Management and Business Administration PhD School

THESES OF PhD DISSERTATION

METHODOLOGICAL DEVELOPMENT OF CONSUMER BEHAVIOR ANALYSIS

Ruff Ferenc

Gödöllő
2014

ABOUT THE PHD SCHOOL

NAME: Management and Business Administration PhD School

DISCIPLINE: management and business studies

HEAD OF SCHOOL: Prof. Dr. István Szűcs
Professor, D.Sc.of MTA
SZIU, Faculty of Economic and Social Sciences,
Institute of Economics and Methodology

SUPERVISOR: Dr. László Szelényi
Associate professor, C.Sc. of agricultural sciences
SZIU, Faculty of Economic and Social Sciences,
Institute of Economics and Methodology

....

Approval of head of school Approval of supervisor

1 The antecedents and objectives

An important area of the marketing research is the grouping of observation units, segmentation of them. The most widely used method to solve this problem is cluster analysis. In connection with this method a further investigation has been carried out of the existing scientific evidence, and has been developed. The essence of the test is to search for the optimal number of clusters created by the cluster analysis (i.e. find the number of clusters that covers the clusters best existing in the database). There are several different methods in the literature, of which perhaps the best known is the BIC index [Schwarz, 1978]. However, there are procedures to be decided on the basis of density tests within and outside the clusters under certain cluster distributions. One of this kind of procedure [Tong, 2009] was studied in the literature (history, current status, past results), and then a proposal has been made for its amendment. Then, the modified method has been compared with the original framework on theoretical and practical tests.

In studying and testing the so-called Tong index and its antecedents has been noticed some problems and has seemed an opportunity to correct it, and because of this, an improvement has been expected in the results. (1st research objective.)

The theme of the second study is forecast of the future consumer behaviour. There are a wide variety of analytical techniques in this field, and some of them can be found in the literature review. One of these models has been chosen [van Oest, 2011], and a modification of the model has been made because the conditions created by the authors of the Oest model is not considered justified. Their work is also a development of a model [Fader, 2005]. The basis of the present research is the latter model, however, the direction of the development differs from the van Oest [2011] model. The description of this latter model [Fader, 2005] can also be found in the literature review.

The essence of changes that has been made is to search the opportunity for increasing the forecasting accuracy by involving other variables to the model. On the one hand the extension of a number of the variables (in the observation period) results an information surplus, while the number of data to be derived from probability distributions (such as the number of the parameters of these distributions) also increases, so the computational requirements and complexity increase, too.

Will these changes have significant effect on the results? If the difference is detected, will the accuracy of the model results increase or decrease? (2nd research objective.)

What kind of inference will be deductible from the comparison of the re-

sults of the modified model and the so-called heuristic one – applied many times in the practice [Wübben, 2008] *– to take into consideration of the usefulness of the application? Can the extra work – which is necessary to apply the probability model – be considered as yielding investment?* (Third research objective.)

2 Material and methods

2.1 The examination of the result of the cluster analysis: a possible solution of the selection of the suitable number of the clusters

2.1.1 The theoretical and empirical analysis of the existing methods

The question of the first part of my paper is, if the analyst has to grant the number of the clusters (as the input of the algorithm), then on what kind of manner may choose the best one from the different results. Liu [2010] examined the impact of the structure of the data (noisy data, density differences, subgroups, asymmetric distribution) to the accuracy of the investigated indices (which was used to determinate the cluster numbers). Only one index – the so-called S_Dbw index – were among the 11 ones that made the right decision for all simulation experiments. The procedure was developed by Halkidi and Vazirgiannis [2001], which is based on the density difference between the clusters. This was further developed by Kim and Lee [2003] and Tong and Tan [2009] in the direction to be more robust[1], and to be able to recognize not only spherical clusters. This section of the paper will be a critical examination of this index on theoretical and empirical manner.

2.1.2 The database used for testing indexes and the methods of the comparisons.

In order to compare the results of the indices, such databases are needed in which the components of the clusters are known (i.e., exists groups and the classification of each object is known). These databases were prepared by using random sampling by normally distributed random variables. Since this paper is dealing with bivariate case, therefore, two values have had to be trained for each observation unit: the values of the first and the second variable. Both of these values are potential values of a normally distributed random variable (random sampling). The creation of the different clusters could be achieved by varying the parameters of the distributions (expected value, standard deviation).

[1] To be less sensitive to outliers.

Indexes have been tested in eight databases. The aspects of the creation of databases have been the following:

- clusters with different numbers of elements,
- sparse and dense clusters,
- well separated and less well-separated clusters.

Table 1 shows the parameters of databases (cluster center, standard deviation, number of elements). On these databases clustering procedures have been run with different parameter settings and the indices have been tested on the resulting clusters. This procedure was followed in all the three articles, which dealt with the development of this index. In the analysis of Halkidi and Vazirgiannis [2001] and Tong and Tan [2009] the so-called DBSCAN [Ester, Kriegel, Sander, and Xu., 1996] algorithm was also applied. This method is based on an examination of the densities, and is very effective to separate non-convex, but well-separated clusters. However, this study focuses on the detection of convex and not necessarily completely distinct groups, so this algorithm is not used in the simulations. In all the three articles also used the K -means clustering method. This procedure is often used in marketing research, so this will not be discussed in this paper. Of course this method has been used in this research (the applied software has worked with the Hartigan - Wong algorithm [Hartigan, 1979]).
Another method has been used in this research is the Ward's method (a hierarchical clustering method), which method is also commonly used in marketing research. This method is useful for identifying compact and spherical clusters. The question is how it will be able to detect clusters that do not have these properties.

Of course, the simulation can not be checked in every possible situation. The aim was to investigate whether there is a difference between the results of the two indices when the clusters are closer together. To illustrate this, eight databases have been created.

For comparison, each database has been prepared 10 times with the same parameter settings (Table 1) and the indices has been tested on these databases. The results have been evaluated with respect to accuracy. Can it be shown that one of the indices gives better results than the other one?

2.2 The prediction of consumer behaviour: change the BG/NBD model

2.2.1 The expansion of the BG/NBD model (1)

van Oest [2011] constructed the expansion of the BG/NBD model – has been mentioned in the literature processing – and a solid presentation of this new

Table 1: The parameters of the database have been used to compare the indexes. Source: own compilation.

	K1			K2			K3			K4		
	$\boldsymbol{v}_1$	$\boldsymbol{\sigma}_1$	N_1	$\boldsymbol{v}_2$	$\boldsymbol{\sigma}_2$	N_2	$\boldsymbol{v}_3$	$\boldsymbol{\sigma}_3$	N_3	$\boldsymbol{v}_4$	$\boldsymbol{\sigma}_4$	N_4
1	(0,0)	(1,1)	500	(7,0)	(1,1)	500	(0,-7)	(1,1)	500	(2,7)	(1,1)	500
2	(0,0)	(1,1)	500	(4,0)	(1,1)	500	(0,-7)	(1,1)	500	(2,5)	(1,1)	500
3	(0,0)	(1,1)	100	(4,0)	(1,1)	100	(0,-7)	(1,1)	100	(2,5)	(1,1)	100
4	(0,0)	(1,1)	500	(4,0)	(1,1)	100	(0,-7)	(1,1)	500	(2,5)	(1,1)	250
5	(2,2)	(1,1)	750	(6,0)	(2,2)	500	(2,-7)	(0.5,0.5)	500			
6	(-4,0)	(1,1)	500	(4,0)	(2,2)	1000	(0,-7)	(3,2)	500	(2,5)	(2,1)	500
7	(-4,0)	(2,2)	500	(4,0)	(2,2)	1000	(0,-7)	(3,2)	500	(2,5)	(2,1)	500
8	(0,0)	(1,1)	500	(4,0)	(1,1)	500	(0,-7)	(1,1)	500	(2,2)	(1,1)	500

K1, K2, K3, K4: cluster ID
$\boldsymbol{v}_i$: the center of the i-th cluster
$\boldsymbol{\sigma}_i$: the standard deviation in the x and the y direction of the elements of the i-th cluster
N_i: the number of the elements of the i-th cluster

method will take place in this section. Since it is the basis of the amendment made by the author of this thesis, it has been placed to this part of the dissertation. It could be considered as a material of the scientific research.

The BG/NBD model uses only the number of transactions and the date of the last transaction to forecast future values. However, the question then arises: if in the CRM systems there is much more data of the various buyers, why not use them also in the forecast? The model – to be presented – is an extension of this simple model.

The created model includes (as inputs) the history of complaints about the purchases as well. It was assumed that these inputs contain information that leads to more accurate results in the forecast.

The model [van Oest, 2011] is based on the following assumptions:

1. As long as the customer is active, the number of purchases follows a Poisson distribution with parameter λ_p, which is the expected number of purchases occurring during a certain period of time.

2. Heterogeneity of λ_p follows a gamma distribution with parameters r and α. [2]

3. In case of complaint-free shopping the customer becomes inactive with probability q_p.

4. Heterogeneity in q_p follows a beta distribution with parameterst u_p and

[2] See previous (BG/NBD) model.

v_p:

$$f(q_p|u_p, v_p) = \frac{q_p^{u_p-1}(1-q_p)^{vp-1}}{\mathrm{B}(u_p, v_p)} \tag{1}$$

5. q_p and λ_p vary independently across customers.
6. A complaint on the same day as a purchase occurs with probability μ.
7. Heterogeneity in μ follows a beta distribution with parameters a and b.
8. While active, the number of complaints not occurring on the same day as a purchase follows a Poisson process with rate λ_c.
9. Heterogeneity in λ_c follows a beta distribution with parameters s and β. λ_c is the expected value of the number of complaints.
10. After any complaint, the customer becomes inactive permanently with probability q_c.
11. Heterogeneity in q_c follows a beta distribution with parameters u_c and v_c.
12. q_c, λ_c and μ vary independently across customers.
13. The purchase-related parameters and complaint-related parameters vary independently across customers.

The following data was needed to construct the model:

T	observation period,
x_p	number of purchases,
$x_{c\|p}$	number of same day complaints,
x_c	number of non same day complaints,
t_x	date of last purchase,
z_c	the number of the complaints generated by the last shopping ($z_c \in \{0, 1\}$).

van Oest [2011] created a model from these data and conditions which is presented in the third chapter. The created model gives better forecasts – according to their examinations – then the model from which it was born, however they signal the opportunities of thinking over.

It is true, that this model uses additional information as well, but the difference between the two types of complaints (same-day and delayed) does not seem clear. Usually more time is available to the customer to validate the complaint. In addition, the date of the complaint may also depend on the distance between the customer's apartment and the shop. Thus, despite the results, the model is not convincing. A possible amendment of this model has been made and is presented in the second part of the Results section.

2.2.2 The databases used for testing the model

The already existing and the created model has been tested on artificially generated databases. The essence of this test, that the results of each model are measured in many databases. The databases have been generated based on the distributions being equal to the experiential facts mostly in such a way that certain parameters of the distributions have been changed. The values of three parameters and the length of the forecast period (t) have been modified. All the three parameters may have recorded 3 different values, like this the models have been tested altogether on $3^4 = 81$ databases. All databases imply the data of 1000 customers, which have been generated by the change of the different customer characteristics (such as parameters). The fundamental distributions were the exponential one and the binomial distributions in generating the databases. Time passed between the purchases can be granted with the exponential distribution, while the churn after all purchases can be modelled with the help of the binomial distribution (the parameters of these distributions change per person, see p. 6.). Of course, other effects involved in the model also play a role, namely, that the purchase happened with positively or negatively evaluated complaints.[3] In these cases – it may be hypothesized – the probability of churn is different.

In the case of the databases the number of purchases, the date of the last purchase, and the number and date of the complaints (per person) is available during the observation (test) period (T). Based on these data the models determine the parameters of the distribution (see maximum likelihood method), and with the help of these estimated parameters the models forecast, e.g. the number of purchases per person for the t period follows the observation T period.

The algorithm for generating the databases can be found in the appendix of the paper.

2.2.3 Methods for evaluating the result of the models

The accuracy of the predictions is examined in the "Results" section according to several aspects. These are determined by a set of indicators and their identity and difference have been measured by using statistical methods.

One of these types of indicators, the Cohen kappa index, was developed to examine the identity of two nominal variable [Cohen, 1960]. This value can

[3] For more details see subsection 3.2.1, p. 14.

be calculated by the following formula:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{2}$$

where
p_0 proportion of matches
p_e proportion of matches with independence assumption.
The index – according to Agresti [2010, p. 250.] – is the most popular concordance index on a nominal scale. It's value is between 0 and 1. The higher the value the closer the similarity is between the two variables. The difference between the actual and predicted values can be measured by this index (e.g. in connection with churn rate).

The indices of customers are different for each model, so the following methods have been used to compare them:

- The results are illustrated on Boxplot, which graphically shows the values of the indices, and suitable for easier comparison.
- The Shapiro-Wilk test has been used for normality examination of the results (based on Razali [2011]).
- The F-test has been performed for the comparison of the standard deviation of the results in the case of each model.
- Comparing averages of the models paired t-test, however, if the necessary conditions have not met, the Wilcoxon paired (non-parametric) test has been used.

3 Results

3.1 Check the results of clustering

3.1.1 Modification of the S_Dbw_{new} index

In the first part of this chapter my results for determining the optimal number of clusters are presented. With the help of correcting errors of the previous methods a new method is presented. Due to the errors described in chapter 3 changing the range[4] has been suggested. Instead of the original proposal the function f^* has been defined as follows (furthermore has been renamed f^{**}):

[4]Range selected around the two cluster centres and around the separating point of the two cluster centres. Based on the number of items – can be found in these three ranges – the two clusters can be considered as really one or two clusters.

$$f^{**}(\boldsymbol{x}_i, \boldsymbol{m}) = \begin{cases} 1 \ , & \text{if } m^{(p)} - \alpha \cdot D^{(p)} \leq x_i^{(p)} \leq m^{(p)} + \alpha \cdot D^{(p)} \ , \\ & \forall p \in \{1, 2, 3, \ldots, k\} \\ 0 \ , & \text{otherwise} \end{cases} \tag{3}$$

where
$\boldsymbol{x}_i$: the i-th observation unit,
$\boldsymbol{m}$: an arbitrary unit,
$x_i^{(p)}$: the value of the p-th variable of the i-th observation unit,
$m^{(p)}$: the value of the p-th variable of an arbitrary unit,
$D^{(p)} = \min_i(\sigma_i^{(p)}) \quad , \quad i \in \{1, 2, \ldots, c\}$, the minimum among the standard deviation of the p-th variable of the cluster elements,
α: a suitably chosen constant.

The essence of the modification that the interval – in which we are looking for the observation units – is independent from n (from the number of cluster elements). On the other hand, in the case of $\boldsymbol{m}_{ij}$ points, the distorting effect as noted earlier terminated.

Using this modified function, sub-index $Dens_{bw}^{**}$ has been received instead of sub-index $Dens_{bw}$, from which the full index has been arisen:

$$S\_Dbw^{**}(c) = Dens_{bw}^{**}(c) + Scat(c) \tag{4}$$

3.1.2 Examination of the structure of the modified index (S_Dbw^{**})

One of the aims of this examination is, that the value of the index can be observe as a function of the two sub-indices.
To model this a database has been made with three clusters, in which the location of two clusters has not been changed, and the third one initially has been placed on one of the other two clusters and then has been removed from it along the first coordinate axis. The two overlapping clusters have been considered as one cluster and as two different clusters, and the value of the index has been studied for both versions.

First, the following parameters were identical for all the three clusters[5]: $\sigma_{1x} = \sigma_{2x} = \sigma_{3x} = \sigma_{1y} = \sigma_{2y} = \sigma_{3y} = 1$, which represents the standard deviations along the first and second coordinate axis for each of the clusters. $\boldsymbol{v_1} = (0, 0)^T$, $\boldsymbol{v_2} = (d, 0)^T$ (where $d \in [0, 7]$) and $\boldsymbol{v_3} = (0, -7)^T$ represent the centres of the clusters. All three clusters contained 1000 observation unit. First, the C_1 and C_2 clusters have been merged to a cluster, and then they have been considered as separate clusters, and in both cases the indices have

[5] C_1, C_2, C_3

been tested, while value d varied from 0 to 7 in certain increments. The results can be found in table 2. The sub-indices and the whole indices have been set to pair in the case of the two-clusters and the three-clusters solutions as well.The comparison of the last two columns shows that the values of the indices change at approximately 3.5-4 unit ($3.5 < d < 4$). From this value of the index the solution containing the three clusters is accepted over the other one, because the minimum value of the index gives the best result [Halkidi, 2001]. That is, if the standard deviation of the two clusters (in a given direction) are 1-1 units, then the distance between the centres should be approximately 4 units, that the two clusters can be distinguished. In other words, it is not necessary to be completely free of overlap ("well separated").

Table 2: The values of the sub-indices and the total index as a function of distance (in case of two and three clusters). Source: own calculations.

Distance d	$Dens\_bw^{**}$ $nc = 2$	$Dens\_bw^{**}$ $nc = 3$	$Scat$ $nc = 2$	$Scat$ $nc = 3$	S_Dbw^{**} $nc = 2$	S_Dbw^{**} $nc = 3$
0.0	0.0053	0.3281	0.0592	0.0776	0.0644	0.4057
0.5	0.0000	0.3076	0.0593	0.0790	0.0593	0.3866
1.0	0.0000	0.2266	0.0608	0.0770	0.0608	0.3036
1.5	0.0093	0.2336	0.0671	0.0792	0.0764	0.3128
2.0	0.0156	0.1911	0.0715	0.0782	0.0872	0.2693
2.5	0.0147	0.1774	0.0779	0.0792	0.0926	0.2566
3.0	0.0294	0.1188	0.0871	0.0776	0.1165	0.1964
3.5	0.0777	0.1004	0.0927	0.0744	**0.1704**	**0.1748**
4.0	0.0437	0.0408	0.1046	0.0723	**0.1483**	**0.1131**
4.5	0.0463	0.0383	0.1140	0.0725	0.1603	0.1108
5.0	0.0756	0.0146	0.1248	0.0693	0.2004	0.0838
5.5	0.1067	0.0099	0.1330	0.0660	0.2397	0.0759
6.0	0.0895	0.0045	0.1444	0.0618	0.2338	0.0662
6.5	0.0806	0.0036	0.1519	0.0600	0.2325	0.0637
7.0	0.1190	0.0056	0.1613	0.0569	0.2803	0.0625

nc : number of clusters

The simulation has been executed in several ways. First, the clusters were the same for all calculations (in the case of all the d values), and only the first variable of cluster C_2 has increased to the specified d value (version "A"). In the second case, for each distance new clusters have been produced according to the appropriate parameters (version "B"). Both cases have been performed under different settings: σ_{1x} and σ_{2x} have been changed, and the other parameters have been constant, as can be seen on Table 3. The values of the two (total) indices have also been changed (increased for the two-cluster version, and decreased for the three-cluster version while the distance of the cluster centres (d) increased) as it has been described above. Of course, due to changes in the value of the standard deviation the changing point is different over different distances.

Table 3: The minimum distances between the centres, according to identify the three clusters in the case of clusters with different standard deviation. Source: own calculations.

Type of experiment	σ_{1x}	σ_{2x}	Number of simulations with the given distance result														
			3,5	4	4,5	5	5,5	6	6,5	7	7,5	8	8,5	9	9,5	10	10,5	11
A	1	1	2	8														
A	1	2			4	6												
A	1	3				2	5	2	1									
A	2	2						1	2	6	1							
A	2	3								2	2	3	3					
A	3	3												1	2	3	3	1
B	1	1	3	7														
B	1	2			2	7	1											
B	1	3					1	7	2									
B	2	2						1	3	4	2							
B	2	3									3	6	1					
B	3	3											1	2	3	3	1	

σ : standard deviation

For each parameter setting 10 executes have been performed, and the following has been examined:

- the value of the index as a function of the distance,
- the distance, where the three-cluster results are accepted instead of the two-cluster results.

Table 3 shows the distances at which the presence of three clusters can be identify in the case of the 10 experiments.

Table 3 also shows that it is not an assumption to recognize the three clusters that the clusters are completely separate. It is also seen, however, that if the standard deviations of the clusters increase the uncertainty grows: the deviation of the detection distance (the distance is necessary to detect the three clusters) is greater.[6]

The role of cluster C_3 was that the index can be calculated when the C_1 and C_2 has been merged. Therefore, it has been placed separately from C_1 and C_2 (since the goal was to examine the overlap between C_1 and C_2.)

3.1.3 Comparision of the S_Dbw_{new} and the S_Dbw^{**} indices.

In this subsection the two indexes have been tested on the eight databases has been described in "Materials and Methods" section. Each database has been divided into groups by two clustering algorithms (K-means, Ward), and the number of groups has varied from 2 to 7. Then, the resulting clusters have been compared to the actual clusters, and the best one has been chosen.

[6]In these studies the number of elements of the clusters have not changed.

The results have been evaluated depending on whether the index has found the best solutions generated by the clustering algorithms. The first database contained well-separated clusters, and each of the indices has achieved good results.

In the case of the 2-nd, 3-rd and 4-th databases the clusters have come closer to each other, and the number of elements have also been changed. It can be observed that the reduced number of elements (3-rd database), and the unequal number of elements (4-th database) decrease the performance of the new index, too. The Tong index, however, has given much worse results, particularly in connection with the fourth database. The difference between the results of the two indices is significant.

In the fifth database there is a significant difference between the density of the clusters, and the cluster K3 separated from the other two ones. The results show that the K-means algorithm for the three-cluster layout has been the best in all the ten simulations while the Ward algorithm has given good solution only in four cases. Looking at the indices on clusters created by K-means, the new index has better results as the Tong index. However, in simulations when the clusters have been produced by Ward's method, the new index is always preferred the two-cluster solution, and only once found the real grouping. It can also be observed that in this database the number of clusters produced by Ward's algorithm was changeable.

The sixth database contains clusters which are not circular. In addition, there are the differences between the number of the observation units and the densities of the clusters. The four clusters are not completely separated from each other. Both the K-means and the Ward method have given a solution with four clusters, as best classification (the original database also contained four clusters). Nevertheless, both indices have essentially determined the wrong classification. The solutions seem to be random. Thus, the applicability of the new index is also questionable for this database.

The seventh database has been generated from the sixth one, so that the standard deviation of cluster K1 doubled in both directions, thus the separation from the other three clusters is less than in the 6-th database. Similarly to the previous experiment, solutions with four clusters has fit best to the original clusters, in both cases, but either of the two indices could not find a consistent solution during the 10 simulations. The results cannot be assessed.

In the eighth database three clusters have been very close to each other, while the fourth (K3) has been well separated from them. The configuration with 4-clusters has been the best classification that suits the actual clusters for both clustering algorithms (true, the Ward method has performed better).

However, the Tong index has been again unable to select the best classification. On the other hand, the new index has found the solution with two-clusters the best one. Appendix A.10 shows that for the three nearest clusters the density between clusters is high, so it cannot be expected that the new index is able to distinguish these groups from each other. So this result meets the expectations.

The results show that none of the experiments Tong index has exceeded the results of the new index, but in many cases has also given much weaker results. Of course, there are point locations, where none of the indices could provide assistance in an appropriate decision. So considering these limitations, we can say that the new index can be widely applied.

3.2 The expansion of the BG/NBD forecasting model, and the results of the tests

3.2.1 The direction of the model expansion, and the justification of this direction

After the critical remarks of the models produced by involving complaints, it has been interesting to expand the scope of the original model in other ways. The inclusion of the complaints to the calculation has been kept. However, has not been concentrated on the date of complaints, but also on solutions of complaints made by the company: whether the complaint was treated or not[7]. Purchase without compliant and purchase with compliant have been considered, and the latter category has also been divided into two groups (treated or not treated). Thus, new information will be incorporated into the model, which may affect the results.

In the new model the probability of churn has been given greater if the complaint has not been treated, even if the complaint was not justified. This assumption can be taken into account by setting the parameters.

3.2.2 Conditions of the model creation

1. While the customer is active, the number of purchases occurring per unit time follows a Poisson distribution with parameter λ.
2. λ follows a gamma distribution with parameters r and α.
3. In case of complaint-free purchase the customer churns with probability q_p.
4. Heterogeneity in q_p follows a beta distribution with parameters u_p and v_p.

[7]Treated complaint means that the complaint of the customer has been corrected, the complaint has been assessed positively

5. The probability of a complaint after a purchase is μ.
6. Heterogeneity in μ follows a beta distribution with parameters a and b.
7. A complaint is treated with probability ϵ.
8. Heterogeneity in ϵ follows a beta distribution with parameters e and f.
9. Churn occurs after a treated complaint with probability q_{c1}.
10. Heterogeneity in q_{c1} follows a beta distribution with parameters u_{c1} and v_{c1}.
11. Churn occurs after a non treated complaint with probability q_{c2}.
12. Heterogeneity in q_{c2} follows a beta distribution with parameters u_{c2} and v_{c2}.
13. The parameters for individual customers vary independently across customers.
14. $\lambda > 0$, furthermore $0 < q_p, q_{c1}, q_{c2}, \mu, \epsilon < 1$.

3.2.3 Input

T	a observation period,
x	the number of purchases during T,
x_{c1}	a number of treated complaints,
x_{c2}	a number of non treated complaints,
t_x	date of last purchase,
z	the last purchase is complaint free (true: $z = 1$, false: $z = 0$),
z_1	the last purchase is followed by treated complaint (true: $z_1 = 1$, false: $z_1 = 0$),
z_2	the last purchase is followed by non treated complaint (true: $z_2 = 1$, false: $z_2 = 0$).

Among z, z_1, z_2 exactly one is 1 and the other two are 0.

3.2.4 Determine the expected value of the number of purchases

We have to determinate the expected value of the number of purchases ($X(t)$) in an arbitrary (observed) period (t) for a given customer. Denote this expected value with $\mathrm{E}(X(t))$. Based on this we will be able to make a forecast for an arbitrary period beyond the observation period (T).

The fourth chapter of the dissertation contains the steps of the model creation, only the final result has been stated here:

$$\mathrm{E}(X(t)|\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon) = \lambda t \cdot \mathrm{P}(\tau > t) + \int_0^t \lambda x \cdot f(x)\,\mathrm{d}x =$$

$$= \lambda t \cdot \mathrm{e}^{-\lambda c t} + \int_0^t \lambda x \cdot \lambda c\,\mathrm{e}^{-\lambda c x}\,\mathrm{d}x =$$

$$= \frac{1 - \mathrm{e}^{-\lambda t[1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})]}}{1-(1-\mu)(1-q_p)-\mu(1-\epsilon)(1-q_{c2})-\mu\epsilon(1-q_{c1})} \quad (5)$$

3.2.5 Prediction of the number of purchases

The purpose of modelling is to determine the expected number of customer purchases – $(Y(t))$ – for t period after the observation period. This will help to apply personalized marketing tools to the customers.

The aim is to determine $\mathrm{E}(Y(t)|\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon, input)$ on individual level, furthermore to determine $\mathrm{E}(Y(t)|r, \alpha, u_p, v_p, a, b, e, f, u_{c1}, v_{c1}, u_{c2}, v_{c2}, input)$ on population level. Let the parameters $\lambda, q_p, q_{c1}, q_{c2}, \mu, \epsilon$ be marked with Φ.

Here only the final result has also been stated:

$$\mathrm{E}(Y(t)|\Theta, input) \approx \frac{1}{L(\Theta|input)} \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1 - \mathrm{e}^{-\lambda_i t c_i}}{c_i} \cdot L_{\text{aktív}}(\Phi_i|input) \right] \quad (6)$$

where
N: the number of elements of the random sample,
Θ: the set of $r, \alpha, u_p, v_p, a, b, e, f, u_{c1}, v_{c1}, u_{c2}, v_{c2}$ parameters,
input: set of the input data of the model,
$c_i = 1 - (1-\mu_i)(1-q_{p_i}) - \mu_i(1-\epsilon_i)(1-q_{c2_i}) - \mu_i\epsilon_i(1-q_{c1_i}), \quad 1 \leqslant i \leqslant N,$
furthermore,
$\lambda_i \sim \Gamma(r, \alpha)$,
$q_{p_i} \sim \mathrm{B}(u_p, v_p)$,
$\mu_i \sim \mathrm{B}(a, b)$,
$\epsilon_i \sim \mathrm{B}(e, f)$,
$q_{c1_i} \sim \mathrm{B}(u_{c1}, v_{c1})$,
$q_{c2_i} \sim \mathrm{B}(u_{c2}, v_{c2})$
are the i-th value of a random variable with the given distribution.

3.2.6 Models involved in the study

Three models have been tested on data presented in the Materials and Methods section: the original BG/NBD model, the new modification of this model (has been made by the author of this dissertation), as well as a so-called heuristic model.

As the description of the first two models has been done, the heuristic method will be presented here.

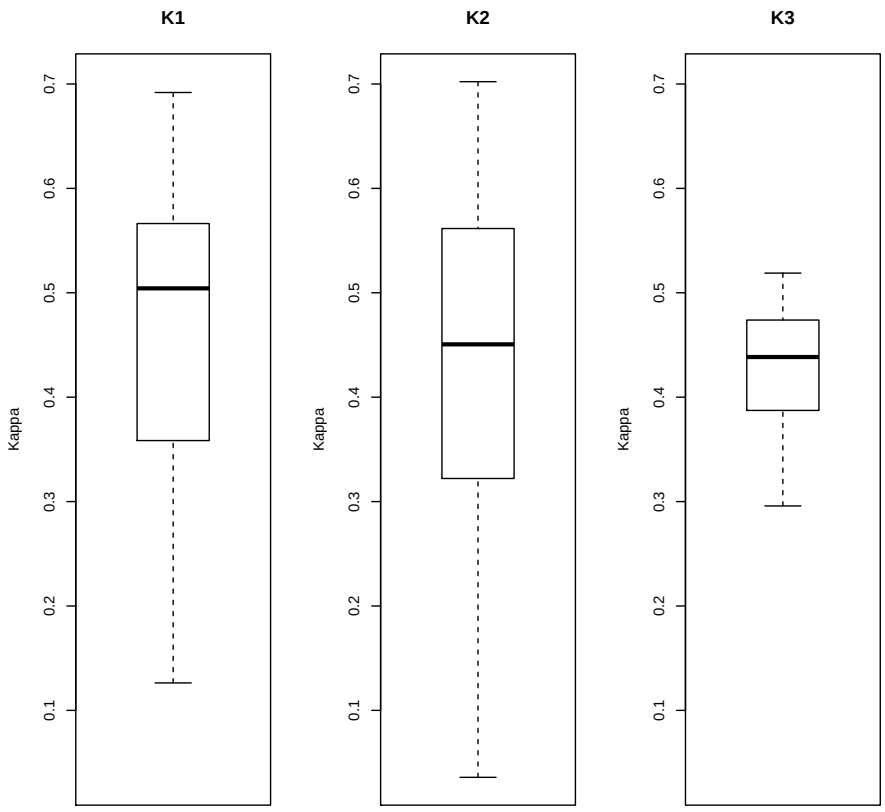
In the case of heuristic model the observation period in each case has been divided into two parts: learning (observation) and test period. In the experiments the observation period (T) has also been divided into two equal $(T/2, T/2)$ parts. Then the average date of the last purchases has been examined for those customers who was inactive for the first $T/2$ period[8], and this average has been chosen as the critical time (which is necessary for the forecast). Of course, the so-called "hiatus" value for the entire T period is the twice of this earlier critical value. Anyone whose last purchase was earlier than this hiatus time (in the observation period), has been taken into consideration as an inactive customer in the forecast (t) period. However, anyone whose last purchase was later than this hiatus time, the number of purchases in the forecast (t) period has been calculated by the number of purchases in the observation period (assuming a linear relationship between the number of purchases and the elapsed time).

3.2.7 Testing the forecast of the customers, who have been active over the observation period

In this subsection the ability of each model to predict the churn of the customer from data of the observation period has been examined. First the database has been prepared and all the three models have been run on them. For each model 10-10 results have been averaged. The values of the Kappa statistics have been examined for each model, the results has been illustrated in a box plot. Figure 1 shows that the best average results is in the case of the new (K1) model, but the difference is not considered statistically validated, which is supported by the paired Wilcoxon test between K1 and K2 ($p = 0.094$).

The results of each model have been expanded according to the ratio of the forecast and observation period ($t/T \in \{0.5, 1, 2\}$), so three groups have been created for each model. Figure 2 shows the result. Comparing the first two models it can be observed that the performance of the second model decreases in the third case. In the third case ($t/T = 2$) the median of K2 is significantly

[8]If the customer did not purchase in the second $T/2$ period, than he/she had become inactive in the first $T/2$ period.



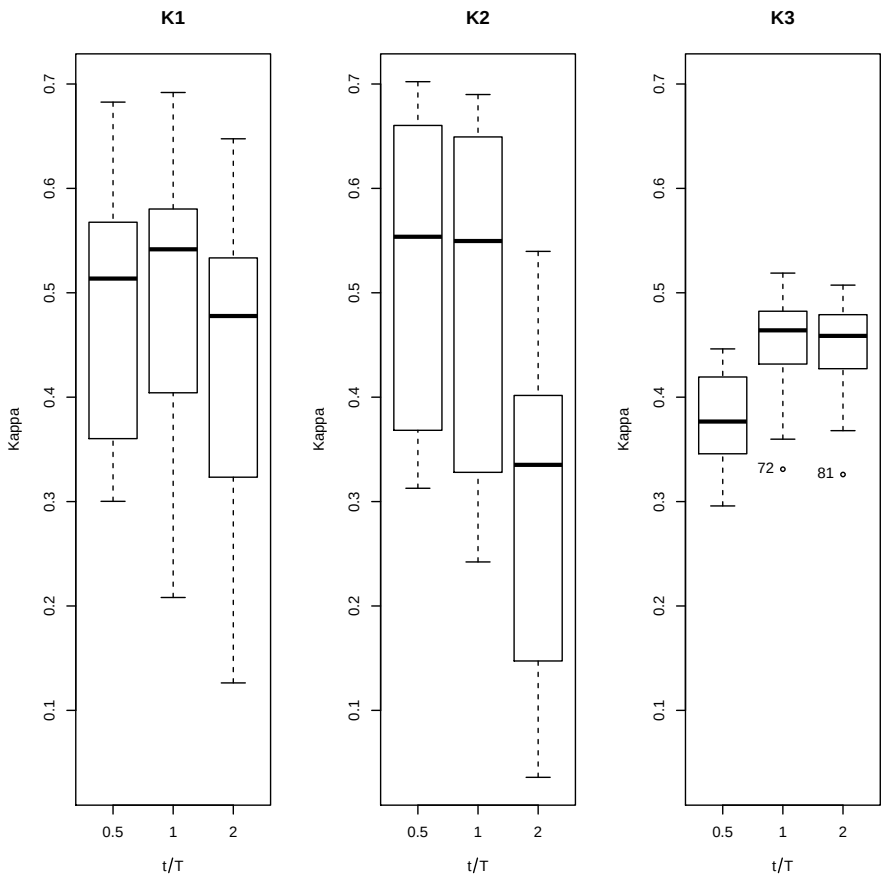
K1: the values of the kappa statistics for the new model,
K2: the values of the kappa statistics for the BG/NBD model,
K3: the values of the kappa statistics for the heuristic model.

Figure 1: The values of the kappa statistics for the three model. Source: own compilation.

different form the median of K1, which has been confirmed by the Wilcoxon test ($p = 7.451\text{e-}08$). In the other two cases ($t/T = 0.5$, and $t/T = 1$): the differences are statistically significant in the first case but in the second case they are not[9].
In other words, the new model is proved to be more reliable for the long term forecast than the BG/NBD model.

Comparing the third model with the other two ones we can see large differences in standard deviations (Figure 2). The results of the heuristic model – in spite of the smaller standard deviation – seem to be poor. (Kappa $\in [0.3; 0.5]$). In contrast, the results of the other two models are spread from very poor (0.1) to good (0.7). If we examine the differences of the means of the new and the heuristic model , the difference which can be observed in Figure 1 is not statistically significant ($p = 0.006$), while among the differences which can be observed in Figure 2 the first one is significant ($p = 6.3\text{e-}05$), but the second and third ones are not ($p = 0.229, p = 0.878$). Based on these results the new model has given better forecasts than the heuristic model.

[9]The p values of the paired Wilcoxon test are $p = 3.1\text{e-}06$ and $p = 0.628$.



K1: the values of the kappa statistics for the new model,
K2: the values of the kappa statistics for the BG/NBD model,
K3: the values of the kappa statistics for the heuristic model.

Figure 2: The values of the kappa statistics for the three model with different t/T ratios. Source: own compilation.

This (first) examination looked at the ability of each model to predict the churn of customers by the end of the observation period. Here, of course, not only the number of churns is important, but also exactly the person who will terminate the connection. There are comparative tests [Persentili Batislam, 2007; Fader, 2005], which have been carried out only group level comparisons instead of individual level one. The good value of such an indicator does not mean necessarily a good solution, since it is possible that there has not been hit on the individual level at all, but good results has been reached on the group level. If our goal is to forecast the future activity of each observation unit (customer), it is necessary to use the individual level indicators.

3.2.8 Comparision of the differences between the estimated and actual number of purchases

In this subsection the models have been examined by an index that represents the hit accuracies for each observation unit and the averages of these values have been compared. There is some index available for this purpose, and one

of them is the mean absolute error (MAE).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_{\text{pred}} - y_{\text{actual}}| \tag{7}$$

where
n: number of the observation units,
y_{pred}: predicted value (number of purchases) for period t,
y_{actual}: actual value (number of purchases) for period t.

The MAE values are represented on boxplot chart again and the groups are split by the t/T rates as factors (Figure 3). In this case the significance of the differences has also been examined. Here the new and BG / NBD models

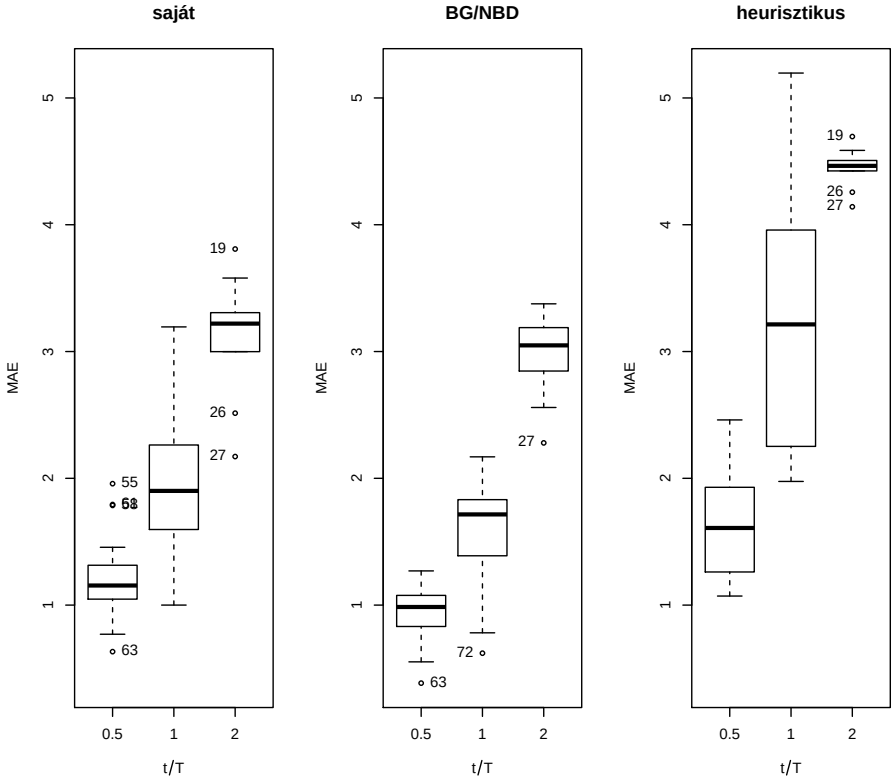


Figure 3: The values of the MAE index in case of different t/T ratios for the three models. Source: own compilation.

have been compared, because the figure shows that the heuristic model gave a weaker result than the other two models.[10]

The paired t-test has been used for the comparison of the two models. In the first and the second case ($t/T = 0.5$ and $t/T = 1$) the differences of the means are significant (at 5% level), while in the third case ($t/T = 2$) – on the basis of the test – the averages can be considered as equals.

[10]The higher the value of the index the less the accuracy of the model. See the definition of MAE.

It can also be observed that the BG/NBD model has not given appreciable (very poor) results in some cases. When considering these cases, the common is that the t/T is equal to 2 for each cases. Which means that the long term forecasts are uncertain. That is, if the result is acceptable in such a case, it is similar to the result of the new model, but also many times the result of the BG/NBD can not be evaluated.

3.2.9 Determination of the best future customers

In this third examination it has been analyzed the ability of each model to predict the future top 200 customers (i.e. the top 20%). The 'top' means those who will have the most purchases in the forecast period (t). The knowledge of the future profitability of the customers is important information for the company[11]. This is supported by Homburg [2008]. They established on the basis of their calculations that the distinction between customers increase the average profitability. In this model the best buyer is who has the most purchases during a specified time period (t).

The collected data contain the number of customers (for all the three models) whom the prediction has been successful, i.e., has been included into the actual top 200. The results of the calculations have also been represented on boxplot chart and the groups have been split by the t/T rates as factors (Figure 4).

To compare the three models the paired Wilcoxon test has also been used. The difference between the medians could be detected statistically in four cases: between the new and the BG/NBD model in the cases $t/T = 0.5$ and $t/T = 1$, and between the heuristic and the BG/NBD model in the same cases. This means that the BG/NBD model has achieved a significantly better average result in forecasting for a relatively shorter period than the new model. However, in the longer-term forecast the results of the new model show a better forecast (although this difference is not statistically significant, $p = 0.1698$).

The results of the test are still important for the comparison of heuristic and probabilistic models. Huang [2012] explored the predictive ability of two such models (namely the heuristic and the Pareto / NBD models). He also performed the calculations with artificial databases, and noted that, in the majority of the calculations the simple heuristics outperform the probability model. The present calculations, however, do not show this. Here, the averages of the results of the probability models are at least as good as the heuristic

[11] Since the model does not include the value of the purchases, so the profitability of the customer has been measured by the number of purchases

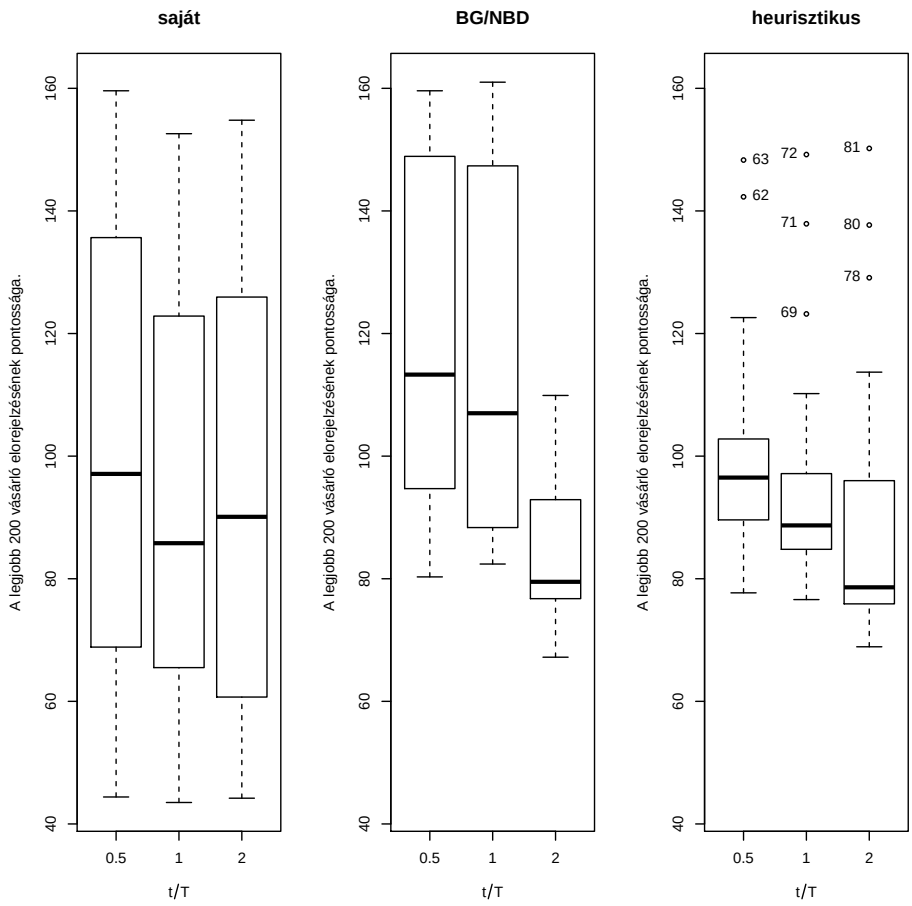


Figure 4: The forecast values of the best 200 customers in case of different t/T ratios for the three models. Source: own compilation.

model.

3.3 New research findings

1. Using empirical and theoretical analyses have been established, that the determination of $\boldsymbol{m}_{ij}$ cut-point developed by Tong [2009] – in cases when the number of the element of the two clusters are substantially different from each other – is inadequate. This is important for the calculation of density related sub-indices ($Dens_{bw}$).

2. Function f^{**} has been created (Equation 3), which determines the number of the observation units in a given environment of the selected points (the cluster centers and the $\boldsymbol{m}_{ij}$ point). By the help of this function the S_Dbw^{**} index (Equation 4) has been created from the S_Dbw_{new} index. The indices have been compared with theoretical and simulation studies. It can be concluded that the new index has given better results in case of overlapping or unequal cluster arrangement (different number of element in the clusters), therefore it is more suitable for decision support.

3. A new model have been created (development of the BG/NBD model) that is suitable for predicting the customer churn and the number of purchases and taking into account the complaints of the acquisition and the management of those as well. The new model have been tested by simulations, using scripts written in the R environment and artificially generated databases. Based on these tests, the new model proved to be accurate on the longer-term forecasts, but the shorter-term forecasts produced similar to or slightly lower results as for the BG/NBD model.

4. The results of the new and the BG/NBD model have been compared with the results of a so-called heuristic model (which are the most used models in practice). The results show, that the predictions of the probability models are more accurate than the predictions of the heuristic model, particularly in the case of the estimated number of purchases. Thus the applicability of the probabilistic models and the importance of such researches has been supported.

4 Conclusions and recommendations

1. The aim of the experiments was to analyse the best solution so far for determination of the number of clusters : can this index provide appropriate assistance to the decision maker in extreme conditions (e.g. overlapping or closely spaced clusters)? The experience was that the authors did not pay attention to this investigation.

   The aim was that the new index can be used for databases that are not completely separated, because usually such databases occur in everyday practice.

2. The new index has given better results on the test databases than the best existing index, even in the case of more realistic cluster configurations. However, the result depends on the clustering algorithm has been chosen, as well. There has been used two clustering methods (K-means and Ward). In most cases they have found the structure of the databases (i.e. the valid number of clusters). The essence of this research was to find the real clusters among the possible results which has been made by clustering algorithms. If the possible results do not contain the real classification, then the index will choose one of these results, which is not the real one.

   In this research the databases contain two variables because of the possibility of the visual verification. If the database contains more than two variables, the calculation of this index is also possible. In this case, of

course, there is no opportunity to verify the result visually. Since the experiments contain paired comparisons (pairs of clusters), the number of the necessary calculations is large when the number of clusters is also large. But in the field of the marketing research the databases do not consists of too many clusters, so this problem has not been examined in this paper.

3. It can be seen by the comparison of the two models (BG/NBD and modified BG/NBD), that the investigation of the new variables have yielded better results partially. The usefulness of the inclusion of additional variables was questionable, because on the one hand the more data allow to understand the reality better, on the other hand, the model is more complex (the number of parameters that need to be determined increasing). The more complex model means, that the parameter estimations may be more uncertain, thus the forecasting power of the model will be weaker.

   The database has been created on the basis of the theory of the new model, so it was assumed that the new model therefore provides a more accurate prediction. That did not happen. That is a simple model was essentially the same effectiveness in the forecast (in the short-term), despite the fact that it used less information.

   On the other hand, the number of complaints has been kept in a realistic range. But it has not been significant effect on the results, that is, without this information the BG/NBD model has led to similar results. Then the connection between the number of complaints and the accuracy of the models have been examined. The correlation between them has not been observed.

4. Since the calculations have been performed on 81 different databases (each model has been run 10 times for all databases), the statement – i.e. the probability model outperforms the heuristic one – has been verified empirically. Of course, there is a further question. Whether there are additional benefits for probabilistic model that let it be worth to use? There is a big difference between the two models (conceptual difficulties, practical difficulties). Since the new model did not consider the value of purchases (only the number of purchases), so this question can not be answered within the scope of this paper.

   As can be seen, the standard deviation of the probability models was greater than the sd. of the heuristic model for both tests, so the results can vary between wide limits. So the conclusion must be considered carefully. If the researcher has a database, then it's possible to do more

databases from this one (e.g. bagging) and performs the calculations on each database. The decision can be taken by evaluating these results.

All the 81 databases contain the data of 1000 customers. The sample was found sufficiently large to accept the results. Similar sized databases are used in marketing research practice.

Related publications of the author

Scientific journal (English language)

- Ruff Ferenc (2012): Empirical comparison of a model based and a non model based clustering methods. *Annals of The Polish Association of Agricultural and Agrobusiness Economists*. Vol. XIV. No.6. 242-246 p.
- Ruff Ferenc (2008): Methodological problems of classification and prediction in food marketing. *Annals of The Polish Association of Agricultural and Agrobusiness Economists*. Vol. X. No.5. 125-129 p. ISSN 1508-3535.

Scientific journal (Hungarian language)

- Ruff Ferenc (2013): Klaszterszámok meghatározásának egy lehetséges megoldása. *Szigma*. XLIV. évf. 3-4. szám. 135-153. p.

Proceedings of scientific conferences, full paper, English language

- Ruff Ferenc (2014): Clustering Methods for Ordinal Variables. Economics Questions, Issues and Problems. Komarno, Konferencia kiadvány 274-279 p. ISBN 978-80-89691-07-4. http://www.irisro.org/economics2014january/55RuffFerenc.pdf
- Ruff Ferenc, Szelényi László (2006): Environmental decision problems and operational research. X. Nemzetközi Agrárökonómiai Tudományos Napok. Gyöngyös, 2006. márc. 30-31. Konferencia CD: \Poszter \krf110. 1-6. p. ISBN 9632296230

Proceedings of scientific conferences, full paper, Hungarian language

- Pitlik László, Ruff Ferenc (2011): Táplálkozási tanácsadó szimulátor fejlesztése, avagy modellezési stratégiák összehasonlító elemzése. IX. Magyar Biometriai, Biomatematikai és Bioinformatikai Konferencia. 2011. július 1., Budapest. Absztrakt: Program, Előadás- és poszterkivonatok, Résztvevők listája (konferencia kiadvány). 20. p.
- Szelényi László, Bedéné Szőke Éva, Ruff Ferenc, Vinogradov Szergej (2004): Agrárökonómiai elemzések többváltozós módszerekkel. XXX. Óvári Tudományos Napok. In: Gazdasági informatika szekció. Mosonmagyaróvár, 2004. október 7. Konferencia CD: aokonomia \ Szelenyi.pdf. 1-5 p.
- Szelényi László, Ruff Ferenc, Bedéné Szőke Éva (2004): Környezetvédelmi mutatók többváltozós elemzése. Környezetgazdálkodási szekció. IX.

Nemzetközi Agrárökonómiai Tudományos Napok, Gyöngyös, 2004. március 25-26. Konferencia CD: 3.Környezetgazdálkodás\6\Szelényi, László - Ruff, Ferenc-Bedéné Szőke, Éva.doc. 1-6. p.

– Szelényi László, Bedéné Szőke Éva, Ruff Ferenc (2003): A vidékfejlesztés helyzetének többváltozós elemzése. Agrárgazdaság, Vidékfejlesztés és Agrárinformatika az évezred küszöbén /AVA nemzetközi konferencia 2003. április 01-02. Debrecen. Konferencia CD: cd\pdf\D098.pdf. 1-6. p.

Proceedings of scientific conferences, abstract, Hungarian language

– Pitlik László, Ruff Ferenc (2011): Táplálkozási tanácsadó szimulátor fejlesztése, avagy modellezési stratégiák összehasonlító elemzése. IX. Magyar Biometriai, Biomatematikai és Bioinformatikai Konferencia. 2011. július 1., Budapest. Absztrakt: Program, Előadás- és poszterkivonatok, Résztvevők listája (konferencia kiadvány). 20. p.

Book, chapter of book, booklet

– Ruff Ferenc (2002): A legjobban illeszkedő függvény-típus kiválasztása. 317-318, 537-539 p. In: Szűcs István (szerk): *Alkalmazott statisztika.* Agroinform Kiadó, Budapest. 551 p.

Other publications

– Pitlik László, Ruff Ferenc (2011): Development of nutrition simulator or comparison modeling approaches. Magyar Internetes Agrárinformatikai Újság. 2011. No 160. 1-33. p. HU-ISSN-1419-1652
http://miau.gau.hu/miau/160/saltseer.doc

– Pitlik László, Ruff Ferenc (2008): „Konzisztencia-gyár", avagy stratégiai és operatív ajánlások a modellezés automatizálásához. Magyar Internetes Agrárinformatikai Újság. 2008. No 119. 1-36. p. HU-ISSN-1419-1652
http://miau.gau.hu/miau/119/cikk_plrf.doc

Research report

– Szűcs István, Farkasné dr. Fekete Mária, Széles Zsuzsanna, Ruff Ferenc: A földhasználat és a földjáradék összefüggései 43362 sz. OTKA kutatási téma zárójelentése 2007. 22 p.

– Szelényi László, Ruff Ferenc, Bedéné Szőke Éva, Vinogradov Szergej: A környezetvédelem jelenlegi helyzetének korszerű többváltozós ökonometriai módszerek felhasználásával történő elemzése és értékelése, a kom-

plex összefüggések feltárása. Közcélú környezet- és természetvédelmi feladat, zárójelentés. Gödöllő, 2005. 55 p.

– Szelényi László, Szűcs István, Ruff Ferenc, Bedéné Szőke Éva, Szergej Vinogradov: Az agrárgazdaság prognosztizálását segítő programozási modellek és termelési függvények kidolgozása, A/0129/2003 sz. OKTK kutatási téma zárójelentése, SZIE, Gödöllő, 2004. 49 p.

– Szűcs István (szerk): Kedvezőtlen adottságú térségek lehatárolásának előkészítése. FVM tanulmány, Gödöllő, 2000. 80 p.