly observed but inferred through a model. We assume that the conditional density of Y_j , where Y_j is a random vector of data points, given $Z_{ji} = 1$ is $f(y_j; \theta_i)$, where θ_i is an unknown vector of parameters for the ith component of the mixture. Also, we make the assumption that the random variables $\{z_j\}$are independent as Picard proposed in [56]. Based on these two assumptions, the random variable z_j, can be seen as a categorical variable that indicates whether the data point belongs to a specific group or, in other words, the labeling of the data points. Thus, this posterior probability of the variable z_j, given the observed value of y_j, will play the most important role in what concerns the clustering purposes.

2.1.5.1 *Brief Introduction to EM Algorithm*

Since the label of each data point is not known, because the random variables that previously mentioned are latent, the estimation of the mixture parameters can be obtained through the observed data. A fundamental approach in the estimation methodology, which has been made very important progress through the years, is the maximum likelihood method. One of the main reasons that this method had such an advance, was the development of the EM algorithm.

The Expectation – Maximization (EM) algorithm is an iterative process which goal is to approximate the maximum likelihood function. Its fundamental idea, is to connect a complete data model with the observed incomplete model, in order to make the computation of the maximum likelihood estimates less complex. Thus. it is often used in the case of incomplete/missing data, or in the case of existence of latent variables. The EM algorithm achieves, through its iterative process to fill the gap of maximum likelihood estimation which does not achieve to approach the 'best fit' of a model, when incomplete data exist.

The algorithm consists of two steps, as its name indicates, Expectation (E) and Maximization (M). In the first – Expectation step, it estimates the missing or latent variables, by computing the expected value of the complete – data log likelihood $l(\theta; X, Y)$, where θ is the unknown parameter vector, given the observed data and the current parameter estimate [29]. The second – Maximization step consists of maximizing the parameters of the model over the expectation computed in the E step. The process is repeated until the sequence of the maximized θ's parameters converges. In the case that the log – likelihood function has multiple local maximums then the algorithm should be put in run for many times, by using a different starting value for the unknown parameter θ , at each iteration [29]. This helps the algorithm not to end up with a local maxima, that will probably not be close to the global one, but after many restarts to arrive to the greatest maximum likelihood.

2.2 ***Probit models for Ranking Data***

The notion of 'Probit models' for ranking data is based on models of choice probabilities, that use a set of random utilities. Choice probabilities are derived from two distributions of the random terms : the extreme value, i.e Logit, and the multivariate normal, i.e Probit [53]. In the next two paragraphs we are going to present the two fundamental classes of probit models, the Multivariate Normal Order Statistics and the Factor Analysis.

2.2.1 Multivariate Normal Order Statistics Models (MVNOS)

This class of models, as their name indicates, are very similar with the Thurstone order statistics models that have been described in section 2.1, in the sense that both classes of models assume that the ranking that a judge gives to a set of objects is determined by the ordering of the corresponding latent utilities for the objects assigned by the judge. The fundamental difference between them, is that the Thurstone models assume independent utilities, in contrast with the Multivariate normal order statistics models that the utilities are possibly correlated.

The probability of a ranking π_j that is given by judge j can be described with the following formula :

$$P(\pi_j) = P\left(y_{[1]_j,j} > y_{[2]_j,j} > \cdots > y_{[t]_j,j}\right),$$

where $[1]_j, \dots, [t]_j$ is the ordering of the t objects corresponding to the ranking π_j. Furthermore, the latent utility vector $y_j = (y_{1j}, \dots, y_{tj})'$ of judge j is supposed to follow multivariate normal distribution with mean utility vector $\mu_j = (\mu_{1j}, \dots, \mu_{tj})'$. A great example where the MVNOS model used for the modeling and clustering of judges, was the Analysis of the APA Election Data. In 1980, the American Psychological Association (APA) conducted an election in which five candidates (A, B, C, D, E) were running for president and voters were asked to rank all of the candidates [45]. Among those voters, 5738 gave complete rankings and those complete rankings were considered in the MVNOS clustering implementation. The results indicated separate groups of voters, where each one had a distinct characteristic.

At this point, it has to be mentioned that the MVNOS model allow the presence of covariates that are associated with the judges and the objects that are modeled. For example, in the case that the FIFA Ballon D' Or data were complete so we could implement the MVNOS model, a possible judge – specific and object – specific covariate could be the country that the judge and the player come from, because this could affect the vote of a judge. Thus, in the MVNOS class of models, a linear model is imposed in order to include these covariates. This linear model is imposed for the mean utility vector μ_j in the following manner : $\mu_j = Z_j\beta$, where Z_j is a t x p matrix of covariates associated with judge j and β is a p x 1 vector of unknown parameters [45]. Then, as previously mentioned, someone could study the impact that the covariates associated with the judge and the objects have in the preferences of the judges. In specific, if we define as s_j the country of the judge and as a_i the country that the ranked player comes from, we would obtain the following model :

$$\mu_{ij} = a'_i\gamma + s'_j\delta_i \, , i = 1, \dots, t,$$

where the parameter vector γ represents the effect of the player's country to all the voters and the vector δ_i represents the country which the voters come from, and may affect their preference to the player i .

In order to test the adequacy of the presented models, a general solution is to group the rankings into a small number of subgroups and examine the fit for each subgroup. The fit can be tested by comparing the observed frequency with the expected frequency of each ranking. In case that the expected frequencies match the observed frequencies, the researcher can claim that the MVNOS model appropriately fits the data.

2.2.2 Factor Analysis

In general, factor analysis is a technique that is used to reduce a large number of variables into fewer number of meaningful factors. It is widely used in social sciences, economic sciences, marketing research etc., for the identification of common characteristics and the construction of groups based on these characteristics, among a set of variables. For the ranking context, by adopting the MVNOS framework with the latent utilities satisfying the general factor model, this model can be generalized in order to be able to analyze ranking data.

So, let's assume that there is a random sample of n individuals that each one is asked to rank t objects. Within the MVNOS framework, the ranking of the t objects given by the individual j in the factor model is determined by the ordering of the t latent utilities $y_{1j}, \dots, y_{tj}$ which satisfies a more general d – factor model :

$y_{ij} = z'_j a_i + b_i + \varepsilon_{ij}\ , \quad j = 1, \dots, n\ ; \ \ i = 1, \dots, t(> d)$ [20]

The terms of the above are explained as follows : $b_i = (b_1, \dots, b_t)'$ is the mean utility vector which depicts the relative significance of the t objects, $a_i = (a_{i1}, \dots, a_{id})'$ represents the factor loadings which provide the variance explained by a variable on that particular factor, $z_j = (z_1, \dots, z_n)$ are the latent common factors which are assumed to be independent and identically distributed according to the standard $d-$ variate normal distribution, ε_{ij} is the error term which represents the unique factor that is assumed to follow a $N(0, \sigma^2)$ distribution, independent of the latent factors z's. The unobservable response utilities and the latent common factors are simulated through the Monte Carlo Expectation – Maximization Algorithm, where the E- step is implemented through the Gibbs sampler.

The d – factor model was proposed concerning complete rankings. But the model can be extended also when incomplete rankings exist. Thus, in the case of top q partial rankings, we can assign objects with ranks respectively, and assign the midrank value to those objects that have not been ranked. The notion of *midrank* value is going to be described on a great scale in further section. As a small note, its formula is

$$[(q+1) + \cdots + t]/(t-q)\ ,$$

where t are the objects that are about to ranked and q are the objects actually ranked. The result of this formula is replaced in every missing position of an incomplete ranking, in the context of factor analysis for partial ranking data. Moreover, when top q partial rankings exist, the process of Monte Carlo Expectation – Maximization Algorithm is implemented for the top q objects and the rest of the objects are simulated by $N(z'_j a_i, {\sigma^2}_i)$.

2.3 ***Decision Tree Models for Ranking Data***

Besides the various types of probability models and the two fundamental types of probit models that have been presented up to know, there is one extra class that is going to be mentioned. The name of this class of models is 'Decision Trees models'. These types of models come to solve the issue of the difficult interpretation of the fitted models coefficients, when nonlinearity or higher – order interactions exist, due to the interaction covariates. The use of decision trees can provide a powerful nonparametric model capable of automatically detecting nonlinear and interaction effects [45]. This could serve, also, as a complement to existing parametric models for ranking data [45]. Thus, since the main advantage of such models is the easiness in the interpretation, they are popular in problems that concern classification or regression. In our case, since the interpretation is not a key fact for clustering, this class of models is not used.

The reason that the decision tree models got this name, is due to the fact that they can be constructed by a set of conditions displayed in a treelike structure. The common procedure for the construction of a decision tree is to start from the root node, that is the entire dataset, and separate the data into two or more child nodes, in a repetitive way. The goal is the new class of nodes to have better performance than the previous – parent node. Thus, in order to make the appropriate split that would achieve this target, in each iteration, a splitting criterion has to be chosen. In what concerns this splitting criterion there are two fundamental approaches. The first approach is the partition based on an impurity function. As ‘impurity’ we could define a metric of how often a randomly selected object from a set would be wrongly labeled, if we assume that it was labeled according to the distribution of labels in the subset. Such functions are the Gini index and the entropy. The second approach is a statistical oriented approach, which does the splitting by applying a statistical test of homogeneity to test whether the split can make the child nodes with significant different distributions of the data [45]. Such independence tests are the chi – square test and the likelihood ratio test.

After mentioned the two main approaches for the splitting of nodes, we will briefly present the two stages for the construction of a decision tree. In a general manner, based on the CART (classification and regression tree) method of Yu et al.(2010), a decision tree is constructed through two stages. The first stage is called ‘tree growing’ and the second stage is called ‘tree pruning’. Before starting the construction of the tree, the ranking dataset is randomly partitioned into a training set and test set. Then, in the ‘tree growing’ stage, the algorithm starts from the whole training set (root node) and through the iterative process that has been described previously, partitions each node to detect the best split according to Gini index or entropy, or according to a statistical test of independence. The procedure stops at the time that some stopping criteria are met.

Someone, can notice that this fact explains the name of the first stage because when these criteria are met, the tree has finally been built. In the ‘tree pruning’ stage, is measured the significant improvement that each branch, of the previously built tree, makes. The branches that show the less significant improvement are removed from the tree. The significance is measured through a cost – complexity metric, based on a ten – fold cross – validation [45]. In order to assess the performance of the decision tree, a very widely used measure is the area under the receiver operating characteristic or ROC curve. Its values fall within the range 0.5 – 1.0, where 0.5 denotes a random prediction and 1.0 indicates perfect accuracy of the prediction. Despite the fact that ROC curve can be implemented only for binary data, Yu et al. generalized the notion into ranking data.

Chapter 3

Manipulation and Exploration of the Datasets

In this chapter, the datasets that are used for the clustering processes are explored. Moreover, the cleaning process and the transformation into compatible ranking format for the use of the clustering algorithms is presented. Finally, some descriptive statistics are given, in order to get a better understanding of the data.

3.1 Presentation of the Datasets

The datasets used for this thesis were retrieved from https://data.opendatasoft.com/explore/dataset/fifa-ballon-dor-2010-2015%40public/table/ , on 22/09/2019. There were retrieved 6 datasets, as .csv files, each one corresponding to a year of period 2010 – 15. The content of the datasets has to do with the FIFA Ballon d'Or votes for this period. Each one of the datasets contain the 6 following columns:

1) Year : The year of period 2010 – 15 for which the dataset is about.

2) Vote : The kind of relationship that the voter has with football. This column can have the three following values : Coach, Captain and Media.

3) Country : The country origin of each voter.

4) Name : The name of the voter.

5) Position : The position that the voter ranked the corresponding player. This column can have three values : First (if the voter ranked first the corresponding player), Second (if the voter ranked second the corresponding player) or Third (if the voter ranked third the corresponding player). Next to each one of these values, there are written off the equivalent points for each of the vales, in a parenthesis. These are : 1 point for the Third rank, 3 points for the Second rank and 5 points for the First rank.

6) Player : The player that is ranked from the corresponding voter in the same row. Some of the players have next to their names a backslash and the country they come from.

The datasets differ in term of the number of rows each one contains. Because the number of players does not change, as it is always 23, that means that only the number of voters changes every year. In particular, the first dataset (2010) consists of 1275 rows, the second (2011) consists of 1387 rows, the third (2012) consists of 1513 rows, the fourth (2013) consists of 1623 rows, the fifth (2014) consists of 1632 rows and the sixth dataset (2015) consists of 1494 rows.

3.2 Data Cleaning Process

The process of cleaning the raw data prior to analysis is necessary and an important before implementing any part of the analysis. The cleaning process that has been implemented, is approximately the same for the 6 datasets because all of them appeared to have similar issues.

The first step is to read the datasets. For this purpose, the function ‘fread’ from the ‘data.table’ package has been used. Thus, we read each of the datasets as data frames and, by using the function ‘fread’, the delimiter of the csv file is detected automatically. Also, we read the datasets with the ‘UTF-8’ encoding, in order to not face any issue with special characters that probably exist in the names of the players or voters.

Then, we check the dataset in order to detect problems that need to be repaired. The columns do not face any concerning issue, except the ‘Player’ column. When we try to crosscheck the number of unique players that exist in this column, the number of players is not 23. Thus, this column needs cleaning in all of the 6 datasets. The names of the players are presented by character vectors. First of all, as it has been mentioned in the ‘Presentation of the Datasets’ section, the names of some of the players are written of with their country and a backslash between the two characters. This leads to the presence of the same player more than one time e.g "Messi Lionel" and "Messi Lionel / Argentina". We deal with this issue by erasing the country and the backslash from each one of the character vectors, by using the ‘gsub’ function. After performing this step, we check again in order to ensure that exist 23 unique names of players, but we observe that there are more than 23 names in the column. By taking a closer look, it is noticed the presence of duplicates. This is due to the format of the character vectors. The main issue is the presence of a space in the tail of a player’s name and the non – presence of a space in the same name, in the list e.g "Sneijder Wesley " and "Sneijder Wesley". Also, there exist double spaces inside the character vector because of the removal of the country and the backslash. These issues are solved by removing the white space in the tail of each character vector through the ‘trimws’ function and by replacing the double whitespace with one space through the ‘gsub’ function. Moreover, in order to ensure that all the character vectors have the same format, we apply the ‘format’ function the ‘Player’ column. After applying these steps, the unique names of players decrease, but still do not reach the 23. The reason is that in the rows of the dataset there is a player’s name which is called ‘invalid vote’. This character, as the name implies, is a vote that does not correspond to any player. The number of rows that contain this value is below 10 in all of the datasets. Thus, we are going to exclude the rows that contain this value in order to not affect the final result.

3.3 Data Transformation Process

After implementing the cleaning process, the next step is to create matrices from the original data, which are going to have the format that is required in the packages that deal with ranking data.

At first, we break the column ‘Position’ of the original data, by creating two new columns. The first one is called ‘Points’ and it contains the points that each voter assigned to the corresponding player and is the number of the part being in the parenthesis of ‘Position’ column, e.g First (5 Points) : Points = 5 . The second one is called ‘Ranks’ and it contains the value that each voter ranked the corresponding player and is the number of the part outside the parenthesis of the ‘Position’ column, e.g First (5 Points) : Ranks = 1 .

After constructing the columns ‘Points’ and ‘Ranks’, a Matrix is constructed. The dimensions of the Matrix are the number of unique voters, for the rows, and the number of unique players, for the columns, which always is 23. Thus, it is a Matrix where in the columns are the ranked players and in the rows are the judges. Now, the goal is to assign the exact rankings that a player received from the corresponding voter. To be more specific, the values (1,2,3) that a voter – row has given to three players – columns must be appear in the cells of the Matrix where this voter - row and these players – columns are met. The other positions in this row are going to be filled by NA value. After constructing the Matrix the NA’s are replaced with zero values, because this is the format that most of the packages that deal with ranking data, require the missing ranks of a partial ranking to be denoted.

At this point, the target is to assign the correct values to the cells of the Matrix. To achieve that, we write SQL queries, inside our R script, through the ‘sqldf’ function. More specifically, we select the columns ‘Name’, ‘Player’ and ‘Ranks’ from the transformed dataset for each one of the 23 players, by adding the SQL statement ‘Where’ in the query. This process is done for the 23 players separately and each one of the results is assigned to a unique data frame. Thus, each of these data frames contain the name of the voter and the corresponding rank, for every row of the transformed data that this player exists. Next, we convert the class of the Matrix from ‘matrix’ to ‘data frame’, in order to be able to use the dollar sign. At this point, for each one of the players – columns of the Matrix, we match the names of the voters to the voters – rows of the Matrix through the ‘match’ function and we assign the ‘Ranks’ of the corresponding player from the Player’s SQL query, that has been stored as data frame in the previous step. Thus, the final Matrix consists of the rankings that each player – columns received from the corresponding voter – row. A snapshot of the first 5 rows and the first 9 columns of such a matrix, for the year 2010 is the following :

	Guyan Asamoah	**Sneijder Wesley**	**Maicon**	**Villa David**	**Alves Daniel**	**Forlan Diego**	**Xavi**	**Iniesta Andres**	**Casillas Iker**
Chamroeun Ung	3	0	0	0	0	0	0	0	1
Patoommawatana Urai	0	1	0	0	0	0	2	3	0
Jonuz Mirsad	0	2	3	0	0	0	1	0	0
Ouk Mic	0	0	0	3	0	0	0	0	1
Colome Jaine	0	0	0	0	3	0	0	0	0

Table 1 : Example of ranking table , with the first 5 rows and 9 columns, for Year 2010.

3.4 Descriptive Statistics

Descriptive statistics present an overall picture of ranking data. It is recommended to be considered before any analysis, in order for the researcher to have a better understanding of the data. The target, in descriptive statistics for ranking data, is to find ways to describe the central tendency of people’s preferences through their ranks. In our case, the goal is to describe the central tendency of the voters preferences to players.

There are three common statistics, that are used to describe ranking data. We start with the mean rank, which is a standard measure to present the central tendency of ranking data.

The mean rank of an object can be defined as

$$m_i = \sum_{j=1}^{t!} n_j \, v_j(i)/n \, ,$$

where m_i is the mean rank of object i, v_j, j = 1,2,…,t! represents all the possible rankings of the t objects, $v_j(i)$ is the rank score given to object i in ranking j, n_j is the observed frequency of ranking j and $n = \sum_{j=1}^{t} n_j$.

Another metric that is also commonly used is the pairwise frequencies measure, that is, the frequency with which object i is more preferred than object j for every possible object pairs (i, j). In the matrix that represents the ranking data, this can be defined as the number of observations which the first item (row) has been ranked higher than the second item (column). These pairwise frequencies can be summarized in a matrix. Let’s suppose that someone would like to make the comparison of received votes between three players (Player 1, Player 2 and Player 3) and Player 1 has ranked higher than Player 2, Player 3 from 9 and 7 voters, correspondingly. Also, we assume that Player 2 has ranked higher than Player 1, Player 3 from 5 and 4 voters, correspondingly. Finally, the Player 3 has ranked higher than Player 1, Player 2 from 12 and 14 voters, correspondingly. Then, based on the above example, the matrix that represents the pairwise frequencies of the three players will be the following :

	Player 1	**Player 2**	**Player 3**
Player 1	0	9	7
Player 2	5	0	4
Player 3	12	14	0

Table 2 : Example of pairwise frequencies table.

In addition to mean ranks and pairwise frequencies, we can look for further insights in the ranking data by studying the marginal distribution of the items. Marden (1995) [43] , defined a matrix with $t \; x \; t$ dimensions, in which the (a, b)th entry equals to

$$M_{ab} = \sum_{j=1}^{t!} n_j I[v_j(a) = b],$$

where M_{ab} is the frequency of object a being ranked bth and $v_j(a)$ is the rank score given to object a in ranking j [45]. Marden named such a matrix as ‘marginal matrix’ because the ath row gives the observed marginal distribution of the ranks assigned to the object a and the bth column gives the marginal distribution of objects given the rank b [45]. In the matrix of ranking data, this can be described as the number of observations which the item i (row) has been ranked j (column).

Based on these statistics we are going to get insights for our datasets. Because of the large size of the datasets, we are going to provide results for some of the years. Before obtaining the results of the descriptive statistics, we have first to define the midrank imputations for the incomplete rankings of the datasets. Because the problem under examination is a ‘top 3 out of 23 objects’ partial ranking problem, the values ‘1’, ‘2’ and ‘3’ have been assigned from the judges to the most, second and third preferred player, respectively. Regarding the less important items, we are going to define the midrank as their rank. Midrank is defined as $\frac{1}{t-q}[(q+1)+..+t]$, where t is the total amount of objects that are about to be ranked and q is the amount of objects that are actually ranked. Thus, in our case, the midrank in all of the 6 datasets can be defined as $\frac{1}{23-3}[(3+1)+\cdots+23]=\frac{270}{20}=13.5$. So, the summary statistics is going to be computed based on the incomplete rankings with their midrank imputations (13.5).

It is necessary to implement this process before moving to the calculation of the statistics, because the measures will provide results that will not have sense, if the computations would include the zero values. Let’s take as example the ranks of Player 1 and Player 2, and assume the Player 1 has been ranked from 10 voters and Player 2 from 200 voters. Because the Player 2 has been preferred much more times than the Player 1, his mean rank is going to be much greater, which implies that he has been ranked lower than Player 2! This is completely wrong and the reason it happens is the fact that, in the calculation of central tendency measures, the votes frequencies of the two players have the same number of rows as numerator, despite the fact that most of the rows of Player 1 contain zero values. To eliminate this phenomenon, we make use of the midranks in the positions of incomplete rankings. Now, the biggest percentage of the numerator of Player 1 will contain the sum of 13.5, instead of zero values, thus his final mean rank is going to be a large value in compare with Player’s 2 mean rank which is going to be much smaller. By implementing this process before the calculations of the descriptive statistics we ensure the credibility of their results. We are going to make use of the imputation of incomplete rankings with the midrank value, also in the construction of visualizations of the partial ranking data and in a clustering method that we implement, afterwards.

3.5 Descriptive Statistics – Application

After presenting the measures that will be used for the exploration of the datasets, we are going to provide results for some of the year of period 2010 – 15. For this purpose, we use the function ‘destat’ from the ‘pmr’ package. Before calling ‘destat’ we transform the input data in an aggregated format, as it is required from the function. To do so, we make use of the ‘unit_to_freq’ function, which constructs the frequency distribution of the distinct observed sequences, from the ‘PLMIX’ package.

The tables below, provide the mean rank for the Top – 6 players, in terms of rankings, for the period 2010 – 15. The tables are sorted by the mean rank value of each player, in descending order. It has to be pointed out, that the position of the players are not the same as at the results. For example, in 2010, Xavi was in the final Top – 3, but as we can see from the table of Year 2010 he is 4th in terms of central tendency. This fact could possibly mean that Sneijder has been ranked more times than Xavi, but Xavi ranked in better positions than Sneijder, which means more points. We evaluate this assumption by looking at the marginal distribution of the ranking matrix for Year 2010 (A.1). We observe from the Matrix that Sneijder – row, column 2 has been ranked first 59 times, second 74 times and third 49 times. On the other hand, Xavi – row, column 7 has been ranked first 88 times, second 51 times and third 36 times. We can notice that, even though Sneijder has been preferred from 7 more voters than Xavi, the Spanish player has received 29 more first - place votes than the Dutch. That is the reason why Sneijder has lower value of mean rank than Xavi, but the second win the third - place in the final results.

Mean Rank for Year 2010

Players	Mean Rank
Messi Lionel	6.825
Iniesta Andres	8.025
Sneijder Wesley	8.551
Xavi	8.642
Forlan Diego	9.88
Ronaldo Cristiano	12.104

Mean Rank for Year 2011

Players	Mean Rank
Messi Lionel	1.989
Ronaldo Cristiano	5.603
Xavi	9.386
Iniesta Andres	10.67
Rooney Wayne	11.997
Suarez Luis	12.53

Mean Rank for Year 2012

Players	Mean Rank
Messi Lionel	2.578
Ronaldo Cristiano	5.247
Iniesta Andres	9.01
Falcao Radamel	11.611
Xavi	11.771
Casillas Iker	12.248

Mean Rank for Year 2013

Players	Mean Rank
Ronaldo Cristiano	4.604
Messi Lionel	5.38
Ribery Franck	6.561
Ibrahimovic Zlatan	11.099
Neymar	12.155
Van Persie Robin	12.716

Mean Rank for Year 2014

Players	Mean Rank
Ronaldo Cristiano	3.716
Messi Lionel	7.496
Neuer Manuel	8.179
Robben Arjen	10.511
Muller Thomas	11.289
Lahm Phillip	12.266

Mean Rank for Year 2015

Players	Mean Rank
Messi Lionel	3.101
Ronaldo Cristiano	4.5
Neymar	9.35
Lewandowski Robert	11.645
Suarez Luis	11.781
Muller Thomas	12.277

Table 3 : Tables of mean ranks for the players in the period 2010 – 15 .

A remarkable fact which can someone observes in the tables above, is the presence of the dipole ‘Lionel Messi – Cristiano Ronaldo’ in 5 out of 6 years of the period under study. The two players are the most preferred in all these years and in some of these years with great difference from the others (2011,2012,2014).

The Year 2011 is the year of Messi’s domination. The Argentinian concentrated the 47.88 % of the total votes. If we take a look at the pairwise frequencies between Messi – row 6 and the other players, for this year (A.2) we observe that the lowest number of voters that preferred Messi against another player is 405 out of 435 and the opponent is Cristiano Ronaldo. This large difference between Messi and the other players, in Year 2011, is displayed in the following bubble plot. The plot represents the mean ranks of the players as bubbles and it uses a different scale and colour, depending on the value of mean rank. As value getting small, the size of the bubble that represents this player getting smaller and its colour getting deep blue. In the x axis of the plot are the mean ranks and in the y axis are the names of the players.

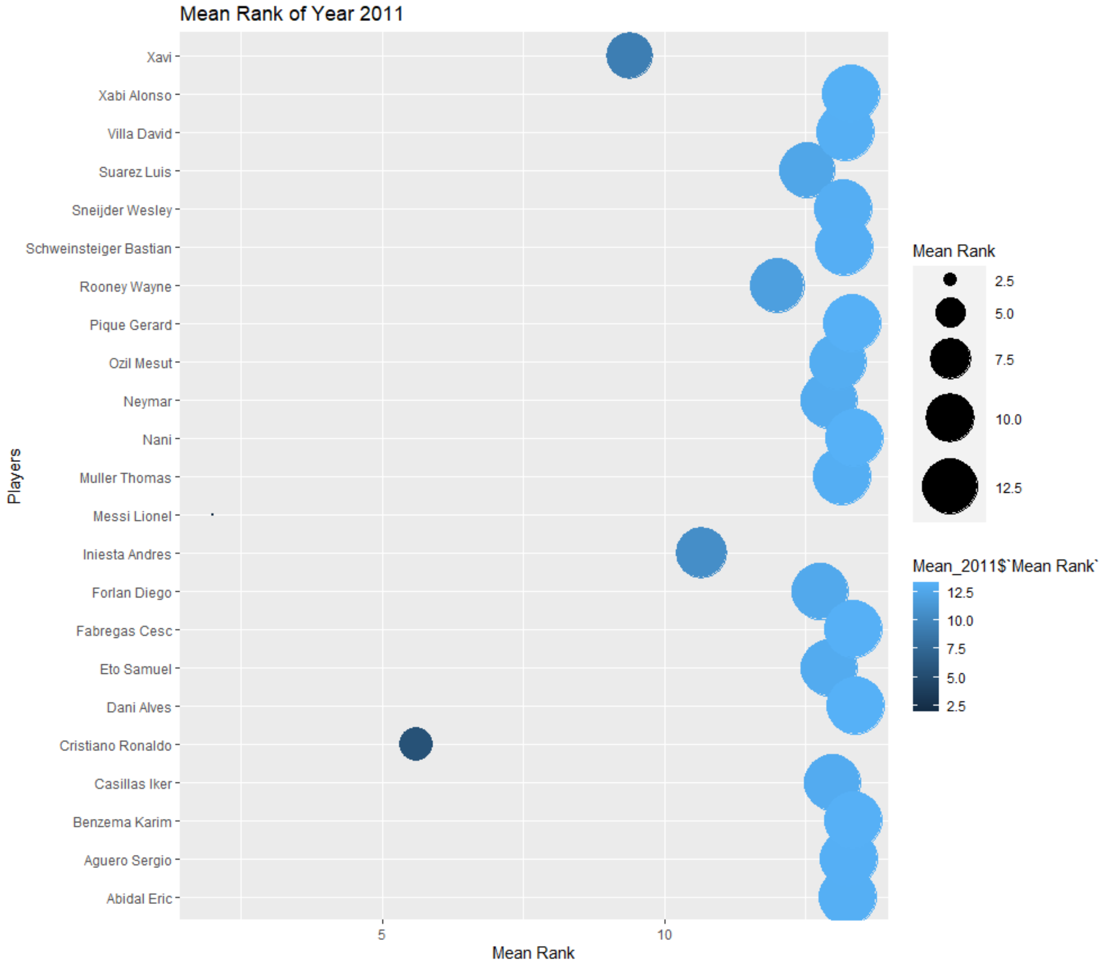


Figure 1 : Bubble plot representing the mean ranks of the players in Year 2011.

As someone can observe from Figure 1, the blue dot that represents Messi can not even be characterized as bubble, because of its very small size, that depicts the tremendous span between Messi and the other players.

Year 2012, does not differ much from 2011, in terms of the winner of the trophy. Lionel Messi wins the award for the third consecutive year and, as in 2011, with large distance from the others. The difference in 2012 is that players that were in the Top-6 in the previous years, like Xavi, Ozil and Casillas, have been ranked in lowest positions.

One could say that 2013 is a very interesting year in the FIFA Ballon d'Or rankings. This is because, after three years of Messi's ascendancy, Cristiano Ronaldo wins the award. It is a year that three players (Ronaldo, Messi, Ribery) are strongly arrogated for gaining the trophy, as we can see from the central tendency of these three players in the table 'Mean Rank of Year 2013'. We can observe that the mean ranks of the Top – 3 players are almost one unit away from each other, in terms of absolute difference. The presence of Frank Ribery in this Top – 3 is not a surprise, if we consider his contribution in the win of Champions League trophy from Bayern Munich, in 2013. It is also observed a big difference of 4.6 units between the mean rank of the third (Frank Ribery) and the fourth (Zlatan Ibrahimovic) player, which does not exist in any of the other years. This is a strong indication, that the first three players have been ranked from the most of the voters. This is ascertained if we take a look at the table of marginal distributions (A.3). It can be noticed that Cristiano has not been preferred from 128 voters, Messi from 162 and Ribery from 222, in the same time that almost all of the other players have not been preferred from more than 500 voters.

In 2014, Ronaldo wins the award again. It is noteworthy the fact that for the first time, in the period under study, there is a goalkeeper in the Top – 3 rankings. Manuel Neuer helped a lot, with his saves, the national team of Germany to win the World Cup of 2014 in Brazil. Moreover, he was awarded with the 'Golden Glove', which is given to the best goalkeeper of the tournament. That's why, it is not a surprise to be in the Top – 3 rankings, for the FIFA Ballon d'Or rankings of this year. From the table containing the marginal distributions we can claim that, despite the fact that Neuer – row, column 1 received more first - place votes (85) than Messi – row, column 5 (55) (A.4) , the Argentinian has a much larger amount on second and third - place votes. Also, Messi has been preferred from more voters, in total, than Neuer. That is the reason that the he has smaller mean rank, thus more dense central tendency, and takes the second - place.

Finally, in 2015, Messi made his comeback by winning the award from Cristiano. Once again, it is outstanding the phenomenon of the total amount of votes that the dipole received. By having a look at the marginal matrix of 2015 (A.5), ones can observe that Messi – row, column 1 has been preferred in the Top – 3 rankings from 425 out of 498 voters and Ronaldo – row, column 4 has been preferred from 386 out of 498 voters. These numbers indicate that, once again, a typical rank in the two first positions is Messi – Ronaldo.

Chapter 4

Visualization Techniques for Ranking data

Visualization of ranking data is an issue of discussion and concern among the statisticians who are involved with the ranking domain. The most basic reason is that the elements of the permutations of the items that are about to be ranked, do not have a natural linear ordering. Thus, traditional methods such as barplots or histograms are not appropriate in this case. Moreover, the size of data in real life examples forbids the drawing of conclusions through descriptive statistics, but a general brief of the data structure. Consequentially, there is a need of visualization methods that tackle such issues and be compatible with the peculiarities of ranking data. These types of methods are going to be presented in these section. More specifically, we are going to describe the permutation polytope method, the multidimensional unfolding technique and the multidimensional preference analysis. Special and more detailed reference is going to be given in the classical – metric and the non – metric multidimensional scaling, since the second is the method that we implement for the visualizations of our data.

First of all, it is important to make clear what questions one expects to answer when implements graphical methods for this kind of data. Types of such questions are :

a) What is the typical ranking of the ranked objects? By typical ranking, we mean the general preference that a ranked object has, in the dataset.

b) How large is the dispersion of votes among the judges? This question is asked in order to provide the agreement among the voters.

c) What are the similarity and dissimilarity among the objects?

So, these are examples of questions that are about to be answered when visualizations of ranking data are implemented. Let's go through these methods in order to understand how they work.

4.1 ***Permutation Polytope***

The idea of using a permutation polytope to visualize ranking data was first proposed by Shulman (1979) [62] and was considered later by McCullagh and Thompson (1993) [46] [69] , who initiated the use of permutation polytopes to display the frequencies of a set of rankings in analogy with histograms for continuous data [45]. If t are the ranked objects, then permutation polytope could be defined as convex hull of $t!$ points in Euclidian space $\mathbb{R}^{t-1}$, which are formed by the set of all $t!$ Rankings [45]. We have to mention at this point that the rankings of these t objects have the ability to be presented as points in the $\mathbb{R}^{t-1}$.

Before explain what we mean by convex hull, we have to define the word convex in terms of geometry.

Definition: A subset of Euclidean space is convex if, with any two points, it contains the whole line segment between these two points and additionally it is able to join them. As line segment, is defined, a part of line that is bounded by two distinct end points and contains every point on the line between its endpoints.

Based on the permutation polytope technique, the frequencies of complete rankings can be visualized as the vertices of such a polytope. On the other hand, partial rankings are represented as a permutation of t non distinct numbers. This is because of the imputation of the incomplete positions.

For example, the top-3 partial ranking (2, 1, - , - , 3), where hyphen denotes a missing position, it can be represented by (2, 1, 4.5, 4.5, 3), and 4.5 is the midrank of this ranking representation ($\frac{1}{t-q}[(q+1)+\cdots+t = \frac{1}{5-3}[(3+1)+5] = 4.5$) . This makes the permutation polytope not applicable for representing partial ranking data. In order to deal with this issue, Thompson (1993) defined a generalized permutation polytope which coordinates are not points in the Euclidian space $\mathbb{R}^{t-1}$, but permutations of t non distinct numbers [45]. In this case, the frequencies of partial rankings can be visualized on the vertices of the generalized permutation polytope.

One could say that this approach could tackle the issue of partial representation, but the drawback in drawing such polytopes is that the generalized permutation polytope has to be drawn in a sphere in a ($t-1$) – dimensional subspace of the set of permutations. In our case, $t = 23$, which means that we have to implement the visualization in a 22-dimensional space, which is not possible to do. In general, despite the fact that the permutation polytopes describe the data from a geometrically point of view, they are not so commonly used because of the difficulties in drawing them.

After navigating the permutation polytope, we are going to present a different class of methods for visualizing ranking data, the Multidimensional class.

4.2 ***Multidimensional Methods***

Multidimensional Scaling or MDS is a big family of graphical methods for representing data which are in the form of measures that provide the proximity or "closeness" between each pair of objects. Examples of such measures are similarity or dissimilarity measures. The basic idea behind MDS, is to search for a low – dimensional space, usually Euclidean, in which each object is represented by a point in the space, such that the distances between the points match as well as possible with the original dissimilarities [45]. Thus, the goal is to find points in a low – dimensional space, that can represent the distances in such a way. This is the main issue in the multidimensional techniques for visualizing such types of data, like ranking data.

Many approaches have been developed, in order to deal with this issue. One approach is to address it like an optimization problem and found the values to formulate the MDS, by minimizing a loss function which is called stress value. We are going to define later on the stress value and the information it provides. Such methods that work with stress value are the metric – classic multidimensional scaling and the non – metric multidimensional scaling, which will be discussed in further section. Another approach that is suitable for the context of ranking data is Kidwell's approach. Kidwell (2008) [37] suggested to use the Kendall distance for the computation of the dissimilarity between two rankings, complete or partial, and then to apply MDS in order to find an integration of a dataset of n rankings assigned by n judges in a two – or three – dimensional space.

Now, let's have a look at the fundamental multidimensional techniques for visualizing ranking data, by presenting in a more explanatory way the metric and non – metric multidimensional scaling methods, which will be implemented for the visualizations of FIFA's datasets.

4.2.1 <u>The Multidimensional Unfolding technique</u>

The unfolding technique was first formulated by Coombs, in 1950 [16], and it belongs to the family of Multidimensional Scaling techniques, for representing ranking data. We are going to see later on the Multidimensional Scaling techniques, in detail, but at this point we will focus on the unfolding one.

The method attempts to visualize a set of points in a low Euclidean space with both judges and objects being represented by the points in the same space [45]. This is the main difference with the other methods that belong to the MDS family, which attempt to visualize only the set of judge points in a low-dimensional Euclidean space. The points that are used for the representation of the rankings are obtained in such a way that the ranked order of the distances from a point representing a judge to the points representing the objects, match as close as possible with the actual rankings that have been assigned from the judge to the objects. Thus, based on the Euclidean distance

$$d_{ij} = \sqrt{(x_i - x_j)'(x_i - x_j)}\,,$$

the goal of multidimensional unfolding method is to find x and y, such that the distances match as much as possible with the ranks of objects given by the judges.

In the case when $d_{ij} = 1$, the unfolding becomes unidimensional unfolding for which objects and judges are represented by points on a straight line. The technique's name, unfolding, has been termed because of the fact that when this straight line is folded from one side to the other side at any judge point, the judge's rankings can be observed.

4.2.2 <u>Multidimensional Preference Analysis</u>

The Multidimensional Preference Analysis method or MDPREF, was introduced by Carroll (1972) [12]. Its fundamental idea is similar to the Multidimensional Unfolding technique. Thus it displays the relationships between judges and the ranked items by reducing the dimensionality of the data, while retaining the main features as many as possible. Moreover, like in MDU, the MDPREF method assumes that the ranking assigned by each judge can be represented in terms of the ordering of distances and projections. The difference with the MDU is that, in the MDPREF's case, the points are replaced from vectors for the representation of judges, in a low dimensional space. The objects are represented, as in previous, by points in the same space. As in all MDS methods, the vectors – judges and the points – objects are chosen in such a way that the projections of the objects to the vector of judges is as closely as possible with the actual rankings of the judges.

4.2.3 <u>Classical – Metric & Non – Metric Multidimensional Scaling</u>

As it has been referred in above section, these are two of the most basic approaches that try to solve the MDS issue. The first is called Metric or Classical, because it attempts to reproduce the original metric or distances of the rankings. The second technique, is called Non – Metric and assumes that only the ranks of the distances are known and not the actual distances. Thus, the Non – Metric approach creates a map which tries to reproduce these ranks.

We are going to explore the two methods separately, starting from the Metric approach.

4.2.3.1 <u>Classical – Metric Multidimensional Scaling</u>

The classical MDS procedures were first introduced from Torgerson (1952) [71]. According to him, the goal in these procedures is to compute a distance matrix which is going to approximate the interpoint distances of a configuration of points X in a low – dimensional space. The interpoint distance is normally taken to be the Euclidean distance, but sometimes we may use the Manhattan distance

$$d(x_1, x_2) = \sum_{j=1}^{p} | x_{1j} - x_{2j} | \, .$$

The classical solution is optimal in the least square sense. That means that when the distance matrix that is used is Euclidean, the solution that is obtained minimizes the sum of squared differences between the elements of the distance matrix. In other words, we could say that the solution minimizes the value of a loss function, called stress value.

Stress value is a goodness-of-fit statistic, for the MDS models, which is based on the differences between the actual distances and their predicted values. The way that the stress value is calculated differs between metric and non - metric approach. For the classical approach, the stress is calculated from the following formula :

$$stress = \sqrt{\frac{\sum (d_{ij} - \widehat{d_{ij}})^2}{\sum {d_{ij}}^2}} \, ,$$

where d_{ij} is the actual distance and $\widehat{d_{ij}}$ is the predicted distance between two points, based on the MDS model. In the case of metric approach, the predicted values depend on the number of dimensions kept and the distance that is used for the calculation of the measure.

We have to mention at this point that the optimal solution always has to be sought in terms of balance between accuracy and parsimony. Thus, an equilibrium point between the smallest stress value and the number of dimensions that are possible for interpretation, has to be reached. It is obvious that, when the dimensionality increases the stress value decreases but, in parallel, the ability of interpretation decreases also.

In order to be able to assess the fit of an MDS model, by interpreting the output of the model's stress values, Kruskal (1964) [38] released a paper which contained a table about the interpretation of stress values, in terms of goodness-of-fit purposes, based on his experience. The table provides the information that stress values below 0.05 indicate a very good fit of the model (0.05 → Good, 0.025 → Excellent, 0 → Perfect) and values above 0.05 indicate a not so good fit of the MDS model, as the values increase (0.1 → Fair , 0.2 → Poor). Kruskal's paper faced backlashes from recent articles, which mentioned that acceptable values of stress depend only from the quality of the distance matrix and the number of objects that are ranked in the matrix [45].

Another way, to check how well the MDS model produces the predicted values in compare with the actual values is the Shepard diagram. The Shepard diagram, like the stress values, can be implemented in both metric and non – metric case.

The Shepard diagram is a scatterplot of the distances between points in the MDS plot against the observed dissimilarities (or similarities). The points in the plot should adhere to a curve or straight line. The plot compares how far apart are the data points before and after the transformation in a scatterplot. A completely straight line is a strong indication that the fitting of the points in a lower dimensional space through MDS is accurate. However, in real life examples since a lot of the information that the data carry is lost during the dimension reduction, Shepard diagrams rarely look completely straight.

4.2.3.2 Non – Metric Multidimensional Scaling

In the above section the classical MDS solution was presented, which assumes that the configuration of points is an Euclidean distance matrix. However, in real life cases, it is more often to use less strict assumptions between the true distances and the observed distances. In such cases, an error parameter is added in order to denote the distortions. Moreover, the distribution is assumed to be unknown and monotonically increasing function. Because of these two reasons, instead of using the actual numerical values of the dissimilarities, the rank order of the dissimilarities between the objects is used.

When the Non – Metric Multidimensional Scaling (NMDS) is used, the configuration between the points is a dissimilarity matrix and not an actual distance matrix. The main difference between the two matrices is that dissimilarity matrices do not require their values (dissimilarities) to be symmetric, in compare with the distance matrices which require for the differences they store to by symmetric.

Thus, the NMDS can be defined as an indirect gradient analysis approach, which produces an ordination based on a dissimilarity matrix [45]. As the classic MDS, the technique attempts to represent as closely as possible the pairwise dissimilarity between objects in a low – dimensional space, but this time, in terms of rank – based approach. It is a robust technique, with tolerance in missing pairwise distances. Also, it is able to use quantitative, qualitative or mixed data. We could claim that, NMDS is more appropriate in our case, because of the fact that it should be used when ordering is more appropriate than actual distances. Moreover, it is preferred when the dataset or the number of ranked items in the dataset is large.

As it was referred in the above section, NMDS makes use of stress values and Shepard diagrams for assessing the goodness of fit. The difference with the classic MDS exists in the formula that the two approaches use for the calculation of stress values. For the non – metric approach, the stress is calculated from the following formula :

$$stress = \sqrt{\frac{\sum(f(x)-d)^2}{\sum d^2}},$$

where x denotes the vector of proximities, $f(x)$ a monotonic transformation of x, and d the point distances. The object of NMDS is the same as the metric MDS, to found the coordinates that minimize the stress function. On the other hand, like in MDS, the increase of dimensionality leads to decrease of interpretation capability. Low – dimensional projections are often better to interpret and are preferable for interpretation issues. Thus, an equilibrium point has to be found, between the goof fit of the original dissimilarities and the interpretation of the dimensions.

At this point, after capturing the theory of the basic methods for visualizing ranking data we are going to present some fundamental R libraries and functions that are appropriate for visualizing ranking data. Furthermore, we will capture the process of implementing visualizations through these functions. At last, visualizations of the ranking datasets that are processed in this thesis, are going to be presented.

4.3 ***Non – Metric Multidimensional Scaling for Ranking data -Application***

We are going to apply Non – Metric Multidimensional Scaling, in order to visualize the ranking matrices that have been obtained from FIFA Ballon d’Or voting datasets, for the period 2010 – 2015. The reason we apply NMDS is that, as have been discussed in the previous chapter, this technique is resorted to when the data are of type that have been observed on a scale (categorical, ranking, etc.) and also, in the case that the ranking between the observations is the important to be computed and not the actual differences. We will provide the results for the year 2010, as an application example of partial rankings visualization, in order to explain the methods and the results.

Before starting the visualization process, the appropriate distance which will be used for the computation of the distance matrix, has to be selected. It is important to choose the correct distance in order to calculate the matrix because the technique is sensitive in the distance that is chosen. It has to be pointed out that, as in the descriptive statistics chapter, the zero values have been replaced by the midrank value, which is 13.5 .

4.3.1 Kendall distance for computing the distance matrix of the rankings

As a starting point, in order to implement NMDS, one has to calculate the distance matrix between the rankings. Thus, the appropriate distance for the calculation of the matrix has to be chosen. One thought could be the Euclidean distance, but it is not a preferable measure when high dimensional data (e.g. partial ranking scores) are analyzed due to the distance's sensitivity to noise of such data. A calculation of the FIFA matrices dissimilarities, based on the Euclidean distance, won't provide a good description of the data, since the rows – rankings containing a bunch of zeros, in our case the whole dataset, will be similar to each other without pointing out the real dissimilarity due to the computational properties of Euclidean distance.

Another option, could be the, commonly used in the context of visualization of high dimensional data, Bray – Curtis distance [8]. The Bray – Curtis or Sorensen distance is a distance measure commonly used in botanology, ecology and environmental sciences. It is a modified Manhattan measurement, where the summed differences between the variables are standardized by the summed variables of the objects. Bray – Curtis is a powerful dissimilarity measure when there is an abundance of dimensions. However, if the objects that are measured are in zero coordinates, the distance is undefined. Zero values, have been imputed by 13.5 and indicate the voter's non – preference for a player. Despite the fact that zero values, have been imputed by the midrank value in the calculation of the distance matrix, we are not going to use Bray – Curtis. Thus, we are going to look for a more rank – based distance which will be able to represent in a more representative way the partial ranks.

Thompson (1993), discovered that Kendall's and Spearman's distance are very powerful in measuring the distances between two rankings. This is due to the fact that these distances are able to provide natural geometric interpretation of the rankings. More specifically, during the initial implementation processes of the permutation polytope, he showed that the minimum number of edges that must be drawn to get from one vertex of a permutation polytope to another reflects the Kendall distance between the two rankings labeled by the two vertices [45]. Furthermore, he showed that the Euclidean distance between any two vertices of a polytope is proportional to the Spearman distance between the two rankings corresponding to the two vertices. Thus, these two distances can be considered as the most suitable for computing the distance matrix of ranking data. Both distances could be used in the computation of distance matrix but, Cabilo and Tiley (1999) [10] observed that when there were no missing observations, Spearman's distance was more powerful than Kendall's. On the other hand, in the incomplete case Kendall's statistic is much more strong, in terms of accuracy of distance calculations and detection of patterns. Thus, Kendall's distance is going to be used for the computation of the distance matrices of our datasets. Based on the fact that, the Kendall's distance counts the pairwise disagreements between the voters in the datasets of FIFA, the pair can be characterized as dissimilar, in terms of preference for the best players, if the computed distance is large in compare with other distances. Thus, the larger the distance between a pair of voters, the more dissimilar preference have these two voters. We are going to give a detailed description of Kendall's distance in further section, where it is going to be used for clustering purposes.

4.3.2 Non – Metric Multidimensional Scaling Process

The starting point of a Non – Metric Multidimensional Scaling Process, is the computation of distance matrix between the rankings. For the reasons explained in the previous section, the optimal distance choice in our case is Kendall distance. Thus, in order to construct it, we make use of the function ‘Dist’ from the ‘amap’ package. The function computes the distance matrix, by taking as input the matrix for which the distances are going to computed and a specified distance.

Choose the number of dimensions

After the computation of the distance matrix, the next step is to determine the number of ordination dimensions with which we are going to present the data. Thus, we plot the stress value for a number of tested dimensions, in order to obtain the optimal number. Such plots are called stress plots or scree plots. Stress plots show the decrease in ordination stress with an increase in the number of ordination dimensions. As it has been discussed in a previous section, stress value depends on dimensionality and it is decreasing with increasing dimensionality. However, higher dimensionality leads to incapability in interpretation, because low – dimensional are often better to interpret and so preferable for interpretation issues. Thus, a stress plot explores both dimensionality and interpretative value and provides dimension – dependent estimations which give indices for meaningful stress reduction in increasing dimensionality.

In order to construct the stress plot, for the year 2010, we are going to call the ‘dimcheckMDS’ function, from the ‘goeveg’ package. This function provides a plot of stress values for a given number of tested dimensions in NMDS. We will make use of the default choice for the tested dimensions, which is 6. The function takes as input, the computed distance matrix and the distance which was used for the computation of the distance matrix. For the year 2010, the stress plot is the following :

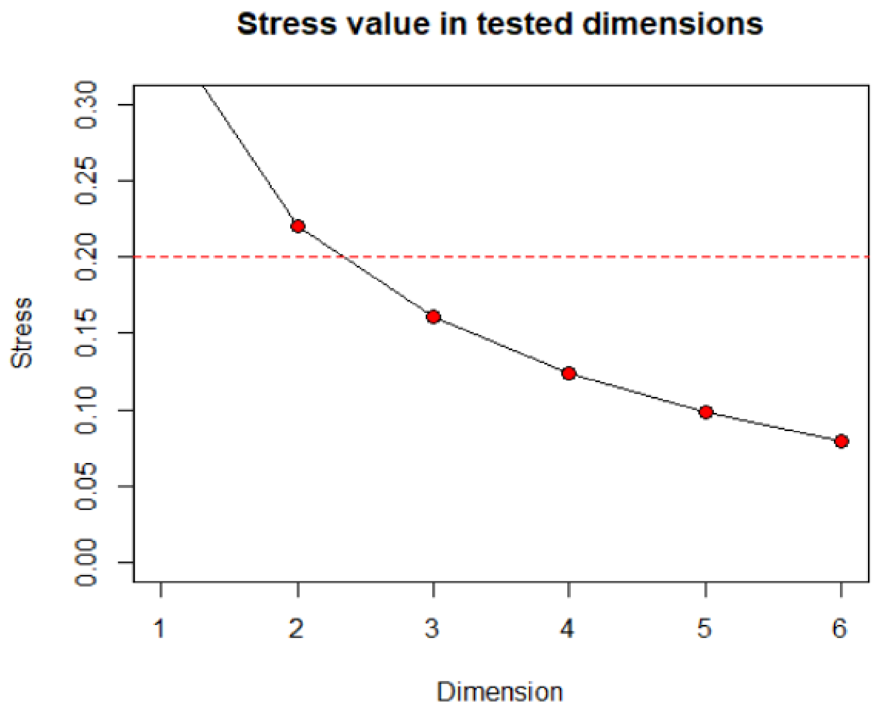


Figure 2 : Stress plot of 6 tested dimensions for the year 2010.

From the stress plot obtained, we observe that stress value surpasses the threshold of 0.20 (when the fit can be characterized as poor but acceptable), in 3 dimensions. Moreover, the absolute differences between the stress values, as the dimensions increase, are not as large as the absolute difference (0.065) from 2 to 3 dimensions. Also, a visualization with 3 dimensions can be interpreted in a more sufficient way, in contrast to visualizations which are constructed with more dimensions. Thus, this is a strong indication to implement NMDS with 3 dimensions.

Implement NMDS with the 'metaMDS' function

After choosing the number of dimensions, let's take a brief look to the way the NMDS works. The algorithm begins by constructing an initial configurations of the samples in the k dimensions. The initial configuration could be based on another ordination or it could consist of an entirely random placement of the samples [67]. The final ordination is partly dependent on the initial configuration, so a variety of approaches are used to avoid the issue of local minimum. One of the approaches is to perform several ordinations, each starting from a different random placement of points, and select the ordination with the best fit [67]. This is how the function that we use for the implementation of NMDS works.

The function's name is 'metaMDS' and is called from the 'vegan' package. It is a wrapper function, that calls several other functions to implement Non – Metric Multidimensional Scaling into one command. The function performs NMDS and tries to find a stable solution using several random starts. For this purpose, it calls the 'monoMDS' function from the same package. It performs several other jobs, but we focus on the implementation of the approach discussed in the previous paragraph. The process of the NMDS by this approach is iterative and can be described as follows. The strategy of ‘metaMDS’ is to first run NMDS with the result of metric scaling as the starting value (it can reach a good solution but often close to the local minimum), or use the option to start searches from a previous solution and take it as a standard. We choose the second option by setting previous.best = True. Then, ‘metaMDs’ starts NMDS from several random starts. If a solution has lower stress than the previous standard, it is taken as the new standard. If the solution is better or close to a standard, ‘metaMDS’ compares these two solutions by the use of Procrustes analysis. Procrustes analysis, is a statistical method which compares a collection of multidimensional shapes by attempting to transform them into a state of maximal superimposition. It does so by attempting to minimise the sum of squared distances between corresponding points in each shape through rotation of their coordinate matrices. It is commonly used in ordination techniques such as NMDS, PCA,etc. Back in the NMDS algorithm, if the reached solutions are very similar in their Procrustes rmse and the largest residual is very small, the solutions are regarded as convergent and the better one is taken as the new standard. By this way, the ‘metaMDS’ finds a stable solution and avoids to stuck on a local maximum.

Assessment of the NMDS result

After performing NMDS to 3 dimensions in the distance matrix obtained by the use of the Kendall distance for the partial ranking matrix, of year 2010, we are going to evaluate the result of the ordination algorithm. For this purpose, the Shepard diagram is used.

The Shepard diagram, represents the actual or transformed proximities versus the predicted proximities. It is a scatterplot of distances between data points. In other way, it could be described as a plot of ordination distances, in the y axis, and monotone or linear fit, line against original dissimilarities, in the x axis. It is analogous to an Actual by Predicted plot, which is a typical plot of the actual response versus the predicted response. Ideally, the points, which are shown in blue, fall on the Y = X line, which is shown in red. The Shepard diagram, also displays two statistics for the evaluation of the fit of the graph. The linear fit is the squared correlation between the fitted values and ordination distances, and the Non – metric fit is based on the stress value and is defined as $R^2 = 1 - S^2$, where S is the obtained stress value. High values in the nonmetric fit indicates high correlation between the observed dissimilarities and the ordination distances. The Shepard plot for the result of the NMDS for year 2010 is the following :

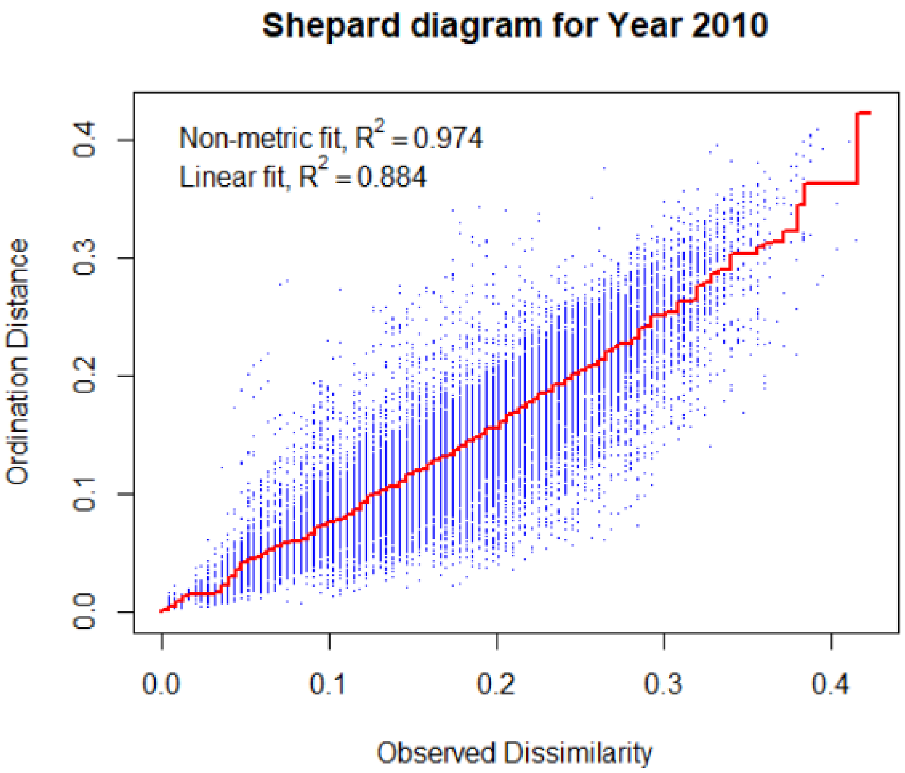


Figure 3 : Shepard diagram for the evaluation of Year's 2010 NMDS result.

As we can observe from the Shepard diagram, the points do not fall exactly on the red line, but are not noticed big deviations from the line. Moreover, the Non – metric fit is near 1 (97.4 %) which indicates high rank correlation between the observed dissimilarities and the ordination distances, thus, good fit of NMDS. We could claim that this confirms the fact that NMDS maintain the distance ranking.

4.3.3 Visualizations based on the results of NMDS

After obtaining the NMDS results for year 2010, we are going to implement visualizations of the ranking data in the dimensional – space, that was reached from the results of NMDS. Thus, because the number of ordination dimensions is 3, we will construct 3D plots, in order to explore the data visually. The packages that are used for the implementation of the 3d visualizations are: the 'MASS' package, the 'vegan3d' package, the 'rgl' package and the 'scatterplot3d' package.

Before beginning the constructions of the visualizations, we have to make groundwork for some of the input parameters that will be used in the functions. First of all, we create references for the resulting points of each one of the 3 dimensions, in order to be used as input data in the implementation of the visuals. The other step of the preprocessing, is to convert the columns – Players of the input matrix which contains the imputed partial rankings to factors and their rankings as levels, in order to be able to make use of colors based on the ranks. After making these steps we are ready to present some visualizations.

The first visualization we provide is a 3D Scatterplot which represents the ranks of Messi by different color, based on the ranks. The plot was constructed through the function ‘plot3d’ of the ‘rgl’ package.

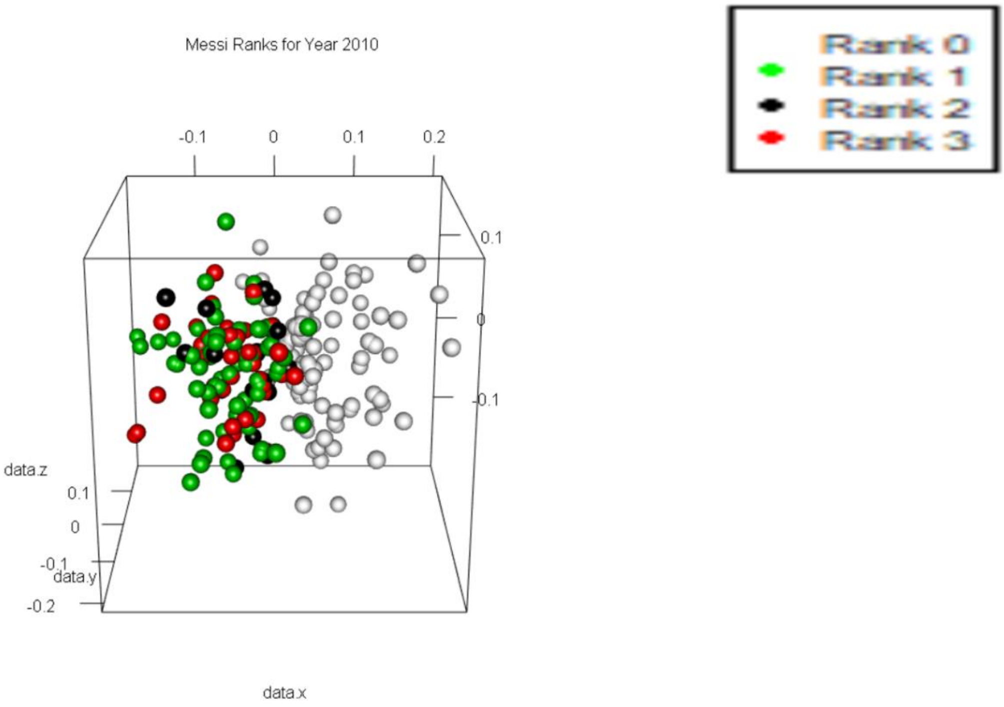


Figure 4 : Static 3D Scatterplot of Messi ranks for year 2010.

As we can observe, in the scatterplot exist white, green, black and red spheres. The white spheres indicate the voters that did not preferred Messi in their Top -3 rankings, the green spheres indicate those who voted him first, the black spheres the judges that voted him second and the red spheres those who preferred him third. By looking at the plot, someone can notice than more than half of the voters have preferred Messi at least one time, because colored spheres appear to be little more than whites. This can be confirmed by checking at the original dataset, where 241 out of 425 judges have put Messi in their Top – 3, at least one time. After taking a closer look at the plot it can be marked that, the green spheres are much more than the red or black spheres. In addition, the red spheres appear to be more than the blacks. Thus, from the total of the voters that rank Messi, the majority ranked him first, a big proportion of voters ranked him second and a little amount of them ranked him third. In order to understand, the importance of the big amount of first votes in Messi’s plot, we are going to provide the same plot for the player that has been ranked second in the final rankings, Andres Iniesta.

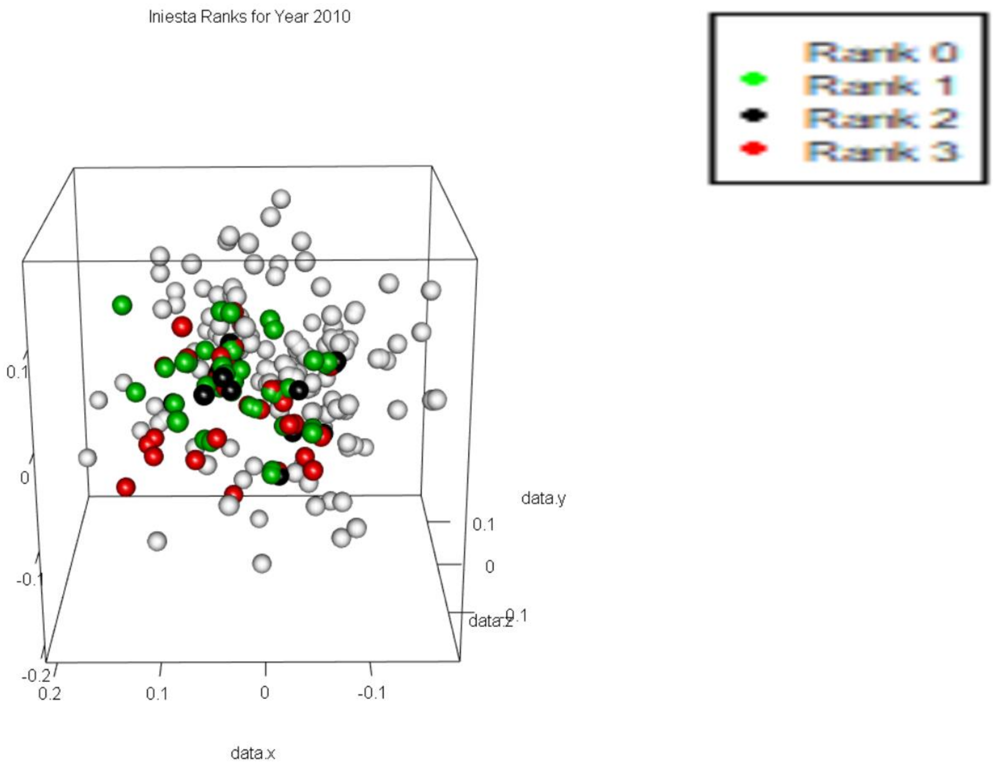


Figure 5 : Static 3D Scatterplot of Iniesta ranks for year 2010.

By taking a careful look at Figure 5, we can observe that the white spheres are more than those in the scatterplot of Messi, which means that more voters preferred Messi in their Top – 3 than Iniesta. The proportion of first, second and third position votes don't seem to differ much in compare with Messi's votes. We notice a large amount of green spheres, in compare with the amount of red and black. The big difference between the two 3D scatterplots is the fact that the one representing the votes that Messi received, has lesser white spheres than the one representing Iniesta's votes. In addition, if we take into account that the proportion of the 1st, 2nd and 3rd ranks between the two players do not differ much, we can claim that Messi wins the trophy because of the numerous amount of voters that preferred him in their top – 3 and not because of the number of first position votes he received, in compare with the other players.

It has to be mentioned at this point that, if someone wants to take a better vision of the actual positions of every sphere in the above scatterplots, the interactive version of the static plots can be used. We observe from the two plots that there are some spheres that are very closed to each other, which may lead to not accurate conclusions if someone does not observe the plots in a careful manner. That is why the interactive 3D scatterplots are also proposed, for a reader who wants to change the orientation of the plots, zoom on it and get a real feeling of the 3D visual. In that case, the functions 'play3d' and 'spin3d' from the 'rgl' package can be used, which allow to reset the viewpoint for a specific number of seconds, set by the user.

Another interesting thing to visualize is the difference in votes between Lionel Messi and Cristiano Ronaldo. The year 2010 can be considered as the beginning of the big rivalry between these two players, in terms of FIFA Ballon d’Or, because Messi won his inaugural award. Having won 11 FIFA Ballon d’Or awards (6 for Messi and 5 for Cristiano), both are widely regarded as the two greatest players of all time. Thus, it would be interesting to have a look at the votes that these two players have received in a different plot than the previous one. Thus, we are going to construct 3d Scatterplots, where the ranks of each player are going to be represented with points. Also, a bar is added in each point in order to visualize the amount of the total ranks and colour of each point in a clearer way. It has to be mentioned that, these plots do not show the votes of the total amount of judges, as in previous, but they provide only the preferences that were in the Top – 3 of the voters, in order to make more clear the difference between the amount of votes and the kind of votes each player received.

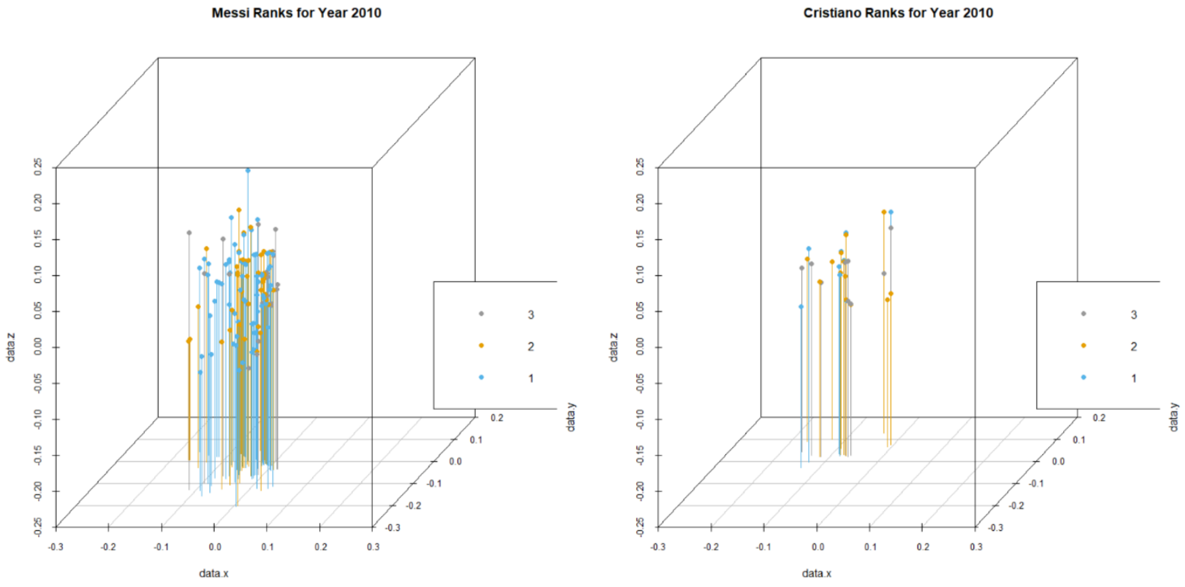


Figure 6 : 3D Scatterplots with bars, comparing Messi and Cristiano ranks for year 2010.

Point and bar with grey colour indicates a 3rd place vote, yellow indicates a 2nd place vote and blue indicates a 1st place vote. The plots were implemented through the function ‘scatterplot3d’ from the homonym package.

From the Figure 6, we can observe that the preference of the voters is clear for the question ‘Who is the best’ for the year 2010. The bars of Messi are far away more than Cristiano’s bars, which demonstrates the fact that Cristiano has been preferred from a very small amount of voters in their Top – 3 players list. Moreover, in Ronaldo’s scatterplot one can notice the presence of many yellow bars, in proportion to the total amount of bars, which implies that those who preferred Cristiano have ranked him in the third - place mostly. On the other hand, Messi has a lot of blue bars and the yellow bars follow, in terms of amount. Thus, this comparison helps to come to a conclusion. It is not only the fact that Messi is the winner of the award in this year, but this wide margin between those two in this path of their career is remarkable. It is a confirming indication of how far away were the two players in their early career stages, in terms of quality, if we also count that Messi won the award for the next two consecutive years. One the other hand, it is strong demonstration of the hard work Cristiano has done, in order to reach the class of the Argentinian and surpass him for some years.

Chapter 5

Cluster Analysis on the Data with Bayesian mixture of Plackett – Luce models

In this chapter, the problem of clustering partial ranking data, has been approached by a Bayesian point of view. In specific, we are going to present a Bayesian finite mixture of Plackett – Luce model, in order to deal with the partial ranked data. Inference is conducted with the combination of the Expectation – Maximization (EM) algorithm for the maximum a posteriori estimation and the Gibbs sampling iterative procedure. The implementation of this approach contains a data augmentation step, with the latent group structure, which allows for approaching the partial top – ordering by a model based aspect. Recent works considering Bayesian mixture modelling based on the PL are Gormley and Murphy (2008) [27], who deal with a grade of membership model where, at each stage of the sequential ranking process, each sample unit has a specific partial membership of each component [49]. Also, Caron et al. (2014), extended their initial work which have been implemented in 2012, and was a Bayesian nonparametric PL based on a Gamma process to account for infinite number of items, to the mixture context in order to cluster partially ranking data. The goal of Caron et al. work was to identify and characterize possible group of rankers with similar preferences/attitude. We could say that the approach that is being presented in this chapter is very similar with the Caron et al. extension, but two main differences are spotted. The first difference is that in the parametric setting of our model, each single component is a standard PL for finite orderings whereas in the Caron et al. approach the ordering of the number of items that are modelled is random [49]. The second difference is that the cardinality of the mixture models, in our case, is finite whereas on the other model is infinite [49]. A fundamental pros of this model in comparison with the MLE frequentist approach, is the ability of addressing the estimation uncertainty in a straightforward way, without relying on large sample approximations. Furthermore, it is much more efficient in terms of the computational time needed for the implementation of the whole process.

5.1 <u>Theoretical Framework of the Method</u>

5.1.1 <u>The *Plackett – Luce model*</u>

The Plackett – Luce model is one of the most popular and frequently applied parametric distributions to analyse rankings of a finite set of items. Also, it is one of the most successful stagewise models for analysing partial ranking data. The model's process could be summarized as a random sampling without replacement from an urn, where at each stage the most – liked item is specified among the alternatives not selected at the previous stages.

The model depends on the Luce's axiom of choice (Luce 1959) [40], which states that the odds of choosing an item over another do not depend on the set of items from which the choice is made [72]. At first, one assumes that there is a set S of K items, $S = \{i_1, i_2, \dots , i_k\}$. Then, under the Luce's axiom, the probability of selecting a k item from S is given by

$$P(k|S) = \frac{w_k}{\sum_{i\in S} w_i},$$

where w_i represents the 'worth' of item i, in terms of ordering. In Plackett – Luce model the ranking of these K items, can be viewed as a sequence of choices, where first is chosen the item with the biggest 'worth' among the items, then is chosen the second ranked item from the remaining items and this process is iterated until all the items being ranked. The 'worth' of each item of the set S is represented by the corresponding support parameter p_i which belongs to the set of the support parameters $p = (p_{i_1}, .., p_{i_k})$, that parametrize the PL model. These support parameters represent a positive constant associated to each item and the higher the value of the support parameter of an item the greater the probability for this item to be preferred at the selection stage. For the final ordering of the items, the probability of the ranking $i_1 > i_2 > \cdots > i_k$ is equal to $\prod_{k=1}^{K} \frac{a_{i_k}}{\sum_{i\in A_k} a_i}$, where A_k is the set of alternatives $\{i_k, i_{k+1}, \dots, i_K\}$, from which the item k is chosen [72].

It can be observed that the ranking probability under such a model can be expressed as a function of top – choice probabilities only. Such samples, that occur from picking one item at a time, out of a set of choice probabilities that satisfy the Luce's axiom, provide a total ordering of items which can be considered as samples from a distribution over all possible orderings. The form of such a distribution was first considered by Plackett (1975) in order to model probabilities in a horse race [33]. Thus, the name of the model has been derived from the independent work by Luce (1959) [40] and Plackett (1975) [57].

At this point, it has to be remarked that, the Luce model satisfies the Independence of Irrelevant Alternatives (IIA) property (Tversky, 1972 [35]) which, in simple words, claims that the choice of a judge between two objects , depends on the preferences of the judge to these objects only and it is irrelevant to the preference on another object. But IIA is not such a good property because, it ignores the fact that a preference of a judge to an object is natural to depend on judge's preference on similar objects. Thus, by ignoring this fact, the estimations of the choice probabilities, that the model obtains, are expected not to be unbiased.

5.1.2 Model's Specification

Based on the PL model that has been discussed above, we are going to conduct inference through the Bayesian approach that was introduced from Caron and Doucet, in 2012 [11]. As it was referred in the introduction, the method has a fundamental step which is the data augmentation step. The data augmentation step, where the Bayesian perspective is taken into account, can be described as follows : Let's suppose that we get a random sample ($\underline{\pi}^{-1}$), of partial top orderings, drawn from a G – component PL mixture. The representation of the sample in symbols is

$$\pi^{-1}{}_1, \dots, \pi^{-1}{}_N | \underline{p}, \underline{\omega} \sim \sum_{g=1}^{G} \omega_g \, P_{PL}(\pi_s{}^{-1} | \underline{p}_g),$$

where $\underline{\pi}^{-1} = \{\pi^{-1}{}_\kappa\}_{s=1}^{N}$ represents a random sample consisting of N partial top orderings of the form $\pi_s^{-1} = (\pi_s^{-1}(1), \dots, \pi_s^{-1}(n_s))$ [45] . The parameter n_s represents the number of objects ranked by the unit s in the top n_s positions. Thus, n_s is going to be always 3 since 3 out of 23 players are ranked from all of the judges of the datasets. The term p_g, is the support parameter vector of the g-th PL mixture components. Thus, the term $P_{PL}(\pi_s{}^{-1} | p_g)$, is the PL likelihood function of the N top partial orderings given the corresponding support parameters. This term multiplied by ω_g, which represents the corresponding weight, is summed for every component of the G – component PL mixture.

After defining the sample of top partial orderings, we introduce the latent feature z_{sg} , which receives the value 1 if the unit s belongs to the g-th mixture component and the value 0 otherwise. Since, the latent feature follows a Bernoulli distribution, then the vector $\underline{z}_s = (z_{s1}, \dots, z_{sG})$, which represents the values of these features for each one of the G values, follows a multinomial distribution with the same weight for each component because the number of ranked players is the same for each judge. These latent features represent the unobserved group labels for each group of judges that probably exists in the dataset. Thus, we include the unobserved group labels $\underline{z}_s$ in such a way so that the labels determine the cluster – specific support parameters on the underlying quantitative variables of the model. These underlying quantitative variables represent the observed variables of the model given by the N partial top orderings ($\underline{\pi}^{-1}$), the unobserved group labels ($\underline{z}$), the support parameters ($\underline{p}$) and the corresponding weights ($\underline{\omega}$). Because there exist both observed and latent variables in the model, we have to elicit the joint prior distribution for the unknown parameters. The most straightforward way is to choose prior distributions with independent support parameters and weights, so that $f\left(\underline{p}, \underline{\omega}\right) = f\left(\underline{p}\right) f(\underline{\omega})$, which can be calculated. In our case, we are going to recover the MLE approach as a special case of the non – informative Bayesian approach, by using flat priors, which means that the priors are going to have negligible information.

5.1.3 MAP Estimation through EM algorithm

After introducing the unobserved group labels, we construct an EM algorithm in order to discover the posterior mode (MAP estimate) and, in general, to optimize the posterior distribution. The EM algorithm was originally introduced by Dempster et al. (1977)

[20] and since then it has been the subject of great research. In general, it is an iterative maximum likelihood procedure, which is usually used in order for the parameters of a mixture model to be estimated. Theoretically, increases in the likelihood function are guaranteed as the algorithm iteratively improves upon previously derived parameter estimates. The iterative procedure is considered to converge when all parameters become stable and no further upgrades can be made to the likelihood value.

The implementation of the EM algorithm includes the iteration of Expectation (E) step and Maximization (M) step. The E – step, which relies on the conditional joint distribution of all the latent variables is given by

$$P(\underline{y}, \underline{z} \mid \underline{\pi}^{-1}, \underline{p}, \underline{\omega}) = f(\underline{y} \mid \underline{\pi}^{-1}, \underline{z}, \underline{p}, \underline{\omega})\, P(\underline{z} \mid \underline{\pi}^{-1}, \underline{p}, \underline{\omega})$$

and returns the objective function with respect to the support parameter and the corresponding weight, where the posterior membership probabilities $\hat{z}_{sg}$ are obtained after the proper calculations [29]. The M – step maximizes the proper objective function each time, with respect to $\left(\underline{p}, \underline{\omega}\right)$. The abiding differentiation of the objective function, with respect to each support parameter of the g – th mixture model, yields the updated support parameters of the M – step. Also, the same process of optimization of the objective function yields the updated mixture weights, with respect to the corresponding weights of the g – th mixture model and the constraint $\sum_{g=1}^{G} \omega_g = 1$. The E- and M- step are repeated alternatively, until there is no further improvement in the likelihood value.

After obtaining the MAP estimations we implement the process of Gibbs sampling in order to learn about the uncertainty associated to the final estimates by drawing a sample from the joint posterior distribution. This is achieved by deriving the full joint density and the posterior conditionals for each of the random variables in the model and simulating samples from the posterior joint distribution based on these posterior conditionals. The Gibbs sampling algorithm is going to be presented in the section of the theoretical framework of the Insertion Sorting Rank method.

5.1.4 Determing the number of components

After performing a separate inference on PL mixtures on different number of components, we are going to choose the model that satisfies in better way among the competing models the corresponding criteria. The Bayesian criteria that are used for the selection of the best model are the Deviance Information Criterion or DIC (Spiegelhalter et al. , 2002) [64], the Bayesian Information Criterion – Monte Carlo or BICM (Raftery et al. , 2007) [60] and the Predictive Information Criterion or BPIC (Ando , 2007) [3]. The main goal for a model to be selected is to minimize those criteria. We consider two alternative versions for each of the criteria, in order to have more variety in the selection criteria and also prevent overfitting. In the next paragraph, where the application of the theory is taking place, we present the 6 different versions – criteria in detailed way. We have to make clear at this point, that the results of the selection criteria may lead to different models as the best choice. This is because some criteria may minimize their value in a specific number of components but other criteria may not with the same number of components. For this reason, we are going to search for the model that does not only minimizes some of the criteria, but also satisfies the purposes of clustering and is able to provide meaningful and useful insights. Thus, the best model has to satisfy simultaneously most of the quantitative selection criteria and the qualitative selection criteria which are going to provide the number of clusters that are going to be useful for the analysis.

5.2 *Application of the Method*

5.2.1 Package Overview

The approach that is going to be presented for clustering the voters of FIFA Ballon d’Or datasets for the period 2010 – 15 is the Bayesian one, where the finite mixtures of the Plackett – Luce model are taken into account by assuming the Bayesian inferential perspective. The method is implemented with the help of PLMIX package. The PLMIX package was first released in 21/12/2016, as the only R package to deal with partial rankings/orderings by obtains inference based on the Bayesian Estimation. In terms of computational time, the PLMIX package outbalances the next application for clustering partial ranking data, the Rankcluster package, as its framework takes into account the computational issues that arise from partial rankings due to complexity of this kind of ranking data structures.

5.2.2 Data Input Format

First of all, it is a need to transform the input data into the proper notation that PLMIX works with. The proper notation for the input data in the PLMIX functions is the ordering notation. This is because, in applications like the one is advanced in this section, there is a lack of ranking elicitation to manage the complexity of ranking sequence when a number of items, which can be considered large, is ranked. If k is the total number of items to be ranked (23 in our case) and t is the number of items that are actually ranked from the voters (3 in our case), the remaining k-t alternatives which have not been ranked by the judges are tactically assumed to be ranked lower. Thus, this is the format that is going to be used in the implementation of clustering with the extensions of Plackett – Luce models, by assuming the inferential perspective, and is called top – t partial ordering. It can be noticed at this point, that this comes in contrast with the Ranklucster package’s input notation (ranking notation),which will be used in order to implement the ISR models in Chapter 8. On the other hand, the missing positions of the matrices are denoted in the same way like in ISR models, thus with zero entries. Moreover, Rank = 1 indicate the most – liked alternative, Rank = 2 indicate the second most preferred item and Rank = 3 the third item in the order of preference.

5.2.3 Estimation of Models

5.2.3.1 MAP Procedure

After transforming the format of input data to the appropriate notation in order to be compatible with the input that the functions of PLMIX require, the next step is to start the procedure of fitting the Baysian G – components Plackett – Luce (PL) mixtures according to Maximum A Posteriori (MAP) estimation procedure via EM algorithm and Gibbs sampling. The functions that are going to be used in order to obtain the models are the following: 1) *mapPLMIX_multistart* , which maximizes the posterior distribution through EM algorithm and returns the MAP point estimate of the PL mixture parameters. The function works by initializing the algorithm many times, with different starting values, in order to address the issue of possible local maxima in the posterior distribution. 2) *gibbsPLMIX* , which implements the MCMC posterior simulation via Gibbs sampling, having the goal to quantify the estimation uncertainty from a fully Baysian perspective. The two functions are going to be applied in sequence, which means that at first, the MAP procedure through 'mapPLMIX_multistart' function is going to be launched and then the resulting MAP estimate is going to be utilized in order to initialize the MCMC chain, via 'gibbsPLMIX' function. The above procedure is going to be implemented for every year of the period under analysis (2010 – 15).

Let's begin with the MAP estimation through the 'mapPLMIX_multistart' function. The arguments that are going to be used as input are :

- *pi_inv* : A data matrix of class 'top_ordering', which contains the partial orderings .
- *K* : The number of alternatives that are going to be ranked .
- *G* : The number of mixture components .
- *n_start* : The number of different starting values .
- *n_iter* : The maxim number of EM algorithm iterations .
- *centered_start* : A logical value which is used to constraint the random starting values to be centered around the observed relative frequency that each alternative has been ranked first .
- *parallel* : A logical value which is set to true in order to be able the parallelization .

In particular :

- ✓ As *pi_inv*, it is used the matrix with the partial orderings for each specific year of the period 2010 – 15 .
- ✓ As *K*, the number of columns – Players that are ranked from the judges (there are 23 number of alternatives in all years) .
- ✓ As *G*, the number of mixture components for the obtained model. For each year have been tried models with 2 components to 8 components .
- ✓ As *n_start* are used large numbers in order to let the algorithm intiallize the procedure many times .
- ✓ As *n_iter*, are used also large numbers in order to run the EM algorithm many times, so to shrink the probability of find a local maxima and not the global maxima .
- ✓ *centered_start* : True
- ✓ *parallel* : True

At this point, it should be emphasized that it is not being put in the input an argument that can be supported from the function, but at the same time, it is not included in the input argument of the function in purpose. The argument is : *hyper*, which is a list of named objects with hyperparameter values that are used for the prior information specification. If the prior setting is noninformative or flat, as it is in this case, the EM algorithm for MAP estimation performs frequentist inference, which means that the MAP solution for estimation coincided with the MLE solution and the best model in terms of maximized posterior distribution is returned.

The output of the function is an object of class 'mapPLMIX' and is the following :

- *mod* : A list of named objects describing the best model in terms of maximized posterior distribution. Two of the derivatives of this argument are used for the Gibbs sampling. These are *mod$P_map*, which is a numeric matrix with the MAP estimates of the component specific support parameters and *mod$class_map*, which is a numeric vector of the mixture component memberships that are based on MAP allocation of the matrix of estimated posterior component membership probabilities. Furthermore, *mod$P_map* and *mod$W_map*, which is a numeric vector with the MAP estimates of the G mixture weights, are used for the comparison of the models obtained with different number of components .
- *max_objective* : Numeric vector of the maximized objective function values for each initialization .
- *convergence* : A binary vector which, for each iteration, indicates if convergence has been achieved (1 if convergence has been achieved, 0 otherwise) .
- *call* : The matched call .

5.2.3.2 Gibbs Sampling

After obtaining MAP estimations for the PL mixtures via EM algorithm, the results of the procedure are going to be used in order for the MCMC chain to be initialized. The goal of the Gibbs sampling, in this stage of the procedure, is to approximate the joint posterior distribution in order to assess the uncertainty of the parameters estimates. This is achieved through 'gibbsPLMIX' function which, as had been referred in the previous paragraph, implements the MCMC posterior simulation via Gibbs sampling.

The input arguments of the function that are going to be used in the implementation of the MCMC chains are the following:

- *pi_inv, K, G* are the same arguments as in MAP procedure .
- *init* : It is a list of named objects which takes two initialization values . 1) *p* : A numeric matrix which contains the binary mixture component memberships, 2) *z* : A numeric G x K matrix of component – specific support parameters ,which is constructed by using the command *binary_group_ind* . The command constructs the binary group membership matrix from the

multinomial classification vector and takes as input a numeric vector with the class memberships and the number of classes .

- *n_iter* : The total number of MCMC iterations .
- *n_burn* : The number of initial burn – in drawings removed from the returned MCMC sample .

It has to be pointed out, that for the ‘n_iter’ arguments big values were chosen, in order to have large difference with the values used for the ‘n_burn’ arguments. The reason is that the difference between ‘n_iter’ and ‘n_burn’ is equal to the size of the final MCMC sample, thus if we want to have large final sample such a difference between these arguments has to be given for input values. Thus, for every model that the Gibbs sampling is performed, the values of ‘n_iter’ and ‘n_burn’ are 22000 and 2000 correspondingly. Another fact that has to be referred is that, in the MAP procedure, the list of named objects with hyperparameter values for the conjugate prior specification is not initialized, in order to not have an informative prior setting.

The output of the Gibbs’s sampling procedure provide the following arguments :

- *W* : A numeric matrix which contain the MCMC samples of the mixture weights. The dimensions of the matrix are LxG, where L is the size of the final MCMC sample and G is the number of mixture components that used for this model.
- *P* : A numeric matrix with MCMC samples of the component-specific support parameters. The dimensions of the matrix are Lx(G*K).
- *log_lik* : Numeric vector of L porsterior log – likelihood values.
- *deviance* : Numeric vector of L posterior deviance values, thus -2*log_lik.
- *objective* : Numeric vector of L objective function values. These values are the kernel of the log – posterior distribution.
- *call* : The matched call.

5.2.3.3 Comparison of the Models

After implementing Gibbs sampling for each of the model obtained from the MAP allocation, for number of components from 2 to 8, it is appropriate to select the best model among the performed fitted models. The model that is going to be selected, will provide the clustering results for the year under study. In order to choose the appropriate one, every model has to be tested according to specific criteria. Thus, the function *selectPLMIX* from PLMIX package is going to be used in order to compute Bayesian selection criteria with the range of the number of components that were used in the implementation of the models.

The input arguments of the ‘selectPLMIX’ function that are going to be used are:

- *pi_inv* : The matrix which contains the partial orderings of the year under study.
- *seq_G* : A numeric vector which contains the number of components of the PL mixtures to be compared. Because the models have been implemented in the range of [2,8] components the numeric vector, in the function, is going to be seq_G = 2:8 .

- *MAPestP* : A list which size is the length of seq_G and contains the MAP estimates of the component specific parameters.
- *MAPestW* : A list which size is the length of seq_G and contains the MAP estimates of the G mixture weights.
- *deviance* : A list which size is the length of seq_G and contains the posterior deviance values obtained from the Gibbs sampling methods for the different number of components.

A table with the six model selection criteria is returned. These are DIC1, DIC2, BPIC1, BPIC2, BICM1, BICM2 . Based on these criteria, we select the estimation with the components that minimize these metrics. The criteria used for the model selection are based on the principle of the trade – off between the fit of the data and the corresponding complexity of the model. Spiegelhalter et al (2002) [64], proposed a model comparison criteria that combined these characteristics. This is DIC (Deviance Information Criterion), which is defined as DIC = ‘goodness of fit’ + ‘model complexity’. The fit is measured through deviance, which is defined as $D(\theta) = -2logL(data|\theta)$ and the complexity is measured through the ‘effective number of parameters’, which is defined as $p_D =$ the posterior mean deviance minus the deviance evaluated at the posterior mean of the parameters. An alternative measure in model complexity which works with negligible prior information is the $p_V = Var(D)/2$. In that case, half the variance of the deviance is an estimate of the number of the free parameters in the model. This estimate, in cases of weak prior information, turns out to be very robust and accurate. Thus, DIC1 corresponds to DIC with p_D as measure of complexity and DIC2 corresponds to DIC with p_V as measure of complexity. BPIC1 and BPIC2 are obtained from the DIC1 and DIC2, respectively. Their difference is that they double the penalty term, in order to prevent the DIC’s tendency to overfit. The last two metrics, BICM1 and BICM2, are the Bayesian variants of the BIC. Their difference is that BICM1 is entirely based on the MCMC sample and, in contrast, BICM2 involves the MAP estimate without the need of its approximation from the MCMC sample. We are looking for the components that minimize these metrics and simultaneously providing a number of clusters that is going to be useful for the analysis.

5.2.3.4 Evaluation of the Models

Sometimes, the criteria do not provide a clear result for the optimal number of components. In that case, we are going to decide the number, based on the p-values of the assessment of the MAP estimates. In specific, we are going to evaluate the mixture – models adequacy, by computing the posterior predictive checks. This is achieved, by making use of the *ppcheckPLMIX* function from the PLMIX package. The function takes as input the *pi_inv* , *seq_G* of the previous functions and two lists, one containing the MCMC samples of the component specific parameters and one containing the MCMC samples of the mixture weights. It returns two posterior predictive p-values, based on two chi – squared discrepancy variables involving : the top – item frequencies and the paired comparison frequencies. The posterior predictive p-values can be compared with a nominal probability, typically set to 0.05, to conclude about the adequacy of the model. Values smaller that 0.05 are typically considered as indication of model’s lack of fit. Thus, as much greater than 0.05 is a p-value, it is an indication of proper fit for this specific model. Based on this assumption, we are going to make use of the p-values in order to obtain the optimal number of components when the model selection criteria will provide a not clear result. In particular, when the number of clusters is going to be chosen between models with similar results, the p-values of the adequacy’s assessment of these models are going to be used in order to obtain the choice of clusters.

5.3 Results

5.3.1 Year 2010

In 2010, the criteria for the comparison of the models with different number of components are presented in the following table.

Models	DIC1	DIC2	BPIC1	BPIC2	BICM1	BICM2
2- **Components**	2037.593	5705.357	-1576.845	5758.678	5920.916	9588.678
3- **Components**	5708.9859	5691.243	5805.616	5770.131	6010.157	5992.415
4- **Components**	5782.0772	5698.763	5950.898	5784.27	6044.437	5961.123
5- **Components**	5833.9357	5701.83	6053.337	5789.125	6054.731	5922.625
6- **Components**	5885.1377	5705.104	6154.534	5794.466	6066.363	5886.329
7- **Components**	5920.8521	5707.823	6224.838	5798.779	6075.526	5862.497
8- **Components**	441.8131	5712.939	-4735.025	5807.227	6094.111	11365.237

Table 4 : The criteria used to obtain the proper number of components, for Year 2010.

Based on Table 3, we reject the MAP estimations with 2 and 8 components, due to their enormous high values in BICM2 and the negative values in BPIC1 which is an indication of a not good estimation. If we check the DIC1 values for the models with these two components, we can observe the DIC tendency to overfit, because of the extremely small values it obtains. Thus, among the rest of the models, the metrics indicate that the appropriate model is the one with 3 components. Moreover, by checking the table (A.6) with the p-values for the models assessment, we notice that the p-value of the ‘paired’ discrepancy variable, for the 3 – components model, is the greatest among the others.

The 3 – components model provide the following number of observations in each cluster. Cluster 1: 121 observations, Cluster 2 : 142 observations and Cluster 3 : 162 observations. In (B.1), someone can observe that the separation between the first and the second cluster is very good but the points of the third cluster are not so concentrated within the cluster and well separated from the other two clusters. Each of the three clusters have a player who is the central person in the cluster. The (A.7) shows that

Cluster 1 belongs to Messi with 86 out of 121 first - place votes, while on the same group his two opponents for the first - place , Iniesta and Xavi, have not been preferred so much from the voters. In Cluster 2 (A.8), it is noticed that there is a strong stream of judges that support Xavi, as he has been preferred in the Top – 3 from all of them. The third Cluster (A.9), belongs to Iniesta because he has received an amount of points that is much greater than his other two opponents. A visualization of the points that each of the three players received in his cluster is presented in the following plot.

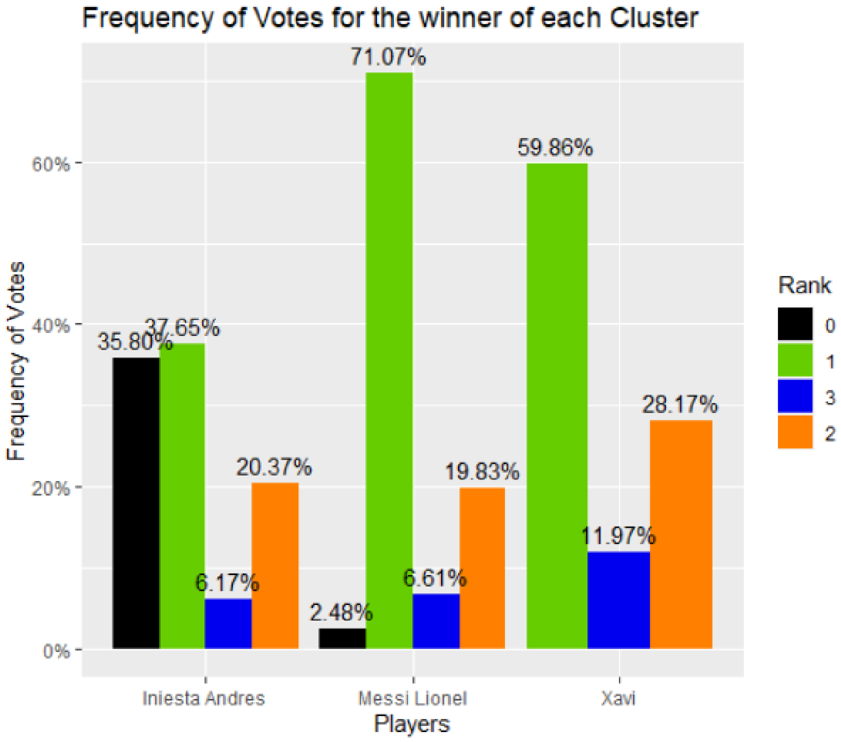


Figure 7 : Bar plot that presents the ranks for each cluster's winner.

From Figure 7, we observe that Messi has been taken a strong lead with a very large amount of 5 – points votes, in the cluster that he was the top vote receiver. On the other hand, one could wonder why Iniesta has been ranked second in the final rankings, despite the fact that he is not such strong as Xavi in his own cluster. The main reason is that the proportion of votes is not the same for Cluster 3 (Iniesta's Cluster) and Cluster 2 (Xavi's cluster). The second is smaller which means that the amount of votes that Iniesta received is bigger than it seems to be, if we consider the actual size of his cluster. Another reason is that Xavi is very weak in Cluster 1 and Cluster 3, by concentrating all his power in Cluster 2. On the other hand, Iniesta has received a respectable amount of votes in the clusters that did not win.

5.3.2 ***Year 2011***

Based on (A.10) we observe that, if we except the models with 2 and 8 components for the same reason as in previous year, the indications for the appropriate model converge to the 4 – components model. Moreover the p-value (A.11) for the 'paired' discrepancy variable, indicates a good fit, as it is 0.43 and surpasses the threshold of 0.05.

From the 4 clusters obtained, the largest and the smallest are the more interesting. Thus, the largest cluster is apart from 232 out of 465 voters of the complete dataset. It is a very well representative of the enormous difference that the winner of the award (Messi) has to the second player (Cristiano), in terms of votes. Also, it represents the difference from the second - place (Cristiano) to the third - place (Xavi), because of the huge amount of second - place votes that Ronaldo received. Someone can observe from the (B.2), that the 93% of the judges in this cluster have voted Messi first and the 91% have voted Cristiano second. One could say that this denotes the start of the greatest rivalry in the history of modern football.

On the other hand, Cluster 4 which apart from 37 observations is also very interesting, but from different scope. The reason is that this cluster could be characterized as a 'Anti – Cristiano' cluster, since Ronaldo has not been preferred in the Top – 3 neither from one voter and in the same time Messi has been ranked first from all of the voters. It could be very interesting to have a more deep look at the kind of relationship that these voters have with football and also the continents that are come from, in order to search for possible patterns.

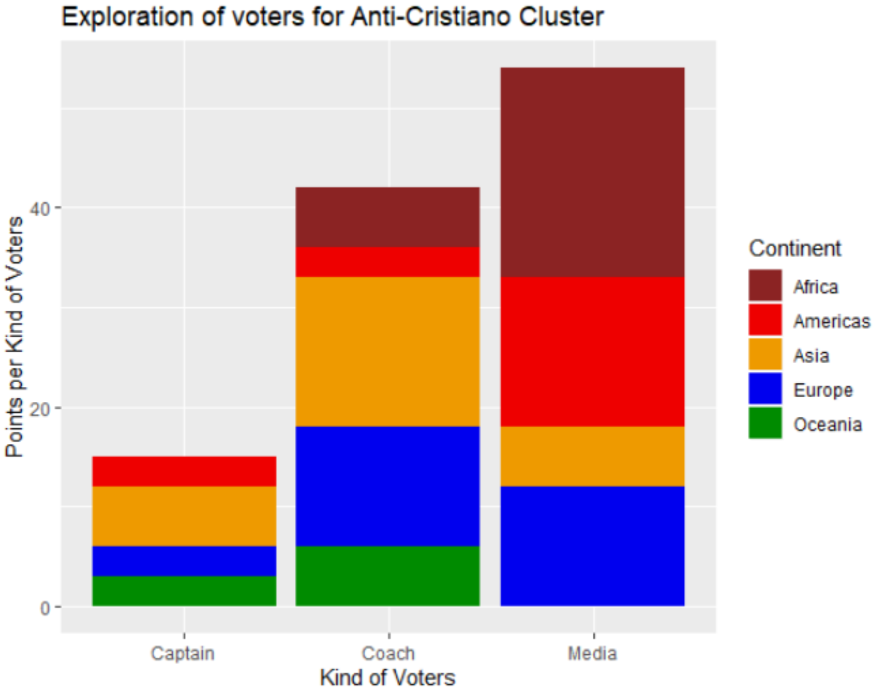


Figure 8 : Stacked bar plot with the job and the continent of the voters in the 'Anti – Cristiano' cluster.

Figure 8, provides insights about the job of the voters and the continent they come from. The continent that each voter comes from, has been obtained by making use of the 'countrycode' function from the 'countrycode' package. Someone can observe from Figure 8, that Media have strong presence on this cluster. On the other hand the players that belong in the cluster are few. We notice that there are not so many European judges in the cluster, in compare with the amount of voters from other continents. This is not a surprise if we take under consideration the origins of the two players (Cristiano from Portugal and Messi from Argentina). On the other hand, there is a strong attendance of African and American voters in the 'Media' bar. It makes sense for American press to prefer Messi than Ronaldo for winner of the FIFA Ballon D'Or, if we consider the origins factor, but the large number of African voters is a surprise. An explanation could be, the influence of Barcelona in Africa, as brand name, in compare with Real Madrid or the assumption that African people would prefer Messi than Ronaldo because the first is a teammate of the great African striker Samuel Eto'o.

5.3.3 Year 2012

In 2012, we perform the clustering with 4 clusters, as the (A.12) and the (A.13) tables indicate. The distribution of the observations in the corresponding clusters is the following, Cluster 1 : 36 observations, Cluster 2 : 213 observations, Cluster 3 : 27 observations, Cluster 4 : 229 observations. We notice that there are two clusters with a large amount of voters and two clusters with a tiny amount.

In Cluster 1 (A.14), Ronaldo has not been ranked neither from one voter while Messi has a large amount of first votes, but without having a head start against his opponent. In Cluster 3 (A.15), the remarkable fact is that the dipole has not been preferred a lot in the first three positions from the voters. On the other hand, in this cluster there is a strong presence of Iniesta. Cluster 2, despite the fact that is the second largest cluster, does not provide an apparent winner. As someone can observe from (A.16), the sum of total points, which obtained from the ranks received, for the two contenders of the award does not differ a lot between them. What Cluster 2 affords to the analysis, is that consolidates the fact that there are only two competitive candidates for this title. Thus, everything depends from the votes exist in Cluster 4. There is no doubt that this cluster is full of voters that nominated Messi as the best football player in 2012. The Argentinian has received the extraordinary amount of 229 first - place votes out of the 229 voters in this cluster. Thus, it is natural corollary for this cluster to be characterized as the 'Total Messi's Cluster'. On the other hand, Ronaldo has ensured the second - place, as he has been ranked in the second position from 139 voters. It is also remarkable the fact that 139 judges out of 229 have been ranked Messi first and Ronaldo second. It would be very interesting to dive into this cluster and search for possible patterns for the voters of this cluster. By obtaining the stacked bar plot (B.3), we do not detect a specific pattern of voters, in terms of continent or the voters job, which could mean that Messi had the vast acceptance of the football audience, no matter other factors.

5.3.4 Year 2013

It is the first year, in the period under study, that the total points of the first three competitors are very close to each other. Also, after three consecutive years of Messi's dominance, Ronaldo wins the award. Another interesting point is the fact that, there is a huge points difference between the third (Frank Ribery) and the fourth (Zlatan Ibrahimovic), in the final rankings. The model comparison criteria (A.17) do not provide an apparent result for choosing the model with the appropriate number of components. Thus, our choice is going to be based on the p-value obtained for the 'paired' discrepancy variables of the models. From (A.18), we can observe that the biggest p-value, thus the best fit among the models, is obtained from the 3 – components estimation. So, we are going to use the 3 – components model for the clustering of voters.

The clusters obtained reveal a voting pattern for the three frontrunners (Ronaldo, Messi, Ribery). Cluster 1 (A.19), which contains 113 observations is a very balanced cluster, in terms of Top – 3 ranks, between Ronaldo and Messi. On the other hand, we notice that Ribery's presence is very 'weak' in this cluster as he has been preferred in the Top – 3, just from 7 out of 113 judges. In Cluster 2 (A.20), which represents the 52% of the total dataset, Messi does not lose his ranking power but Ronaldo makes the difference which indicates that he could be the possible winner on this run. In particular, Cristiano has been preferred in the first three positions from 251 out of 282 voters and especially in the first - place , from 143 voters. On the other side, Messi, has been preferred in the first three places 184 times and from them the 63 are first - place positions. Ribery, has increased his rates in this cluster, but the small boost he received does not enable him to reach the other two candidates. The (B.4) provides a visualization of the second cluster's results. After analysing the outcomes of the first two clusters, one could say that Ronaldo is the winner with small difference from Messi and with an enormous difference of both from the third player. But after looking at the third cluster the conclusions are very different. Cluster 3 (A.21), contains the supporters of Ribery and could be characterized as a 'Ribery – centric' cluster. The French player has received the enormous proportion of 137 first - place votes out of the 149 voters included in this cluster. It seems strange for a player, which is outside of the dipole, to receive such an amount of votes but we have to take under consideration that this group of voters rewarded Ribery for the outstanding performance he had in that specific year and for his enormous contribution in the conquering of the Champions League trophy from Bayern Munich. In (B.5), someone can observe that the most of the voters in this cluster are journalists. Also, there are many Europeans and Asians in the 'Captain' and the 'Media' bar and not so many from Africa.

To conclude, someone could argue that two voting types have been detected in this year. Ronaldo and Messi had a consistent and dense amount of votes in the first two clusters (almost 70% of the dataset in total), with the difference that Cristiano had much more first position ranks and that's the reason he won the award. On the other hand, Ribery has an extreme ranking behaviour if we consider that he has very small presence in the biggest part of the dataset, without receiving a respectable amount of second or third - place ranks, and has been ranked first from a whole cluster. Thus, the largest percentage of his ranks are either 0 or 1. Based on the final rankings, we could argue that the consistent ranking behaviour of Messi is preferred than Ribery's extreme ranking behaviour.

5.3.5 Year 2014

In 2014, Ronaldo wins his second consecutive award, with a large difference from Messi. Moreover, this year is very interesting because the difference of points between the second and third - place is negligible.

Based on (A.22), we notice that the largest absolute difference, between different number of components, exists in the 6 – component model. Moreover, the p-value (0.504) (A.23), of the model's assessment indicates a very good fit. Thus, we are going to work with the 6 – component model for the clustering. The observations are distributed raggedly in the clusters. More specifically, Cluster 1 : 93 observations, Cluster 2 : 163 observations, Cluster 3 : 54 observations, Cluster 4 : 44 observations, Cluster 5 : 15 observations and Cluster 6 : 175 observations. We are going to present the clusters that provide interesting insights, in terms of the analysis.

Cluster 1 could be characterized as 'Anti – Messi' cluster, since the Argentinian has not been preferred neither from one voter in the cluster. On the other hand, Cristiano has been voted first, 83 times. It would be interesting to have a look at the people that apart this cluster. Based on (B.6), we can not observe a special group of voters. On the other hand, it is a surprise the strong presence of American voters in the cluster, which was not expected, if we consider the origins of Messi and the votes of people from this continent in previous years. Cluster 2 is also very interesting, since it is the first time in the period 2010 – 15, that a goalkeeper receives so many votes in a cluster. From (A.24), we can notice that Neuer is the winner in this group, by overcoming even Ronaldo. In contrast with the voters in Cluster 1, where we could not justify the dense presence of American people, there is a pattern that someone could investigate in Cluster 2. By looking at (B.7), we are able to notice the strong presence of European people and the fact that are very few voters from Africa or America. We could assume that Neuer does not reach so much audience in Africa and America, because people in these continents like offensive football so they would prefer a striker than a goalkeeper. Small, but interesting, is the fourth cluster where Cristiano has not been preferred from any voter despite his big win. From (B.7), we can observe that mostly coaches exist in this cluster. Also, we can observe that are very few Europeans in the cluster. Finally, Cluster 6 (A.25), is the largest group of voters in 2014. It is a cluster that represents the breadth of Ronaldo's win, who receives 168 first - place ranks out of 175. It is noticed that Neuer has not been preferred from a large amount of voters in the Top – 3. On the other hand, Messi may not have been ranked first, in the total dataset, as many times as Neuer has, but his presence in the Top – 3 ranking is more frequent and robust than the German's. This reminds us the previous year where Ribery had similar ranking behaviour with Neuer, while Messi had a more consistent ranking behaviour. In both cases, the final result is that, Messi ranked above his opponents.

5.3.6 Year 2015

In 2015, Messi makes his comeback by receiving almost the half of the total points and winning the award. The second - place belongs to Ronaldo with a large difference from the third, Neymar.

The output of the model comparison (A.26), is very clear and provides the 2 – components model, as the best one among the others. After obtaining the results of clustering with the 2 – components model, we observe two clusters with 385 and 113 observations, respectively. Cluster 1, which constitutes the 77 % of the dataset, provides a typical representation of the final result. The following plot is presented in order to visualize this representation.

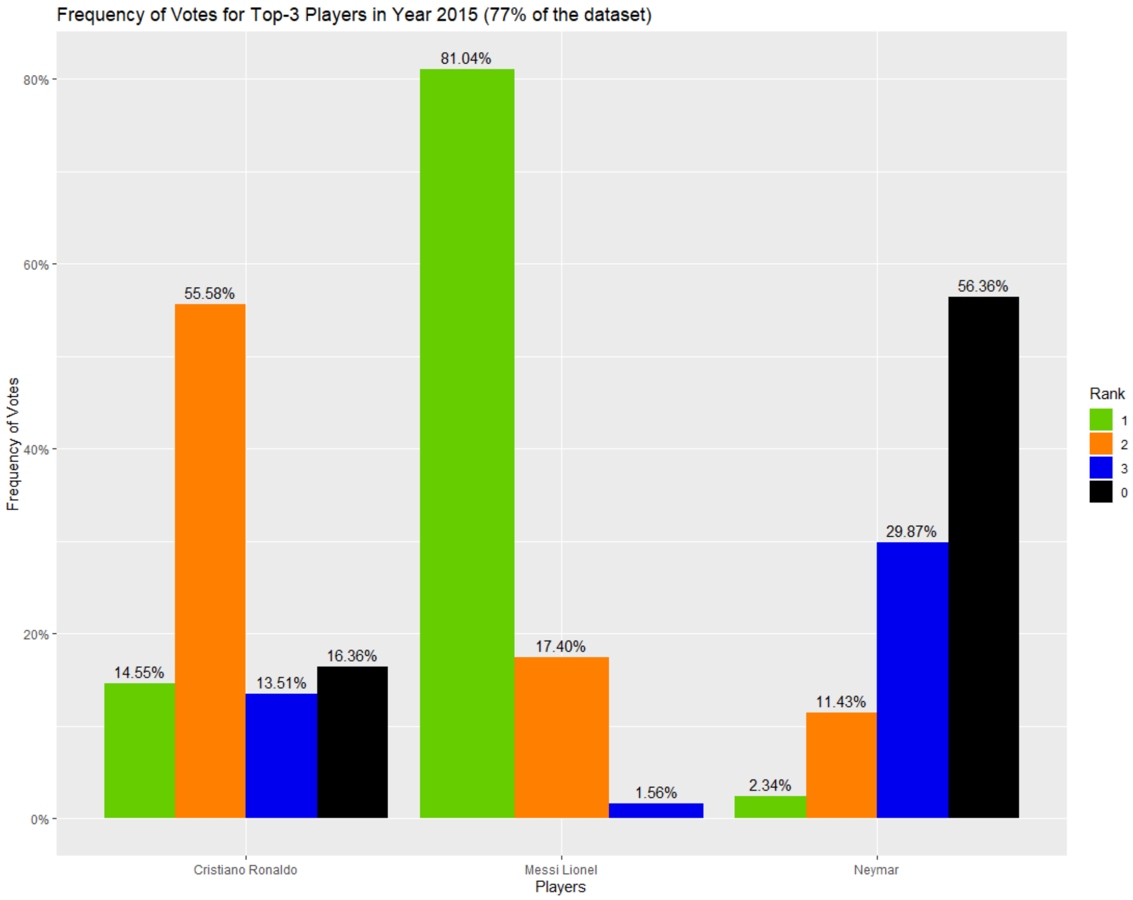


Figure 9 : Bar plot for the frequency of Top – 3 rankings, that apart the 77% of the total dataset, in Year 2015.

As we can notice from Figure 9, the proportion of first position votes for Messi, 81.04 %, is enormous and indicates his domination against the other players. Moreover, the proportions of second - place ranks for Ronaldo and third - place ranks for Neymar, are also indicate the players that are in the final second and third - place, respectively. We could say that this cluster does not reveal any peculiar pattern, but captures the vox pop of 2015. In the second cluster, Ronaldo has received more points than Messi but without making any change in the final result. Also, there are some players like Lewandowski, Muller and Benzema that have received little more points in compare with Cluster 1.

Chapter 6

Cluster Analysis on the Data with K – medoids algorithm

In this chapter, it is presented an alternative approach for the clustering of the FIFA Ballon D'Or datasets. In specific, we separate and group the voters, based on their preferences on the players that are about to be ranked, through the k – medoids or partitioning around medoids clustering method. The distance metric that is used in order to calculate the distance between the points – voters, is the Kendall's distance. The implementation of the k – medoids method contains two steps (Build and Swap), which are repeated until there is no change in the clusters. Thus, in this chapter, we present a distance – based clustering method, where the similarity quantification between the ranking objects is based on the Kendall's distance. As already been said in the previous chapter, the more similar objects in terms of preference, have a closer distance and vice versa. So, after the distance is defined a partitioning algorithm is applied, in order to achieve the clustering purpose.

6.1 Theoretical Framework of the Method

6.1.1 ***K – medoids or PAM algorithm***

One of the most popular partitioning algorithm, in the unsupervised learning field, is the k – means algorithm. The key idea of this algorithmic approach is to partition the sample space in separate parts. Someone could describe the k – means algorithm as an iterative procedure, where at each iteration each observation is assigned to the closest cluster and afterwards the cluster centers are updated. The assignment of the observations to the clusters is based on Euclidean distance. The number of clusters in k–means have to be known, before the start of the iterative process. Thus, the final clustering results depend on the initial values.

However, the k – means algorithm are not appropriate when the data that are about to be clustered are categorical. This is due to the fact that, it has no sense to calculate the distance between two categorical variables with the Euclidean distance. The point is that the k – means algorithm use numerical distances (e.g Euclidean distance), so the result would consider close two probably distant objects that would have been assigned two close numbers. Thus, in our case, the k – means method is not applicable, as the data are categorical. Moreover, it is not wise to consider the option of transforming our categorical data to numerical data in order to perform k – means.

Thus a solution in order to perform the clustering, in such cases, lies in the k–medoids or partition around medoids (PAM) algorithm. The PAM algorithm is a well – known clustering algorithm, which aims to find k medoids and assigns every point to the nearest medoid that is the point with the shortest distance to the other points in the cluster. In k – medoids algorithm the data objects are chosen to be the medoids, in compare with the k – means algorithm where the means are chosen to be the centroids. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is the minimum one [73]. Medoids are similar in concept to means or centroids, but medoids are always members of the dataset.

6.1.2 K – medoids procedure

The fundamental idea of the algorithm is to first define the ‘centers’ of the clusters, which are the medoids. After finding the set of medoids, each object of the dataset is assigned to the medoid from which has the shortest distance, according to the used distance measure. So, in other words, object i is put into cluster v_i, when medoid mv_i is nearest than any other medoid m_k, where k indicates the number of medoids that have been defined [63]. The procedure of the algorithm can be described in two steps : 1) the ‘Build’ step and 2) the ‘Swap’ step. In the first – ‘Build’ step, k centrally located objects are chosen, sequentially, to be used as initial medoids. These k objects of the dataset are chosen randomly. In the second – ‘Swap’ step, the non – chosen objects are assigned to the nearest representative objects according to a distance metric [63]. In our approach, the distance metric is the Kendall’s distance. After that, for each pair of non – selected object and selected object, the total swapping cost is calculated. If the total swapping cost is smaller than 0, the initially selected point is replaced by the initially non – selected. This procedure is repeated until there is no change of the medoids.

6.1.3 Kendall’s distance

The distance metric that is used in this approach in order to calculate the distance between the objects of the dataset, is the Kendall’s distance. As discussed in the ‘Non -Metric Multidimensional Scaling’ section , the Kendall’s distance is very powerful, in compare with other distances that are proper for the calculation of dissimilarities between ranking data, when missing data exist.

Someone could define the Kendall’s distance as the metric that counts the pairwise disagreements between rankings. Thus, the larger the distance the more dissimilar are the preferences of the two voters, and vice versa. It is also a metric distance. That means that it satisfies the triangle inequality, which states that the sum of the lengths of any two sides of a triangle is greater than the length of the remaining side or $(\mathrm{d}(\mu,\nu) <= \mathrm{d}(\mu,\sigma) + \mathrm{d}(\sigma,\nu))$ [45]. Because we want to find the ‘shortest paths’ between the data points, the distances that capture the notion of triangle inequality enable to define these distances to be the length of the ‘shortest path’ without having to define things like path, or length of a path. Another important property that the Kendall’s distance captures is the right – invariance property. A distance measure is defined as right – invariant if for any permutation of the rankings σ, μ, ν the following property is satisfied : $d(\mu,\nu) = d(\mu\ o\ \sigma, \nu\ o\ \sigma)$, where $\mu\ o\ \sigma(i) = \mu(\sigma(i))$ [45]. More specifically, right invariance assures that the distances which have this property remain immutable under any possible permutation relabeling of the objects.

If a distance measure is right invariant, for a set of permutations, it allows us to rewrite the ranking tables in a different, more convenient way.

The Kendall's distance is calculated by the following formula : $\tau = \frac{n_c - n_d}{n(n-1)/2}$. The term n_c is the number of concordant pairs and the terms n_d is the number of discordant pairs. The term $n(n-1)/2$ is the normalizing term of the distance, where n is the number of the listed ranked objects – players. A concordant pair can be defined as a pair of players that both have been ranked in the same order, or in other words that they both moved in the same direction. For example, Lionel Messi and Gyan Asamoah are a concordant pair of players because Messi was consistently ranked higher than Asamoah. Conversely, two players can be characterized as discordant because the voters have ranked them in opposite directions. Such an example of discordant players are Messi and Ronaldo, because other voters rank higher Messi than Ronaldo and vice versa.

There are also other distance metrics that are appropriate for ranking datasets and satisfy the above properties (triangle inequality, right invariance), such as Spearman distance or Hamming distance. The reasons that the Kendall's distance has been chosen for the calculations of dissimilarities between the preferences of the voters, lie on the fact that it satisfies the above properties and, also, it is very powerful when the dataset contains partial rankings. As has been discussed in previous chapter, it has been proved from Tilley and Cabilio in 1999, that when there exist missing observations the Kendall's statistic has more power, in compare with other distances, in identifying more patterns.

6.1.4 Determine and Assess the final model

6.1.4.1 Average Silhouette Value

After obtaining the method for different number of clusters, we are going to select and assess the final model with the help of the Silhouette method . Silhouette refers to a method of interpretation and validation of clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster. The measure that is going to determine the number of clusters and to provide how appropriately all the data has been clustered, is the average silhouette value [36]. A high average silhouette indicates a good clustering. The optimal number of clusters k is the one that maximizes the average silhouette value over a range of possible clusters. Thus, if there are a lot of 'secure' clustered values, it is expected a big average silhouette value.

The definition of silhouette value for a datum i is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

and the possible outcome lies in the range of [-1,1]. The term $b(i)$ represents the lowest average dissimilarity of the datum i to any other cluster which i is not a member. The average dissimilarity of a datum i to a cluster c can be defined as the average of the distance from i to points in c [36]. The term $a(i)$ represents the average dissimilarity of i with all other data within the same cluster [36]. A value of +1 indicates that the sample is far away from its neighboring cluster and very close to the cluster its assigned. Similarly, value of -1 indicates that the point is close to its neighboring cluster than to the cluster its assigned. And, a value of 0 means its at the boundary of the distance between the two cluster. Thus, it can be followed that the value of 1 is ideal and the value of -1 is least preferred.

6.1.4.2 Elbow Method

Besides the average silhouette value, the Elbow method [70] is going to play a helpful role in choosing the optimal number of clusters for the model of each year. The Elbow method is a very well known method in which the sum of squares at each number of clusters is calculated and graphed. Based on this graph, a steep change of slope, that looks like an ‘elbow’, indicates where the optimal number of clusters might be. It is logical that as the number of clusters increase, the fit is improved because more of the overall variation is explained. In the same time, since more clusters are added to the model there is the danger of overfit. Thus, the method tries to identify a ‘knee point’ where the variation that is explained from the model’s parameters is acceptable and the increasing of the parameters is not going to reflect to the predictive ability of the model to other data. So, it is expected that the first clusters are necessary since they explain a lot of the variation and the data consist of that many groups. But at the time that the number of added parameters exceeds the actual number of groups in the data, the added information will drop sharply, since it separates the actual groups. Based on this fact, it is assumed to be a knee point in the graph of explained variation versus the clusters, since the line is going to increase rapidly and then increase slowly. It has to be said at this point, that the Silhouette method is going to play the conclusive role in determine the optimal number of clusters and the Elbow method is going to be used for confirmation or reconsideration of the ‘appropriate’ number.

6.2 *Application of the Method*

6.2.1 Package Overview

In this chapter, we put into practice the theoretical framework that has been provided in the previous chapter. First of all, we have to mention the packages that are used in the application of the method. So, in order to compute the distance matrix between the points of each dataset, we make use of the ‘amap’ package and the ‘Dist’ function. The ‘Dist’ function computes and returns the distance matrix computed by using the specified distance measure, which is the Kendall’s distance in this case, to compute the distances between the rows of a data matrix. Afterwards, for the implementation of the partitioning around medoids, based on the computed distance matrix, is used the ‘pam’ function from the ‘cluster’ package. This function partitions the data into the number of cluster that the user specifies, around the medoids. We are going to see in detail the input that the function requires. Finally, in order to determine the optimal number of components for the model of each year and evaluate it, we make use of the ‘silhouette’ function from the ‘cluster’ package and the ‘fviz_nbclust’ function from the ‘factoextra’ package for the Elbow method.

6.2.2 Data Input Format

Before starting the implementation of the clustering process, we have to transform the input data in order to be compatible with the requirements of the package and meaningful, in terms of the analysis purposes. Thus, before computing the distance matrix, with Kendall’s distance as the distance measure, the positions that have not been ranked are going to be imputed by the midrank. The notion of midrank has been discussed in the Chapters 2,3 and 4, as an appropriate way for imputing missing positions in partial ranking data. It has to be pointed out that in the previous method, the missing positions were part of the estimation process. Thus, the first reason of the midrank imputation in the k – medoids method is the fact that it is a suboptimal choice to ignore the missing data when testing for trend. Besides that, the function ‘pam’ from the ‘cluster’ package, that is going to be used for the clustering process, does not allow the presence of missing data when the input matrix is a dissimilarity matrix. It has been stated that the whole procedure starts with the computation of the dissimilarity matrix with the Kendall’s distance as the distance measure, and this matrix is used as input for the ‘pam’ algorithm.

6.2.3 Estimation of Models

As previously been stated we are going to perform the clustering around medoids with the help of ‘pam’ function. The function implements the process of the pam algorithm, which means that tries to look for k objects or medoids among the observations of the dataset, which are representative of the structure of the data. After finding these k medoids, it assigns each observation to the nearest medoid in order to construct k separate clusters. The overall target is to find k representative objects which minimize the sum of dissimilarities of the observations of the dataset to their closest representative object. The whole process is implemented though the Build and Swap phases that described in the theoretical part of the method.

Let’s have a look at the input parameters that are required in any case :

- *x* : A data matrix, data frame or dissimilarity matrix, depending on the value of the *diss* argument that follows. In case of a dissimilarity matrix, the missing values are not allowed.
- *k* : A positive integer specifying the number of clusters. It is important that the user is able to specify the number of clusters under examination, without a restriction. The clusters have to be less than the number of observations.
- *diss* : It is a logical flag that plays the most important role in the output because it is determined if the input object is a dissimilarity object or not. If TRUE, then the input matrix is considered as a dissimilarity matrix. If FALSE, the input matrix is considered as a matrix of observations by variables.

- *metric* : A character string that specifies the metric to be used for calculating dissimilarities between the observations. If *x* is already a dissimilarity matrix, the argument is ignored.
- *medoids* : A length k – vector of integer indices, which specifies initial medoids instead of using the 'Build' algorithm. The default choice is NULL.
- *stand* : A logical value, that in case it is TRUE, the measurements in the input matrix are standardized before calculating the dissimilarities. If the input matrix is already a dissimilarity matrix, the argument is ignored.

These are the basic input arguments that the 'pam' function requires in order to be implemented. Based on these arguments, the input values that have been given to the function in order to run the algorithm for our case, are the following :

- ✓ As input matrix *x* , is used the dissimilarity matrix that has been computed with the help of 'Dist' function. Thus, the *diss* argument has been set to TRUE and the *metric*, *stand* arguments have been ignored.
- ✓ For the number of clusters *k*, it has been chosen a vector of 2:10 clusters for the algorithm to run. This is due to the fact that there is no restriction for the number of clusters that the algorithm can take, so it is feasible to test a big number of clusters.
- ✓ For the *medoids* argument, the value has been set to the default (NULL), in order for the algorithm not to have specified initial medoids and be able to run the first – Build phase of the process.

The output of the function is an object of class 'pam' and provides information about various aspects regarding the results of the clustering process. A typical example is a matrix where each row corresponds to numerical information for a cluster e.g the number of observations, the maximum and average dissimilarity between the observations in the cluster and the cluster's medoid, etc. Another useful object provided in the output is a clustering vector with the number of observations in each cluster, a list with the silhouette information, a total dissimilarity matrix between the objects of the dataset, a vector with the medoids or representative objects of the clusters, etc.

6.3 Results

6.3.1 Year 2010

As in the previous methods, we are going to present the results of the clustering for each specific year of the period 2010 – 15, starting from the Year 2010. In the average silhouette plot (B.8) someone can observe that the silhouette values, for the examined clusters, vary from 0.19 to 0.23. Despite the fact that the largest average silhouette value is in the model with the 6 components, we spot the largest difference of value between the models of 2 and 3 components. In particular, the model with 2 clusters has a value below 0.2 and the 3 – clusters model has an average silhouette value between 0.22 and 0.23. Thus, because the difference in average silhouette value between the two models is very small, we are going to select the model with 3 clusters, in order to prevent overfitting by separating the data in more groups. Moreover, in the plot of Elbow method (B.9), it cannot be spotted a 'breakpoint' where the total within sum of square is minimized sharply. The silhouette plot for the 3 different clusters of the final model, is presented below. We can notice that the third cluster is the most appropriate clustered, as it has no point below zero and also having the largest average silhouette value (0.47). On the other hand, some points seem that have wrongly been placed in the first cluster. The second cluster, which has the most observations, seems more robust than the first but it includes points that are not appropriately clustered.

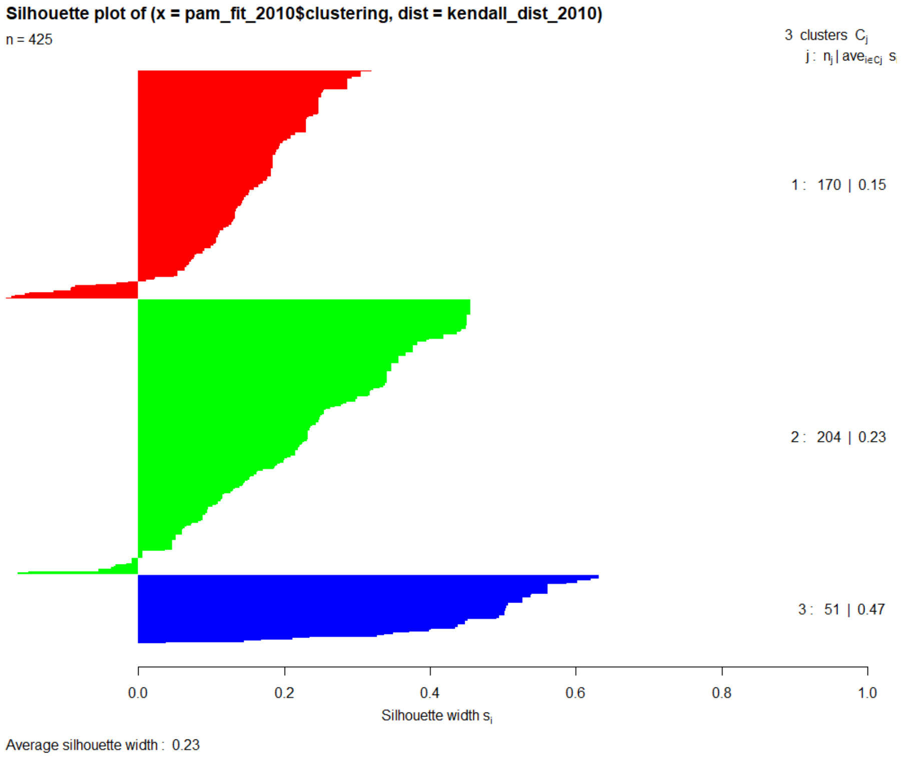


Figure 10 : Silhouette plot for each cluster of the final 3 – components model.

The above observations can be confirmed by looking at the visualization of the clustered data, that is presented in (B.10).

Now let’s have a look at some information about the clusters. The first one contains 170, the second 204 and the third 51 observations. As previously stated, through the ‘silinfo’ function, can be provided information about the clusters such as the medoids of each cluster. In Cluster 1, the object that represents the votes in this cluster is Van Marwijk Bert, who has voted Messi and Xavi in the two of the three places, as his third vote was invalid. As we can see from (A.27), Messi and Xavi are the players with the most votes in Cluster 1. In Cluster 2, the medoid turns out to be Danilevicius Tomas. The former Lithuanian player has ranked first Iniesta, second Sneijder and third Forlan, which is interesting because the preferences of Cluster’s 2 medoid are quite similar with the final results of Cluster 2. In specific, from (A.28), someone can observe that the player with the most first - place votes is Iniesta, the one with the most second - place votes is Xavi and the player with the most third - place votes is Diego Forlan. What is really strange in this cluster is the fact that Messi has been ranked from 36 out of 204 voters and only 7 times first. If we take under consideration that the Argentinian won the trophy that year, these rankings are peculiar. Thus, it would be interesting to check the composition of this voting group.

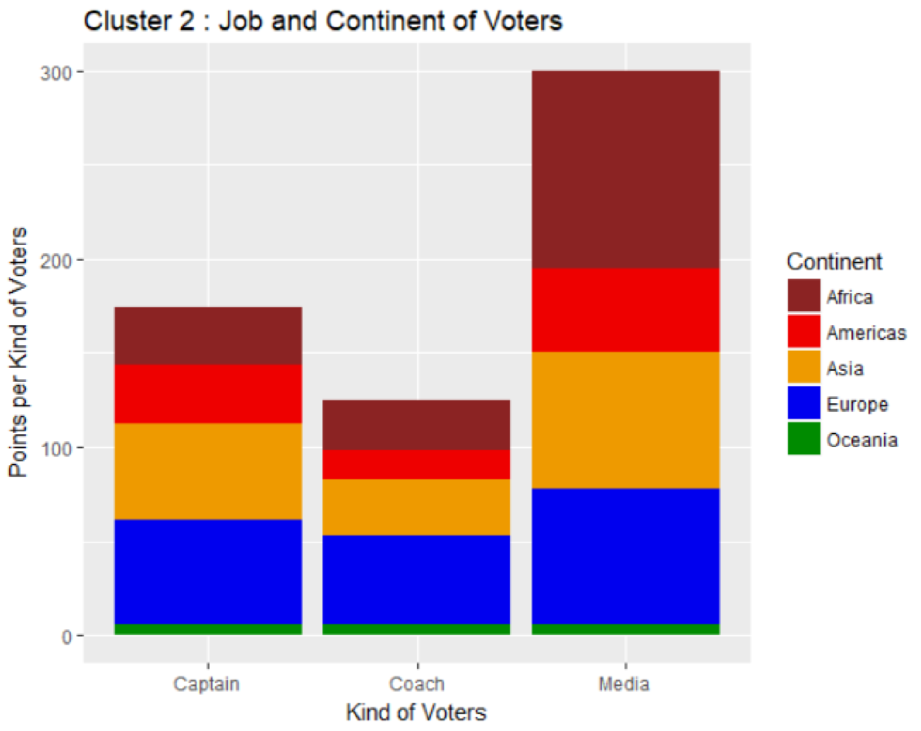


Figure 11 : Job and Continent of Voters for Cluster 2 in Year 2010.

From Figure 11, it can be observed that journalists are the main population of this group and many of them are from Africa. On the other hand, the percentage of coaches that constitutes this cluster is small, while the bar of players is also relative small but not that minor as the one of coaches.

Cluster 3 (A.29) could be characterized as the Messi's group, because 47 of the 51 voters have preferred the Argentinian and 31 of them have ranked him first. After a more careful look, Cluster 3 could be characterized as a 'Messi – Cristiano' cluster because it is observed that it is the first group in 2010 that the Portuguese receives a respectable amount of votes. Moreover, if we dig into the structure of the voting group, the pattern that was detected in the second cluster where Media have not preferred Messi too much in 2010 is confirmed. More specifically, it can be observed from the (B.11) that the presence of Media in this cluster is fractional, while the presence of coaches and players is very strong. One could say that the two bar plots are completely opposite.

6.3.2 Year 2011

In 2011, the average silhouette plot (B.12) indicates very clear that the optimal number of clusters is 2, since it has the largest value (0.4) and after that the line falls very sharply and starts to increase in a very slow manner. Thus, we fit the model with two clusters.

This year could be described as the first year that Messi starts to branch off his opponents. From the total rankings that are displayed in (A.30, A.31), it can be easily observed that Messi has received a tremendous amount of votes and the largest proportion of this amount are first - place votes. Besides that, 2011 is the year that starts to be configured the dipole Messi – Ronaldo, which conquers in the wins of the trophy the next years, as it is the first year that Cristiano receives a respectable amount of votes that ranked him second with a large difference behind Messi. It would be interesting to observe the two clusters that occurred and to spot their differences.

Cluster 1 is consisted of 322 observations of the total 465 that apart the dataset of Year 2011. Based on the cluster's information, the voter that represents this voting group, in the best way, is Anthony Griffith. His preferences were : Lionel Messi first, Cristiano Ronaldo second and Xavi third. Someone could say that the medoid is very accurate if the total preferences of the voters in Cluster 1 be taken under consideration. A graphical representation of these preferences is presented in the next figure.

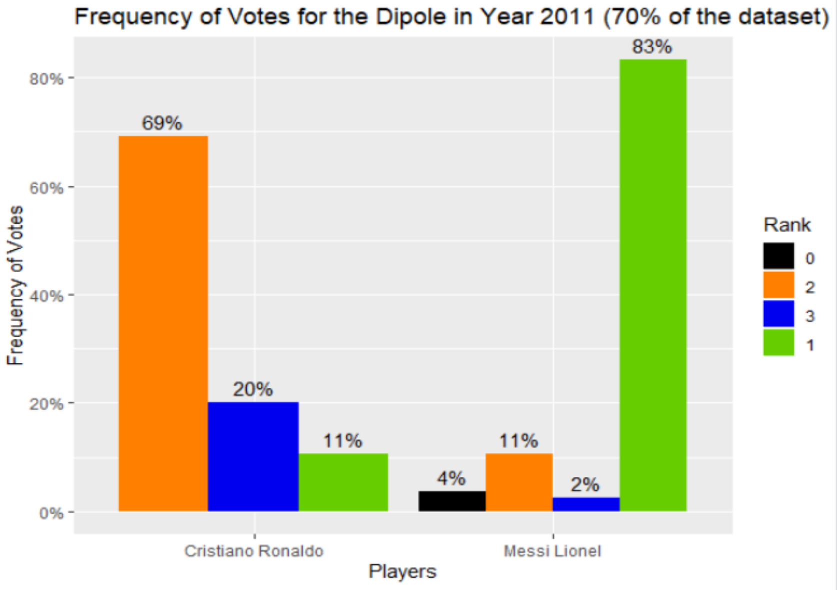


Figure 12 : Frequency of Votes for Messi and Ronaldo in Cluster 1, Year 2011.

Based on Figure 12, we can observe that the 83% of the Messi's votes are first - place votes and the 69% of Cristiano's votes are second - place votes. Thus, it is obvious the ascendance of Messi in this group, which composes the 70% of the dataset.

The Cluster 2 is consisted of 143 voters. The preferences regarding the winner are not different in compare with the first group, since Messi has been ranked first from 96 out of 143 voters. This is depicted also from the medoid, Ernst Hasler, who has ranked also Messi. The difference between the two groups is that in the second one Cristiano has not been preferred even from one voter. Also, Xavi and Iniesta have larger percentages of votes in compare with Cluster 1. Thus, in Cluster 2, is observed a preference in Barcelona, which has been spotted also in the Bayesian approach in 2011. If we search the job and the continent of this voting group (B.13), it is not noticed any specific pattern except the fact that it contains many journalists from Africa and coaches from Asia.

6.3.3 <u>Year 2012</u>

Year 2012 is the second consecutive year that Lionel Messi wins the trophy, having a great lead from his opponents. Based on the average silhouette plot (B.14) is indicated clearly that the optimal number of components for the clustering model is 4. This is due to the fact that the average silhouette width when the line reaches 4 number of components, is 0.56 which is the highest value of the line chart. Thus, the clustering in 2012 takes place on a 4 components model. It can be observed from the silhouette plot (B.15) that most of the observations are appropriately clustered as the average silhouette value for each specific cluster is the following : Cluster 1 (0.54), Cluster 2 (0.66), Cluster 3 (0.58) and Cluster 4 (0.67). It has to be mentioned at this point that it is reasonable for Cluster 1 to have smaller silhouette value than the rest of the groups, as it is the largest one by having 410 observations, when Cluster 2 holds 52, Cluster 3 holds 25 and Cluster 4 holds 18 observations. Thus, the smaller clusters are most solid since they have a common characteristic, in compare with a bigger cluster which is reasonable to contain some noise.

The first cluster constitutes the 68% of the total dataset. Thus, its information is very important. By having look in the frequency table obtained for the votes of Cluster 1 (A.32) it is obvious that Messi has the absolute control of the rankings. In particular, he has been voted 256 times first out of 410 and he has not been preferred from any voter only 33 times. On the other side, his opponent Cristiano, has been voted with 5 points from 70 judges and he has not been ranked 103 times. Moreover, the fact that he has been preferred in the second - place from 160 voters is a strong indication for the possession of the second - place in the final rankings. In (B.16), it is presented a graphical representation of these conclusions for Cluster 1.

After observing the rest of the clusters someone can notice that the main preferences of the voters are not different than those of Cluster 1. Again, in these clusters Messi receives the largest amount of first votes among his opponents. What is really interesting in those clusters is the fact that all of them have been created depending on a specific player, for each one of the three, who has not ranked in such high positions in the final rankings. That means that each of these three clusters can be characterized based on the corresponding player. To be more specific, in Cluster 2 (A.33), Andrea Pirlo has been preferred from all of the judges and he received 5 first - place votes, 15 second - place votes and 28 third - place votes. Moreover, this has been indicated from the medoid of Cluster 2, Siamak Rahmani, who ranked the Italian legend in the second - place. Cluster 3 (A.34) could be characterized as Zlatan's cluster. Zlatan Ibrahimovic has been preferred from all of the 25 objects of the cluster and has been ranked 5 times first, 6 times second and 14 times third. Also, in this cluster, the representative voter, Rat Razvan, has ranked Ibrahimovic second. In the final group of voters, someone can observe the strong presence of Neymar. The Brazilian football player has been ranked 5 times second and 13 times third and he has been chosen in the rankings from all of the voters. Felipe Baloy, who is the medoid of Cluster 4 can be an strong indication as he has ranked Neymar third.

It would be very interesting to dig into the clusters and trying to identify a possible pattern between the vote of a judge to a player that is not such possible to win the Ballon D' Or. Thus, the attempt is to spot a relationship between the job or the continent of the voter and his vote, in such cases. After observing the stacked bar charts, which represent the job and the continent for each specific cluster, we are able to identify a pattern between the continent that a voter comes from and his vote. More specifically, in Pirlo’s cluster (B.17), someone can observe that the largest proportion of voters, regarding the actual size of the continent, comes from Europe. Moreover, in Ibrahimovic’s cluster (B.18), someone can notice that more than a half of the voters come from Europe. Finally, in Neymar’s cluster (B.19), it is easy to observe that the 2/3 of the voters come from America. If we take under consideration that Pirlo is Italian, Ibrahimovic is Swedish and Neymar is Brazilian, we can state that it is observed a pattern between the vote, the continent of the player and the continent that the voter comes from.

6.3.4 Year 2013

The average silhouette plot for the different number of examined clusters (B.20) indicates that the 3 – components model is the most appropriate, in terms of the criteria that have been discussed. This is due to the fact that the value in the 3 clusters is slight larger than the one of the 2 clusters and the largest among the values. Moreover, the line drops suddenly from 3 to 4 clusters. Besides that, the plot of the Elbow method (B.21) indicates an ‘angle’ from 3 to 4 clusters, where the fall of within sum square starts to be smoother. Thus, based on these indications we are going to implement the clustering process with 3 groups.

Year 2013 was the first year that Cristiano Ronaldo won the award. Besides that, it was the first year among the examined period, that the margins between the first, the second and third - place were such small. The table of the results of Cluster 1 (A.35) indicate a voting pattern that exists in this group. In specific, the first group is consisted of 468 observations, from which the 395 voters have preferred either Cristiano or Ribery or Messi in the first - place (B.22). Because of the size of this group (86% of the dataset), it is obvious that the voting pattern that this group provides is a very strong indication for the final rankings. The peculiar here is that Ribery has been preferred first more times than Cristiano, despite the fact that he has been placed third in the final rankings. This can be occurred from this cluster’s medoid, Cunliffe Jason, who ranked Cristiano first, Ribery second and Messi third. The reason why the Frenchman has been finally ranked third is the ‘extreme’ ranking behavior that the Ribery’s voters have, which has been presented also in the Bayesian approach in the same year. By looking at the tables of the Clustering results and sum the first - place votes for each of the two players, it occurs that Messi has received in total 119 first - place votes and Ribery 163 first - place votes. But if we sum the total second - place votes that each one has received, it occurs that Messi has received 175 second - place votes and Ribery 78. The third - place rankings are quite similar with the previous case. Thus, it is observed that the total voters that preferred Messi are more than those who preferred Ribery. As in the Bayesian approach, the voters of Messi are observed to be more robust and have been split in first and second - place, in contrast to Ribery’s voters who have been gathered in first - place without a mass participation in the second one. Since this specific ranking behavior and its results are confirmed also in this approach, it would be interesting for someone to implement a research in such ranking behaviors in more topics like elections that make use of rankings, surveys etc.

Besides Cluster1, the results of the third cluster are also interesting. Despite the fact that Cluster 3 is small (37 observations), Messi receives more first - place votes than his two opponents (A.36) and also the key player in this cluster is Iniesta who has been preferred first from 12 voters, second from 9 voters and third from 16 voters despite the fact that he has been placed 17th in the final rankings. Thus, someone could state that this is a group of Barcelona fans. By looking at the stacked bar chart in (B.23), that the presence of Media in this groups is negligible. Also, most of the voters are coaches that come from Africa, America and players that come from Asia.

6.3.5 <u>Year 2014</u>

It is the second consecutive year that Cristiano Ronaldo is awarded with the 'FIFA Ballon D'Or' trophy, with a great lead to the other challengers of the title. Beside that it is the first year that we are going to perform clustering with many groups. In specific, the portioning around the medoids is going to be implemented for 9 groups. This occurs from the line chart of the average silhouette vale for the corresponding clusters (B.24), which peak is in the 9 components having almost 0.4 as value. It can not be observed any 'breakpoint' before the 9 components, in order to perform the clustering with less number of groups, thus it is going to be performed with the number of components that have the highest silhouette value. Moreover, the Elbow method (B.25) does not provide us either with a sufficient 'breakpoint', except a small one in the point of two clusters.

After performing the k – medoids algorithm, we observe that the output provides one large cluster which contains 280 observations, a smaller one which contains 97 observations and seven other small clusters. The visualization of the clustered points is provided in (B.26) accompanied with the silhouette plot of each specific cluster (B.27), with the average silhouette value being 0.39 .

After observing the created clusters and the medoids of them, someone can notice that there is one big cluster (A.37) that consists the 50% of the total dataset and is the depiction of the win of Cristiano against his opponents. A graphical representation of this cluster is presenting in the following bar chart.

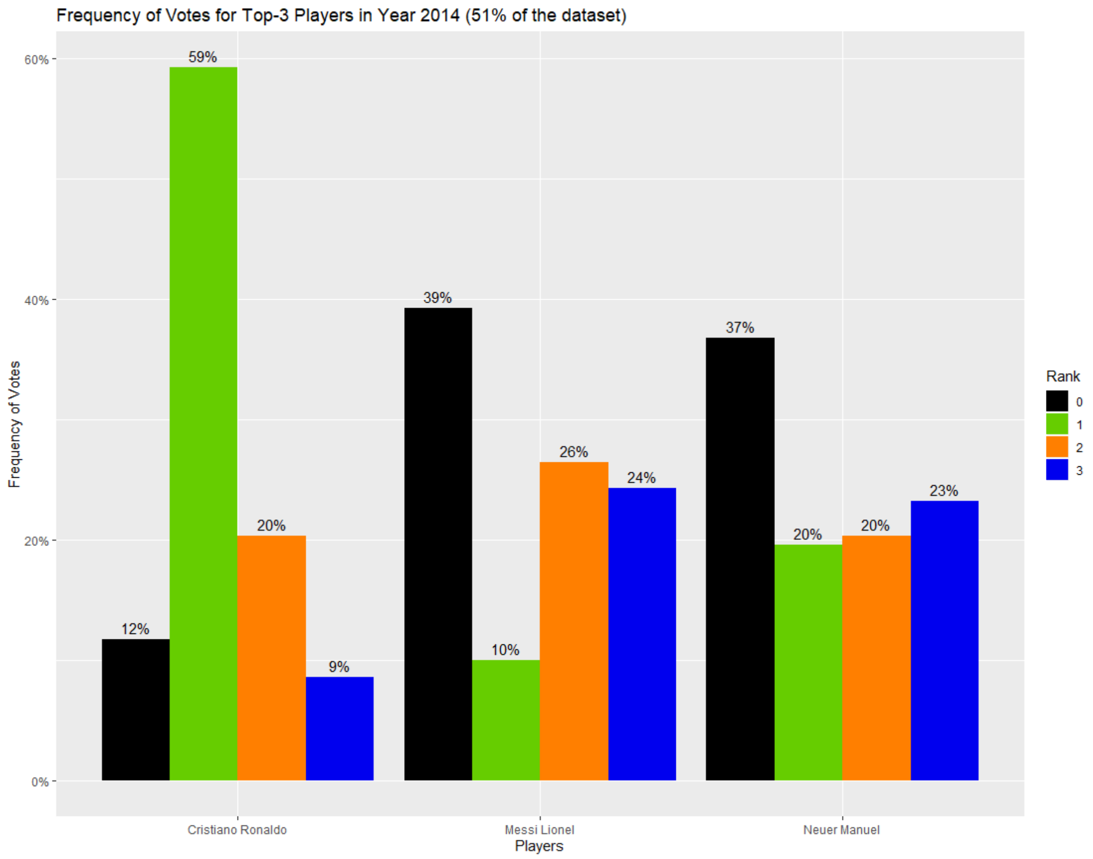


Figure 13 : Frequency of Votes for the first three ranked players in Cluster 1, for the Year 2011.

In Figure 13, it can be observed that Cristiano holds the lion's share in terms of the overall and first - place votes. Also in this figure it can be noticed that Neuer has been preferred in the first - place from exactly the double voters than Messi. If the final rankings, where Messi placed 2nd and Neuer 3rd ,would be taken under consideration we could obtain that it is noted a similar ranking behavior like this one in 2013 between Messi and Ribery. More specifically, in the other 8 small groups Messi receives much more second - place votes than Neuer, who keeps having more first - place votes than the Argentinian. In the overall sum Messi ends up with more points than Neuer, as his votes are more normally distributed than the German goalkeeper's whose votes are mainly skewed in the first - place . Thus, the conclusion in both cases (2013 and 2014) is that Messi is almost always in the top three, despite the actual rank, while on the other hand Neuer and Ribery have not high voting frequency in 2nd and 3rd place.

Besides Cluster 1, in the remaining 8 clusters a player holds a key role for the creation of each cluster. Thus, in each cluster there is always a player that have been preferred in the Top – 3 from all the voters of the cluster. Because this case is quite similar to the case of the small clusters in 2012, it would be very interesting to identify if the pattern between the origins of the player and the origins of the voter, that has been observed in 2012, is applied also in 2014. After implementing the stacked bar charts for each of the 8 remaining clusters, we observe that this allegation is also applied in most of these clusters. To be more specific, in Cluster 3 (A.38), the key player is Neymar who comes from Brazil. If someone looks on (B.28), it is evident that most of the voters come from

America. Moreover, in Cluster 5 (A.39), the key player is James Rodriguez who comes from Colombia. By observing the (B.29) , one can observe that despite the job each one has, almost the 2/3 of the voters come from America. Exactly the same thing happens in Cluster 6 (B.30), where the key player is the Argentinian midfielder Di Maria. Also, almost the total amount of votes that Yaya Toure got in Cluster 7 (B.31) are from people that come from Africa. Finally, in Cluster 8 (B.32), the German Bastian Schweinsteiger, has been ranked mainly from Europeans. Thus, from all the above mentioned, there is a strong evidence which indicates that when a big cluster, which provides the information about the winner of the year exists, then the remaining small clusters are shaped driven by the origins relationship between the player to be ranked and the voter.

6.3.6 Year 2015

In 2015, both the Silhouette method and the Elbow method provide the information that the optimal number of clusters is 2. This is due to the fact that in the point of 2 clusters the average silhouette plot (B.33) for the different number of components indicates the highest value (0.51) and the line chart of the Elbow method (B.34) indicates a ‘breakpoint’ in 2 clusters. Thus, we are going to analyze the clustering results of the two components model.

The average silhouette plot for each specific cluster (B.35), denotes a good clustering (Cluster 1 : 0.50, Cluster 2 : 0.62). On the other hand, the two clusters are completely different in terms of the size. Cluster 1, consists of 473 observations and Cluster 2, 25 observations. In Cluster 1, which is the 95% of the total dataset, is displayed the absolute ascendance of Lionel Messi and his return to the awards. In (B.36), someone can observe that the green bar which represents the first - place votes that Messi has received, is more than 3 times bigger than Cristiano’s green bar. In specific, the 64.69% of the total votes that the Argentinian player has received were first - place , in compare to the 19.45% of Cristiano. Thus, the figure that represents the votes of Cluster 1 for the Top – 3 does not allow for any doubt about the winner of trophy. Messi is the complete preponderant of the competition. The other cluster of the final output is very small and not able to figure a different result. But it has to be pointed out that it reveals a pattern for a specific group of voters who have preferred Alexis Sanchez. The Chilean footballer, was selected in the Top – 3 from all of the persons that exist in Cluster 2. In specific, he received 2 first – place votes, 10 second – place votes and 13 third – place votes, while his position in the final rankings was the 10th . Thus, this cluster has been curved with Alexis Sanchez at its center. This is also confirmed from the medoid of Cluster 2, Mahamud Raihan, who has ranked third Alexis Sanchez behind Messi and Cristiano Ronaldo.

Chapter 7

Cluster Analysis on the Data with the Insertion Sorting Rank (ISR) algorithm

Jacques and Biernacki (2012) [13], proposed an alternative model, which scope is much more wider than the algorithms that were presented before. The Insertion Sorting Rank (ISR) is an algorithm, with a model - based approach, which has a very wide application scope in the context of clustering ranking data. The ISR model is set up by modelling the ranking generating process, assumed to be a sorting algorithm in which a stochastic event has been introduced at each comparison between two objects [14].

Like the methods that have been presented up to now, the ISR algorithm is about clustering data that they are not able to be modelled in a straight forward nonparametric way. Such categories are the ranking data that occur from heterogeneous populations (different political meaning, different strategies in marketing research, etc.), partial and multivariate ranking data. The purpose of this algorithm is to cluster ranking datasets which contain complete or incomplete ranking, multivariate or univariate. In our case, there exist partial ranking data. The missing entries, in the case of partial ranking data, are considered as missing values and inferred in the estimation process. In the case of multivariate ranking data, the algorithm is based on an extension of the ISR model, which allows the presence of multivariate ranking data under a conditional independence assumption on the components of these data. The algorithm was first introduced for univariate rankings and by using of the extension of ISR and the conditional assumption can take into account the multivariate case, where many dimensions are tanking part in the analysis.

7.1 ***Theoretical Framework of the Method***

7.1.1 ***Model based algorithms, Finite Mixture Models, Latent Class Models***

It was referred in the above paragraph, that the ISR model is a model – based clustering approach. This is for the multivariate case of ranking data, because in that case a finite mixture model is taken under consideration. In the Finite Mixture Models method, the distribution f of the variable X is considered as a mixture of K distributions $(f_1, \dots, f_k)$: $f(x, \theta) = \sum_{k=1}^{K} \pi_k f_k(x, \theta_k)$, where θ is a vector of parameters $\theta = (\pi', \theta_1', \dots, \theta_k')'$, π_k is a proportion of k_{th} distribution in the mixture and θ_k is a parameter of f_k distribution [54]. So, in this case the clusters are not being found by a chosen distance measure, like in distance – based models, but a probabilistic model is obtained to describe the structure of the ranking data. Thus, the name of this approach is "model – based clustering". We are going to see, in the multivariate instance of ranking data, the way that the analysis is going to be performed with the use of the Finite Mixture Models approach.

The finite mixtures models, as have been described in a previous section, are type of latent variable models, where the heterogeneity of the population is assumed to be resulted from the existence of two or more distinct homogeneous subgroups or latent classes. The latent class model, is a specific case of multivariate discrete categorical data, where a set of observed multivariate discrete categorical variables is related to a set of latent variables [48]. The latent class analysis, tries to find patterns, groups or subtypes of cases in such data (e.g multivariate ranking data). A class is characterized by a pattern of conditional probabilities that indicate the chance that variables take on certain values.

7.1.2 Univariate and Multivariate ISR model

Let’s first explain the difference between univariate and multivariate ranking data in the case of the FIFA Ballon d’or data. In the case of clustering the judges – voters for each specific year we construct a matrix which represents the ranking of each player that was assigned by the corresponding voter each time, for this specific year (e.g Year 2010). In this instance, the analysis is under the univariate ranking data case because the preferences of one year are taken into account in the analysis. In contrast, if the purpose of the analysis was to detect patterns, in a period of years (e.g 2010 - 15) and not for one specific year, then the ranking dataset that is going to be analyzed is composed of multivariate rankings, where each dimension represents a year and the input dataset contains the rankings for each dimension.

7.1.2.1 The Univariate case

The ISR model for the case of univariate ranking data is obtained when the assumption that a rank datum is the result of sorting algorithms based on pair comparisons, is taken under consideration. Then the formula of the ISR model is the following :

$$\mathrm{p}(x;\mu,\pi)=\frac{1}{m!}\sum_{y\in\mathcal{P}_m}\mathrm{p}(x|y;\mu,\pi)=\frac{1}{m!}\sum_{y\in\mathcal{P}_m}\pi^{G(x,y,\mu)}(1-\pi)^{A(x,y)-G(x,y,\mu)};$$

In order to make clear the formula of ISR model for univariate rankings, the explanation of the parameters of the formula has to be given.

$x = (x^1, \dots, x^m) \in P_m$, where x is the ordering representation of the resulting ranking of the objects $O_1, \dots, O_m$.

P_m is the set of permutations of the first m integers.

$\mu \in P_m$ is the modal ranking or reference/central ranking. Modal ranking is the sequence of ranks that has the highest probability to occur. If we denote the modal ranking as π_0 , the rankings that are most observed in a dataset are close to π_0 .

$\pi \in [\frac{1}{2}, 1]$ is the probability of good pair comparison according to μ .

The sum over $y \in P_m$ represents all the possible initial presentations orders to rank , with identical prior probabilities equal to 1/m! .

$G\,(x\,, y, \mu\,)$ is equal to the number of good paired comparisons during the sorting process

$A\,(x, y)$ corresponds to the total number of paired comparisons that have been implemented .

7.1.2.2 The Multivariate case

In the case of multivariate ranking data, the multivariate rank $x = (x^1, \dots, x^p)$ is a vector where each component of the multivariate rank is a vector $x^j = \left(x^{j1}, \dots, x^{jmj}\right), 1 \leq j \leq p$ which corresponds to the ranking of each one of the p dimensions. The conditional independence assumption of the ISR model for multivariate ranking data is that the population of the multivariate ranks is composed of K groups in proportions p_k, where the sum of the proportions to the K groups is equal to 1 [15]. Thus, based on this assumption, the components of x can be assumed to be sampled from independent ISR distributions with a corresponding modal ranking and good paired comparison probability for each one of the p dimensions [15] . This conditional independence assumption is called latent class model and it can be considered such as, since rankings are a specific category of categorical data, as we have seen in previous paragraph.

7.1.3 Estimation of the Model

The ISR model, uses Maximum Likelihood Estimation (MLE) in order to obtain inference. The MLE method, estimates the parameters of the probability distribution by maximizing a likelihood function, so that under the statistical model that is assumed in this specific case, the observed data are most probable.

In the case of multivariate rankings, we can approach the estimation of the groups by assuming a binary latent variable which records the group membership of the observations of the dataset and takes the value 1 if the observation belongs to a certain group and 0 otherwise [13]. So, for each one of the observations of a set there is a latent variable which demonstrates if the observation belongs or not to a specific group. Let's assume that $x = \{x_1, \dots, x_n\}$ is a sample of multivariate rankings, $z = (z^1, \dots, z^K)$ is the set of the corresponding latent variables for K groups and $y = \{y_1, \dots, y_n\}$, where $y_i = (y_i^1, \dots, y_i^P) \in P_{m_1} \times \dots \times P_{m_p}$, are the presentation orders of the objects for the i_{th} observation [13]. Assuming the triplets (x_i, y_i, z_i) arise independently, the data log likelihood model is obtained. The problem is that the maximization of this likelihood is not easy to be done, because of missing data. Then, a solution in order to deal with the missing data, is to consider an Expectation – Maximization (EM) algorithm. The EM algorithm has the advantage of stability in terms of occurrence of missing data and it requires the computation of the conditional expectation of the complete – data log – likelihood function given the observed data, at E step, and then the maximization of the likelihood function with respect to the parameters of interest, at M step [29]. But, the complete log – likelihood function is not linear for the types of x_i, y_i, z_i , which is an obstacle for the Expectation step of EM algorithm.

7.1.4 SEM – Gibbs algorithm

In order to overcome the issue in the E step of EM algorithm, the SEM – Gibbs algorithm is used. The SEM algorithm, generates the latent variables y_i, z_i and the unobserved positions of x_i ($\hat{x}_i$)from the conditional probabilities that were computed in E step, in the stochastic step (S step). The advantage of the SEM – Gibbs algorithm in contrast to the EM algorithm, is that these latent variables are generated without calculating conditional probabilities at E step, which leads to reducing the computational complexity by removing the complicated use of the products of missing data. The algorithm achieves this result because of the use of a Gibbs sampling.

Gibbs sampling is a randomized algorithm, which is used especially when the direct sampling is not a straightforward process and consists of drawing samples ($\hat{x}_i, y_i, z_i$) consecutively from the full conditional posterior probabilities [31]. The generic idea is to resample one variable at a time conditional to the others, by initializing the algorithms with random numbers. It is a Markov chain Monte Carlo algorithm (MCMC), for obtaining a sequence of observation which are approximately from a specified multivariate probability distribution [31].

So, the SEM -Gibbs algorithm consists of two steps (SE – Gibbs step and M step), from which the first step is consisted of three sub – steps . In the first step (SE – Gibbs step), is considered a Gibbs sampler generating a chain e.g for generating y_i, the chain is $y_i^{j\{q,0\}}, \ldots, y_i^{j\{q,R_j\}}$, in which the last value $y_i^{j\{q,R_j\}}$ is retained for $y_i^{j\{q\}}$. In order for the $y_i^{j\{q,R_j\}}$ value to be retained, the size of $R_j$has to be greater than $\frac{m_j\,(m_j-1)}{2}$, which is the maximum Kendall distance between two ranks of size m_j, so that any rank of P_{m_j} can be reached with non – null probability for any arbitrary initialization. Starting from $y_i^{j\{q,0\}} = y_i^{j\{q-1\}}$, the Gibbs sampler generates $r \in \{1, \ldots, R_j\}$ sequences $y_i^{j\{q,r\}}$. Thus, for the incomplete rankings, which are considered as missing data in the algorithm's procedure, the corresponding full rankings are estimated by using this Gibbs chain. The algorithm runs for a number of iterations. The same process is followed for the other two sub – steps, for y_i and $\hat{x}_i$. The second step (M step) of the algorithm consists in computing the parameter value $\theta^{\{q\}}$ which maximizes the completed log – likelihood computed at the previous step. The parameter value $\theta^{\{q\}}$ is defined as $\theta^{\{q\}} = argmax_{\theta\in\Theta} l_c\big(\theta; \{x, \hat{x}^{\{q\}}\}, y^{\{q\}}, z^{\{q\}}\big)$, where $\hat{x}^{\{q\}}, y^{\{q\}}, z^{\{q\}}$ are simulated in the first step (E step) .

7.1.5 Determine the final model

The total procedure is going to run for a number of clusters in order to choose among the models with the different clusters which model is more ‘appropriate’, for each specific year. In other words, the number of groups that will separate the voters, have to be defined. Thus, in order to detect the optimal number of clusters, the Bayesian Information Criterion is going to be used.

The Bayesian Information Criterion or BIC is a very well-known criterion for model selection among a finite set of models. When fitting models, in order to increase the likelihood, it is possible to add more parameters in the model which may lead to overfit. The BIC introduces a penalty term for the number of parameters that exist in a model, in order to protect it by overfitting. It is closely related to Akaike Information Criterion or AIC, as they both are penalized – likelihood criteria. The AIC measure tries to balance between the model accuracy and model complexity, as it uses the maximum likelihood estimate and the number of parameters, in order to estimate the information lost in the model. It can be observed that the goal of the two criteria is pretty similar, as both try to prevent from overfitting. The BIC measure can be defined as BIC = $-2\ln(\hat{L}) + \ln(n)\,k$, where $\hat{L}$ is the maximized value of the likelihood function of the model, n is the number of data points and k is the number of free parameters to be estimated [58]. In order to detect the optimal number of clusters for the models, we are going to create line plots where in the x – axis are going to be the number of examined clusters and in the y – axis are going to be the values of BIC. The classical decision, in terms of the BIC value, is to choose the number of clusters that minimize the BIC value. But it has to be pointed out that the choice of the number of clusters for each year’s model has to be an integration of small BIC and proper number of groups, in terms of the interpretation of the voting behavior. In other words, if a model with 2 groups has slight smaller BIC value than a model with 3 groups, we are not going to select the model with the 2 groups as a straightforward process, because the model with 3 groups could reveal one more pattern of voting behavior which may be useful for the analysis. For that reason, in such cases where the absolute differences of the values between models of different number of components are not great, we will try to find a ‘knee – point’ where after this point the BIC values will increase more sharply, in compare with the previous number of groups. Thus, in such cases we will based on this point and to a number of components that will be helpful for the voting patterns detection.

7.2 ***Application of the Method***

7.2.1 ***Package Overview***

In this chapter, is presented the Application of the model – based approach that has been discussed in the previous section. Thus, in order to implement the Insertion Sorting Rank algorithm in the FIFA Ballon d’Or data, for the period 2010 – 15, we make use of the ‘Rankcluster’ package. The ‘Rankcluster’ package, was first released in 02/09/2013 as a package that could take into account both multivariate and partial ranking data, through the implementation of a model – based clustering algorithm. The algorithm is working by taking into account the heterogeneity of the rank population that is modelled. This is achieved with a conditional independence assumption that is considered for the multivariate rankings.

7.2.2 Data Input Format

The input data of the 'Rankcluster' package, have to be given in specific representation. The ranking representation $r = (r_1, \dots, r_m)$, where r_m is the rank of the m – th object, contains the ranks assigned to the objects from one judge, and means that the $i-$ th object is in the $r_i -$ th position. In Rankcluster's functions, ranks have to be given in the ranking notation. Thus, the input data parameter must be a matrix, with every row corresponding to a rank. The missing positions that occurred because of the partial ranks, should be denoted by 0. Also, 1 indicates the most – liked alternative, 2 indicates the second most – liked object and 3 denotes the third most – preferred player.

7.2.3 Estimation of the Model

In order to perform this model – based clustering method to the partial ranking data and obtain estimations for the potential group that each voter belongs to, based on a mixture of ISR model that was proposed in the theoretical part of the application, we are going to use the 'rankclust' function from the 'Rankcluster' package.

The arguments that are used in the function, in order to obtain the result, are the following :

- *data* : A data matrix, where each row is a ranking and the missing elements are denoted with 0 or NA. As it has been already mentioned previously, the data must be in the ranking notation.
- *m* : The number of columns of the data matrix.
- *K* : An integer or a vector of integers with the number of clusters that are going to be obtained. The algorithm is going to obtain clustering results for each one of the number of desired clusters.
- *criterion* : The penalty criterion that is going to be used in order to the appropriate number of cluster being chosen. The possible choices are the 'BIC' and the 'ICL' criterion.
- *Qsem* : The total number of iterations for which the SEM algorithm is going to be repeated.
- *Bsem* : The value of burn – in period for SEM algorithm. As burn – in period is described the practice of throwing away some iterations, before the algorithm is going to run normally by using each iteration in the calculations [7]. The name 'burn – in' comes from electronics, where many electronic components fail quickly and those which don't, is a more reliable subset. Thus, a burn – in is done in the factory to eliminate the worst ones [7].
- *Ql* : The number of iterations of the Gibbs sampler for estimation of log – likelihood.
- *Bl* : The burn – in period for the estimation of the log – likelihood.

- *maxTry* : The maximum number that the algorithm is being put for restart, in case of non convergence.
- *run* : The number of runs of the algorithm for each number of clusters is given by the value of K.

In our case, the values that are being set as input are the following :

- ✓ As *data*, has been set the matrix with the partial rankings of the players, for each of the years 2010 – 15. The matrix that is used each time, corresponds to the year for which the algorithm is implemented.

- ✓ As *K*, we set the vector 1:5. That means that the algorithm is going to obtain results for 1 cluster, 2 clusters, etc., up to 5 clusters. The reason that these integers have been selected, is the computational burden of the algorithm. As the number of clusters, that are included in the runs, increases, the computation of the results tends to become slower and the endurance of the machine used for the computation decreases. Thus the number of times - number of clusters that the ISR algorithm is going to run, should be in compliance with these constraints.

- ✓ The criterion that has been chosen for the selection of the best model, among the models with the different number of clusters, is the BIC penalty criterion.

- ✓ For the rest of the arguments, have been used values that are compiled with the restricted resources due to the computational complexity of the algorithm and, at the same time, are able to produce results that correspond to the goals of the analysis. The algorithm has been put to run 2 times, for each specific number of clusters. Also, the algorithm has been set to restart up to 3 times, in case of non convergence.

The output of the run is stored in a different variable, for each of the years 2010 – 15. These outputs contain a bunch of information for the clustering results and the different distances between the estimations and the current values, and can be approached from the slots of the output's class. Among this information, the summary of the clustering result contains the observations with the highest probability and highest entropy, for each cluster. The probability is estimated by using the last simulation of the presentation orders used for the likelihood approximation and its output exhibits the best representative of each cluster. On the other hand, the entropy output illustrates the less confidence in the clustering of each observation. Thus, the observations with the highest probability can be considered as the voting representatives of the rest of the voters in the cluster and the observations with the highest entropy can be considered as the less confident and representative voters in that specific cluster.

7.2.4 <u>Computational Time</u>

Before providing the algorithm's obtained results, it has to be mentioned the computational time needed for the implementation of the ISR method. This stands for the general evaluation of the clustering method by taking under consideration the difficulties in the implementation of the algorithm, in terms of the machine e.g execution time, requirements of machine's capabilities.

From the computational point of view, the FIFA Ballon D'Or datasets are challenging since the size of objects that are about to be ranked is large (= 23) and the presence of partial rankings is also sizeable (the percentage of the ranking elements, that is missing, is greater than 80% of each dataset). For this reason, a small number of iterations (Qsem = 100, Ql = 300) has been chosen with a respectively 1:5 clusters, in order to eliminate the computational time and the sources that were required for the drawing of the results. With these iteration numbers and clusters that the algorithm run for, took about 8 – 9 hours per run (laptop 1.30GHz CPU). At the same time, the implementation of another process in the machine while the algorithm was still running, was very slow and almost unachievable.

Most of this computing time is consumed in the likelihood approximation, at each run of the algorithm. The reason is the high proportion of missing elements, which leads to a large number of different modal rankings, simulated during the SEM – Gibbs algorithm and then to a large number of likelihood approximations. It has to be mentioned at this point that, the retained parameters at the end of the estimation algorithm are those leading the highest approximated likelihood. Another fact that is has to be referred is that the small number of iterations probably makes the implementation of the algorithm feasible but, at the same time, the variabilities of parameters estimations are expected to be larger.

7.3 *Results*

7.3.1 Year 2010

In Year 2010, the plot of BIC for the different number of groups is the following :

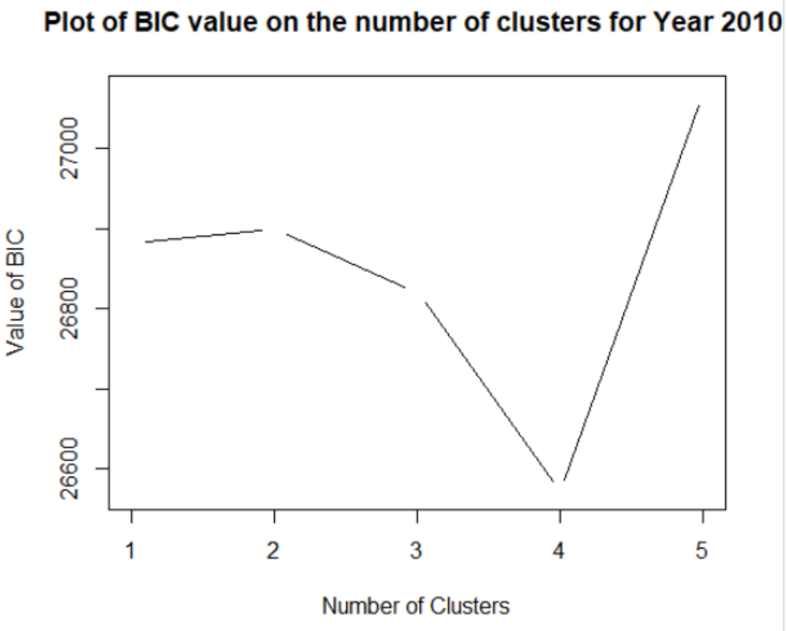


Figure 14 : The BIC plot used to determine the final model's number of clusters.

By having a look on the Figure 14, someone can observe that the model with 4 components provides the smallest BIC value. Moreover, we can notice that the absolute differences between the values of BIC for different number of components are small. Based on the smallest value of BIC, the 4 – components model is going to be selected as the final model.

The proportion of observations in each cluster is the following: Cluster 1 – 6.35% (23 observations), Cluster 2 – 34% (150 observations), Cluster 3 – 15.06% (65 observations) and Cluster 4 – 44.47% (187 observations). It has to be mentioned that, Year 2010 is very different in compare with other years. It is the only year, in the period under study, that the dipole Messi – Ronaldo did not exist and Cristiano was not even in the Top – 5 players. This is the reason that justifies the presence of players that are not in the Top – 3 in most of the remaining years. An example of such a player is Diego Forlan, who has been ranked fifth in the final rankings of 2010, and has received the largest amount of votes in Cluster 1 (A.40).

By looking at the tables that represent the votes in this year, someone can observe that the voting difference between the three claimants of the title is small. This can be also viewed by observing the four different clusters that have been shaped and illustrate the preferences of the voters. Especially, in the bar chart that represents the percentages of votes for Messi, Iniesta and Xavi in Cluster 4 (B.37), is depicted the small difference in votes. The group that displays the win of Messi is the second (B.38), where the Argentinian has much bigger proportion of first - place votes in compare with his opponents, while the amount of second and third - place votes are not such different between them. Besides the fact that the winner of the trophy was a Barcelona player, 2010 is a year that all the Top – 3 players were playing for Barcelona. The preference stream for Barcelona players is not a surprise if someone considers that the season 2009 – 10, was a very good seasons for the Catalan team. This can be asserted also from the fact that in all the plots that represent the job of the voters and the continent that the voters come from, has not been detected a pattern that reveals any relationship between the vote and someone of these two factors (B.39, B.40, B.41, B.42). This can be interpreted as a strong indication for the assumption that in year 2010, the job and the origins of the voters did not affect their final preferences.

7.3.2 <u>Year 2011</u>

Based on the BIC plot of Year 2011 (B.43), for the different number of components, we conclude that the optimal number of clusters to fit the model is 2, because the line reaches the smallest value of BIC when the examined groups are two. Thus, we perform the ISR approach for two clusters.

The lead of Messi, in terms of votes, in this year is extraordinary. In the following figure is presented the distribution of votes for the Top - 6 nominees for the 2011 FIFA Ballon D’Or. The plot has been constructed to describe the Cluster 2 which consists the 73% of the total dataset.

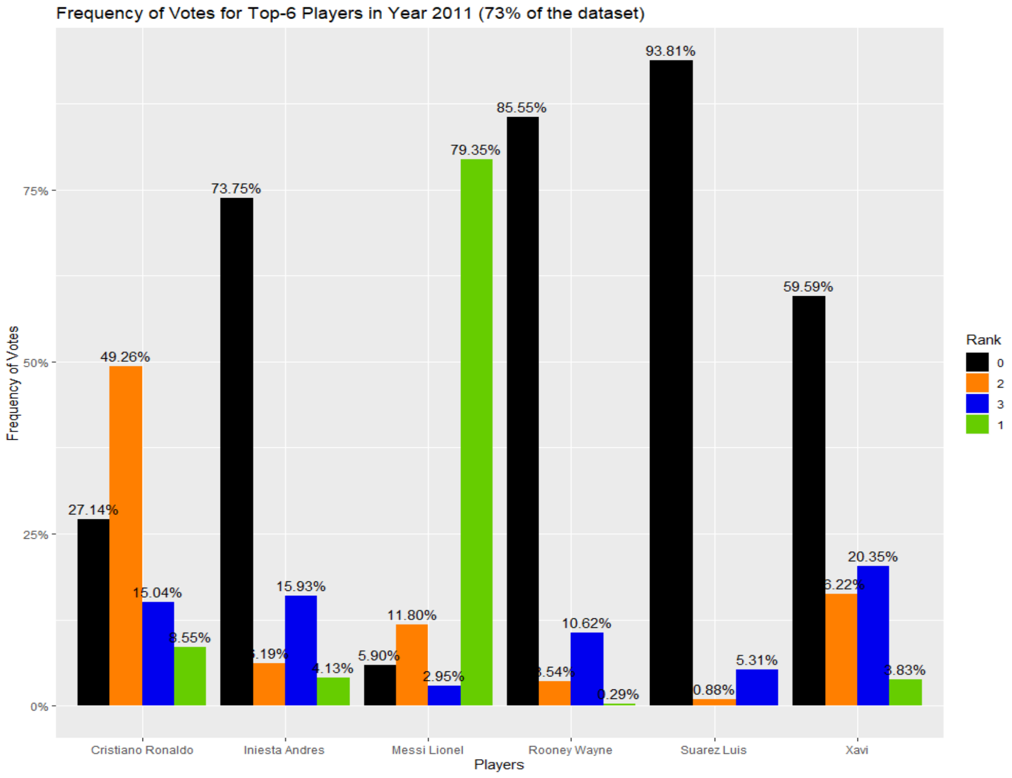


Figure 15 : Frequency of Votes for the Top - 6 in Cluster 2, Year 2011.

From Figure 15, someone can observe that Messi has got an amount of first - place votes which is more than the lion’s share. The orange bar of Cristiano Ronaldo illustrates that he has been ranked after Messi from almost the 50% of the Cluster. Moreover, the black bars of the rest 4 players indicate that, despite the fact that they frame the Top – 6, the sum of total points they have got is very poor. The ascendancy of Messi in Year 2011 is confirmed also from the objects that have the highest probability and highest entropy. It is observed that, despite the fact that the voters with highest probability are the most representative objects of a cluster and the voters with highest entropy are the less representative objects of a cluster, the 4 different objects that have the highest probability and entropy for each of the two groups contain the same values in the first two positions and differ in the third. Not surprisingly, in the first position is ranked Lionel Messi and in the second one Cristiano Ronaldo. This indicates that the influence of Messi in the first position and Cristiano in the second, is that big, that neither the less representative object of a cluster has voted in a different way for these two places.

7.3.3 Year 2012

As in previous years, in order to decide the appropriate number of clusters for which we are going to implement the Insertion Sorting Rank algorithm, we observe the plot of BIC versus the candidate components. Based on it (B.44), the model is fit in 3 clusters. This happens because from the point of 3 clusters to the one of the 4 clusters it is observed a sharp increase of the BIC value, which becomes sharper from the point of 4 clusters to the one of 5 clusters. Thus, we choose the 3 components model as the best, in terms of the BIC increase.

It has been seen also in the previous approaches that, 2012 was the third consecutive year that Lionel Messi won the award and the second in a row with such a big lead from his opponents. Besides that, the dipole Messi – Ronaldo is getting more and more stable in the first two positions of the final rankings. By looking at the tables (A.41, A.42, A.43) that present the votes for all the players in each cluster, it is very easy to observe that the captain of Barcelona is first in the preferences in all of these clusters and Cristiano is also the second choice for the largest amount of voters. Moreover, from the bar charts that present the votes frequency for each of the Top – 4 players (B.45, B.46, B.47), in every cluster, it can be seen that the amount of voters that preferred a player in the first - place , second - place, etc. and the amount of those that they did not prefer a player in the first three positions, are proportionally similar in all the three clusters. What is really remarkable in this year, is the fact that the people that prefer a player who is outside the dipole are very few. Even the black bars of Iniesta and Xavi, who were ranked, correspondingly, third and fourth in the final rankings, are very large. Especially the amount of voters that did not prefer Xavi (4th in the final rankings) is greater than 80% of the total voters, in each cluster. This fact accompanied with the astonishing amount of first - place votes for Messi, in second consecutive year, bare the lack of competition between the dipole and the rest of the players.

7.3.4 Year 2013

As in the previous years, on the line chart that exists in the Appendix (B.48), is presented the change of BIC value while the number of clusters are increasing. Based on this figure, the line that depicts the trend of the BIC value through the different number of components, is track to have a sharp decrease from the point of 2 clusters to the one of 3 clusters and after this point starts to increase again. Thus, it is observed a 'knee point' when the line approaches the 3 clusters, which provides an indication that the optimal number of groups for fitting the ISR algorithm, is 3. For that reason, we implement the clustering of the voters on 3 clusters.

The results of the clustering show that the voters in 2013 are separated in 2 large groups, that consist almost the 93% of the dataset, and 1 small group. More specifically, the first cluster contains the 43.8% of the dataset (235 observations), the second cluster contains the 6.83% of the dataset (43 observations) and the third cluster contains the 49.35% of the dataset (263 observations). It is clear from the results of the first cluster (A.44), that the difference in the first - place votes between Ronaldo and Ribery is very small. On the other hand, the Frenchman has much fewer votes than Cristiano and Messi, in terms of second - place votes. Let's have a look of the votes on the third cluster.

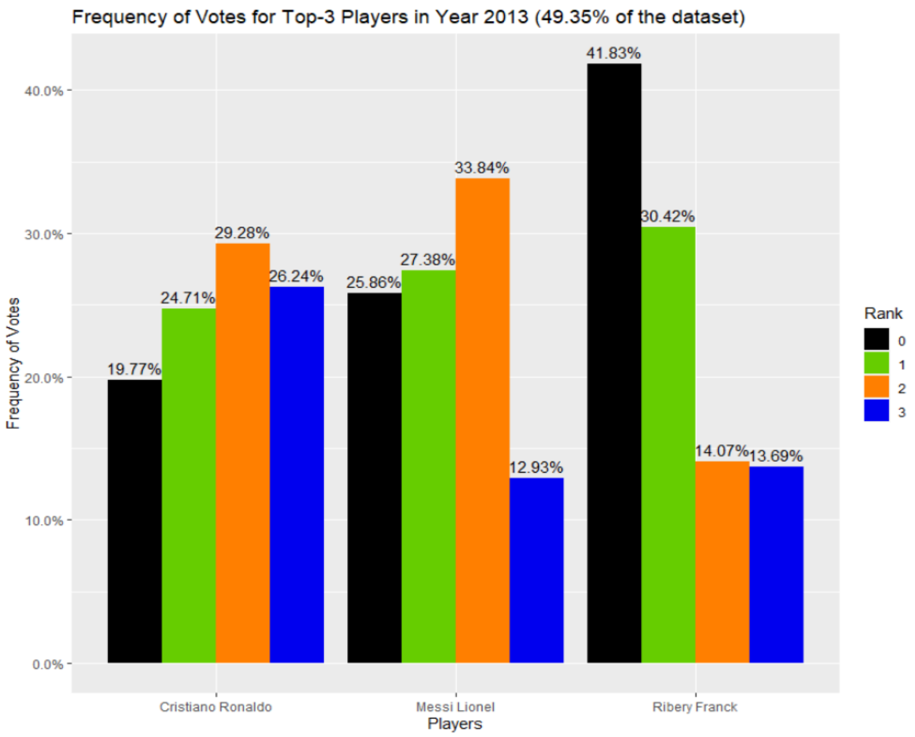


Figure 16 : Frequency of Votes for the Top - 3 in Cluster 3, Year 2013.

The same voting behavior pattern that we have spotted in the first cluster, is revealed also in the third cluster. From the above figure, someone could assume that the main reason that Ribery did not win the trophy was the large amount of voters that did not prefer him in compare with his opponents, despite the fact that he overcomes the other two candidates in first – place votes, in this cluster. At this point, it has to be reminded to the reader that Cristiano Ronaldo won the award in 2013, with very small lead of points to the second Lionel Messi and the third Franck Ribery. Thus, in such a year when the competition is that fragile every vote matters. This bar chart is an earmark of such a situation, because it can be observed that even Ribery had the most first - place votes, he came third because he had also the largest percentage of voters that did not rank him at all. At the same time, despite the fact that Lionel Messi has a small lead in first - place and second - place votes, he got left behind in third - place votes, in compare with the corresponding votes of Cristiano and that is the reason he remained second.

As in previous methods, Cluster 3 has been chosen on purpose for presentation because it does not depicts a clear ascendance of a player and ,thus, it is a great confirmation of the theory which claims that when the difference of points between some candidates is marginal, the candidate who would have been preferred from larger amount of voters, wins the award, no matter the genre (1st place, 2nd place, 3rd place) of the vote. Moreover, the fact that Ribery has been preferred many times first and very few times in the second or third - place, is a strong indication of how difficult is for a player to break the dipole apart and make his position more robust throughout the years. It is also a verification of the voting behavior that has been detected in the previous methods, for the same year.

7.3.5 Year 2014

Based on the BIC plot for 2014 (B.51), it is observed that the smallest BIC value is in 2 clusters. Furthermore, in that point, the line starts to increase, so it can be considered also as a ‘knee point’ of a small BIC value. We are able to notice, at this point, that the approach of ISR provides us with a model which number of clusters are relatively small, in compare to the Bayesian and the K – medoids model which have provided a large number of clusters, with an individual information in each one. This is an indication that ,probably, should have been examined more clusters in the fit of the model, in the ISR approach for 2014, which have not been examined due to the computational burden of many clusters in this method.

Besides that, it is clear that the voting behavior that has been detected in the previous methods, for Neuer’s voters, is ascertained also in this approach. If Cristiano Ronaldo would be excluded from the analysis of the two clusters, as his proportions of votes are similar in the two groups, we are able to spot the difference that works as the ‘separation cause’ between the two groups and that is the votes for Messi and Nuer. From the bar chart that represents the preferences for the Top – 3 in the first cluster (B.52), which consists the 75.73% of the dataset, someone can observe that the first – place votes that were assigned to the two players are equal, but Messi has been voted much more times in the other two places in compare with the German player. On the other hand, in Cluster 2 (B.53), it is observed a clear preference to the goalkeeper of Bayern Munich. The first – place votes for Neuer were the 21% of the total first – place votes, while on the same time, the corresponding votes for Messi were 6%. Furthermore, this can be evaluated from the votes of the most representative objects of the two clusters that have got the highest probability value. For the first cluster, Mahamud Raihan ranked first Cristiano, second Messi and third Neuer. Contrarily in the second cluster, Vladimir Petkovic has ranked first Cristiano, second Neuer and third Messi. Based on the bar plots, accompanied with the votes of the most representative objects for each cluster, the difference in the two clusters is evident. But, despite the fact that Neuer is more preferable in this cluster than Messi, the second – place and third – place points that have been assigned to the Argentinian are somewhat more. Thus, the voting pattern that has been detected in the previous chapters, for the same year but also in 2013, is confirmed after the analysis of the clustering findings.

7.3.6 Year 2015

From the plot of BIC for the Year 2015 (B.54), it is observed a ‘knee point’, where the value of the criterion starts to increase sharply. This event happens when the line approaches the 4 clusters. Thus, we fit the ISR model on 4 clusters, for 2015.

After running the algorithm and looking at the output, the proportions of observations that each cluster has are the following : Cluster 1 → 122 observations (23.9%), Cluster 2 → 102 observations (20.28%), Cluster 3 → 166 observations (34.13%), Cluster 4 → 108 observations (21.68%). We notice that the observations are close to uniformly distributed throughout the clusters. By looking at the table of Cluster 1 (A.46), it is observed that Messi is the total winner in this group, by having an enormous amount of first – place votes. Besides him, Cristiano has been voted many times in the second - place. Besides the dipole, someone can notice that the ranking rates of Suarez, Muller and Neymar are high, in compare with the rest of the players. In Cluster 2 (A.47), Messi retains his outstanding power in points. Also, Cristiano’s second – place rates are preserved in a high level. The difference with Cluster 1, is found on the much larger amount votes that Neymar has received from the audience that consists this group. The rankings in the fourth cluster (A.48) are very similar to those of the second cluster. In Cluster 3 (A.49), while the rates of Top – 3 (Messi, Ronaldo, Neymar) are proportional to those of Cluster 2, it is observed an increase in votes of players that have not been preferred almost at all in the previous clusters. Such players are Neuer, Lewandowski and Hazard.

After analyzing the results in each cluster, we construct stacked bar plots to detect indications of possible affect, to the preference of a group of voters, from the origins of these voters. By looking at the figures (B.55, B.56, B.57, B.58), we can observe that there has not been identified a specific pattern in any cluster. This fact supports the theory that has been presented in a previous section, which claims that when a player or some players have a great performance in a season, the votes that he is going to received do not depend on the country that the player comes from. It has to be reminded that this theory is assumed only for the players that exist in Top – 3, for each year.

Chapter 8

Conclusions

The purpose of this Thesis was to present the fundamental notions in the context of ranking data by providing ways to visualizing, modeling and clustering partial ranking data. Also, by applying three different clustering methods in the manipulated FIFA Ballon D’Or datasets in order to separate the voters in different groups, according to their preference, and trying to detect possible voting behavioral patterns that could affect the final votes.

In the first Chapter it was conducted a brief description of the topics that would be represented in the next chapters. These were the concepts regarding the presentation of FIFA Ballon D’Or voting method, the notion of ranking data and clustering. Also, were presented the main goals of the Thesis.

In Chapter 2, we described the basic classes of ranking data models. After presenting the first basic separation between probability models and probit models, we further expanded the frame to the basic categories of such models. Especially, the focus was put on the categories of probability models, which are mainly used in the clustering applications of the Thesis. Also, we highlighted some of the pros and cons, that these models have, regarding their compatibility with partially ranked data. Chapter 2 ends with a brief description of the decision trees for ranking data.

In the third Chapter, we presented the FIFA raw datasets in the format that were retrieved from the source. Moreover, we showed the cleaning and the manipulation process that the datasets went through in order to be transformed in a ranking format. Furthermore, some fundamental descriptive statistics for ranking data were presented and implemented, in order to get a better understanding of the data.

In Chapter 4, we dealt with the concept of graphical representation for ranking data. This topic is very interesting and challenging, especially when partial ranking data exist because most of the techniques are different, in order for the missing positions to be ‘faced’. The main techniques that were presented for visualizing ranking data are the permutation polytope, multidimensional methods such as the multidimensional unfolding technique, the multidimensional preference analysis, the classic – metric and the non – metric multidimensional scaling method, by mentioning also how suitable each method is for partially ranked datasets. We focused on the non – metric multidimensional scaling approach, since it was applied to the FIFA partial ranking datasets.

In Chapter 5, we applied cluster analysis to the FIFA datasets based on the Bayesian mixture of Plackett – Luce models. In the first part of the chapter, some fundamental concepts of the approach, such as the notion of Plackett – Luce model, the maximum a posteriori estimation of the model's parameters through EM algorithm and Gibbs sampling, and the Bayesian criteria that were used for the selection of the 'appropriate' model, were covered. In the second part, we described the main features of the method's application in R and in the third part, we applied the method in the FIFA datasets for each year of the period 2010 – 15. The structure of Chapter 5, was also followed in the sixth and seventh chapter. In Chapter 6, we worked with the K – medoids algorithm in order to cluster the voters of the FIFA datasets. In the theoretical framework of the method, we described the K – medoids or Partitioning Around Medoids (PAM) procedure. Also, the Kendall's distance and its properties, were also presented, as it is the distance metric that is used in the algorithm. Besides these, we presented the Average Silhouette value approach and the Elbow method, in order to determine the 'best' model among the candidates and evaluate it. After a description of the basic features of the method's application in R, we applied it for each specific year.

Finally, in Chapter 7, the Insertion Sorting Rank (ISR) algorithm was taken into account for the clustering of the voters. The univariate and multivariate ISR models, and the models estimation through the Maximum Likelihood Estimation approach and the SEM – Gibbs algorithm, were presented. Besides these, the Bayesian Information Criterion (BIC) was also described as a criterion for selecting the 'appropriate' number of clusters for the final model. Afterwards, as in the previous methods, we presented the main features of the method's application in R and we applied the ISR clustering approach on the FIFA partial ranking datasets.

From the clustering applications that were implemented in the above chapters, we managed to group the voters according to their preference, compare the results and abstract some very interesting conclusions. We detected a fascinating voting behavioral pattern, which was spotted in 2013 and 2014. In particular, it was identified that when two players are very close, in terms of amount of votes, it is preferable for a player to be ranked in a more normal distributed way in the three places (1st ,2nd,3rd), rather than ranked many times in the first - place and simultaneously, very few times, in the other two places. This conclusion was distinguished in 2013 and 2014, in all the three methods and is about the 'battle' for the second - place that had been taken place between Messi – Ribery, in 2013, and Messi – Neuer, in 2014.

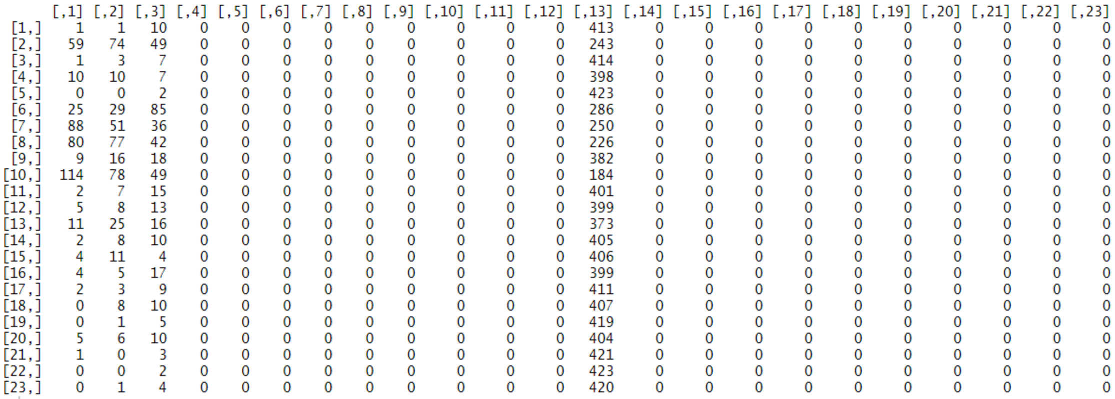
Another fascinating pattern that was detected, is about the relationship between the origins of a voter and his final vote. More specifically, we explored the voters in many clusters, in terms of the continent that they come from and their job. It was observed that in cases where the winner of the award is obvious and is represented by a large cluster, the remaining smaller clusters are shaped driven by the origins relationship between the player that is voted and the voter. This indication was not proved statistically, but it was confirmed as a strong indication of pattern existence in such cases, through the analysis that was implemented in the corresponding years in all the methods.

Regarding the above conclusions, it would be very interesting if someone could prove statistically these voting behavioral patterns and the origins relationship that was indicated that affects the vote, in specific cases. Moreover, we encourage the reader to improve the above – mentioned approaches by trying to reduce the computational time in the ISR method or by using different distance, that is going to be compatible with the partial rankings, than the Kendall's distance in the K – medoids method or by trying to apply visualization methods that are about complete rankings (e.g permutation polytope) to incomplete rankings. Finally, we hope that this Thesis is going to be a very useful tool for someone who wants to learn about the fundamental notions of ranking data and work with clustering methods in the context of partial ranking data.

APPENDICES

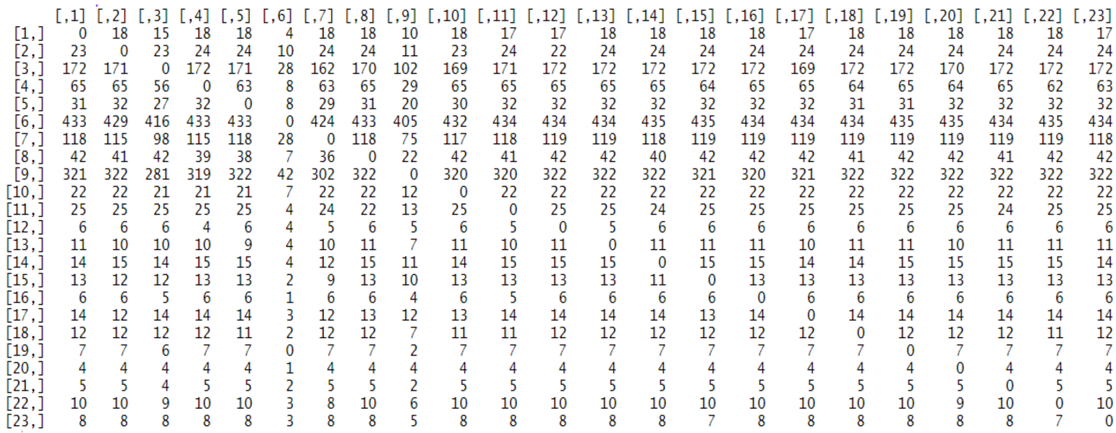
APPENDIX A : TABLES

Chapter 4



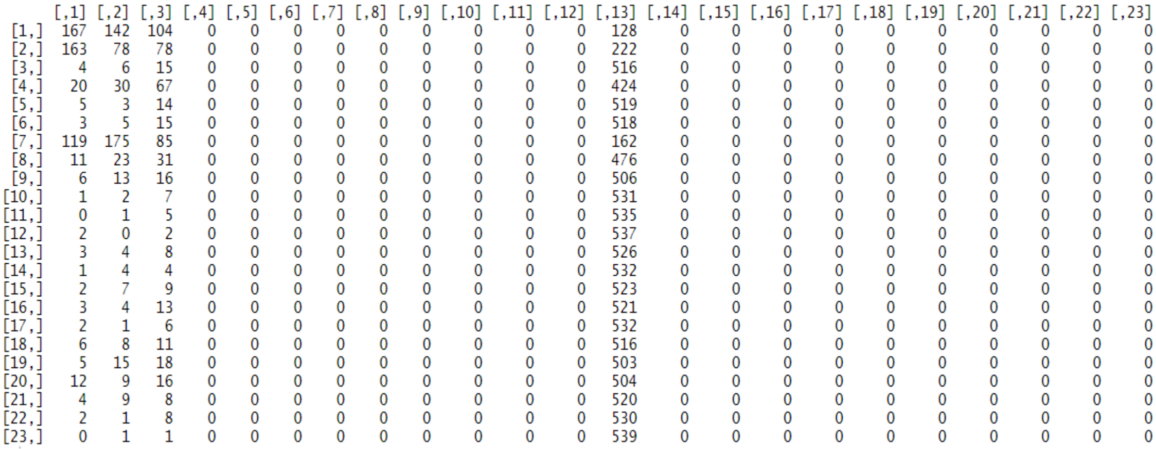
```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23]
 [1,]    1    1   10    0    0    0    0    0    0     0     0     0   413     0     0     0     0     0     0     0     0     0     0
 [2,]   59   74   49    0    0    0    0    0    0     0     0     0   243     0     0     0     0     0     0     0     0     0     0
 [3,]    1    3    7    0    0    0    0    0    0     0     0     0   414     0     0     0     0     0     0     0     0     0     0
 [4,]   10   10    7    0    0    0    0    0    0     0     0     0   398     0     0     0     0     0     0     0     0     0     0
 [5,]    0    0    2    0    0    0    0    0    0     0     0     0   423     0     0     0     0     0     0     0     0     0     0
 [6,]   25   29   85    0    0    0    0    0    0     0     0     0   286     0     0     0     0     0     0     0     0     0     0
 [7,]   88   51   36    0    0    0    0    0    0     0     0     0   250     0     0     0     0     0     0     0     0     0     0
 [8,]   80   77   42    0    0    0    0    0    0     0     0     0   226     0     0     0     0     0     0     0     0     0     0
 [9,]    9   16   18    0    0    0    0    0    0     0     0     0   382     0     0     0     0     0     0     0     0     0     0
[10,]  114   78   49    0    0    0    0    0    0     0     0     0   184     0     0     0     0     0     0     0     0     0     0
[11,]    2    7   15    0    0    0    0    0    0     0     0     0   401     0     0     0     0     0     0     0     0     0     0
[12,]    5    8   13    0    0    0    0    0    0     0     0     0   399     0     0     0     0     0     0     0     0     0     0
[13,]   11   25   16    0    0    0    0    0    0     0     0     0   373     0     0     0     0     0     0     0     0     0     0
[14,]    2    8   10    0    0    0    0    0    0     0     0     0   405     0     0     0     0     0     0     0     0     0     0
[15,]    4   11    4    0    0    0    0    0    0     0     0     0   406     0     0     0     0     0     0     0     0     0     0
[16,]    4    5   17    0    0    0    0    0    0     0     0     0   399     0     0     0     0     0     0     0     0     0     0
[17,]    2    3    9    0    0    0    0    0    0     0     0     0   411     0     0     0     0     0     0     0     0     0     0
[18,]    0    8   10    0    0    0    0    0    0     0     0     0   407     0     0     0     0     0     0     0     0     0     0
[19,]    0    1    5    0    0    0    0    0    0     0     0     0   419     0     0     0     0     0     0     0     0     0     0
[20,]    5    6   10    0    0    0    0    0    0     0     0     0   404     0     0     0     0     0     0     0     0     0     0
[21,]    1    0    3    0    0    0    0    0    0     0     0     0   421     0     0     0     0     0     0     0     0     0     0
[22,]    0    0    2    0    0    0    0    0    0     0     0     0   423     0     0     0     0     0     0     0     0     0     0
[23,]    0    1    4    0    0    0    0    0    0     0     0     0   420     0     0     0     0     0     0     0     0     0     0
```

Table A.1 : Marginal matrix of the players for Year 2010.



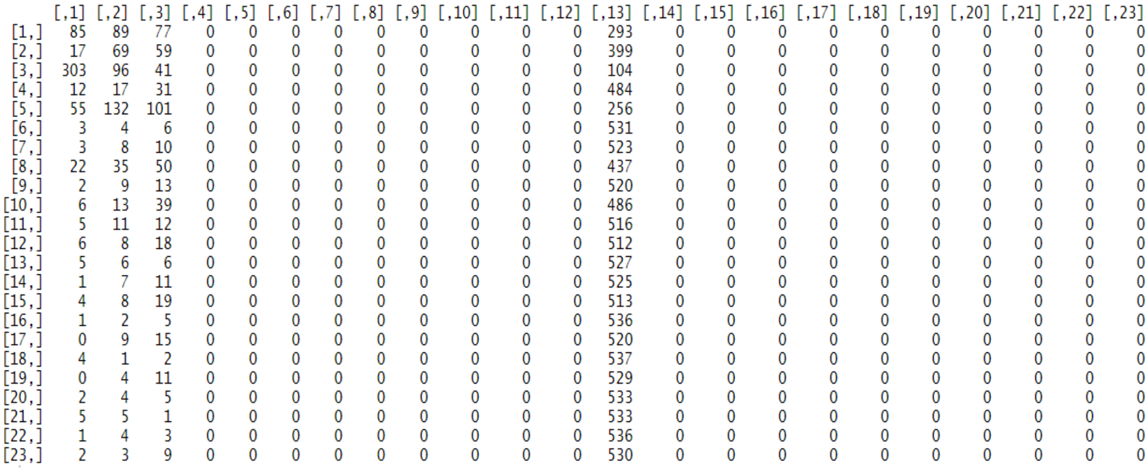
```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23]
 [1,]    0   18   15   18   18    4   18   18   10    18    17    17    18    18    18    18    17    18    18    18    18    18    17
 [2,]   23    0   23   24   24   10   24   24   11    23    24    22    24    24    24    24    24    24    24    24    24    24    24
 [3,]  172  171    0  172  171   28  162  170  102   169   171   172   172   172   172   172   169   172   172   170   172   172   172
 [4,]   65   65   56    0   63    8   63   65   29    65    65    65    65    65    64    65    65    64    65    64    65    62    63
 [5,]   31   32   27   32    0    8   29   31   20    30    32    32    32    32    32    32    32    31    31    32    32    32    32
 [6,]  433  429  416  433  433    0  424  433  405   432   434   434   434   435   435   434   434   434   435   435   434   435   434
 [7,]  118  115   98  115  118   28    0  118   75   117   118   119   119   118   119   119   119   119   119   119   119   119   118
 [8,]   42   41   42   39   38    7   36    0   22    42    41    42    42    40    42    42    42    41    42    42    41    42    42
 [9,]  321  322  281  319  322   42  302  322    0   320   320   322   322   322   321   320   321   322   322   322   322   322   322
[10,]   22   22   21   21   21    7   22   22   12     0    22    22    22    22    22    22    22    22    22    22    22    22    22
[11,]   25   25   25   25   25    4   24   22   13    25     0    25    25    24    25    25    25    25    25    25    24    25    25
[12,]    6    6    6    4    6    4    5    6    5     6     5     0     5     6     6     6     6     6     6     6     6     6     6
[13,]   11   10   10   10    9    4   10   11    7    11    10    11     0    11    11    11    10    11    11    10    11    11    11
[14,]   14   15   14   15   15    4   12   15   11    14    15    15    15     0    15    15    14    14    15    15    15    15    14
[15,]   13   12   12   13   13    2    9   13   10    13    13    13    13    11     0    13    13    13    13    13    13    13    13
[16,]    6    6    5    6    6    1    6    6    4     6     5     6     6     6     6     0     6     6     6     6     6     6     6
[17,]   14   12   14   14   14    3   12   13   12    13    14    14    14    14    13    14     0    14    14    14    14    14    14
[18,]   12   12   12   12   11    2   12   12    7    11    11    12    12    12    12    12    12     0    12    12    12    11    12
[19,]    7    7    6    7    7    0    7    7    2     7     7     7     7     7     7     7     7     7     0     7     7     7     7
[20,]    4    4    4    4    4    1    4    4    4     4     4     4     4     4     4     4     4     4     4     0     4     4     4
[21,]    5    5    4    5    5    2    5    5    2     5     5     5     5     5     5     5     5     5     5     5     0     5     5
[22,]   10   10    9   10   10    3    8   10    6    10    10    10    10    10    10    10    10    10    10     9    10     0    10
[23,]    8    8    8    8    8    3    8    8    5     8     8     8     8     8     7     8     8     8     8     8     8     7     0
```

Table A.2 : Matrix of the players pairwise frequencies for Year 2010.



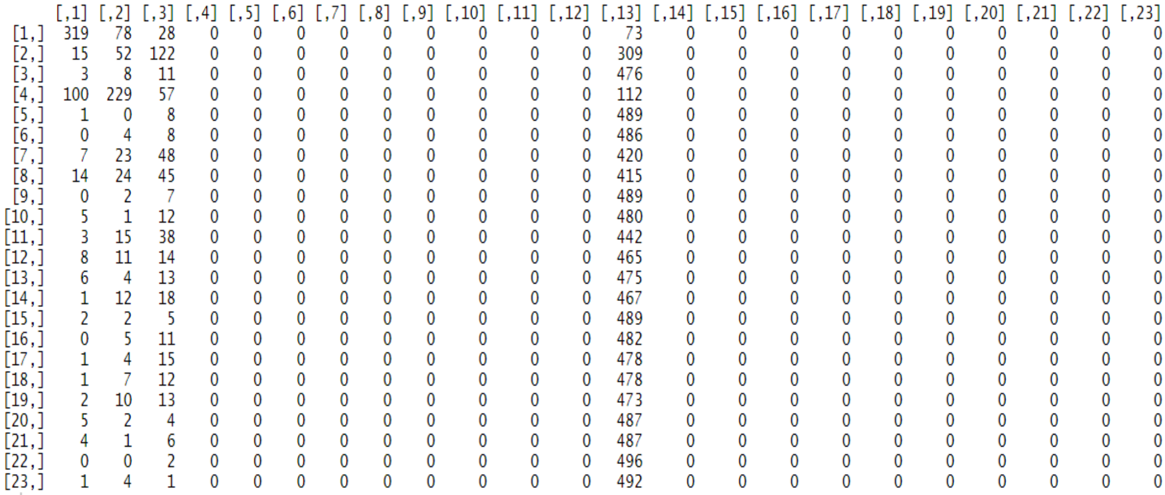
```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23]
 [1,]  167  142  104    0    0    0    0    0    0     0     0     0   128     0     0     0     0     0     0     0     0     0     0
 [2,]  163   78   78    0    0    0    0    0    0     0     0     0   222     0     0     0     0     0     0     0     0     0     0
 [3,]    4    6   15    0    0    0    0    0    0     0     0     0   516     0     0     0     0     0     0     0     0     0     0
 [4,]   20   30   67    0    0    0    0    0    0     0     0     0   424     0     0     0     0     0     0     0     0     0     0
 [5,]    5    3   14    0    0    0    0    0    0     0     0     0   519     0     0     0     0     0     0     0     0     0     0
 [6,]    3    5   15    0    0    0    0    0    0     0     0     0   518     0     0     0     0     0     0     0     0     0     0
 [7,]  119  175   85    0    0    0    0    0    0     0     0     0   162     0     0     0     0     0     0     0     0     0     0
 [8,]   11   23   31    0    0    0    0    0    0     0     0     0   476     0     0     0     0     0     0     0     0     0     0
 [9,]    6   13   16    0    0    0    0    0    0     0     0     0   506     0     0     0     0     0     0     0     0     0     0
[10,]    1    2    7    0    0    0    0    0    0     0     0     0   531     0     0     0     0     0     0     0     0     0     0
[11,]    0    1    5    0    0    0    0    0    0     0     0     0   535     0     0     0     0     0     0     0     0     0     0
[12,]    2    0    2    0    0    0    0    0    0     0     0     0   537     0     0     0     0     0     0     0     0     0     0
[13,]    3    4    8    0    0    0    0    0    0     0     0     0   526     0     0     0     0     0     0     0     0     0     0
[14,]    1    4    4    0    0    0    0    0    0     0     0     0   532     0     0     0     0     0     0     0     0     0     0
[15,]    2    7    9    0    0    0    0    0    0     0     0     0   523     0     0     0     0     0     0     0     0     0     0
[16,]    3    4   13    0    0    0    0    0    0     0     0     0   521     0     0     0     0     0     0     0     0     0     0
[17,]    2    1    6    0    0    0    0    0    0     0     0     0   532     0     0     0     0     0     0     0     0     0     0
[18,]    6    8   11    0    0    0    0    0    0     0     0     0   516     0     0     0     0     0     0     0     0     0     0
[19,]    5   15   18    0    0    0    0    0    0     0     0     0   503     0     0     0     0     0     0     0     0     0     0
[20,]   12    9   16    0    0    0    0    0    0     0     0     0   504     0     0     0     0     0     0     0     0     0     0
[21,]    4    9    8    0    0    0    0    0    0     0     0     0   520     0     0     0     0     0     0     0     0     0     0
[22,]    2    1    8    0    0    0    0    0    0     0     0     0   530     0     0     0     0     0     0     0     0     0     0
[23,]    0    1    1    0    0    0    0    0    0     0     0     0   539     0     0     0     0     0     0     0     0     0     0
```

Table A.3 : Marginal matrix of the players for Year 2013.



```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23]
 [1,]   85   89   77    0    0    0    0    0    0     0     0     0   293     0     0     0     0     0     0     0     0     0     0
 [2,]   17   69   59    0    0    0    0    0    0     0     0     0   399     0     0     0     0     0     0     0     0     0     0
 [3,]  303   96   41    0    0    0    0    0    0     0     0     0   104     0     0     0     0     0     0     0     0     0     0
 [4,]   12   17   31    0    0    0    0    0    0     0     0     0   484     0     0     0     0     0     0     0     0     0     0
 [5,]   55  132  101    0    0    0    0    0    0     0     0     0   256     0     0     0     0     0     0     0     0     0     0
 [6,]    3    4    6    0    0    0    0    0    0     0     0     0   531     0     0     0     0     0     0     0     0     0     0
 [7,]    3    8   10    0    0    0    0    0    0     0     0     0   523     0     0     0     0     0     0     0     0     0     0
 [8,]   22   35   50    0    0    0    0    0    0     0     0     0   437     0     0     0     0     0     0     0     0     0     0
 [9,]    2    9   13    0    0    0    0    0    0     0     0     0   520     0     0     0     0     0     0     0     0     0     0
[10,]    6   13   39    0    0    0    0    0    0     0     0     0   486     0     0     0     0     0     0     0     0     0     0
[11,]    5   11   12    0    0    0    0    0    0     0     0     0   516     0     0     0     0     0     0     0     0     0     0
[12,]    6    8   18    0    0    0    0    0    0     0     0     0   512     0     0     0     0     0     0     0     0     0     0
[13,]    5    6    6    0    0    0    0    0    0     0     0     0   527     0     0     0     0     0     0     0     0     0     0
[14,]    1    7   11    0    0    0    0    0    0     0     0     0   525     0     0     0     0     0     0     0     0     0     0
[15,]    4    8   19    0    0    0    0    0    0     0     0     0   513     0     0     0     0     0     0     0     0     0     0
[16,]    1    2    5    0    0    0    0    0    0     0     0     0   536     0     0     0     0     0     0     0     0     0     0
[17,]    0    9   15    0    0    0    0    0    0     0     0     0   520     0     0     0     0     0     0     0     0     0     0
[18,]    4    1    2    0    0    0    0    0    0     0     0     0   537     0     0     0     0     0     0     0     0     0     0
[19,]    0    4   11    0    0    0    0    0    0     0     0     0   529     0     0     0     0     0     0     0     0     0     0
[20,]    2    4    5    0    0    0    0    0    0     0     0     0   533     0     0     0     0     0     0     0     0     0     0
[21,]    5    5    1    0    0    0    0    0    0     0     0     0   533     0     0     0     0     0     0     0     0     0     0
[22,]    1    4    3    0    0    0    0    0    0     0     0     0   536     0     0     0     0     0     0     0     0     0     0
[23,]    2    3    9    0    0    0    0    0    0     0     0     0   530     0     0     0     0     0     0     0     0     0     0
```

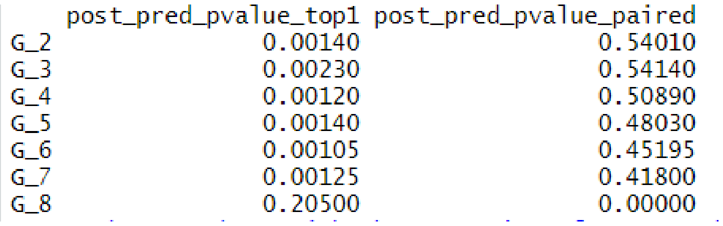
Table A.4 : Marginal matrix of the players for Year 2014.



```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23]
 [1,]  319   78   28    0    0    0    0    0    0     0     0     0    73     0     0     0     0     0     0     0     0     0     0
 [2,]   15   52  122    0    0    0    0    0    0     0     0     0   309     0     0     0     0     0     0     0     0     0     0
 [3,]    3    8   11    0    0    0    0    0    0     0     0     0   476     0     0     0     0     0     0     0     0     0     0
 [4,]  100  229   57    0    0    0    0    0    0     0     0     0   112     0     0     0     0     0     0     0     0     0     0
 [5,]    1    0    8    0    0    0    0    0    0     0     0     0   489     0     0     0     0     0     0     0     0     0     0
 [6,]    0    4    8    0    0    0    0    0    0     0     0     0   486     0     0     0     0     0     0     0     0     0     0
 [7,]    7   23   48    0    0    0    0    0    0     0     0     0   420     0     0     0     0     0     0     0     0     0     0
 [8,]   14   24   45    0    0    0    0    0    0     0     0     0   415     0     0     0     0     0     0     0     0     0     0
 [9,]    0    2    7    0    0    0    0    0    0     0     0     0   489     0     0     0     0     0     0     0     0     0     0
[10,]    5    1   12    0    0    0    0    0    0     0     0     0   480     0     0     0     0     0     0     0     0     0     0
[11,]    3   15   38    0    0    0    0    0    0     0     0     0   442     0     0     0     0     0     0     0     0     0     0
[12,]    8   11   14    0    0    0    0    0    0     0     0     0   465     0     0     0     0     0     0     0     0     0     0
[13,]    6    4   13    0    0    0    0    0    0     0     0     0   475     0     0     0     0     0     0     0     0     0     0
[14,]    1   12   18    0    0    0    0    0    0     0     0     0   467     0     0     0     0     0     0     0     0     0     0
[15,]    2    2    5    0    0    0    0    0    0     0     0     0   489     0     0     0     0     0     0     0     0     0     0
[16,]    0    5   11    0    0    0    0    0    0     0     0     0   482     0     0     0     0     0     0     0     0     0     0
[17,]    1    4   15    0    0    0    0    0    0     0     0     0   478     0     0     0     0     0     0     0     0     0     0
[18,]    1    7   12    0    0    0    0    0    0     0     0     0   478     0     0     0     0     0     0     0     0     0     0
[19,]    2   10   13    0    0    0    0    0    0     0     0     0   473     0     0     0     0     0     0     0     0     0     0
[20,]    5    2    4    0    0    0    0    0    0     0     0     0   487     0     0     0     0     0     0     0     0     0     0
[21,]    4    1    6    0    0    0    0    0    0     0     0     0   487     0     0     0     0     0     0     0     0     0     0
[22,]    0    0    2    0    0    0    0    0    0     0     0     0   496     0     0     0     0     0     0     0     0     0     0
[23,]    1    4    1    0    0    0    0    0    0     0     0     0   492     0     0     0     0     0     0     0     0     0     0
```

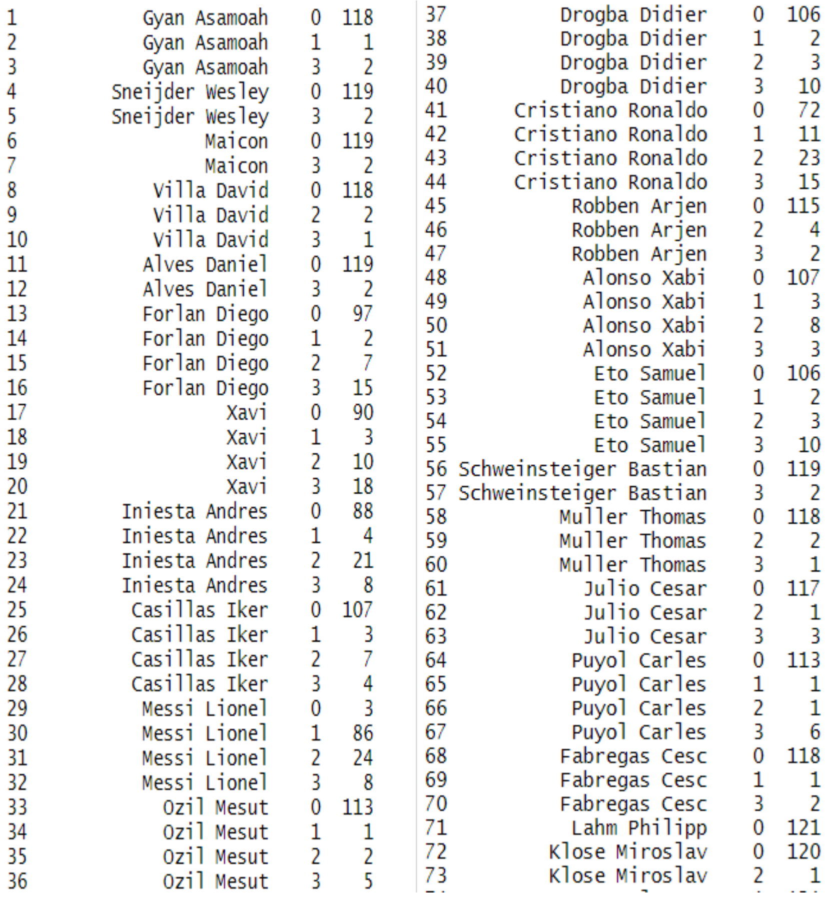
Table A.5 : Marginal matrix of the players for Year 2015.

Chapter 5



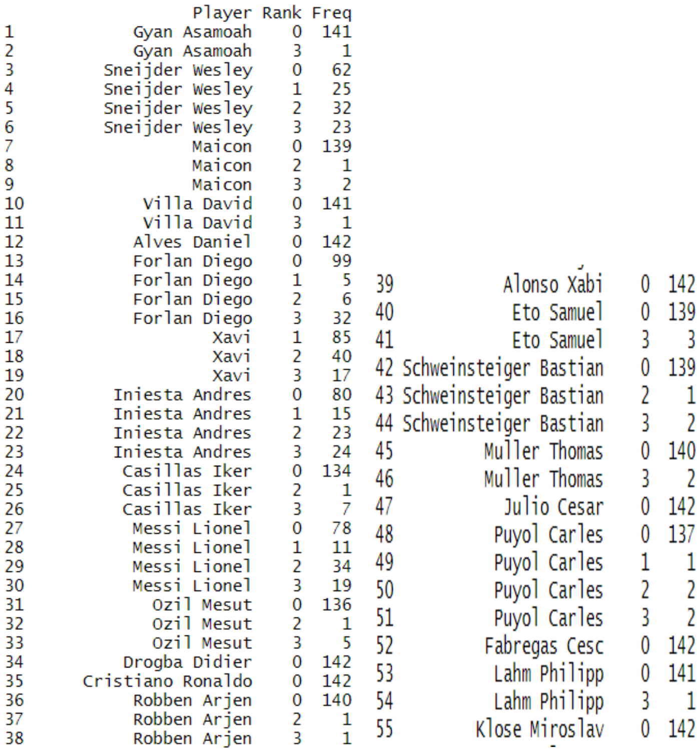
```
    post_pred_pvalue_top1 post_pred_pvalue_paired
G_2               0.00140                 0.54010
G_3               0.00230                 0.54140
G_4               0.00120                 0.50890
G_5               0.00140                 0.48030
G_6               0.00105                 0.45195
G_7               0.00125                 0.41800
G_8               0.20500                 0.00000
```

Table A.6 : P-values for the Bayesian PL models assessment, Year 2010



```
1          Gyan Asamoah  0  118    37          Drogba Didier  0  106
2          Gyan Asamoah  1    1    38          Drogba Didier  1    2
3          Gyan Asamoah  3    2    39          Drogba Didier  2    3
4       Sneijder Wesley  0  119    40          Drogba Didier  3   10
5       Sneijder Wesley  3    2    41     Cristiano Ronaldo  0   72
6                Maicon  0  119    42     Cristiano Ronaldo  1   11
7                Maicon  3    2    43     Cristiano Ronaldo  2   23
8           Villa David  0  118    44     Cristiano Ronaldo  3   15
9           Villa David  2    2    45          Robben Arjen  0  115
10          Villa David  3    1    46          Robben Arjen  2    4
11         Alves Daniel  0  119    47          Robben Arjen  3    2
12         Alves Daniel  3    2    48           Alonso Xabi  0  107
13         Forlan Diego  0   97    49           Alonso Xabi  1    3
14         Forlan Diego  1    2    50           Alonso Xabi  2    8
15         Forlan Diego  2    7    51           Alonso Xabi  3    3
16         Forlan Diego  3   15    52            Eto Samuel  0  106
17                 Xavi  0   90    53            Eto Samuel  1    2
18                 Xavi  1    3    54            Eto Samuel  2    3
19                 Xavi  2   10    55            Eto Samuel  3   10
20                 Xavi  3   18    56 Schweinsteiger Bastian  0  119
21       Iniesta Andres  0   88    57 Schweinsteiger Bastian  3    2
22       Iniesta Andres  1    4    58         Muller Thomas  0  118
23       Iniesta Andres  2   21    59         Muller Thomas  2    2
24       Iniesta Andres  3    8    60         Muller Thomas  3    1
25        Casillas Iker  0  107    61           Julio Cesar  0  117
26        Casillas Iker  1    3    62           Julio Cesar  2    1
27        Casillas Iker  2    7    63           Julio Cesar  3    3
28        Casillas Iker  3    4    64          Puyol Carles  0  113
29         Messi Lionel  0    3    65          Puyol Carles  1    1
30         Messi Lionel  1   86    66          Puyol Carles  2    1
31         Messi Lionel  2   24    67          Puyol Carles  3    6
32         Messi Lionel  3    8    68         Fabregas Cesc  0  118
33           Ozil Mesut  0  113    69         Fabregas Cesc  1    1
34           Ozil Mesut  1    1    70         Fabregas Cesc  3    2
35           Ozil Mesut  2    2    71          Lahm Philipp  0  121
36           Ozil Mesut  3    5    72        Klose Miroslav  0  120
                                   73        Klose Miroslav  2    1
```

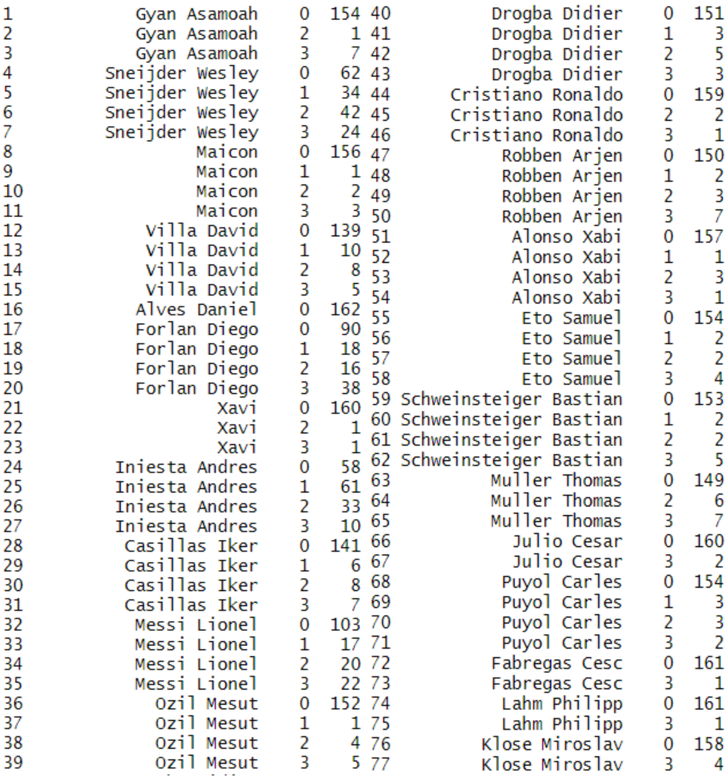
Table A.7 : The amount of votes that each player received in Cluster 1, Year 2010.



```
                  Player Rank Freq
1           Gyan Asamoah    0  141
2           Gyan Asamoah    3    1
3        Sneijder Wesley    0   62
4        Sneijder Wesley    1   25
5        Sneijder Wesley    2   32
6        Sneijder Wesley    3   23
7                 Maicon    0  139
8                 Maicon    2    1
9                 Maicon    3    2
10           Villa David    0  141
11           Villa David    3    1
12          Alves Daniel    0  142
13          Forlan Diego    0   99
14          Forlan Diego    1    5
15          Forlan Diego    2    6
16          Forlan Diego    3   32
17                  Xavi    1   85
18                  Xavi    2   40
19                  Xavi    3   17
20        Iniesta Andres    0   80
21        Iniesta Andres    1   15
22        Iniesta Andres    2   23
23        Iniesta Andres    3   24
24         Casillas Iker    0  134
25         Casillas Iker    2    1
26         Casillas Iker    3    7
27          Messi Lionel    0   78
28          Messi Lionel    1   11
29          Messi Lionel    2   34
30          Messi Lionel    3   19
31            Ozil Mesut    0  136
32            Ozil Mesut    2    1
33            Ozil Mesut    3    5
34         Drogba Didier    0  142
35     Cristiano Ronaldo    0  142
36          Robben Arjen    0  140
37          Robben Arjen    2    1
38          Robben Arjen    3    1
```

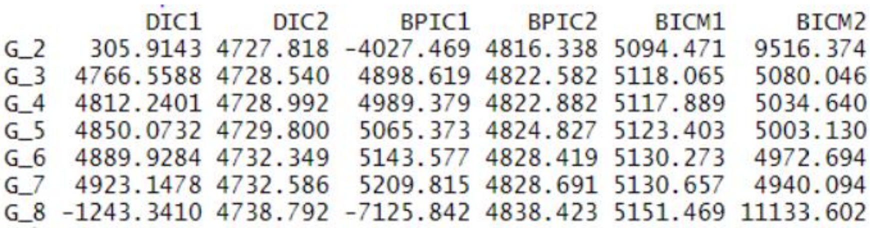
```
39            Alonso Xabi    0  142
40             Eto Samuel    0  139
41             Eto Samuel    3    3
42 Schweinsteiger Bastian    0  139
43 Schweinsteiger Bastian    2    1
44 Schweinsteiger Bastian    3    2
45          Muller Thomas    0  140
46          Muller Thomas    3    2
47            Julio Cesar    0  142
48           Puyol Carles    0  137
49           Puyol Carles    1    1
50           Puyol Carles    2    2
51           Puyol Carles    3    2
52          Fabregas Cesc    0  142
53           Lahm Philipp    0  141
54           Lahm Philipp    3    1
55          Klose Miroslav    0  142
```

Table A.8 : The amount of votes that each player received in Cluster 2, Year 2010.



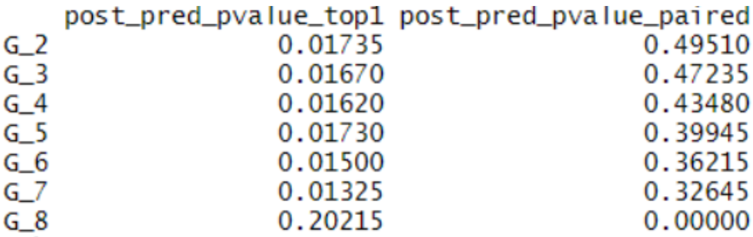
```
1             Gyan Asamoah    0  154 40           Drogba Didier    0  151
2             Gyan Asamoah    2    1 41           Drogba Didier    1    3
3             Gyan Asamoah    3    7 42           Drogba Didier    2    5
4          Sneijder Wesley    0   62 43           Drogba Didier    3    3
5          Sneijder Wesley    1   34 44       Cristiano Ronaldo    0  159
6          Sneijder Wesley    2   42 45       Cristiano Ronaldo    2    2
7          Sneijder Wesley    3   24 46       Cristiano Ronaldo    3    1
8                   Maicon    0  156 47            Robben Arjen    0  150
9                   Maicon    1    1 48            Robben Arjen    1    2
10                  Maicon    2    2 49            Robben Arjen    2    3
11                  Maicon    3    3 50            Robben Arjen    3    7
12             Villa David    0  139 51             Alonso Xabi    0  157
13             Villa David    1   10 52             Alonso Xabi    1    1
14             Villa David    2    8 53             Alonso Xabi    2    3
15             Villa David    3    5 54             Alonso Xabi    3    1
16            Alves Daniel    0  162 55              Eto Samuel    0  154
17            Forlan Diego    0   90 56              Eto Samuel    1    2
18            Forlan Diego    1   18 57              Eto Samuel    2    2
19            Forlan Diego    2   16 58              Eto Samuel    3    4
20            Forlan Diego    3   38 59  Schweinsteiger Bastian    0  153
21                    Xavi    0  160 60  Schweinsteiger Bastian    1    2
22                    Xavi    2    1 61  Schweinsteiger Bastian    2    2
23                    Xavi    3    1 62  Schweinsteiger Bastian    3    5
24          Iniesta Andres    0   58 63           Muller Thomas    0  149
25          Iniesta Andres    1   61 64           Muller Thomas    2    6
26          Iniesta Andres    2   33 65           Muller Thomas    3    7
27          Iniesta Andres    3   10 66             Julio Cesar    0  160
28           Casillas Iker    0  141 67             Julio Cesar    3    2
29           Casillas Iker    1    6 68            Puyol Carles    0  154
30           Casillas Iker    2    8 69            Puyol Carles    1    3
31           Casillas Iker    3    7 70            Puyol Carles    2    3
32            Messi Lionel    0  103 71            Puyol Carles    3    2
33            Messi Lionel    1   17 72           Fabregas Cesc    0  161
34            Messi Lionel    2   20 73           Fabregas Cesc    3    1
35            Messi Lionel    3   22 74            Lahm Philipp    0  161
36              Ozil Mesut    0  152 75            Lahm Philipp    3    1
37              Ozil Mesut    1    1 76          Klose Miroslav    0  158
38              Ozil Mesut    2    4 77          Klose Miroslav    3    4
39              Ozil Mesut    3    5
```

Table A.9 : The amount of votes that each player received in Cluster 3, Year 2010.



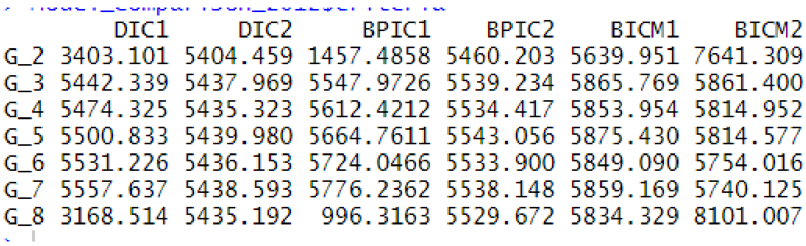
```
             DIC1      DIC2      BPIC1     BPIC2     BICM1      BICM2
G_2   305.9143 4727.818 -4027.469 4816.338 5094.471   9516.374
G_3  4766.5588 4728.540  4898.619 4822.582 5118.065   5080.046
G_4  4812.2401 4728.992  4989.379 4822.882 5117.889   5034.640
G_5  4850.0732 4729.800  5065.373 4824.827 5123.403   5003.130
G_6  4889.9284 4732.349  5143.577 4828.419 5130.273   4972.694
G_7  4923.1478 4732.586  5209.815 4828.691 5130.657   4940.094
G_8 -1243.3410 4738.792 -7125.842 4838.423 5151.469  11133.602
```

Table A.10 : Model comparison criteria values, Year 2011.



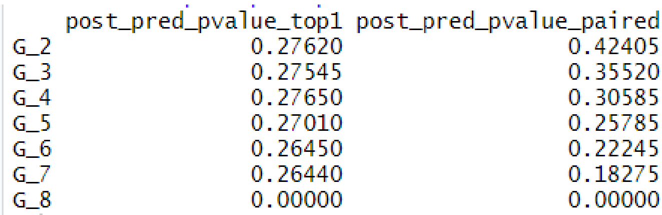
```
    post_pred_pvalue_top1 post_pred_pvalue_paired
G_2               0.01735                 0.49510
G_3               0.01670                 0.47235
G_4               0.01620                 0.43480
G_5               0.01730                 0.39945
G_6               0.01500                 0.36215
G_7               0.01325                 0.32645
G_8               0.20215                 0.00000
```

Table A.11 : P-values for the Bayesian PL models assessment, Year 2011.



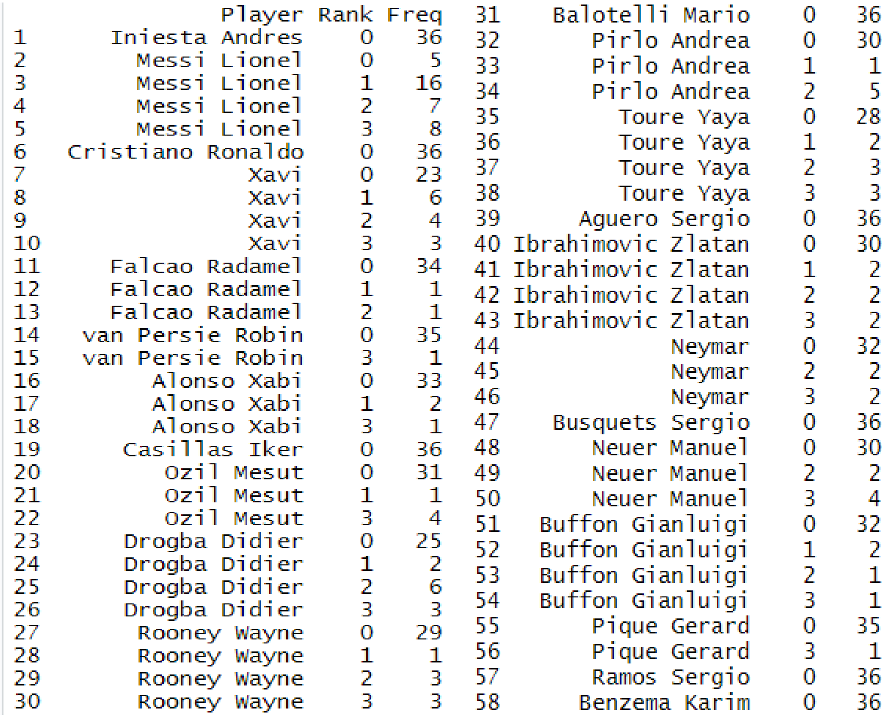
```
        DIC1     DIC2     BPIC1    BPIC2    BICM1    BICM2
G_2 3403.101 5404.459 1457.4858 5460.203 5639.951 7641.309
G_3 5442.339 5437.969 5547.9726 5539.234 5865.769 5861.400
G_4 5474.325 5435.323 5612.4212 5534.417 5853.954 5814.952
G_5 5500.833 5439.980 5664.7611 5543.056 5875.430 5814.577
G_6 5531.226 5436.153 5724.0466 5533.900 5849.090 5754.016
G_7 5557.637 5438.593 5776.2362 5538.148 5859.169 5740.125
G_8 3168.514 5435.192  996.3163 5529.672 5834.329 8101.007
```

Table A.12 : Model comparison criteria values, Year 2012.



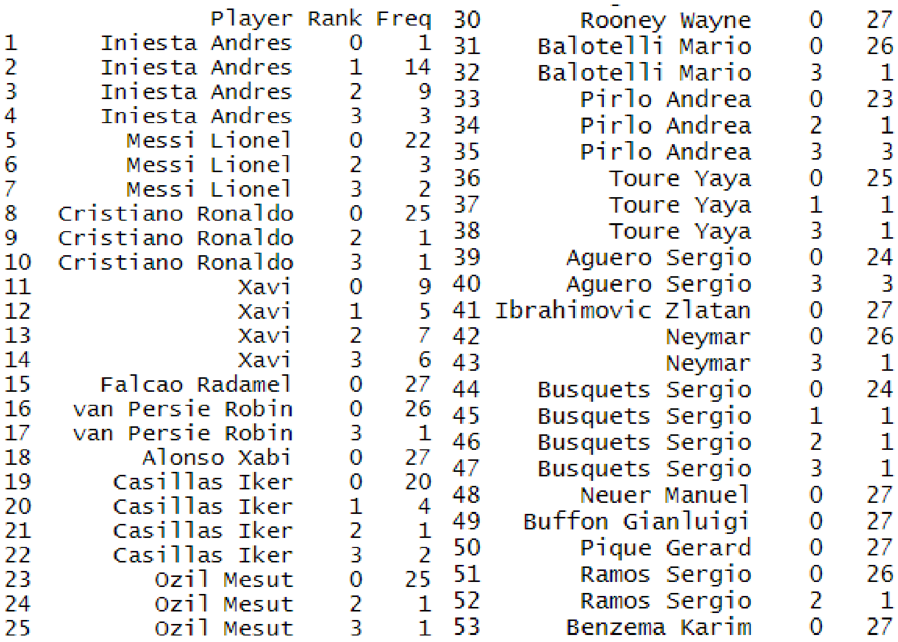
```
    post_pred_pvalue_top1 post_pred_pvalue_paired
G_2               0.27620                 0.42405
G_3               0.27545                 0.35520
G_4               0.27650                 0.30585
G_5               0.27010                 0.25785
G_6               0.26450                 0.22245
G_7               0.26440                 0.18275
G_8               0.00000                 0.00000
```

Table A.13 : P-values for the Bayesian PL models assessment, Year 2012.



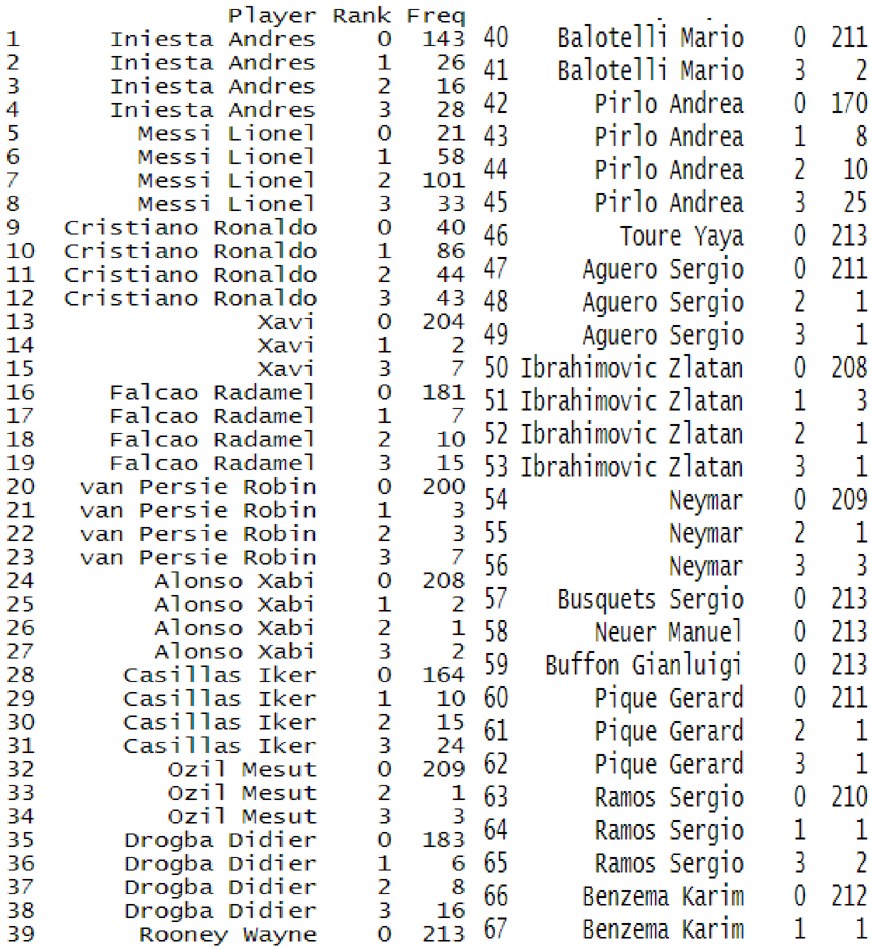
```
             Player Rank Freq  31     Balotelli Mario    0   36
1      Iniesta Andres    0   36  32        Pirlo Andrea    0   30
2         Messi Lionel    0    5  33        Pirlo Andrea    1    1
3         Messi Lionel    1   16  34        Pirlo Andrea    2    5
4         Messi Lionel    2    7  35          Toure Yaya    0   28
5         Messi Lionel    3    8  36          Toure Yaya    1    2
6   Cristiano Ronaldo    0   36  37          Toure Yaya    2    3
7                Xavi    0   23  38          Toure Yaya    3    3
8                Xavi    1    6  39       Aguero Sergio    0   36
9                Xavi    2    4  40 Ibrahimovic Zlatan    0   30
10               Xavi    3    3  41 Ibrahimovic Zlatan    1    2
11     Falcao Radamel    0   34  42 Ibrahimovic Zlatan    2    2
12     Falcao Radamel    1    1  43 Ibrahimovic Zlatan    3    2
13     Falcao Radamel    2    1  44              Neymar    0   32
14   van Persie Robin    0   35  45              Neymar    2    2
15   van Persie Robin    3    1  46              Neymar    3    2
16        Alonso Xabi    0   33  47     Busquets Sergio    0   36
17        Alonso Xabi    1    2  48        Neuer Manuel    0   30
18        Alonso Xabi    3    1  49        Neuer Manuel    2    2
19      Casillas Iker    0   36  50        Neuer Manuel    3    4
20         Ozil Mesut    0   31  51    Buffon Gianluigi    0   32
21         Ozil Mesut    1    1  52    Buffon Gianluigi    1    2
22         Ozil Mesut    3    4  53    Buffon Gianluigi    2    1
23      Drogba Didier    0   25  54    Buffon Gianluigi    3    1
24      Drogba Didier    1    2  55        Pique Gerard    0   35
25      Drogba Didier    2    6  56        Pique Gerard    3    1
26      Drogba Didier    3    3  57        Ramos Sergio    0   36
27       Rooney Wayne    0   29  58       Benzema Karim    0   36
28       Rooney Wayne    1    1
29       Rooney Wayne    2    3
30       Rooney Wayne    3    3
```

Table A.14 : The amount of votes that each player received in Cluster 1, Year 2012.



```
             Player Rank Freq  30        Rooney Wayne    0   27
1      Iniesta Andres    0    1  31     Balotelli Mario    0   26
2      Iniesta Andres    1   14  32     Balotelli Mario    3    1
3      Iniesta Andres    2    9  33        Pirlo Andrea    0   23
4      Iniesta Andres    3    3  34        Pirlo Andrea    2    1
5        Messi Lionel    0   22  35        Pirlo Andrea    3    3
6        Messi Lionel    2    3  36          Toure Yaya    0   25
7        Messi Lionel    3    2  37          Toure Yaya    1    1
8   Cristiano Ronaldo    0   25  38          Toure Yaya    3    1
9   Cristiano Ronaldo    2    1  39       Aguero Sergio    0   24
10  Cristiano Ronaldo    3    1  40       Aguero Sergio    3    3
11               Xavi    0    9  41 Ibrahimovic Zlatan    0   27
12               Xavi    1    5  42              Neymar    0   26
13               Xavi    2    7  43              Neymar    3    1
14               Xavi    3    6  44     Busquets Sergio    0   24
15     Falcao Radamel    0   27  45     Busquets Sergio    1    1
16   van Persie Robin    0   26  46     Busquets Sergio    2    1
17   van Persie Robin    3    1  47     Busquets Sergio    3    1
18        Alonso Xabi    0   27  48        Neuer Manuel    0   27
19      Casillas Iker    0   20  49    Buffon Gianluigi    0   27
20      Casillas Iker    1    4  50        Pique Gerard    0   27
21      Casillas Iker    2    1  51        Ramos Sergio    0   26
22      Casillas Iker    3    2  52        Ramos Sergio    2    1
23         Ozil Mesut    0   25  53       Benzema Karim    0   27
24         Ozil Mesut    2    1
25         Ozil Mesut    3    1
```

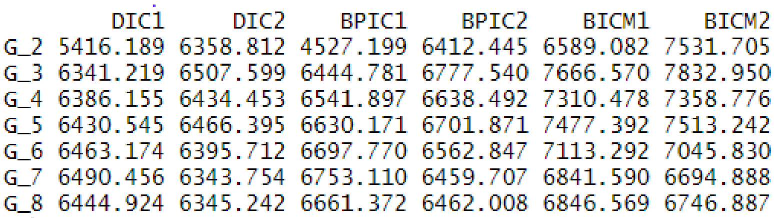
Table A.15 : The amount of votes that each player received in Cluster 3, Year 2012.



```
               Player Rank Freq
1      Iniesta Andres    0  143
2      Iniesta Andres    1   26
3      Iniesta Andres    2   16
4      Iniesta Andres    3   28
5        Messi Lionel    0   21
6        Messi Lionel    1   58
7        Messi Lionel    2  101
8        Messi Lionel    3   33
9   Cristiano Ronaldo    0   40
10  Cristiano Ronaldo    1   86
11  Cristiano Ronaldo    2   44
12  Cristiano Ronaldo    3   43
13               Xavi    0  204
14               Xavi    1    2
15               Xavi    3    7
16     Falcao Radamel    0  181
17     Falcao Radamel    1    7
18     Falcao Radamel    2   10
19     Falcao Radamel    3   15
20    van Persie Robin    0  200
21    van Persie Robin    1    3
22    van Persie Robin    2    3
23    van Persie Robin    3    7
24        Alonso Xabi    0  208
25        Alonso Xabi    1    2
26        Alonso Xabi    2    1
27        Alonso Xabi    3    2
28      Casillas Iker    0  164
29      Casillas Iker    1   10
30      Casillas Iker    2   15
31      Casillas Iker    3   24
32         Ozil Mesut    0  209
33         Ozil Mesut    2    1
34         Ozil Mesut    3    3
35      Drogba Didier    0  183
36      Drogba Didier    1    6
37      Drogba Didier    2    8
38      Drogba Didier    3   16
39       Rooney Wayne    0  213
```

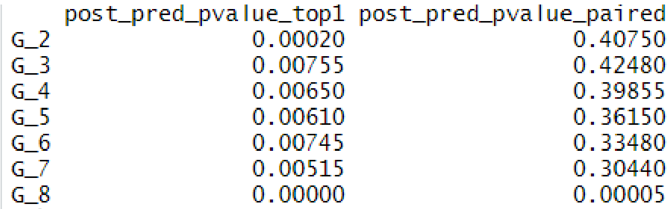
```
40     Balotelli Mario  0  211
41     Balotelli Mario  3    2
42        Pirlo Andrea  0  170
43        Pirlo Andrea  1    8
44        Pirlo Andrea  2   10
45        Pirlo Andrea  3   25
46          Toure Yaya  0  213
47       Aguero Sergio  0  211
48       Aguero Sergio  2    1
49       Aguero Sergio  3    1
50 Ibrahimovic Zlatan  0  208
51 Ibrahimovic Zlatan  1    3
52 Ibrahimovic Zlatan  2    1
53 Ibrahimovic Zlatan  3    1
54              Neymar  0  209
55              Neymar  2    1
56              Neymar  3    3
57     Busquets Sergio  0  213
58        Neuer Manuel  0  213
59    Buffon Gianluigi  0  213
60        Pique Gerard  0  211
61        Pique Gerard  2    1
62        Pique Gerard  3    1
63        Ramos Sergio  0  210
64        Ramos Sergio  1    1
65        Ramos Sergio  3    2
66       Benzema Karim  0  212
67       Benzema Karim  1    1
```

Table A.16 : The amount of votes that each player received in Cluster 2, Year 2012.



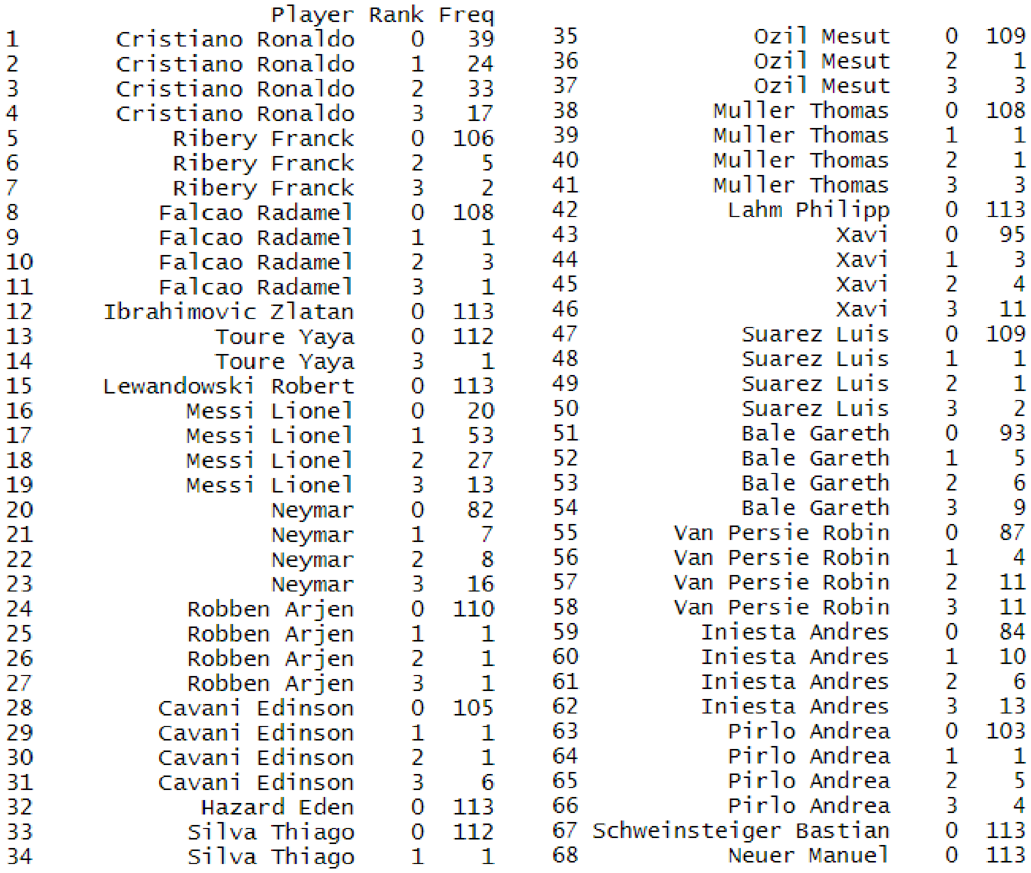
```
        DIC1     DIC2    BPIC1    BPIC2    BICM1    BICM2
G_2 5416.189 6358.812 4527.199 6412.445 6589.082 7531.705
G_3 6341.219 6507.599 6444.781 6777.540 7666.570 7832.950
G_4 6386.155 6434.453 6541.897 6638.492 7310.478 7358.776
G_5 6430.545 6466.395 6630.171 6701.871 7477.392 7513.242
G_6 6463.174 6395.712 6697.770 6562.847 7113.292 7045.830
G_7 6490.456 6343.754 6753.110 6459.707 6841.590 6694.888
G_8 6444.924 6345.242 6661.372 6462.008 6846.569 6746.887
```

Table A.17 : Model comparison criteria values, Year 2013.



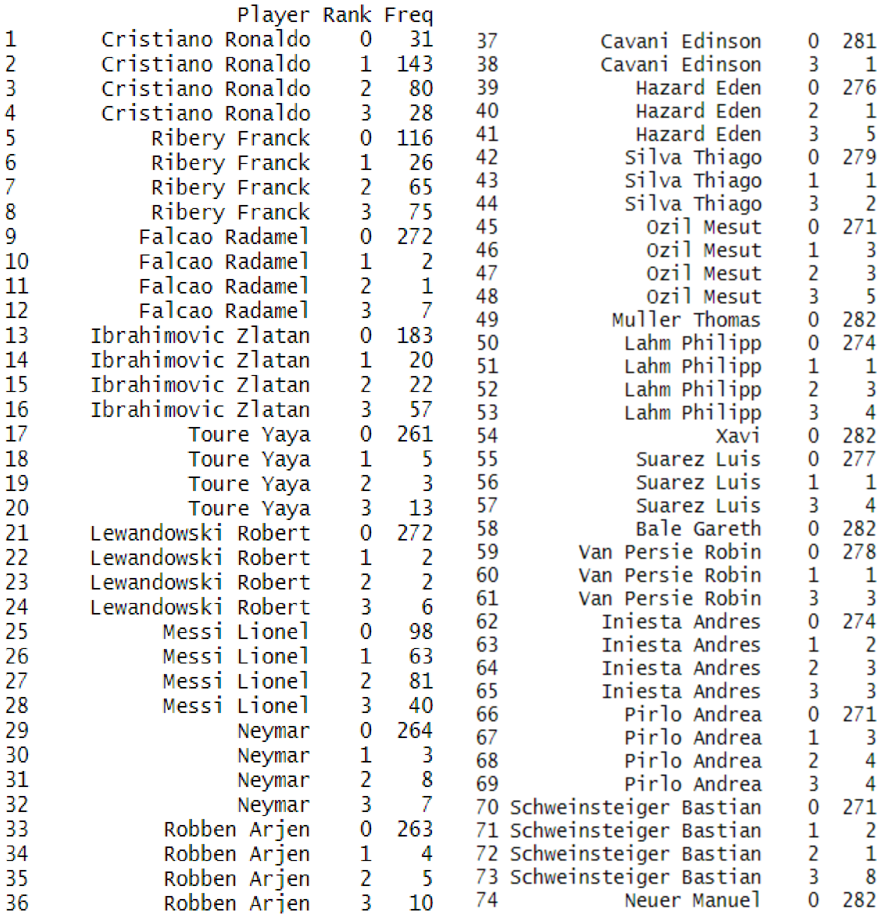
```
    post_pred_pvalue_top1 post_pred_pvalue_paired
G_2               0.00020                 0.40750
G_3               0.00755                 0.42480
G_4               0.00650                 0.39855
G_5               0.00610                 0.36150
G_6               0.00745                 0.33480
G_7               0.00515                 0.30440
G_8               0.00000                 0.00005
```

Table A.18 : P-values for the Bayesian PL models assessment, Year 2013.



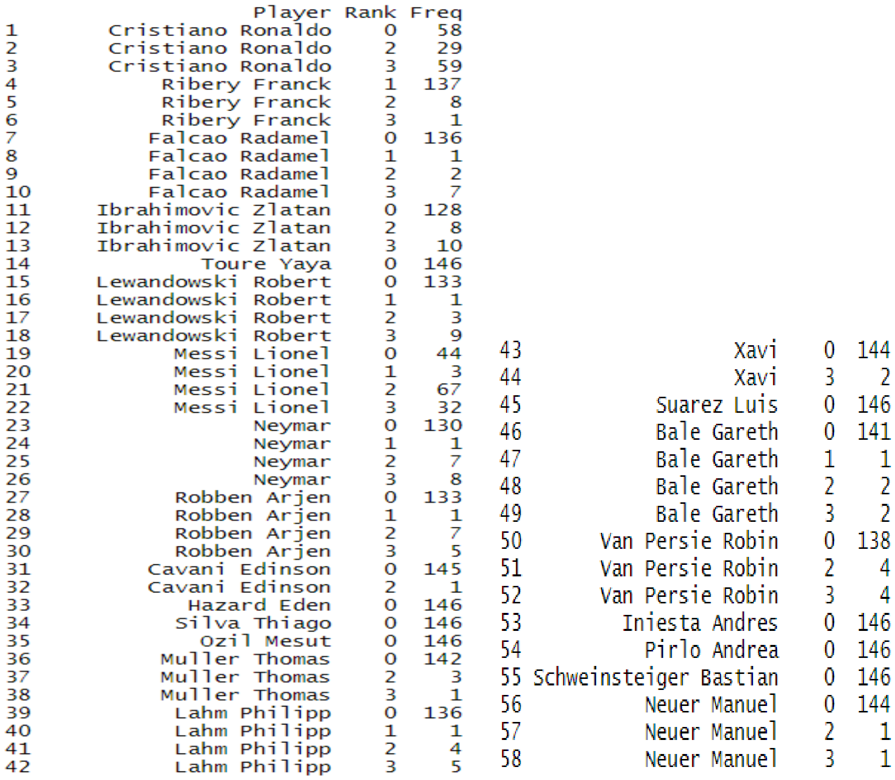
```
                Player Rank Freq    35             Ozil Mesut    0  109
1    Cristiano Ronaldo    0   39    36             Ozil Mesut    2    1
2    Cristiano Ronaldo    1   24    37             Ozil Mesut    3    3
3    Cristiano Ronaldo    2   33    38          Muller Thomas    0  108
4    Cristiano Ronaldo    3   17    39          Muller Thomas    1    1
5        Ribery Franck    0  106    40          Muller Thomas    2    1
6        Ribery Franck    2    5    41          Muller Thomas    3    3
7        Ribery Franck    3    2    42           Lahm Philipp    0  113
8       Falcao Radamel    0  108    43                   Xavi    0   95
9       Falcao Radamel    1    1    44                   Xavi    1    3
10      Falcao Radamel    2    3    45                   Xavi    2    4
11      Falcao Radamel    3    1    46                   Xavi    3   11
12   Ibrahimovic Zlatan   0  113    47            Suarez Luis    0  109
13          Toure Yaya    0  112    48            Suarez Luis    1    1
14          Toure Yaya    3    1    49            Suarez Luis    2    1
15   Lewandowski Robert   0  113    50            Suarez Luis    3    2
16        Messi Lionel    0   20    51            Bale Gareth    0   93
17        Messi Lionel    1   53    52            Bale Gareth    1    5
18        Messi Lionel    2   27    53            Bale Gareth    2    6
19        Messi Lionel    3   13    54            Bale Gareth    3    9
20              Neymar    0   82    55       Van Persie Robin    0   87
21              Neymar    1    7    56       Van Persie Robin    1    4
22              Neymar    2    8    57       Van Persie Robin    2   11
23              Neymar    3   16    58       Van Persie Robin    3   11
24        Robben Arjen    0  110    59         Iniesta Andres    0   84
25        Robben Arjen    1    1    60         Iniesta Andres    1   10
26        Robben Arjen    2    1    61         Iniesta Andres    2    6
27        Robben Arjen    3    1    62         Iniesta Andres    3   13
28      Cavani Edinson    0  105    63           Pirlo Andrea    0  103
29      Cavani Edinson    1    1    64           Pirlo Andrea    1    1
30      Cavani Edinson    2    1    65           Pirlo Andrea    2    5
31      Cavani Edinson    3    6    66           Pirlo Andrea    3    4
32         Hazard Eden    0  113    67 Schweinsteiger Bastian    0  113
33        Silva Thiago    0  112    68           Neuer Manuel    0  113
34        Silva Thiago    1    1
```

Table A.19 : The amount of votes that each player received in Cluster 1, Year 2013.



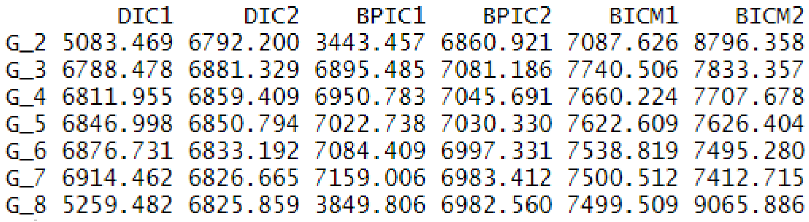
```
                   Player Rank Freq
1       Cristiano Ronaldo    0   31    37         Cavani Edinson    0  281
2       Cristiano Ronaldo    1  143    38         Cavani Edinson    3    1
3       Cristiano Ronaldo    2   80    39            Hazard Eden    0  276
4       Cristiano Ronaldo    3   28    40            Hazard Eden    2    1
5           Ribery Franck    0  116    41            Hazard Eden    3    5
6           Ribery Franck    1   26    42           Silva Thiago    0  279
7           Ribery Franck    2   65    43           Silva Thiago    1    1
8           Ribery Franck    3   75    44           Silva Thiago    3    2
9          Falcao Radamel    0  272    45             Ozil Mesut    0  271
10         Falcao Radamel    1    2    46             Ozil Mesut    1    3
11         Falcao Radamel    2    1    47             Ozil Mesut    2    3
12         Falcao Radamel    3    7    48             Ozil Mesut    3    5
13     Ibrahimovic Zlatan    0  183    49          Muller Thomas    0  282
14     Ibrahimovic Zlatan    1   20    50           Lahm Philipp    0  274
15     Ibrahimovic Zlatan    2   22    51           Lahm Philipp    1    1
16     Ibrahimovic Zlatan    3   57    52           Lahm Philipp    2    3
17             Toure Yaya    0  261    53           Lahm Philipp    3    4
18             Toure Yaya    1    5    54                   Xavi    0  282
19             Toure Yaya    2    3    55            Suarez Luis    0  277
20             Toure Yaya    3   13    56            Suarez Luis    1    1
21     Lewandowski Robert    0  272    57            Suarez Luis    3    4
22     Lewandowski Robert    1    2    58            Bale Gareth    0  282
23     Lewandowski Robert    2    2    59       Van Persie Robin    0  278
24     Lewandowski Robert    3    6    60       Van Persie Robin    1    1
25           Messi Lionel    0   98    61       Van Persie Robin    3    3
26           Messi Lionel    1   63    62         Iniesta Andres    0  274
27           Messi Lionel    2   81    63         Iniesta Andres    1    2
28           Messi Lionel    3   40    64         Iniesta Andres    2    3
29                 Neymar    0  264    65         Iniesta Andres    3    3
30                 Neymar    1    3    66           Pirlo Andrea    0  271
31                 Neymar    2    8    67           Pirlo Andrea    1    3
32                 Neymar    3    7    68           Pirlo Andrea    2    4
33           Robben Arjen    0  263    69           Pirlo Andrea    3    4
34           Robben Arjen    1    4    70 Schweinsteiger Bastian    0  271
35           Robben Arjen    2    5    71 Schweinsteiger Bastian    1    2
36           Robben Arjen    3   10    72 Schweinsteiger Bastian    2    1
                                       73 Schweinsteiger Bastian    3    8
                                       74           Neuer Manuel    0  282
```

Table A.20 : The amount of votes that each player received in Cluster 2, Year 2013.



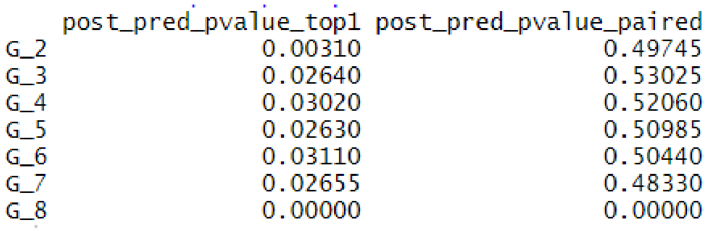
```
                   Player Rank Freq
1       Cristiano Ronaldo    0   58
2       Cristiano Ronaldo    2   29
3       Cristiano Ronaldo    3   59
4           Ribery Franck    1  137
5           Ribery Franck    2    8
6           Ribery Franck    3    1
7          Falcao Radamel    0  136
8          Falcao Radamel    1    1
9          Falcao Radamel    2    2
10         Falcao Radamel    3    7
11     Ibrahimovic Zlatan    0  128
12     Ibrahimovic Zlatan    2    8
13     Ibrahimovic Zlatan    3   10
14             Toure Yaya    0  146
15     Lewandowski Robert    0  133
16     Lewandowski Robert    1    1
17     Lewandowski Robert    2    3
18     Lewandowski Robert    3    9
19           Messi Lionel    0   44    43                   Xavi    0  144
20           Messi Lionel    1    3    44                   Xavi    3    2
21           Messi Lionel    2   67    45            Suarez Luis    0  146
22           Messi Lionel    3   32    46            Bale Gareth    0  141
23                 Neymar    0  130    47            Bale Gareth    1    1
24                 Neymar    1    1    48            Bale Gareth    2    2
25                 Neymar    2    7    49            Bale Gareth    3    2
26                 Neymar    3    8    50       Van Persie Robin    0  138
27           Robben Arjen    0  133    51       Van Persie Robin    2    4
28           Robben Arjen    1    1    52       Van Persie Robin    3    4
29           Robben Arjen    2    7    53         Iniesta Andres    0  146
30           Robben Arjen    3    5    54           Pirlo Andrea    0  146
31         Cavani Edinson    0  145    55 Schweinsteiger Bastian    0  146
32         Cavani Edinson    2    1    56           Neuer Manuel    0  144
33            Hazard Eden    0  146    57           Neuer Manuel    2    1
34           Silva Thiago    0  146    58           Neuer Manuel    3    1
35             Ozil Mesut    0  146
36          Muller Thomas    0  142
37          Muller Thomas    2    3
38          Muller Thomas    3    1
39           Lahm Philipp    0  136
40           Lahm Philipp    1    1
41           Lahm Philipp    2    4
42           Lahm Philipp    3    5
```

Table A.21 : The amount of votes that each player received in Cluster 3, Year 2013.



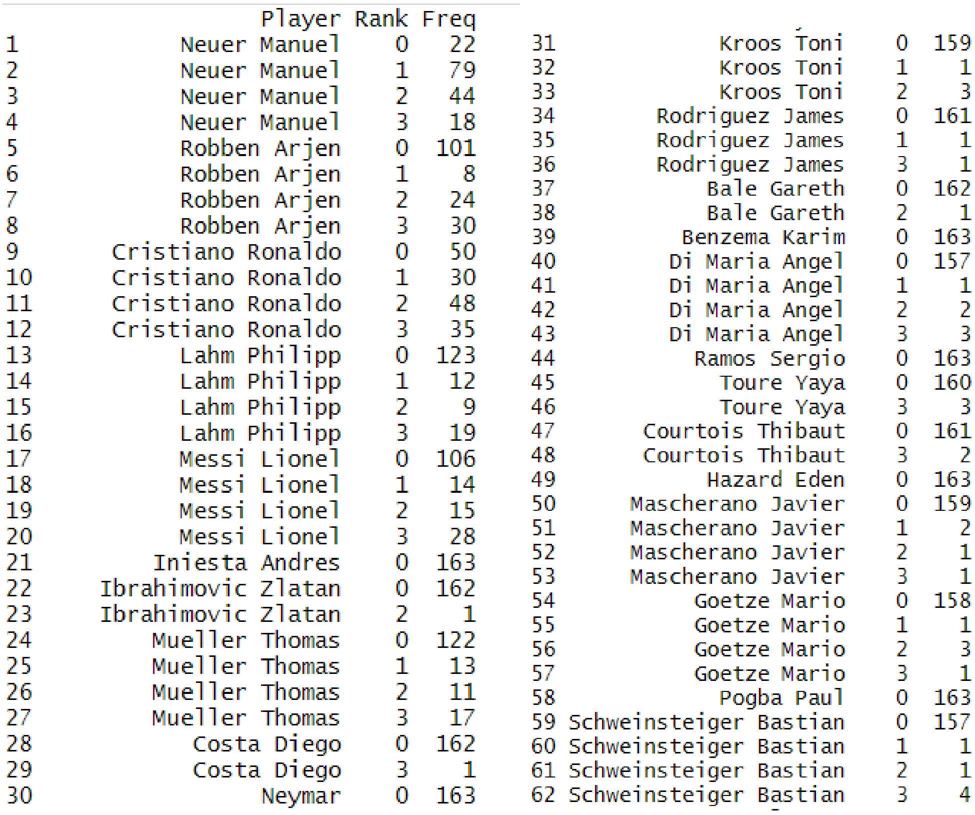
```
        DIC1     DIC2    BPIC1    BPIC2    BICM1    BICM2
G_2 5083.469 6792.200 3443.457 6860.921 7087.626 8796.358
G_3 6788.478 6881.329 6895.485 7081.186 7740.506 7833.357
G_4 6811.955 6859.409 6950.783 7045.691 7660.224 7707.678
G_5 6846.998 6850.794 7022.738 7030.330 7622.609 7626.404
G_6 6876.731 6833.192 7084.409 6997.331 7538.819 7495.280
G_7 6914.462 6826.665 7159.006 6983.412 7500.512 7412.715
G_8 5259.482 6825.859 3849.806 6982.560 7499.509 9065.886
```

Table A.22 : Model comparison criteria values, Year 2014.



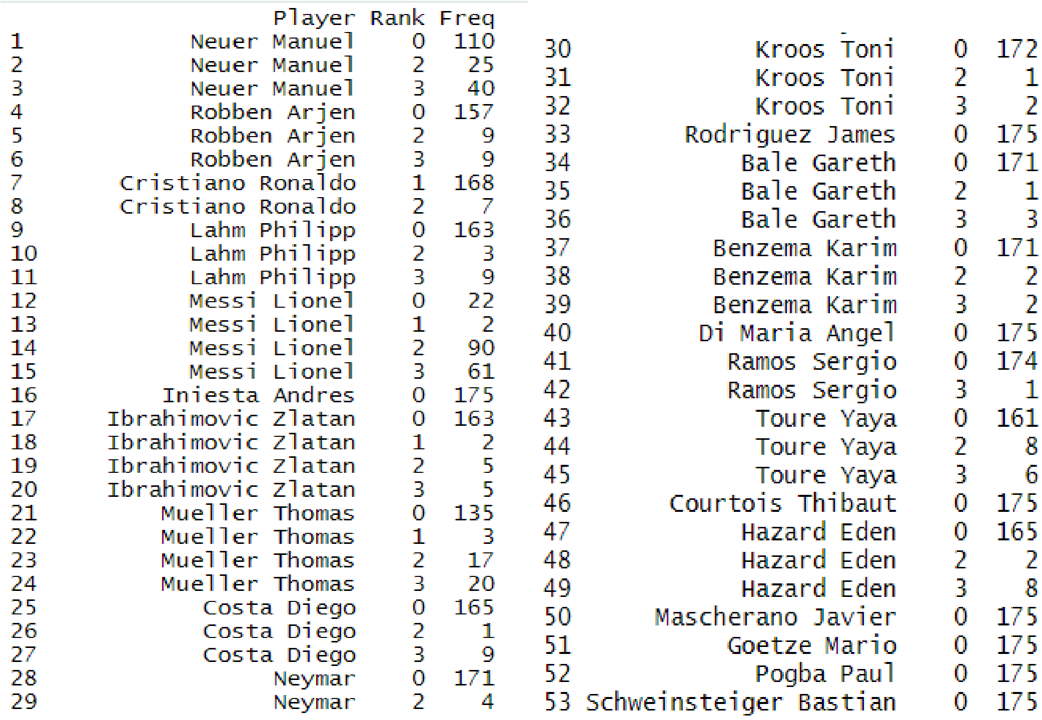
```
    post_pred_pvalue_top1 post_pred_pvalue_paired
G_2               0.00310                 0.49745
G_3               0.02640                 0.53025
G_4               0.03020                 0.52060
G_5               0.02630                 0.50985
G_6               0.03110                 0.50440
G_7               0.02655                 0.48330
G_8               0.00000                 0.00000
```

Table A.23 : P-values for the Bayesian PL models assessment, Year 2014.



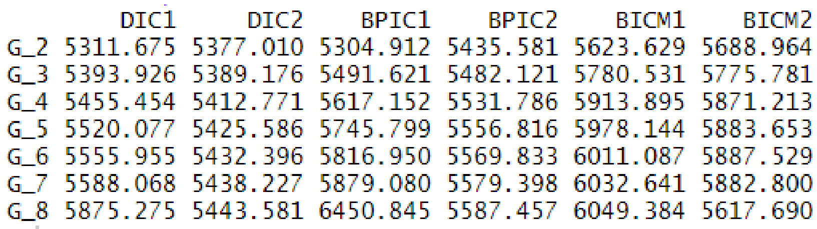
```
                   Player Rank Freq
1            Neuer Manuel    0   22      31           Kroos Toni    0  159
2            Neuer Manuel    1   79      32           Kroos Toni    1    1
3            Neuer Manuel    2   44      33           Kroos Toni    2    3
4            Neuer Manuel    3   18      34      Rodriguez James    0  161
5            Robben Arjen    0  101      35      Rodriguez James    1    1
6            Robben Arjen    1    8      36      Rodriguez James    3    1
7            Robben Arjen    2   24      37          Bale Gareth    0  162
8            Robben Arjen    3   30      38          Bale Gareth    2    1
9       Cristiano Ronaldo    0   50      39        Benzema Karim    0  163
10      Cristiano Ronaldo    1   30      40       Di Maria Angel    0  157
11      Cristiano Ronaldo    2   48      41       Di Maria Angel    1    1
12      Cristiano Ronaldo    3   35      42       Di Maria Angel    2    2
13           Lahm Philipp    0  123      43       Di Maria Angel    3    3
14           Lahm Philipp    1   12      44         Ramos Sergio    0  163
15           Lahm Philipp    2    9      45           Toure Yaya    0  160
16           Lahm Philipp    3   19      46           Toure Yaya    3    3
17            Messi Lionel    0  106      47     Courtois Thibaut    0  161
18            Messi Lionel    1   14      48     Courtois Thibaut    3    2
19            Messi Lionel    2   15      49          Hazard Eden    0  163
20            Messi Lionel    3   28      50    Mascherano Javier    0  159
21         Iniesta Andres    0  163      51    Mascherano Javier    1    2
22      Ibrahimovic Zlatan    0  162      52    Mascherano Javier    2    1
23      Ibrahimovic Zlatan    2    1      53    Mascherano Javier    3    1
24         Mueller Thomas    0  122      54         Goetze Mario    0  158
25         Mueller Thomas    1   13      55         Goetze Mario    1    1
26         Mueller Thomas    2   11      56         Goetze Mario    2    3
27         Mueller Thomas    3   17      57         Goetze Mario    3    1
28            Costa Diego    0  162      58           Pogba Paul    0  163
29            Costa Diego    3    1      59 Schweinsteiger Bastian    0  157
30                 Neymar    0  163      60 Schweinsteiger Bastian    1    1
                                         61 Schweinsteiger Bastian    2    1
                                         62 Schweinsteiger Bastian    3    4
```

Table A.24 : The amount of votes that each player received in Cluster 2, Year 2014.



```
                  Player Rank Freq
1          Neuer Manuel    0  110    30          Kroos Toni    0  172
2          Neuer Manuel    2   25    31          Kroos Toni    2    1
3          Neuer Manuel    3   40    32          Kroos Toni    3    2
4          Robben Arjen    0  157    33     Rodriguez James    0  175
5          Robben Arjen    2    9    34         Bale Gareth    0  171
6          Robben Arjen    3    9    35         Bale Gareth    2    1
7     Cristiano Ronaldo    1  168    36         Bale Gareth    3    3
8     Cristiano Ronaldo    2    7    37       Benzema Karim    0  171
9          Lahm Philipp    0  163    38       Benzema Karim    2    2
10         Lahm Philipp    2    3    39       Benzema Karim    3    2
11         Lahm Philipp    3    9    40      Di Maria Angel    0  175
12         Messi Lionel    0   22    41        Ramos Sergio    0  174
13         Messi Lionel    1    2    42        Ramos Sergio    3    1
14         Messi Lionel    2   90    43          Toure Yaya    0  161
15         Messi Lionel    3   61    44          Toure Yaya    2    8
16       Iniesta Andres    0  175    45          Toure Yaya    3    6
17    Ibrahimovic Zlatan   0  163    46    Courtois Thibaut    0  175
18    Ibrahimovic Zlatan   1    2    47         Hazard Eden    0  165
19    Ibrahimovic Zlatan   2    5    48         Hazard Eden    2    2
20    Ibrahimovic Zlatan   3    5    49         Hazard Eden    3    8
21        Mueller Thomas   0  135    50   Mascherano Javier    0  175
22        Mueller Thomas   1    3    51        Goetze Mario    0  175
23        Mueller Thomas   2   17    52          Pogba Paul    0  175
24        Mueller Thomas   3   20    53 Schweinsteiger Bastian 0  175
25          Costa Diego    0  165
26          Costa Diego    2    1
27          Costa Diego    3    9
28               Neymar    0  171
29               Neymar    2    4
```

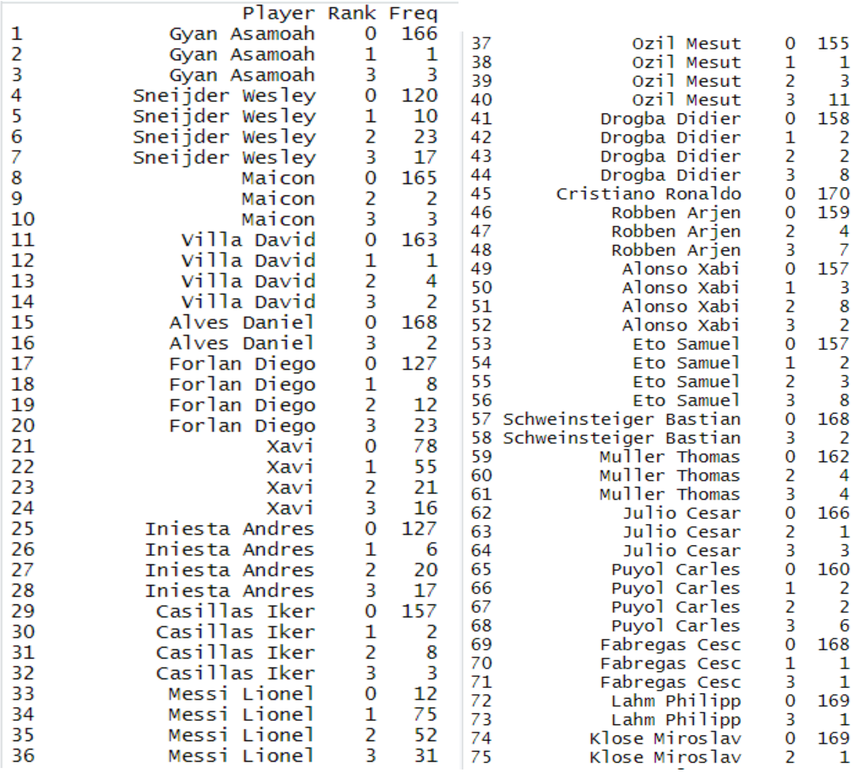
Table A.25 : The amount of votes that each player received in Cluster 6, Year 2014.



```
        DIC1     DIC2    BPIC1    BPIC2    BICM1    BICM2
G_2 5311.675 5377.010 5304.912 5435.581 5623.629 5688.964
G_3 5393.926 5389.176 5491.621 5482.121 5780.531 5775.781
G_4 5455.454 5412.771 5617.152 5531.786 5913.895 5871.213
G_5 5520.077 5425.586 5745.799 5556.816 5978.144 5883.653
G_6 5555.955 5432.396 5816.950 5569.833 6011.087 5887.529
G_7 5588.068 5438.227 5879.080 5579.398 6032.641 5882.800
G_8 5875.275 5443.581 6450.845 5587.457 6049.384 5617.690
```

Table A.26 : Model comparison criteria values, Year 2015.

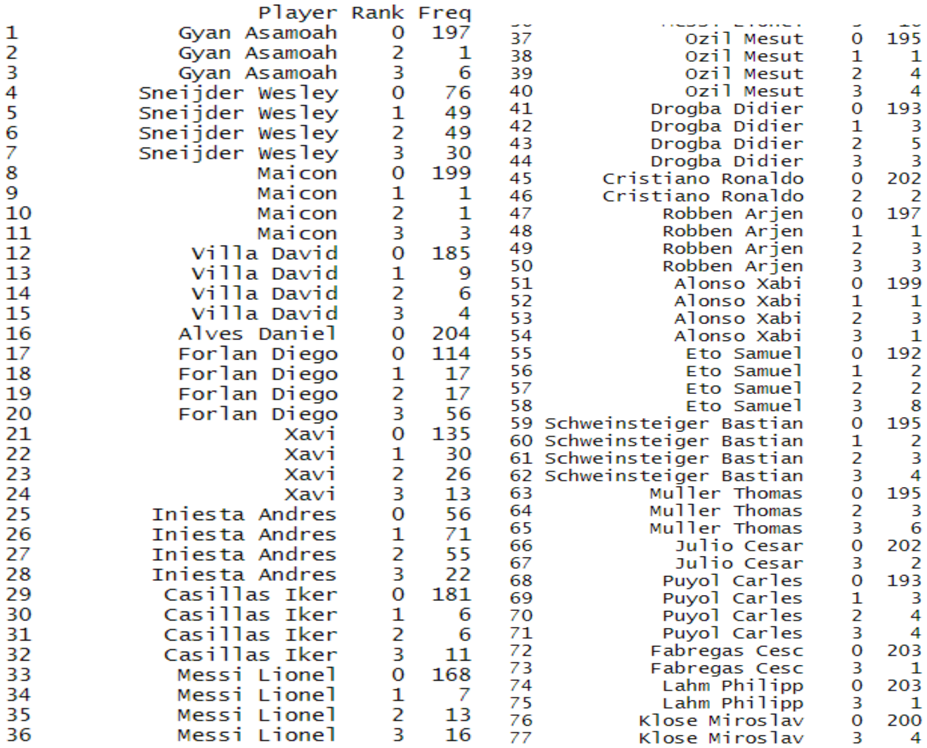
Chapter 6



```
                   Player Rank Freq
1            Gyan Asamoah    0  166
2            Gyan Asamoah    1    1
3            Gyan Asamoah    3    3
4         Sneijder Wesley    0  120
5         Sneijder Wesley    1   10
6         Sneijder Wesley    2   23
7         Sneijder Wesley    3   17
8                  Maicon    0  165
9                  Maicon    2    2
10                 Maicon    3    3
11            Villa David    0  163
12            Villa David    1    1
13            Villa David    2    4
14            Villa David    3    2
15           Alves Daniel    0  168
16           Alves Daniel    3    2
17           Forlan Diego    0  127
18           Forlan Diego    1    8
19           Forlan Diego    2   12
20           Forlan Diego    3   23
21                   Xavi    0   78
22                   Xavi    1   55
23                   Xavi    2   21
24                   Xavi    3   16
25         Iniesta Andres    0  127
26         Iniesta Andres    1    6
27         Iniesta Andres    2   20
28         Iniesta Andres    3   17
29          Casillas Iker    0  157
30          Casillas Iker    1    2
31          Casillas Iker    2    8
32          Casillas Iker    3    3
33           Messi Lionel    0   12
34           Messi Lionel    1   75
35           Messi Lionel    2   52
36           Messi Lionel    3   31
```

```
37             Ozil Mesut    0  155
38             Ozil Mesut    1    1
39             Ozil Mesut    2    3
40             Ozil Mesut    3   11
41          Drogba Didier    0  158
42          Drogba Didier    1    2
43          Drogba Didier    2    2
44          Drogba Didier    3    8
45      Cristiano Ronaldo    0  170
46           Robben Arjen    0  159
47           Robben Arjen    2    4
48           Robben Arjen    3    7
49            Alonso Xabi    0  157
50            Alonso Xabi    1    3
51            Alonso Xabi    2    8
52            Alonso Xabi    3    2
53             Eto Samuel    0  157
54             Eto Samuel    1    2
55             Eto Samuel    2    3
56             Eto Samuel    3    8
57 Schweinsteiger Bastian    0  168
58 Schweinsteiger Bastian    3    2
59          Muller Thomas    0  162
60          Muller Thomas    2    4
61          Muller Thomas    3    4
62            Julio Cesar    0  166
63            Julio Cesar    2    1
64            Julio Cesar    3    3
65           Puyol Carles    0  160
66           Puyol Carles    1    2
67           Puyol Carles    2    2
68           Puyol Carles    3    6
69          Fabregas Cesc    0  168
70          Fabregas Cesc    1    1
71          Fabregas Cesc    3    1
72           Lahm Philipp    0  169
73           Lahm Philipp    3    1
74         Klose Miroslav    0  169
75         Klose Miroslav    2    1
```

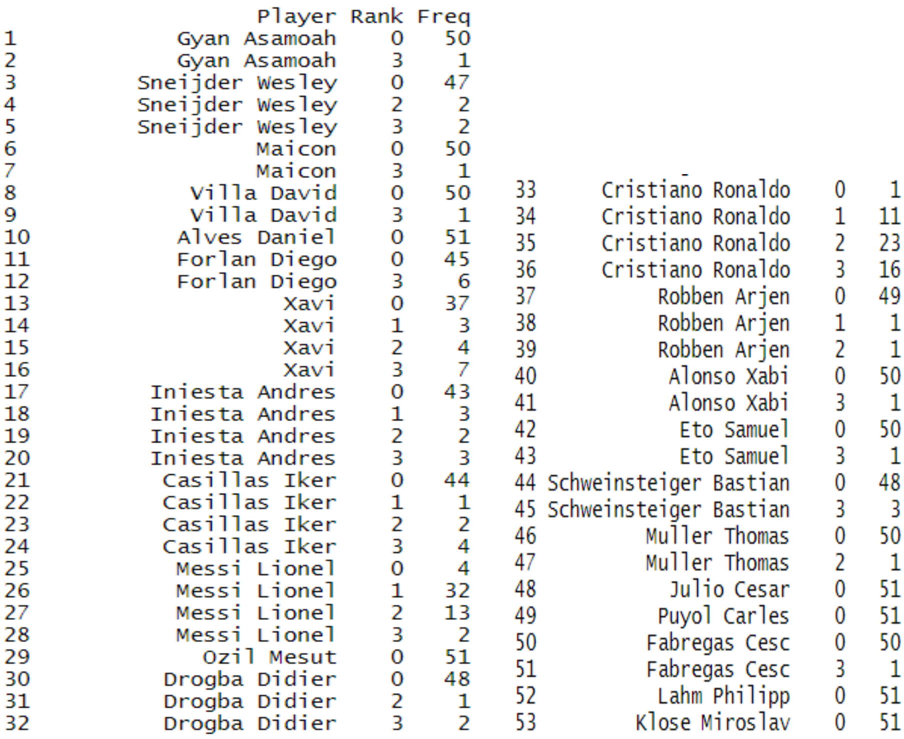
Table A.27 : The amount of votes that each player received in Cluster 1, Year 2010.



```
                   Player Rank Freq
1            Gyan Asamoah    0  197
2            Gyan Asamoah    2    1
3            Gyan Asamoah    3    6
4         Sneijder Wesley    0   76
5         Sneijder Wesley    1   49
6         Sneijder Wesley    2   49
7         Sneijder Wesley    3   30
8                  Maicon    0  199
9                  Maicon    1    1
10                 Maicon    2    1
11                 Maicon    3    3
12            Villa David    0  185
13            Villa David    1    9
14            Villa David    2    6
15            Villa David    3    4
16           Alves Daniel    0  204
17           Forlan Diego    0  114
18           Forlan Diego    1   17
19           Forlan Diego    2   17
20           Forlan Diego    3   56
21                   Xavi    0  135
22                   Xavi    1   30
23                   Xavi    2   26
24                   Xavi    3   13
25         Iniesta Andres    0   56
26         Iniesta Andres    1   71
27         Iniesta Andres    2   55
28         Iniesta Andres    3   22
29          Casillas Iker    0  181
30          Casillas Iker    1    6
31          Casillas Iker    2    6
32          Casillas Iker    3   11
33           Messi Lionel    0  168
34           Messi Lionel    1    7
35           Messi Lionel    2   13
36           Messi Lionel    3   16
```

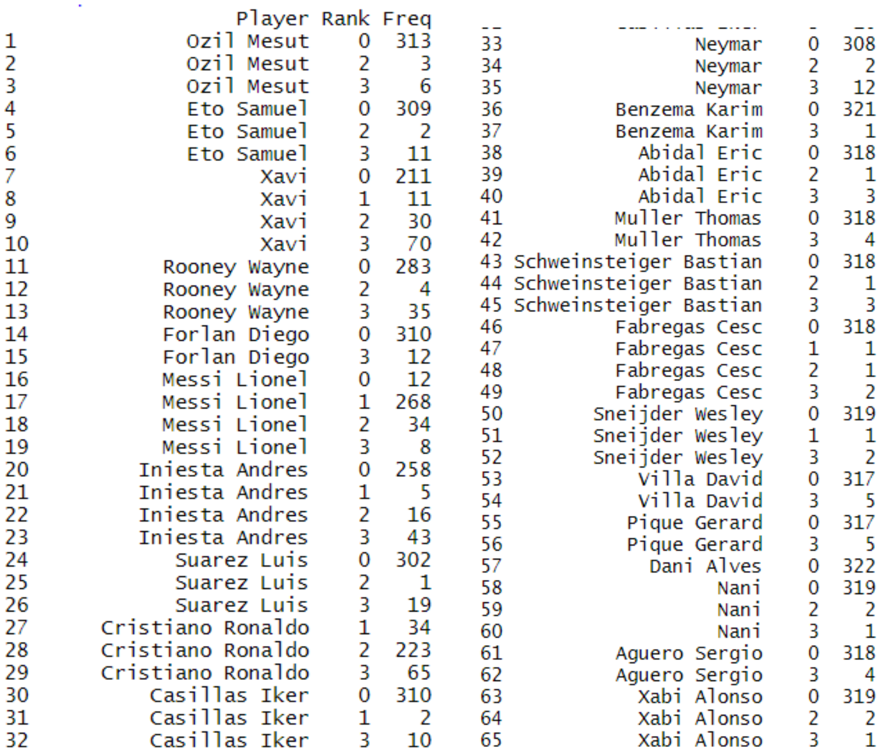
```
37             Ozil Mesut    0  195
38             Ozil Mesut    1    1
39             Ozil Mesut    2    4
40             Ozil Mesut    3    4
41          Drogba Didier    0  193
42          Drogba Didier    1    3
43          Drogba Didier    2    5
44          Drogba Didier    3    3
45      Cristiano Ronaldo    0  202
46      Cristiano Ronaldo    2    2
47           Robben Arjen    0  197
48           Robben Arjen    1    1
49           Robben Arjen    2    3
50           Robben Arjen    3    3
51            Alonso Xabi    0  199
52            Alonso Xabi    1    1
53            Alonso Xabi    2    3
54            Alonso Xabi    3    1
55             Eto Samuel    0  192
56             Eto Samuel    1    2
57             Eto Samuel    2    2
58             Eto Samuel    3    8
59 Schweinsteiger Bastian    0  195
60 Schweinsteiger Bastian    1    2
61 Schweinsteiger Bastian    2    3
62 Schweinsteiger Bastian    3    4
63          Muller Thomas    0  195
64          Muller Thomas    2    3
65          Muller Thomas    3    6
66            Julio Cesar    0  202
67            Julio Cesar    3    2
68           Puyol Carles    0  193
69           Puyol Carles    1    3
70           Puyol Carles    2    4
71           Puyol Carles    3    4
72          Fabregas Cesc    0  203
73          Fabregas Cesc    3    1
74           Lahm Philipp    0  203
75           Lahm Philipp    3    1
76         Klose Miroslav    0  200
77         Klose Miroslav    3    4
```

Table A.28 : The amount of votes that each player received in Cluster 2, Year 2010.



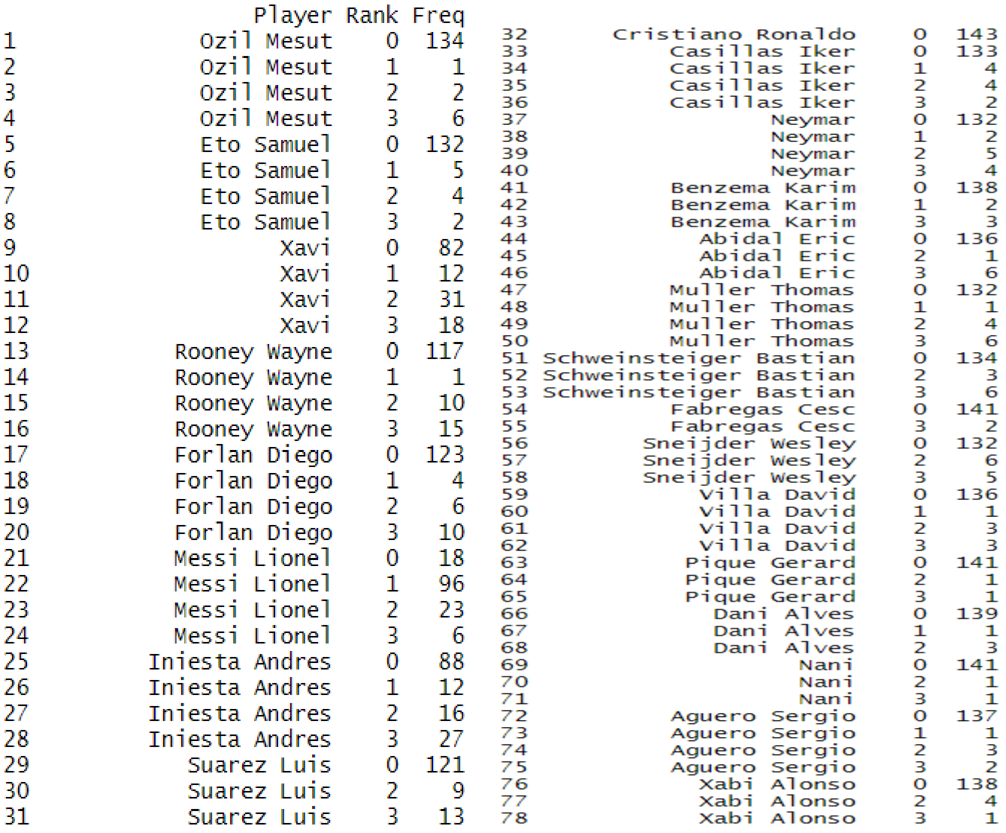
```
              Player Rank Freq
1        Gyan Asamoah    0   50
2        Gyan Asamoah    3    1
3     Sneijder Wesley    0   47
4     Sneijder Wesley    2    2
5     Sneijder Wesley    3    2
6              Maicon    0   50
7              Maicon    3    1
8         Villa David    0   50
9         Villa David    3    1
10       Alves Daniel    0   51
11       Forlan Diego    0   45
12       Forlan Diego    3    6
13               Xavi    0   37
14               Xavi    1    3
15               Xavi    2    4
16               Xavi    3    7
17     Iniesta Andres    0   43
18     Iniesta Andres    1    3
19     Iniesta Andres    2    2
20     Iniesta Andres    3    3
21      Casillas Iker    0   44
22      Casillas Iker    1    1
23      Casillas Iker    2    2
24      Casillas Iker    3    4
25       Messi Lionel    0    4
26       Messi Lionel    1   32
27       Messi Lionel    2   13
28       Messi Lionel    3    2
29         Ozil Mesut    0   51
30      Drogba Didier    0   48
31      Drogba Didier    2    1
32      Drogba Didier    3    2
33      Cristiano Ronaldo    0    1
34      Cristiano Ronaldo    1   11
35      Cristiano Ronaldo    2   23
36      Cristiano Ronaldo    3   16
37            Robben Arjen    0   49
38            Robben Arjen    1    1
39            Robben Arjen    2    1
40             Alonso Xabi    0   50
41             Alonso Xabi    3    1
42              Eto Samuel    0   50
43              Eto Samuel    3    1
44 Schweinsteiger Bastian    0   48
45 Schweinsteiger Bastian    3    3
46           Muller Thomas    0   50
47           Muller Thomas    2    1
48             Julio Cesar    0   51
49            Puyol Carles    0   51
50           Fabregas Cesc    0   50
51           Fabregas Cesc    3    1
52            Lahm Philipp    0   51
53           Klose Miroslav    0   51
```

Table A.29 : The amount of votes that each player received in Cluster 3, Year 2010.



```
              Player Rank Freq
1          Ozil Mesut    0  313
2          Ozil Mesut    2    3
3          Ozil Mesut    3    6
4          Eto Samuel    0  309
5          Eto Samuel    2    2
6          Eto Samuel    3   11
7                Xavi    0  211
8                Xavi    1   11
9                Xavi    2   30
10               Xavi    3   70
11       Rooney Wayne    0  283
12       Rooney Wayne    2    4
13       Rooney Wayne    3   35
14       Forlan Diego    0  310
15       Forlan Diego    3   12
16       Messi Lionel    0   12
17       Messi Lionel    1  268
18       Messi Lionel    2   34
19       Messi Lionel    3    8
20     Iniesta Andres    0  258
21     Iniesta Andres    1    5
22     Iniesta Andres    2   16
23     Iniesta Andres    3   43
24        Suarez Luis    0  302
25        Suarez Luis    2    1
26        Suarez Luis    3   19
27  Cristiano Ronaldo    1   34
28  Cristiano Ronaldo    2  223
29  Cristiano Ronaldo    3   65
30      Casillas Iker    0  310
31      Casillas Iker    1    2
32      Casillas Iker    3   10
33                 Neymar    0  308
34                 Neymar    2    2
35                 Neymar    3   12
36          Benzema Karim    0  321
37          Benzema Karim    3    1
38            Abidal Eric    0  318
39            Abidal Eric    2    1
40            Abidal Eric    3    3
41          Muller Thomas    0  318
42          Muller Thomas    3    4
43 Schweinsteiger Bastian    0  318
44 Schweinsteiger Bastian    2    1
45 Schweinsteiger Bastian    3    3
46          Fabregas Cesc    0  318
47          Fabregas Cesc    1    1
48          Fabregas Cesc    2    1
49          Fabregas Cesc    3    2
50        Sneijder Wesley    0  319
51        Sneijder Wesley    1    1
52        Sneijder Wesley    3    2
53            Villa David    0  317
54            Villa David    3    5
55           Pique Gerard    0  317
56           Pique Gerard    3    5
57             Dani Alves    0  322
58                   Nani    0  319
59                   Nani    2    2
60                   Nani    3    1
61          Aguero Sergio    0  318
62          Aguero Sergio    3    4
63            Xabi Alonso    0  319
64            Xabi Alonso    2    2
65            Xabi Alonso    3    1
```

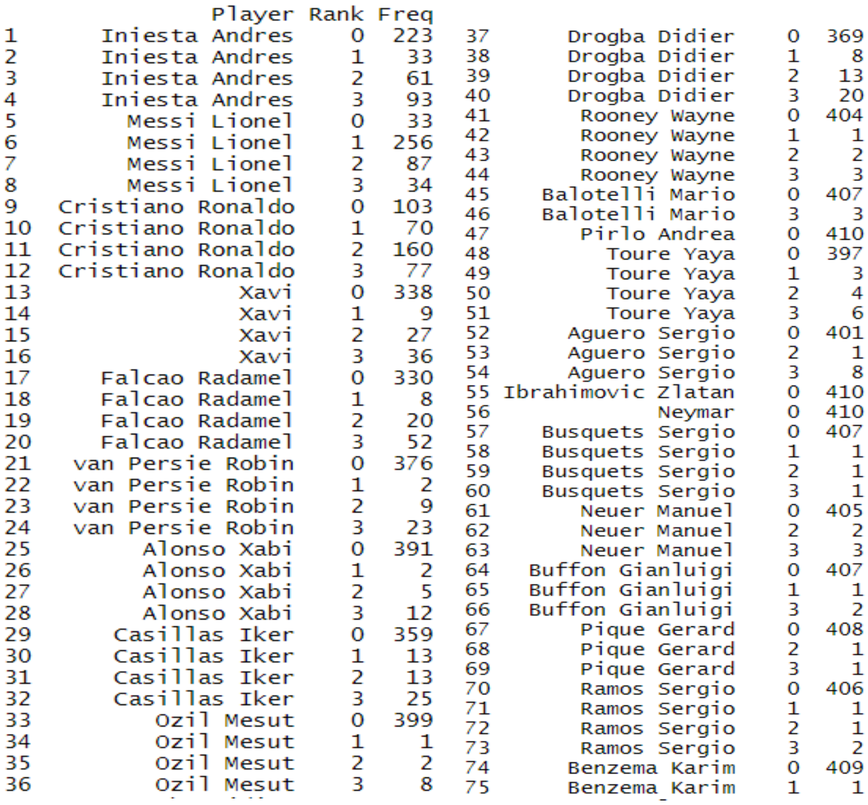
Table A.30 : The amount of votes that each player received in Cluster 1, Year 2011.



```
                Player Rank Freq
1           Ozil Mesut    0  134
2           Ozil Mesut    1    1
3           Ozil Mesut    2    2
4           Ozil Mesut    3    6
5           Eto Samuel    0  132
6           Eto Samuel    1    5
7           Eto Samuel    2    4
8           Eto Samuel    3    2
9                 Xavi    0   82
10                Xavi    1   12
11                Xavi    2   31
12                Xavi    3   18
13        Rooney Wayne    0  117
14        Rooney Wayne    1    1
15        Rooney Wayne    2   10
16        Rooney Wayne    3   15
17        Forlan Diego    0  123
18        Forlan Diego    1    4
19        Forlan Diego    2    6
20        Forlan Diego    3   10
21        Messi Lionel    0   18
22        Messi Lionel    1   96
23        Messi Lionel    2   23
24        Messi Lionel    3    6
25      Iniesta Andres    0   88
26      Iniesta Andres    1   12
27      Iniesta Andres    2   16
28      Iniesta Andres    3   27
29         Suarez Luis    0  121
30         Suarez Luis    2    9
31         Suarez Luis    3   13
```

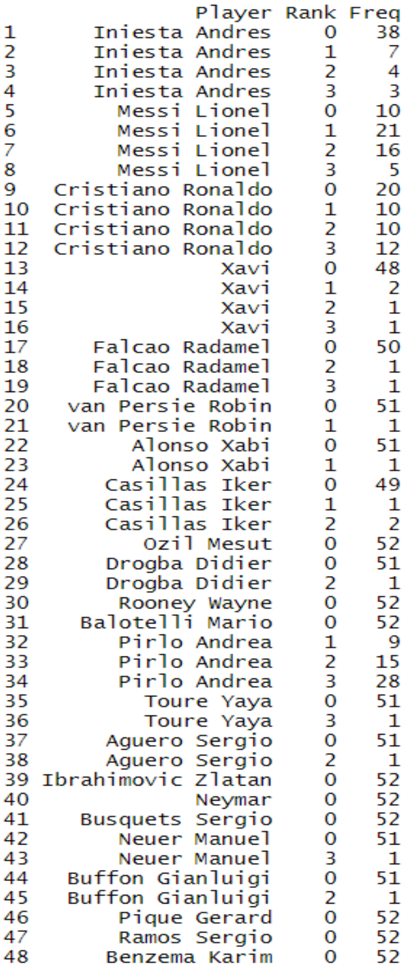
```
32       Cristiano Ronaldo    0  143
33           Casillas Iker    0  133
34           Casillas Iker    1    4
35           Casillas Iker    2    4
36           Casillas Iker    3    2
37                  Neymar    0  132
38                  Neymar    1    2
39                  Neymar    2    5
40                  Neymar    3    4
41           Benzema Karim    0  138
42           Benzema Karim    1    2
43           Benzema Karim    3    3
44             Abidal Eric    0  136
45             Abidal Eric    2    1
46             Abidal Eric    3    6
47           Muller Thomas    0  132
48           Muller Thomas    1    1
49           Muller Thomas    2    4
50           Muller Thomas    3    6
51 Schweinsteiger Bastian    0  134
52 Schweinsteiger Bastian    2    3
53 Schweinsteiger Bastian    3    6
54           Fabregas Cesc    0  141
55           Fabregas Cesc    3    2
56          Sneijder Wesley    0  132
57          Sneijder Wesley    2    6
58          Sneijder Wesley    3    5
59             Villa David    0  136
60             Villa David    1    1
61             Villa David    2    3
62             Villa David    3    3
63            Pique Gerard    0  141
64            Pique Gerard    2    1
65            Pique Gerard    3    1
66              Dani Alves    0  139
67              Dani Alves    1    1
68              Dani Alves    2    3
69                    Nani    0  141
70                    Nani    2    1
71                    Nani    3    1
72           Aguero Sergio    0  137
73           Aguero Sergio    1    1
74           Aguero Sergio    2    3
75           Aguero Sergio    3    2
76             Xabi Alonso    0  138
77             Xabi Alonso    2    4
78             Xabi Alonso    3    1
```

Table A.31 : The amount of votes that each player received in Cluster 2, Year 2011.



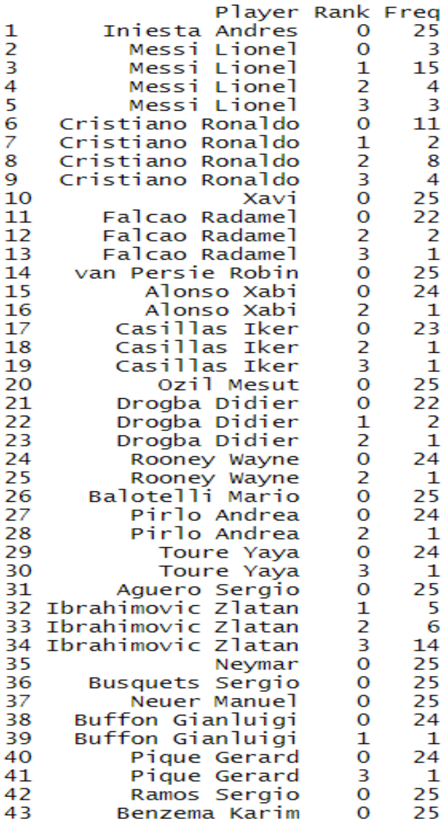
```
              Player Rank Freq
1     Iniesta Andres    0  223   37      Drogba Didier    0  369
2     Iniesta Andres    1   33   38      Drogba Didier    1    8
3     Iniesta Andres    2   61   39      Drogba Didier    2   13
4     Iniesta Andres    3   93   40      Drogba Didier    3   20
5       Messi Lionel    0   33   41       Rooney Wayne    0  404
6       Messi Lionel    1  256   42       Rooney Wayne    1    1
7       Messi Lionel    2   87   43       Rooney Wayne    2    2
8       Messi Lionel    3   34   44       Rooney Wayne    3    3
9  Cristiano Ronaldo    0  103   45    Balotelli Mario    0  407
10 Cristiano Ronaldo    1   70   46    Balotelli Mario    3    3
11 Cristiano Ronaldo    2  160   47      Pirlo Andrea    0  410
12 Cristiano Ronaldo    3   77   48         Toure Yaya    0  397
13              Xavi    0  338   49         Toure Yaya    1    3
14              Xavi    1    9   50         Toure Yaya    2    4
15              Xavi    2   27   51         Toure Yaya    3    6
16              Xavi    3   36   52      Aguero Sergio    0  401
17    Falcao Radamel    0  330   53      Aguero Sergio    2    1
18    Falcao Radamel    1    8   54      Aguero Sergio    3    8
19    Falcao Radamel    2   20   55 Ibrahimovic Zlatan    0  410
20    Falcao Radamel    3   52   56             Neymar    0  410
21  van Persie Robin    0  376   57    Busquets Sergio    0  407
22  van Persie Robin    1    2   58    Busquets Sergio    1    1
23  van Persie Robin    2    9   59    Busquets Sergio    2    1
24  van Persie Robin    3   23   60    Busquets Sergio    3    1
25       Alonso Xabi    0  391   61      Neuer Manuel    0  405
26       Alonso Xabi    1    2   62      Neuer Manuel    2    2
27       Alonso Xabi    2    5   63      Neuer Manuel    3    3
28       Alonso Xabi    3   12   64   Buffon Gianluigi    0  407
29     Casillas Iker    0  359   65   Buffon Gianluigi    1    1
30     Casillas Iker    1   13   66   Buffon Gianluigi    3    2
31     Casillas Iker    2   13   67       Pique Gerard    0  408
32     Casillas Iker    3   25   68       Pique Gerard    2    1
33        Ozil Mesut    0  399   69       Pique Gerard    3    1
34        Ozil Mesut    1    1   70       Ramos Sergio    0  406
35        Ozil Mesut    2    2   71       Ramos Sergio    1    1
36        Ozil Mesut    3    8   72       Ramos Sergio    2    1
                                 73       Ramos Sergio    3    2
                                 74      Benzema Karim    0  409
                                 75      Benzema Karim    1    1
```

Table A.32 : The amount of votes that each player received in Cluster 1, Year 2012.



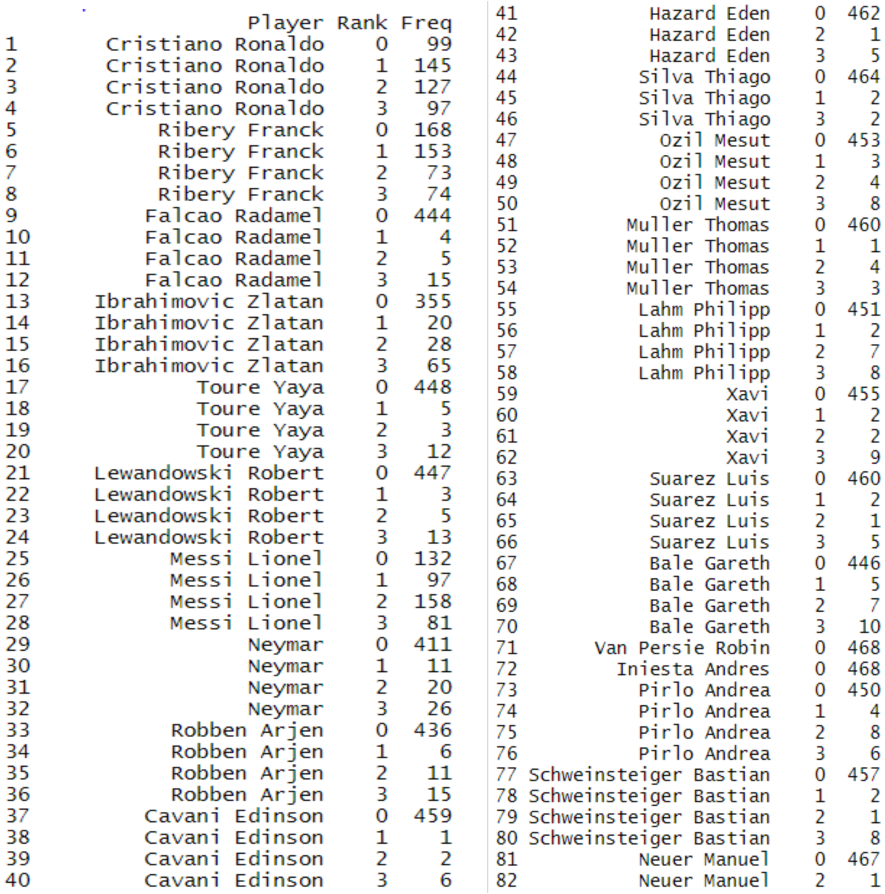
```
              Player Rank Freq
1     Iniesta Andres    0   38
2     Iniesta Andres    1    7
3     Iniesta Andres    2    4
4     Iniesta Andres    3    3
5       Messi Lionel    0   10
6       Messi Lionel    1   21
7       Messi Lionel    2   16
8       Messi Lionel    3    5
9  Cristiano Ronaldo    0   20
10 Cristiano Ronaldo    1   10
11 Cristiano Ronaldo    2   10
12 Cristiano Ronaldo    3   12
13              Xavi    0   48
14              Xavi    1    2
15              Xavi    2    1
16              Xavi    3    1
17    Falcao Radamel    0   50
18    Falcao Radamel    2    1
19    Falcao Radamel    3    1
20   van Persie Robin    0   51
21   van Persie Robin    1    1
22       Alonso Xabi    0   51
23       Alonso Xabi    1    1
24      Casillas Iker    0   49
25      Casillas Iker    1    1
26      Casillas Iker    2    2
27        Ozil Mesut    0   52
28     Drogba Didier    0   51
29     Drogba Didier    2    1
30      Rooney Wayne    0   52
31   Balotelli Mario    0   52
32     Pirlo Andrea    1    9
33     Pirlo Andrea    2   15
34     Pirlo Andrea    3   28
35        Toure Yaya    0   51
36        Toure Yaya    3    1
37     Aguero Sergio    0   51
38     Aguero Sergio    2    1
39 Ibrahimovic Zlatan    0   52
40            Neymar    0   52
41   Busquets Sergio    0   52
42      Neuer Manuel    0   51
43      Neuer Manuel    3    1
44   Buffon Gianluigi    0   51
45   Buffon Gianluigi    2    1
46      Pique Gerard    0   52
47      Ramos Sergio    0   52
48     Benzema Karim    0   52
```

Table A.33 : The amount of votes that each player received in Cluster 2, Year 2012.



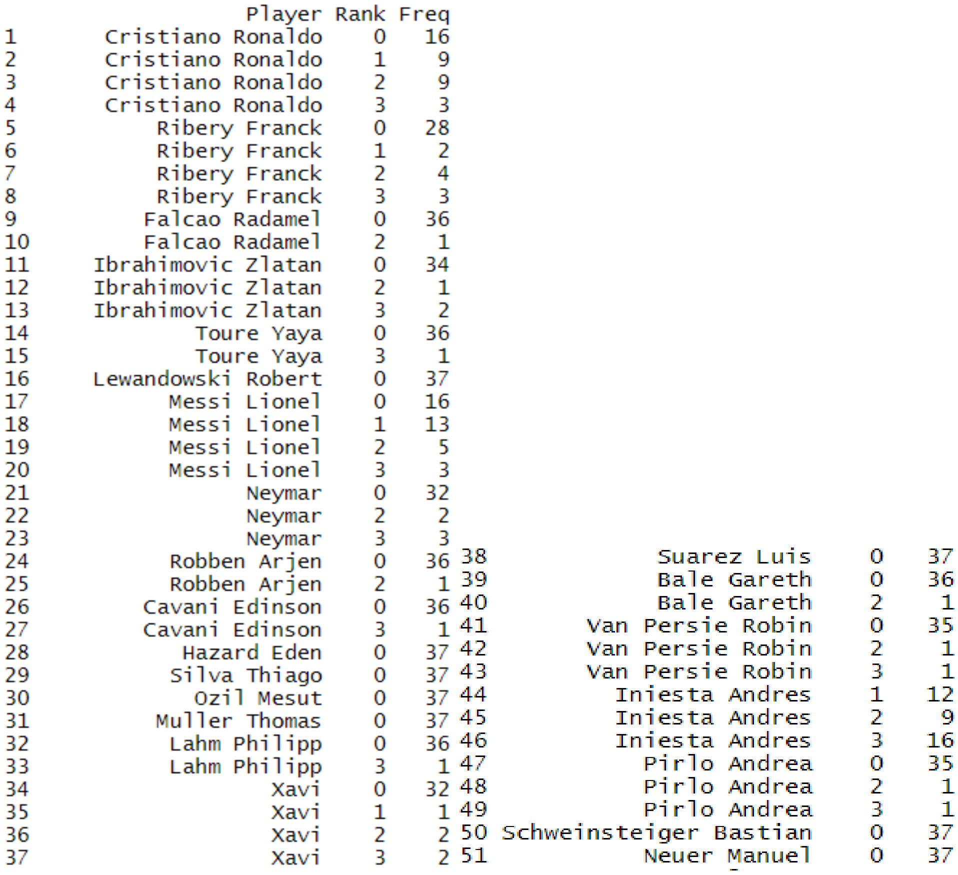
```
          Player Rank Freq
1     Iniesta Andres    0   25
2       Messi Lionel    0    3
3       Messi Lionel    1   15
4       Messi Lionel    2    4
5       Messi Lionel    3    3
6  Cristiano Ronaldo    0   11
7  Cristiano Ronaldo    1    2
8  Cristiano Ronaldo    2    8
9  Cristiano Ronaldo    3    4
10              Xavi    0   25
11    Falcao Radamel    0   22
12    Falcao Radamel    2    2
13    Falcao Radamel    3    1
14  van Persie Robin    0   25
15       Alonso Xabi    0   24
16       Alonso Xabi    2    1
17      Casillas Iker    0   23
18      Casillas Iker    2    1
19      Casillas Iker    3    1
20        Ozil Mesut    0   25
21     Drogba Didier    0   22
22     Drogba Didier    1    2
23     Drogba Didier    2    1
24      Rooney Wayne    0   24
25      Rooney Wayne    2    1
26   Balotelli Mario    0   25
27      Pirlo Andrea    0   24
28      Pirlo Andrea    2    1
29        Toure Yaya    0   24
30        Toure Yaya    3    1
31     Aguero Sergio    0   25
32 Ibrahimovic Zlatan    1    5
33 Ibrahimovic Zlatan    2    6
34 Ibrahimovic Zlatan    3   14
35            Neymar    0   25
36    Busquets Sergio    0   25
37      Neuer Manuel    0   25
38   Buffon Gianluigi    0   24
39   Buffon Gianluigi    1    1
40      Pique Gerard    0   24
41      Pique Gerard    3    1
42      Ramos Sergio    0   25
43     Benzema Karim    0   25
```

Table A.34 : The amount of votes that each player received in Cluster 3, Year 2012.



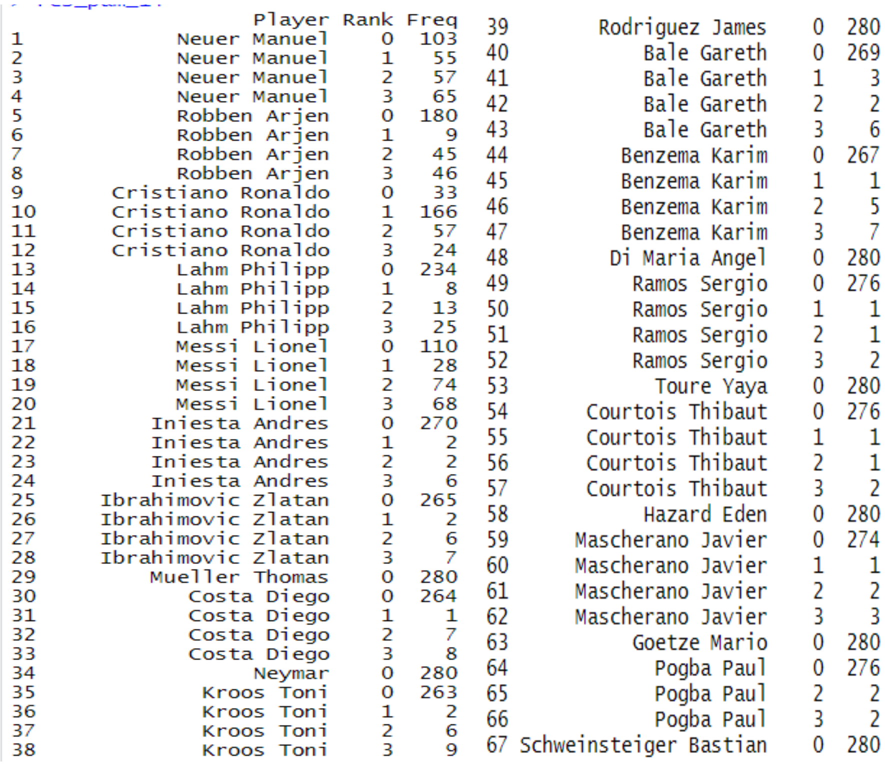
```
                Player Rank Freq
1    Cristiano Ronaldo    0   99
2    Cristiano Ronaldo    1  145
3    Cristiano Ronaldo    2  127
4    Cristiano Ronaldo    3   97
5        Ribery Franck    0  168
6        Ribery Franck    1  153
7        Ribery Franck    2   73
8        Ribery Franck    3   74
9       Falcao Radamel    0  444
10      Falcao Radamel    1    4
11      Falcao Radamel    2    5
12      Falcao Radamel    3   15
13  Ibrahimovic Zlatan    0  355
14  Ibrahimovic Zlatan    1   20
15  Ibrahimovic Zlatan    2   28
16  Ibrahimovic Zlatan    3   65
17          Toure Yaya    0  448
18          Toure Yaya    1    5
19          Toure Yaya    2    3
20          Toure Yaya    3   12
21  Lewandowski Robert    0  447
22  Lewandowski Robert    1    3
23  Lewandowski Robert    2    5
24  Lewandowski Robert    3   13
25        Messi Lionel    0  132
26        Messi Lionel    1   97
27        Messi Lionel    2  158
28        Messi Lionel    3   81
29              Neymar    0  411
30              Neymar    1   11
31              Neymar    2   20
32              Neymar    3   26
33        Robben Arjen    0  436
34        Robben Arjen    1    6
35        Robben Arjen    2   11
36        Robben Arjen    3   15
37      Cavani Edinson    0  459
38      Cavani Edinson    1    1
39      Cavani Edinson    2    2
40      Cavani Edinson    3    6
41             Hazard Eden    0  462
42             Hazard Eden    2    1
43             Hazard Eden    3    5
44            Silva Thiago    0  464
45            Silva Thiago    1    2
46            Silva Thiago    3    2
47              Ozil Mesut    0  453
48              Ozil Mesut    1    3
49              Ozil Mesut    2    4
50              Ozil Mesut    3    8
51           Muller Thomas    0  460
52           Muller Thomas    1    1
53           Muller Thomas    2    4
54           Muller Thomas    3    3
55           Lahm Philipp    0  451
56           Lahm Philipp    1    2
57           Lahm Philipp    2    7
58           Lahm Philipp    3    8
59                    Xavi    0  455
60                    Xavi    1    2
61                    Xavi    2    2
62                    Xavi    3    9
63             Suarez Luis    0  460
64             Suarez Luis    1    2
65             Suarez Luis    2    1
66             Suarez Luis    3    5
67             Bale Gareth    0  446
68             Bale Gareth    1    5
69             Bale Gareth    2    7
70             Bale Gareth    3   10
71        Van Persie Robin    0  468
72          Iniesta Andres    0  468
73            Pirlo Andrea    0  450
74            Pirlo Andrea    1    4
75            Pirlo Andrea    2    8
76            Pirlo Andrea    3    6
77 Schweinsteiger Bastian    0  457
78 Schweinsteiger Bastian    1    2
79 Schweinsteiger Bastian    2    1
80 Schweinsteiger Bastian    3    8
81            Neuer Manuel    0  467
82            Neuer Manuel    2    1
```

Table A.35 : The amount of votes that each player received in Cluster 1, Year 2013.



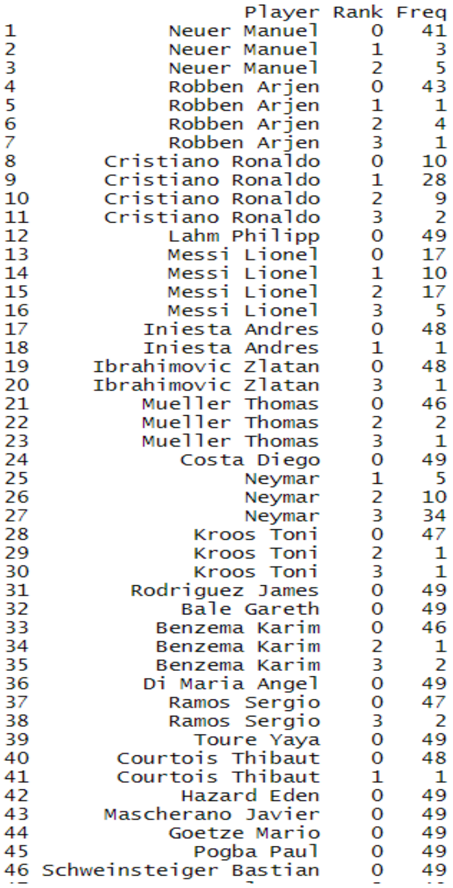
```
                    Player Rank Freq
1        Cristiano Ronaldo    0   16
2        Cristiano Ronaldo    1    9
3        Cristiano Ronaldo    2    9
4        Cristiano Ronaldo    3    3
5            Ribery Franck    0   28
6            Ribery Franck    1    2
7            Ribery Franck    2    4
8            Ribery Franck    3    3
9           Falcao Radamel    0   36
10          Falcao Radamel    2    1
11      Ibrahimovic Zlatan    0   34
12      Ibrahimovic Zlatan    2    1
13      Ibrahimovic Zlatan    3    2
14              Toure Yaya    0   36
15              Toure Yaya    3    1
16      Lewandowski Robert    0   37
17            Messi Lionel    0   16
18            Messi Lionel    1   13
19            Messi Lionel    2    5
20            Messi Lionel    3    3
21                  Neymar    0   32
22                  Neymar    2    2
23                  Neymar    3    3
24            Robben Arjen    0   36 38          Suarez Luis    0   37
25            Robben Arjen    2    1 39          Bale Gareth    0   36
26          Cavani Edinson    0   36 40          Bale Gareth    2    1
27          Cavani Edinson    3    1 41     Van Persie Robin    0   35
28             Hazard Eden    0   37 42     Van Persie Robin    2    1
29            Silva Thiago    0   37 43     Van Persie Robin    3    1
30              Ozil Mesut    0   37 44       Iniesta Andres    1   12
31           Muller Thomas    0   37 45       Iniesta Andres    2    9
32            Lahm Philipp    0   36 46       Iniesta Andres    3   16
33            Lahm Philipp    3    1 47         Pirlo Andrea    0   35
34                    Xavi    0   32 48         Pirlo Andrea    2    1
35                    Xavi    1    1 49         Pirlo Andrea    3    1
36                    Xavi    2    2 50 Schweinsteiger Bastian    0   37
37                    Xavi    3    2 51         Neuer Manuel    0   37
```

Table A.36 : The amount of votes that each player received in Cluster 3, Year 2013.



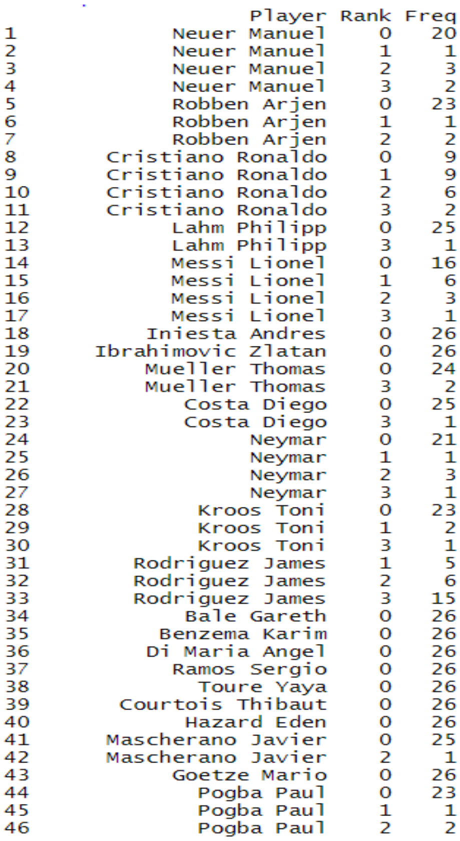
```
                    Player Rank Freq
1             Neuer Manuel    0  103
2             Neuer Manuel    1   55
3             Neuer Manuel    2   57
4             Neuer Manuel    3   65
5             Robben Arjen    0  180
6             Robben Arjen    1    9
7             Robben Arjen    2   45
8             Robben Arjen    3   46
9        Cristiano Ronaldo    0   33
10       Cristiano Ronaldo    1  166
11       Cristiano Ronaldo    2   57
12       Cristiano Ronaldo    3   24
13            Lahm Philipp    0  234
14            Lahm Philipp    1    8
15            Lahm Philipp    2   13
16            Lahm Philipp    3   25
17            Messi Lionel    0  110
18            Messi Lionel    1   28
19            Messi Lionel    2   74
20            Messi Lionel    3   68
21          Iniesta Andres    0  270
22          Iniesta Andres    1    2
23          Iniesta Andres    2    2
24          Iniesta Andres    3    6
25      Ibrahimovic Zlatan    0  265
26      Ibrahimovic Zlatan    1    2
27      Ibrahimovic Zlatan    2    6
28      Ibrahimovic Zlatan    3    7
29          Mueller Thomas    0  280
30             Costa Diego    0  264
31             Costa Diego    1    1
32             Costa Diego    2    7
33             Costa Diego    3    8
34                  Neymar    0  280
35               Kroos Toni    0  263
36               Kroos Toni    1    2
37               Kroos Toni    2    6
38               Kroos Toni    3    9

39        Rodriguez James    0  280
40            Bale Gareth    0  269
41            Bale Gareth    1    3
42            Bale Gareth    2    2
43            Bale Gareth    3    6
44         Benzema Karim    0  267
45         Benzema Karim    1    1
46         Benzema Karim    2    5
47         Benzema Karim    3    7
48          Di Maria Angel    0  280
49            Ramos Sergio    0  276
50            Ramos Sergio    1    1
51            Ramos Sergio    2    1
52            Ramos Sergio    3    2
53              Toure Yaya    0  280
54        Courtois Thibaut    0  276
55        Courtois Thibaut    1    1
56        Courtois Thibaut    2    1
57        Courtois Thibaut    3    2
58             Hazard Eden    0  280
59       Mascherano Javier    0  274
60       Mascherano Javier    1    1
61       Mascherano Javier    2    2
62       Mascherano Javier    3    3
63            Goetze Mario    0  280
64              Pogba Paul    0  276
65              Pogba Paul    2    2
66              Pogba Paul    3    2
67 Schweinsteiger Bastian    0  280
```

Table A.37 : The amount of votes that each player received in Cluster 1, Year 2014.



```
                  Player Rank Freq
1           Neuer Manuel    0   41
2           Neuer Manuel    1    3
3           Neuer Manuel    2    5
4           Robben Arjen    0   43
5           Robben Arjen    1    1
6           Robben Arjen    2    4
7           Robben Arjen    3    1
8      Cristiano Ronaldo    0   10
9      Cristiano Ronaldo    1   28
10     Cristiano Ronaldo    2    9
11     Cristiano Ronaldo    3    2
12          Lahm Philipp    0   49
13          Messi Lionel    0   17
14          Messi Lionel    1   10
15          Messi Lionel    2   17
16          Messi Lionel    3    5
17        Iniesta Andres    0   48
18        Iniesta Andres    1    1
19    Ibrahimovic Zlatan    0   48
20    Ibrahimovic Zlatan    3    1
21        Mueller Thomas    0   46
22        Mueller Thomas    2    2
23        Mueller Thomas    3    1
24           Costa Diego    0   49
25                Neymar    1    5
26                Neymar    2   10
27                Neymar    3   34
28            Kroos Toni    0   47
29            Kroos Toni    2    1
30            Kroos Toni    3    1
31       Rodriguez James    0   49
32           Bale Gareth    0   49
33         Benzema Karim    0   46
34         Benzema Karim    2    1
35         Benzema Karim    3    2
36        Di Maria Angel    0   49
37          Ramos Sergio    0   47
38          Ramos Sergio    3    2
39            Toure Yaya    0   49
40      Courtois Thibaut    0   48
41      Courtois Thibaut    1    1
42           Hazard Eden    0   49
43     Mascherano Javier    0   49
44          Goetze Mario    0   49
45            Pogba Paul    0   49
46 Schweinsteiger Bastian    0   49
```

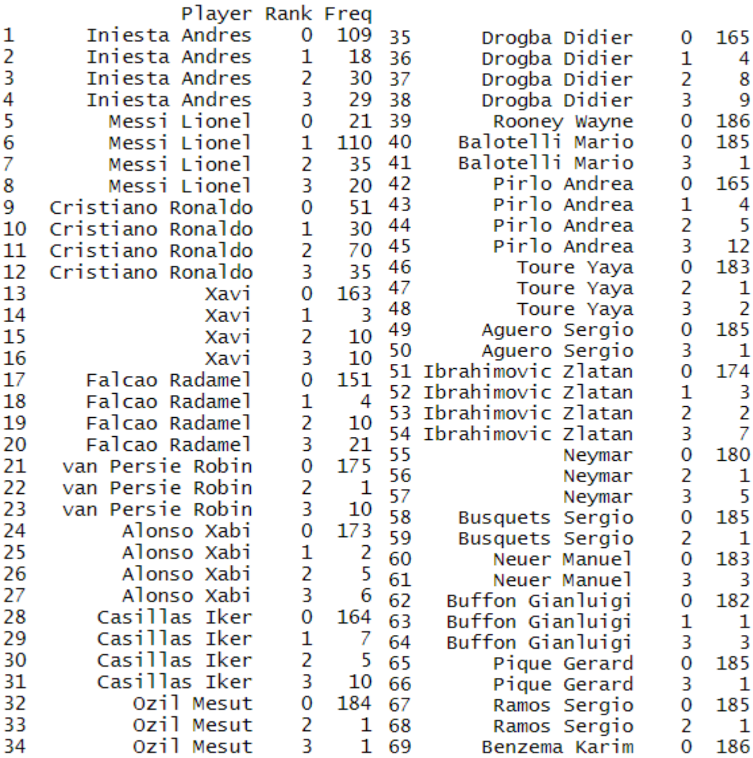
Table A.38 : The amount of votes that each player received in Cluster 3, Year 2014.



```
                 Player Rank Freq
1          Neuer Manuel    0   20
2          Neuer Manuel    1    1
3          Neuer Manuel    2    3
4          Neuer Manuel    3    2
5          Robben Arjen    0   23
6          Robben Arjen    1    1
7          Robben Arjen    2    2
8     Cristiano Ronaldo    0    9
9     Cristiano Ronaldo    1    9
10    Cristiano Ronaldo    2    6
11    Cristiano Ronaldo    3    2
12         Lahm Philipp    0   25
13         Lahm Philipp    3    1
14         Messi Lionel    0   16
15         Messi Lionel    1    6
16         Messi Lionel    2    3
17         Messi Lionel    3    1
18       Iniesta Andres    0   26
19   Ibrahimovic Zlatan    0   26
20       Mueller Thomas    0   24
21       Mueller Thomas    3    2
22          Costa Diego    0   25
23          Costa Diego    3    1
24               Neymar    0   21
25               Neymar    1    1
26               Neymar    2    3
27               Neymar    3    1
28           Kroos Toni    0   23
29           Kroos Toni    1    2
30           Kroos Toni    3    1
31      Rodriguez James    1    5
32      Rodriguez James    2    6
33      Rodriguez James    3   15
34          Bale Gareth    0   26
35        Benzema Karim    0   26
36       Di Maria Angel    0   26
37         Ramos Sergio    0   26
38           Toure Yaya    0   26
39     Courtois Thibaut    0   26
40          Hazard Eden    0   26
41    Mascherano Javier    0   25
42    Mascherano Javier    2    1
43         Goetze Mario    0   26
44           Pogba Paul    0   23
45           Pogba Paul    1    1
46           Pogba Paul    2    2
```

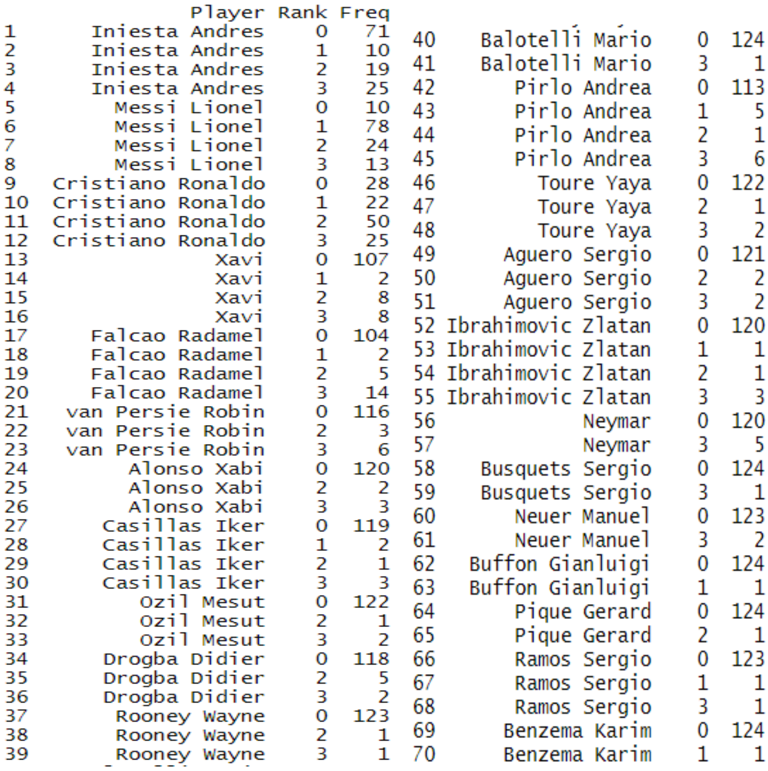
Table A.39 : The amount of votes that each player received in Cluster 5, Year 2014.

Chapter 7



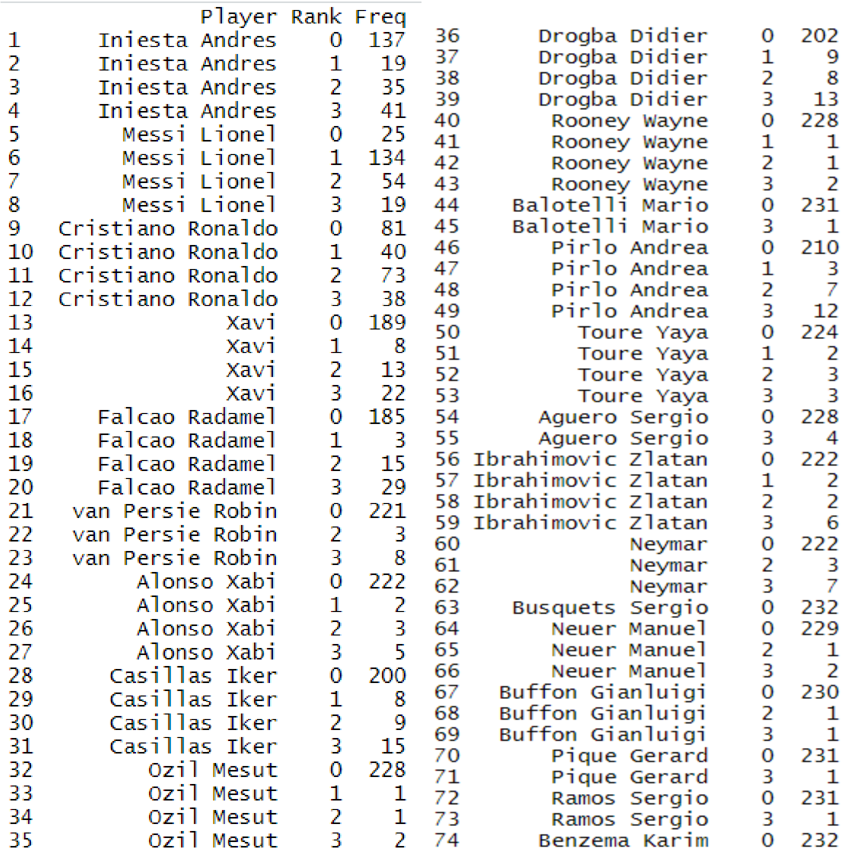
```
              Player Rank Freq
1      Iniesta Andres    0  109 35      Drogba Didier     0  165
2      Iniesta Andres    1   18 36      Drogba Didier     1    4
3      Iniesta Andres    2   30 37      Drogba Didier     2    8
4      Iniesta Andres    3   29 38      Drogba Didier     3    9
5        Messi Lionel    0   21 39       Rooney Wayne     0  186
6        Messi Lionel    1  110 40    Balotelli Mario     0  185
7        Messi Lionel    2   35 41    Balotelli Mario     3    1
8        Messi Lionel    3   20 42       Pirlo Andrea     0  165
9   Cristiano Ronaldo    0   51 43       Pirlo Andrea     1    4
10  Cristiano Ronaldo    1   30 44       Pirlo Andrea     2    5
11  Cristiano Ronaldo    2   70 45       Pirlo Andrea     3   12
12  Cristiano Ronaldo    3   35 46         Toure Yaya     0  183
13               Xavi    0  163 47         Toure Yaya     2    1
14               Xavi    1    3 48         Toure Yaya     3    2
15               Xavi    2   10 49      Aguero Sergio     0  185
16               Xavi    3   10 50      Aguero Sergio     3    1
17     Falcao Radamel    0  151 51 Ibrahimovic Zlatan     0  174
18     Falcao Radamel    1    4 52 Ibrahimovic Zlatan     1    3
19     Falcao Radamel    2   10 53 Ibrahimovic Zlatan     2    2
20     Falcao Radamel    3   21 54 Ibrahimovic Zlatan     3    7
21   van Persie Robin    0  175 55             Neymar     0  180
22   van Persie Robin    2    1 56             Neymar     2    1
23   van Persie Robin    3   10 57             Neymar     3    5
24        Alonso Xabi    0  173 58    Busquets Sergio     0  185
25        Alonso Xabi    1    2 59    Busquets Sergio     2    1
26        Alonso Xabi    2    5 60       Neuer Manuel     0  183
27        Alonso Xabi    3    6 61       Neuer Manuel     3    3
28      Casillas Iker    0  164 62   Buffon Gianluigi     0  182
29      Casillas Iker    1    7 63   Buffon Gianluigi     1    1
30      Casillas Iker    2    5 64   Buffon Gianluigi     3    3
31      Casillas Iker    3   10 65       Pique Gerard     0  185
32         Ozil Mesut    0  184 66       Pique Gerard     3    1
33         Ozil Mesut    2    1 67       Ramos Sergio     0  185
34         Ozil Mesut    3    1 68       Ramos Sergio     2    1
                                69      Benzema Karim     0  186
```

Table A.40 : The amount of votes that each player received in Cluster 1, Year 2010.



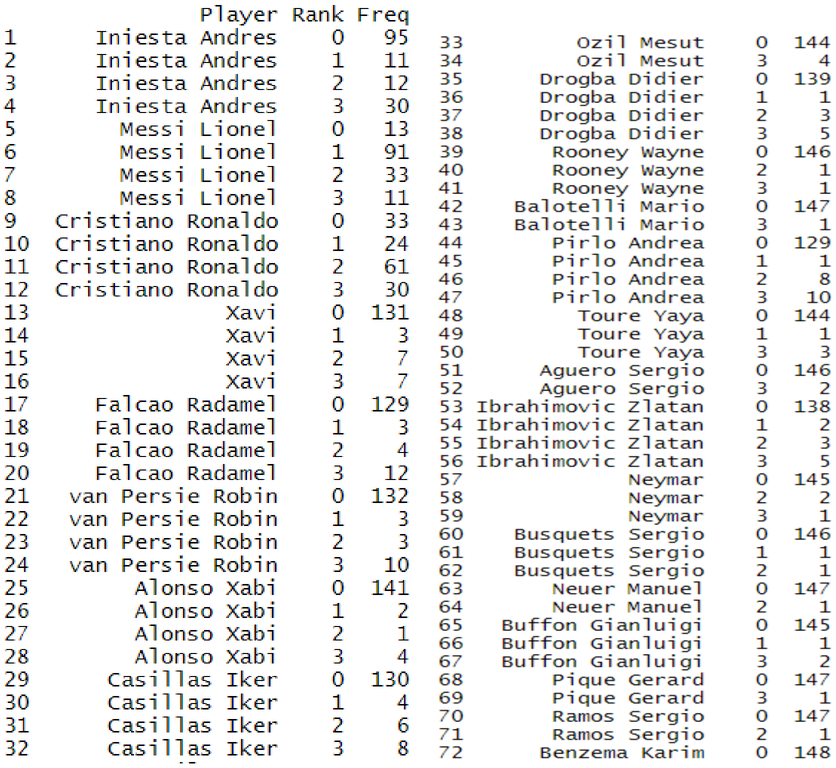
```
              Player Rank Freq
1      Iniesta Andres    0   71 40    Balotelli Mario     0  124
2      Iniesta Andres    1   10 41    Balotelli Mario     3    1
3      Iniesta Andres    2   19 42       Pirlo Andrea     0  113
4      Iniesta Andres    3   25 43       Pirlo Andrea     1    5
5        Messi Lionel    0   10 44       Pirlo Andrea     2    1
6        Messi Lionel    1   78 45       Pirlo Andrea     3    6
7        Messi Lionel    2   24 46         Toure Yaya     0  122
8        Messi Lionel    3   13 47         Toure Yaya     2    1
9   Cristiano Ronaldo    0   28 48         Toure Yaya     3    2
10  Cristiano Ronaldo    1   22 49      Aguero Sergio     0  121
11  Cristiano Ronaldo    2   50 50      Aguero Sergio     2    2
12  Cristiano Ronaldo    3   25 51      Aguero Sergio     3    2
13               Xavi    0  107 52 Ibrahimovic Zlatan     0  120
14               Xavi    1    2 53 Ibrahimovic Zlatan     1    1
15               Xavi    2    8 54 Ibrahimovic Zlatan     2    1
16               Xavi    3    8 55 Ibrahimovic Zlatan     3    3
17     Falcao Radamel    0  104 56             Neymar     0  120
18     Falcao Radamel    1    2 57             Neymar     3    5
19     Falcao Radamel    2    5 58    Busquets Sergio     0  124
20     Falcao Radamel    3   14 59    Busquets Sergio     3    1
21   van Persie Robin    0  116 60       Neuer Manuel     0  123
22   van Persie Robin    2    3 61       Neuer Manuel     3    2
23   van Persie Robin    3    6 62   Buffon Gianluigi     0  124
24        Alonso Xabi    0  120 63   Buffon Gianluigi     1    1
25        Alonso Xabi    2    2 64       Pique Gerard     0  124
26        Alonso Xabi    3    3 65       Pique Gerard     2    1
27      Casillas Iker    0  119 66       Ramos Sergio     0  123
28      Casillas Iker    1    2 67       Ramos Sergio     1    1
29      Casillas Iker    2    1 68       Ramos Sergio     3    1
30      Casillas Iker    3    3 69      Benzema Karim     0  124
31         Ozil Mesut    0  122 70      Benzema Karim     1    1
32         Ozil Mesut    2    1
33         Ozil Mesut    3    2
34      Drogba Didier    0  118
35      Drogba Didier    2    5
36      Drogba Didier    3    2
37       Rooney Wayne    0  123
38       Rooney Wayne    2    1
39       Rooney Wayne    3    1
```

Table A.41 : The amount of votes that each player received in Cluster 1, Year 2012.



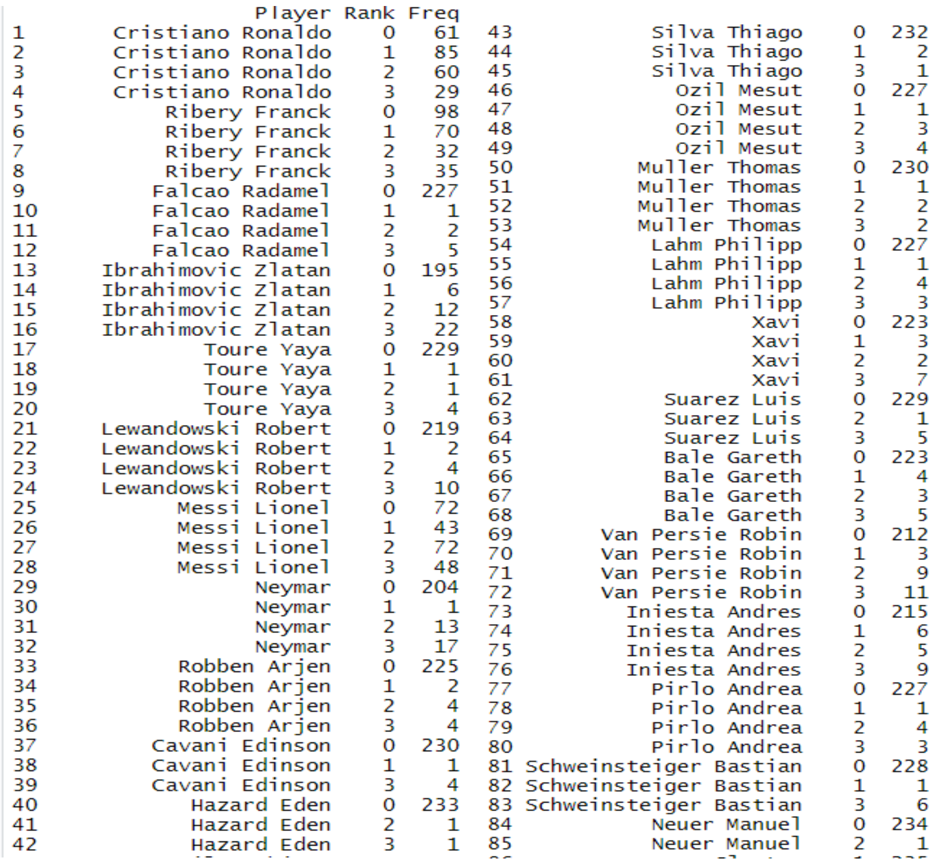
```
               Player Rank Freq
1      Iniesta Andres    0  137    36       Drogba Didier    0  202
2      Iniesta Andres    1   19    37       Drogba Didier    1    9
3      Iniesta Andres    2   35    38       Drogba Didier    2    8
4      Iniesta Andres    3   41    39       Drogba Didier    3   13
5        Messi Lionel    0   25    40        Rooney Wayne    0  228
6        Messi Lionel    1  134    41        Rooney Wayne    1    1
7        Messi Lionel    2   54    42        Rooney Wayne    2    1
8        Messi Lionel    3   19    43        Rooney Wayne    3    2
9   Cristiano Ronaldo    0   81    44     Balotelli Mario    0  231
10  Cristiano Ronaldo    1   40    45     Balotelli Mario    3    1
11  Cristiano Ronaldo    2   73    46        Pirlo Andrea    0  210
12  Cristiano Ronaldo    3   38    47        Pirlo Andrea    1    3
13               Xavi    0  189    48        Pirlo Andrea    2    7
14               Xavi    1    8    49        Pirlo Andrea    3   12
15               Xavi    2   13    50          Toure Yaya    0  224
16               Xavi    3   22    51          Toure Yaya    1    2
17     Falcao Radamel    0  185    52          Toure Yaya    2    3
18     Falcao Radamel    1    3    53          Toure Yaya    3    3
19     Falcao Radamel    2   15    54       Aguero Sergio    0  228
20     Falcao Radamel    3   29    55       Aguero Sergio    3    4
21   van Persie Robin    0  221    56 Ibrahimovic Zlatan    0  222
22   van Persie Robin    2    3    57 Ibrahimovic Zlatan    1    2
23   van Persie Robin    3    8    58 Ibrahimovic Zlatan    2    2
24        Alonso Xabi    0  222    59 Ibrahimovic Zlatan    3    6
25        Alonso Xabi    1    2    60              Neymar    0  222
26        Alonso Xabi    2    3    61              Neymar    2    3
27        Alonso Xabi    3    5    62              Neymar    3    7
28       Casillas Iker    0  200   63     Busquets Sergio    0  232
29       Casillas Iker    1    8   64        Neuer Manuel    0  229
30       Casillas Iker    2    9   65        Neuer Manuel    2    1
31       Casillas Iker    3   15   66        Neuer Manuel    3    2
32          Ozil Mesut    0  228   67    Buffon Gianluigi    0  230
33          Ozil Mesut    1    1   68    Buffon Gianluigi    2    1
34          Ozil Mesut    2    1   69    Buffon Gianluigi    3    1
35          Ozil Mesut    3    2   70        Pique Gerard    0  231
                                   71        Pique Gerard    3    1
                                   72        Ramos Sergio    0  231
                                   73        Ramos Sergio    3    1
                                   74       Benzema Karim    0  232
```

Table A.42 : The amount of votes that each player received in Cluster 2, Year 2012.



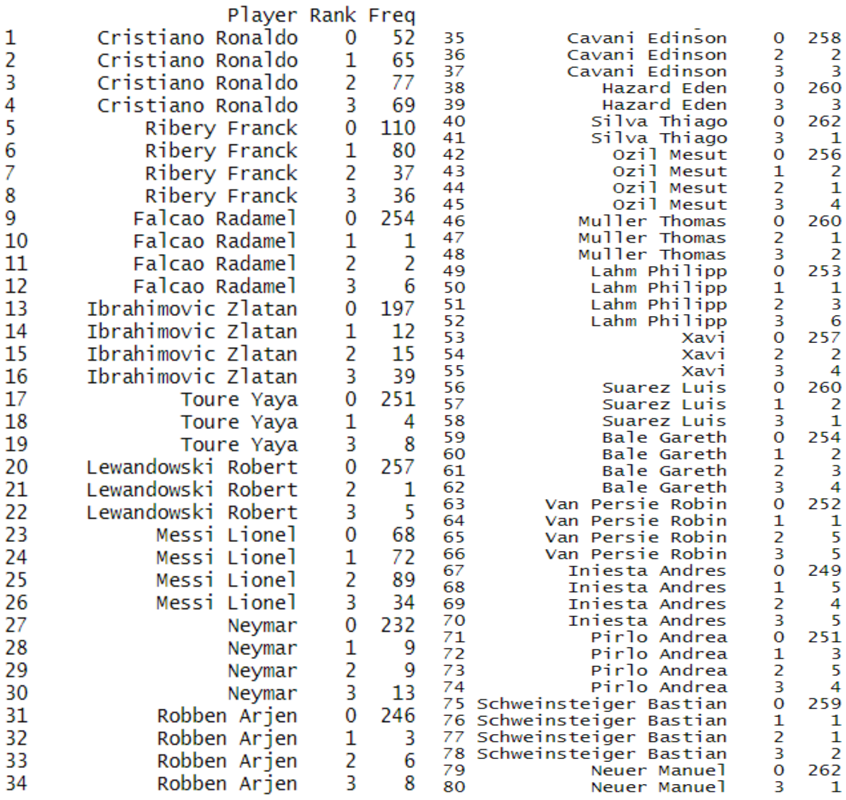
```
               Player Rank Freq
1      Iniesta Andres    0   95
2      Iniesta Andres    1   11    33          Ozil Mesut    0  144
3      Iniesta Andres    2   12    34          Ozil Mesut    3    4
4      Iniesta Andres    3   30    35       Drogba Didier    0  139
5        Messi Lionel    0   13    36       Drogba Didier    1    1
6        Messi Lionel    1   91    37       Drogba Didier    2    3
7        Messi Lionel    2   33    38       Drogba Didier    3    5
8        Messi Lionel    3   11    39        Rooney Wayne    0  146
9   Cristiano Ronaldo    0   33    40        Rooney Wayne    2    1
10  Cristiano Ronaldo    1   24    41        Rooney Wayne    3    1
11  Cristiano Ronaldo    2   61    42     Balotelli Mario    0  147
12  Cristiano Ronaldo    3   30    43     Balotelli Mario    3    1
13               Xavi    0  131    44        Pirlo Andrea    0  129
14               Xavi    1    3    45        Pirlo Andrea    1    1
15               Xavi    2    7    46        Pirlo Andrea    2    8
16               Xavi    3    7    47        Pirlo Andrea    3   10
17     Falcao Radamel    0  129    48          Toure Yaya    0  144
18     Falcao Radamel    1    3    49          Toure Yaya    1    1
19     Falcao Radamel    2    4    50          Toure Yaya    3    3
20     Falcao Radamel    3   12    51       Aguero Sergio    0  146
21   van Persie Robin    0  132    52       Aguero Sergio    3    2
22   van Persie Robin    1    3    53 Ibrahimovic Zlatan    0  138
23   van Persie Robin    2    3    54 Ibrahimovic Zlatan    1    2
24   van Persie Robin    3   10    55 Ibrahimovic Zlatan    2    3
25        Alonso Xabi    0  141    56 Ibrahimovic Zlatan    3    5
26        Alonso Xabi    1    2    57              Neymar    0  145
27        Alonso Xabi    2    1    58              Neymar    2    2
28        Alonso Xabi    3    4    59              Neymar    3    1
29      Casillas Iker    0  130    60     Busquets Sergio    0  146
30      Casillas Iker    1    4    61     Busquets Sergio    1    1
31      Casillas Iker    2    6    62     Busquets Sergio    2    1
32      Casillas Iker    3    8    63        Neuer Manuel    0  147
                                   64        Neuer Manuel    2    1
                                   65    Buffon Gianluigi    0  145
                                   66    Buffon Gianluigi    1    1
                                   67    Buffon Gianluigi    3    2
                                   68        Pique Gerard    0  147
                                   69        Pique Gerard    3    1
                                   70        Ramos Sergio    0  147
                                   71        Ramos Sergio    2    1
                                   72       Benzema Karim    0  148
```

Table A.43 : The amount of votes that each player received in Cluster 3, Year 2012.



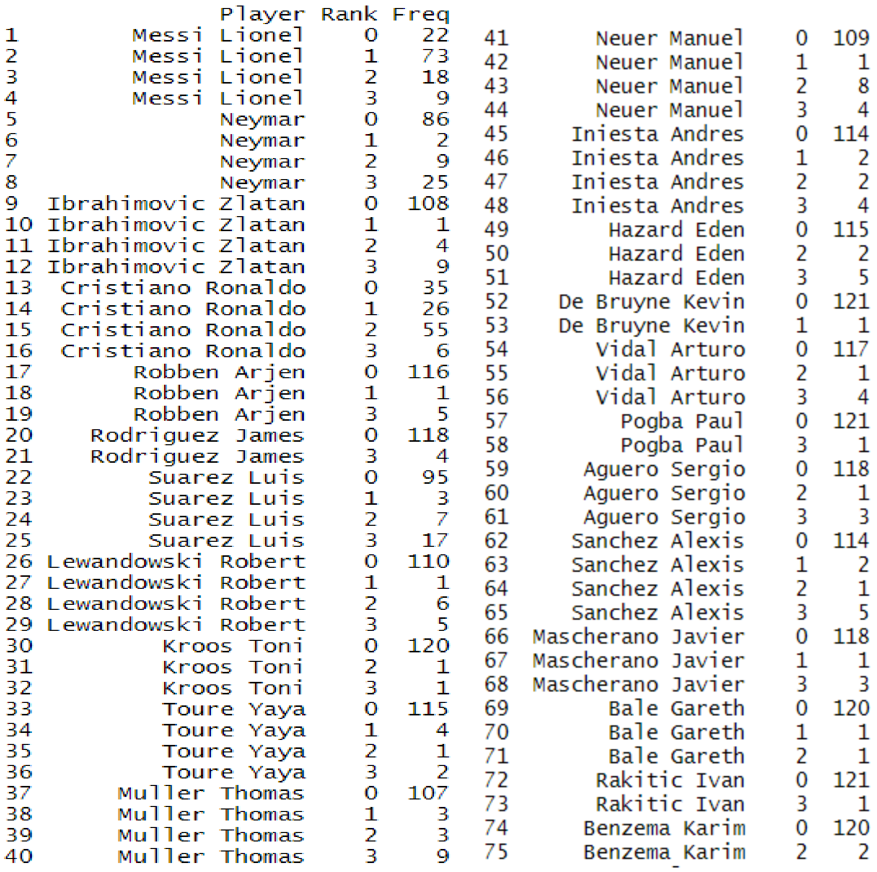
```
                    Player Rank Freq
1       Cristiano Ronaldo    0   61   43             Silva Thiago    0  232
2       Cristiano Ronaldo    1   85   44             Silva Thiago    1    2
3       Cristiano Ronaldo    2   60   45             Silva Thiago    3    1
4       Cristiano Ronaldo    3   29   46              Ozil Mesut    0  227
5           Ribery Franck    0   98   47              Ozil Mesut    1    1
6           Ribery Franck    1   70   48              Ozil Mesut    2    3
7           Ribery Franck    2   32   49              Ozil Mesut    3    4
8           Ribery Franck    3   35   50           Muller Thomas    0  230
9          Falcao Radamel    0  227   51           Muller Thomas    1    1
10         Falcao Radamel    1    1   52           Muller Thomas    2    2
11         Falcao Radamel    2    2   53           Muller Thomas    3    2
12         Falcao Radamel    3    5   54            Lahm Philipp    0  227
13     Ibrahimovic Zlatan    0  195   55            Lahm Philipp    1    1
14     Ibrahimovic Zlatan    1    6   56            Lahm Philipp    2    4
15     Ibrahimovic Zlatan    2   12   57            Lahm Philipp    3    3
16     Ibrahimovic Zlatan    3   22   58                    Xavi    0  223
17             Toure Yaya    0  229   59                    Xavi    1    3
18             Toure Yaya    1    1   60                    Xavi    2    2
19             Toure Yaya    2    1   61                    Xavi    3    7
20             Toure Yaya    3    4   62             Suarez Luis    0  229
21     Lewandowski Robert    0  219   63             Suarez Luis    2    1
22     Lewandowski Robert    1    2   64             Suarez Luis    3    5
23     Lewandowski Robert    2    4   65             Bale Gareth    0  223
24     Lewandowski Robert    3   10   66             Bale Gareth    1    4
25           Messi Lionel    0   72   67             Bale Gareth    2    3
26           Messi Lionel    1   43   68             Bale Gareth    3    5
27           Messi Lionel    2   72   69       Van Persie Robin    0  212
28           Messi Lionel    3   48   70       Van Persie Robin    1    3
29                 Neymar    0  204   71       Van Persie Robin    2    9
30                 Neymar    1    1   72       Van Persie Robin    3   11
31                 Neymar    2   13   73         Iniesta Andres    0  215
32                 Neymar    3   17   74         Iniesta Andres    1    6
33           Robben Arjen    0  225   75         Iniesta Andres    2    5
34           Robben Arjen    1    2   76         Iniesta Andres    3    9
35           Robben Arjen    2    4   77           Pirlo Andrea    0  227
36           Robben Arjen    3    4   78           Pirlo Andrea    1    1
37         Cavani Edinson    0  230   79           Pirlo Andrea    2    4
38         Cavani Edinson    1    1   80           Pirlo Andrea    3    3
39         Cavani Edinson    3    4   81 Schweinsteiger Bastian    0  228
40            Hazard Eden    0  233   82 Schweinsteiger Bastian    1    1
41            Hazard Eden    2    1   83 Schweinsteiger Bastian    3    6
42            Hazard Eden    3    1   84           Neuer Manuel    0  234
                                      85           Neuer Manuel    2    1
```

Table A.44 : The amount of votes that each player received in Cluster 1, Year 2013.



```
                    Player Rank Freq
1       Cristiano Ronaldo    0   52   35         Cavani Edinson    0  258
2       Cristiano Ronaldo    1   65   36         Cavani Edinson    2    2
3       Cristiano Ronaldo    2   77   37         Cavani Edinson    3    3
4       Cristiano Ronaldo    3   69   38            Hazard Eden    0  260
5           Ribery Franck    0  110   39            Hazard Eden    3    3
6           Ribery Franck    1   80   40           Silva Thiago    0  262
7           Ribery Franck    2   37   41           Silva Thiago    3    1
8           Ribery Franck    3   36   42             Ozil Mesut    0  256
9          Falcao Radamel    0  254   43             Ozil Mesut    1    2
10         Falcao Radamel    1    1   44             Ozil Mesut    2    1
11         Falcao Radamel    2    2   45             Ozil Mesut    3    4
12         Falcao Radamel    3    6   46          Muller Thomas    0  260
13     Ibrahimovic Zlatan    0  197   47          Muller Thomas    2    1
14     Ibrahimovic Zlatan    1   12   48          Muller Thomas    3    2
15     Ibrahimovic Zlatan    2   15   49           Lahm Philipp    0  253
16     Ibrahimovic Zlatan    3   39   50           Lahm Philipp    1    1
17             Toure Yaya    0  251   51           Lahm Philipp    2    3
18             Toure Yaya    1    4   52           Lahm Philipp    3    6
19             Toure Yaya    3    8   53                   Xavi    0  257
20     Lewandowski Robert    0  257   54                   Xavi    2    2
21     Lewandowski Robert    2    1   55                   Xavi    3    4
22     Lewandowski Robert    3    5   56            Suarez Luis    0  260
23           Messi Lionel    0   68   57            Suarez Luis    1    2
24           Messi Lionel    1   72   58            Suarez Luis    3    1
25           Messi Lionel    2   89   59            Bale Gareth    0  254
26           Messi Lionel    3   34   60            Bale Gareth    1    2
27                 Neymar    0  232   61            Bale Gareth    2    3
28                 Neymar    1    9   62            Bale Gareth    3    4
29                 Neymar    2    9   63       Van Persie Robin    0  252
30                 Neymar    3   13   64       Van Persie Robin    1    1
31           Robben Arjen    0  246   65       Van Persie Robin    2    5
32           Robben Arjen    1    3   66       Van Persie Robin    3    5
33           Robben Arjen    2    6   67         Iniesta Andres    0  249
34           Robben Arjen    3    8   68         Iniesta Andres    1    5
                                      69         Iniesta Andres    2    4
                                      70         Iniesta Andres    3    5
                                      71           Pirlo Andrea    0  251
                                      72           Pirlo Andrea    1    3
                                      73           Pirlo Andrea    2    5
                                      74           Pirlo Andrea    3    4
                                      75 Schweinsteiger Bastian    0  259
                                      76 Schweinsteiger Bastian    1    1
                                      77 Schweinsteiger Bastian    2    1
                                      78 Schweinsteiger Bastian    3    2
                                      79           Neuer Manuel    0  262
                                      80           Neuer Manuel    3    1
```

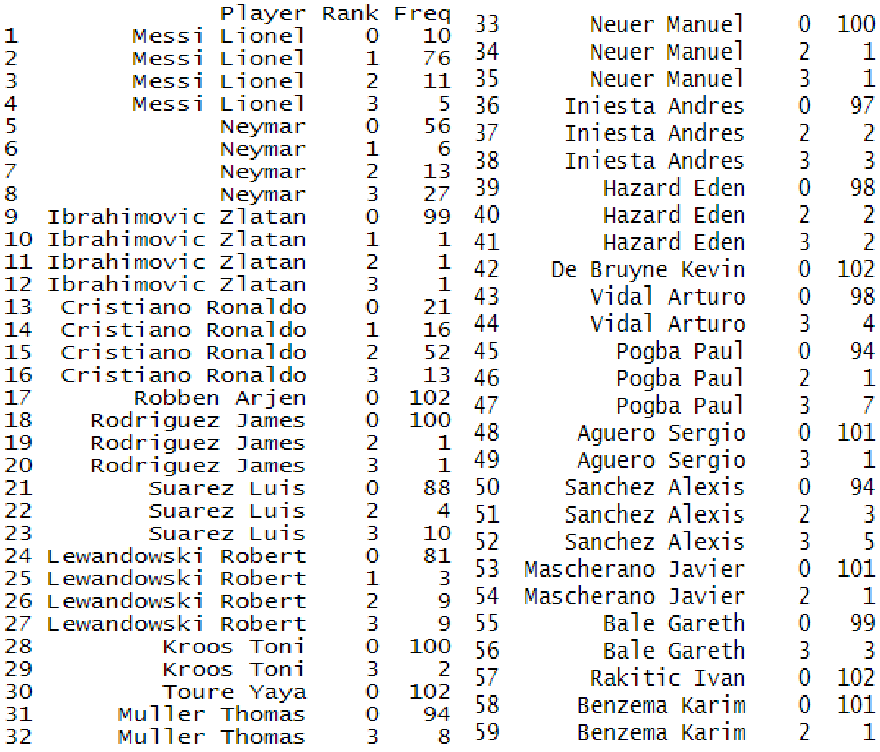
Table A.45 : The amount of votes that each player received in Cluster 3, Year 2013.



```
                 Player Rank Freq
1          Messi Lionel    0   22
2          Messi Lionel    1   73
3          Messi Lionel    2   18
4          Messi Lionel    3    9
5                Neymar    0   86
6                Neymar    1    2
7                Neymar    2    9
8                Neymar    3   25
9   Ibrahimovic Zlatan    0  108
10  Ibrahimovic Zlatan    1    1
11  Ibrahimovic Zlatan    2    4
12  Ibrahimovic Zlatan    3    9
13   Cristiano Ronaldo    0   35
14   Cristiano Ronaldo    1   26
15   Cristiano Ronaldo    2   55
16   Cristiano Ronaldo    3    6
17        Robben Arjen    0  116
18        Robben Arjen    1    1
19        Robben Arjen    3    5
20     Rodriguez James    0  118
21     Rodriguez James    3    4
22         Suarez Luis    0   95
23         Suarez Luis    1    3
24         Suarez Luis    2    7
25         Suarez Luis    3   17
26  Lewandowski Robert    0  110
27  Lewandowski Robert    1    1
28  Lewandowski Robert    2    6
29  Lewandowski Robert    3    5
30          Kroos Toni    0  120
31          Kroos Toni    2    1
32          Kroos Toni    3    1
33          Toure Yaya    0  115
34          Toure Yaya    1    4
35          Toure Yaya    2    1
36          Toure Yaya    3    2
37      Muller Thomas    0  107
38      Muller Thomas    1    3
39      Muller Thomas    2    3
40      Muller Thomas    3    9
```

```
41        Neuer Manuel    0  109
42        Neuer Manuel    1    1
43        Neuer Manuel    2    8
44        Neuer Manuel    3    4
45      Iniesta Andres    0  114
46      Iniesta Andres    1    2
47      Iniesta Andres    2    2
48      Iniesta Andres    3    4
49         Hazard Eden    0  115
50         Hazard Eden    2    2
51         Hazard Eden    3    5
52     De Bruyne Kevin    0  121
53     De Bruyne Kevin    1    1
54        Vidal Arturo    0  117
55        Vidal Arturo    2    1
56        Vidal Arturo    3    4
57          Pogba Paul    0  121
58          Pogba Paul    3    1
59       Aguero Sergio    0  118
60       Aguero Sergio    2    1
61       Aguero Sergio    3    3
62      Sanchez Alexis    0  114
63      Sanchez Alexis    1    2
64      Sanchez Alexis    2    1
65      Sanchez Alexis    3    5
66   Mascherano Javier    0  118
67   Mascherano Javier    1    1
68   Mascherano Javier    3    3
69         Bale Gareth    0  120
70         Bale Gareth    1    1
71         Bale Gareth    2    1
72        Rakitic Ivan    0  121
73        Rakitic Ivan    3    1
74       Benzema Karim    0  120
75       Benzema Karim    2    2
```

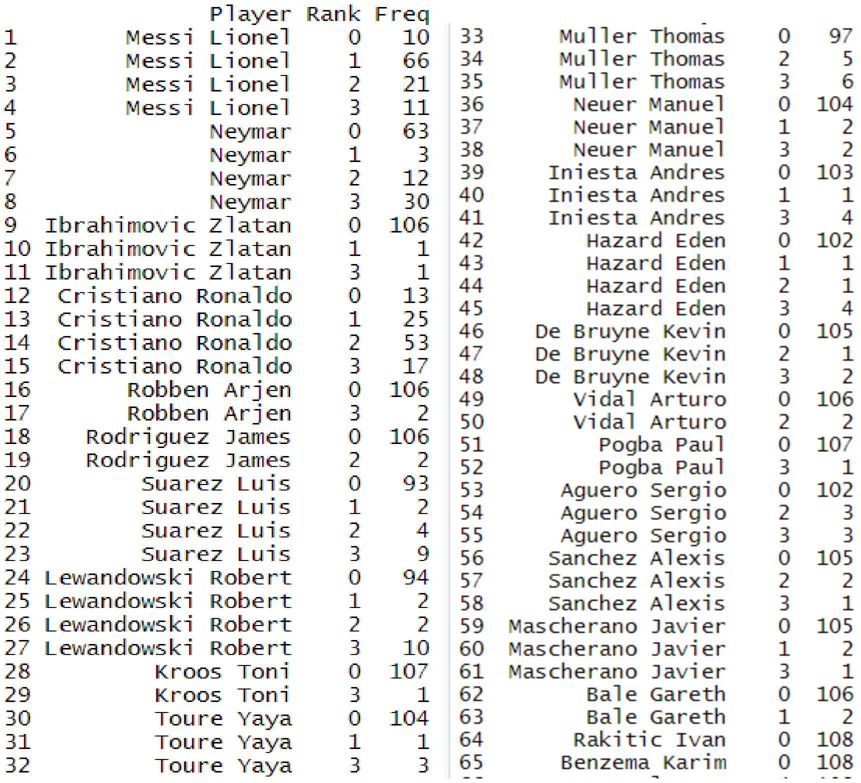
Table A.46 : The amount of votes that each player received in Cluster 1, Year 2015.



```
                 Player Rank Freq
1          Messi Lionel    0   10
2          Messi Lionel    1   76
3          Messi Lionel    2   11
4          Messi Lionel    3    5
5                Neymar    0   56
6                Neymar    1    6
7                Neymar    2   13
8                Neymar    3   27
9   Ibrahimovic Zlatan    0   99
10  Ibrahimovic Zlatan    1    1
11  Ibrahimovic Zlatan    2    1
12  Ibrahimovic Zlatan    3    1
13   Cristiano Ronaldo    0   21
14   Cristiano Ronaldo    1   16
15   Cristiano Ronaldo    2   52
16   Cristiano Ronaldo    3   13
17        Robben Arjen    0  102
18     Rodriguez James    0  100
19     Rodriguez James    2    1
20     Rodriguez James    3    1
21         Suarez Luis    0   88
22         Suarez Luis    2    4
23         Suarez Luis    3   10
24  Lewandowski Robert    0   81
25  Lewandowski Robert    1    3
26  Lewandowski Robert    2    9
27  Lewandowski Robert    3    9
28          Kroos Toni    0  100
29          Kroos Toni    3    2
30          Toure Yaya    0  102
31      Muller Thomas    0   94
32      Muller Thomas    3    8
```

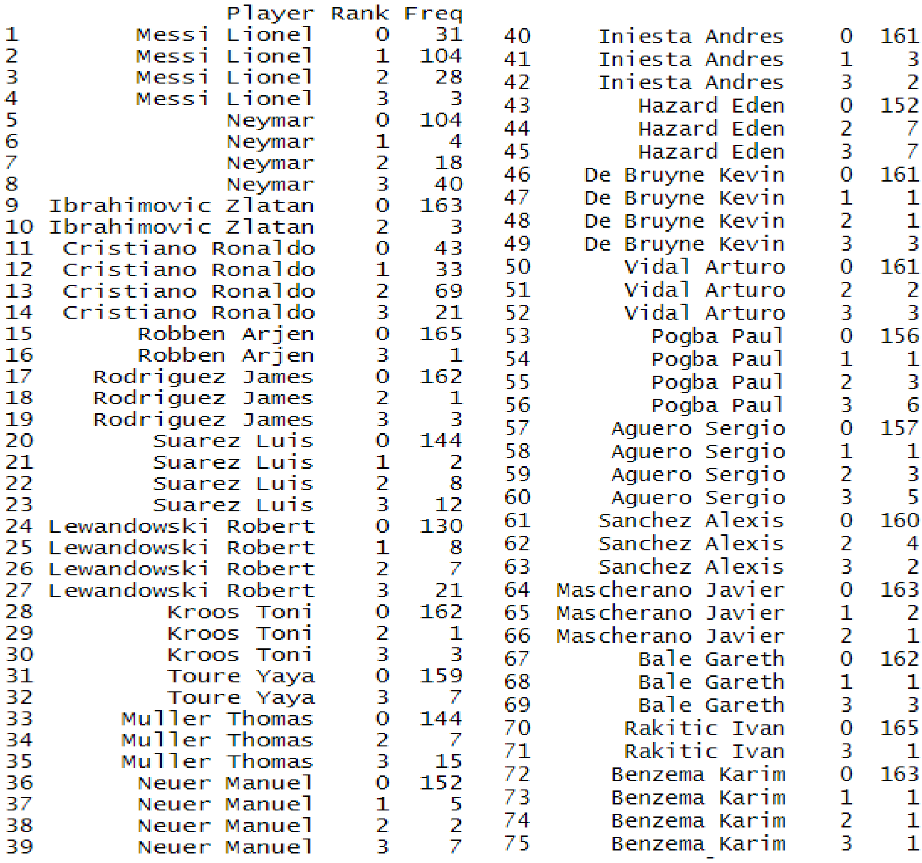
```
33        Neuer Manuel    0  100
34        Neuer Manuel    2    1
35        Neuer Manuel    3    1
36      Iniesta Andres    0   97
37      Iniesta Andres    2    2
38      Iniesta Andres    3    3
39         Hazard Eden    0   98
40         Hazard Eden    2    2
41         Hazard Eden    3    2
42     De Bruyne Kevin    0  102
43        Vidal Arturo    0   98
44        Vidal Arturo    3    4
45          Pogba Paul    0   94
46          Pogba Paul    2    1
47          Pogba Paul    3    7
48       Aguero Sergio    0  101
49       Aguero Sergio    3    1
50      Sanchez Alexis    0   94
51      Sanchez Alexis    2    3
52      Sanchez Alexis    3    5
53   Mascherano Javier    0  101
54   Mascherano Javier    2    1
55         Bale Gareth    0   99
56         Bale Gareth    3    3
57        Rakitic Ivan    0  102
58       Benzema Karim    0  101
59       Benzema Karim    2    1
```

Table A.47 : The amount of votes that each player received in Cluster 2, Year 2015.



```
              Player Rank Freq   33       Muller Thomas    0   97
1        Messi Lionel    0   10  34       Muller Thomas    2    5
2        Messi Lionel    1   66  35       Muller Thomas    3    6
3        Messi Lionel    2   21  36        Neuer Manuel    0  104
4        Messi Lionel    3   11  37        Neuer Manuel    1    2
5              Neymar    0   63  38        Neuer Manuel    3    2
6              Neymar    1    3  39      Iniesta Andres    0  103
7              Neymar    2   12  40      Iniesta Andres    1    1
8              Neymar    3   30  41      Iniesta Andres    3    4
9  Ibrahimovic Zlatan    0  106  42         Hazard Eden    0  102
10 Ibrahimovic Zlatan    1    1  43         Hazard Eden    1    1
11 Ibrahimovic Zlatan    3    1  44         Hazard Eden    2    1
12  Cristiano Ronaldo    0   13  45         Hazard Eden    3    4
13  Cristiano Ronaldo    1   25  46    De Bruyne Kevin    0  105
14  Cristiano Ronaldo    2   53  47    De Bruyne Kevin    2    1
15  Cristiano Ronaldo    3   17  48    De Bruyne Kevin    3    2
16       Robben Arjen    0  106  49       Vidal Arturo    0  106
17       Robben Arjen    3    2  50       Vidal Arturo    2    2
18    Rodriguez James    0  106  51         Pogba Paul    0  107
19    Rodriguez James    2    2  52         Pogba Paul    3    1
20        Suarez Luis    0   93  53      Aguero Sergio    0  102
21        Suarez Luis    1    2  54      Aguero Sergio    2    3
22        Suarez Luis    2    4  55      Aguero Sergio    3    3
23        Suarez Luis    3    9  56     Sanchez Alexis    0  105
24 Lewandowski Robert    0   94  57     Sanchez Alexis    2    2
25 Lewandowski Robert    1    2  58     Sanchez Alexis    3    1
26 Lewandowski Robert    2    2  59  Mascherano Javier    0  105
27 Lewandowski Robert    3   10  60  Mascherano Javier    1    2
28         Kroos Toni    0  107  61  Mascherano Javier    3    1
29         Kroos Toni    3    1  62        Bale Gareth    0  106
30         Toure Yaya    0  104  63        Bale Gareth    1    2
31         Toure Yaya    1    1  64        Rakitic Ivan    0  108
32         Toure Yaya    3    3  65      Benzema Karim    0  108
```

Table A.48 : The amount of votes that each player received in Cluster 4, Year 2015.



```
              Player Rank Freq   40      Iniesta Andres    0  161
1        Messi Lionel    0   31  41      Iniesta Andres    1    3
2        Messi Lionel    1  104  42      Iniesta Andres    3    2
3        Messi Lionel    2   28  43         Hazard Eden    0  152
4        Messi Lionel    3    3  44         Hazard Eden    2    7
5              Neymar    0  104  45         Hazard Eden    3    7
6              Neymar    1    4  46     De Bruyne Kevin    0  161
7              Neymar    2   18  47     De Bruyne Kevin    1    1
8              Neymar    3   40  48     De Bruyne Kevin    2    1
9  Ibrahimovic Zlatan    0  163  49     De Bruyne Kevin    3    3
10 Ibrahimovic Zlatan    2    3  50        Vidal Arturo    0  161
11  Cristiano Ronaldo    0   43  51        Vidal Arturo    2    2
12  Cristiano Ronaldo    1   33  52        Vidal Arturo    3    3
13  Cristiano Ronaldo    2   69  53          Pogba Paul    0  156
14  Cristiano Ronaldo    3   21  54          Pogba Paul    1    1
15       Robben Arjen    0  165  55          Pogba Paul    2    3
16       Robben Arjen    3    1  56          Pogba Paul    3    6
17    Rodriguez James    0  162  57       Aguero Sergio    0  157
18    Rodriguez James    2    1  58       Aguero Sergio    1    1
19    Rodriguez James    3    3  59       Aguero Sergio    2    3
20        Suarez Luis    0  144  60       Aguero Sergio    3    5
21        Suarez Luis    1    2  61      Sanchez Alexis    0  160
22        Suarez Luis    2    8  62      Sanchez Alexis    2    4
23        Suarez Luis    3   12  63      Sanchez Alexis    3    2
24 Lewandowski Robert    0  130  64   Mascherano Javier    0  163
25 Lewandowski Robert    1    8  65   Mascherano Javier    1    2
26 Lewandowski Robert    2    7  66   Mascherano Javier    2    1
27 Lewandowski Robert    3   21  67         Bale Gareth    0  162
28         Kroos Toni    0  162  68         Bale Gareth    1    1
29         Kroos Toni    2    1  69         Bale Gareth    3    3
30         Kroos Toni    3    3  70        Rakitic Ivan    0  165
31         Toure Yaya    0  159  71        Rakitic Ivan    3    1
32         Toure Yaya    3    7  72       Benzema Karim    0  163
33      Muller Thomas    0  144  73       Benzema Karim    1    1
34      Muller Thomas    2    7  74       Benzema Karim    2    1
35      Muller Thomas    3   15  75       Benzema Karim    3    1
36       Neuer Manuel    0  152
37       Neuer Manuel    1    5
38       Neuer Manuel    2    2
39       Neuer Manuel    3    7
```

Table A.49 : The amount of votes that each player received in Cluster 3, Year 2015.

Packages	**Description**	**Citation**
data.table	The package was first released on 01/02/2017. It provides fast aggregation on large datasets, fast ordered joins, fast modifications of columns by group. Also, it offers a natural and flexible syntax, for faster development.	Matt Dowle and Arun Srinivasan (2017). data.table: Extension of `data.frame`. R package version 1.10.4. https://CRAN.R-project.org/package=data.table
pmr	The package was first released on 14/05/2010. It provides descriptive statistics (mean rank, pairwise frequencies, marginal matrix), probability models (Luce models, distance – based models, rank – ordered logit models) and visualization with multidimensional preference analysis, for ranking data. Currently, only complete rankings are supported by this package.	Paul H. Lee and Philip L. H. Yu (2015). pmr: Probability Models for Ranking Data. R package version 1.2.5.https://CRAN.R-project.org/package=pmr
PLMIX	The package was first released on 21/12/2016. It provides functions to fit and analyze finite mixtures of Plackett – Luce models, for partial top rankings/orderings within the Bayesian framework. It provides MAP estimates via EM algorithm and posterior MCMC simulations via Gibbs sampling. It also fits MLE as a special case of the noninformative Bayesian analysis with negligible priors.	Mollica, C., Tardella, L. (2016). Bayesian Plackett-Luce mixture models for partially ranked data. Psychometrika
amap	The package was first released on 17/12/2014.	Antoine Lucas (2014). amap: Another Multidimensional

	It includes standard hierarchical clustering and k – means. It optimizes the implementation, with a parallelized hierarchical clustering, and allows the possibility of using different distances like Euclidean or Spearman (rank – based metric). It also offers the implementation of principal component analysis.	Analysis Package. R package version 0.8-14.https://CRAN.R-project.org/package=amap
goeveg	The package was first released on 24/01/2017. It can be described as a collection of functions useful in (vegetation) community analyses. It includes automatic species for ordination diagrams, species response curves and rank – abundance curves as well as calculation and sorting of synoptic tables.	Friedemann Goral and Jenny Schellenberg (2017). goeveg: Functions for Community Data and Ordinations. R package version 0.3.3. https://CRAN.R-project.org/package=goeveg
vegan	The package was first released on 17/01/2017. It provides tools for descriptive community ecology. It contains fundamental functions of diversity analysis, community ordination and dissimilarity analysis. Most of its multivariate tools can be used for other data types as well.	Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R.Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2017). vegan: Community Ecology Package. R package version 2.4-2. https://CRAN.R-project.org/package=vegan
vegan3d	The package was first released on 15/06/2016. It provides static and dynamic 3D plots to be used with ordination results and in diversity analysis, especially with the vegan package.	Jari Oksanen, Roeland Kindt and Gavin L. Simpson (2016). vegan3d: Static and Dynamic 3D Plots for the 'vegan' Package. R package version 1.0-1. https://CRAN.R-project.org/package=vegan3d

MASS	The package was first released on 10/11/2015. It contains functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S'.	Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN
rgl	The package was first released on 08/03/2017. It provides medium to high level functions for 3D interactive graphics, including functions modelled on base graphics (plot3d(), etc.), as well as functions for constructing representations of geometric objects (cube3d(), etc.). Output may be on screen using OpenGL, or to various standard 3D formats.	Daniel Adler, Duncan Murdoch and others (2017). rgl: 3D Visualization Using OpenGL. R package version 0.98.1. https://CRAN.R-project.org/package=rgl
scatterplot3d	The package was first released on 05/01/2017. It provides the ability of plotting a three dimensional (3D) point cloud perspectively (3D scatterplot).	Ligges, U. and Machler, M. (2003). Scatterplot3d - an R Package for Visualizing Multivariate Data. Journal of Statistical Software 8(11), 1-20.
Rankcluster	The package was first released on 21/07/2016. It provides the implementation of a model – based clustering algorithm for ranking data, where multivariate rankings and partial rankings are also taken into account. The algorithm is based on an extension of the Insertion Sorting Rank (ISR) model for ranking data.	Quentin Grimonprez and Julien Jacques (2016). Rankcluster: Model-Based Clustering for Multivariate Partial Ranking Data. R package version 0.94. https://CRAN.R-project.org/package=Rankcluster
countrycode	The package was first released on 06/02/2017. The package can convert country names and country codes. The	Vincent Arel-Bundock (2017). countrycode: Convert Country Names and Country Codes. R package version

	fundamental property is the one that standardizes country names, converts them into one of 40 different coding schemes and assign region descriptors. It uses regular expressions to convert country names into any of those coding schemes , or into standardized country names in several languages. It can create variables with the name of the continent and several regional groupings to which each country belongs.	0.19.https://CRAN.R-project.org/package=countrycode
cluster	The package was first released on 16/09/2016. It implements different methods for cluster analysis (Hierarchical methods, Fuzzy analysis, Clustering large applications, etc.).	Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2016).cluster: Cluster Analysis Basics and Extensions. R package version 2.0.5.
factoextra	The package was first released on 08/01/2017. It provides some easy – to – use functions to extract and visualize the output of multivariate data analyses, including many kind of analyses (e.g ‘PCA’), by combining functions from different R packages. It contains also functions for simplifying some clustering analyses steps and provides ‘ggplot2’ – based data visualization.	Alboukadel Kassambara and Fabian Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.4. https://CRAN.R-project.org/package=factoextra
sqldf	The package was first released on 31/10/2014. Its main purpose is to run SQL statements on R data frames. The user specifies an SQL	G. Grothendieck (2014). sqldf: Perform SQL Selects on R Data Frames. R package version 0.4-10.https://CRAN.R-project.org/package=sqldf

	statement in R using data frame names in place of table names, a database with appropriate table schema is automatically created, the data frames are automatically loaded into the database, the specified SQL statement is performed, the result is read back into R and the database is deleted all automatically, making the database's transparent to the user who only specifies the SQL statement.	

Table A.50 : The R packages that are mentioned in the main report of the Thesis.

APPENDIX B : FIGURES

Chapter 5

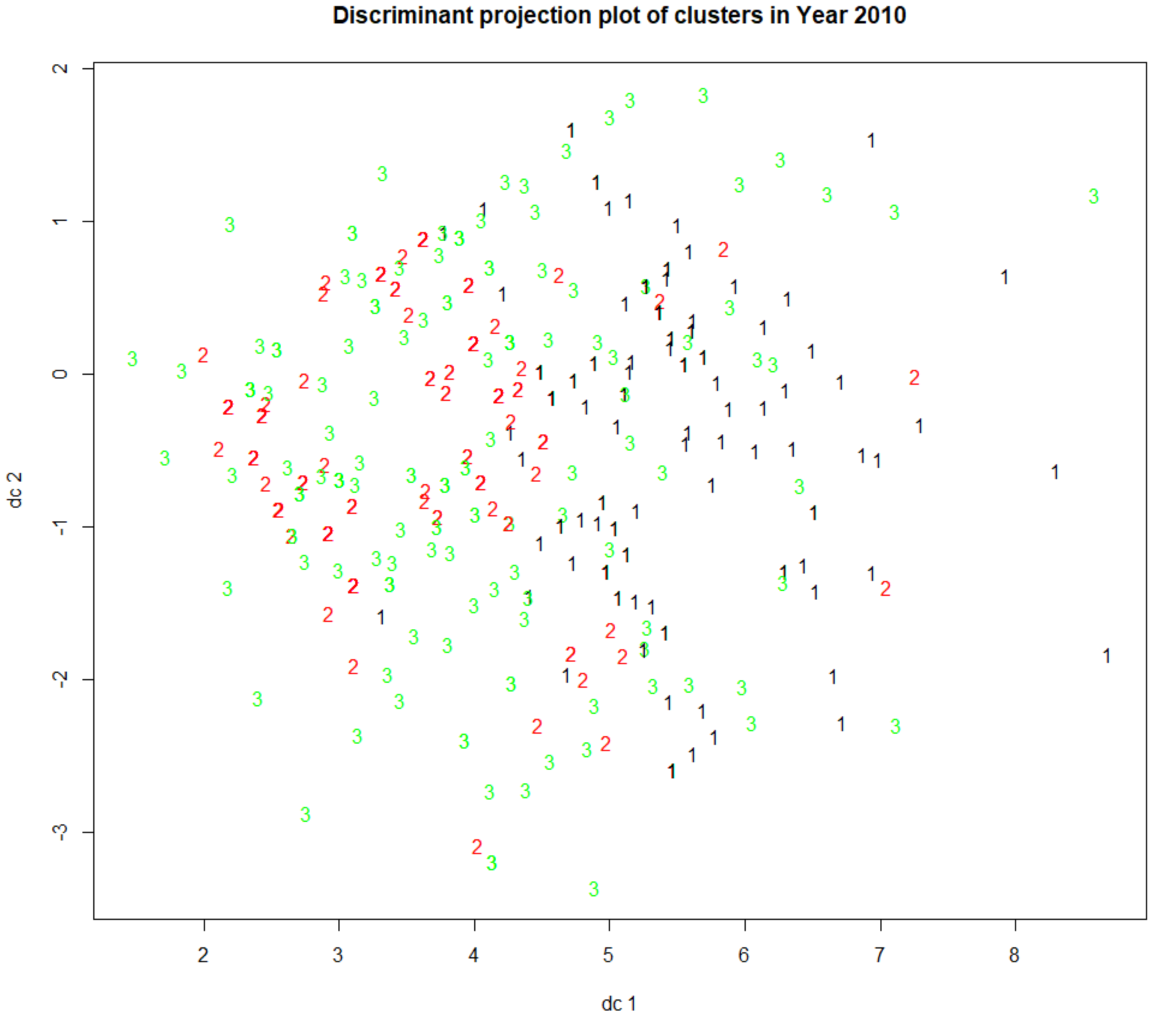


Figure B.1 : 2 – D representation of the observations distribution in each cluster, Year 2010.

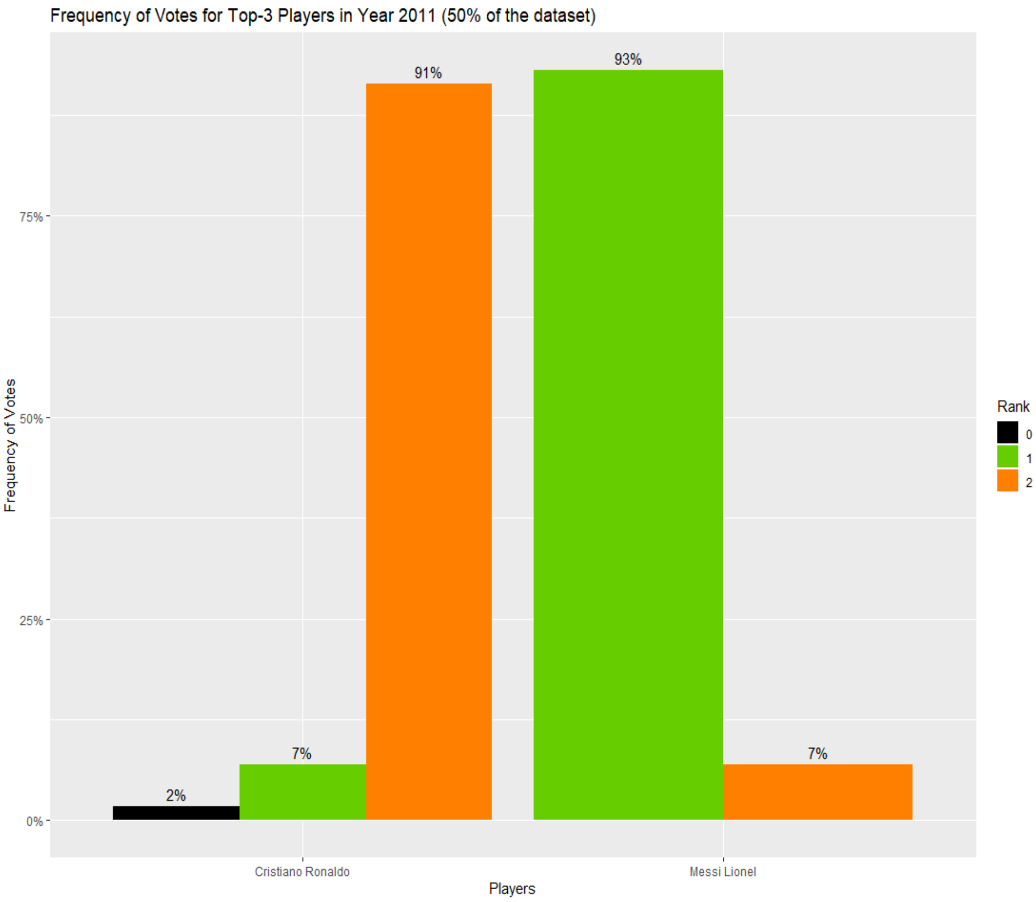


Figure B.2 : Bar plot, representing the percentage of places that Cristiano Ronaldo and Lionel Messi have been ranked in Cluster 1, Year 2011.

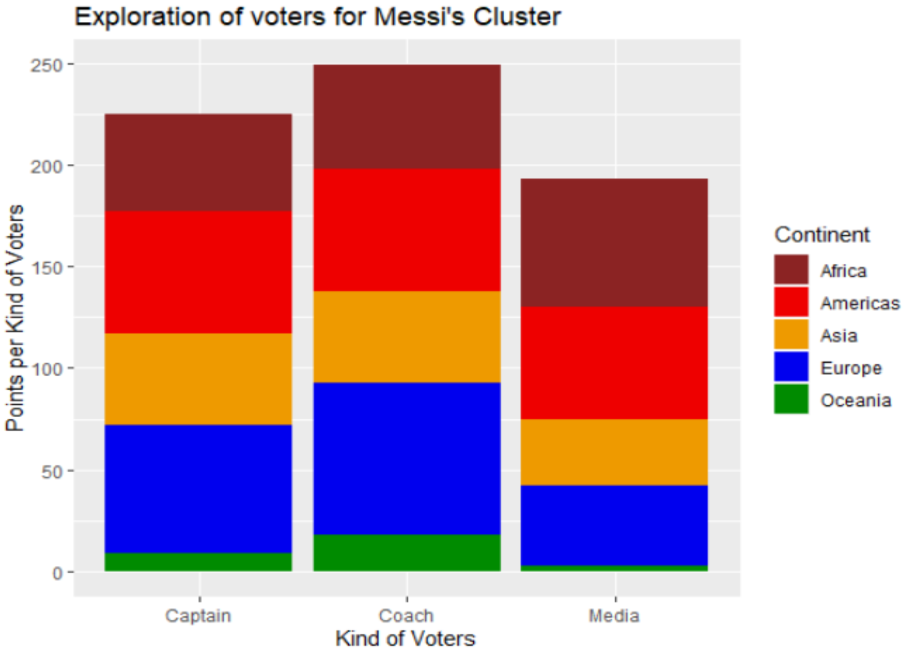


Figure B.3 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 2, Year 2012.

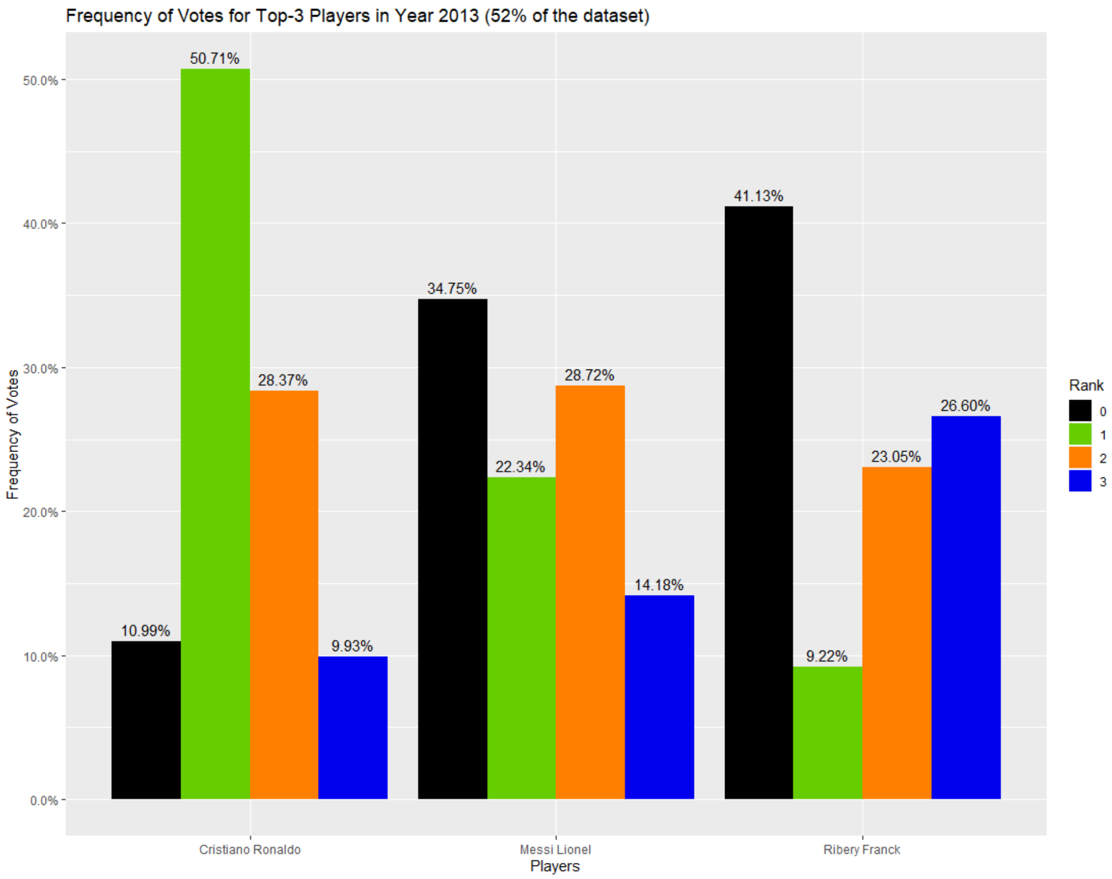


Figure B.4 : Bar plot, representing the percentage of places that Cristiano Ronaldo, Lionel Messi and Ribery Franck have been ranked in Cluster 2, Year 2013.

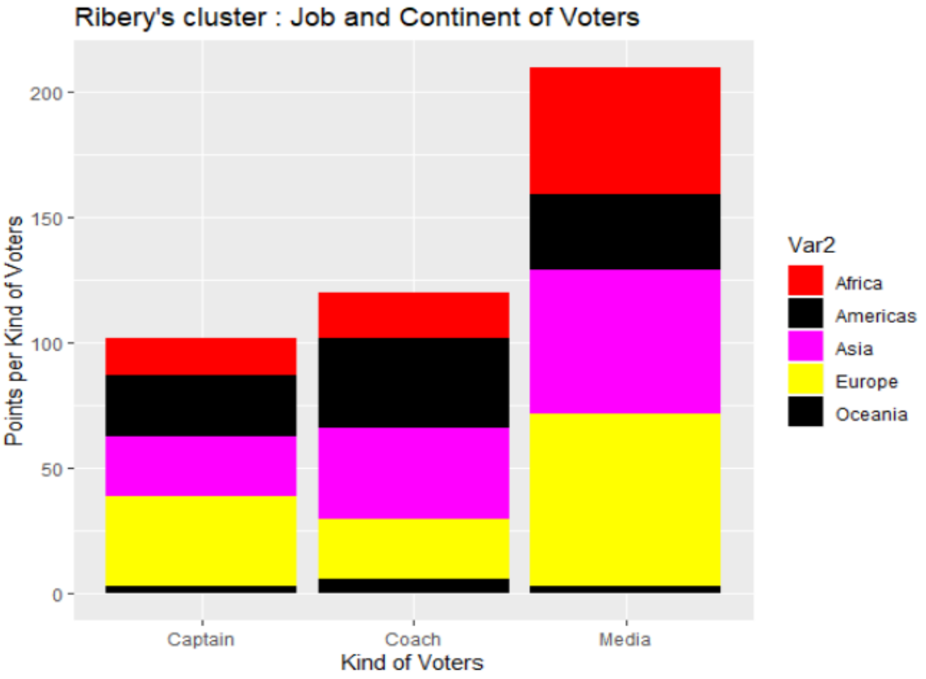


Figure B.5 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 3, Year 2013.

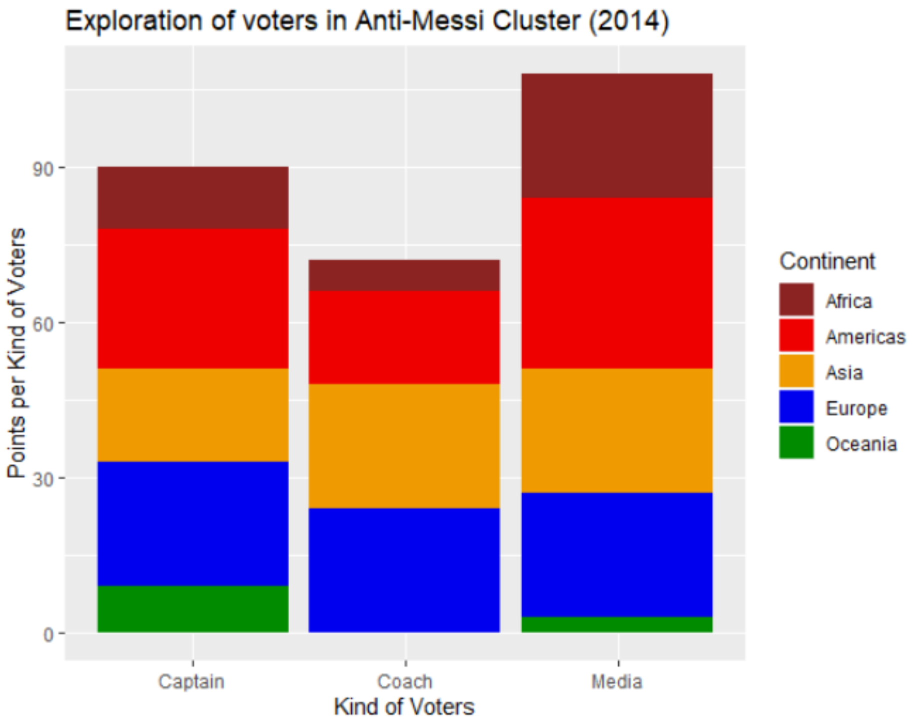


Figure B.6 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 1, Year 2014.

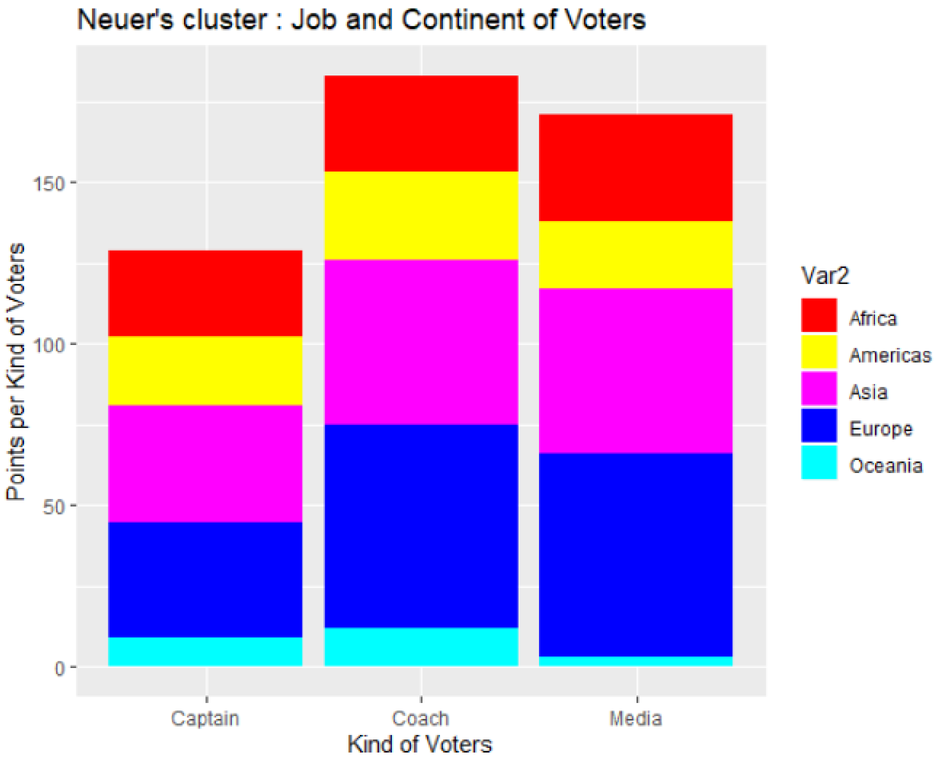


Figure B.7 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 2, Year 2014.

Chapter 6

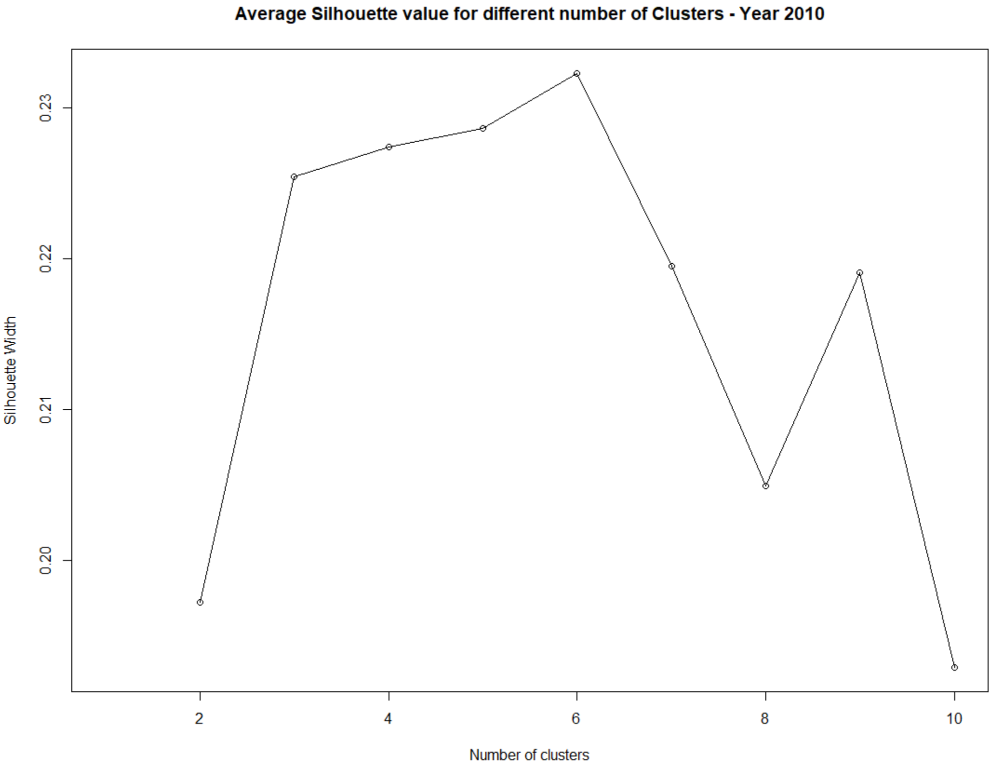


Figure B.8 : Average silhouette plot for the different number of components, Year 2010.

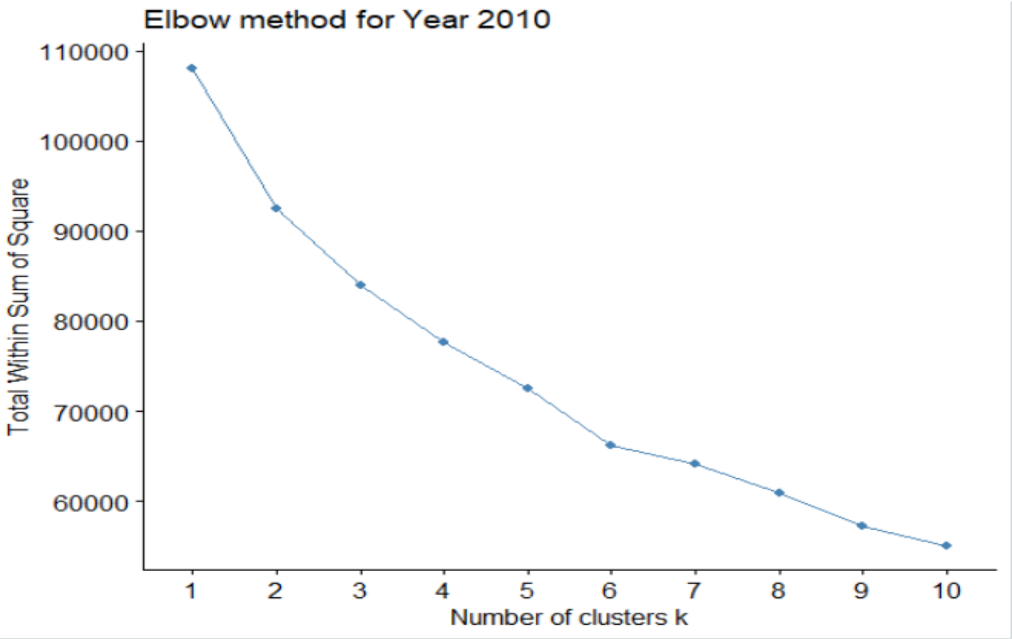


Figure B.9 : The Elbow method using the Total Within Sum of Square for the different number of components, Year 2010.

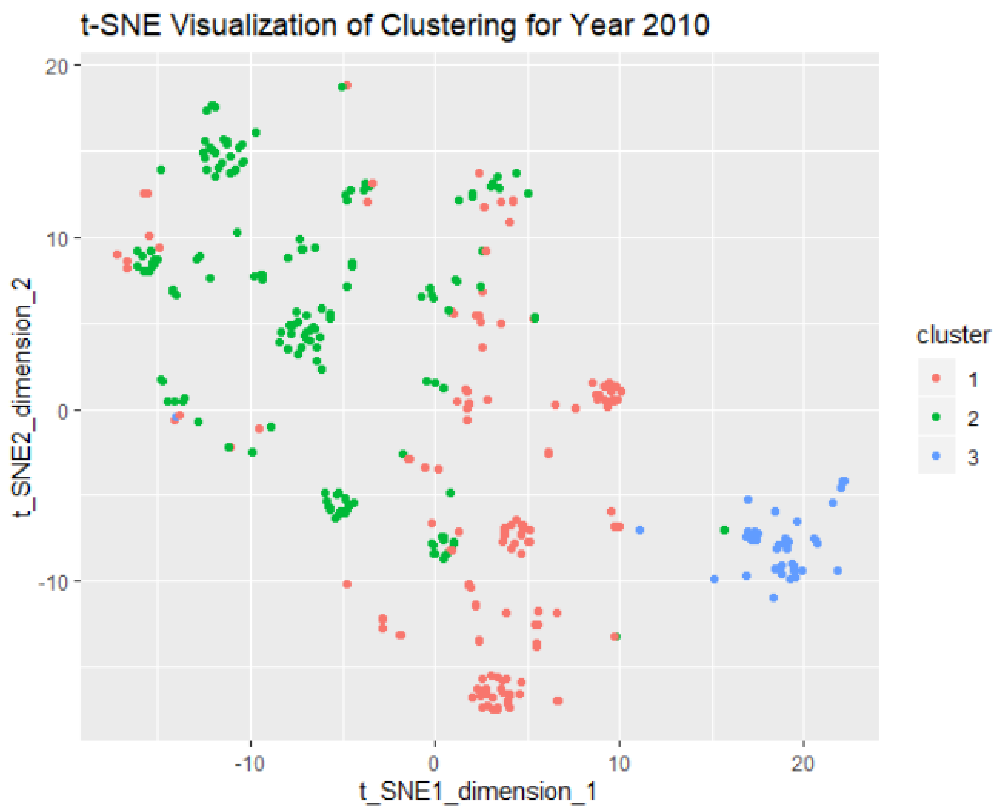


Figure B.10 : t – SNE graphical representation of the three clusters, Year 2010.

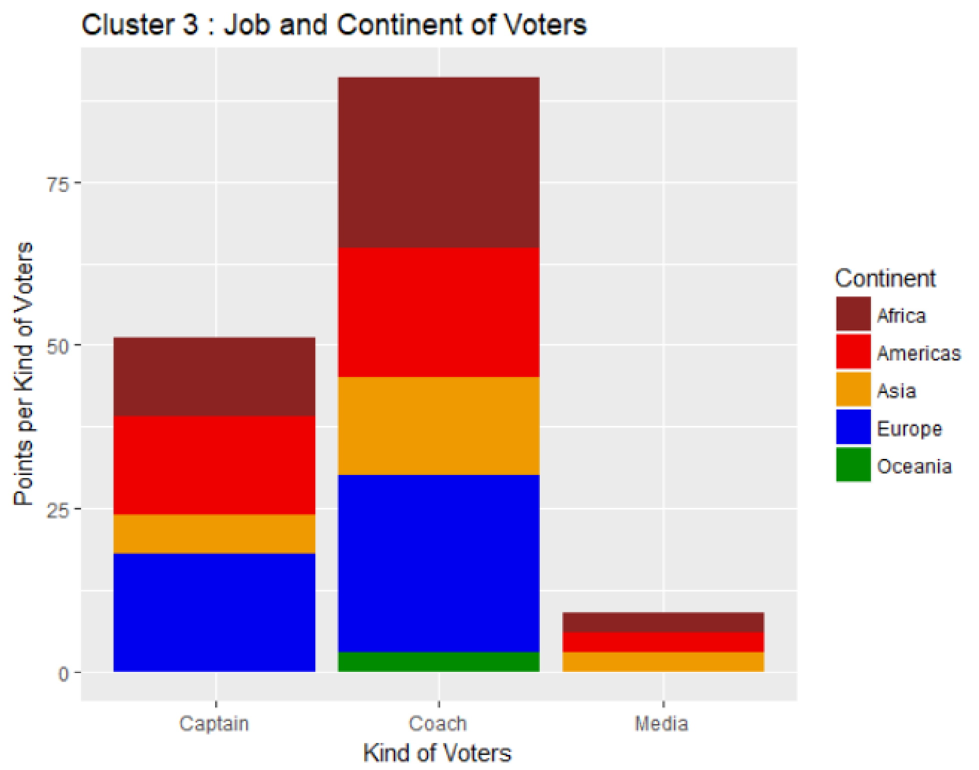


Figure B.11 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 3, Year 2010.

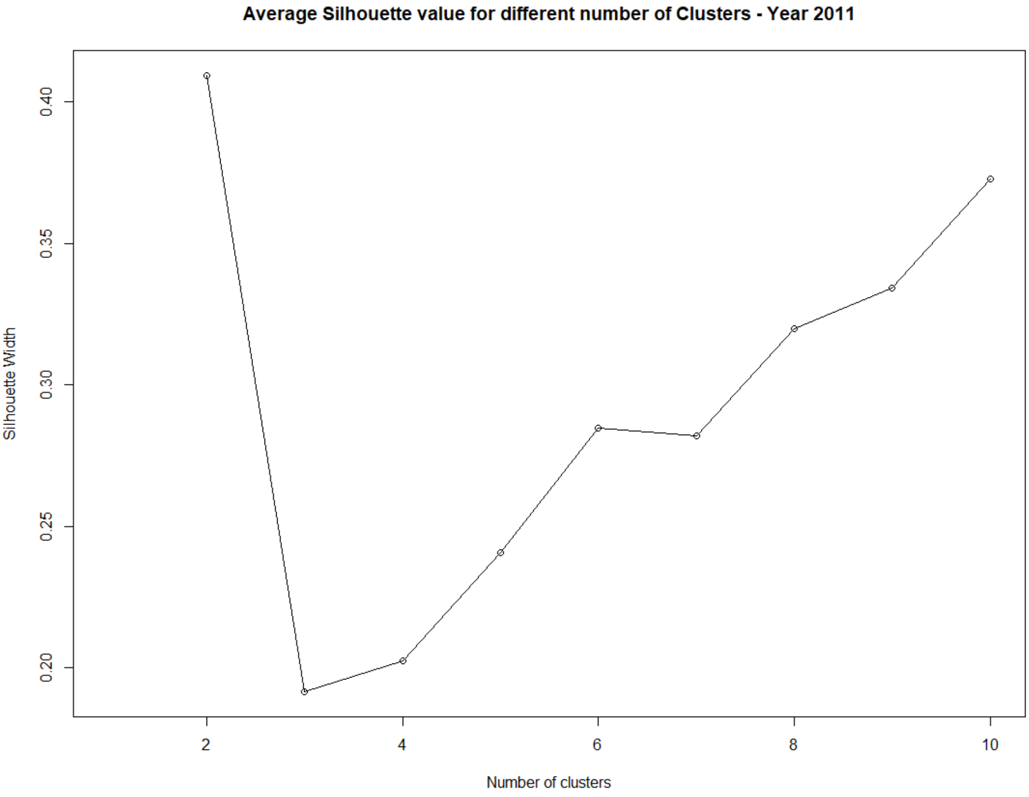


Figure B.12 : Average silhouette plot for the different number of components, Year 2011.

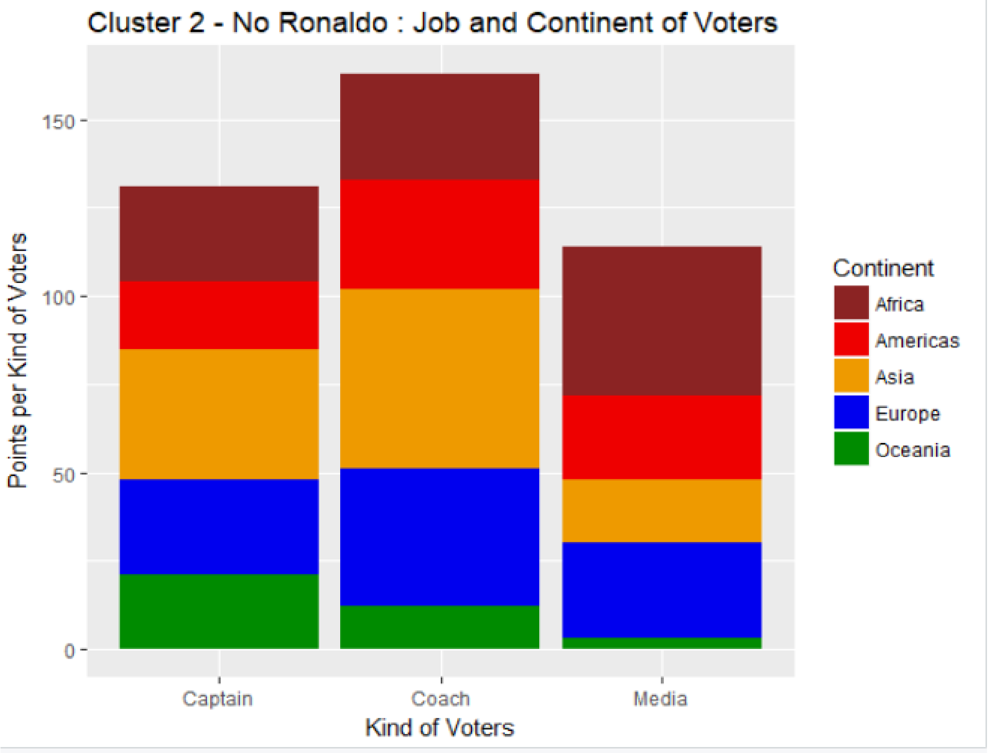


Figure B.13 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 2, Year 2011.

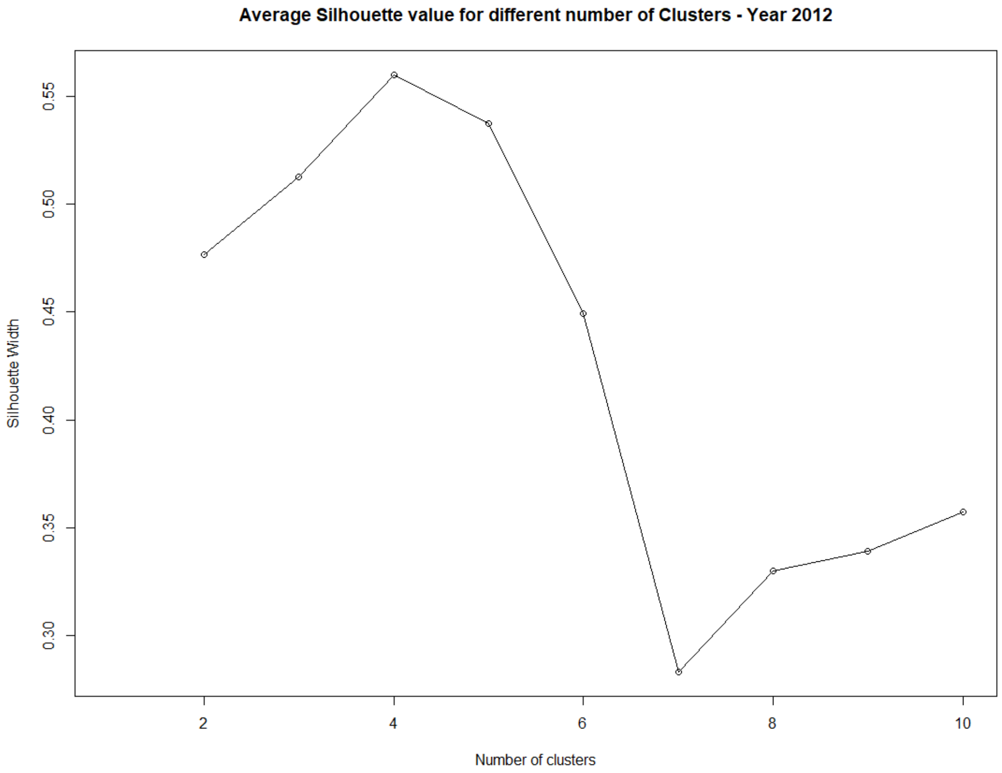


Figure B.14 : Average silhouette plot for the different number of components, Year 2012.

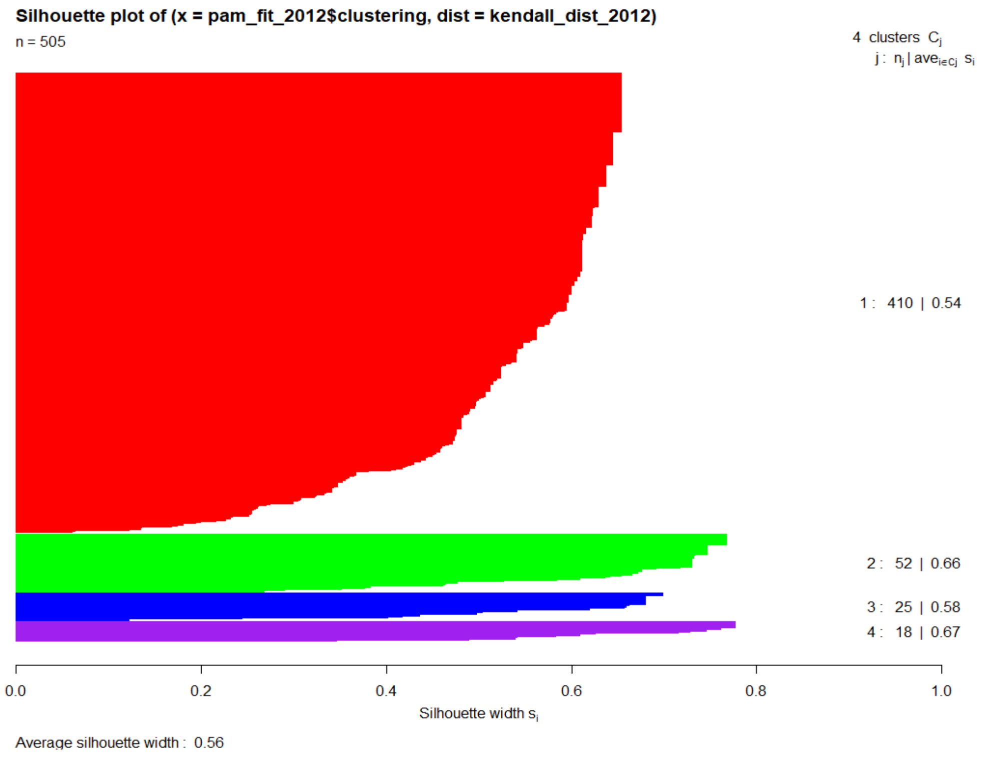


Figure B.15 : Silhouette plot for each cluster of the final 4 – components model, Year 2012.

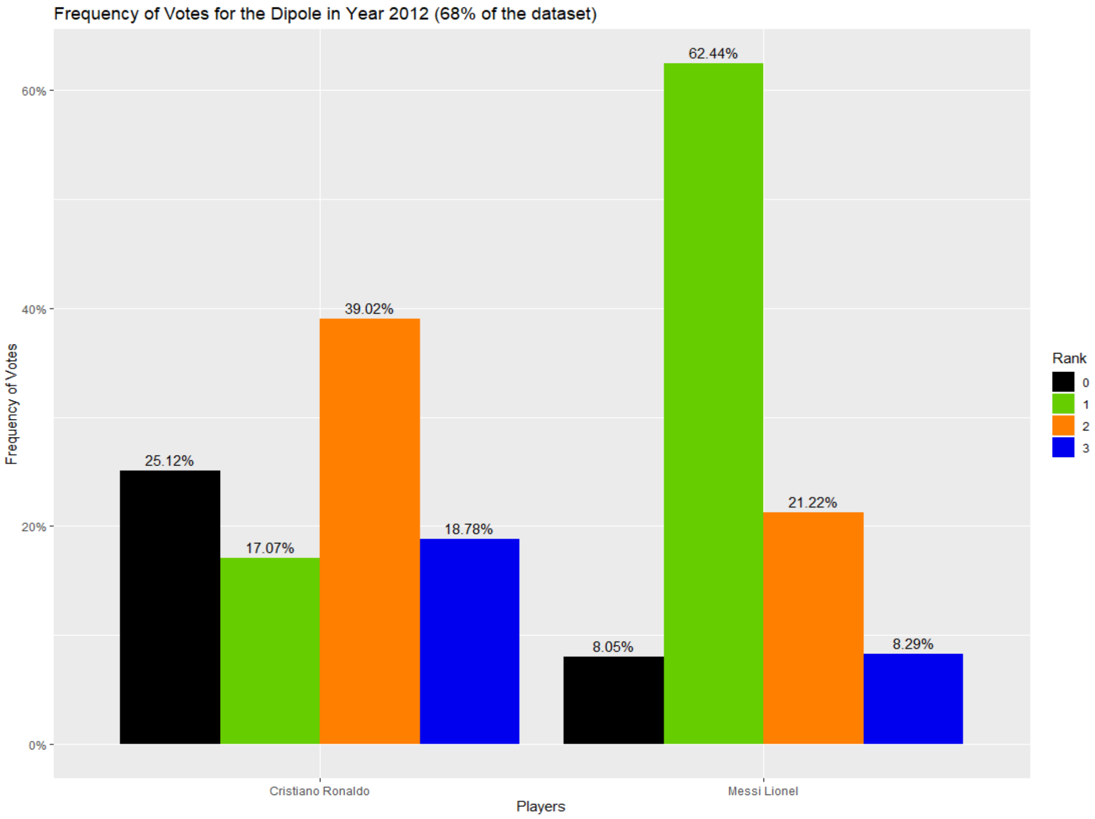


Figure B.16 : Bar plot, representing the percentage of places that Cristiano Ronaldo and Lionel Messi have been ranked in Cluster 1, Year 2012.

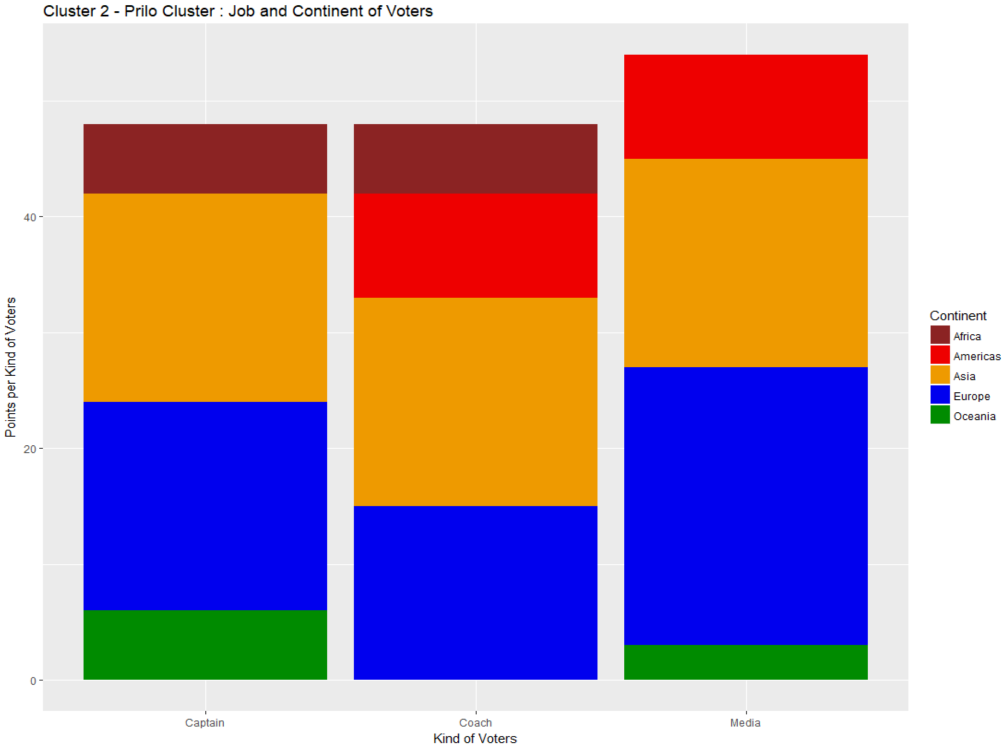


Figure B.17 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 2, Year 2012.

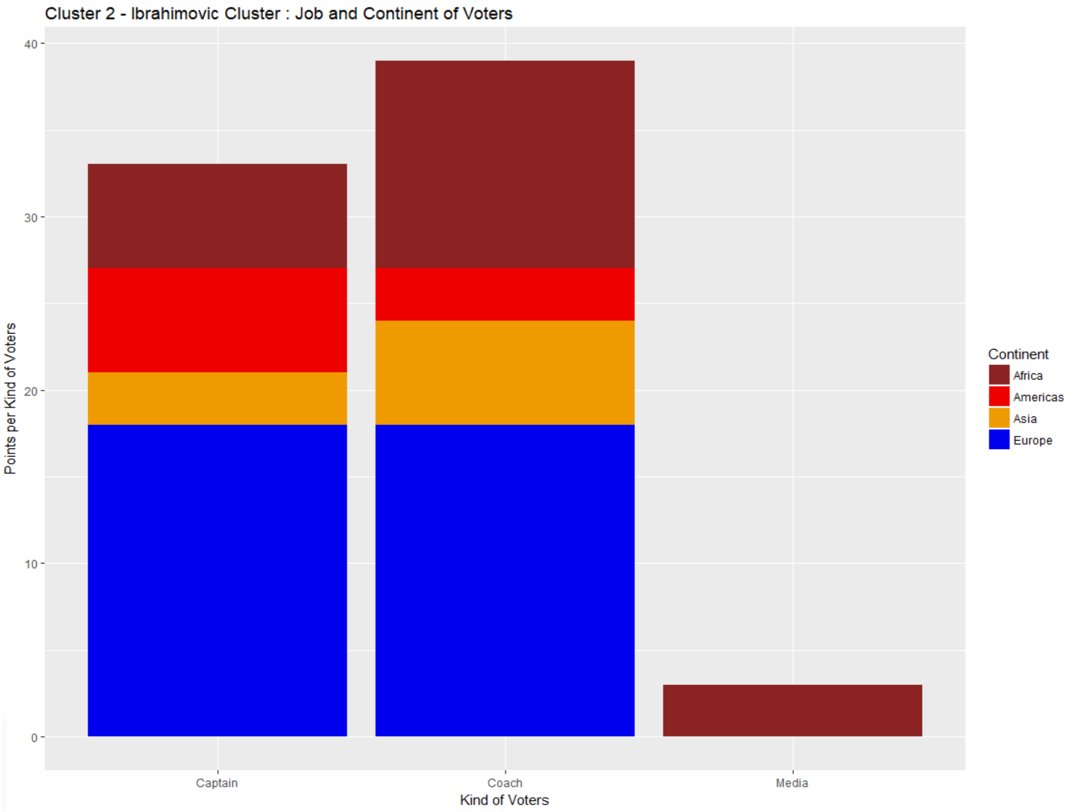


Figure B.18 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 3, Year 2012.

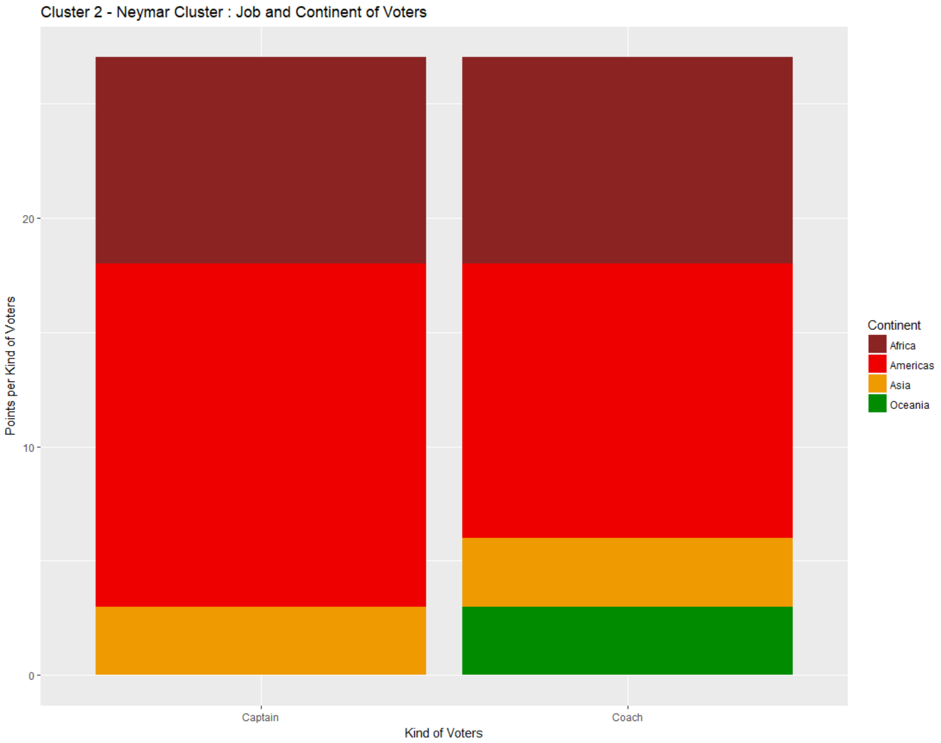


Figure B.19 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 4, Year 2012.

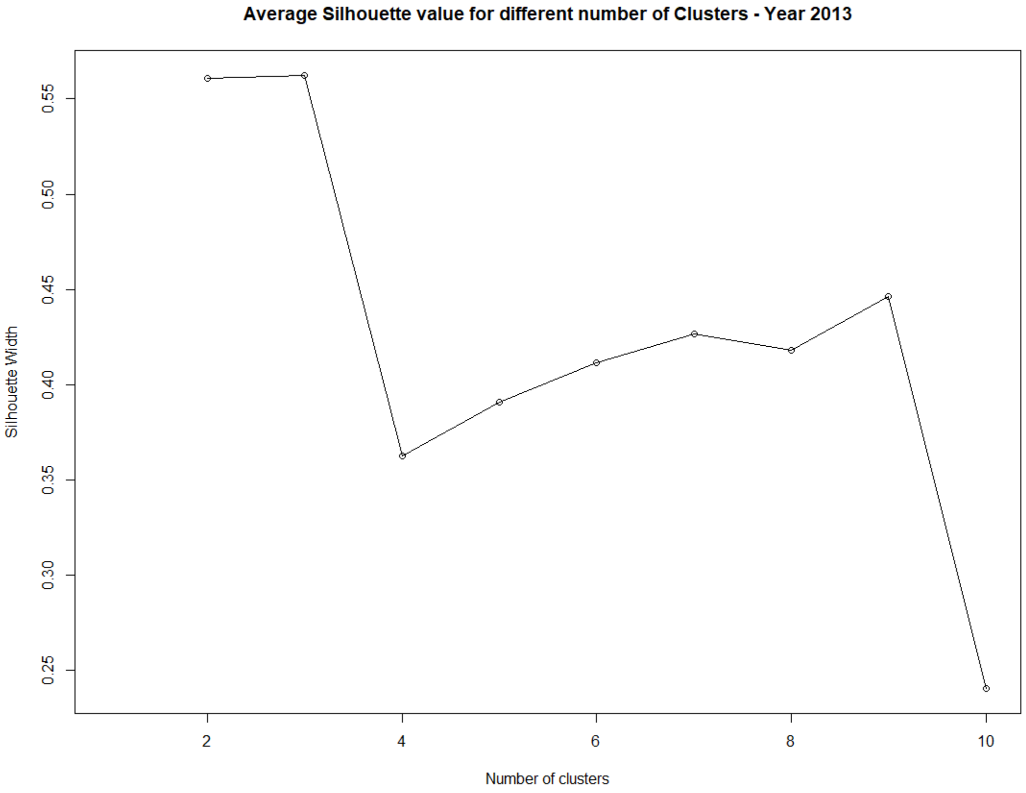


Figure B.20 : Average silhouette plot for the different number of components, Year 2013.

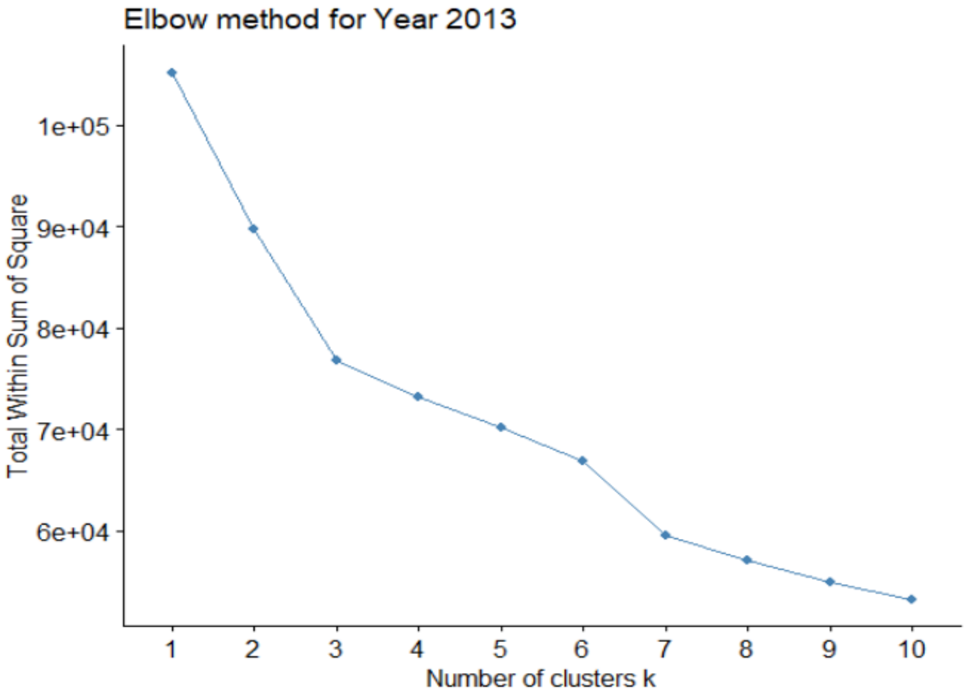


Figure B.21 : The Elbow method using the Total Within Sum of Square for the different number of components, Year 2013.

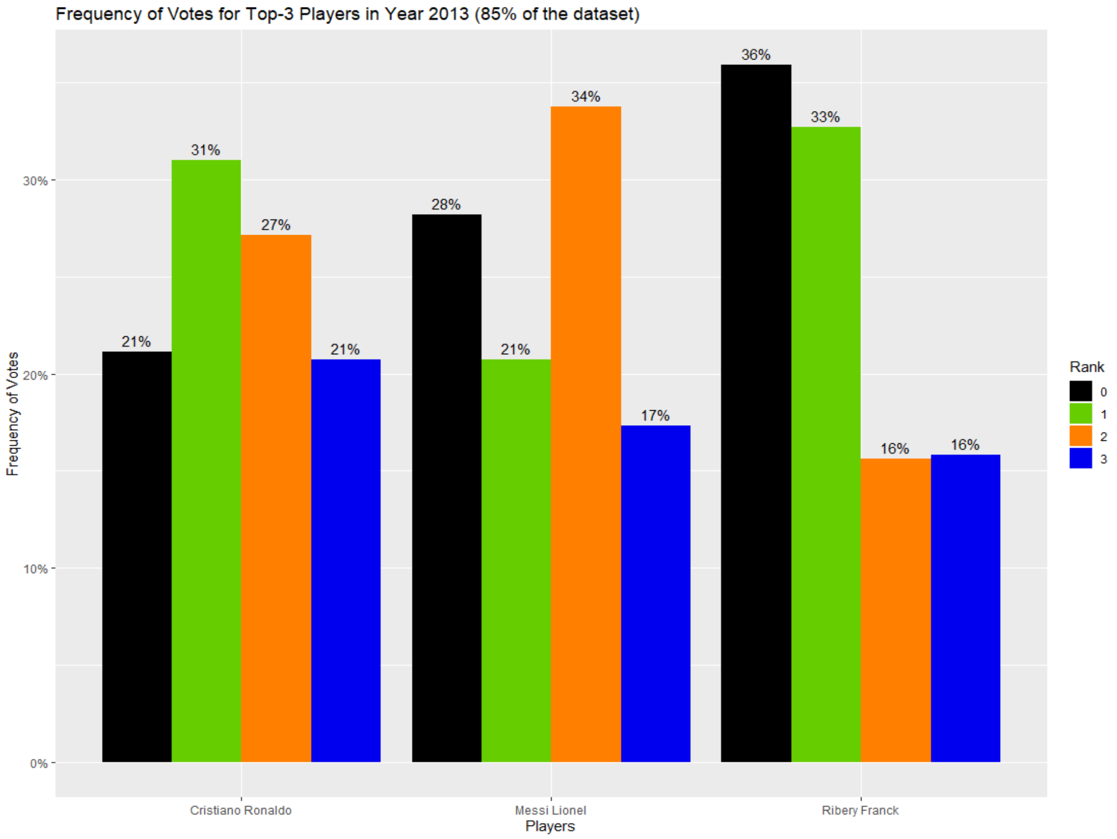


Figure B.22 : Bar plot, representing the percentage of places that Cristiano Ronaldo, Lionel Messi and Ribery Franck have been ranked in Cluster 1, Year 2013.

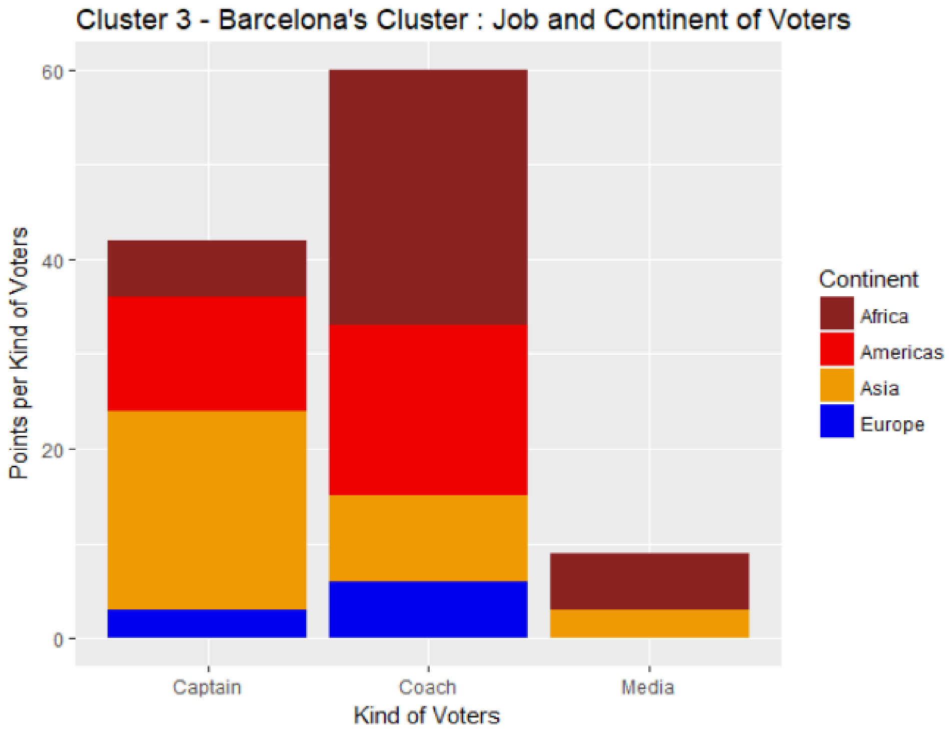


Figure B.23 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 3, Year 2013.

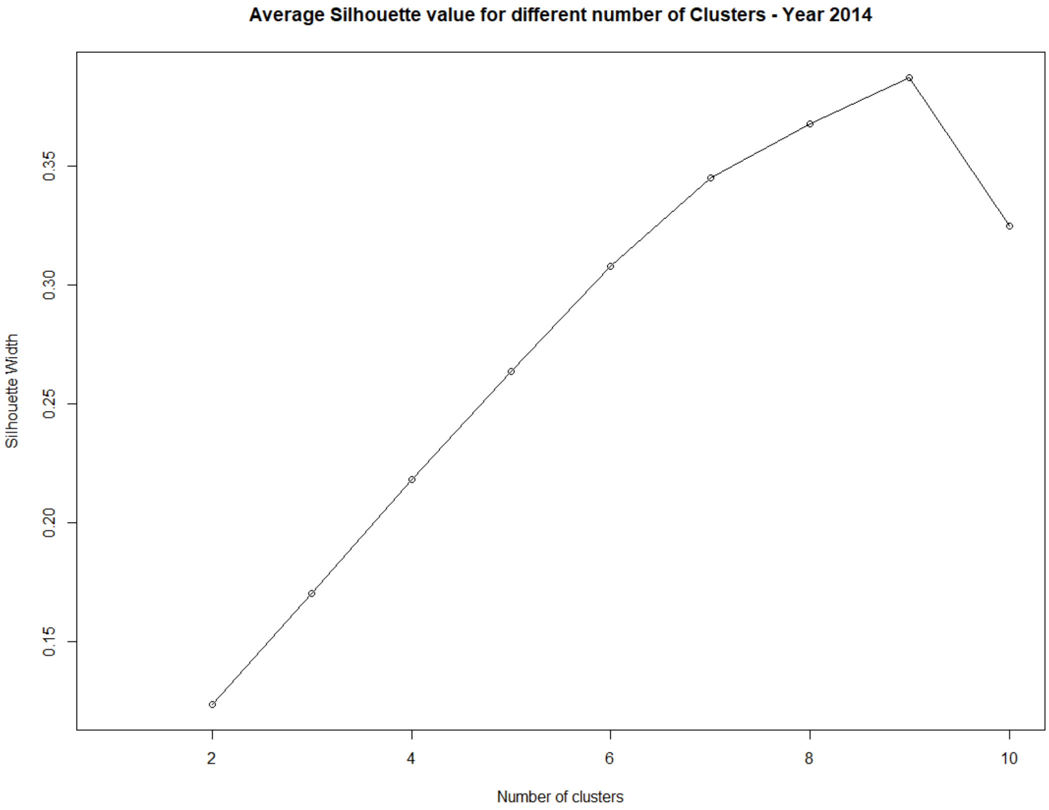


Figure B.24 : Average silhouette plot for the different number of components, Year 2014.

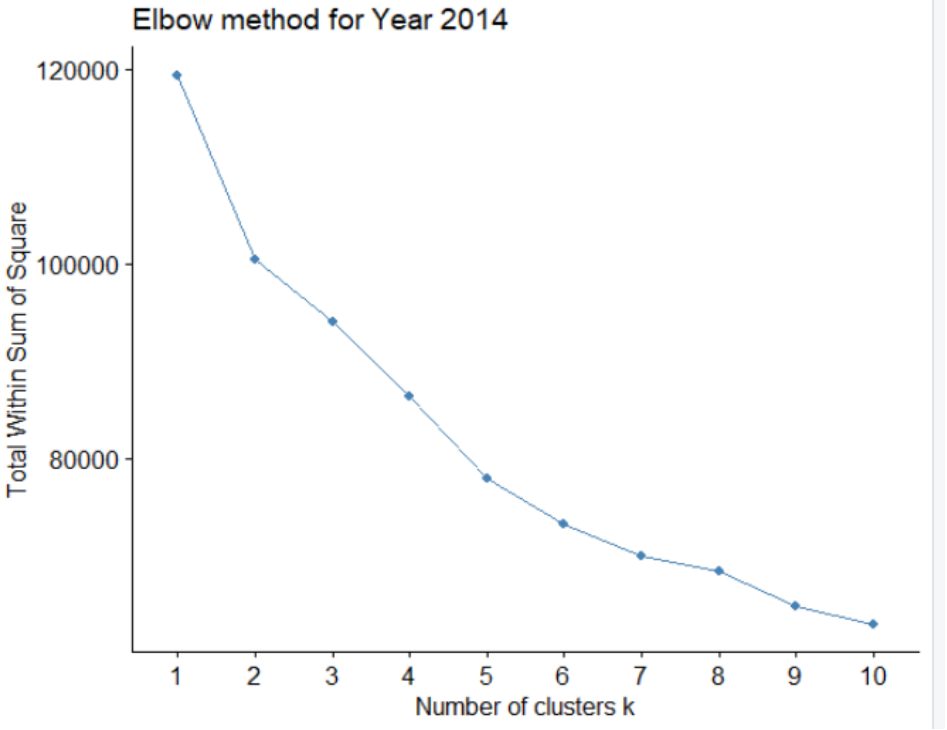


Figure B.25 : The Elbow method using the Total Within Sum of Square for the different number of components, Year 2014.

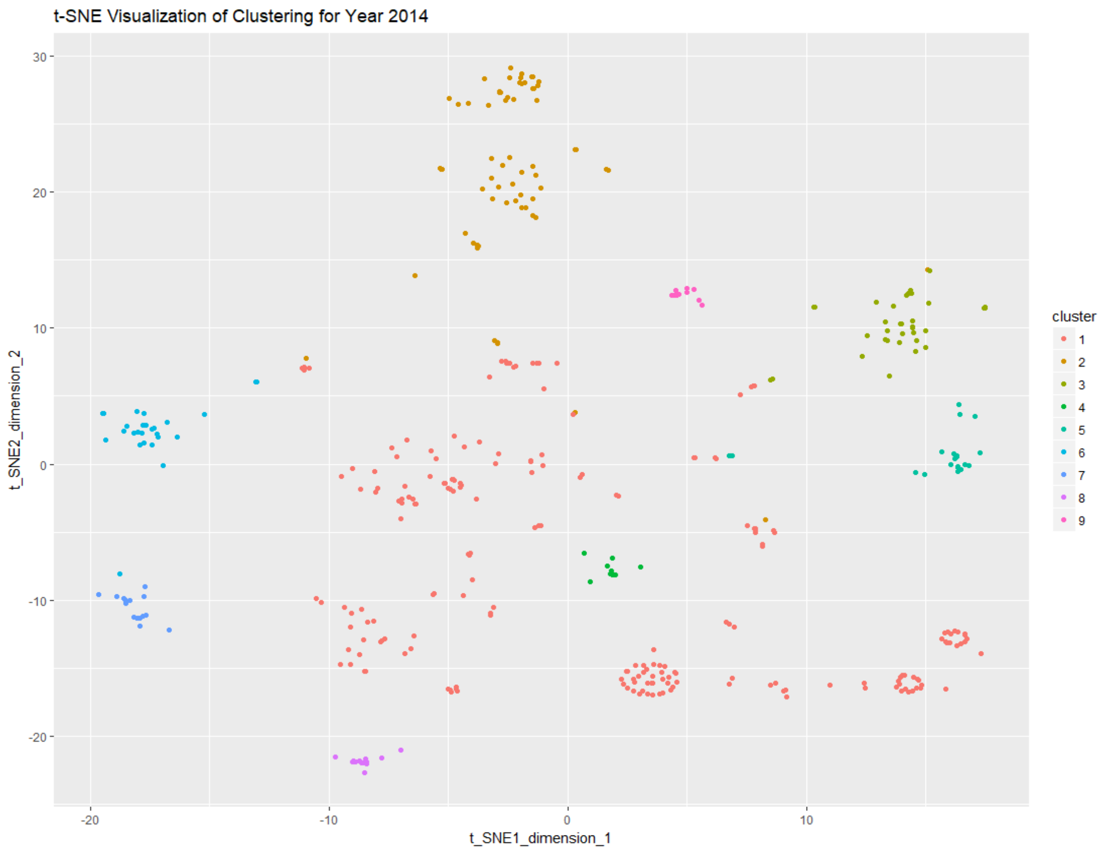


Figure B.26 : t – SNE graphical representation of the nine clusters, Year 2014.

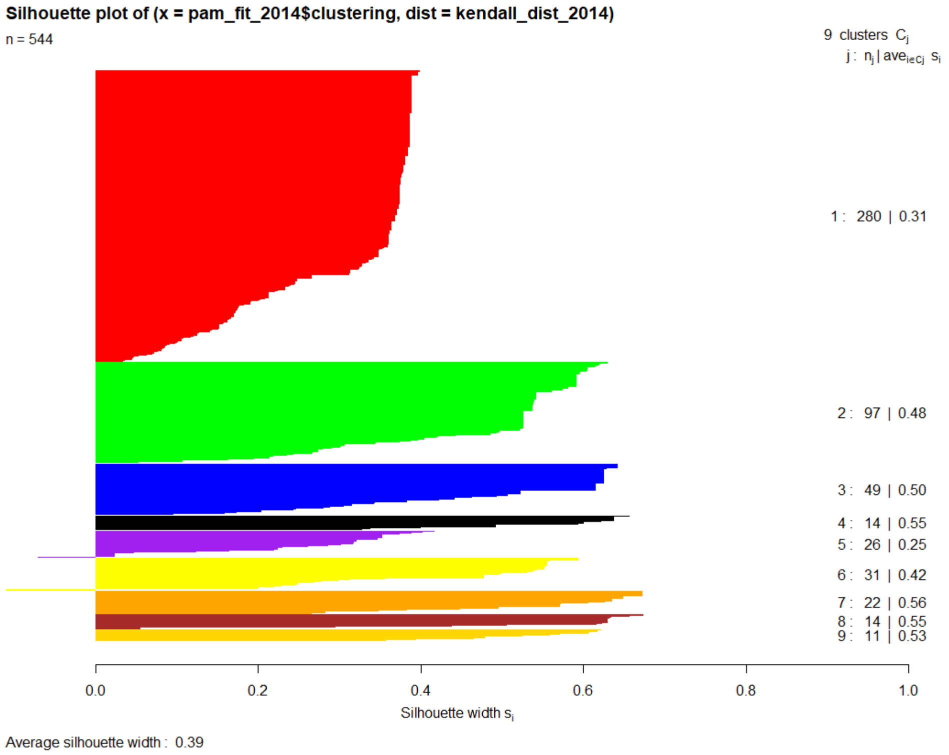


Figure B.27 : Silhouette plot for each cluster of the final 9 – components model, Year 2014.

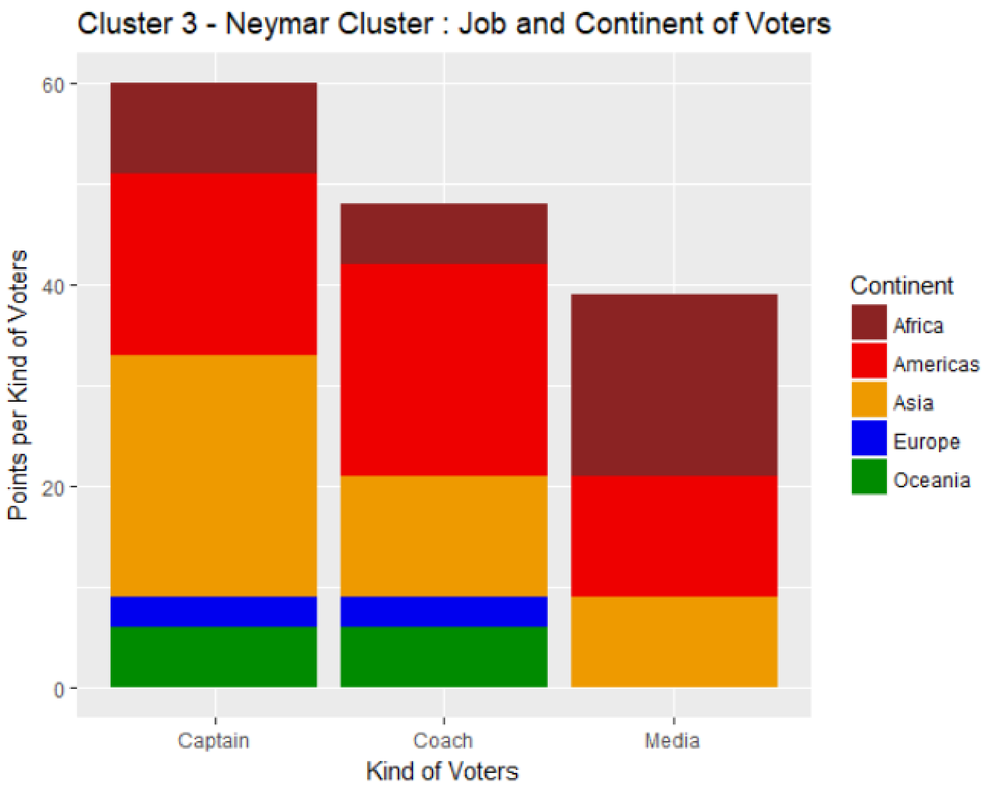


Figure B.28 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 3, Year 2014.

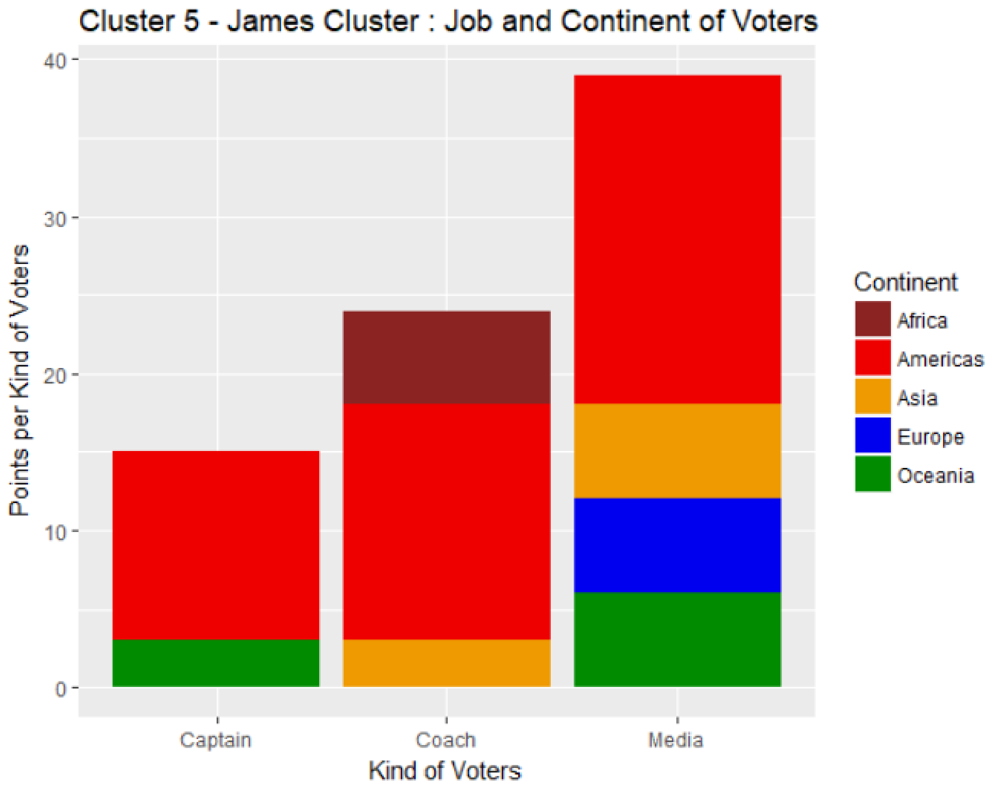


Figure B.29 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 5, Year 2014.

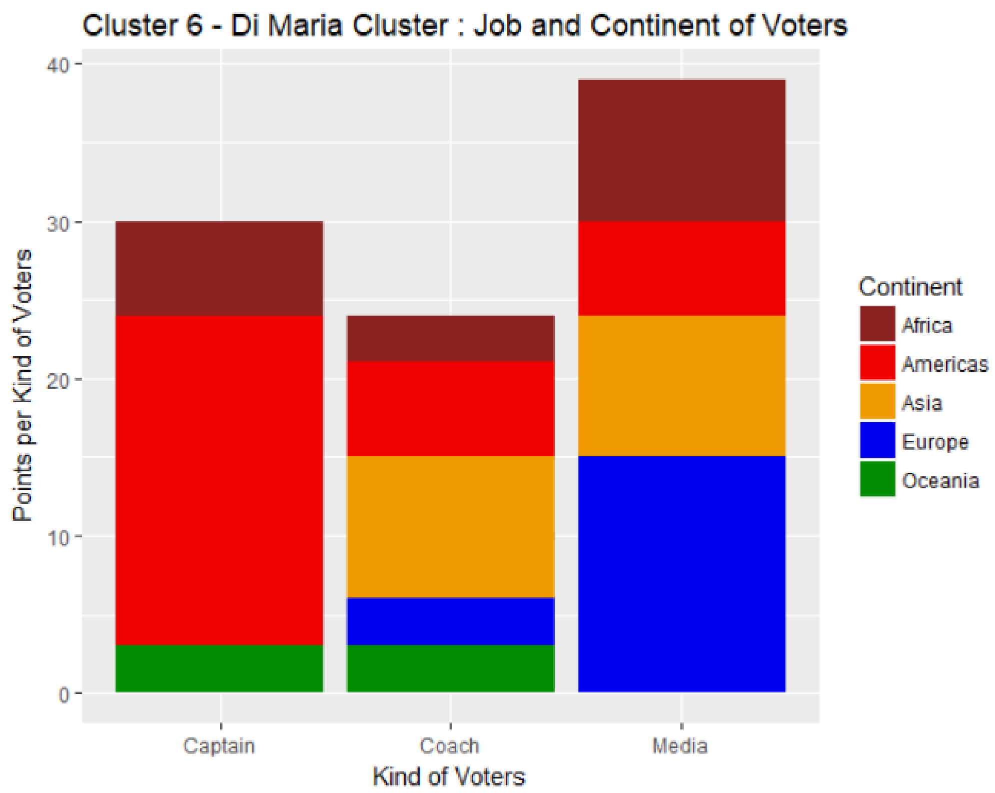


Figure B.30 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 6, Year 2014.

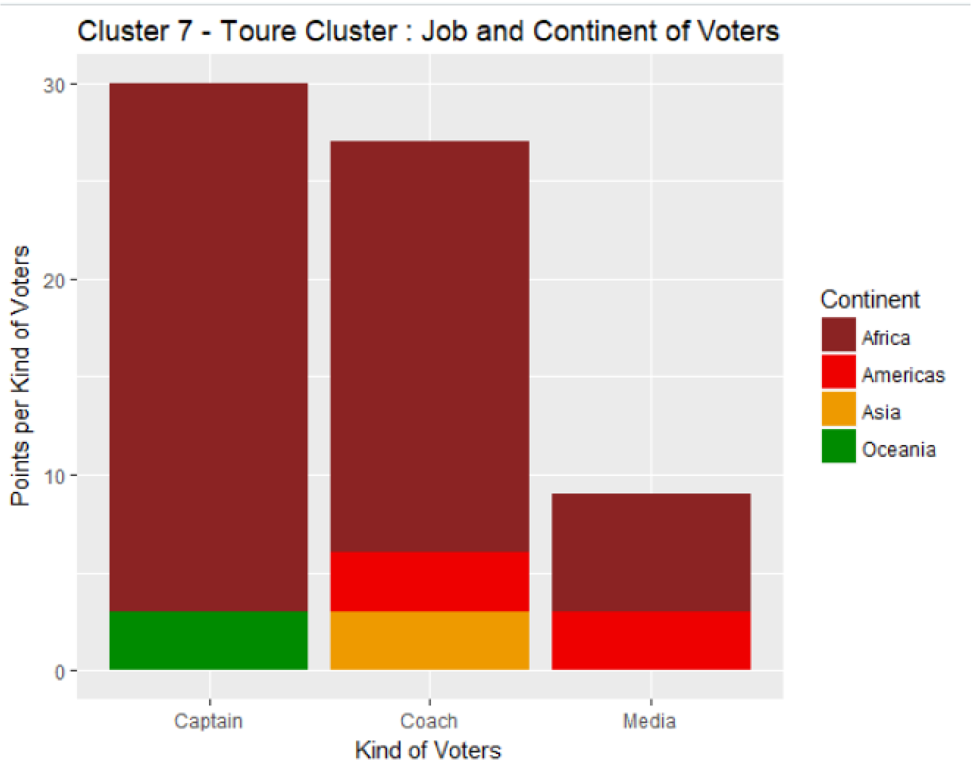


Figure B.31 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 7, Year 2014.

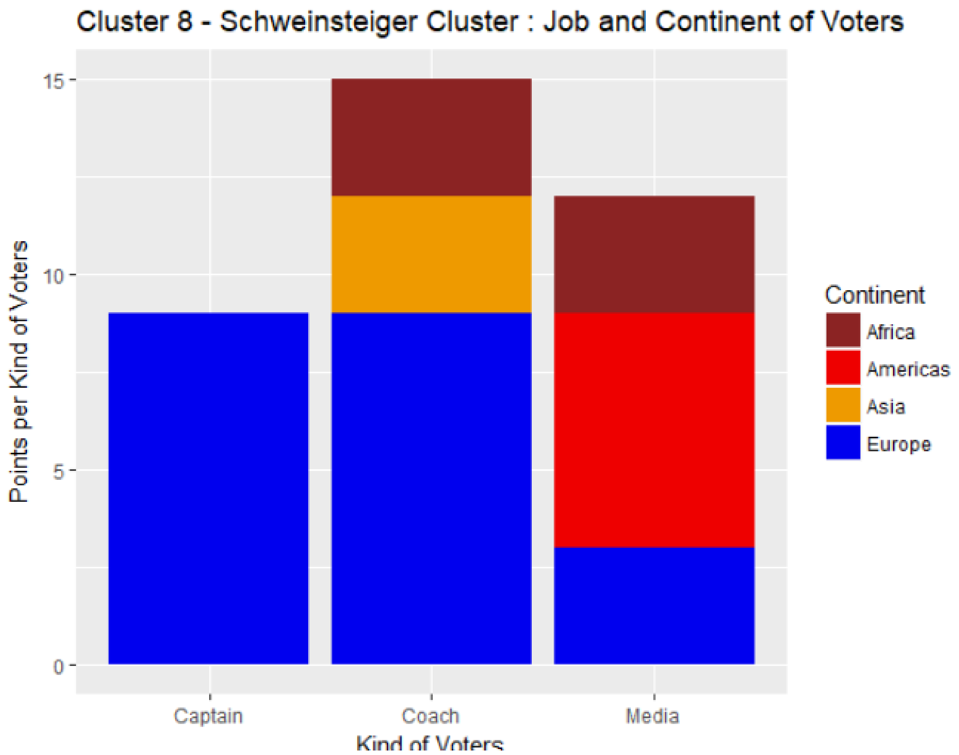


Figure B.32 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 8, Year 2014.

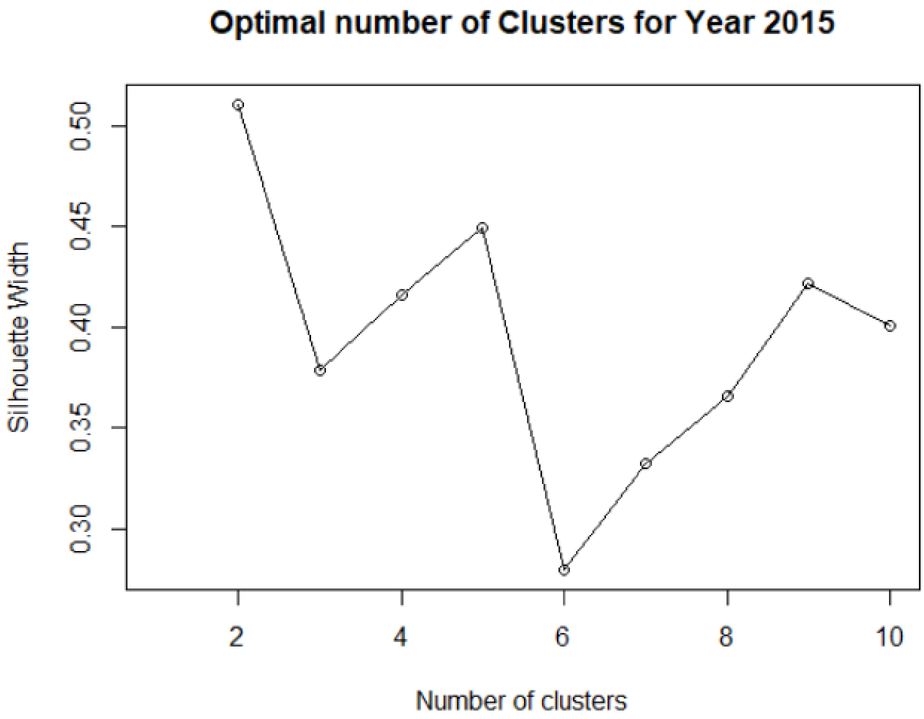


Figure B.33 : Average silhouette plot for the different number of components, Year 2015.

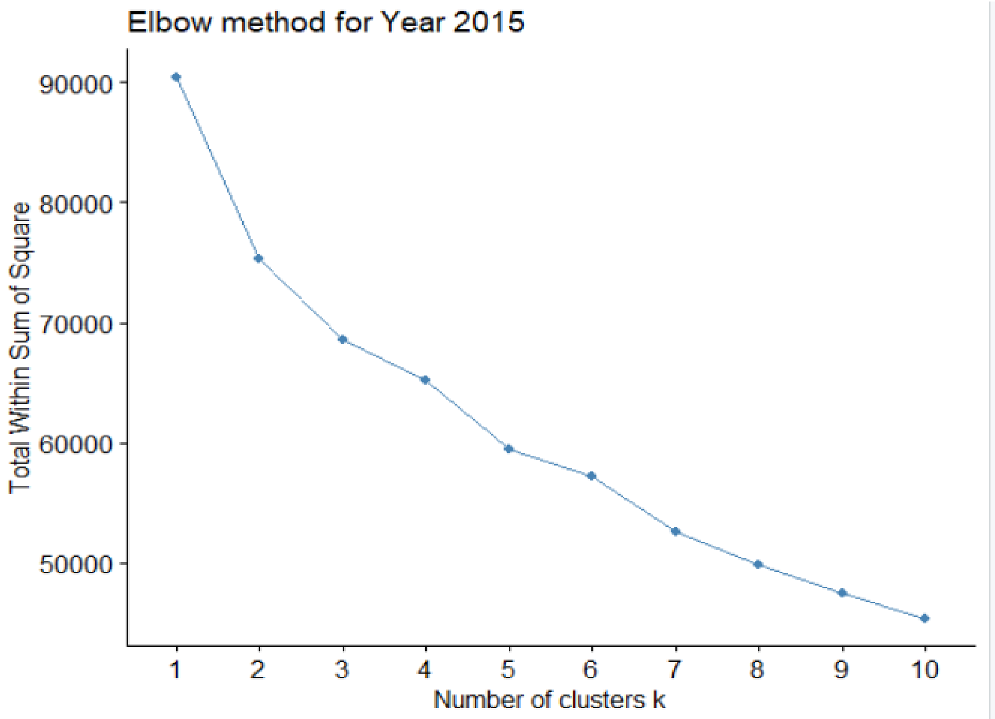


Figure B.34 : The Elbow method using the Total Within Sum of Square for the different number of components, Year 2015.

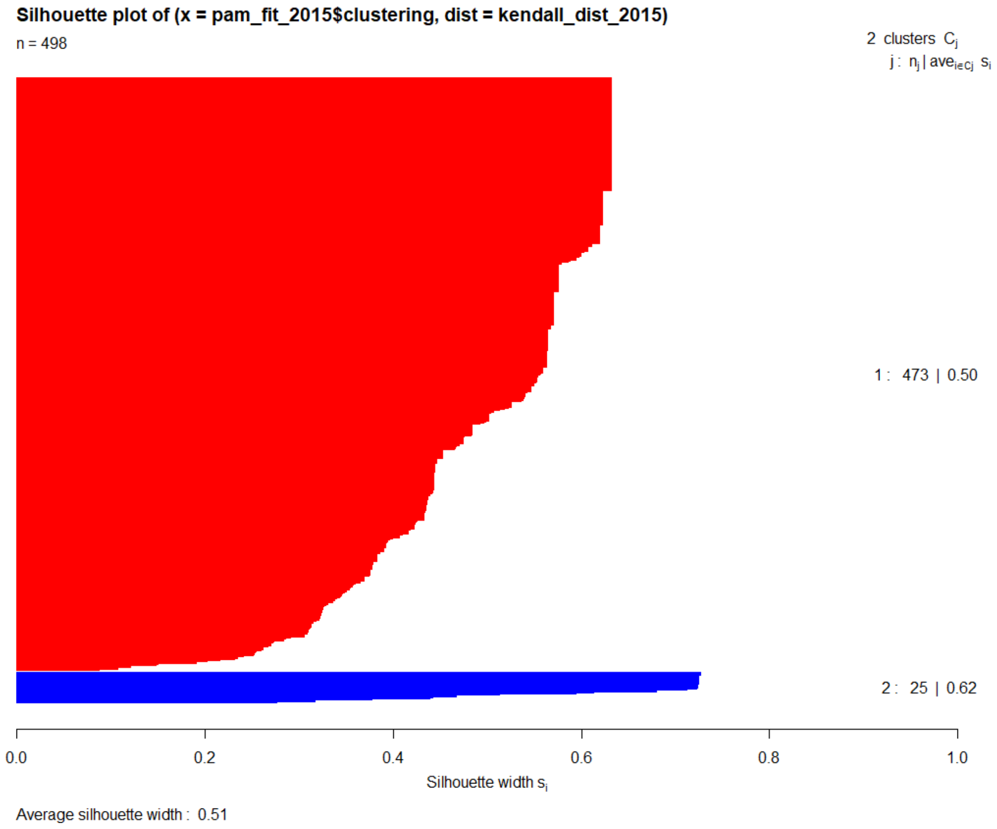


Figure B.35 : Silhouette plot for each cluster of the final 2 – components model, Year 2015.

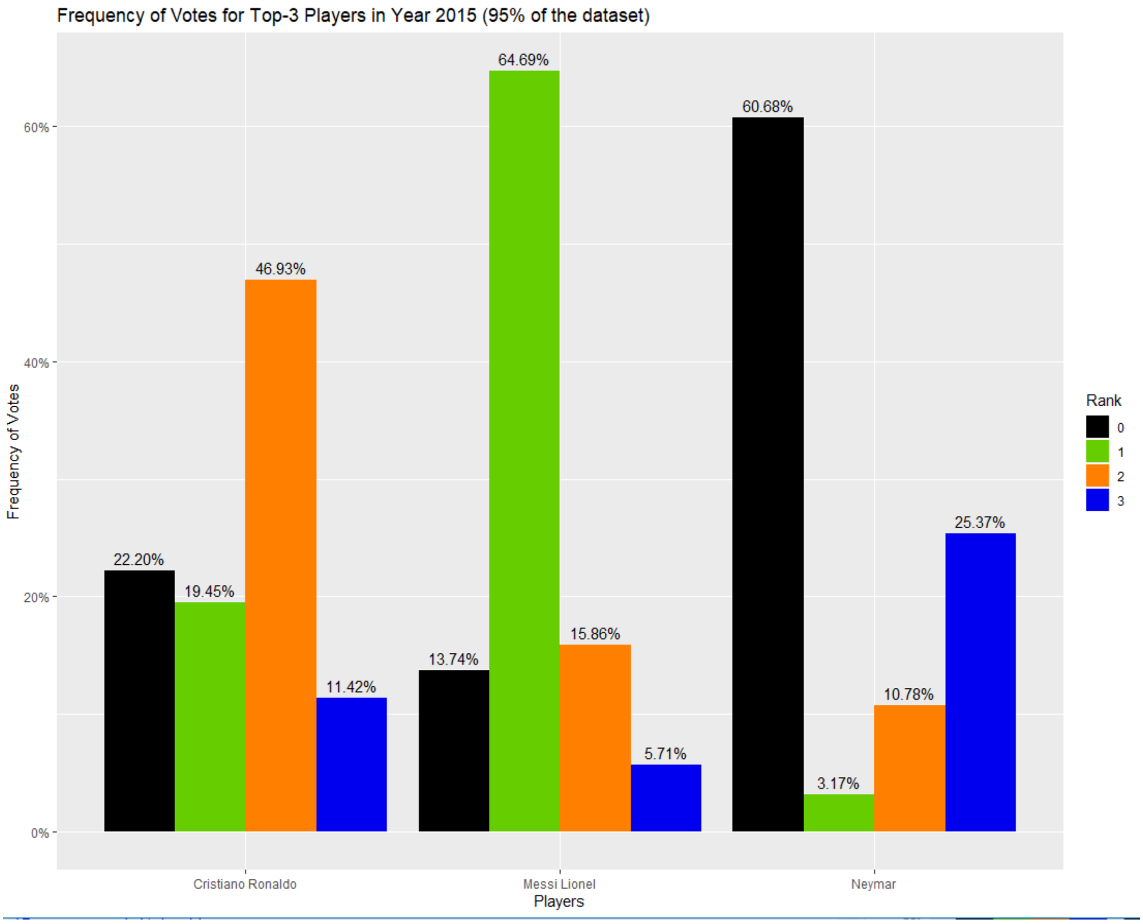


Figure B.36 : Bar plot, representing the percentage of places that Cristiano Ronaldo, Lionel Messi and Neymar have been ranked in Cluster 1, Year 2015.

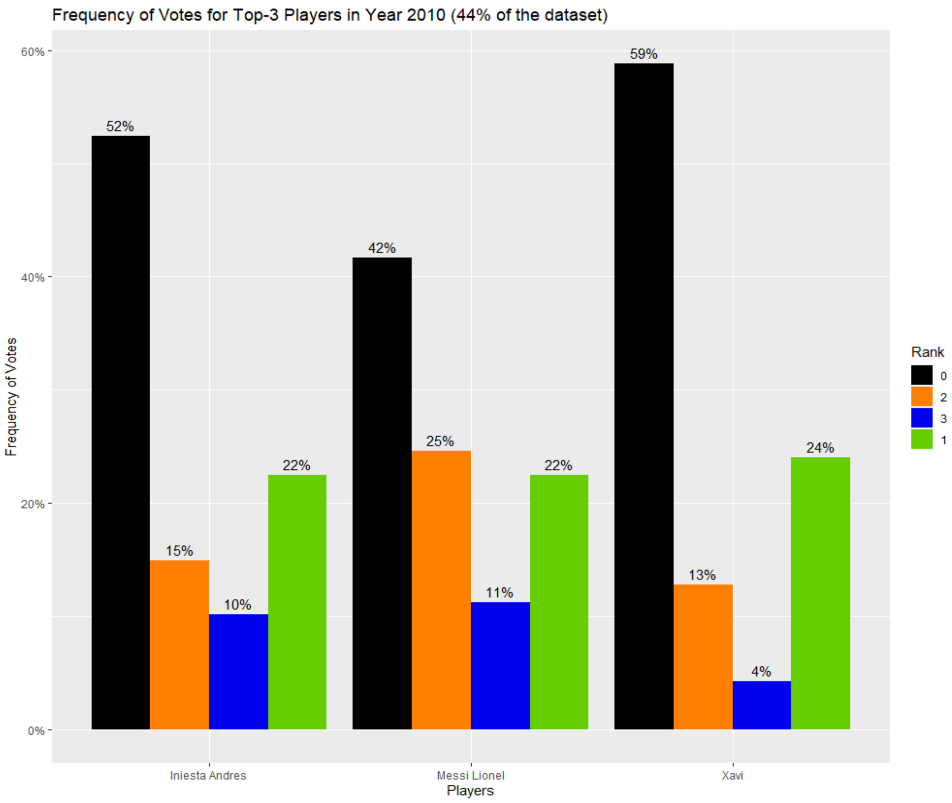


Figure B.37 : Bar plot, representing the percentage of places that Iniesta Andres, Messi Lionel and Xavi have been ranked in Cluster 4, Year 2010.

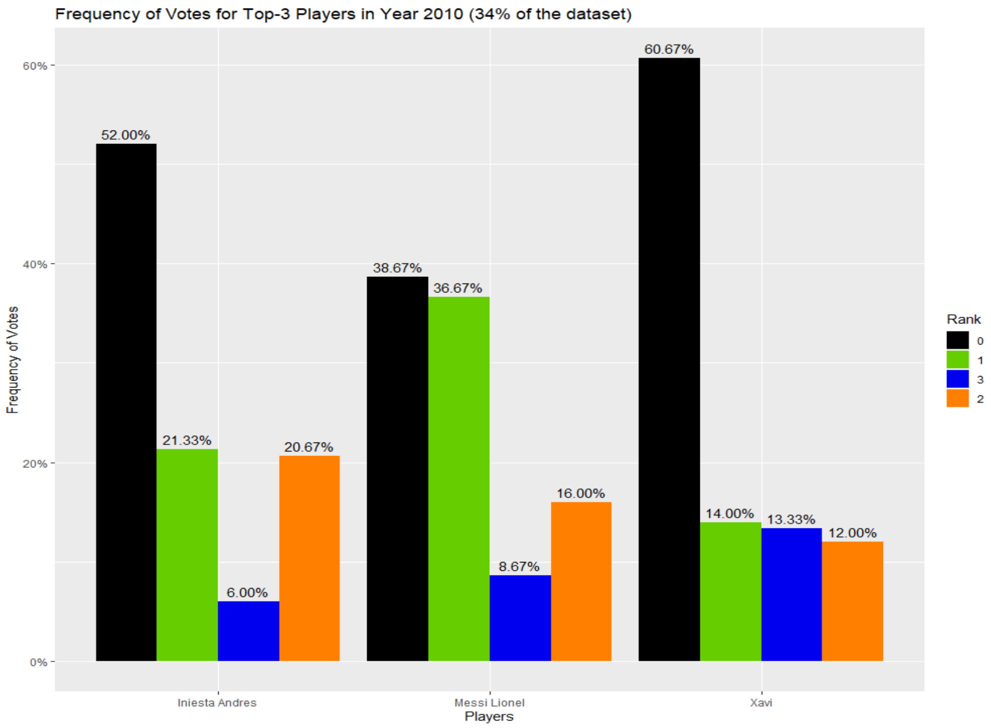


Figure B.38 : Bar plot, representing the percentage of places that Iniesta Andres, Messi Lionel and Xavi have been ranked in Cluster 2, Year 2010.

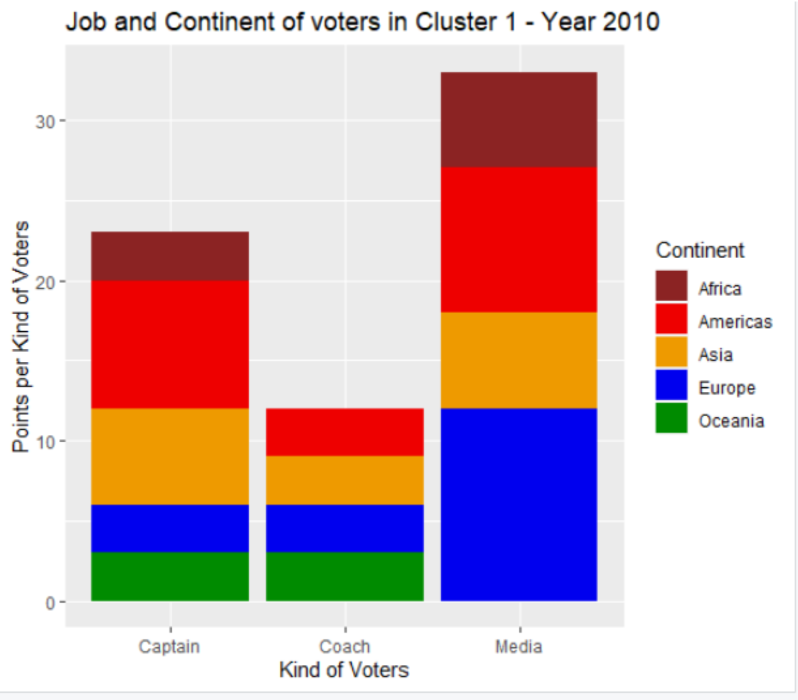


Figure B.39 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 1, Year 2010.

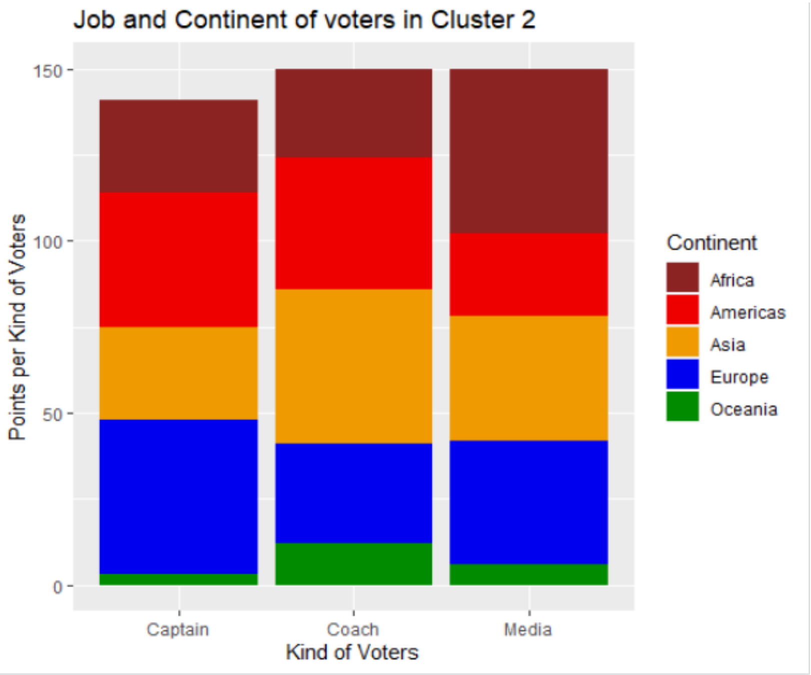


Figure B.40 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 2, Year 2010.

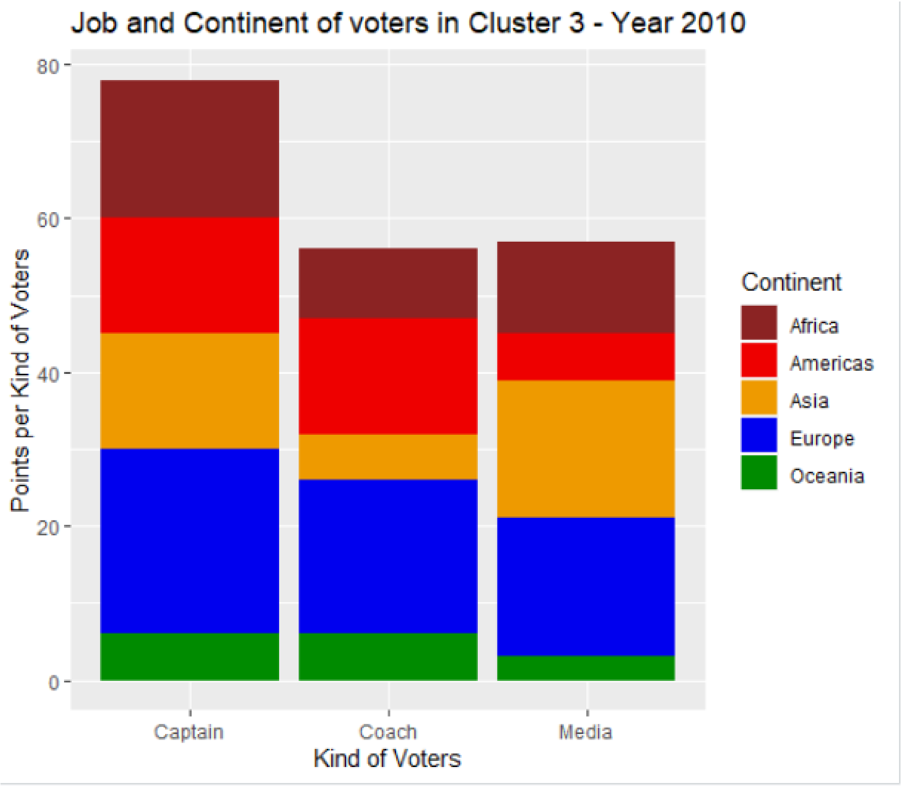


Figure B.41 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 3, Year 2010.

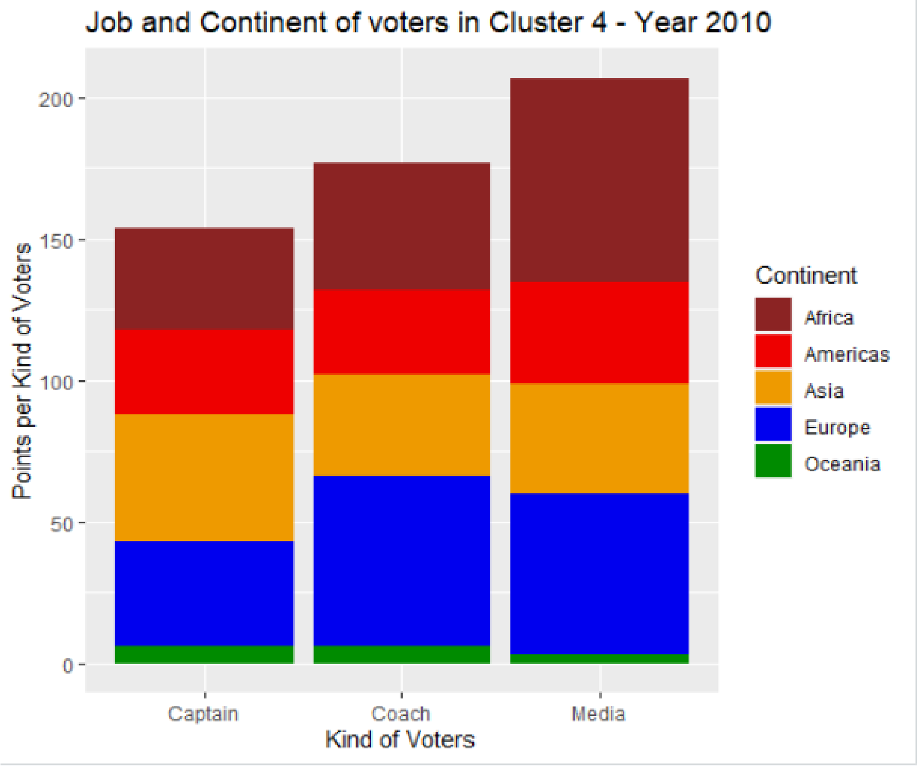


Figure B.42 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 4, Year 2010.

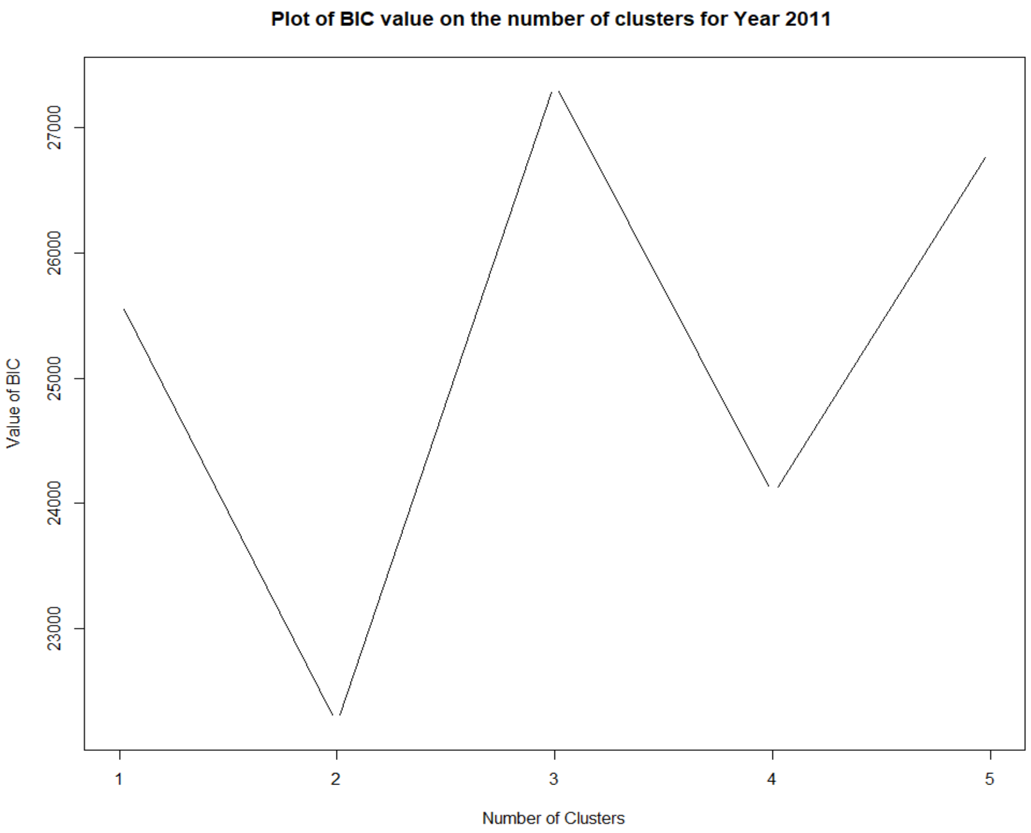


Figure B.43 : Plot of the BIC value on the different number of components, Year 2011.

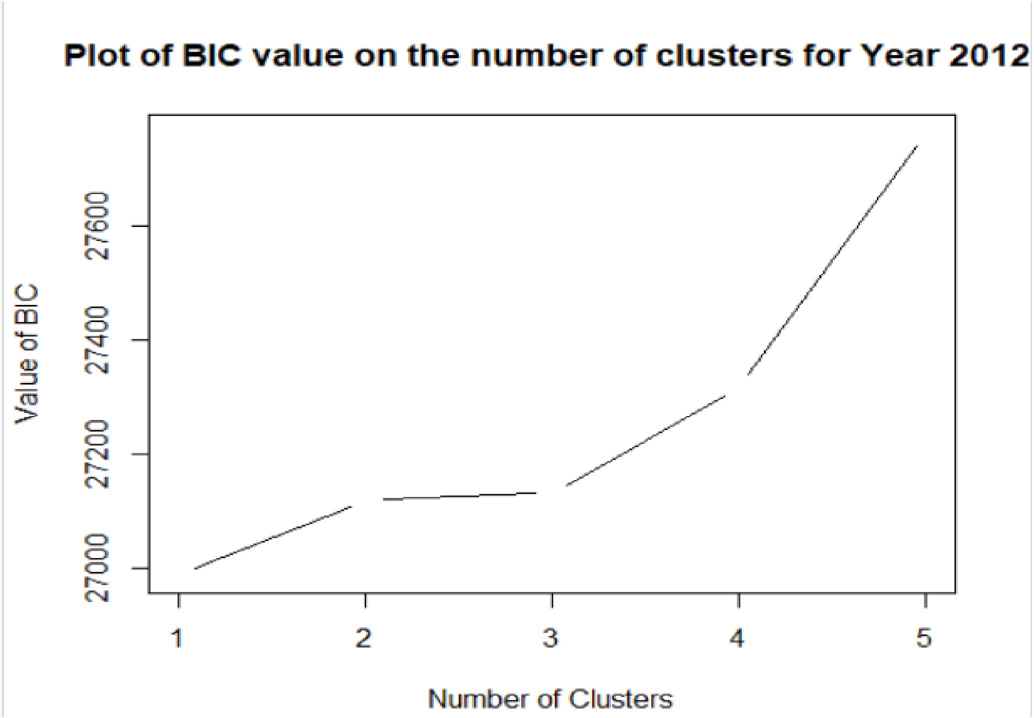


Figure B.44 : Plot of the BIC value on the different number of components, Year 2012.

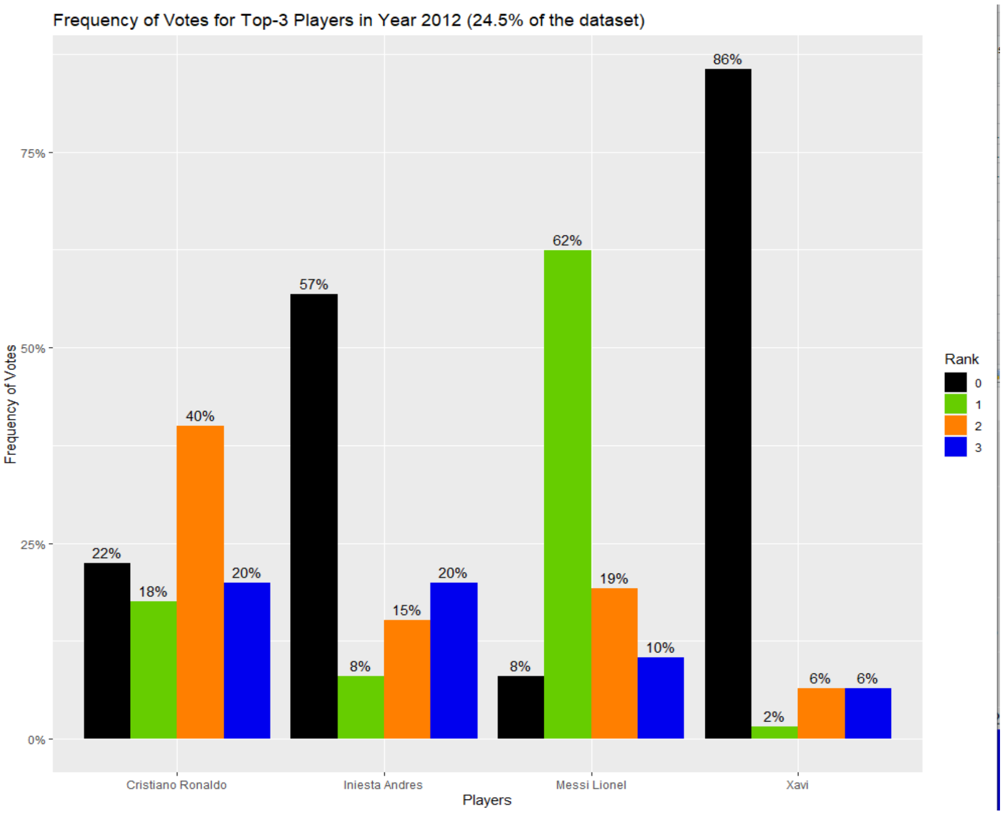


Figure B.45 : Bar plot, representing the percentage of places that Cristiano Ronaldo, Iniesta Andres, Lionel Messi and Xavi have been ranked in Cluster 1, Year 2012.

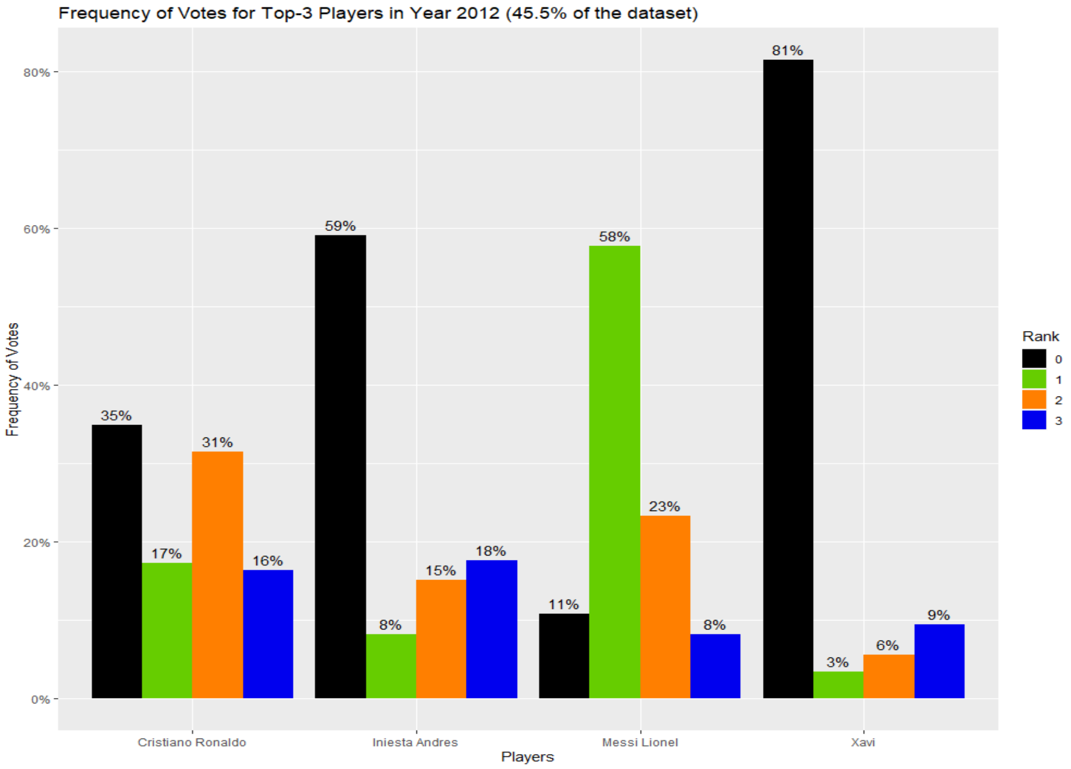


Figure B.46 : Bar plot, representing the percentage of places that Cristiano Ronaldo, Iniesta Andres, Lionel Messi and Xavi have been ranked in Cluster 2, Year 2012.

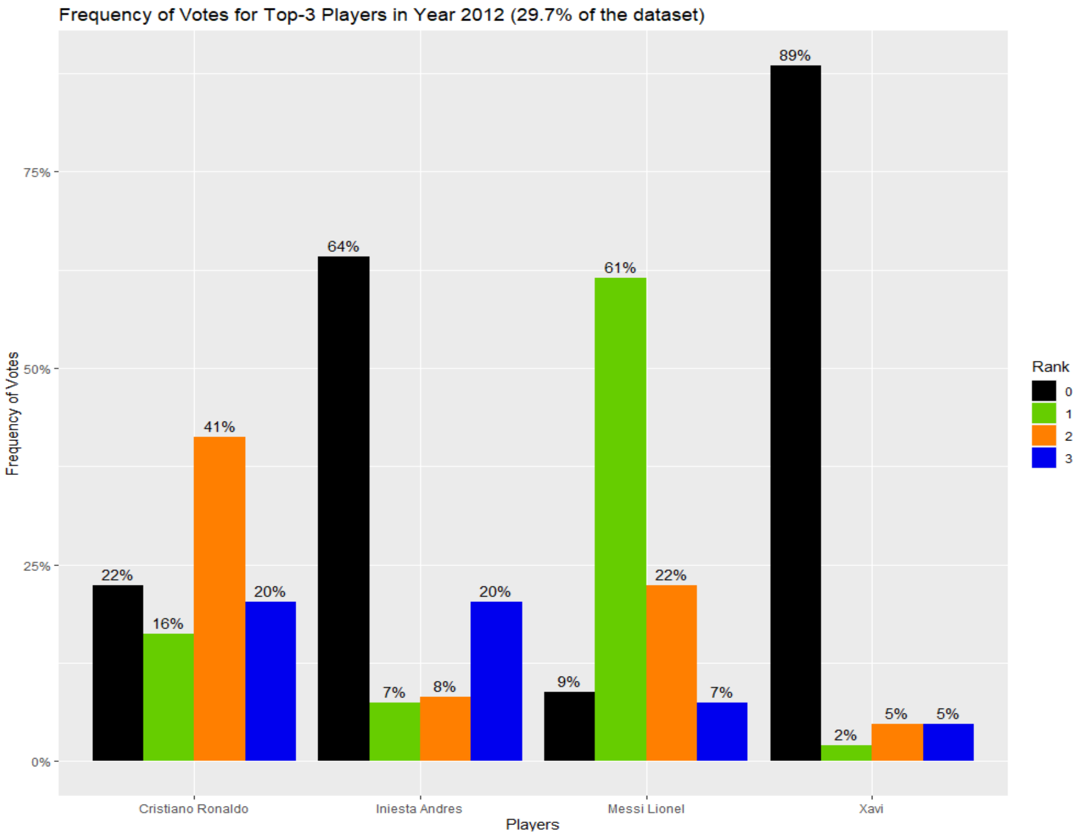


Figure B.47 : Bar plot, representing the percentage of places that Cristiano Ronaldo, Iniesta Andres, Lionel Messi and Xavi have been ranked in Cluster 3, Year 2012.

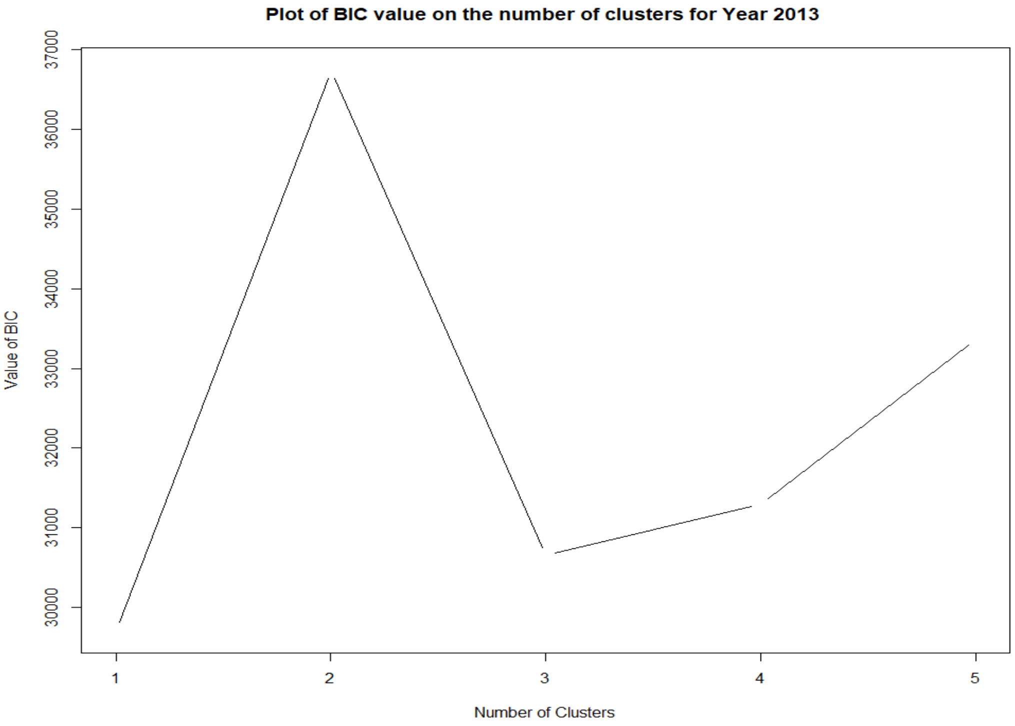


Figure B.48 : Plot of the BIC value on the different number of components, Year 2013.

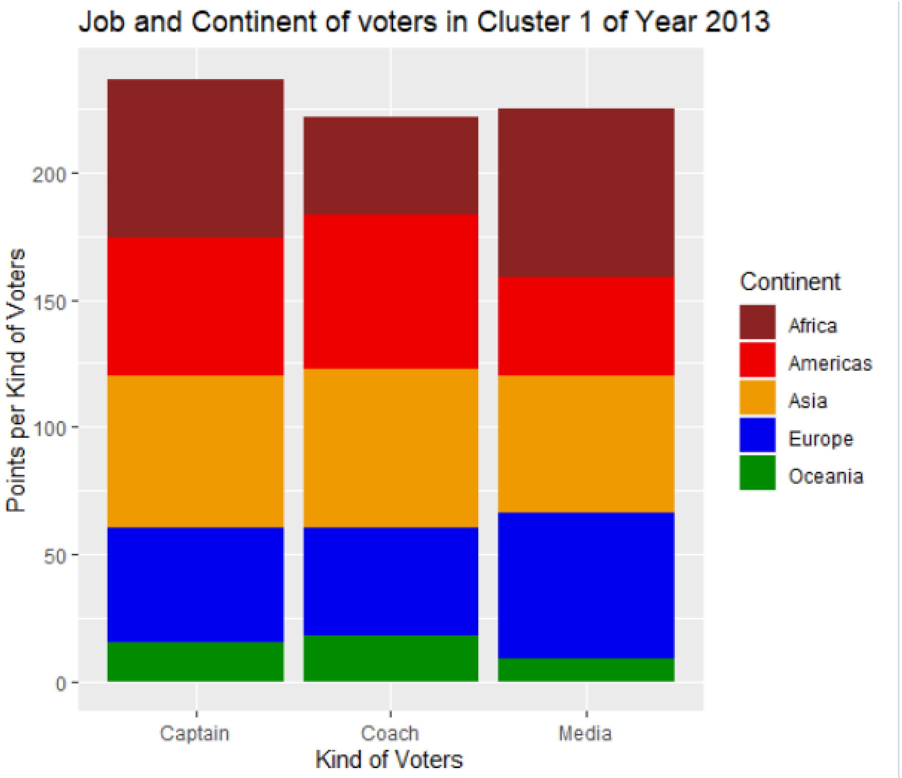


Figure B.49 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 1, Year 2013.

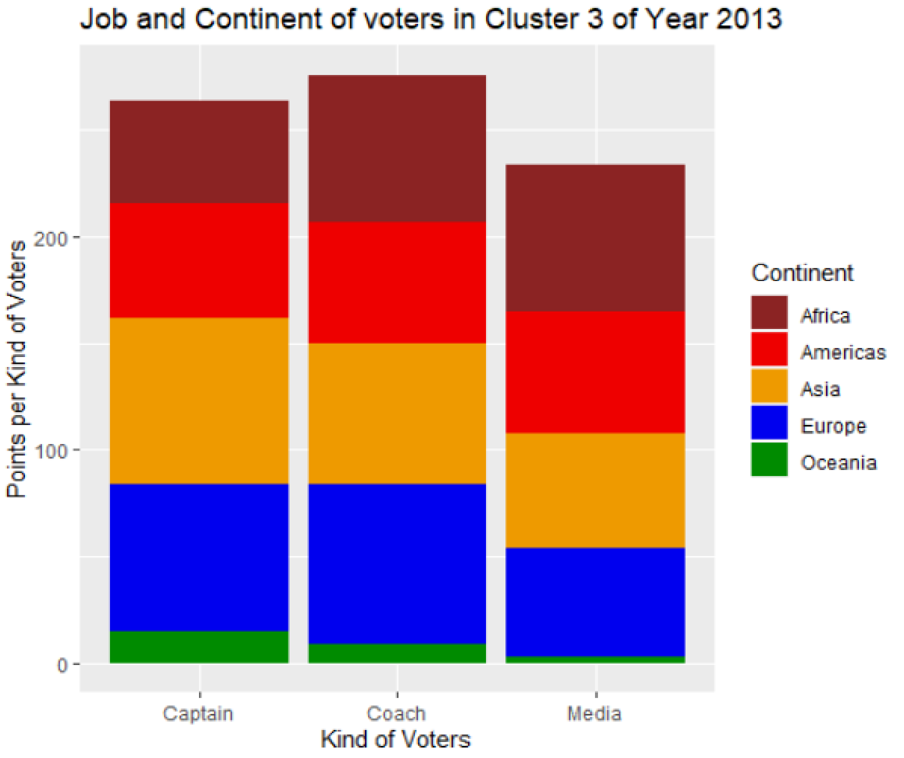


Figure B.50 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 3, Year 2013.

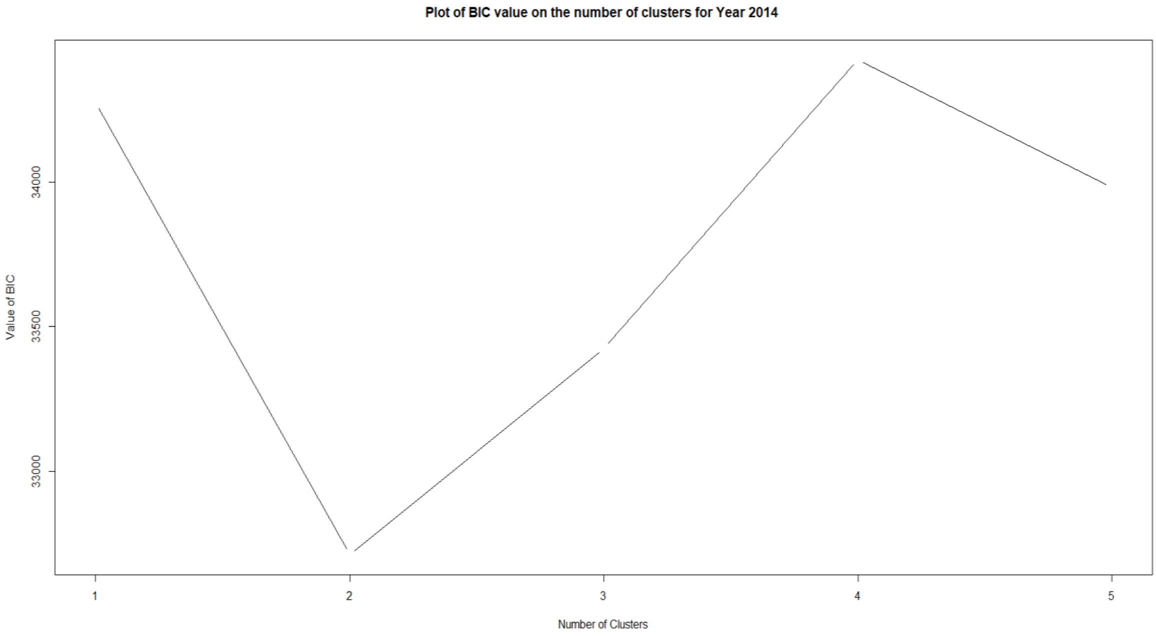


Figure B.51 : Plot of the BIC value on the different number of components, Year 2014.

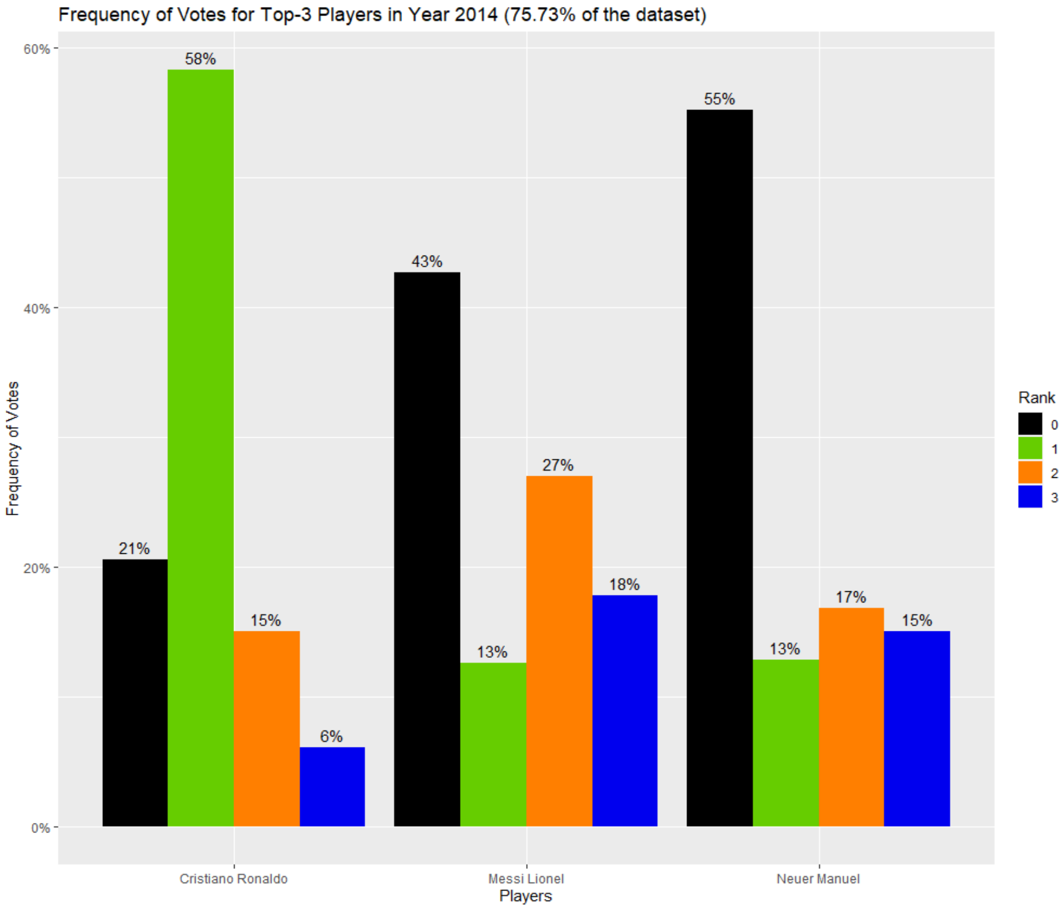


Figure B.52 : Bar plot, representing the percentage of places that Cristiano Ronaldo, Iniesta Andres, Lionel Messi and Xavi have been ranked in Cluster 1, Year 2014.

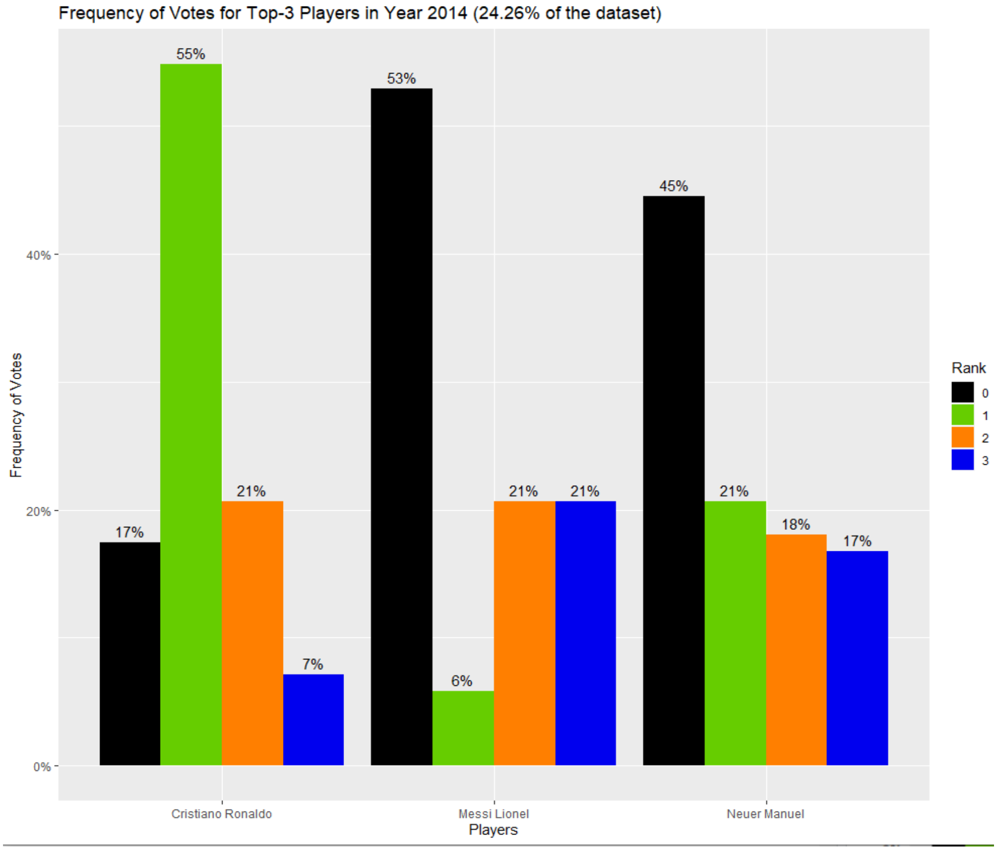


Figure B.53 : Bar plot, representing the percentage of places that Cristiano Ronaldo, Iniesta Andres, Lionel Messi and Xavi have been ranked in Cluster 2, Year 2014.

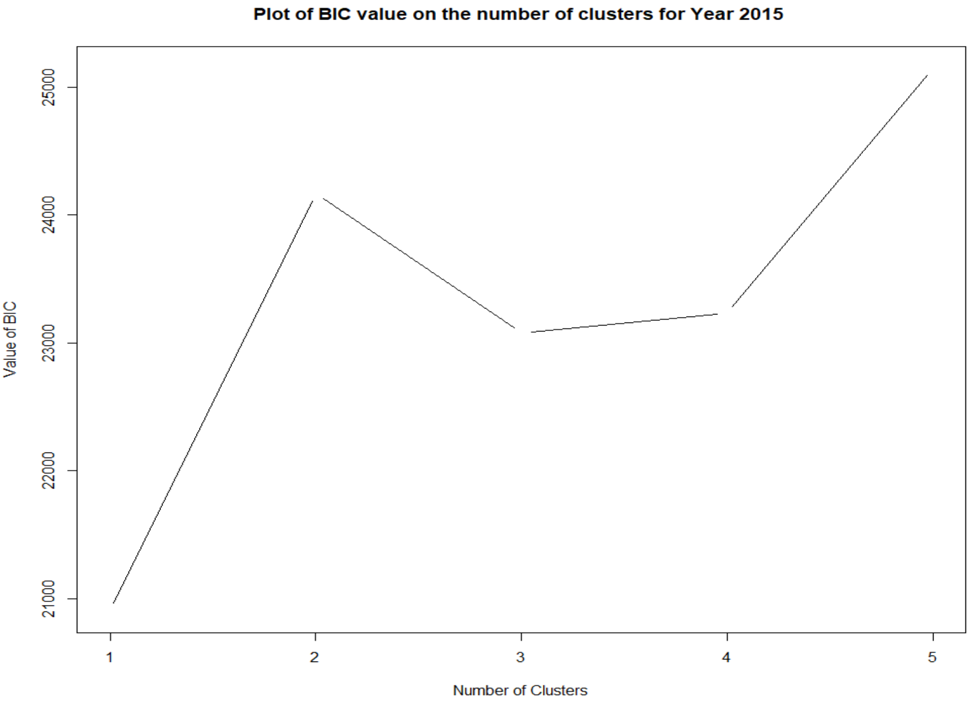


Figure B.54 : Plot of the BIC value on the different number of components, Year 2015.

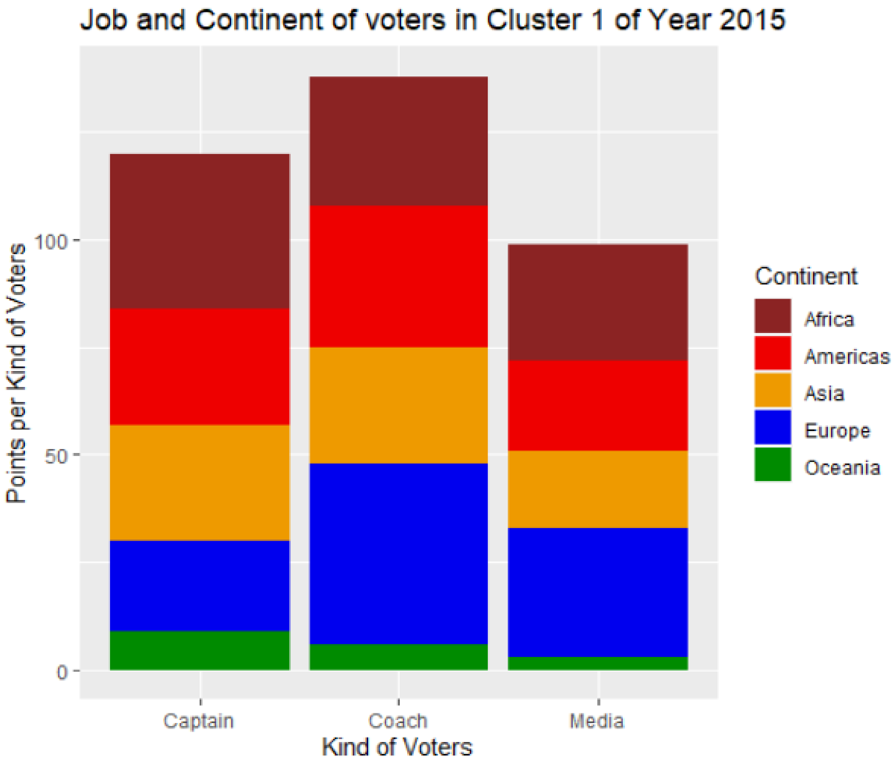


Figure B.55 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 1, Year 2015.

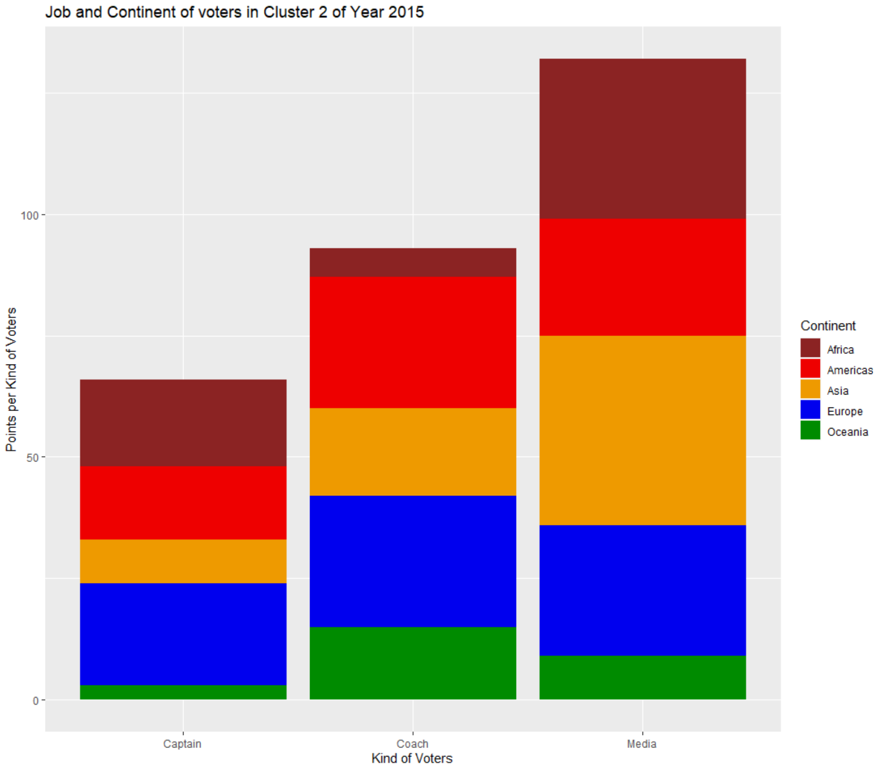


Figure B.56 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 2, Year 2015.

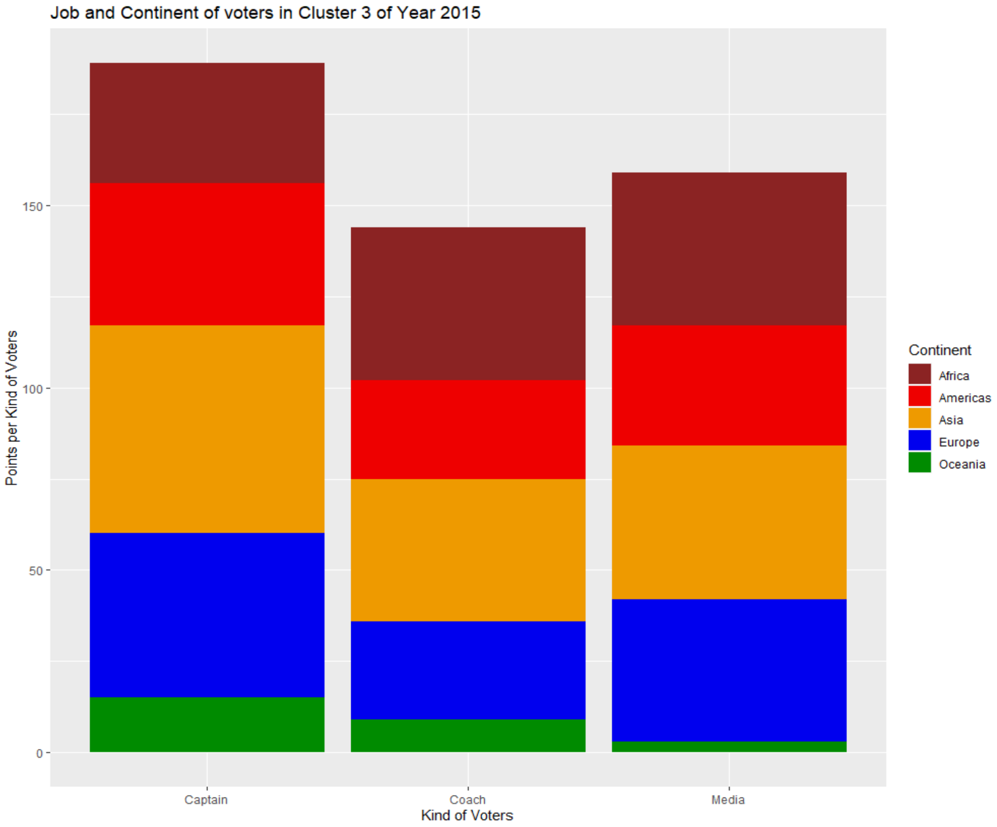


Figure B.57 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 3, Year 2015.

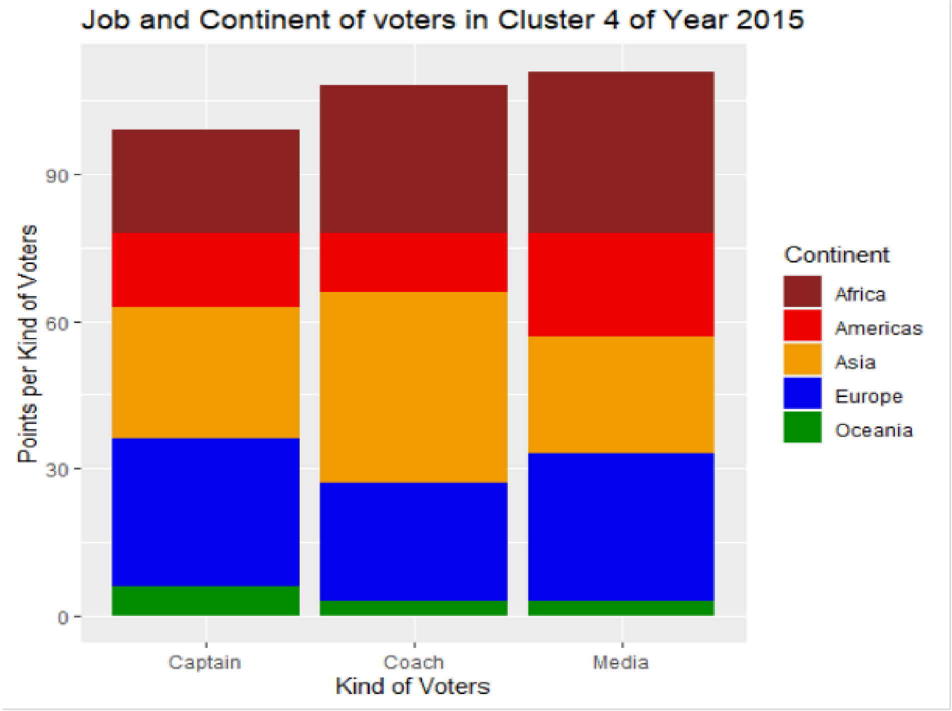


Figure B.58 : Stacked bar plot, representing the job and the continent where the voters come from in Cluster 4, Year 2015.

BIBLIOGRAPHY

[1] Alexey Raskin (2014). Comparison of Partial Orders Clustering Techniques, National Research Nuclear University MEPhI.

[2] Alvas College of Education (2018). The PAM Clustering Algorithm, Notes, pages : 1 – 4

[3] Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94, 443–458.

[4] B. Babington Smith. Discussion of Professor Ross's paper. *J. Roy. Statist. Soc. B*, **12**:53–56, 1950.

[5] Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs, I. The method of paired comparisons. *Biometrika*, 39:324–345.

[6] Buttigieg PL, Ramette A (2014). A Guide to Statistical Analysis in Microbial Ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol Ecol*, 90, 543–550.

[7] Beckett L (1993). Maximum likelihood estimation in Mallow's model using partially ranked data. *Flinger M., Verducci J. (Eds.), Probability Models and Statistical Analyses for Ranking Data, Springer (1993),* pp. 92 – 107.

[8] Bray RJ, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr* 27: 325–349.

[9] Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review, 58*(6), 413–423.

[10] Cabilio Paul, Tilley Jessica (1999). Power calculations for tests of trend with missing observations. *Environmetrics*, 10, 803 – 816.

[11] Caron Francois, Doucet Arnaud. Efficient Bayesian Inference for Generalized Bradley-Terry Models. *Journal of Computational and Graphical Statistics, Taylor & Francis*, 2012, 21 (1), pp.174-196.

[12] Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, R. A. Kimball, & S.B. Nerlove (Eds.). *Multidimensional scaling : Theory and applications in the behavioral sciences, Volume I : Theory*. New York : Seminar Press.

[13] Christophe Biernacki, Julien Jacques (2012). Model-based clustering for multivariate partial ranking data. *Journal of Statistical Planning and Inference, Elsevier*, 149, 201-217.

[14] Christophe Biernacki, Julien Jacques (2013). A generative model for rank data based on an insertion sorting algorithm. *Computational Statistics and Data Analysis*, 58, 162 – 176.

[15] Christophe Biernacki, Julien Jacques,Quentin Grimonprez (2014). Rankcluster: An R package for clustering multivariate partial rankings. *The R Journal, R Foundation for Statistical Computing*, 6 (1).

[16] Coombs, C. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57, 147 – 158.

[17] Critchlow, D. E., Fligner, M. A., Verducci, J. S. (1991). Probability models on rankings. *Journal of Mathematical Psychology*, *35*, 294–318.

[18] Damien de Graav. Fifa Ballon d’or voting method. Thesis at Economics and Business Economics, University of Rotterdam.

[19] Daniels, H.E. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society*, B, 12, 171-181.

[20] Dempster A. P., Laird N. M., Rubin D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological),* Vol.39, No.1., pp. 1 – 38.

[21] Diaconis P (1988). Group Representations in Probability and Statistics. Institute of Mathematical Statistics, Hayward.

[22] Fligner, M. A., & Verducci, J. S. (1986). Distance based rank models. *Journal of the Royal Statistical Society* B, 48, 359–369.

[23] Fligner, M. A & Verducci, J. S. (1988) Multistage Ranking Models. *Journal of the American Statistical Association*, 83:403, 892-901.

[24] Fritsch Arno, Ickstadt Katja (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4 (2), 367 – 391.

[25] Gilles Celeux, Gérard Govaert (1989). Clustering criteria for discrete data and latent class models, Research Report.

[26] Gopindra S. Nair, Chandra R. Bhat, Ram M. Pendyala, Becky P.Y. Loo, William H.K. Lam (2019). On the use of Probit based models for ranking data analysis. *Transportation Research Record*, 2673 (4), 229 – 240.

[27] Gormley, I. C., & Murphy, T. B. (2008). Exploring voting blocs with the Irish electorate : A mixture modeling approach. *Journal of the American Statistical Association*, 103, 1014 – 1027.

[28] Guy Brock, Vasyl Pihur, Susmita Datta, Somnath Datta (2008). clValid : An R package for Cluster Validation. *Journal of Statistical Software*, 25 (4).

[29] Haugh Martin (2015). The EM Algorithm. *Machine Learning for OR & FE*, Department of Industrial Engineering and Operations Research Columbia University.

[30] Ioannis Caragiannis, Ariel D. Procaccia, Nisarg Shah (2008). A Uniquely Robust Voting Rule, Carnegie Mellon University.

[31] Ilker Yildirim (2012). Bayesian Inference : Gibbs Sampling. University of Rochester.

[32] Joe H (2001). Multivariate Extreme Value Distributions and Coverage of Ranking Probabilities. *Journal of Mathematical Psychology*, 45(1), 180–188.

[33] John Guiver, Edward Snelson (2009). Bayesian inference for Plackett-Luce ranking models. *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, 377 – 384.

[34] Jose A. Lozano, Ekhine Irurozki (2012). Probabilistic Modeling on Rankings, *Intelligent Systems Group*, The University of the Basques Country.

[35] Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

[36] Karlis Dimitris (2016). Notes on *Statistics for Business Analytics II, Part V, Clustering*, MSc in Business Analytics, Athens University of Economics and Business.

[37] Kidwell, P., Lebanon, G., & Cleveland, W. S. (2008). Visualizing incomplete and partially ranked data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6), 1356 – 1363.

[38] Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1-27.

[39] Lee Paul & Yu Philip (2012). Mixtures of weighted distance – based models for ranking data with applications in political studies. *Computational Statistics & Data Analysis*, 56, 2486 – 2500.

[40] Luce, R. D. (1959). *Individual choice behavior*. *A Theoretical Analysis*. New York: Wiley.

[41] Ludwig M. Busse, Peter Orbanz, Joachim M. Buhmann (2007). Cluster Analysis of Heterogeneous Rank Data, *ICML '07 : Proceedings of the 24th international conference on Machine Learning*, 113 – 120.

[42] C. L. Mallows. Non-null ranking models. *Biometrika*, 44:114–130, 1957.

[43] Marden, J. I. (1995). *Analyzing and modeling rank data*. New York: Chapman Hall.

[44] Marley, A. A. J (1968). Some probabilistic models of simple choice and ranking. *Journal of Mathematical Psychology*, 5, 311 – 332.

[45] Mayer Alvo, Philip L.H Yu (2014). *Frontiers in Probability and Statistical Sciences : Statistical Methods for Ranking Data*, Springer.

[46] McCullagh, P (1993a). Models on spheres and models for permutations. In M. A. Flinger & J. S. Verducci (Eds.), *Probability models and statistical analyses for ranking data* (pp. 278 – 283), New York : Springer.

[47] McFadden, D. (1978). Modelling the choice of residential location, in A. Karlquist et al. (ed.). *Spatial interaction theory and residential location*, North-Holland, Amsterdam, pp. 75 - 96.

[48] Melnykov Volodymyr, Maitra Ranjan (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, Volume 4, 80 – 116.

[49] Mollica Cristina, Tardella Luca (2016). Bayesian Plackett-Luce Mixture Models for Partially Ranked Data. *Psychometrika*, 82(2), 442-458.

[50] Mollica Cristina, Tardella Luca (2016). PLMIX : An R package for modeling and clustering partially ranked data. *Journal of Statistical Computation and Simulation*, Volume 90, 925 – 959.

[51] Murphy, T. B., & Martin, D. (2003). Mixtures of distance-based models for ranking data. *Computational Statistics and Data Analysis*, 41, 645–655.

[52] Newcomb S (1886) A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics* 8:343-366.

[53] Paolo Delle Site, Karim Kilani, Valerio Gatta, Edoardo Marcucci, André de Palma (2018). *Estimation of Logit and Probit models using best, worst and best-worst choices,* hal-01953581.

[54] Partha Deb (2008). *Finite Mixture Models*, Summer North American Stata Users, Stata Users Group.

[55] Philip L. H. Yu, K. F. Lam, S. M. Lo (2005). *Factor analysis for ranked data with application to a job selection attitude survey*. *Journal of the Royal Statistical Society Series A, Royal Statistical Society*, vol. 168(3), 583 – 597.

[56] Picard F. (2007). An introduction to mixture models. *Statistics For Systems Biology Group*, Research Report No.7 .

[57] Plackett, R. L. (1975). The Analysis of Permutations. *Journal of the Royal Statistical Society Series C, Royal Statistical Society*, vol. 24(2), pages 193-202.

[58] Qinpei Zhao, Ville Hautamaki, Pasi Franti (2008). *Knee Point Detection in BIC for Detecting the Number of Clusters*, ACIVS '08: Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems, 664 – 673.

[59] Qian Zhao, Yu L. H. P (2019). Weighted Distance-Based Models for Ranking Data, Using the R Package rankdist. *Journal of Statistical Software*, 90(5), 1-31.

[60] Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsiy, P. (2007) *Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion).* In Bayesian Statistics 8 (eds J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 1–45. Oxford: Oxford University Press.

[61] Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar Erik Vee (2006). Comparing Partial Rankings. *SIAM Journal on Discrete Mathematics*, pages: 20 (3) : 628 – 648.

[62] Schulman, R. S. (1979). A geometric model for rank correlation. *The American Statistician*, 33(2), 77 – 80.

[63] Seyed Mohammad Razavi Zadegan , Mehdi Mirzaie, Farahnaz Sadoughi (2012). Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge – Based Systems*, Volume 39, 133 – 143

[64] Spiegelhalter D. J., Best N. G., Carlin B. P., A. van der Linde (2002). Bayesian measures of complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B.* 64, Pp. 1 – 34.

[65] Spiegelhalter, D. J., N. G. Best, B. P. Carlin, A. van der Linde (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society*, B(76), 485 – 493.

[66] Srinath Sampath, Joseph S. Verducci (2012). " Is there a particular consensus ordering between rankings ? " *Proc Am Stat Assoc*. 2012 August ; 2012: 2941–2947.

[67] Steven Holland. *Data Analysis in Geosciences*, Lecture Notes, University of Georgia.

[68] Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286.

[69] Thompson, G. L. (1993a). Generalized permutation polytopes and exploratory graphical methods for ranked data. *The Annals of Statistics*, 21, 1401 – 1430.

[70] Thorndike, Robert L. (1953). “ Who belongs in the family ? ”. *Psychometrica*, 18 (4).

[71] Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17, 401-419.

[72] Turner, Heather L, van Etten, Jacob, Firth, David, Kosmidis, Ioannis (2018). Modelling rankings in R: the PlackettLuce package. arXiv preprint arXiv:1810.12068, 2018.

[73] Weksi Budiaji, Friedrich Leisch (2019). Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. *Algorithms*, 2019, 12, 177.

[74] Werry Priy, Kaptein Rianne (2016). *Clustering Ordinal Survey Data in a Highly Structured Ranking*, Research Paper BA, Vrije Universiteit Amsterdam.

[75] Wikipedia contributors. (2020). Ballon d'Or. *Wikipedia The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Ballon_d%27Or&oldid=962637702.

[76] William M. Rand (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66:336, 846-850.

[77] Yellott, J. I., Jr. The relationship between Luce’s Choice Axiom, Thurstone’s theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology*, 1977, 15, 109 – 144.

[78] http://eacharya.inflibnet.ac.in/data-server/eacharya-documents/53e0c6cbe413016f23443704_INFIEP_33/93/LM/33-93-LM-V1-S1__kmedoids.pdf (2019)