Finding Consumption Patterns Through the Use of Bankcards

Ana Ferreira, *Instituto Superior Técnico, Departamento de Matemática*

Abstract

Banks gave up mass marketing strategy, where market segment differences are ignored and followed strategies where market are divided into different homogeneous groups of customers. This process is known as market segmentation and consists of analysing and identifying groups of customers who have similar product or service needs and preferences. This process allows a better understanding of people's requirements and expectations, as well as their answers to the existent or potential commercial offers. In this work we obtain a segmentation of a bank customers set, according to their bankcard transactions for a year and more specifically according to their bankcard purchases. Customers segmentation was achieved, perfoming a cluster analysis. Principal components (PC) are used as a preprocessing set that allows to identify clusters based on consumption profiles, downweighting the impact of total number of transactions as a cluster criteria. This is obtained by projecting the data in all PCs, except the first one that, in both cases, can be understood as an overall measure of the number of transactions of all types under study.

Introduction

Marketing is one of the important functions in any company. Marketing has become important in banking industry because of the existence of intense competition and due to the increasing requirements of the clients. This has resulted in banks paying increasing attention to marketing techniques [3].

Banks deal with various types of customers and no bank can afford to assess the need of each and every individual customer separately. To overcome this problem, banks adopt a market segmentation strategy, which recognises the wisdom of finding homogeneous groups of clients with common interests and needs to whose they can offer specific business purpose rather than trying to address the requirements of each and every customer separately. Market segmentation divides the whole market into groups of customers who have the requirement of similar kinds of products and services [3].

The increase of products diversity has also contributed to the rise of market segmentation since it is rare for mass marketing to be a profitable strategy. Market segmentation enables more accurate and effective communication of benefits in relation to needs. It can also helps to identify growth opportunities for the bank [3].

Market segmentation must have certain qualities that make it possible to specialise the marketing approaches. The segmentation must be measurable in terms of the criteria used for segmentation, accessible by comunication and distribution channels and

substantial enough to be profitable. This segmentation can be performed in a number of different ways. Market of banking products can be segmented on the basis of demography, geography, psychography or behavior [6]. In this work it is carried out a behavioral segmentation, where market is divided based on the type of bankcard transations and more specifically based on the kind of purchases performed by the customers. Behavioral segmentation is considered most favorable segmentation tool as it uses those variables that are closely related to the product itself [6].

As discussed above, a bank segment aims to be an homogeneous group in terms of the clients needs and expectations from the banking industry. Each segment of the market may demand different products and require different treatment to address the demand. The bank should, therefore, develop the profile of different market segments. Then the targeted market segments should be selected based on their attractiveness. Once the bank has identified the market segments, the next steps will be positioning the product into the targeted market segment [3].

Despite the wide variety of techniques available for grouping individuals into market segments on the basis of multivariate survey information (such as, latent class analysis, neural networks or decision trees) clustering remains the most popular and most widely applied method [1]. Nevertheless, a review of the application of such clustering techniques reveals that questionable standards have emerged. For instance, the exploratory nature of clustering methods is typically not accounted for. Crucial parameters of the algorithms used are ignored, thus leading to a dangerous black-box approach, where the reasons for particular results are not fully understood and pre-processing techniques are applied uncritically leading to segmentation solutions in an unnecessarily transformed data space [1].

In this work it was carried out a cluster analysis grouping the customers of a bank into market segments based on the type of bankcard transations and more specifically based on the kind of purchases performed by them. In the next sections it will be explained the problem in detail, the results and conclusions achieved, as well as the whole procedure of data processing.

Data

The customers data used in this work was provided by a bank. These data contains information on customers and on bankcard transactions made by them. The data set consists of $57\,233$ customers that have made a total of $2\,678\,889$ transactions. This information was collected over a year. A transaction, in reality, represents a set os transactions that were made in the same place along the year under study. Each customer is characterized by several attributes, such as: gender, age, marital status, educational qualification, profession, income and the link strength to the bank. Transactions are caracterized by three attributes: type, CAE and money spent. The transaction types considered are: *cash-advance, purchases, special purchases, deposits, withdrawals, ATM movements* and *services payments*, while CAE is the economic activity code that allows you to have the notion of the purchase type, since it reflects the branch of economic activity of the entity where the purchase was made[1]. In this work, we are interested in grouping customers according to their purchasing patterns, so we have agglomerated the CAE, which can take a very large number of values, in more general classes that can reflect broadly the customers purchasing patterns. Sixteen classes were considered, namely: *food, children, sport, education, housing, personal care, technologies, leisure, jewellery, health, supermarkets, telecommunications, tourism, vehicles, clothing* and *others*.

The initial variables considered to perform both segmentations were the number of transactions or, in the latter case, the number of purchases of each type, made by each customer. Tables 1 e 2 show same summary statistics on these variables, respectively.

Table 1: *Summary statistics on the number of transactions made by each customer.*

	Mean	Median	Max	Variance
Cash-Advance	0.04	0	3	0.04
Purchases	36.57	24	445	1 600.81
Special Purch.	0.00	0	3	0.00
Deposits	0.20	0	9	0.29
Withdrawals	6.67	5	54	31.50
ATM Movements	0.78	0	26	2.15
Serv. Payments	2.54	2	33	6.54

[1] You can find out more information about this code in http://www.sicae.pt/faqs.aspx.

Table 2: *Summary statistics on the number of purchases made by each customer.*

	Average	Maximum	Variance
Food	8.30	130	94.99
Children	1.98	26	1.39
Sport	2.22	25	2.56
Education	1.30	7	0.12
Housing	3.08	33	6.66
Personal Care	2.52	39	3.57
Technologies	1.41	11	0.45
Leisure	3.46	46	8.87
Jewellery	1.52	16	0.57
Health	3.92	44	11.43
Supermarkets	7.02	90	32.82
Telecommunic.	1.33	11	0.38
Tourism	2.00	39	1.46
Vehicles	6.29	110	38.38
Clothing	7.59	105	64.27
Others	6.44	67	33.84

As can be seen, variables have very different variances.

Cluster analysis

Cluster analysis is a multivariate exploratory technique aiming to find groups of observations such that observations within each cluster are more similar to each other than to those in different clusters. In this work, cluster analysis is used to determine groups of bank customers according to their bankcard transactions carried.

The wide range of clustering methods makes it difficult to select one to the problem at hand. This choice should depend on the scope of the data, as well as on the purposes of the analysis. Generally, we want to cluster the data, such that, the members of the same group are similar than members of different groups. So, to obtain clusters it is necessary to define a proximity measure, not only between objects, but also between groups. Moreover, objects are characterized by features whose nature (qualitative or quantitative) will play a crucial role in the choice of these measures. These and other relevant issues that generally arise during the cluster analysis, are adressed in this paper.

In this work, an important step in the early processing of the data was the Principal Components Analysis (PCA), using standardised variables. Principal Components Analysis searchs for linear combinations of the original variables with maximum variance uncorrelated with each other. The new variables obtained in this way are called principal components (PC) and are often used to reduce the dimensionality of the original problem since it is expect that a small number of PCs can explain a large percentage of the data variability.

PCA was used in this work with a slightly different purpose,as we want groups of clients based on their consumption profile not based on the total number of transactions made by each client. Note that the first PC was, in both data sets, an overall measure of the variables under study, *i.e.*, a global measure of the total number of transactions/purchases made by each customer. Thus, by projecting the data in directions orthogonal to the one that defines the first PC (*i.e.* in the directions associated with the remaining PCs), we have downweighted this effect of "size", which was the dominating one in the initial partitions obtained based on (standardized) initial variables.

In this work we used the CLARA (Clustering LARge Applications), a K-medoids method, since it is known that it is more robust than K-means and was constructed to deal with large data sets. Being a partitioning method, CLARA requires the specification of the number of clusters. This problem was solved based on graphical analysis. To this end, we analysed the graphics of SSE, R^2_K and the objective function as functions of the number of clusters.

Having determined both segmentations, clusters were interpreted in terms of the original variables and also based on customer's information, such as: gender, age, marital status, educational qualification, profession, income and the link strength to the bank. Additionally, the money spent on transactions was also used to charaterized the clusters. The interpretation of the clusters was made comparing the average number of transactions or purchases. In order to validate this comparisons, statistical tests were performed and the so-called *effect size*[2] was computed . The same procedure has been applied to the comparisons of median amounts of money spent. Finally, clusters were validated using a cross validation technique and the simple matching coefficient.

Results

In this section, we present the interpretation and caracterization of clusters achieved for both seg-

[2]The effect size is a family of indices that measures the strength of the result established by the null hypothesis, associated with a certain statistical test.

mentations.

Segmentation by transactions type

The segmentation based on transaction types results in six clusters. Figure 1 shows mean of transactions of each type, per cluster.

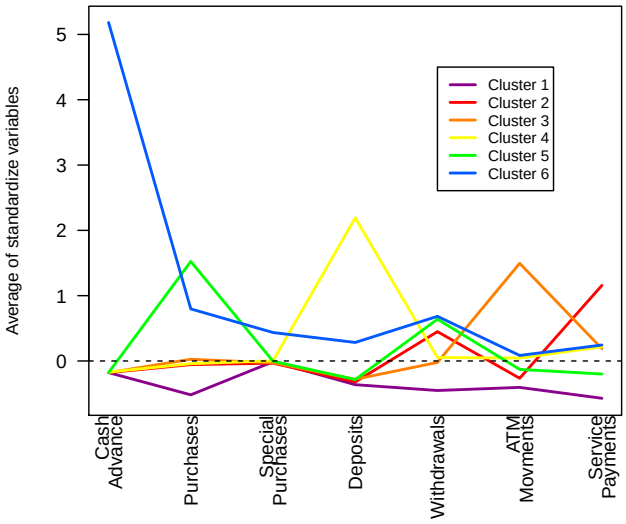

Figure 1: *Average of standardized variables per cluster.*

To compare the cluster's average numbers of transactions of each type and support the conclusions drawn from Figure 1, some statistical procedures were performed. First, we have carried out an analysis of variance (ANOVA), to test whether the differences among the means are statistically significant. The transaction types *Cash-Advance* and *Special Purchases* were not considered because transactions of *Cash-Advance* were made essentially by customers of cluster 6 and *Special Purchases* were only made by 144 customers that are distributed by all clusters. For the remain variables, the p-values are very small. Given the large sample sizes involved, it was not clear of the rejection of the null hypothesis were due to real differences among clusters or just a side effect of the sample size. Thus to find practical evidence against the null hypothesis, the effect size was computed. The index used to estimate the effect size, when comparing several means (using ANOVA) is defined by [2],

$$w^2 = \frac{SSB - (K-1)MSE}{SST + MSE}, \qquad (1)$$

where K is the number of clusters, $MSE = SSE/(n-K)$ and n is the total number of customers. An effect size below 0.01 is considered small [2]. Table 3 shows these values, when comparing the number of transactions of each type among clusters.

Table 3: *Estimates of effect size when comparing clusters means of transaction types.*

	Effect Size
Purchases	0.317
Deposits	0.003
Withdrawals	0.106
ATM Movements	0.008
Services Payments	0.009

Table 3 shows that, with the exception of *Purchases* and *Withdrawals*, the other values are smaller than 0.01, which means that, pratical evidence leads to the conclusion that, on average there are no differences among clusters for *Deposits*, *ATM Movements* and *Services Payments*. In oposition we accept that for *Purchases* and *Withdrawals*, on average, there are differences among the clusters. To verify which clusters have, effectively, different transactions means, Student's t-tests were performed [5] and an estimate of the associated effect size was computed. The index used to estimate the effect size, when comparing two sample means is defined by [2],

$$ES_1 = \frac{\bar{x}_{ij} - \bar{x}_{kj}}{s_p}, \qquad (2)$$

where $\bar{x}_{ij}$ is the sample mean of transactions of the type j made by customers of cluster i (considering only the customers who have performed at least one transaction of this type) and s_p^2 is the pooled variance of the sample, which is estimated by,

$$s_p^2 = \frac{s_{ij}^2(n_{ij}-1) + s_{kj}^2(n_{kj}-1)}{n_{ij} + n_{kj} - 2},$$

where s_{ij}^2 and n_{ij} are the sample variances and the samples sizes. As in ([2]), an effect size below 0.5 is considered small.

Simultaneously, Bonferroni correction was carried out, to counteract the problem of multiple comparisons [6]. This correction consists simply on dividing the significance level by the number of tests. In this case, we have $0.05/15 = 0.003$. Table 4 shows conclusions drawn from this procedure. The notation $Cli \prec (\approx) Clj$ means that mean of cluster i is equal to (smaller than) that of the cluster j for a specific transaction.

According to the obtained p-values and effect size

Table 4: *Clusters sorted by average numbers of transactions of Purchases and Withdrawals.*

Purchases	$C1 \prec C2 \approx C4 \approx C3 \prec C6 \prec C5$
Withdrawals	$C1 \prec C3 \approx C4 \prec C2 \approx C5 \approx C6$

esstimates, it is not possible in some cases, to guarantee a transitivity relation between clusters means. For instance, we can say that the average number of *Withdrawals* made by customers of cluster 2 is equal to that of cluster 5 and this one is equal to that of cluster 6. However, we can not say that the average number of *Withdrawals* made by customers of cluster 2 is equal to that of cluster 6.

We can verify that the types of transactions for which clusters have meaningfull differences between expected numbers of performed transactions, are the ones with more information. Since, *Purchases* and *Withdrawals* were the transactions made in larger quantities and by more customers.

To the best of our knowledge, effect size is an open topic where questions like the threshold value that separates a small effect size from medium to high in the context of segmentation have not been adressed. As a result, there is no history pointing for what can be small in practical sense. In this paper, we consider the threshold suggested by Cohen, proposed for psychology and social science areas. So, we have considered that there may be no practical distinction among clusters based on the expected numbers of transactions in some transaction types, even in cases where a statistical significance evidence to reject null hypothesis (equalitty of means) was found.

To study money spent by customers of each cluster, on each type of transaction, we analysed the medians of these quantities, which are shown in Table 5.

The medians of all clusters were computed simultaneously for each type of transactions, using the Kruskal Wallis test [5], the p-values obtained were all very small. Since we were incapable of finding specific indexes defined to compute the effect size for this test, it was decided to use Mann-Whitney-Wilcoxon test [5] to do multiple comparisons of clusters medians and then compute the effect size. The index used to estimate the effect size, when comparing several medians, is defined by [4],

$$ES_2 = \frac{U}{n_{ij} n_{kj}},$$

where n_{ij} is the length of the sample and U is the statistic of the Mann-Whitney-Wilcoxon test. Note that the index ES_2 coincides with the so-called probability of superiority, which is defined by $P(M_{ij} > M_{kj}) + 0.5P(M_{ij} = M_{kj})$, where M_{ij} is the variable that measures the amount of money spent in transaction type j by customers of cluster i. We have $0 \leq ES_2 \leq 1$ and $ES_2 = 0.5$ when the medians are iqual. So, the effect size around 0.5 is considered small [4]. As there are not explicitly defined thresholds, in this work an effect size between 0.45 and 0.55 is considered small.

Moreover, Bonferroni correction was carried out and p-values were compared to 0.003. Table 6 shows the achieved conclusions. The notation $Cli \prec (\approx) Clj$ means that median money spent on cluster i is equal to (smaller than) that of cluster j.

Table 6: *Clusters sorted by medians of money spent on transactions of each type.*

Purchases	$Cl1 \prec Cl4 \prec Cl2 \approx Cl3 \prec Cl6 \prec Cl5$
Deposits	$Cl1 \approx Cl2 \approx Cl5 \approx Cl3 \approx Cl4 \prec Cl6$
Withdrawals	$Cl1 \approx Cl4 \prec Cl3 \prec Cl5 \approx Cl6 \approx Cl2$
ATM Movmen.	$Cl1 \approx Cl2 \prec Cl4 \approx Cl6 \approx Cl5 \prec Cl3$
Ser. Payments	$Cl1 \prec Cl5 \prec Cl4 \approx Cl6 \approx Cl3 \approx Cl2$

Below, we have a summarised description of the clusters in terms of customer information (not used to performed the cluster analysis).

- Cluster 1

  This is the biggest cluster and it has $23\,142$ customers (40% of total sample). It gathers the customers who have performed, on average, fewer transactions. This cluster has the lowest percentage of women, the highest average age, the lowest proportion of single customers and it has the highest proportion of customers with professions less well paid (housewives, pensioners, unemployed, students, technical and agricultural workers and workers). On the other hand, it has the highest proportion of customers with weak link to the bank. The median amount of money spent by these customers is also small.

- Cluster 2

  This cluster has $9\,622$ customers (17% of total sample). It gathers the customers who have performed more transactions of the type *Service Payments* and the money spent in this type of transaction was also high. The average number of *Withdrawals* was also relatively high. In this cluster, the proportion of customers with

Table 5: *Medians of the amount of money spent in each transaction type, in descending order.*

Cash-Advance	Compras	Comp. Especiais	Depósitos	Levantam.	Movim. ATM	Pag. Serviços	
1 120.00	7 277.13	501.43	837.50	3 580.00	2 285.00	565.02	Global
545.50	4 393.31	385.54	600.00	3 314.11	1 100.00	556.44	Cluster 1
540.00	2 600.89	345.89	600.00	3 260.00	956.96	414.32	Cluster 2
—[1]	2 166.09	327.28	477.50	2 530.00	948.00	400.14	Cluster 3
—[1]	2 113.37	310.27	400.00	2 507.50	800.00	329.88	Cluster 4
—[1]	1 731.37	273.63	360.00	1 970.00	560.00	263.98	Cluster 5
—[1]	1 030.72	65.87	295.00	1 770.00	500.00	136.53	Cluster 6

low educational qualification is the largest one and the proportion of customers with less well paid professions is also high.

- Cluster 3

  This cluster has $8\,282$ customers (14% of total sample). It is the cluster where customers carry out more *ATM Movements* (on average) and spent high amounts of money. For the other types of transactions, the average number is around the global average with exception of *Cash-Advance* that there does not exist in this cluster. The proportion of women and married customers and the average age of customers are relatively high. The remainder proportions have intermediate values.

- Cluster 4

  This cluster has $7\,004$ customers (14% of total sample). It is the cluster where customers carry out more *Deposits* (on average) and spent high amounts of money. This cluster has the smallest age sample mean and the highest proportions of single customers and customers with incomes smaller than or equal to $1\,000$ euros.

- Cluster 5

  This cluster has $7\,292$ customers (13% of total sample). In this cluster, customers carry out more *Purchases* (on average) and spent the highest amount of money. It has the highest proportions of women and costumers with college education. It also has the highest proportions of customers with well paid professions (senior management, self employed, technical and middle management, commercial agents and administrative employees) and with high incomes.

- Cluster 6

  The last cluster is the smallest one, with $1\,891$ customers (3% of total sample). Costumers perform a high average number of all transactions, especially of *Cash-Advance*. However, they move the highest amount of money in *Deposits* when compared to the other customers. The proportion of women is relatively low, as well as the age sample mean of customers. The proportions of customers with professions less well paid and incomes less than or iqual to $1\,000$ euros are relatively low.

[1] No single customer of this cluster have made transactions of this type.

Segmentation by purchases type

Segmentation based on purchases types results in seven clusters. Figure 2 shows the average number of purchases of each type, per cluster.

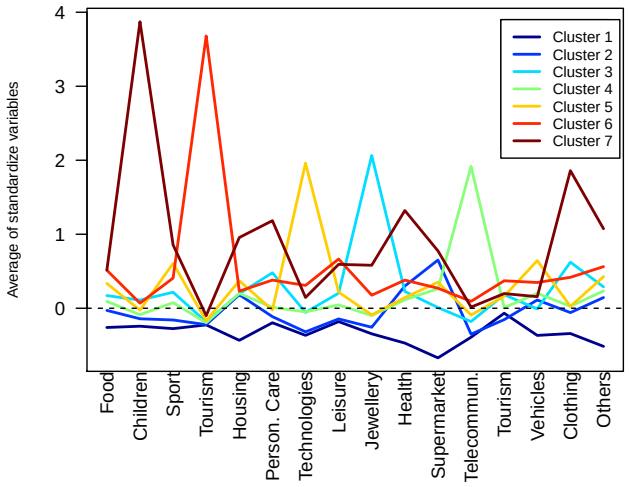

Figure 2: *Average of standardized variables per cluster.*

As in the previous subsection, to compare the clusters average numbers of purchases of each type, we performed an analysis of variance (ANOVA). The purchase type *Education* was not considered because the purchases of this type were made essentially by customers of cluster 6. All p-values obtained were very small, with the exception of *Jew-*

ellery and *Tourism*, so, we reject the hypothesis that there are no significative differences between the clusters average numbers of this purchases type. For the remain types of purchases, the effect size was computed. Table 7 shows these values.

Table 7: *Estimates of effect size when comparing clusters means of purchases types.*

	E. Size		E. Size
Food	0.026	*Health*	0.051
Children	0.262	*Superm.*	0.088
Sport	0.036	*Telec.*	0.001
Housing	0.043	*Vehicles*	0.045
P. Care	0.009	*Clothing*	0.064
Technol.	0.035	*Others*	0.105
Leisure	0.008		

The table above shows that *Personal Care*, *Leisure* and *Telecommunications* have efect sizes smaller than 0.01. So, for the other types of purchases, we carried out Student's t-tests and computed the effect size to verify which clusters have, effectively, diffent average numbers of purchases. Moreover, Bonferroni correction was carried out, using as threshold, $0.05/21 = 0.002$. Table 8 shows conclusions drawn from this procedures.

Table 8: *Clusters sorted by average numbers of purchases.*

Food	$Cl1 \approx Cl2 \approx Cl4 \prec Cl3 \prec Cl5 \approx Cl7 \approx Cl6$
Sport	$Cl2 \approx Cl1 \approx Cl4 \approx Cl3 \approx Cl6 \approx Cl5 \approx Cl7$
Housing	$Cl1 \prec Cl2 \approx Cl3 \approx Cl4 \approx Cl5 \approx Cl6 \prec Cl7$
Technol.	$Cl1 \prec Cl2 \prec Cl4 \approx Cl3 \approx Cl7 \approx Cl6 \prec Cl5$
Health	$Cl1 \prec Cl4 \approx Cl5 \approx Cl2 \approx Cl3 \approx Cl6 \prec Cl7$
Superm.	$Cl1 \prec Cl3 \prec Cl4 \approx Cl6 \approx Cl5 \prec Cl2 \prec Cl7$
Vehicles	$Cl1 \prec Cl3 \approx Cl7 \approx Cl2 \approx Cl4 \approx Cl6 \approx Cl5$
Clothing	$Cl1 \approx Cl2 \approx Cl5 \approx Cl4 \prec Cl6 \approx Cl3 \prec Cl7$
Others	$Cl1 \prec Cl2 \approx Cl4 \approx Cl3 \approx Cl5 \approx Cl6 \prec Cl7$

To study the money spent by customers of each cluster, on each type of purchase, we analysed the medians of these quantities, which are shown in Table 9.

The medians of all clusters were computed simultaneously for each type of purchases, using the Kruskal Wallis test, the and p-values obtained were all very small. After that, it was carried out Mann-Whitney-Wilcoxon tests to do multiple comparisons of clusters medians and then effect size was computed. Moreover, Bonferroni correction was carried out and p-values were compared with 0.002. Table 10 shows the conclusions achieved.

Table 10: *Clusters sorted by medians of money spent on purchases of each type.*

Food	$Cl1 \prec Cl2 \approx Cl4 \approx Cl3 \approx Cl6 \approx Cl5 \approx Cl7$
Children	$Cl2 \approx Cl4 \approx Cl1 \approx Cl5 \approx Cl6 \approx Cl3 \prec Cl7$
Sport	$Cl2 \prec Cl4 \approx Cl1 \approx Cl3 \approx Cl6 \approx Cl7 \approx Cl5$
Housing	$Cl1 \approx Cl2 \approx Cl4 \approx Cl6 \approx Cl3 \approx Cl5 \prec Cl7$
P. Care	$Cl2 \approx Cl5 \approx Cl4 \approx Cl1 \approx Cl6 \prec Cl3 \approx Cl7$
Technol.	$Cl2 \prec Cl1 \approx Cl4 \approx Cl6 \approx Cl3 \approx Cl7 \prec Cl5$
Leisure	$Cl2 \approx Cl4 \approx Cl1 \approx Cl5 \approx Cl3 \approx Cl6 \approx Cl7$
Jewelle.	$Cl2 \prec Cl1 \approx Cl4 \approx Cl5 \approx Cl7 \approx Cl6 \prec Cl3$
Health	$Cl1 \prec Cl4 \approx Cl5 \approx Cl2 \approx Cl6 \approx Cl3 \prec Cl7$
Superm.	$Cl1 \prec Cl6 \approx Cl3 \approx Cl4 \approx Cl5 \approx Cl2 \prec Cl7$
Telecom.	$Cl2 \prec Cl1 \approx Cl6 \approx Cl7 \approx Cl5 \approx Cl3 \approx Cl4$
Tourism	$Cl2 \approx Cl6 \approx Cl4 \approx Cl5 \approx Cl7 \approx Cl1 \approx Cl3$
Vehicles	$Cl1 \prec Cl2 \approx Cl3 \approx Cl4 \approx Cl7 \approx Cl6 \prec Cl5$
Clothing	$Cl1 \approx Cl2 \approx Cl4 \approx Cl5 \prec Cl6 \prec Cl3 \prec Cl7$
Others	$Cl1 \prec Cl2 \approx Cl4 \approx Cl3 \approx Cl6 \approx Cl5 \prec Cl7$

This section ends with a brief characterization of the clusters based on customers information.

- Cluster 1

  This is the biggest cluster and it has $19\,834$ customers (38% of total sample). It gathers the customers who have performed, on average, fewer purchases. The amount of money spent is also low, with exception of *Tourism*. This cluster has low proportion of customers with professions less well paid and the lowest proportion of customers with incomes smaller than or equal to $1\,000$ euros. It has the lowest proportion of customers with weak link to the bank.

- Cluster 2

  This cluster has $11\,612$ customers (22% of total sample). Customers of this cluster have performed an average number of purchases in *Supermarkets* relatively high. The amount of money spent is essentially low. This cluster has the highest age sample mean, the lowest proportion of single customers and it has the highest proportion of customers with less well paid professions. It has, also, the smallest proportion of customers with college education. On the other hand, it has the highest proportion of customers with weak link to the bank.

- Cluster 3

  This cluster has $4\,605$ customers (9% of total sample). These customers have performed an average number of purchases in those types which are traditionally related to women,

Table 9: *Medians of the amount of money spent in each purchases type, in descending order.*

Food	Children	Sport	Education	Housing	P. Care	Technolog.	Leisure
308,04	224,26	115,70	236,44	220,33	161,10	66,02	118,47
232,45	57,89	112,52	220,00	160,00	128,20	52,93	100,10
215,15	55,00	92,25	177,20	145,18	89,90	51,00	99,90
204,60	52,20	91,84	165,00	129,20	77,50	48,48	83,30
172,85	49,75	78,25	136,33	121,00	75,00	45,00	75,60
157,53	44,00	72,93	125,50	119,52	68,00	41,08	72,80
150,59	42,00	72,47	100,00	108,72	67,75	39,97	62,00
89,00	37,90	54,00	__(1)	88,00	59,05	29,92	50,83
Jewellery	Health	Supermar.	Telecomm.	Tourism	Vehicles	Clothing	Others
114,00	421,64	1551,71	69,00	200,00	549,81	920,60	664,84
83,10	219,47	893,31	61,48	173,95	404,09	442,75	517,20
80,00	192,45	766,37	59,90	171,50	375,10	308,52	478,61
79,25	153,75	666,23	59,90	149,50	334,53	221,66	438,70
77,00	145,47	581,02	59,00	140,75	309,10	205,46	384,18
66,45	145,46	568,38	57,10	136,00	273,17	192,07	325,91
65,00	127,01	502,26	51,20	131,40	266,12	158,00	310,35
50,00	87,81	139,91	39,90	107,00	128,34	136,35	168,00

Global
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7

namely, *Personal Care*, *Jewellery*, *Clothing*, higher than the global means. The medians of the amount of money spent in this type of purchases are also high. The proportions of women and single customers are the second largest.

- Cluster 4

  This cluster has $6\,713$ customers (13% of total sample). These customers have performed an average number of purchases around the global average in almost all types of purchases, except in *Technologies*, which have the highest average number. The medians of money spent follow the same pattern.

- Cluster 5

  This cluster has $5\,330$ customers (10% of total sample). It gathers the customers who have made large average number of purchases of the types *Sport, Technologies* and *Vehicles* and the amount of money spent in these purchases is also high. This clusters has the biggest proportion of men and the age sample mean is relatively low.

- Cluster 6

  This cluster has $2\,795$ customers (5% of total sample). It gathers the customers who have made large average number of purchases traditionally related to young people and students, like *Education* and *Leisure* and the same hapen for the money spent. The age sample mean of the cluster is the smallest one and it also has the lowest proportion of single customers.

- Cluster 7

  This cluster has $1\,679$ customers (3% of total sample). These customers have performed the higest average number of purchases of the following types: *Food, Children, Sport, Housing, Personal Care, Health, Supermarkets, Clothing* and *Others*. They have spent an huge amount of money in these type of purchases. This cluster has the highest percentage of women and it has the second lowest age sample mean. The proportion of married customers is the biggest one (the same of cluster 2), as well as, the proportion of customers with college education. It also has the smallest proportion of customers with less well paid professions and with incomes smaller than or equal to $1\,000$ euros. The proportion of customers with strong link to the bank is the highest one.

[1] No single customer of this cluster have made purchases of this type.

Conclusions

The market segments identified express, on the one hand, the usage profiles over the seven types of transactions and, on the other hand, the profiles and consumption habits relating to the perfomed purchases. These segments can be explored by the bank determining which ones are worth betting on them and what is the best way to meet the cus-

tomers demands. The bank can develop a separate product and marketing program to fit more exactly the needs of one or more segments. This approach can provide significant benefits, not only in his own profit but also, to customers.

At a technical level, the main achievement in this work was a novel approach to the use of principal components. In that way, it was possible to smooth the impact of the total number of transactions/purchases in the formation of clusters, whose methodology can be replicated for an update sample of clients.

References

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

[2] M. Rosário Oliveira, Rui Valadas, Marcin Pietrzyk, and Denis Collange. Portability of statistical classifiers for the identification of internet applications. *Submetido para publicao*, 20xx.

[3] OPPapers. Argument for market segmentation, March 2011. URL `http://www.oppapers.com/essays/Argument-For-Market-Segmentation/642235`.

[4] Nick Redfern. An empirical approach to film studies, May 2011. URL `http://nickredfern.wordpress.com/2011/05/12/the-mann-whitney-u-test/`.

[5] David J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC, 5^{th} edition, 2011.

[6] Jeff Tanner and Mary Anne Raymond. *Principles of marketing*. Flat World Knowledge, 2010.