Classification of the electronic retail core banking market consumers

Ivan Soukal[1], Martina Hedvicakova[2]

Abstract. This paper is focused on the retail core banking services market. In the Czech Republic the consumers can use web accessible free of charge comparison tool for current account offer called the retail bank charges calculator. This system's database holds day-to-day usage patterns of every consumer using this calculator. The aim is to classify the Czech e-banking clients that used the calculator during the year 2011. As a data preparations there were performed verification-validation phase, transformation and dimension reduction. The two-step cluster analysis was performed on 11 255 records classified by 19 variables regarding the type of the service, moth usage frequency or average amount of money, communication channel used to establish or order the service and the criterion if the service falls within or outside the single provider. Analysis indentified as an optimal number of clusters 3. There were identified mainstream client, active client and the client with the mixed communication channel preference.

Keywords: consumer, cluster, retail core banking services, calculator.

JEL Classification: C38, D12
AMS Classification: 91C20

1 Introduction

It is not a simple task to get the market overview in the retail core banking services (thereinafter abbreviated as RCBS) offer. The opportunity costs are high and so the Retail bank charges calculator system (thereinafter abbreviated as Calculator) was created as a web accessible free of charge comparison tool for current account offer. The database of the Calculator now holds the tariff data of 13 banks (more that 98% of the RCBS market in the Czech Republic) and their current accounts. Consumer can use the Calculator to gain individualized market price overview. The collateral benefit to information asymmetry reduction on the RCBS is the dataset of consumer usage patterns.

This dataset will be used to classify the consumers however the users of the Calculator cannot be considered as the adequate sample from the Czech Republic RCBS consumer population. Still the data are the description of the major group of consumers (consumer s with activated electronic banking) and so it can be used for marketing or research purposes. The goal of the paper is to identify these usage patterns regarding the automated teller machine (thereinafter abbreviated as ATM), direct payments, standing orders, encashment, turnover or account balance.

2 Methodology

2.1 The goal and previous problems

The aim of the analysis is to classify the electronic retail core banking services consumers in the Czech Republic into the basic groups by the usage patterns. This paper continues and complements the effort from. [6] There the k-means method was applied to classify the respondent answers gathered during the pilot run of the Calculator and so we demonstrated one of the possible utilization of the Calculator output. The number of clusters determination was based on G5 criterion computation. [4] One of the problems, that occurred, was the asymmetry between the number of transactions and concentration of low frequency of transactions and low turnover consumers. This asymmetry resulted in the computation of one main, or there can be said "mainstream", cluster that contained almost 70 % of the analyzed members. The rest consisted of several smaller, less significant clusters mainly containing consumers showing extremely high numbers of transactions and high turnover on the current account. Part of the problem was wide interval of allowed values, respectively usage frequencies and inclusion

[1] University of Hradec Králové, Faculty of Informatics and Management, Rokitanského 62, Hradec Králové 500 03, Czech Republic, ivan.soukal@uhk.cz.
[2] University of Hradec Králové, Faculty of Informatics and Management, Rokitanského 62, Hradec Králové 500 03, Czech Republic, martina.hedvicakova@uhk.cz.

of consumers that do not have electronic banking activated. However those facts did not caused the large mainstream cluster issue.

2.2 Data source

Data gathering process is conducted by the Calculator's respondent database that can be exported to .csv file. In this analysis the respondent for the year 2011 were analyzed. The Calculator's GUI for consumers can be found at bankovnipoplatky.com server. The consumer enters his or hers usage of the RCBS into the online form and then the individual market price overview is computed and displayed. Therefore it is necessary to presume that the findings of our analysis are limited to consumers with access to the Internet and possessing at least basic level of ICT literacy. Furthermore only the consumers with PC access to account were analyzed (regular reports of [2] show that the share of these accounts is rising and we presume that now it is more than 65 %). Also there has to mentioned that the form for the consumers does not verify the uniqueness of the respondents and so we cannot avoid repeated access of the same respondent, however our number of respondents is high enough to set this potential problem aside. For more information about the Calculator, please see [6] or [7].When the Calculator's form is filled and the price commutation starts, the inputted figures are saved. From the marketing research point of view there are gathered data:

- Multivariate – there has been monitored 53 variable concerning RCBS usage, 2 system variables for respondent identification and 45 variables containing the calculated costs for each of monitored RCBS product,
- Primary – data were gathered directly from the client,
- Subjective – data came from respondent himself, respectively it is his or hers subjective seem.

It is clear that our sample cannot represent the entire RCBS consumer population in the Czech Republic however the Calculator's database holds unique data set that is practically impossible to obtain by usual market survey methods due to the high costs. Moreover there has to be taken into the consideration the motivation, to be more specific the value added. The market overview is highly personalized and so it is in the best interest of the consumer to fill in the form as accurate figures and answers as possible, otherwise the Calculator's output will not help at all.

2.3 Data preparation

Data gathered using the online forms contain a number of faults. Even though the form was improved since the pilot run, yet it fails in data validation e.g. data include obvious errors, blank values or wrong data type. The data preparation was necessary to:

1. decide what variables are suitable for the analysis – due to wide range of the monitored services there has to be determined what to include into the analysis. Considering rarely used services it can be assumed that such services characterize the consumer only superficially and the use of such variables in the cluster analysis could result in the creation of artificial clusters, where the cause criterion would be those services. For each service, respectively variable, the frequency table and mean are computed and evaluated.
2. remove the records with invalid values – number of records in the database contained blank or invalid values. At first there were marked for removal the records containing text values in numeric variables. The correction in some cases is possible. Nevertheless this loss in the file does not mean a significant change in the structure of the clients. Then the interval of illegal values is set because the analysis in [6] confirmed the hypothesis that RCBS products can be also used for business purposes. This is not only in violation of the focus of current research, but mainly in violation of retail bank–client contract. With this in mind frequency tables were again checked and individually assessed for the cut-off frequency.
3. decide the missing values treatment – the literature is usually recommending:

a) replacement by the mean value of the variable,
b) generating a random number from the distribution of the variable,
c) using predicted values from regression model based on other variables,
d) delete the whole record. [3]

In our specific case the variables should be assessed individually with regard to their role and significance in the model. It has to be taken into account the fact that missing values in the questionnaire might have different reasons. Therefore it is necessary to distinguish whether the value is blank because the client does not use the service, or whether the value is blank due to omission or otherwise. That is why it was chosen to treat missing values by:

a) the replacement value of zero – although this operation is not a standard one, in this case it is tenable. It is expected that if the client uses for direct payments telephone and electronic banking, then for the same service the “at the desk” channel is not demanded. So we presume for certain variables that blank values mean: “I do not use.”. Therefore the missing values should be replaced by zero. Some services in the calculator, [6] or [7], are obviously complementary services, where the task of filling in all the fields, would only hinder the consumer.

b) the regression model – not all variables can be substituted by zero. There it is used the regression model where independent variable will be determined from the correlation matrix and predicted value is saved. Using multivariate regression in our case is inappropriate.

c) exclusion of the record – in case the number filled in values is insufficient, the record is removed from the file.

2.4 Data transformation

As mentioned earlier, the analysis published in [6] showed problem of clusters characterized by extreme values and very large mainstream cluster. Use of appropriate transformation will help the symmetrization of the distribution and so to suppress importance of high values of individual variables and to be able of more detailed assessment of the major cluster. Certain number of the variables showed single peak distribution skewed to left. For these variables a logarithmic transformation showed to suitable:

$$X^* = \ln(x+1) \tag{1}$$

2.5 Multicollinearity elimination

Due to the nature and number of analyzed variables can be expected interdependence of the variables. Multicollinearity could significantly affect the output of clustering. To eliminate multicollinearity there was used the principal component analysis method. This method computes a new set of linearly uncorrelated variables from the set of possibly correlated, while reducing the dimension of vector space. The analysis is in the IBM 18PASW (formerly SPSS) included the factor analysis.

There cannot be presumed that all variables have the same weight because the variability differs greatly, there was chosen the extraction based on correlation matrix. Using the correlation matrix extraction usually makes harder to interpret the component, however in this case the interpretation is not needed at all. When determining the optimal number of new variables (there was used non-rotated solution), there can be followed several rules and recommendations [4], [5] such as:

- the value of eigenvalues should be 1.0 and higher however the interval of eigenvalues (0,7; 1) can be consider if revision needed,
- described cumulative variability should be at least 70 % of the original,
- the number of components should be significantly lower than the number of original variables.

Considering our data collection gathered using the questionnaire survey, there can be assumed higher number of components, or a lower level of cumulative variability described. Principal component analysis also solves the problem of scale. In our case there is used its transformation function where e.g. the encashment has naturally lower frequency (mainly considering SIPO payment the bank considers as one payment with no regard to how many orders are made under one SIPO number) than direct payment usage and extremely lower considering significantly different scale of month turnover and average balance in €.

2.6 Cluster analysis

The aim of this paper is to determine the major consumer groups by their usage patterns – what and how often they demand. Given this goal there has been chosen the method of cluster analysis. Compared to the [6] where the k-means algorithm was used, here we used two-step cluster analysis that allows using log-likelihood or euclidean distance measurement (K-means is Euclidean based only).

The pre-cluster step uses a sequential clustering approach. It scans the data records one by one and decides if the current record should be merged with the previously formed clusters or starts a new cluster based on the distance criterion. The procedure is implemented by constructing a modified cluster feature (thereinafter abbreviated as CF) tree, for CF tree and BIRCH algorithm details see [8]. This method allows working with large datasets because using CF characteristics that represent separated data groups this method abstract from single records (cases) inside the group. This reduces computation time and resources.

In two-step cluster analysis we decided to use log-likelihood distance measurement, based on the decrease in the likelihood function when merging clusters. This method was introduced in [1]. The algorithm can estimate the optimal number of clusters by use one of BIC (Bayessian's Information Criterion) or AIC (Akaike's Information Criterion). The BIC characteristic was used. The BIC for each number of clusters within a specified range is calculated and used to find the initial estimate for the number of clusters. Then the initial estimate is refined by finding the largest relative increase in distance between the two closest clusters in each hierarchical clustering stage. The BIC is described e.g. in [4]. Noise reduction was not used because of the risk of possible suppression, although small but valid cluster of clients preferring a e-banking however using the desk services too.

3 Analysis

3.1 Raw data treatment

There were declared types and scale of the variables and so the interval of valid values. Then there was analyzed which variables will be taken into consideration for the upcoming analysis (descriptives and frequency of missing values). All qualitative variables such as card type, homebank, the type of statement etc. were excluded. From qualitative variables there were excluded communication channels of telebanking and collection box because of very low usage and the same case was cash-back service, deposit and withdrawal of high amount of cash, ATM usage abroad. Total of 21 variables was chosen for further analysis, for the list of the variables please see tab. 1 at the end analysis chapter.

The choice of variables was followed by the recode of missing values but in chosen variables only, please see the methodology. Regarding the specific methods of recoding the missing values by zero, there has to be mentioned that in previous research [6] there was identified that most respondents are interested in only less than half the questions, meaning the empty field can be without a doubt considered as unused service. For the variable of average turnover and minimal balance no recode was performed. The variable “ATM withdrawal amount, other bank” there was used the regression of function with independent variable “ATM withdrawal amount, other bank” (the rest of the variable showed only low correlation from -0,18 to 0,23 value). At the end of this phase there was conducted the search for the records service relation errors e.g. when no card selected, then no ATM withdrawal can be made. After this phase there was left 11 255 records that were taken into the next analysis by listwise missing (missing value is blank value or used defined illegal value) value treatment. It lowered the initial number of records from 17 402 to 11 255. It seems as a high number of lost records however there has to be taken into consideration very specific form of data acquisition via the Internet.

3.2 Transformation and dimension reduction

After the verification/validation phase the 21 selected variables was transformed using the formula 1 and saved as a new ones. Due to high dimensionality and the presumption of the cluster analysis of uncorrelated variables there has to be treated the possible interdependence. In the correlation matrix there was found that there is correlation between some variables, to be more specific Direct payments to other bank at desk – Encashment to other bank at desk, Incoming payment from other bank – Incoming payment from own bank. Those correlation coefficients were around 0,45 value. That is why we decided to use principal component analysis.

On the scree plot chart the 4th and 6th components were obvious elbows but were followed by components with eigenvalue higher than one and moreover the total variance explained was lower than 70 %. The last component with eigenvalue higher than one is 8th one and the total variance explained in non-rotated solution is almost 70 and it can be considered almost as an “elbow” point by sight. Still there has to mentioned that both sources [4], [5] that scree plot interpretation is very individual. New variables were saved and were used for the last phase of the analysis.

3.3 Cluster analysis

Due to high number of records only nonhierarchical or modified hierarchical methods were taken into consideration. We decided to use two-step cluster analysis, please see methodology. 8 components were used as continuous variables for clustering using log-likelihood metric and BIC optimal cluster criterion. The CF tree was extended to 5 levels because initial 3 level settings showed unsatisfactory results.

There was computed optimal number of clusters 3. For each cluster there was computed the centroid in initial scale for analyzed variables. There was computed also 4 cluster analysis however the next cluster did not show enough of interpretation value. For 3 cluster computation output see the table below, all frequencies of usage are per mensum, own/other bank reflects that payments can be made between the accounts provided by the same bank or by two independent providers. Amount of money are converted by exchange rate EUR/CZK = 24,29 where EUR is the base currency.

Variable/cluster	**1**	**2**	**3**
Relative cluster size in %	8,0	34,7	57,3

Minimum account turnover in €	709	838	801
Average account balance in €	1 108	1 005	1 098
Domestic ATM withdrawal, own bank	3,34	3,13	2,90
Domestic ATM withdrawal, other bank	0,67	1,85	0
Domestic ATM withdrawal amount, other bank	31,09	94,25	0
Incoming payment from other bank	2,39	2,67	2,37
Incoming payment from own bank	1,26	1,19	1,00
Direct payments to own bank at desk	0,77	0	0
Direct payments to own bank Internet	1,96	2,24	1,98
Direct payments to other bank at desk	0,96	0	0
Direct payments to other bank Internet	2,79	4,54	3,96
Standing orders to own bank at desk	0,62	0	0
Standing orders to own bank Internet	0,97	1,15	1,05
Standing orders to other bank at desk	1,13	0	0
Standing orders to other bank Internet	1,33	2,72	2,53
Encashment to own bank at desk	0,37	0	0
Encashment to own bank Internet	0,40	0,44	0,41
Encashment to other bank at desk	0,56	0	0
Encashment to other bank Internet	0,55	1,02	0,98
Cash deposit at desk	0,69	0,44	0,29
Cash withdrawal at desk	0,54	0,29	0,09

Table 1 average RCBS usage of indentified clusters

In the table there are three usage patterns that can be identified as:

1. mixed preference consumer – even this consumer has active electronic banking, he or she prefers for mostly standing orders and about one third of direct payments at the desk service. Due to higher price of these services it is unusual however the share is low and so generally we can presume that most of consumers in analyzed population prefer strongly e-banking. Considering the overall activity of 4 ATM withdrawals, more than 3 incoming payments, more than 6 direct payments and 6 regular payments it is slightly above average consumer mainly in ATM usage confirming certain preference of cash (and cash services).
2. active e-banking preferring consumer – the third of the analyzed sample is more active then the major one but shares almost total preference of electronic banking. Overall activity varies from 40 % above average to 10 %, 5 ATM withdrawals without regards to the ATM owner, almost 4 incoming payments, almost 7 direct payments and more than 5 regular payments.
3. average consumer – almost total preference of electronic banking and compared to the rest also of ATM of own provider where withdrawals from the other provider ATM is more charged. Average activity is then 3 ATM withdrawals slightly more than 3 incoming payments, 6 direct payments and 5 regular payments.

The analysis also showed that turnover of more active consumer is not significantly higher meaning without noticeable interpretation value. If e.g. the ANOVA would be performed, than due to high number of records statistical significance would be likely found still interpretation value is almost none. This is surprising and it deserves further analysis because it is expectable dependence we are about to analyze in future.

4 Conclusion

The analysis proved that there can be identified specific usage patterns in the dataset of the respondent answers in the Calculator. There were identified 3 clusters in the sample of the electronic banking activated consumers with Internet access to account and basic ICT literacy: mixed preference consumer, active consumer and the average consumer. The difference between the clusters is not considering just the frequency of usage but also the range of used services. The mixed preference consumer uses the branch services however the electronic banking is activated. The average client differs from the active one not just by the frequency but also by the preference of ATMs of own provider. These usage patterns can be used in the further research not just to describe the demand side of the market but also to be the base of the price register. The consumer would just identify the closest cluster. For each cluster the bank would have to state the price without the product tying but fixed one.

Acknowledgements

This paper is written in the frame of specific research project *“Vyspělé technologie pro podporu efektivního podnikového řízení a bankovnictví”*, translated as *“Advanced technologies for effective business management support and banking”*, project order 2108 on the University of Hradec Králové Faculty of Informatics and management, funded by Czech Republic Ministry of Education, Youth and Sport.

References

[1] Chiu, T. D.: A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*. San Francisco, 2001.

[2] Czech national bank. *Financial market supervision reports*. [online]. 2010, [cit. 2012-04-01]. From Financial market supervision reports - Czech National Bank: http://www.cnb.cz/miranda2/export/sites/www.cnb.cz/cs/dohled_financni_trh/souhrnne_informace_fin_trhy/zpravy_o_vykonu_dohledu/download/dnft_2010_en.pdf.

[3] Hebák, P., Hustopecký, J., Jarošová, E., Pecáková, I.: *Vícerozměrné statistické metody. vol. 1.* Informatorium, Prague, 2004.

[4] Hebák, P., Hustopecký, J., et Pecáková, I. [eds.]: *Vícerozměrné statistické metody. vol. 3.* Informatorium, Prague, 2005.

[5] Meloun, M., Militký, J., Hill, M.: *Počítačová analýza vícerozměrných dat v příkladech.* Academia, Prague, 2005.

[6] Soukal, I., Hedvičáková, M.: Retail core banking services e-banking client cluster identification. *Procedia Computer Science Journal* **3** (2010), 1205–1210.

[7] Soukal, I., Hedvičáková, M.: Retail core banking services costs optimization. *Procedia Computer Science Journal* In print

[8] Zhang, T., Ramakrishan, R., Livny, M. Birch, R.: An Efficient Data Clustering Method for Very Large Databases, In *Proceedings of the ACM SIGMOD Conference on Management of Data*. Montreal, 1996.