Mat-2.4177 Seminar on Case Studies in Operations Research

Midterm report 5.4.2013
Determinants of cash usage in Baltic countries

Client: SEB

Project team
Lahtinen, Tuomas
Remes, Sami
Repo, Anna
Roponen, Juho (project manager)

Contents

1. Introduction

This midterm report presents the current status of project Determinants of cash usage in Baltic countries. The aim of this project is to investigate the factors that determine cash use in Baltic countries using Estonia as an example. This information may be used in targeting marketing to promote card use in the Baltics. In this project we address the following main research questions:

1) What factors affect cash usage in Estonia?
2) Can these factors be used to identify customer segments that differ in their cash usage?
3) Have these customer segments responded differently to changes in, for example, ATM network?
4) Where the customer segments use their cards?

The expected result of the study is a description of the factors that impact cash use in Estonia, and a segmentation of customers by different cash use profiles. This project is carried out within the course Mat-2.4177 Seminar on Case Studies in Operations Research during January 2013 – May 2013. The client of the project is SEB Estonia.

2. Materials and methods

2.1. Project tasks and status of the tasks

The project tasks were introduced in the project plan. These tasks and their current status are presented in the Table 1. The project has proceeded according to the project plan, and at this point, no modifications to the main research questions have been made. So far, the most challenging and time consuming task has been the identification of customer segments. The trade-offs between different segmentation methods are discussed in more detail in section 2.3. Completion of this task is crucial for the tasks 4-6.

The literature review has revealed that the not much scientific research on the factors that determine cash use have been conducted with large data sets. Previous studies have focused in describing the trends in different payment methods in different countries (e.g. Amromin and Chakravorti, 2007), investigating the effect of debit cards on cash demand (e.g. . Stix 2004) or in understanding the reasons why people adopt and use electronic payment instruments (e.g. Bounie and Francois 2006).

Task	Subtasks	Status
1.Familiarization with data and problem	literature review on previous reseach and trends	in progress
	descriptive statistics of the data	completed
2.Identification of customer segments	familiarization with segmentation methods	in progress
	segmentation by common sense	completed
	segmentation by statistical methods	in progress
3.Comparison of customer segments by cash usage	graphics of customer segments by cash usage	in progress
	comparison with statistical methods	in progress
4.Analysis of card usage in different merchant categories	distribution of card usage in different categories	in progress
	comparidon of card usage in different categories between customer segments	waiting for the segmentation
5.Analysis of the effect of bank action on cash usage	familiarization and data processing of bank action data	in progress
	analysis of changes in pricelist	not started
	analysis of changes in the ATM network	not started
6.Validation and verification of the	validation of methods and assumptions	in progress

Table 1: Current status of project tasks

2.2. Data

The data used in this project consists of real customer and transactions data of SEB. The data consists of personal characteristics of the customers, and the number and the volume of their transactions. The data on personal characteristics describe age, mother tongue, gender, location, income/salary, liabilities, and usage of bank channels, possible unemployment, and children. The data also includes payment merchant category codes (MCCs), which indicate where cards have been used for purchases. The transactions data is aggregated monthly for each customer, and the time-series covers years 2007–2012. Only customers that did transactions every month between January 2007 and December 2012 and were private customers that had not moved to different region or changed living place type during the study period were included in the dataset. The data consists of approximately 71000 SEB private customers that were permanent residents in one of the fifteen regions of Estonia. Customers are between 15 to 80 years of age.

To study the factors that determine cash use in Estonia, we have focused on card purchases and cash withdrawals to limit the analysis to transactions that may involve cash. Electronic payments from one account to another have been excluded from the analysis.

2.3. Identification of customer segments

A customer segment is a collection of customers that can be characterized by some common features. The features that we can use are presented in section 2.2. Our main goals in the identification of segments are relevance and simplicity of the information gained.

- Relevance: The main goal of the project is to study the determinants of cash usage. Ultimately the information could be useful in promoting the use of cards over cash. Therefore, it would be good to find segments that differ in cash usage. The information would be useful, for example if one wanted to target marketing efforts to the most problematic segments. For the information to be practical, the segments should be large enough.

- Simplicity: The segments should be as homogenous as possible. This helps to communicate the information to the client and makes the information more useful, as it is easier to target efforts in order to change the current situation. Additionally, the number of segments should be low, such that the information is easy to comprehend.
-

The applicable statistical methods to categorizing data points into groups usually go by the name Cluster Analysis (see, e.g. Hastie et al. 2009). Broadly, the objective could be to minimize the ratio of “within the groups variance” to “between the groups variance” subject to a constraint that a certain amount of groups must be formed and that every data point belongs to one group. Hastie et al. (2009) presents several algorithms that are commonly used. The crucial part in any approach is determining the “distance” between two data points. In our case, the features that describe the customers are mainly categorical data. A possible distance function between two customers could be the number of features they differ in. However, some features might be more important than the others, making the choice of a distance function difficult. As a status quo approach, we will use equal weights for differences in each feature. Nevertheless, we suspect that by picking a somewhat arbitrary distance function and by resorting to some of the commonly used clustering algorithms could yield an outcome that is not easy to interpret. For example, if the segments are characterized by distributions with respect to each feature, the result may be difficult to understand and use. Ideally, each cluster should contain people that are equal in every feature. This outcome can be achieved by grouping the people with respect to all combinations of the features. As a disadvantage, the number of clusters grows exponentially in the number of features used. On the other hand, if sufficiently many customers belong to a reasonable number (e.g. 20) of clusters, only those clusters can be analysed to find the interesting clusters with the least or the most cash usage.

We decided to proceed as follows. As a starting point we cluster the people with respect to the following features: income class (low = lowest 25%, mid = 25%-75%, high=highest 25%), loan (yes= had loan at any period / no), gender, urban or rural, age (young= born after 1979, mid= born in 1950-1979, old= born before 1950), credit card (yes=had card over 10% of the periods / no), receives children support benefits (yes = received children support benefits over 10% of the periods / no), receives unemployment benefits (yes = received the benefits in over 1% of the periods / no), language (Estonian/ Russian). Using these features to group a random sample of 10000 subjects (out of 71000) results in approximately 500 unique clusters. (This was done with an R-clustering tool hclust.) Selecting the clusters with over 1% of the sample resulted in approximately 20 groups that contain 25% people in the sample. We feel that this number is too low for two reasons 1) it does not necessarily give a good view of the big picture, 2) the effectiveness of targeting these groups might be hindered because the number of people is too low. We used the complete-link method (see, e.g. Hastie et al. 2009) to reduce the number of clusters to around 200. We used the distance function that calculates the number of features in which two persons differ. The resulting clusters seem understandable, which we feel is a key requirement for usability of the results. Additionally, 27 largest clusters contain approximately 64% of all subjects. The median proportions of cash usage in the 27 largest clusters show a range 0.3-0.9. The preliminary results are discussed in more detail in the section 3.1.

2.4. Analysis of card usage in different merchant categories

Merchant category codes are codes assigned to each individual payment terminal depending on what kind of business the terminal belongs to. We have data on all customers that shows in which MCCs they have made payments. What we did with this information was that we plotted histograms of different codes in customer groups that selected based on in how many different kinds of places customers had made purchases. The goal of this was to identify in which kinds of businesses customers used card most or where they were used first.

Next thing we are going to do is have a look at different customer identified segments and their MCC usage and also try to find out if number of MCCs used by a customer correlates with their card usage ratio.

3. Preliminary results

3.1. Customer segments

The Figure 1 shows the proportion of cash usage in the 27 largest (out of 200) customer segments that cover 64 % of the used sample (10000 out of 71000). The range in median proportions of cash usage [0.3, 0.9] is large. The most interesting segments are those that have either especially low or high proportion of cash usage. The five segments that use least cash mostly involve Estonian speaking middle-aged females that live in the urban area. The three segments that use t most cash involve mostly old persons living in urban area, coming from the low income class that do not have loan or credit card.

The characteristic features in the segments that use least cash (median < 0.33, mean <0.39):

- **Segment 1 (1.36% of the sample)**: Lives in rural area, Estonian speaking middle-aged females who are employed, have loan, own credit card and receive children support payment. Mostly middle income class (67% middle, 24% high).
- **Segment 2 (3.7% of the sample):** Lives in urban area, Estonian speaking middle-aged females who are employed, have loan, own credit card. Half receives children support payment, mostly high income class.
- **Segment 3 (1.17% of the sample):** Lives in urban area, Estonian speaking young females who are employed, mostly have loan (80%) do not own credit card, have children. Mostly mid income class (80% mid, 20% high).

Additionally the segment 4 (1.4%) and the segment 5 (2.54%) are quite similar to segments 2 and 3, they contain urban, mostly middle aged, middle income Estonian females. In total, the segments 2-4 cover 8.7% of total population.

The characteristic features in the segments that use most cash (median > 0.73, mean >0.68) are**:**

- **Segment 25 (4.8% of the sample):** lives in mostly urban area (77%), Estonian speaking, mostly old (81% old, 19% mid) females from low income class who are employed, do not have loan or credit card, do not receive children benefits.
- **Segment 26 (3% of the sample):** lives in urban area, mostly Estonian speaking (71%), old males from mostly low income class (57% low, 34% mid), who are employed, do not have loan or credit card, do not receive children benefits
- **Segment 27 (3.5% of the sample):** lives in urban area, Russian speaking mostly old (69% old, 26% mid), females from low income class, who are employed, do not have loan or credit card, mostly do not receive children benefits (91%).

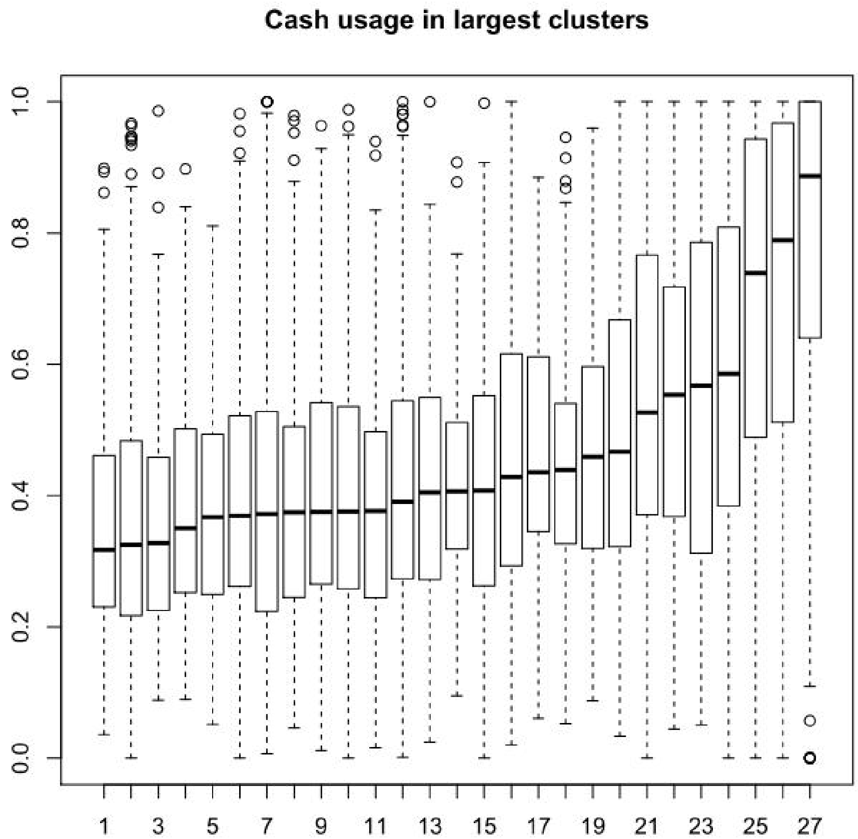

Figure 1: Boxplot of 27 largest clusters. Y-axis shows the proportion of cast usage relative to use of debit card.

3.2. Card usage in different merchant categories

By simply plotting a histograms of the MCCs for the different customer groups separated by MCC usage we found which MCCs were used most by all customers and which were used most by customers who only used cards in very limited kinds of places. Figure 2 for example shows that out of people who used card in less than 10 different MCCs most people used cards in different kinds of food stores and supermarkets(two first bars) and pharmacies (third bar). The fourth bar is building material stores and fifth service stations but they are significantly smaller than the first three.

This kind of information is useful for example because it isn't really that obvious that people use cards to pay in pharmacies even though they don't use them in too many other places. Reasons for this might be numerous. Pharmacies can be viewed as highly reliable stores in which it is safe to use a card without having to fear them cheating. Whatever the reasons might be the information could be interesting for banks because it might give them some kind of chance to increase their customer's card usage.

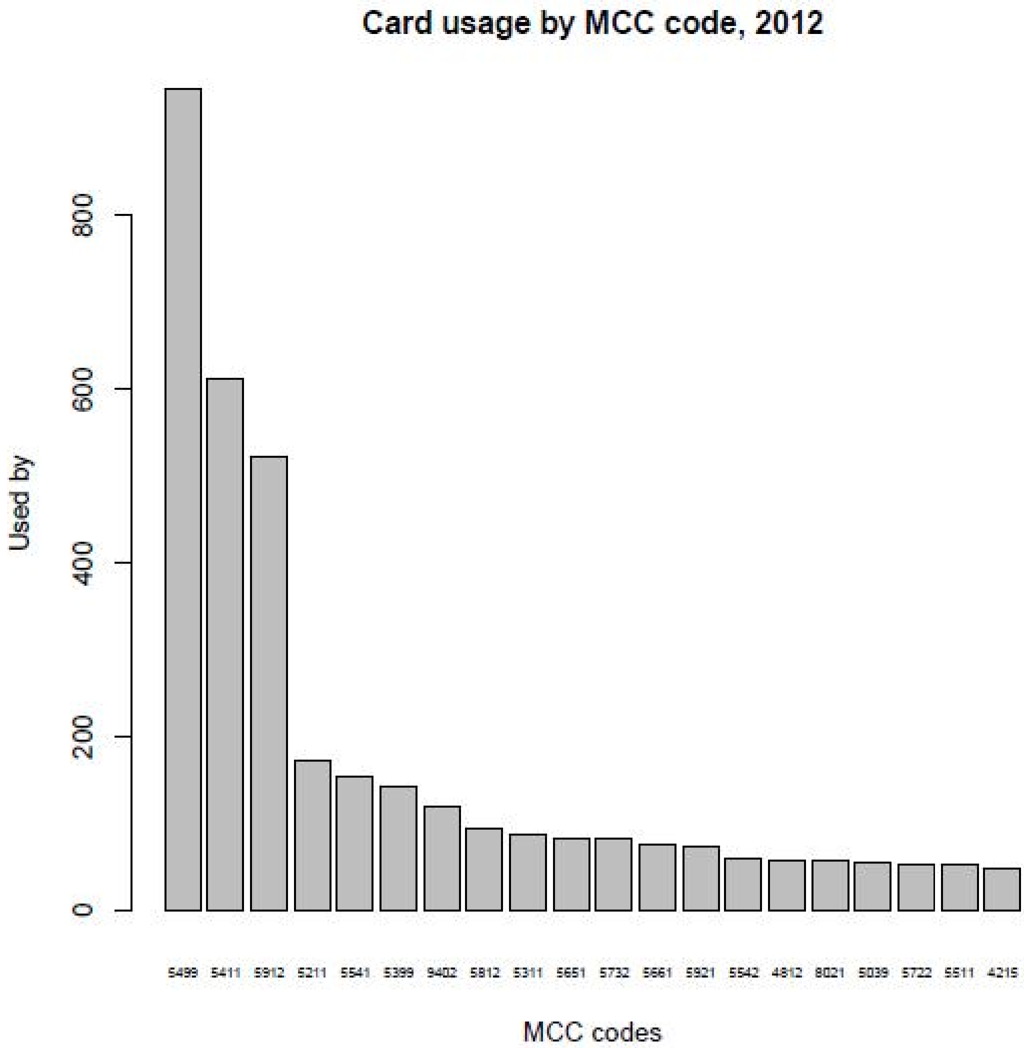

Figure 2: MCCs used by people who used card in less than 10 MCCs in the year2012

4. Project management

4.1 Task allocation and schedule

Our original project plan really did not have a real task allocation. Since then a task allocation has formed. Analysing the information merchant category codes can give has become a task for Juho, while Tuomas and Sami have been working on the customer segmentation and Anna has been doing some valuable background research on other related studies and helping all around.

What comes to schedule then is that we have been managing to hold on to it quite well. The literature review has produced quite nice results and can be considered complete.

The customer segmentation has had some delays but has made nice progress in the last few weeks and we are confident that it will be ready at most two weeks late.

Analysis on the MCC data has been progressing pretty much as planned and should be done on schedule if no surprising delays arise.

Analysis on the price list changes was not started on time but the data was of lower quality than expected, so we will just use less time on it.

Task	February			March					April				May	
week	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Familiarization with data and topic														
Methods for customer segmentation														
Segmentation														
Literature review														
Card use analysis														
Analysis on the effects of the price list changes														
Validation, verification														
Project plan														
Midterm report														
Final report														
Presentations														

Table 2: Project schedule

4.2 Risks revisited

The troubles in customer identification were noted as a risk already in the project plan. As it stands it has taken a bit longer to finish than anticipated, but the situation is not too bad since we already do have some results so a complete failure is no longer a risk.

Timetable was also mentioned as a risk in the project plan and it could be said that we have been able to manage that risk quite well. Even though all parts of the project have not started and finished on schedule, communication within the group and with our customer has worked well enough that everyone is on the same page regarding the progress of the project.

There has not been any problems within the project group thus far either and everyone has been committed and active in the project. The risk of someone quitting right now because of some great dissatisfaction with the project is very low. The possibility of a sudden illness however still remains a risk.

References

Amromin G, Chakravorti, S (2007) Debit Card and Cash Usage: A Cross-Country Analysis. Federal Reserve Bank of Chicago.

Bounie D, Francois A (2006) Cash, Check or Bank Card? The Effects of Transaction Characteristics on the Use of Payment Instruments.

Hastie, T., Tibshirani, R., Friedman, J. 2009. The Elements of Statistical Learning, Data Mining, Inference, and Prediction, 2nd edition, Springer

Stix H (2004) How Do Debit Cards Affect Cash Demand? Survey Data Evidence. Empirica, 31, 93–115.