QUANTITATIVE EVALUATION OF CLUSTERINGS FOR SEGMENTATION:

AN APPLICATION TO THE BANKING SECTOR

by

Ana Isabel Quinteiro Lopes Gonçalves Loureiro

Thesis of MADSAD on Data Mining

Supervised by
Doctor Carlos Manuel Milheiro de Oliveira Pinto Soares
Doctor Peter van der Putten

2012

I was born in Viseu - Portugal in 1988, where I lived until I went to the university in OPorto. I studied economics during my bachelor and then started my master in Data Mining and Decision Support Systems.

I was part and spent most of my free time working for ten year on a scout group. This experience helped me growing and learning about team work.

When I was asked what did want to study at college my choice for economics appeared as something natural in my life, even if without a true explanation. Data Mining did not come so naturally, but in the today's world, where everything is transformed in a dataset, the capacity of learning something from that did look like an interesting challenge.

Going to Leiden - Netherlands was something that I had never thought about. However, when the possibility came the challenge was so big and appeared like such a good idea that I simple could not avoid.

Acknowledgments

Hopping not to forget anyone!

First of all I want to thank my supervisor Carlos Soares for the help on my thesis and most of all for the possibility of going to Leiden. For the same reasons, I want to thank Petter van der Putten for accepting to supervise me in Leiden, making it the possibility for me to go there.

Thinking about all the Leiden experience, as work and life experience, I have now to thanks a lot of people. First of all to Geraldina Ribeiro for sharing with me these three months and helping to make this an unique and unforgettable experience. To Ricardo Cachucho for all the help and guidance before and during our stay in the Netherlands. To Arno Knobbe for letting me use his offices even without any obligation. For the same reason, I need to thank to Joaquin Vanschoren, Marvin Meeng, Michael Mampaey, Rob Konijn, Shengfa Miao, Siegfried Nijssen, Ugo Vespier and Wouter Duivesteijn. Thank you all for the help and for the knowledge sharing, transforming my short passage through LIACS in something so good and helpful. Out of LIACS there are more people that I would like to thank. However, since I do not have much space, I would like to thank everyone that made part of my life during these three months, with a special reference to Jimita Prashant for all the good moments.

To all my friends in Portugal, for supporting me in the decision of going to Netherlands and for being there in all the good and bad moments.

Last but definitely not least, to my family. My parents, António and Manuela, for all the support that they gave me during all my life, making it possible to achieved all the things that I wanted and also for not cutting my dreams short. To my brothers, Miguel and Nuno, for making part of my life and for sharing a lot of important moments, always trying to understand me. To the rest of my family, that has always been present in my life, specially to those who have helped me to accomplish my dream of finishing my thesis in the Netherlands and with that helping me to grow in several aspects. With a special reference to Andreia and her “4 millimeters”.

Resumo

Um *cluster* é um grupo de objectos que têm quase as mesmas características. Avaliação de *clusters* é um grande problema neste campo de estudos. Várias tentativas foram feitas com o objectivo de superar este problema. Existem três formas de abordar este problema: interna, externa e relativa. A interna usa medidas que estão unicamente relacionadas com os dados em estudo. A externa usa classes para tentar resolver o problema. A relativa usa algum tipo de critério para escolher a melhor divisão. O último caso é com que vamos trabalhar, usando a ideia do marketing sobre o que é um bom *cluster*.

O Marketing tem com objectivo atrair e manter os clientes. Para isso é importante conhece-los e satisfaze-los. A segmentação ajuda neste objectivo. Seis critérios definem o que é um bom *cluster* na perspectiva do Marketing: *identifiability*, *accessibility*, *substantiality*, *responsiveness*, *stability* e *actionability*. Contudo estes critérios estão somente formalmente definidos. O objectivo final deste trabalho é encontrar uma forma de transformar estes critérios formais em medidas que nos permitam avaliar quantitativamente o problema.

As medidas construídas analisam a definição formal de três dos critérios. Ao juntar as definições do marketing com as medidas, queremos não só quantificá-los, mas também explicar a melhor divisão usando os resultados. Isto foi algo interessante que se alcançou com as nossas medidas, quando as mesmas nos permitem usar as definições do marketing para explicar os resultados quantitativos.

Abstract

A cluster is a group of objects that have almost the same characteristics. Clustering evaluation is a big problem in this field of studies. Several attempts were made in order to overcome the problem. Three ways exist to try to approach the problem: internal, external and relatively. The internal approach use measures that are only related with the data in study. The external way use labels to try to solve the problem. The relatively way use some kind of criterion to choose the best division. This last case is the one we are going to work with, using the marketing idea about what a good cluster is.

Marketing has as final goal attract and retain costumers. For that it is important to know and satisfy them. Segmentation helps in this goal. Six criterions define what a good cluster is in the marketing perspective: identifiability, accessibility, substantiality, responsiveness, stability and actionability. However these criterions are only formal defined. The final goal of this work is to try to find a way to transform these formal criterions in measures that allow us to evaluate the problem in a quantitative way.

The measures built analyze the formal definition of three of the criterions. In joining the marketing definitions with the measures we want not only to quantify them but also to explain the best division using the results. That is an interesting achievement of our measures when they allow the use of the marketing definitions to explain the quantitative result.

Contents

1 Introduction 1

1.1 Overview 3

2 Segmentation and Clustering 4

2.1 Description of segmentation in the marketing context 4

2.2 Illustration of segmentation applications 5

2.3 Segmentation methods (manual and automatic) 7

2.4 Segmentation evaluation 15

2.5 Clustering: Formal definition of the problem 17

2.5.1 Hierarchical method 18

2.5.2 Model-based method 20

2.5.3 Partitional method 21

2.6 Clustering evaluation 22

2.7 Internal validation 23

3 Quantitative Measures to Evaluate Clustering for Segmentation 26

3.1 Criteria 26

3.2 Substantiality 27

3.2.1 Previous work 27

3.2.2 Proposed measure 27

3.2.3 Illustrative examples 30

3.3 Accessibility 38

3.3.1 Previous work 38

3.3.2 Proposed measure 38

3.3.3 Illustrative examples 40

3.4 Identifiability 44

3.4.1 Previous work 44

3.4.2 Proposed measure 44
3.4.3 Illustrative examples 45
3.5 Overview 46
4 Experimental Results 49
4.1 Exploratory data analysis 49
4.1.1 Dataset 49
4.1.2 Correlation 51
4.1.3 Subgroup discovery 52
4.1.4 Data preparation 56
4.1.5 Preliminary clustering experiments 57
4.2 Testing the measures 61
4.2.1 Comparing results 69
5 Conclusion 71
5.1 Review 71
5.2 Future work 73

Ilustration index

2-1 First example data 7
2-2 Plot of the first example data 8
2-3 Division into two groups 9
2-4 Division into four groups 9
2-5 Division into three groups 10
2-6 Division into three groups 10
2-7 Plot of a random example 11
2-8 Second example data 12
2-9 Division into two groups 13
2-10 Division into three groups 14
2-11 Division into four groups 14
2-12 Dendogram to the real data with Euclidean distance and ward method 19
2-13 Different results to the real data to understand the differences in the hierarquical method 20
2-14 Results of the model-based method to the real data 21
2-15 Different results to the real data to understand the differences in the partitional method 22
3-1 First example to explain the substantiality 30
3-2 Percentages for the population problem 31
3-3 Percentages for the high income clients problem 31
3-4 Scores for three groups 31
3-5 Scores for two groups 32
3-6 Second example to explain the substantiality 33
3-7 Division into three groups (graphic) 3-8 Division into three groups (table) 33
3-9 Division into two groups (graphic) 3-10 Division into two groups (table) 34
3-11 Division into four groups (graphic) 3-12 Divion into four groups (table) 34
3-13 Percentages for the population problem 35
3-14 Scores for two groups 35
3-15 Scores for three groups 35
3-16 Scores for four groups 36
3-17 Comparing results table 37
3-18 Example for explaining the accessibility 41
3-19 Cross table to the division into two groups 42
3-20 Averages of the spending of the groups 3-21 Averages of the spending of the groups 42
3-22 Comparing results table 43
3-23 Variance of each variable in each group 45
3-24 Comparing results table 46

3-25 Overview 47
4-1 Age division 4-2 Spending division 50
4-3 Age and spending scale division 50
4-4 Division of the global spendings of people 51
4-5 Correlation between the variables 52
4-6 Target type = numeric; quality measure = Z-score 53
4-7 Clients division 53
4-8 Target type = numeric; quality measure = Z-score; refining the search 54
4-9 Clients division 54
4-10 Target type = nominal – spend.scale=5; quality measure = WRAcc 55
4-11 ROC curve 56
4-12 Dendrogram for the real data 57
4-13 Description of the groups using hierarquical method 58
4-14 Comparing table of hierarquical method results 59
4-15 Description of the groups using model-based method 59
4-16 Description of the groups using partitional method, considering 2 groups 60
4-17 Description of the groups using partitional method, considering 9 groups 61
4-18 Division of the population with k-means, considering 2 groups 62
4-19 Division of the population with k-means, considering 3 groups 62
4-20 Division of the population with k-means, considering 4 groups 62
4-21 Division of the population with k-means, considering 5 groups 63
4-22 Percentages to the population problem 4-23 Percentages to the high income problem 64
4-24 Substantiality results 64
4-25 Modified table to evaluate the substantiality 65
4-26 Results to the substantiality according with the modified table 65
4-27 Accessibility results 66
4-28 Averages of spending with and without "food" variable for two groups 66
4-29 Averages of spending with and without "food" variable for three groups 66
4-30 Averages of spending with and without "food" variable for two groups 66
4-31 Averages of spending with and without "food" variable for three groups 67
4-32 Results to the identifiability 67
4-33 Description of the population considering two groups 68
4-34 Comparing results 69
4-35 Comparing results 69

1 Introduction

All the companies want to approach their clients with offers that interest them, thus, helping the company to improve its profits (David Jobber 2009). The final goal of marketing is to attract and retain clients (David Jobber 2009). To achieve this, the company needs to know the clients and their needs in order to satisfy them (Smith 1956). Knowing every customer individually is impossible in most businesses, simply because there are too many of them. However, the behavior of customers can typically be typified. Segmentation is the division of the clients in groups with the ultimate goal of building a strategy that is used to approach clients with the right offer (Pratter 1997).

One approach to segmentation consists of analyzing the data about the clients and their behavior, to find patterns that represent interesting segments (Liu et al. 2010). Different clustering techniques can be used for this task. Therefore, many different clusterings (or segmentations) can be obtained for the same set of clients. So the question is which one is the best (Maulik & Bandyopadhyay 2002).

Several approaches have been proposed to evaluate segmentations from a marketing perspective (Halkidi et al. 2002a; Halkidi et al. 2002b). It is possible to find in the marketing theory various criterions to evaluate the segments, however they are theoretical definitions (Wedel & Kamakura 1998). The subjectivity of these criterions complicates their quantification, further making difficult that after the quantification the application can be general, meaning that the measures can be used in different problems. The criteria define theoretically which characteristics a segment should have to be considered a good segment in the marketing perspective. However, they are hard to quantify and, thus, are not very helpful to select the best from a large number of clusterings. The challenge addressed in this project is how to quantify the quality of clusterings from a marketing perspective, to support that decision. That way we can transform this decision into an automatic and objective process.

The criteria that will be addressed are: substantiality, accessibility and identifiability. The first criterion is related with the size of the segment. A segment must be large enough to be useful. Previously this criterion was studied while trying to obtain groups not too different in terms of number of clients. However this is an approach that is

arguable in terms of marketing. We do not want a group that is too small because a marketing campaign has costs. On the other hand we do not want to judge a group without looking to the clients in it. It may be interesting to consider a small group if it contains the most important clients to the company. We propose a measure that takes into account both the size of the clusters as well as the value of the clients it contains.

The second criterion, accessibility, tries to give some information about the type of campaign the company should do to reach the clients in the corresponding group. The previous approach to this criterion was to look at to the compactness of the decision tree that is obtained from the data describing the clients in the group. This analysis does not work well with small clusters. Moreover, we can say that in terms of marketing analysis we do not obtain much information with this measure and it is not simple to compare the results. Our proposal is to study which variables make the clients in a group stay together. Learning if the group is supported by a single variable or by several variables helps to understand which campaign should be targeted at them.

The final criterion, identifiability evaluates how well the company can describe the clients in the group. Differently from the previously criterion, the goal here is not to understand which campaign to address to each group, but really to understand which type of clients are in each group. The previous approach to this problem used the accuracy of a classification model. This has the problem of the influence of the choice of the method used to obtain the model. Our proposal is to look to the behavior of the clients, trying to estimate how heterogeneous they are.

The development of the measures was focused on a particular problem, the segmentation of the customers of a bank. The data contains two types of variables. Variables that allow us to frame the client according with age and amount of spends and variables that show us where the clients spend the money allowing us to know their habits.

The results obtained with the measures proposed in this project are similar to the ones obtained with the measures that were previously proposes (Rebelo et al. 2007). This is not surprising as the two approaches have the same goals. However, our approach is

better suited for marketing as it considers information that is more relevant for that purpose and it is adaptable to different businesses.

1.1 Overview

We will now give a brief idea about all the work. In chapter 2 it will be done a literature review in the segmentation and clustering fields. Section 2.1 presents the concepts of segmentation and marketing. In section 2.2 will be presented two study cases of marketing segmentation. To better explain the complexity of the problem, in section 2.3 it will be shown how complicated it is to do segmentation without automatic methods. In section 2.4 will be presented the six marketing criterions to evaluate the segmentation. Section 2.5 presents the formal definition of the problem. Different methods of cluster will be presented in this section too. Finally section 2.6 is an introduction of the cluster evaluation problem.

In chapter 3 we can find the description of the measures that will be applied, each sections are dedicated to a criterion and is divided in three distinct parts: previous work, proposed measure and illustrative example. Section 3.1 analyzes the substantiality. Section 3.2 is dedicated to accessibility In section 3.3 we can find the analysis of the identifiability, also with the same division.

Chapter 4 will be dedicated to present and study the dataset as well as the application of the measures presented in chapter 3 to the real problem. Section 4.1 has an exploration of the data in study: description of the dataset and variables, subgroup discovery analysis and data preparation. In section 4.2 we can find a review of the measures but applied to a real data. In the previous chapter the data used are controlled. We will also compare our results with results of previous measures used to solve this problem.

Chapter 5 conducts us through a quick passage of all the work and presents the principal conclusions. In this chapter we also have the future work.

2 Segmentation and Clustering

This chapter will be dedicated to present the work already done in the segmentation and clustering fields. To better approach our problem, it is important to have a better knowledge about the definitions of segmentation and marketing. It is also significant to understand why it is so important to do the segmentation as an automatic procedure instead of a manual analysis. In this chapter we will also present the definition of the six marketing criterions: identifiability, substantiality, accessibility, responsiveness, stability and actionability. We will not study all of them, however, it is important to have a global idea of all.

Concerning to clustering, it is not only important to know what it is but also to know the different types of methods to do it. To conduct our work it is also important to know how cluster is evaluated and what is the most common measure. In this chapter these problems will be discussed.

2.1 Description of segmentation in the marketing context

In this section, we will focus in two different points: segmentation and marketing, and then to understand how they work together.

Marketing has as final goal attracting and retaining clients (David Jobber 2009). The idea is not to fool a client, just to sell something now and then forget about him. The idea is to have new clients but then satisfying their needs and keeping them faithful to the company. According with the same author, the cost of attracting a new client is six times higher than maintaining an older one. Therefore, the investment that a company does in that combination needs to be well prepared and studied. For that, the company has to know the clients and their needs, to better satisfy them (Smith 1956). We can assume that satisfied clients will stay, the company will profit and the clients will be satisfied.

Given that concept, it is possible to implement the same campaign for all clients or to make a division. Segmentation will be the division of the clients into groups aiming a

strategy with more proximity between the product and the client, making the marketing campaign more attractive (Pratter 1997).

Marketing segmentation describes the division of a group of people in smaller groups, taking in consideration their characteristics. We want to define targets in order to do a good marketing approach. The goal is to know our client and label him. This way we will have the capacity to apply the better marketing strategy (Rebelo et al. 2006).

The success of a company is in great measure defined by its capacity to obtain and maintain the clients (Radosavljevik et al. 2011). The better way to do it is to know them as well as possible. However a big company cannot do this without help. Marketing segmentation is a good way to achieve this goal. Having satisfied clients is a big step to success. Therefore, marketing segmentation is a very important tool to accomplish this goal and all the ways to precede them are important to the businessmen (Rebelo et al. 2006). With it, we can better know and group the clients having in account their characteristics and needs. The ideal would be to know and to approach every client as an individual person but since that is not possible, the marketing segmentation concept helps us to find an interesting solution to the problem (Pratter 1997).

The study case that will be used in this work shows the importance of marketing segmentation. In the presence of all transactions made by the clients, the bank wants to be able to classify them into groups. With that, it is possible to offer the client a specific product (Dolnicar 2003). Without this tool, in presence of a client, the bank does not know which product is more interesting to a specific client. That way, we are able to do marketing directly to the specific type of client, making him feel more important and not conduct him to boring situations. It is a win win situation.

2.2 Illustration of segmentation applications

In this section we will briefly present two case studies in order to clarify the importance of segmentation. All the study cases that we can read about have one thing in common: the client interest is the most important thing to the company (Radosavljevik et al.

2011). Whether we are reading about beauty products or air travels, the only important thing is: what could we do to have a satisfied client? The process of discovering the different groups and finding the best strategy is what we call marketing segmentation. Another way to see the problem is: I have a product, let us find the people that are interested in that kind of product and do the marketing approach taking those people in account (The Times 2011).

One example of marketing segmentation that we can present here is the case of BIC (The Times 2011). In our days, the market is so competitive that it is impossible to produce hundreds of products in many different categories. In this sense, BIC felt the necessity of reducing the offer of products (that reduction consisted in a significant passage of 9000 to 150 products). To be able to do that, the company had to make a study of the market. They focused their attention in two different types of clients, the retailers and the final client. Knowing the best mix of products that the first could sell and the desires of the second, they were able to undertake such restructuring. Meanwhile, the company used that opportunity to renew the production process having profits with it.

The other case study that will be presented is United Airlines (The Times 2011). In that case, we are talking about providing a service instead of products. The great difference is, in that case, we do not really know why the clients need the service. In the previous case, we knew why a client needed a pen or a lighter. Whereas here, we can have in the same plane one client that is in a business trip and a family taking a weekend off. Those two types of clients can be viewed as an example. They have different needs, require different types of attention, probably they even pay in a different way. All those things are important to have the client satisfied. Offering the clients what they need gives the company their preference and loyalty. So the goal of this company is to identify the different types of clients and how they are divided. To then have products to satisfy their needs, to build the perfect strategy of communication to go to them and distribute the money to develop the different markets.

2.3 Segmentation methods (manual and automatic)

Study the problem of segmentation is very interesting because we can extract knowledge from a very large quantity of data (Flynn 1999). The capacity of the human being to see a certain default hosted in the same data is very reduced in face of the capacity of a computer. So, this possibility is the starting point of the interest of the study problem. To do an exhaustive search for the relevant segments is an impossible task (Kulkar et al. 2003).

To have an idea about the complexity of this problem, we will look at two examples. In both we will imagine that we want to divide the clients of a company in the best way to do the best marketing campaign.

In the first case, we have 20 clients characterized by two variables and they are distributed like showed in figure 2-2.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
V1	4	2	10	8	3	4	1	4	3	8
V2	66	57	72	66	62	73	59	58	60	77

2-1 First example data

	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
V1	3	5	4	6	5	3	5	5	4	2
V2	29	23	26	22	22	26	29	24	29	30

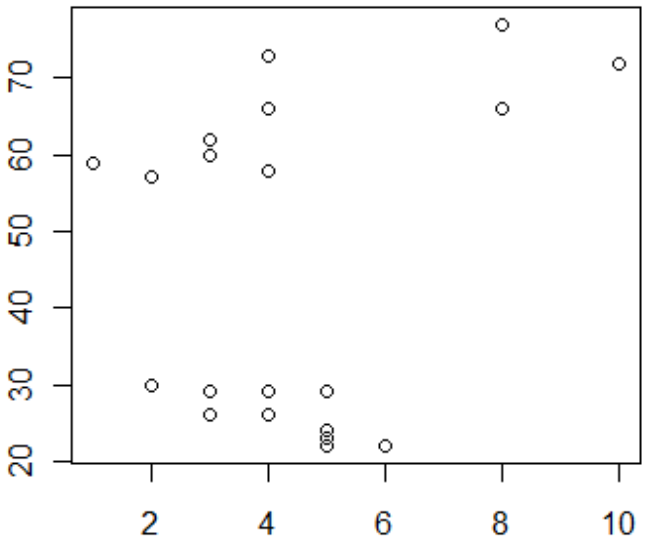
2-1 First example data - continuation

Looking to table 2-1, we are not able to see how many types of clients we have or which client belongs to which partition. There are a lot of processes to separate them. In the next section we will learn more about this.

In figure 2-2, we can observe the data. Intuitively, we can say that there exist three types of clients in that company. However, if we run different algorithms based on the

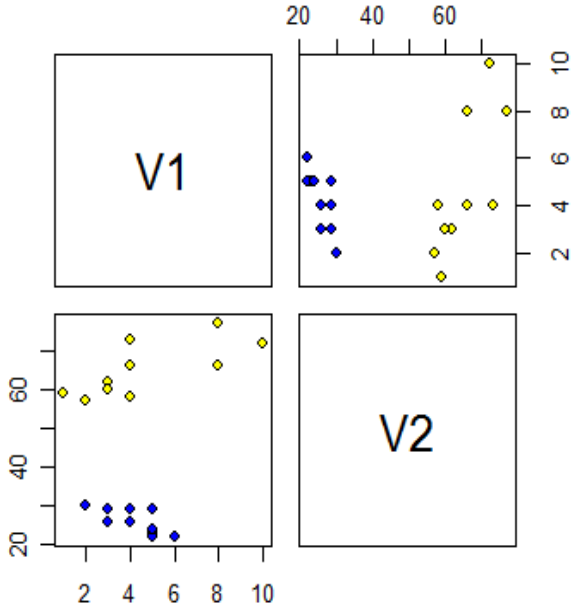
Euclidean distance that is expressed like in the formula below, we can find several different separations.

$$d_{ih} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{hj})^2}$$

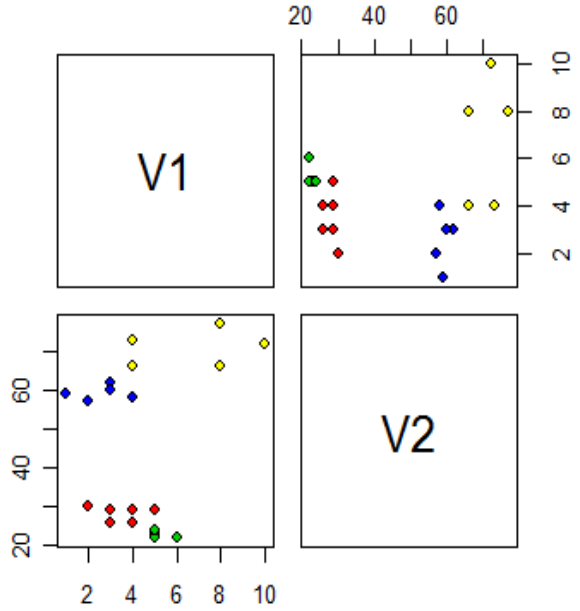


2-2 Plot of the first example data

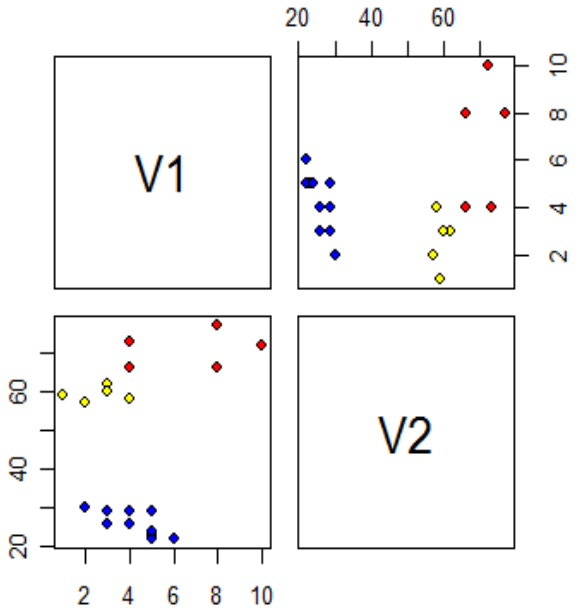
Whether we considered 3 groups but running the algorithm different times or the separation of the dataset in 2 or 4 groups, the main problem will always be the interpretation of the results and the choice of the better division. In the next graphics, we find different results to that problem.



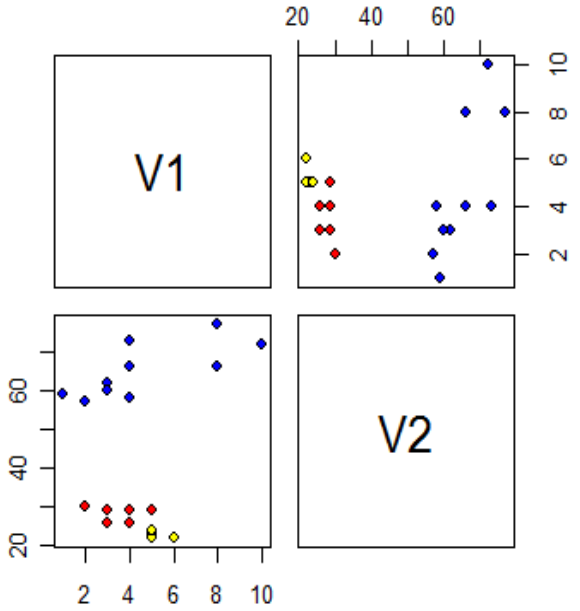
2-3 Division into two groups



2-4 Division into four groups

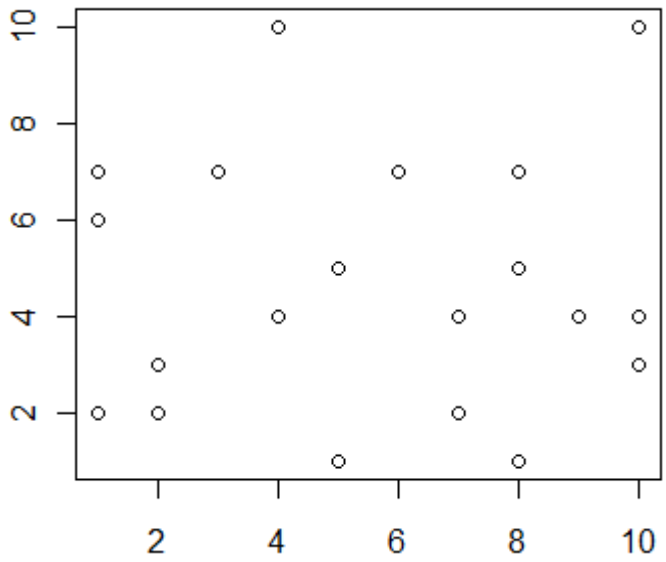


2-5 Division into three groups



2-6 Division into three groups

Figures 2-3 and 2-4 show us the division of the clients into two and four groups, respectively. Figures 2-5 and 2-6 give us the division into three groups. However, the groups are not formed by the same clients. In the presence of those results, we could simply pick one randomly, without any criterion, or we could analyze all of them and try to characterize the clients in every group and then chose the better division. However, that would be a hard work. We are looking for a small dataset with few records and only two variables. But the complexity of the problem increases considerably just by adding one variable and a few more records. Even considering the same number of records and variables, if we look to a problem like in figure 2-7, how will we deal?



2-7 Plot of a random example

In the previous example, intuitively, we could say that a division in three groups was good. In this example, we are not able to say a number of groups that could work. And like mentioned, if we add more records and more variables it starts to be impossible to analyze the problem manually. Even if we want to do it, the amount of time, money and

human resources necessary are so huge that it probably would not compensate. We can have a small idea of this problem by looking at table 2-8.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
V1	1	1	9	9	9	6	4	4	4	10
V2	1	3	6	3	4	1	2	1	9	2
V3	4	2	6	4	3	3	2	3	5	2

2-8 Second example data

	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
V1	4	1	6	7	7	1	2	10	8	8
V2	10	3	5	2	3	7	10	9	5	4
V3	5	4	6	4	3	2	5	3	6	3

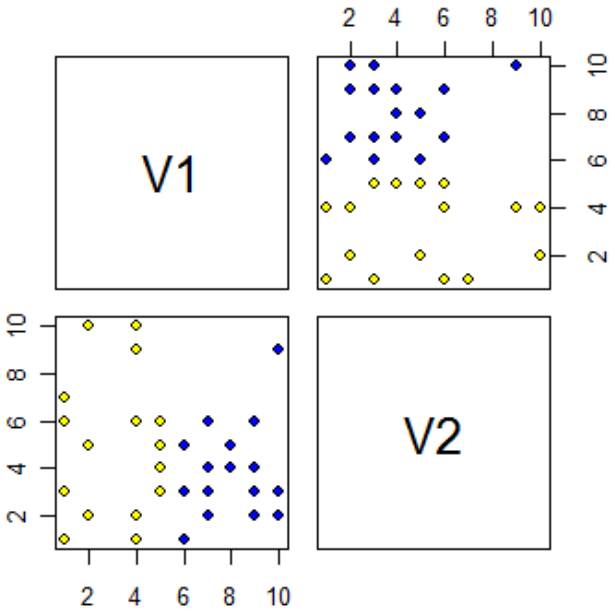
2-8 Second example data - continuation

	21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
V1	7	9	1	6	7	5	5	6	10	5
V2	4	2	6	5	2	4	6	3	3	5
V3	4	6	4	6	3	2	5	6	4	3

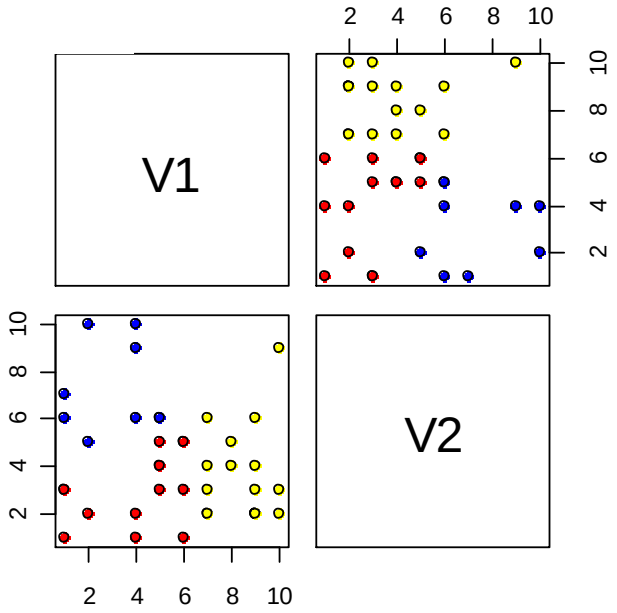
2-8 Second example data - continuation

	31	32	33	34	35	36	37	38	39	40
V1	9	2	5	2	9	5	9	7	5	4
V2	2	2	4	5	2	6	2	6	3	6
V3	2	6	4	6	3	4	2	2	5	2

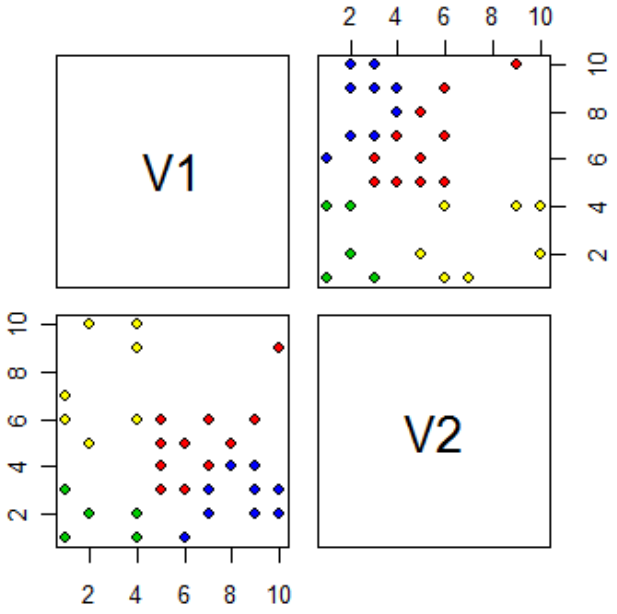
2-8 Second example data - continuation



2-9 Division into two groups



2-10 Division into three groups



2-11 Division into four groups

Figure 2-9 has the division of the second example into two groups, figure 2-10 into three and figure 2-11 into four. We only have three results and we run the algorithm only one time for each number of groups. However, we can see that it is possible to find a rational explanation for choosing one over another for any division that we want. It is possible to chose two groups and to say that in one cluster we have the records with the higher values for the variable one or chose three groups and say that it makes sense to divide the clients with lower values in variable one in two groups.

The idea of this section was to explain how hard it can be to find the better division in the presence of a dataset. We only look for small examples, but even like this it is difficult to answer the question: how many clusters should we consider?

2.4 Segmentation evaluation

In the marketing perspective, there are six criteria to evaluate a segmentation which are: identifiability, substantiality, accessibility, responsiveness, stability and actionability (Wedel & Kamakura 1998). Since a unique criteria to evaluate them does not exist, the authors proposed different measures to use them in a quantitative way (Rebelo et al. 2007). However, at this point, we will just present the formal definition of this criteria and how it could be related with the problem in study.

The identifiability is related with the capacity of well distinguishing the different groups obtained (Wedel & Kamakura 1998). In the study case, that criterion is verified if the division of the clients shows perfectly the differences between the clients in each group. The bank should be able to characterize every group.

The substantiality is the interest of considering a segment based on the number of elements in it (Wedel & Kamakura 1998). In other words, a segment should be considered if the number of elements is significant. In our case, if a group is very small, maybe the cost that the bank will have with direct marketing to them is not justified. However, that is a sensible point since the group could only be constituted for big clients and, in that case, the cost will be more than justified.

The accessibility reflects the ease of creating a marketing campaign that reaches the consumers (Wedel & Kamakura 1998). Looking to our problem we could say that a segment meets this criterion if the bank understands how to approach the group. In other words, it should be easy to identify which are the needs of the clients in the different groups in order to be possible for the company to satisfy them.

The responsiveness is related, like the name suggests, with the future acceptance of the group to the offer that will be made to them (Wedel & Kamakura 1998). In the case of the bank, that criterion is related with the acceptance of the elements of the group to the marketing campaign made directly to them.

The stability represents the importance of time in marketing problems (Wedel & Kamakura 1998). In the study case, for example, the bank needs time to prepare and execute a marketing strategy. If the groups found are not solid, in other words, if the elements of one group easily change to other, the campaign is not going to be effective but the money and the time was already spent.

Finally, the actionability is related with the capacity of the company to handle the groups (Wedel & Kamakura 1998). In our case, the segmentation will be good if the bank has the resources and the capacity to produce and release the perfect marketing campaign to each group. It is different from accessibility because it is related with the capacity of the bank to approach the group and not with the capacity to comprehend the needs of the group.

The analysis that are made in this area are in a great measure a manual process (Rebelo et al. 2006). Like mentioned in the previous section, we can easily obtain a lot of classifications to a dataset. However, to do manual interpretations of the results in order to understand if the criterions are satisfied is a hard work. The main idea is to transform this into an automatic and numeric process.

2.5 Clustering: Formal definition of the problem

A cluster is a group of objects that have almost the same characteristics. In our study case, a cluster will be a group of similar clients (Jain & Lansing 2010). For example, if we are doing the separation by age, one cluster will be the group of the youngest clients. We call cluster analysis to the process through which we obtain clusters, when the number and form of groups are unknown. The quality of a good cluster can be seen when looking for the associations between the elements on the same cluster and the weak links between elements of different clusters (Liu et al. 2010). This association is based on distance measurements between attributes of two objects. It will be the use of these measures that will select the relevant segments. However, this can be a really difficult job since clusters can differ in shape, size and density and the analysis can be even more difficult when it exists noise in the data (Liu et al. 2010).

At this point, it is important to make the distinction between what is and what is not clustering (Jain & Lansing 2010). Supervised classification is not clustering. In this type of classification we have labels to some objects and we learn a model based on that information (Tan et al. 2005). Then we use that model to classify unlabeled examples. Clustering is learning from a data set without knowing any information before we start. In other words, we do not know labels of any example before starting. That is the reason why the selection of the number of clusters is one of the biggest challenges in that field.

Comparing this kind of learning with predictive learning, where the goal is to classify a new example with a label previously defined, we can find a problem, which is the difficulty of defining a measure to evaluate the quality of the structure of clusters. However, the quality of a good cluster, in the end, is related with its capacity to discover unknown patterns in data (Liu et al. 2010).

There are three kinds of clusters: hierarchical, partition and model based methods. It is interesting, for now, to refer the main differences between them. When we use the partition method, k-means being one of the most known, we need to define a-priori the number of clusters that we want to use. That choice could be difficult since most of the times the user does not have a big knowledge of the problem at the beginning, so this is

a disadvantage. This problem does not exist in the hierarchical or the model based method.

2.5.1 Hierarchical method

This method creates groups, without the necessity of the user to define their number a priori and represents them in a dendrogram, which has the same structure of a tree (Flynn 1999). At the root of the tree is the group that contains all the observations. In the leaves are the more specific groups, each leaf containing just one example. In the analysis of the tree, the distance between a node and another one represents the dissimilarity between groups. Therefore, a split in the tree must be made in such cases where this distance is greatest, since we want that the elements of the groups to be as similar as possible within the group and that the differences be between elements of different groups.

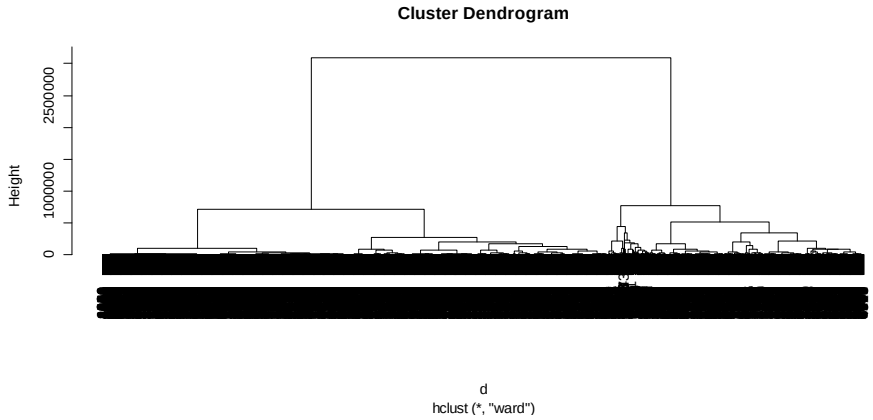
In hierarchical methods, there are two ways to proceed in finding groups. From general to specific or from specific to general (Flynn 1999). In the first case, top-down method, initially all examples belong to the same group and then go by divididing into groups looking at the examples which are the most dissimilar. These divisions will happen as long as there are groups with more than one observation. In the second case, bottom-up method, the procedure is the opposite. Initially each example belongs to a group and will be joining the examples that are more similar to each other until all the examples belong to the same group.

Analysis of similarity / dissimilarity can be done through various techniques (Flynn 1999). The choice of the technique will influence the groups found. Some examples of these techniques are: single linkage, complete linkage, average linkage and linkage wards.

All of them form the group by analyzing which are the examples that have the smaller distance. The difference is in calculating the distance of the groups that are formed and other examples. For example, when the technique used is single linkage distance between groups and the other examples, what is considered is the smallest of the

distances between group members and the example considered, while in complete linkage is considered the largest of these distances (Maimon & Rokach 2010). Although we then choose the shorter distance, the way the table of distances is constructed will influence the analysis. This issue will be discussed later in more detail. At this point, only the influence that this choice can have in groups found is exemplified.

With figure 2-12 it is possible to exemplify this type of method. For obtaining the results present in the dendrogram we will use the Euclidian distance and the ward method. But in the table 2-13, we can see the results with other methods and it is possible to confirm that this choice has an influence on the results.



2-12 Dendogram to the real data with Euclidean distance and ward method

It is impossible to understand which elements belong to which group, but it is clear that the bigger distance is when we have two groups. So, cutting the tree in k=2 (where k means number of groups) we will obtain a group with 1677 elements and other with 3323.

Distance/Method	Ward	Single
Euclidian	2 (1677,3323)	2(4999,1)
Maximum	2 (2211,2789)	2 (4999,1)

2-13 Different results to the real data to understand the differences in the hierarquical method

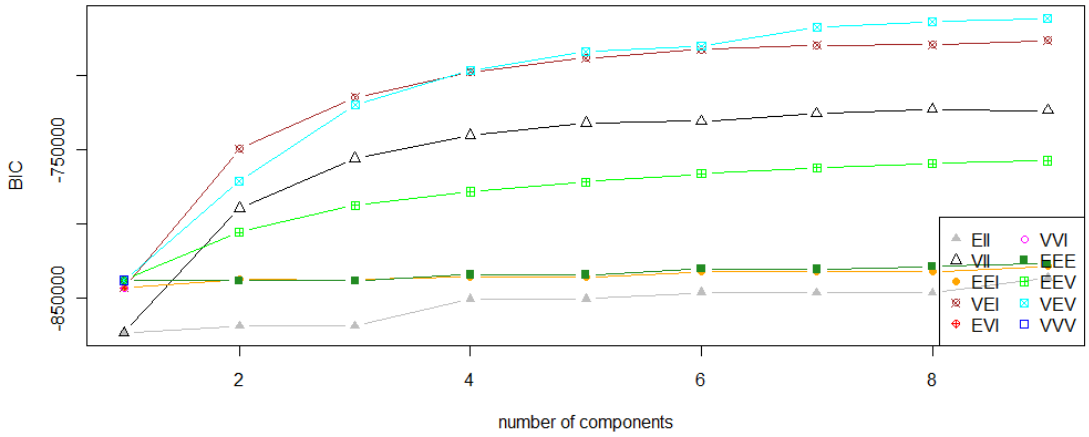
In table 2-13, there are just four combinations of measures, but the difference between the results is already clear. However it is important to try different combinations to be sure of the consistency of the partition that we will be considering.

2.5.2 Model-based method

That type of approach is based on probability models instead on heuristic methods like the previous (Chris Fraley & Adrian E Raftery 2007). The methodology is built based on the idea that the data is formed for a subjacent probability distribution (especial attention to Gaussian) where each component is a cluster (C Fraley & A E Raftery 1998). A maximum-likelihood method is used to form the groups and expectation-maximization is used to reallocate the observations. Having that in account, different models that vary in volume, shape and form are running and in every state a pair of clusters is combined in order to maximize the probability.

After that process a matrix with the BIC values will be calculated and the choice of the model and the number of clusters will be made where this value is higher (Chris Fraley & Adrian E Raftery 2007).

This methodology does not work well in datasets with noise (C Fraley & A E Raftery 1998). However there are several ways to remove the noise of the data and then it is possible to follow the same steps to find a solution.



2-14 Results of the model-based method to the real data

In our case study, as we can see in figure 2-14, the result is to consider 9 components with a model with equal shape.

2.5.3 Partitional method

In this method, the user defines at the beginning how many different groups he wants to find, which is a disadvantage compared to the methods presented previously (Flynn 1999). This method can then work through an exhaustive search or a heuristic search. The first possibility is however very difficult to achieve since the complexity of this problem has an exponential growth.

We will analyze this method from a heuristic method, the k-means.

After setting the desired number of groups, points in the study are associated with the nearest center (Jain & Lansing 2010). After defining these groups, the average of points assigned to each group is calculated again and a new center is set. Points are associated with a new center and this process is repeated until none of the points change.

Though it is possible to prove that this algorithm always converges to a solution, it is not guaranteed to be the optimal solution, since it is a heuristic search method (Jain & Lansing 2010). Another limitation of this algorithm is the fact of being affected by the

choice of the initial configuration. This limitation can be partly overcome by running the algorithm several times and assigning the example to the group in which it was ranked more often.

Looking to the results of the hierarquical method, we will choose 2 clusters to apply to the partitional methodology. But like before, we will try to compare this with other results to have a better idea of how our initial choice influences the final results.

Experience	Partion
1	(797,4203)
2	(4200,800)
3	(4203,797)
4	(797,4203)
5	(4203,797)

2-15 Different results to the real data to understand the differences in the partitional method

With this algorithm we found results very similar between them, as we can see in table 2-15. However, they are very different from the results of the hierarquical method. How to choose between them?

2.6 Clustering evaluation

This question is the main difficulty in the study of clustering. Since this is an unsupervised study, there is not the possibility to compare the results of a model that we can learn with the reality. Therefore, in order to compare different results of clusters, it is very difficult and it turns out to be subjective to say that one clustering is better than another (Halkidi 2001b). We know that the elements in a group should be similar and that the elements should be as different as possible of the members of other groups. That evaluation can be done based on distances. But how to compare two clustering?

Like mentioned before, one of the main objectives is to combine the marketing idea of segmentation with measures that are possible to apply to the cluster's segmentation and that way to be able to evaluate the quality of the cluster.

The validation of a cluster can be done on three ways: internal, relative and external (Halkidi et al. 2002b; Halkidi et al. 2002a). In the first approach the evaluation is made based on the fitting obtained with the clustering compared with the data only. So the measures used to apply this criterion can only be related with the dataset in analysis, because we do not have any further information than the dataset. The second criterion looks for different results of clustering and uses some kind of criterion to choose the best division. That is the idea of this work, to use the marketing criteria to evaluate. That means that in this criterion we will run different clusters methods, or the same cluster method, but with different initial parameters and find the best algorithm and the best number of clusters to consider. The last criterion is related with the label. However, if the real label is available, to study the problem like a clustering problem is not interesting. This last approach is just interesting to discover which algorithm we should use, knowing the number of clusters that we should considerer, since we know the labels. In the two first cases, the validation could help in the choice of the best algorithm as well in the optimal number of clusters that we should considerer.

Another way to see the problem is to look at the stability of the results (Jain & Lansing 2010). In the case of cluster, we can use a measure to evaluate how strong the relation between the members of the same group is. If that connection is strong, we can say that it is a stable cluster because if we run a different algorithm the probability of that cluster remaining together is higher.

2.7 Internal validation

In the study of Liu and others (Liu et al. 2010) we can see the comparison of eleven internal measures. Knowing how clustering works and the final goal of this methodology, internal measures usually try evaluating two different things in a clustering, which are the compactness and the separation. The first one is related with

the idea already presented about the elements of the group that should be as close as possible. It is possible, for instance, to look at the variance and, if this number is small, we can think of the group as a good one. The second thing that we try to evaluate is the separation between the groups, meaning that we expect the groups to be as far as possible of each other. To evaluate this, we can look at the distance between the center of the clusters or at the distance between the closer elements between two groups. According with the same study (Liu et al. 2010), the better measure or, in other words, the measure that could deal with all the problems that the authors studied, like monotonicity, noise or subgroups among others, is S_Dbw (Scatter and Density between clusters) presented by Halkidi (Halkidi 2001a). This measure evaluates the compactness and the separation and sums the results to these two problems turning into a problem of minimization. Comparing with other measures, this has an advantage. Not only summing the results but trying to understand if the separation of two groups is really better than having the two of them forming a unique group. This is done in the part of the separation, measuring the density of the two clusters and comparing with the density of the midpoint. At least for one of them the quality measure needs to be better because otherwise we are not improving our results doing the separation of the groups. The compactness in this measure is evaluated through the variance. Then, like mentioned, it is a problem of minimization, the best number of clusters is when we have the smaller value of S_Dbw which is given by the formula below.

$$Scatt = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{||\sigma(v_i)||}{||\sigma(x)||}$$

The term n_c is the number of clusters, $\sigma(v_i)$ is the variance of the cluster i and $\sigma(x)$ is the variance of a dataset.

$$dens_{bw} = \frac{1}{n_c(n_c-1)}\sum_{\substack{j=1 \\ i \stackrel{\text{def}}{=} j}}^{n_c}\frac{density\ (u_{ij})}{max\{density(v_i), density(v_j)\}}$$

The terms v_i e v_j represent the centers of the cluster i and j respectively and u_{ij} represent the middle point of the line segment defined between the two centers. The density is calculated as $density\ (u) = \sum_{i=1}^{n_{ij}} f(x_i, u)$ where n_{ij}= number of tuples.

With the first formula, we can evaluate the intra class problem and with the second one the inter class problem (Halkidi et al. 2002b). In the end, it is possible not only to sum these two measures but to consider different weights for the two parts.

For this work it could be interesting to consider this measure to compare the results with the ones that we will try to find, according with the marketing theory.

3 Quantitative Measures to Evaluate Clustering for Segmentation

This chapter will be dedicated to the exploration of the three criteria that will be studied in this work: substantiality, accessibility and identifiability. For those criteria, we will present the measures previously used and present our own measure. The measures will be explained and discussed. Each one will also be applied to a small problem to better understand how they work and which are the implications that they have, as well as compared with the previous measures.

3.1 Criteria

The criteria that we are studying were presented, as refered in the previous chapter by Wedel and Kamakura (Wedel & Kamakura 1998). We used this work as reference because the authors gave us a good formal definition of the rules that a good cluster should obey to be considered a good cluster in the marketing perspective. In this sense the six criteria already presented are: substantiality, accessibility, identifiability, responsiveness, stability and actionability. With these criteria we look from the identification of the clients in a group until their reaction to the marketing campaign made directly to them. Having the data in study in account, that will be presented in the next chapter, we are not able to study all of them. We have data about how and where people spend their money. With that we will be able to study three of the criteria. The substantiality that is concerned with the number of clients in a group can be easily studied by us with the data available. The accessibility tries to understand if it is easy to the company to create a marketing campaign that reaches the clients in the groups. In that sense it is possible with our data to try to understand the needs of the clients in order to built the perfect campaign to them. The last criterion in study, identifiability, is related with the ability to distinguish the clients in the different groups. With our data it is possible to answer this question finding a way of producing a characterization of the clients in the groups based in their spending behavior. The other three criteria will not be studied in this work. For study the responsiveness, we would need data related with the acceptance of the clients to a marketing campaign. About the stability, we would need data in different points of time, not only in one point since those criteria study the

importance of time in marketing problems. To build a campaign the company needs time, so it is important that the clients identified in a group stay in that group for a while, otherwise when the campaign is released the groups are already different. To study the actionability we needed data about the availability of the bank to spend in marketing campaigns. We could use data about money, human resources and time, among others.

3.2 Substantiality

3.2.1 Previous work

The first criterion for which we will try to find a measure is substantiality. This problem has been treated before (Rebelo et al. 2007) using maximum and minimum as expressed in the next formulas. The authors proposed two measures, $S_1 = \frac{\max(n_i) - \min(n_i)}{\bar{n}}$ and $S_2 = \frac{\min(n_i)}{\bar{n}}$, where n_i represents the size of the segment i and $\bar{n}$ the average size of the segments. In the first measure, when S_1 decreases, the segments are more balanced and the criterion increases. Concerning the second measure, when S_2 increases the criterion increases too, showing more balanced segments as the previous one.

3.2.2 Proposed measure

Generically, this definition is related with finding groups big enough to be considered, meaning having a significant number of records, as present in the previous chapter. In our problem, like mentioned before, the bank has interest in considering a group if what they spend in the marketing campaign allows improving their results. To do that, we can look at the problem from two sides. One is the number of clients in the group and the other is how much they spend or, in other words, we need to know if the clients in the group are big clients in the context of the bank. Because, like mentioned before, even if the group of the best clients is small, probably it is interesting to consider them from the

marketing point of view. With the previous measures only the first part of the problem was covered. In the previous measures used this second approach was not taken in account. The concern was to find balanced groups, meaning that a group should not be too big in relation to another. However from the marketing point of view it could be interesting, as referred, to consider groups with very different sizes having in account the type of clients in there.

We can see that the first approach to the problem is general, meaning that it is possible to apply it to all problems in study. However, the second part can only be applied if a specific scale is available. In this case, and having in account the problem in study, a scale will classify the clients in terms of income. The measure will only consider the clients with the higher score in that scale, representing the high income clients.

To evaluate this, we will work with two tables. At the first one, the user defines how many people need to be in a group for it to be considered. At the second one, the same is defined but for the income scale. Meaning that the user defines how many high income clients should be in the group for it to be considered. Based on these two tables, a score will be assigned to each group for each problem. The result will be a weighing of the two scores for all groups or zero if at least one of the groups had zero in both scores. We want to maximize this measure once that higher score means best performance.

$P - \ population$

$k - \ clusters$

C= $\bigcup_{i=1}^{k} C_i$

$n_i, i \in \{1, \dots, k\}\ n_i = \#C_i$

$H \subseteq P\ high\ income\ costumers$

The process of segmentation to obtain the high income clients is based on single variable. In our study case, it is especially easy once it is defined by domain knowledge.

$s_i - score\ for\ segment\ i\ in\ the\ population\ problem$

$s_i^h - score\ for\ segment\ i\ in\ the\ high\ come\ costumers\ problem$

$\propto - importance\ of\ the\ high\ income\ segment$

$$Substantiality\ (Sub) = \begin{cases} 0\ \ if\ \exists_{i=1}^{k} s_i^h = 0 \wedge s_i = 0 \\ \propto \sum_{i=1}^{k} s_i^h + (1-\propto) \sum_{i=1}^{k} s_i \end{cases}$$

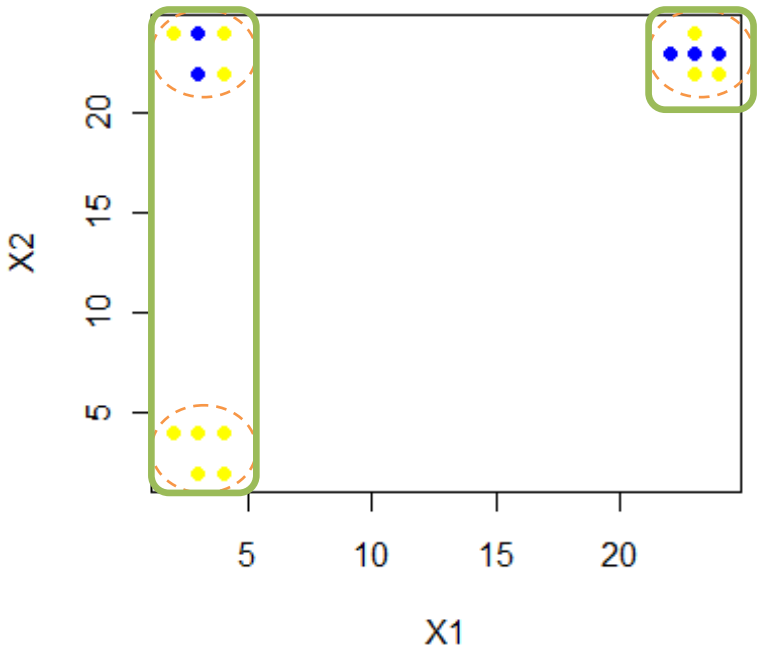
When we are using this measure in other datasets, we can have a few problems. As already mentioned, we need a specific scale to apply the second part of the measure. Also, depending on the problem that we are studying, it may not be wise to take only the higher results of this scale in consideration. The definition of the tables can be a problem too. When we are looking at our measure, a problem that can appear is the fact that we are not explicitly considering the number of clusters in the measure. However, this problem is taken in account in the scores. Since we attribute a score to the group, according with percentages, that problem is immediately covered when we pass from one for two to three groups, for example, at least one of the groups needs to decrease. That means that even if the number of the groups is not considered in the measure, the impact of that number is.

Another aspect that can be discussed is the possibility of obtaining a better classification than when we are considering only one group. However we will see in the next section that it is possible to obtain better results dividing the clients in groups.

3.2.3 Illustrative examples

Let us have in account an example where all the points are perfectly separate and the distribution of rich people helps us to better understand the measure.

In figure 3-1, we can see the data and perfectly understand which groups we can find. The blue points represent the high income clients. So the idea now is to prove that, according with the theory that we have learned, it is better to consider 2 groups than 3. Why? Thinking in the bank perspective, we can see that to consider a group of 5 people (around 31% of the population), where no one has a high income, and to do a marketing campaign for them, it could be too expensive considering the results that they will get in the end.



3-1 First example to explain the substantiality

The tables that we will consider in this example, for the population problem and the high income clients problem, are 3-2 and 3-3 respectively. The weight will be 70%

$(1-\propto)$ for the total of people in the group and 30% $(\propto)$ for the percentage of high income clients.

	0	1	2	3
Total	35%	50%	70%	100%

3-2 Percentages for the population problem

	0	1	2	3
High income	20%	75%	90%	100%

3-3 Percentages for the high income clients problem

Considering three groups, the scores for each group according to tables 3-2 and 3-3 are present in table 3-4.

	G1	**G2**	**G3**
Total	1	1	1
High income	0	1	1

3-4 Scores for three groups

In this case the result to substantiality is:

$$Sub = 2.7$$

Table 3-5 gives us the scores for the division into two groups, according with tables 3-2 and 3-3.

	G1	G2
Total	3	1
High income	1	1

3-5 Scores for two groups

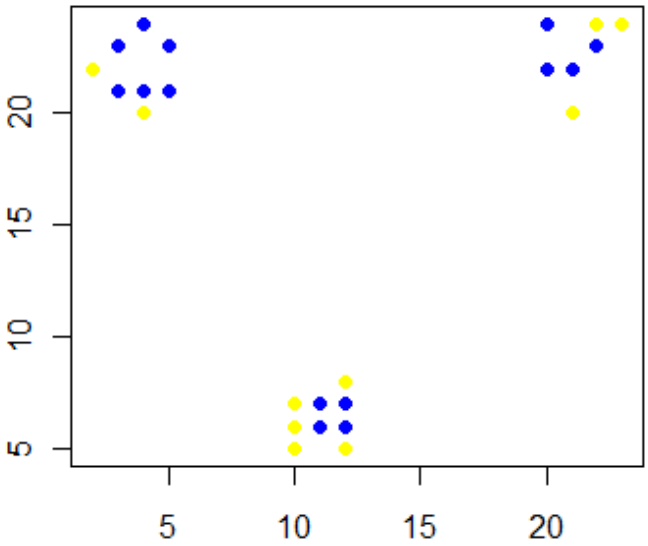
Having in account the scores previously presented, the result for substantiality is:

$$Sub = 3.4$$

We can see that is better to consider 2 than 3 groups once the score is higher. As expected from the brief analysis already made, there is no benefit for the bank to split the clients into three groups since the results are worse for both of the problems addressed in this measure.

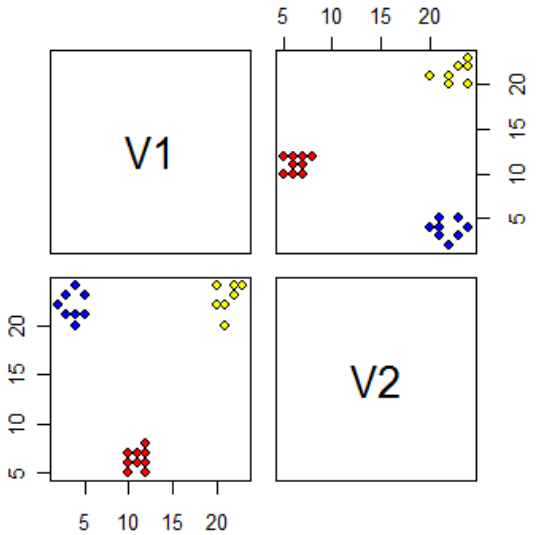
We can also see what happens when we obtain the same score. Let's imagine that in table 3-2 the score 2 began at 80%. In that case the identifiability for 2 groups would be also equal 2.7. In this case it is better to consider only two groups instead of three even if having the same score. The main idea is to have a good approach to the clients but using the minimal amount of resources. So if the score is the same and we can only do two campaigns, rationally we will choose that.

To make sure that the measure really works, we need a bigger example where we can get the inflection point of the measure. For that, we need to create an example where, intuitively, we figured out that three groups are better than two or four. The figure 3-6 represents that example. Like before, the blue points represent the high income clients.



3-6 Second example to explain the substantiality

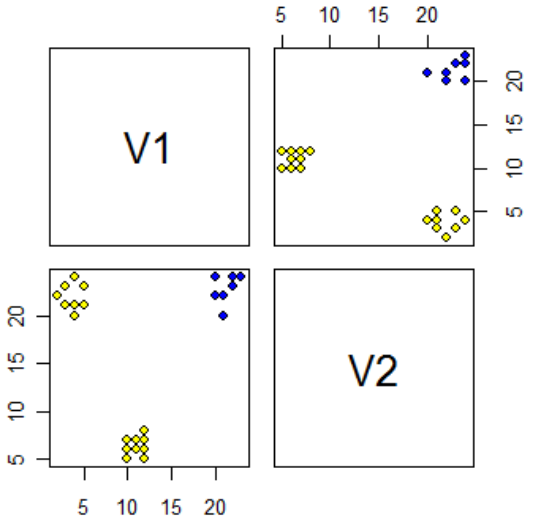
In figures 3-7, 3-9 and 3-11, we can see the division of the clients from the example in figure 3-6 in three, two and four groups respectively. Tables 3-8, 3-10 and 3-12 show us the number of clients with low and high income in each group.



3-7 Division into three groups (graphic)

	Low income	High income
G1 – blue	2	**6**
G2 – yellow	3	**4**
G3 – red	5	**4**

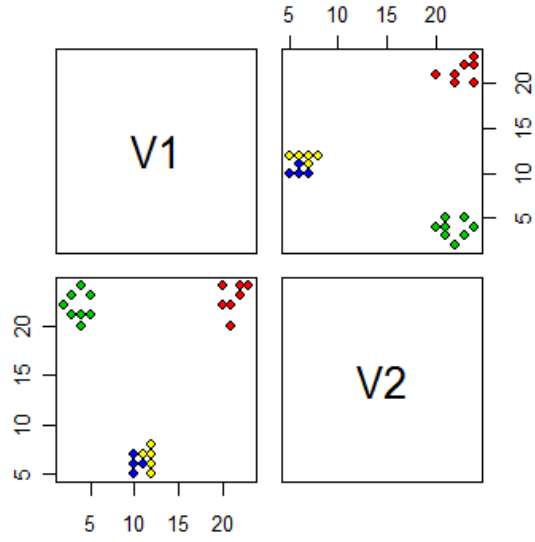
3-8 Division into three groups (table)



3-9 Division into two groups (graphic)

	Low income	High income
G1 – blue	3	**4**
G2 – yellow	7	**10**

3-10 Division into two groups (table)



3-11 Division into four groups (graphic)

	Low income	High income
G1 – blue	3	**1**
G2 – yellow	2	**3**
G3 – red	3	**4**
G4 – green	2	**6**

3-12 Divion into four groups (table)

	0	1	2	3
Total	25%	50%	80%	100%

3-13 Percentages for the population problem

	G1	**G2**
Total	1	2
High income	1	1

3-14 Scores for two groups

In table 3-14, based on tables 3-13 for the population problem and 3-3 for the income problem, we have the scores for each group in the two problems studied for substantiality when we divide the clients into two groups. Having these scores in account, the result for substantiality is:

$$Sub = 2.7$$

	G1	**G2**	**G3**
Total	1	1	1
High income	1	1	1

3-15 Scores for three groups

In the same way, table 3-15 gives us the scores when the clients are divided in three groups. The result to the measure is in this case:

$$Sub = 3$$

	G1	G2	G3	G4
Total	0	0	1	1
High income	0	1	1	1

3-16 Scores for four groups

Table 3-16 has the scores of each group when we are considering a division in four groups. Since we have one group with score zero in s_i and s_i^h, we immediately know that our measure will be zero.

$$Sub = 0$$

Like expected, to consider three groups is better in this example. When we have two groups we are considering groups with more people which could be a good thing in a way, but we lose the importance of the high income clients in the middle of the growth. Indeed, it is better to have bigger groups but it is important to do a marketing approach that really reaches the target population. When we pass from three to four groups, we find groups too small that are not interesting in the sense of containing high income clients. Therefore, the separation does not improve the quality measure.

Concerning the problem discussed in the previous section, we can observe that we find better situations than with one group. In the last case, the substantiality will be equal to 3, therefore when we divide the clients in groups we have a better score.

To have a better feeling about our measures, we can compare them with the measures previously used.

We can see in table 3-17 that for both of the previous measures the indication is to consider three groups. However, the second choice is four groups. For four groups our measure is 0, which means that at least one of the groups has 0 in both of the problems in analysis, according with the opinion of the user. In the context of our study, we consider that our results are better. The analyses that are made for the previous measures embrace more cases since it analyzes the problem trying to find a more equal distribution of the records in the groups (Rebelo et al. 2007). However, when the goal is to build a marketing campaign, one group could be much bigger than another and that difference could be interesting. As already explained, we can have a small interesting group for the bank, because it only has high income clients, or a big group that is not that important, since it is constituted only by low income clients.

	2 groups	**3 groups**	**4 groups**
Sub	2.7	3	0
S_1	0.83	0.25	0.67
S_2	0.58	0.88	0.67

3-17 Comparing results table

Those reasons help us to understand why our measure is an improvement in relation with the previous ones. We have in consideration not only the size of the group, according with the opinion of the user, but also the type of clients in that group. The possibility of considering different opinions in the measure is also important. Different users can build different measures. A small bank and a big bank can have different opinions about percentage of clients in the groups and that different opinions can have a big impact in the measure. This impact will be studied with the real data in the next chapter.

3.3 Accessibility

3.3.1 Previous work

Two measures have been considered (Rebelo et al. 2007) to evaluate this criterion: the eigen values of the discriminating function obtained by the linear discriminate and the compactness of the decision tree. In the first case the higher the value of the eigen values the better is the criterion. The second measure is evaluated based on the number of variables used to build the tree and the number of leaves. The criterion improves when these numbers fall.

3.3.2 Proposed measure

Like mentioned before, this criterion is fulfilled if it is possible to identify the needs of the clients in the group in order to find the best marketing campaign. Meaning that what we want to evaluate is the strength of the group. For that reason, we will evaluate the capacity of understanding the position of a client, in relation to a certain variable when we do not use this variable to find the group.

The idea of this measure is to realize if the relation between the groups is the same when we run the algorithm without the most important variable to separate them. With this, we can try to understand if the relation between the clients in the same group is strong enough not to lose this relation when the variable is not there. To be able to do this, we have to study our data and understand which the most important variable to separate the clients is. After knowing that, we will run the algorithm with and without that variable. And then, analyze the distance of the means of total spending in all variables and without the most important variable for each group for both results. The final measure will be the distances between this means of spending. The idea behind this measure is to understand the changes in the groups if we do not consider the most important variable for the separation. If the groups are the same without it, we can use other types of campaigns to reach the clients. Otherwise we will need to focus our campaign in that variable.

$k_w -$ *clusters without the most important variable* (w)

When we are applying this measure, the difference between k and k_w is only the variables considered to run the algorithm because the number of clusters has to be the same.

$t_{mv} -$ *spent of client m in variable v*

$T = \#t_{mv}$

$t_i \; i \in \{1, \dots, k\}$

$t_{iw} \; i \in \{1, \dots, k\}, v/w$

$t_i^w \; i \in \{1, \dots, k_w\}$

$t_{iw}^w \; i \in \{1, \dots, k_w\}, v/w$

$$Auxiliary_i(Aux_i) = \left| |\overline{t_\iota} - \overline{t_{\iota w}}| \right|$$

$$Auxiliary_{iw}(Aux_{iw}) = \left| |\overline{t_\iota^w} - \overline{t_{\iota w}^w}| \right|$$

$$Acessibility\ (Acc) = |mean(Aux_i) - mean(Aux_{iw})|$$

When looking at this measure, it is possible to think that it does not make sense to look at the results without the most important variable. Imagining that we are studying a problem where age is the most important variable to separate the groups and when we run the algorithm without that variable, we find a completely different result. Why should we forget about this variable? In most of the cases, it probably does not make

sense to do this. However, we are looking at a marketing problem and the final goal is to produce a campaign that reaches the highest possible number of clients. It is possible to do the campaign based on the strongest variable, but this may not be the only important link between the clients and consequently we will lose opportunities. The marketing strategy should be a balance between the cost and the goal. We want to reach our clients but we have restrictions of money. So, when we realize a campaign we need to be sure that we will receive the profits of it. That is why evaluating if the clients stay together with and without the most important variable could be important since that allows us to do a larger and more elaborate campaign. It is not only based on one variable but there are other aspects that make the group to stay together and we can use that to approach them.

3.3.3 Illustrative examples

For the study of this measure we need a more significant example. In table 3-18, we have the data that will be used to explore this measure. As we can see, there will be considered 20 clients characterized by 10 variables. The last column, "T", indicates the position of the client in relation to the income. The clients with the lowest income are classified with 1 and the clients with the high income with 5. This scale will be used to evaluate the study about the most important variable. This study will be done with the help of subgroup discovery. Subgroup discovering is a different way to approach the problem. In this methodology, we need a target to run the experiments and the goal is to find groups with a very different behavior from the data in general (Pieters et al. 2010). For example, imagine that in a dataset like the one in table 3-18 we find a subgroup where 60 percent of the people have variable T equal 5 and actually we know that in all the data we only have 20 percent of the people equal 5 in this variable. With that subgroup maybe we could understand something that we had not noticed before. For instance, this type of analysis could help us to find the more important variables to understand the data, since that they appear in the first places as rules to split the data. For these reasons we will do an analysis with subgroup discovering, that will be presented in the next chapter using the real data. To present the measure we will not do

a deep analysis in the data present in table 3-18. In this dataset, using subgroup discovery analysis, we can see that the most important variable is "V6".

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	T
1	69	17	60	9	68	25	5	43	29	64	**1**
2	47	48	87	28	14	70	60	50	75	74	**1**
3	46	18	43	93	60	66	34	72	18	78	**2**
4	62	90	62	11	73	63	81	62	68	8	**5**
5	44	32	61	83	68	2	5	96	51	70	**4**
6	66	78	96	87	90	92	55	90	97	79	**2**
7	76	2	79	19	51	30	63	28	23	12	**4**
8	56	51	18	77	45	26	46	76	80	63	**5**
9	51	66	98	91	71	7	8	11	50	23	**3**
10	80	79	61	63	39	77	99	87	49	55	**1**
11	79	50	5	54	23	51	70	29	38	88	**1**
12	76	95	68	87	3	69	39	73	38	11	**2**
13	14	77	62	33	84	77	35	55	34	45	**2**
14	88	9	9	60	21	7	74	67	85	82	**3**
15	68	58	76	80	64	53	83	15	100	6	**4**
16	8	12	59	95	47	39	8	5	42	65	**5**
17	71	81	71	46	65	100	18	81	32	68	**1**
18	4	26	68	72	67	94	12	26	45	36	**1**
19	73	22	27	3	5	52	23	94	76	45	**5**
20	93	36	95	94	75	21	1	55	44	38	**4**

3-18 Example for explaining the accessibility

We will demonstrate the measure to this criterion, comparing the division of this dataset in two, three and four groups with and without the variable 6.

Concerning the number of elements in the groups, we can find the division present in table 3-19 when we are considering 2 groups.

All variables\without 6	**Group 1**	**Group 2**
Group 1	1	10
Group 2	9	0

3-19 Cross table to the division into two groups

When looking at the table 3-19, we see that only one of the records changes the group. We can observe the mean of each group for all variables and without "V6" and see if the groups maintain their identity. If that happens, we can say that the group is strong because it is not only dependent of one important variable. The records appear together because they are in fact a group.

In the following tables we can see the means of the groups for all variables and without "V6". The results in table 3-20 were found with the results of k-means when all the variables were used to found the clusters. Table 3-21 shows us the results when we run the algorithm without "V6", meaning the clusters were found when k-means did not have the variable "V6" to found the groups.

	Group 1	**Group 2**
All variables	58.13	46.48
Without 6	57.92	47.19

3-20 Averages of the spending of the groups with all variables

	Group 1	**Group 2**
All variables	57.61	48.16
Without 6	57.79	48.39

3-21 Averages of the spending of the groups without "V6"

In table 3-20, we find a difference of 0.91 and in table 3-20 is 0.41. Having our measure in account, the result to this criterion will be:

$$Acc = 0.25$$

Thinking in the same way, for three and four groups we have 0.75 for three and 0.73 for four.

Looking at the results, we can see that it is better to consider two groups, since it is when the groups maintain more accessibility according with our measure. We know that using the average income as variable to study the homogeneity is arguable since the clients can have very different profiles having the same spending. However we assume that this is a social economical indicator.

Once again we will compare the results of our measure with a previous measure used. In this case, it will be the compactness of the decision tree in terms of variables and leaves.

	2 groups	**3 groups**	**4 groups**
Acc	0.25	0.75	0.73
Compactness	(1,2)	(1,2)	(1,2)

3-22 Comparing results table

In this case, the analysis of the compactness of the decision tree in terms of variables and leaves does not help us to decide. For all the cases we found a tree based in one variable and with two leaves. We should decide for two groups, since the final goal is to do a marketing campaign. As already mentioned, we want to spend the less money possible having the higher profit possible. Since all the divisions are equal for this criterion we chose the cheapest one. Comparing with our measure, we think that we are improving. Our measure works in small or big examples when the other has troubles

with small examples, since the decision tree will always be small. We can see that with our measure, it is easier to compare the results for the different groups. Concerning the interpretation of the measure we can say that the compactness of the decision tree can be more intuitively. However with our measure, we can find an explanation closer to the data. For instance in the case in analysis, when we have two groups, the information of the previous measure is, only, that the tree is small. This was expected since that we only have a few records and they are described for few variables. However, our measure allows us to say that when we run the algorithm with and without the most important variable, the changes in the groups are not too significant. In fact the groups remained almost the same as we can also tell from the analysis in table 3-19. That means that the campaign that will be built to reach those clients can embrace more of their interests. From the point of view of marketing, we can say that it is a more interesting analysis.

3.4 Identifiability

3.4.1 Previous work

Previously (Rebelo et al. 2007) this criterion was evaluated using the accuracy of a classification model. The idea is to learn a model with different algorithms and measure the capacity of assigning the clients to a segment. In terms of the marketing perspective, the authors considered that it was more interesting to use the most common models and with an easy interpretation. The results showed we have high identifiability when we have high accuracy. However low accuracy does not mean necessarily low identifiability since the choice of the method can be influencing the results.

3.4.2 Proposed measure

In the previous criterion, accessibility, the idea was to understand if the best campaign to a group was strongly related with the most important variable or if that group would still be a group without it. Here we want to know how well we can characterize a group.

Meaning that when we look at a group, we want to be able to understand with which type of clients we are dealing with. For the problem in study, it means that it is important for the bank to know how much the group spends, giving the possibility of characterize the groups in terms of income, and where they spend it, enabling to characterize the groups in terms of spending behavior. So, to evaluate this criterion, the idea is to find a measure which shows us a partition where the groups can explain themselves in the best way.

To analyze this concept we will use a well known statistical measure. Since we want to evaluate the homogeneity of the groups, we will consider the smaller mean of distances between the variables with bigger and smaller variance in each group.

$$Identifiability\ (Iden) = \frac{\sum_{i=1}^{k} \left| \min(\sigma_{iv}^2) - \max\ (\sigma_{iv}^2) \right|}{k}$$

3.4.3 Illustrative examples

Analyzing the problem for two groups, using the data in table 3-18, we have the results for the variances in table 3-23.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
G1	135.85	762.49	1036.82	868.96	873.56	726.20	656.89	641.76	585.89	957.25
G2	986.25	619.19	323.28	1218.86	132.50	1022.61	428.00	864.78	141.50	512.86

3-23 Variance of each variable in each group

Using the measure presented, we obtain 993.66 as a result to this criterion when we consider two groups. With the same procedure we have 1083.97 and 1168.68 for three and four groups respectively.

Looking at the results, the best choice in this case will be two groups. The distance between the highest and the smallest variance of the variables is the lowest. That means that the spending of the clients in different variables is more similar in that case.

	2 groups	3 groups	4 groups
Iden	993.66	1083.97	1168.68
Accuracy (decision tree)	0.33	0.50	0.50
Accuracy (neural network)	1	1	0.67

3-24 Comparing results table

In table 3-24, we can see the comparison with the results of the previous measure used. As mentioned in the section of the previous work the results can be influenced by the choice of the method. The accuracy of the decision tree tells us that the best division is three or four groups. In this case, for the reasons already mentioned, it would be better to choose three groups. If we look the accuracy using the neural network, the best choice will be two groups. These two different results show us that the previous measure can be harder to use. The choice of the method can be supported by the type of data that we are using. However, in our case, both methods are correct and they give us different results. Our measure does not have this limitation.

3.5 Overview

To finalize the analysis of our measures, we can do a brief overview of all. However, in order to do it we need to apply the measures to the same data. The results in table 3-25 are, for our measures, applied to the dataset from the table 3-18. To explain the substantiality we used an easy example to better understand the measure. However, to do the overview of the measures, we needed to apply them to the same data. We will maintain the percentages in tables 3-2 and 3-3 to evaluate the population and the high income problems. The weights to each problem will be the same used, 30% for the problem of the high income population and 70% for the total population.

	2 groups	3 groups	4 groups
Sub	2.7	0.9	0
Acc	0.25	0.75	0.73
Iden	993.66	1083.97	1168.68

3-25 Overview

Looking at the indication of all measures in analysis, we will consider two groups in this data, as already discussed. The challenge of this work was to quantify the quality of clusterings from a marketing perspective based in a specific dataset. So as expected, our results are really close to the data in study. Would be good to work the measures in order to generalize them to more problems. However we consider that they are helpful since the idea behind them is different from the previous.

In relation with the substantiality, this criterion is concerned with the problem of records in the groups. Previously this criterion was evaluated trying to find a division that splits the groups in a way that one group was not too much bigger than another (Rebelo et al. 2007). However, as analyzed earlier in this chapter, a group can be big and not be important, or small and very important. Our measure takes this in consideration. The need for a specific scale to apply the measure can be a disadvantage. However, since the measure is design for marketing purposes, we can assume that a measure of the value of costumers is available. This measure can be used as the score for the measure purposed here. For instance, in our case, this scale is based on a single variable and defined by domain knowledge. In our study, we are looking only at the records with the highest results in this scale, but it is possible to have other situations in consideration. The capacity of the user to interact with the results, having the specific problem in attention, is also an improvement.

Concerning accessibility, we are trying to evaluate the capacity of understanding the best marketing campaign to each group. We need to know what we can use to reach the subjects in the group. The previous measures used to evaluate this criterion tried to do it looking at the centers of the groups (Rebelo et al. 2007). We think that with our approach we will be able to improve the analysis. Since we are trying to understand, not only how different the groups are but also which are the variables that maintain the

groups together. From the point of view of this analysis, that is using the marketing ideas, this can be very helpful. However, the theoretical concept is not restricted to the marketing idea. Perhaps, the measures as they are built cannot be used directly in other problems, but the idea can.

The last criterion in analysis is identifiability. As explained before, it is different from the previous one since we are not trying to understand the best marketing campaign to reach a group, but the type of records in the group. Previously this criterion was evaluated with accuracy. However, the choice of the method can have a big influence in the results (Rebelo et al. 2007). Our measure tries to identify the spending behavior of our clients. From the point of view of marketing this can be an interesting analysis, since that gives to the bank an important knowledge of the clients. In our case, all of the variables are in the same unit and the kind of spending gives us information about the clients. This cannot be true in other data, however the unit problem can be solved with normalization.

4 Experimental Results

In this chapter, we will firstly present the dataset in study and do an exploration of it. Through different graphics and tables, we will analyze different specificities of the data. Some issues that could interfere in the results will be discussed and solved. A few experiments with different methods of cluster will be run in the data, in order to have a first felling of the behavior of the data. Then, the measures presented and discussed in the previous chapter will be applied on the real data. In the end, we will compare the results of our measure with the results of the previous measures used.

4.1 Exploratory data analysis

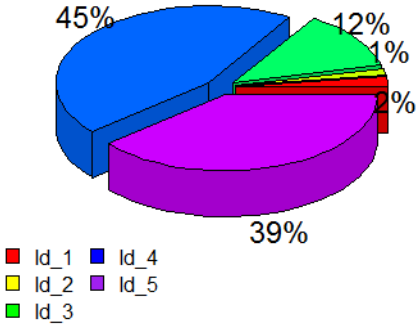
4.1.1 Dataset

Our dataset consists in a set of records discribing the way as people spend their money. We have 5000 entries characterized by 12 variables.

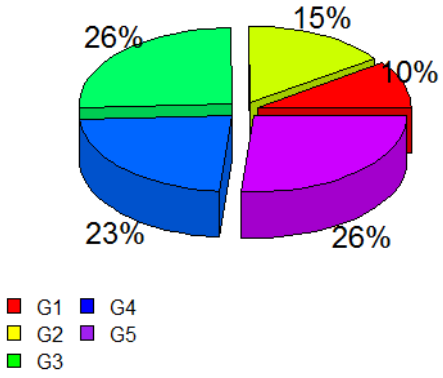
The variables are divided into two different types. We have variables that allow us to characterize the individual and variables describing how much money they spend on different types of products.

Concerning the variables that characterize the individual, we have one about the age scale and other about the scale of spending. Both vary in a range between 1 and 5.

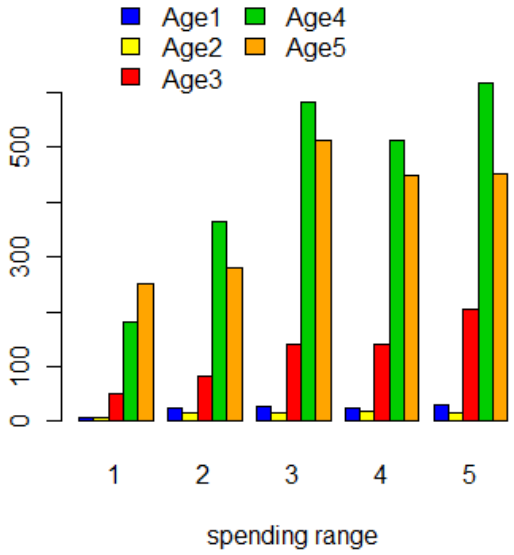
Figure 4-1 shows us the division of the records in the age scale. We can see that most of the clients are in range 4. The ranges with less influence are 1 and 2 with 2% and 1% respectively. Figure 4-2 gives us the same type of analysis but for the spending scale. We can see that most of the clients are in the highest ranges, meaning higher spendings. Figure 4-3 is a representation of both scales. It is possible to understand that the most important age ranges, in all the spending ranges, are 4 and 5. In relation with the age range 4, the spending range where they are in bigger quantity is 3. For the age range 5 it is spending range 5.



4-1 Age division

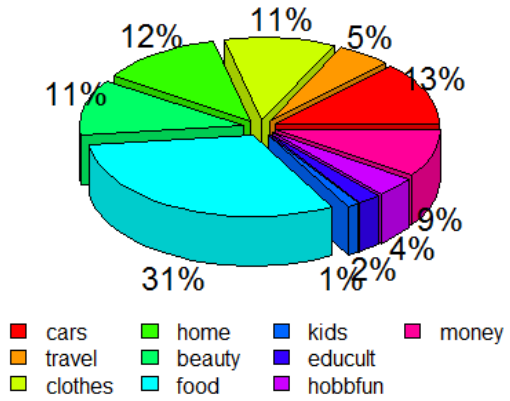


4-2 Spending division



4-3 Age and spending scale division

The others 10 variables allow us to know how people spend their money. We have: cars, travel, clothes, home, beauty, food, kids, educulture (education and culture), hobbiesfun (hobbies and fun) and money, divided like we can see in figure 4-4. The variable with more weight is "food" and with less importance is "kids".



4-4 Division of the global spendings of people

4.1.2 Correlation

In table 4-5, we can see the correlation between the variables. When we look to this coefficient, what we are trying to know is the strength and the direction of the relation between the variables. If the coefficient is positive we have a direct relation, so when one of them grows the other grows in the same direction. When this relation is negative they grow in different directions. The higher correlation that we can have is 1 and the minimum is -1, meaning perfect relation in positive or negative side. If this coefficient is 0, we have no relation between the variables.

The highest relation that we can find, in table 4-5, is 0.50. We can see this value between food and scale of spending and between clothes and beauty. Both of them could be easily explained and make sense. In relation with the first pair, like we showed previously, "food" is the variable where the clients spend more, so this relation makes perfect sense. If "food" is the variable where the clients spend more money, we expected to find a high and positive relation between this variable and the scale that classifies the clients according with their spends. The second can be easily understood when thinking the relation between the variables. It is normal that people that spend money in "beauty" also spend it in "clothes" because, in a way, one helps the other. We can also refer that only nine of the relations are negative and all of them are between the age scale and other variable.

	Age.scale	Spend.scale	Cars	Travel	Clothes	Home	Beauty	Food	Kids	Educulture	Hobbiesfun	Money
Age.scale	1.00	-0.06	-0.03	-0.03	-0.04	-0.03	0.03	-0.08	-0.06	-0.09	-0.01	0.01
Spend.scale		1.00	0.17	0.20	0.40	0.38	0.42	0.50	0.14	0.26	0.12	0.23
Cars			1.00	0.09	0.03	0.09	0.04	0.12	0.01	0.08	0.03	0.04
Travel				1.00	0.13	0.20	0.12	0.42	0.09	0.29	0.08	0.04
Clothes					1.00	0.33	0.50	0.28	0.19	0.20	0.07	0.09
Home						1.00	0.29	0.31	0.10	0.28	0.13	0.11
Beauty							1.00	0.28	0.13	0.17	0.06	0.08
Food								1.00	0.13	0.33	0.10	0.13
Kids									1.00	0.08	0.03	0.01
Educulture										1.00	0.06	0.06
Hobbiesfun											1.00	0.03
Money												1.00

4-5 Correlation between the variables

4.1.3 Subgroup discovery

As explained in the previous chapter, to use subgroup discovery we need a variable to use as target. For running the experiments the spending scale will be used as target and the other variables will be used as description attributes. The analysis that will be made here will not be very deep it is just to have some results that could help us understand a little bit better the dataset in study and find the most important variable to use in accessibility measure. All the results in this section will be obtained from the implementation of the data in Cortana (Meeng & Knobbe 2011).

Nr	Coverage	Quality	Average	St.dev.	conditions
1	1875	37,33614	4,5184	0,631659	FOOD >= '1126.76'
2	1250	36,089211	4,7232	0,481853	FOOD >= '1855.97'
3	2500	34,861168	4,3052	0,75687	FOOD >= '651.89'
4	1875	32,563538	4,376	0,784493	HOME >= '290.0'
5	1250	31,973495	4,5728	0,663247	HOME >= '557.92'

4-6 Target type = numeric; quality measure = Z-score

First condition analysis

3	4	5
140	623	1112
(7%)	(33%)	(60%)

4-7 Clients division

Looking at the first condition, from table 4-6, we obtain a group with 1875 clients that are distributed in the spending scale like we can see in table 4-7. In the quality column, from table 4-6, we can see the score of the subgroup based in the measure that we choose. In this case is z-score which gives us the idea of how this group behaves in relation to the whole dataset. Meaning how far the mean of the subgroup is away from the mean of the whole dataset measured in standard deviations. We can find in a group a behavior above or below the mean. The z-score for this group is 37.33 meaning that this group is above the mean of the whole dataset in this value, measured in standard deviations. In the end, we can say that in this group (that respects the condition of "FOOD >=1126.76") we have 60% of the data with 5 in the spending scale. Comparing with the whole data, it is something interesting since that the 5's only represent 26% of the total records. In the other two cases, we have a big difference in relation to the 3's because in the whole data they represent 26% of the data and the 4's in the whole data

represent 23% and in this group we found 33% of the records. So, in the end the average of the target in this group is around 4.51 as in the whole group is 3.04.

We can also see that food and home are the two most important variables to split the dataset. As explained in the previous chapter, the variables that appear first are the ones that better split the groups.

Nr	Coverage	Quality	Average	St.dev	Conditions
1	1644	37,40	4,60	0,58	FOOD >= '1126.76' AND HOME >= '59.0'
2	1641	37,38	4,60	0,58	FOOD >= '1126.76' AND FOOD >= '1363.44'
3	1875	37,34	4,52	0,63	FOOD >= '1126.76'
4	1875	37,34	4,52	0,63	FOOD >= '651.89' AND FOOD >= '1126.76'
5	1563	37,29	4,62	0,55	FOOD >= '651.89' AND FOOD >= '1457.99'

4-8 Target type = numeric; quality measure = Z-score; refining the search

First condition analysis

3	4	5
77	510	1057
(5%)	(31%)	(64%)

4-9 Clients division

Keeping the same target and in the numeric form, but refining the search using two conditions instead of one, we can see that the quality did not improve a lot. However, the weight of the 5's in these group increases, as showed in table 4-9. Once again, it is possible to see the importance of the two variables to split the records.

Once more we can confirm the importance of the "food" variable to split the groups, looking at the results in table 4-8. All of the rules include this variable and in three of

the cases the two conditions represent the same thing. Like in the case of the last one, being higher than 1457.99 it is the same than being higher than 651.89.

Nr	Coverage	Quality	Probability	Positives	Conditions
1	1875	0,12355	0,593067	1.112	FOOD >= '1126.76'
2	1250	0,1189	0,7392	924	FOOD >= '1855.97'
3	2500	0,1094	0,4824	1.206	FOOD >= '651.89'
4	1875	0,10775	0,550933	1.033	HOME >= '290.0'
5	1250	0,1009	0,6672	834	HOME >= '557.92'

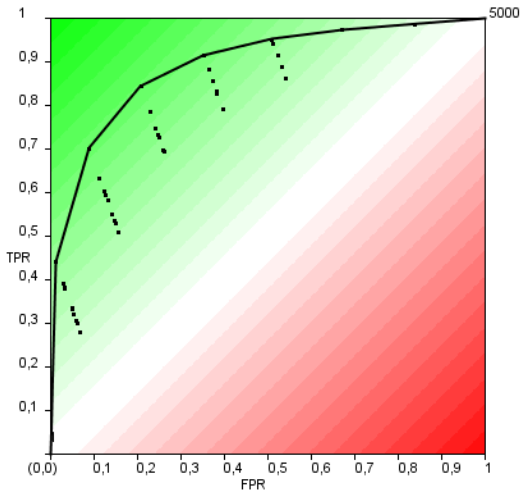
4-10 Target type = nominal – spend.scale=5; quality measure = WRAcc

First condition analysis

Like mentioned before in the whole dataset, we have 26% of the total records equal 5 in the spend scale, this means that 1318 clients are in this range.

Once again the first variable used to do the split was the "food", and actually the split point is the same. So we have the same 1875 clients in this group and in table 4-10 we can analyze directly that 1112 of them are 5 in the target which means that 60% of the groups are in this range. The difference here is the quality measure. In this table, we are using WRAcc (weighted relative accuracy) measuring the accuracy but having in account the size of the group (Lavrac et al. 1999). The difference is that in accuracy when we have a small group it is easier to have a high score, but with this measure is not like that, because the size of the group is taken in account. So, the quality is equal at 12.35% meaning that this percentage of 5's in the spending scale is covered by the rule.

Another possibility in this problem is look to the ROC curve that let us visualize the quality of our group in terms of 5's, comparing the tax of false positives and true positives (Ling & Zhang 2003). As higher is the value under the curve the better the result is. We can see the representation of the ROC curve in figure 4-11 and we have 89.2% of values under the curve, which is a very good result.



4-11 ROC curve

4.1.4 Data preparation

Looking at the data and having in account the problem, one of the first things that we can do is to delete the records of people that do not do any transaction. This makes sense in the context of the analysis since we are trying to find the groups that the bank should consider to approach with new products or marketing campaigns. In this sense, these clients need to be treated in a different way, so it makes sense not to consider them in the analysis. Only a record is deleted in this process, so there is no problem for the analysis.

Another important thing is to analyze the noise in the dataset. Considering as noise the records above, the measure is defined in the formula below:

$$OS = Q3 + 3 * AIQ$$

Q3 represents the third quartile and AIQ the difference between the third and the first quartile.

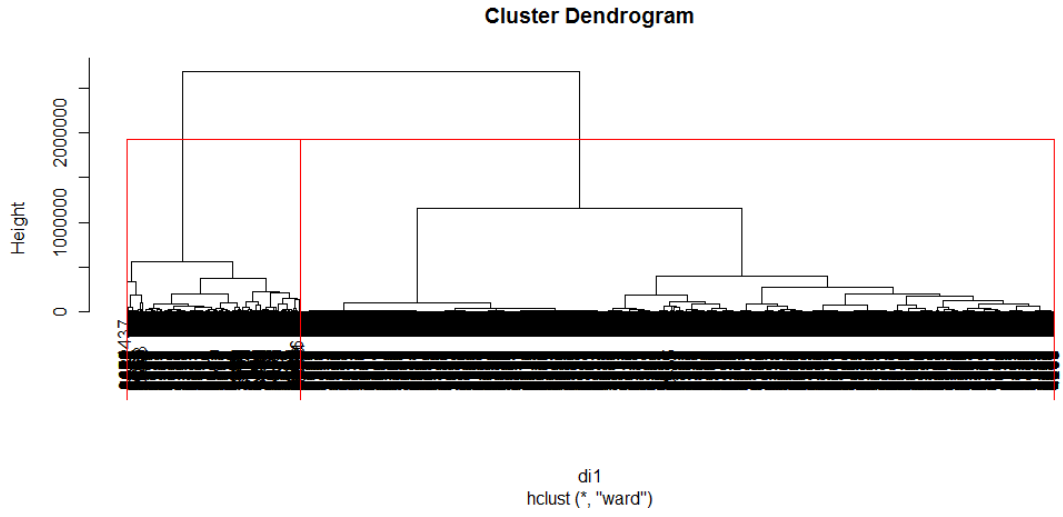
If we consider every record above this measure, in at least one variable, we will delete 2252 records. Since we only have 4999 records we will not do that. We will only delete the records that are outliers in more than half of the variables, meaning 6 or more. In that case, we will delete 56 records leaving us to work with 4943 records in the end.

4.1.5 Preliminary clustering experiments

In this section the intention is to run a first set of experiences with the three algorithms already presented to have a better feeling of the data. That way we hope to find interesting details that help us to better solve the problems that we will face later.

Hierarquical cluster

In order to have one first idea about what type of clusters we can find, in figure 4-12, we can see a representation of the division of the records. Like mentioned above, the first approach will be to run the hierarchical and the model based algorithms and then, based in the heuristics already described, run the partitional method.



4-12 Dendrogram for the real data

In these experiments we are only considering the ten variables about how the clients spend their money.

To build the dendrogram we used the Euclidean distance and the ward method. For the reasons already mentioned, we cut the tree in the separation of two clusters. In table 4-13, we can find a small characterization of the groups, with the average of spending of which group in which variable and the number of elements of each group.

	Group1	Group2	Global
Cars	1372.33	336.60	530.42
Travel	583.88	108.76	197.67
Clothes	1190.35	226.06	439.03
Home	1385.62	281.32	487.97
Beauty	1258.04	282.86	465.35
Food	3739.15	718.83	1284.03
Kids	90.48	33.39	44.07
Educulture	212.43	66.11	93.49
Hobbiesfun	443.36	86.65	153.40
Money	1314.08	184.65	396.00
ELEMENTS	**925**	**4018**	

4-13 Description of the groups using hierarquical method

Looking at the results, it is easy to understand the difference between the two groups. In the first one, we have the clients that spend more money. That is true in all the variables. However, like it was already discussed, we cannot just focus one result.

	Euclidean	Maximum	Manhattan	Canberra	Binary	Minkowski
Ward	(925;4018)	(2155;2788)	(2728;2215)	(2377;2566)	(2559;2384)	(925;4018)
Single	(4942;1)	(4942;1)	(4942;1)	(4942;1)	(4941;2)	(4942;1)

4-14 Comparing table of hierarquical method results

To make sure that the groups found are good ones, we need to look at other results and compare them. Even with the same algorithm but using different measures, we can find very different results, meaning the groups are not natural ones. We can confirm that with the results in table 4-14.

Model-based method

The model based algorithm advices us to use 9 groups in this data.

	Group1	Group2	Group3	Group4	Group5	Group6	Group7	Group8	Group9
Cars	993.46	666.63	912.54	1042.11	391.13	23.28	3779.85	146.66	677.05
Travel	885.21	303.86	230.07	403.81	0.01	0.01	728.71	3.31	333.54
Clothes	907.01	194.43	1160.57	1118.86	382.42	10.24	2479.91	69.12	1410.88
Home	1265.10	512.31	893.93	1119.17	302.59	5.91	2877.92	80.65	886.53
Beauty	920.96	287.40	1078.28	1372.98	431.23	17.24	2339.99	103.29	853.53
Food	2485.01	1221.44	2390.76	2424.13	1287.38	74.27	4185.20	368.53	2595.45
Kids	6.94	0.02	110.33	314.16	13.67	0	213.39	0.85	1124.34
Educulture	335.74	27.81	3.81	776.65	37.71	0.99	973.94	0	226.88
Hobbiesfun	416.58	57.86	193.95	175.69	89.86	0.30	1334.79	14.93	1585.64
Money	1398.05	582.95	487.48	719.76	153.59	2.21	1810.50	37.65	265.68
ELEMENTS	**576**	**802**	**373**	**127**	**1412**	**572**	**75**	**920**	**86**

4-15 Description of the groups using model-based method

Table 4-15 shows us the average of spending of each group in each variable. In this case, the analysis is not so intuitively. However, it can be interesting to refer a couple of

things about these results. For all the groups the variable where they spend more money is "food". But, when we look to the variable where groups spend less money we found very different answers. The variable where more groups spend less money is kids and this happens in groups 1, 2, 6 and 7. In relation with the groups that spend more and less money, we have group 7 and group 6, respectively. Concerning group 6, we can see that only by spending in "education and culture" it is not the group that spends less. However, the difference is less than one unit. Calculating the average of spending of group 6 and 7, we have 13.45 and 2072.42 unities. This shows us that the profiles of the groups are very different.

Partitional method

To obtain the results to this method we will use the suggestions from the two previous methods to choose the number of clusters.

	Group1	Group2
Cars	334.53	1384.70
Travel	129.22	496.18
Clothes	248.91	1268.14
Home	292.27	1341.47
Beauty	288.23	1237.80
Food	670.56	3959.50
Kids	26.93	118.84
Educulture	60.90	235.63
Hobbiesfun	99.81	387.13
Money	216.43	1179.14
ELEMENTS	**4021**	**922**

4-16 Description of the groups using partitional method, considering 2 groups

The results that we find in table 4-16 are close to the ones that we saw in table 4-13. The group with more records is constituted by the clients that spend less money in every variable and are closer to the average of spending. However, the results in table 4-17 give us very different groups when compared with the ones obtained for the model based algorithm. Looking at the records in each group we can say that. A significant difference that we can refer from the previous division is that the variable "food" is not the variable where all groups spend more money. That is only true for four of the groups. On the other hand, the variable where groups spend less money is "kids", which was already verified in the analysis.

	Group1	Group2	Group3	Group4	Group5	Group6	Group7	Group8	Group9
Cars	1216.36	351	887.28	587.38	221.02	176814.9	1311.56	915.84	1045.39
Travel	353	0	222.13	377.49	66.72	0	355.90	5668.73	217.53
Clothes	102.86	0	592.76	4473.91	168.18	0	922.44	792.49	440.19
Home	5693.58	0	648.42	1509.08	181.48	0	1022.12	677.69	912.14
Beauty	773.30	0	683.53	4101.43	189.54	0	955.62	757.83	655.54
Food	2291.89	69090.18	2177.47	2875.46	369.40	0	5742.32	1775.98	1832.17
Kids	181.06	0	67.89	182.77	19.01	0	119.28	32.02	21.43
Educulture	350.62	0	150.59	304.26	39.49	0	234.08	228.26	148.08
Hobbiesfun	463.49	0	194.33	365.85	60.30	0	365.05	2322.93	364.42
Money	813.76	6729.72	458.74	733.32	167.93	0	572.84	370.14	15112.28
ELEMENTS	**88**	**1**	**1239**	**107**	**3112**	**1**	**307**	**54**	**34**

4-17 Description of the groups using partitional method, considering 9 groups

These few results show us the difficulty of the problem in analysis. We only considered two possibilities of division and even like that we found different results.

4.2 Testing the measures

In this section we will apply the measures present in the previous chapter to the real data in study. As mentioned in the previous section, after data preparation we have 4943

records. We will work with the ten spending variables already presented and the spending scale that classifies the clients between 1 and 5 according with their spending.

Let us consider splitting the data in two, three, four and five groups. The tables below show us the division of the clients running the k-means algorithm one time for the different number of clusters. As referred in the second chapter this method does not guarantee the best result and is highly related with the initial point chosen to start it (Jain & Lansing 2010). However, it is one of the most known and used methods to solve that kind of problems, so we will apply it here. We can also see the number of 5's in each group and the mean of total expenses since that information will be important to the measures in a way.

	Group 1	**Group 2**
Number of clients	4021	922
5's	385	877
Mean of expenses	2367.78	11608.53

4-18 Division of the population with k-means, considering 2 groups

	Group 1	**Group 2**	**Group 3**
Number of clients	959	28	3956
5's	872	28	362
Mean of expenses	10731.19	27514.20	2316.07

4-19 Division of the population with k-means, considering 3 groups

	Group 1	**Group 2**	**Group 3**	**Group 4**
Number of clients	3280	1320	316	27
5's	73	846	316	27
Mean of expenses	1624.51	7147.29	14881.72	28094.62

4-20 Division of the population with k-means, considering 4 groups

	Group 1	Group 2	Group 3	Group 4	Group 5
Number of clients	148	27	3281	248	1239
5's	148	27	86	248	753
Mean of expenses	15723.00	28094.62	1651.52	13139.40	6829.07

4-21 Division of the population with k-means, considering 5 groups

Table 4-18 contains the results of the division in 2 groups, which are the same already presented in table 4-16, in relation with the number of records. In this case, we also have the division of the 5's through the groups. These analyses show us that we have more 5's in the smallest group. That was expected having in account the pattern of spending already analyzed and the analysis made with Cortana in a previous section. Table 4-19 has the same type of analysis but for the division into three groups. The indications are the same as the previous one. In this it is actually interesting that the second group is only constituted by clients in range 5 in terms of spending. Table 4-20 is the analysis of the division into four groups. Once more the smallest groups are constituted for the clients that spend more money and in this case the two smaller groups have only clients from range 5 in the spending scale. The same conclusions can be seen in table 4-21 that gives us the results to the division into 5 groups.

In relation to the first criterion in study, substantiality, we are playing with the number of clients in the group but paying attention in the position of those same clients in the spending scale at the same time. Therefore, as mentioned, we do not want a lot of small groups. However, we do not want to ignore a small group if it is constituted only for big clients. In this criterion, we want to find equilibrium between these two realities. For example when we pass from two to three groups, we are confronted with a group with only 28 clients. The question in this criterion is if it is important to consider this division and have a group so small since the clients in the group are all big clients in the context of the bank. The main problem is, once we are trying to define groups to do a marketing campaign, the resources that we have are limited to do it (Smith 1956). So the main question is if the return that we have will compensate the cost that we are having now.

The introduction of this idea in the measure appears in the possibility of defining the tables that will influence the results. Having the possibility of choosing the importance of the percentage of people in the groups, in terms of total population and high income people, the user can work with these concepts.

In table 4-24, we have the results of that measure to the data. The intervals defined to assign the scores are in tables 4-22 and 4-23 for the population problem and the high income problem respectively. We will analyze the impact of change in these tables, to understand the different results that we can obtain.

	0	1	2	3
Total	40%	60%	75%	100%

4-22 Percentages to the population problem

	0	1	2	3
High income	15%	30%	60%	100%

4-23 Percentages to the high income problem

	2 groups	3 groups	4 groups	5 groups
Substantiality	3.6	0	0	0

4-24 Substantiality results

According to the intervals defined for the scores, we should consider two groups in this dataset. In fact when we passed to three groups we lose, according the intervals defined, the importance in the two measures at least for one group. In the case of three groups the smallest is not important in any of the measures. In terms of population we only have 0.6% of the records and in the income part they only represent 2.2% of the records.

To better understand how that measure works, we can change the initial tables and see how that influences the results. Let us imagine that the high income clients are really important to the bank. When we look to the mean of spending of the 28 persons when we are considering three groups, we can see that importance. So, the user defines table 4-25 to rank the income and maintained the table 4-23 to the population problem.

	0	1	2	3
High income	1%	15%	50%	100%

4-25 Modified table to evaluate the substantiality

	2 groups	3 groups	4 groups	5 groups
Substantiality	3.6	3.9	3.5	3.8

4-26 Results to the substantiality according with the modified table

The results that we find are now completely different. We do not have any group with two zeros for both of the measures. The reason is the importance that the bank gives to the high income clients. Changing the table of percentages to the high income problem, what the bank is doing is change its perspective relatively to these clients. For instance, the passage from 15% to 1% means that the user is giving much more importance to the high income clients. They want to pay them a lot of attention in the marketing approach, Meaning that they considered important to do a specific approach to those clients to then have profits with that.

With the second measure in study, the accessibility, we are trying to understand how well we can define the best marketing campaign that we should build to each group. The idea is to evaluate the capacity of the bank to understand how to approach the clients and if they have the capacity of doing a campaign that really reaches the clients. To accomplish this goal the bank has to understand what the clients want and what they need. We can also say that it is important to understand the needs of the clients that they do not know they have. With the measure used we can understand why the group is a group and if it depends only of one variable or if we can use other points of interest to approach them. For instance if "food" is not the only variable that maintains the group together, we can do a broader campaign to them. If the clients are similar, even if one of them does not have a certain need, the campaign can make him pay attention to

something new and create more profits to the bank. However, if only one variable makes them be together we need to explore only that point.

The indication of this criterion is to consider two groups, as we can see in table 4-27.

	2 groups	3 groups	4 groups	5 groups
Accessibility	98.74	141.86	1087.38	1048.88

4-27 Accessibility results

We can now analyze the results to two and three groups to better understand why two groups are better.

	Group 1	Group 2
All variables	2367.78	11608.53
Without 6	1697.24	7649.04

4-28 Averages of spending with and without "food" variable for two groups

	Group 1	Group 2	Group 3
All variables	10731.19	27514.20	2316.07
Without 6	6782.65	26031.14	1679.36

4-29 Averages of spending with and without "food" variable for three groups

	Group 1	Group 2
All variables	3231.35	16434.50
Without 6	2086.97	13146.30

4-30 Averages of spending with and without "food" variable for two groups

	Group 1	Group 2	Group 3

All variables	12180.33	2666.87	27329.04
Without 6	9227.34	1646.67	25659.51

4-31 Averages of spending with and without "food" variable for three groups

Tables 4-28 and 4-29 give us the results for the clusters obtained with all the variables and tables 4-30 and 4-31 give us the average of the global spending when we run k-means without the variable “food”. Looking at the averages it is possible to understand that we have closer results when we consider only two groups. For instance the group 1, when we are considering 3 groups, is very different in terms of average spending when we have in account the variable “food” and when we do not have it. That occurs when we run the algorithm with or without variable “food”, meaning that the group is mostly based on that variable.

The last criterion analyzed in this work is the identifiability. As already mentioned, it is distinct from the previous one in the way that here we want to know how well we can describe the groups as in the previous one we wanted to know how to approach a certain group. It is a problem of knowing and understanding the people in the group but the previous one is also a problem of capacity to create something that really reaches the population in the group.

	2 groups	**3 groups**	**4 groups**	**5 groups**
Identifiability	18145376	370538413	2912244196	238128246

4-32 Results to the identifiability

According with this criterion the best solution is to consider two groups, as we can observe in table 4-32. That means that when we have two groups the population in them

is more homogeneous. Here, we are comparing the distance between the two most distant variances of variables in each group. That let us understand if the population has the same habits of consumes or if they have it in one variable but not in the others. Once again we can understand that the goal is to know the population in the group, but not in a way that let us understand their needs rather than characterize them. For instance we can use table 4-33 to do it.

	Cars	Travel	Clothes	Home	Beauty
1	334.53	129.22	248.91	292.27	288.23
2	1384.70	496.18	1268.14	1341.47	1237.80

4-33 Description of the population considering two groups

	Food	Kids	Educulture	Hobbies/fun	Money
1	670.56	26.93	60.89	99.81	216.43
2	3959.50	118.90	235.63	387.13	1179.14

4-33 (continuation) Description of the population considering two groups

We can easily see that in group 1 the average of the spending is lower. The spending with the kids is the less important in both of the cases and, as expected, the food is the variable where both of them spend more money. Actually it is interesting to find that the priority in the spending is almost the same in both of the groups, the difference is the quantity.

Since all of the measures used point to two groups, we can do a more deep analysis of them. Regarding the spending scale, we can see that in the second group we only have clients in the 4^{th} and 5^{th} range and only 45 are in the 4^{th}. In the first group most of the clients belong to the 3^{rd} and 4^{th} range. In the age scale in both of the groups the highest

percentage of the clients is in 4th range. It is also interesting to refer that in group 1 we have several clients from the 5th range.

4.2.1 Comparing results

Another way to analyze the results of our measures is comparing them with previous results. We will apply measures used in the study of a web portal (Rebelo et al. 2007) and the internal measure already presented (Kovács et al. 2006).

In the first case we will use the S_1 measure to evaluate the substantiality, the compactness of the decision tree, in terms of variables and leaves, to the accessibility and the accuracy to the identifiability. The results are present in table 4-34.

	2 groups	3 groups	4 groups	5 groups
Substantiality	1.25	2.38	2.63	1.23
Accessibility	(5,10)	(4,8)	(4,7)	(6,13)
Identifiability	0.97	0.96	0.92	0.86

4-34 Comparing results

The results to the internal measure S_Dbw are presented in table 4-35.

	2 groups	3 groups	4 groups	5 groups
S_Dbw	3.32	42.31	33.50	27.63

4-35 Comparing results

Looking at the results in table 4-34, we have different indications about the number of groups that we should considerer according with different criteria. If we look at the substantiality, the indication is considerer 5 groups. The indication of our measure was

two groups. In this case, two groups are in second place, not too far from the best result. Concerning the accessibility, the indication in table 4-34 is to consider 4 groups. In this case, the decision tree has 7 leaves and use 4 variables. Comparing with our results we had the indication to consider two groups. The accessibility in table 1 for two groups is in third place. For two groups the decision tree has 10 leaves and 5 variables that are used. The increase of the variables is not too significant but we have 3 more leaves. The last criterion, identifiability, in table 2 indicates 2 groups. In our measure we had the same indication. Although the classification in table 1 is not unanimous, two groups have the lower rank average.

According with the internal measure, in table 4-15, the indication is to consider two groups. We can also see that the other results are quite higher than this result, which means a worse distribution of the clients. In this table, the second best result is to consider 5 groups. This is different from the previous results, once that division is the worse according with table 1 and the third according with our measures.

It is now important to understand why our measures can be interesting since that the results are similar with previous ones. With our measures we give to the user the possibility of defining certain variables. With that we allow the approach of the results with the specific problem. For instance, in the case of substantiality, the importance that a bank attributes to the percentage of population in a group could be very different from another bank. This difference is important for this problem and our results reflect it. Another advantage is the very straight relation of the measures with the marketing. For each measure we can explain our results with the marketing theory. For that case study in particular this is very interesting. From the point of view of the bank it is very interesting to explain its decision based on marketing theories once that it is the main goal of the division. Concerning the accessibility the way that our measures are built permits to the user to know if he can use only one variable to reach the population or if there are more behind the first idea. This gives to the marketing department more possibilities to work, which can help the bank to improve the profits. As discussed before, knowing which variables make a group be a group allows the marketing department to make a more interesting and rich campaign. Not only focus the most

important variable that made the group be a group but also using the other interests that they have in common.

5 Conclusion

In this chapter, a summarization of the work developed and a brief indication of future work will be presented.

5.1 Review

Many tools can be used to cluster data about customers, to support the segmentation process of companies. However, this raises the problem of selecting which clustering to use for a given application. In this project we proposed measures to quantify the quality of clustering from a marketing perspective. The main goal of this work was to combine the marketing theory about what a good cluster is with cluster evaluation. Two problems were the start point to this work. In the first place, how to divide the clients of a bank according to their transactions in order to find the best marketing approach? For solving this problem, we had six criteria of marketing. However, they are formal definitions. As showed in chapter 2, to do an evaluation manually could be very hard. That led us to the second problem. How to evaluate the results of clustering? How can we decide if two segments are better than three? The intention was to find a way to transform those criteria in measures. Having in account the specific problem in study and the six criteria, we built three measures to do this evaluation. The addressed criteria were: substantiality, accessibility and identifiability.

The first criterion is related with the size of the segment. A segment must be large enough to be useful. Previously this criterion was studied while trying to obtain groups not too different in terms of number of clients. However this is an approach that is arguable in terms of marketing. We do not want a group that is too small because a marketing campaign has costs. On the other hand we do not want to judge a group without looking to the clients in it. It may be interesting to consider a small group if it contains the most important clients to the company. We proposed a measure that takes into account both the size of the clusters as well as the value of the clients it contains.

Having in account the goal of the segmentation this is an important factor. However, when we consider applying the measures to other datasets the second part cannot make sense. But as referred, to the purposes of this work, this measure works well.

The second criterion, accessibility, tries to give some information about the type of campaign the company should do to reach the clients in the corresponding group. The previous approach to this criterion was to look at to the compactness of the decision tree that is obtained from the data describing the clients in the group. This analysis does not work well with small clusters. Moreover, we can say that in terms of marketing analysis we do not obtain much information with this measure and it is not simple to compare the results. Our proposal was to study which variables make the clients in a group stay together. In that sense, our measure reflects this idea when it compares the results of the clustering with and without the more important variable to split the groups. Learning if the group is supported by a single variable or by several variables helps to understand which campaign should be targeted at them. This comparison makes sense when we do not forget that we are evaluating our capacity of understanding a group to build a marketing campaign. It is important to know if the clients are together only because one of the variables or if there are more things that we can use in the campaign, helping us to improve profits.

The final criterion, identifiability evaluates how well the company can describe the clients in the group. Differently from the previously criterion, the goal here is not to understand which campaign to address to each group, but really to understand which type of clients are in each group. The previous approach to this problem used the accuracy of a classification model. This has the problem of the influence of the choice of the method used to obtain the model. Our proposal was to look to the behavior of the clients, trying to estimate how heterogeneous they are. The measure used reflects this idea because it analyzes the distance between variables. As higher is this distance more separated are the groups. As bigger are the distances better is the capacity of explaining the groups and the differences between them.

The development of the measures was focused on a particular problem, the segmentation of the customers of a bank. The data contains two types of variables. Variables that allow us to frame the client according with age and amount of spends and variables that show us where the clients spend the money allowing us to know their habits.

In the end of the work, and after the comparison with previous measures we considerer that the results are helpful. Although the results were very similar to the previous measures, the idea behind our measures is different. They are very close to the problem in study and the application to other datasets needs to be worked first. However, these measures allow the user to explain the results in a marketing sense. The way that the measures were built and explained allows to the user to look at the results and to understand not only how many groups he should considerer but the relation between those same results and the marketing criterion behind.

5.2 Future work

Two points can be developed in the future to improve the work done until now. In this work only three criteria were deep analyzed. It could be interesting to build measures to quantify the results with the other three criterions: responsiveness, stability and actionability. Another challenge for the future could be to find a way of applying these measures to other type of problems. As explained during this work, these measures are well related with the problem in study. Even in other marketing problems some parts of the measures cannot work. For example, in order to apply the second part of the substantiality measures we need a scale like our spending scale.

Concerning the measures present in this work, it would be interesting to work the accessibility in order to find a better way to express the homogeneity. We used the income but as referred it is arguable that this is a good measure. In relation to the identifiability, a more common measure of homogeneity could work too, making the interpretation easier.

David Jobber, 2009. *Principles and Practice of Marketing* M. G. Hill, ed.,

Dolnicar, S., 2003. Using cluster analysis for market segmentation - typical misconceptions , established methodological weaknesses and some recommendations for improvement. *Australasian Journal of Market Research*, 11(2), pp.5-12.

Flynn, P.J., 1999. Data Clustering : A Review. *ACM Computing Surveys (CSUR)*, 31(3), pp.264 - 323.

Fraley, C & Raftery, A E, 1998. How Many Clusters ? Which Clustering Method ? Answers Via Model-Based Cluster Analysis 1. *The Computer Journal*, pp.578-588.

Fraley, Chris & Raftery, Adrian E, 2007. Model-based Methods of Classification : Using the mclust Software in Chemometrics. *Journal Of Statistical Software*, 18(6).

Halkidi, M., 2001a. Clustering validity assessment: Finding the optimal partitioning of a data set. *Proceedings IEEE International Conference on Data Mining, 2001. ICDM 2001*, pp.187 - 194. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=989517 [Accessed March 26, 2012].

Halkidi, M., 2001b. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2-3), pp.107-145.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M., 2002a. Cluster Validity M e t h o d s : Part I. *ACM SIGMOD Record*, 31(2), pp.40-45.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M., 2002b. Clustering Validity Checking Methods : Part II. *ACM SIGMOD Record*, 31(3), pp.19-27.

Jain, A.K. & Lansing, E., 2010. Data Clustering : 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31(8), pp.651 - 666.

Kovács, F., Legány, C. & Babos, A., 2006. Cluster Validity Measurement Techniques. *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Beses AIKED'06*, pp.388 - 393.

Kulkar, P. et al., 2003. Finding Effective Optimization Phase Sequences. *Proceedings of the 2003 ACM SIGPLAN conference on Language, compiler, and tool for embedded sustems LCTES'03*, pp.12 - 23.

Lavrac, N., Flach, P. & Zupan, B., 1999. Rule Evaluation Measures : A Unifying View. *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP-99)*, 1634, pp.174-185.

Ling, C.X. & Zhang, H., 2003. AUC: a Statistically Consistent and more Discriminating Measure than Accuracy. *Proceedings of the Eighteenth International Joint Conference of Artificial Intelligence (IJCAI)*.

Liu, Y. et al., 2010. Understanding of Internal Clustering Validation Measures. *2010 IEEE International Conference on Data Mining*, pp.911-916. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5694060 [Accessed March 10, 2012].

Maimon, O. & Rokach, L., 2010. *Data Mining and knowledge discovery handbook* Springer, ed.,

Maulik, U. & Bandyopadhyay, S., 2002. Performance Evaluation of Some Clustering Algorithms and Validity Indices. *Analysis*, 24(12), pp.1650-1654.

Meeng, M. & Knobbe, A., 2011. Flexible Enrichment with Cortana – Software Demo. In *Proceedings 20th Anual Belgian-Dutch Conference on Machine Learning (BENELEARN 2011)*. pp. 117-120.

Pieters, B.F.I., Knobbe, A. & Dzeroski, S., 2010. Subgroup Discovery in Ranked Data , with an Application to Gene Set Enrichment. *Proceedings Preference Learning workshop (PL2010) at ECML PKDD*.

Pratter, F., 1997. Clustering for Market Segmentation. , pp.1-10.

Radosavljevik, D., Putten, P. van der & Larsen, K.K., 2011. Customer Satisfaction and Network Experience in Mobile Telecommunications. *International Network Economics*.

Rebelo, C., Brito, P.Q. & Soares, C., 2006. Factor Analysis to Support the Visualization and Interpretation of Clusters of Portal Users. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. pp. 987-990.

Rebelo, C., Brito, P.Q. & Soares, C., 2007. Quantitative Evaluation of Clusterings for Marketing Applications : A Web Portal Case Study. *Proceedings of the artificial intelligence 13th Portuguese conference on Process in artificial intelligence EPIA'07*, pp.437-448.

Smith, W.R., 1956. PRODUCT DIFFERENTIATION AND MARKET SEGMENTATION AS ALTERNATIVE MARKETING STRATEGIES. *Journal of Marketing*, 21(1), pp.3-8.

Tan, P.-N., Steinbach, M. & Kumar, V., 2005. *Introduction to Data Mining* ADDISON-WESLEY, ed.,

The Times, 100, 2011. http://businesscasestudies.co.uk/ consulta: 07-02-2012. *The Times 100 Teaching business studies by example*, p.2012. Available at: http://businesscasestudies.co.uk [Accessed February 7, 2012].

Wedel, M. & Kamakura, W.A., 1998. *Market Segmentation - Conceptual and Methodological Foundations* K. ACADEMIC, ed.,