Comparative Analysis of Two Approaches to the Clustering of Respondents (based on Survey Results)

Iryna Strutynska, Halyna Kozbur, Lesia Dmytrotsa, Olha Hlado, Liliya Melnyk

Ternopil Ivan Puluj National Technical University, Ternopil, Ukraine
{ringtons999, dmytrotsa.lesya, kozbur.galina}@ gmail.com

Abstract. This paper proposes an algorithm for solving the survey respondents' clustering problem, including the steps of collecting, preparing data, summarizing key results, and developing future goals. The research consists of two approaches to clustering: iterative and hierarchical in order to produce consistent and comprehensible results. The iterative method is implemented in MS Excel using the Data Mining add-in, hierarchical one is used with the help of writing code and using Python libraries. Hard clusters with sufficient degree of similarity within the cluster and differences from others were distinguished, the main characteristics of the obtained clusters were described as well. It has been experimentally established that the method of agglomerative hierarchical clustering is more effective for solving the problem of clustering of mixed-type data obtained from the survey of respondents.

Keywords: digital maturity, clustering methods, mixed-type data

1 Introduction

The fastest and the most convenient way to get any information you need today is to directly interview your target audience on a specific topic. With the development of information technology, such questionnaires are increasingly shifting from personal or telephone communication to online questionnaires. This allows you to reach a larger audience in a shorter time span and with fewer human resources. The positive aspects of such surveys are: convenience of expression; partial or complete anonymity of results; the ability to complete a survey in any convenient for the respondent way; no need to communicate with the employees of the survey organization, etc. Online surveys are a particularly effective way of retrieving information if your target audience is the users of the web. Data collection is only part of the complex task of getting the information you need. Further processing and analysis of data with conclusions and recommendations make the data cycle complete. Segmentation or clustering is one of the most important and interesting tasks of data analysis. This paper offers an algorithm for solving the problem of clustering respondents by online survey, including the steps of collecting, preparing data, summarizing key findings, and developing future goals.

The problem of clustering of numerical data as a result of a series of measurements was described by [1]. A similar grouping of respondents was addressed in the works of [2]. The main purpose of the work was to provide recommendations on the results of determining the respondents' political preferences and to compare the clustering method with other generally accepted methods of providing such recommendations. The problem of clustering of categorical data using probabilistic approach and GACUC algorithm was investigated by [3]. The main statements regarding the clustering of mixed-type data and applying of chosen in the paper algorithms and metrics were discussed in [4,5,6,7]. However, the problem of processing and clustering of mixed data obtained from the questionnaire has not been researched so far.

This paper studies clustering of respondents using two approaches: iterative and hierarchical in order to produce consistent and comprehensible results. The iterative method is implemented in MS Excel using the Data Mining add-in, hierarchical one is used with the help of writing code and using Python libraries.

The survey, the results of which were taken as inputs to the clustering task, was conducted among small and medium-sized enterprises in Ternopil region for the use of digital technologies and tools in their business activities [8].

2 The problem of clustering. Approach typing and solution algorithms

Clustering or cluster data analysis is one of the machine learning tasks of splitting multiple objects into subsets (clusters) so that the objects assigned to one cluster are as similar as possible to each other and the objects referred to different kinds are as different as possible. This approach does not require a labeled data.

One of the most common contemporary tasks that uses cluster analysis is text analysis for news broadcasting, image grouping, consumer segmentation, community identification on social networks, etc.

The variability of tasks, types of datasets and expected results has led to the formation of a large number of methods and approaches to clustering, which differ in their understanding of the concept of "cluster", as well as adjusting the parameters of algorithms (number of expected clusters, density threshold, distance metrics, etc.) the specifics of the dataset and the subsequent use of the results. Thus, this makes it difficult to uniquely select the algorithm of operation and its parameters for each type of task.

Due to this, clustering can also be called an interactive task of machine learning "with reinforcement", which provides repeated experimental correction of algorithm parameters for obtaining stable and interpretative results [9,10].

There is no single common way to classify clustering methods and algorithms. One approach is to distinguish clustering methods by cluster models used (connectivity-based or hierarchical, centroid-based, distribution-based, density-based overlapping clustering etc.). Another approach uses grouping of methods based on their key characteristics (probabilistic, logical, graph-theoretic, hierarchical, neural, frequency algorithms, etc.) [10].

One of the simplest approaches to clustering methods is to divide them into two groups: hierarchical and non-hierarchical. Hierarchical cluster analysis methods are divided into ascending or descending and can be represented graphically in the form of dendrograms. At the same time with each subsequent step the number of clusters increases or decreases depending on the chosen method: divisional or agglomerative respectively.

The largest group among non-hierarchical methods is iterative. In an iterative approach, they define cluster centers and redistribute the elements of the data set by proximity to the selected centers. These include algorithms k-means, Expectation-Maximization method, mean-shift and others [11].

The similarity of cluster elements and the closeness of clusters are determined by predefined metrics.

3 Data Collection

The input data set was obtained through Google Forms questionnaires from executives in various companies and enterprises (limited liability companies (LLC), individual entrepreneur) and businesses (construction, trade, repair, logistics, services, etc.)

Respondents answered 35 questions regarding two main aspects of running their business:

- forms, organizations and spheres of activity;
- level of informatization of business activities (use of digital tools in their work, work with social networks, planning services, analytics or advertising).

Due to the specifics of the information requested and the use of different categories of questions, both quantitative and categorical data were received. For example, information about the number of employees in an organization was obtained in the form of natural numbers, and information about the presence of a business model was presented as a binary “yes” or “no” answer. There were some open-ended questions regarding the respondent’s attitude to a particular problem related to informatization of the business structure. Such responses were excluded from the general clustering dataset.

Numerous mechanical errors and blank answers were found in the data retrieval. These problems were solved with manual processing. However, as the number of respondents increases, such processing will require the unification of the possible answer options for each of the questions or the reduction of all possible answers only to the choice of the suggested ones.

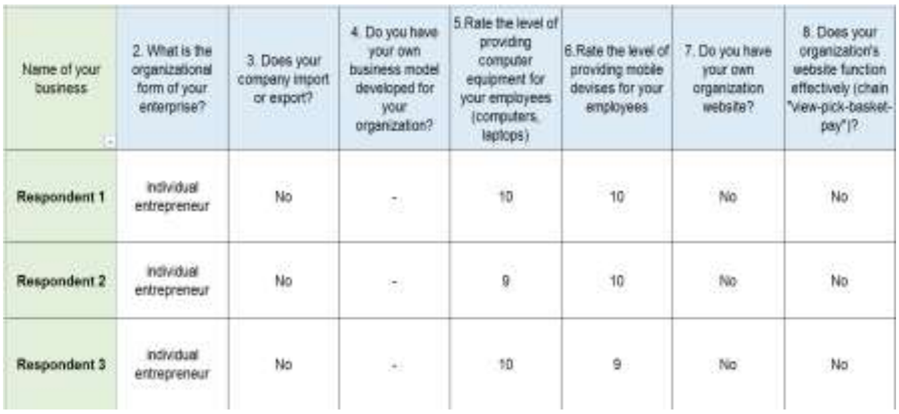
4 Data Preparation

Data preparation consisted of the steps of clearing and encoding data, missing values were not identified in this study.

4.1 Clearing data

Both the hierarchical agglomerative algorithm and the EM method were run for the same data set, so pre-processing due to data cleaning was performed equally for both methods.

The answers to the open-ended questions were reduced to a specific template, for example, only “yes” or “no”, or otherwise unified, for example how is shown in Fig 1. Attributes of the “automatically calculated questionnaire time” or “respondent’s personal attitude” were marked as informative and removed from the task input. All manipulations were performed manually due to the small dimension of the task.



Name of your business	2. What is the organizational form of your enterprise?	3. Does your company import or export?	4. Do you have your own business model developed for your organization?	5.Rate the level of providing computer equipment for your employees (computers, laptops)	6.Rate the level of providing mobile devises for your employees	7. Do you have your own organization website?	8. Does your organization's website function effectively (chain "view-pick-basket-pay")?
Respondent 1	individual entrepreneur	No	-	10	10	No	No
Respondent 2	individual entrepreneur	No	-	9	10	No	No
Respondent 3	individual entrepreneur	No	-	10	9	No	No

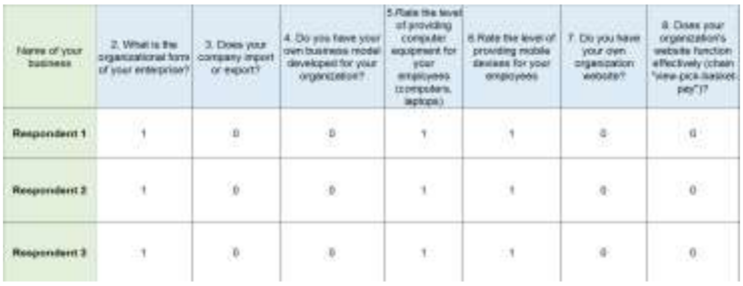
Fig.1. Respondents' answers to the questionnaire

4.2 Data encoding

Using the MS Excel add-on requires no special training and accepts a simple spreadsheet of values of any type. Therefore, data encoding was not performed for clustering with the Data Mining add-in in MS Excel.

Data encryption was required to work with Python machine learning libraries since most algorithms use mathematical operations on quantitative data. For those questions where it was possible to rank more or less or better or worse, ranked value coding was used. Responses were coded from 0 to some positive number, where 0 meant a single answer “no” or a number close to 0, and other values were ranked according to the increase in manifestation of the sign.

Non-ranking answer options are nominal type data and have been indicated by some character numbers. For further computational work of the algorithm with such data, the Hower metric was used, which makes it possible to work with both quantitative and categorical numerical data at the same time. Respondents' coded answers are shown in Fig. 2.



Name of your business	2. What is the organizational form of your enterprise?	3. Does your company import or export?	4. Do you have your own business model developed for your organization?	5.Rate the level of providing computer equipment for your employees (computers, laptops)	6.Rate the level of providing mobile devices for your employees	7. Do you have your own organization website?	8. Does your organization's website function effectively (chain "view-pick-basket-pay")?
Respondent 1	1	0	0	1	1	0	0
Respondent 2	1	0	0	1	1	0	0
Respondent 3	1	0	0	1	1	0	0

Fig 2. The survey respondents' coded answers

5 Solving the clustering problem

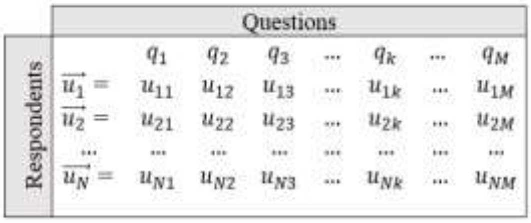
The main objectives of the study were:

- experimental finding of the optimal number of clusters and their characteristic features for the interpreted (understandable) segmentation of business structures according to the level of digital maturity by several methods;
- comparing the results obtained by different methods and determining the most effective for a particular data analysis task.

The study used two methods of clustering:

1. Using the Data Mining add-in for MS Excel spreadsheets [13]. Clustering capabilities in MS Excel are represented by iterative algorithms: k-means and Expectation-Maximization. For the reference, it was determined EM-algorithm;
2. Using the functions of libraries for machine learning Python programming language [14,15].

To describe how it works with two algorithms, we have introduced the notation: N respondents $U = \{\overrightarrow{u_1}, \overrightarrow{u_2}, \dots, \overrightarrow{u_N}\}$ and M questions $Q = \{q_1, q_2, \dots, q_M\}$. Every participant $\overrightarrow{u_i} \in U$ $(l \in \overline{1, N})$ answered each of the questions $q_k \in Q$ $(k \in \overline{1, M})$, so the result is a matrix of responses with dimension$(N \times M)$, in which each respondent is represented as follows: $\overrightarrow{u_l} = \{u_{l1}, u_{l2}, \dots, u_{lk}, \dots, u_{lM}\}$, where u_{lk} is an answer l-respondents to k-question (Fig. 3). In the future, we call this tuple a point [8].



$$\begin{array}{c|ccccccc} & \multicolumn{7}{c}{\text{Questions}} \\ \hline \text{Respondents} & & q_1 & q_2 & q_3 & \dots & q_k & \dots & q_M \\ & \overrightarrow{u_1} = & u_{11} & u_{12} & u_{13} & \dots & u_{1k} & \dots & u_{1M} \\ & \overrightarrow{u_2} = & u_{21} & u_{22} & u_{23} & \dots & u_{2k} & \dots & u_{2M} \\ & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & \overrightarrow{u_N} = & u_{N1} & u_{N2} & u_{N3} & \dots & u_{Nk} & \dots & u_{NM} \end{array}$$

Fig. 3. The matrix of answers

Let us consider the principles of the selected methods.

5.1 Method of hierarchical agglomeration

The principle of operation of the modified agglomerative method is described in detail by [8]. According to the agglomerative approach to clustering, each point is considered to be a separate cluster at the beginning. As the algorithm works, each of the two closest clusters is merged at each step, eventually forming a predefined number of clusters or merging into one. To get started with the agglomerative algorithm, we build a matrix of pairwise distances between the objects of the cluster. In the context of the problem, the Hower metric (1) proposed in [8] was used to calculate the distance matrix.

$$d\left(\overrightarrow{u_\imath},\overrightarrow{u_\jmath}\right) = \frac{1}{M}\sum_{k=1}^{M} d_{ijk}, \tag{1}$$

where $d_{ijk} = d\left(u_{ik},\ u_{jk}\right)$ – the distance between the answers in the k-th question, M is the number of answers to the query in the tuple. The distance matrix D_k for the k-th question is symmetric:

$$D_k = \begin{array}{|c|c|c|c|c|}\hline 0 & d_{12k} & d_{13k} & \dots & d_{1Nk} \\ \hline & 0 & d_{23k} & \dots & d_{2Nk} \\ \hline & & 0 & \dots & d_{3Nk} \\ \hline & & & \dots & \dots \\ \hline & & & \dots & 0 \\ \hline \end{array}$$

The symmetric matrix D for distances between individual points of the cluster looks like:

$$D = \begin{array}{|c|c|c|c|c|}\hline 0 & d\left(\overrightarrow{u_1},\overrightarrow{u_2}\right) & d\left(\overrightarrow{u_1},\overrightarrow{u_3}\right) & \dots & d\left(\overrightarrow{u_1},\overrightarrow{u_N}\right) \\ \hline & 0 & d\left(\overrightarrow{u_2},\overrightarrow{u_3}\right) & \dots & d\left(\overrightarrow{u_2},\overrightarrow{u_N}\right) \\ \hline & & 0 & \dots & d\left(\overrightarrow{u_3},\overrightarrow{u_N}\right) \\ \hline & & & \dots & \dots \\ \hline & & & \dots & 0 \\ \hline \end{array}$$

The elements of the matrix D are the averaged values of the pairwise values of the distances calculated by the formula (1). All questionnaire weights are taken to be 1.

The way if measuring distances d_{ijk} depend on the type of data in k question. If u_{ik} and u_{jk} quantitative, then distance d_{ijk} is expressed by the formula (2):

$$d_{ijk} = \frac{\left|u_{ik}-u_{jk}\right|}{\max\left(u_k\right)-\min\left(u_k\right)}. \tag{2}$$

In this case, $d(\vec{u_i},\vec{u_j}) \in [0;1]$. If u_{ik} and u_{jk} – nominal data that cannot be ordered, then the distance is calculated by the formula (3):

$$d_{ijk} = \begin{cases} 0, & u_{ik} = u_{jk}, \\ 1, & u_{ik} \neq u_{jk}. \end{cases} \tag{3}$$

In both cases $d_{ijk} = 0$ means identical answers of the respondents u_j to k question, and $d_{ijk} = 1$ – maximal difference. As a consequence, for the averaged distances calculated by formula (1), all values$d(\vec{u_i},\vec{u_j}) \in [0;1]$.

The distance between the individual clusters was by the distance neighbour method. Clusters closest to the selected metric are merged, distances from newly created to other clusters are recalculated, the distance matrix is automatically updated, and clustering continues. The method of the far neighbour allows to allocate rather compact and stable structures corresponding to the task.

5.2 Expectation-Maximization method

In contrast to the proposed modification of the agglomerative method, the fuzzy clustering EM algorithm presented in the Data Mining Add-in for Microsoft Excel was selected among the iterative algorithms. In this case, the main idea of the method is to assume that the elements of the input data set are independent random variables distributed by a law, in most cases a normal Gaussian distribution [16,17,18,19].

When using the EM method, any object in the dataset is considered to belong to all clusters with different probabilities. Before starting the algorithm, the number of *K* clusters and the initial approximate parameters for each of the *K* distributions of the input data are specified. Iterations incrementally improve the distribution parameters to a predetermined level of model accuracy. Upon completion of the algorithm, each object will be assigned to a cluster with the highest probability of belonging. Thus, two successive steps are performed at each iteration:

Further, the algorithm is based on an iterative repetition of two consecutive steps as shown in the Fig. 4.

1. Expectation is calculating the probability (plausibility) of the points belonging to each of the clusters;
2. Maximization is improvement of distribution parameter values to maximize the likelihood of points belonging to clusters.

5.3 Adjustment of algorithm parameters

The MS Cluster Task Wizard in MS Excel allows you to select the desired parameters and adjust their values [13]. For this task, the list of questions that will affect the result was changed in the algorithm settings, the value of the number of clusters and the cluster seed of the EM clustering method were set.

Referring to the sklearn.cluster.AgglomerativeClustering method to create a clustering model using Python generally involves specifying 3 parameters: number of clusters, intra-cluster distance and inter-cluster distance metrics. The metric of the distance metric between the elements may be one of those proposed in [21] or otherwise calculated. Calling the function of creating a cluster model:

model = AgglomerativeClustering (number of clusters = m, metric = “precomputed”, linkage = complete)

labels = model.fit_predict(distances),

where m ─ predefined number of clusters, distances – distance matrix previously calculated by the Hower metric (1) – (3).

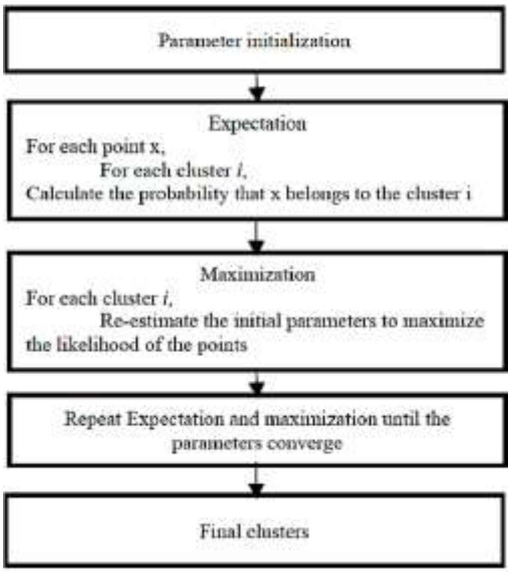

Parameter initialization
Expectation
For each point x,
For each cluster *i*,
Calculate the probability that x belongs to the cluster i
Maximization
For each cluster *i*,
Re-estimate the initial parameters to maximize the likelihood of the points
Repeat Expectation and maximization until the parameters converge
Final clusters

Fig. 4. Functional diagram of the EM algorithm, borrowed from [20]

6 Analysis of the results

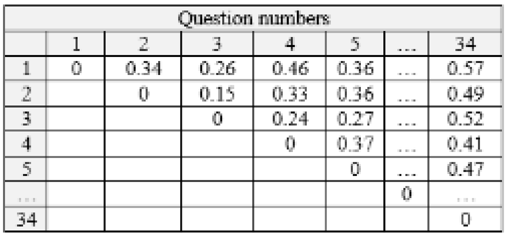
Let us consider the clustering results by each method and compare the results obtained. The clustering output of MS Excel's Data Mining add-on provides a breakdown of the dataset into clusters with the ability to visualize, view statistics, and cluster profiles.

Using the clustering methods of the sklearn library to analyse data on Python output provides a one-dimensional numeric array indicating which cluster each input tuple belongs to. Further analysis and visualization of the obtained result is carried out additionally. The algorithm for selecting the optimal number of clusters is described in [8].

6.1 The results of hierarchical agglomerative clustering

The matrix of distances *D* between the points of the input set now looks like this (the fragment of the matrix is shown in Table 1):

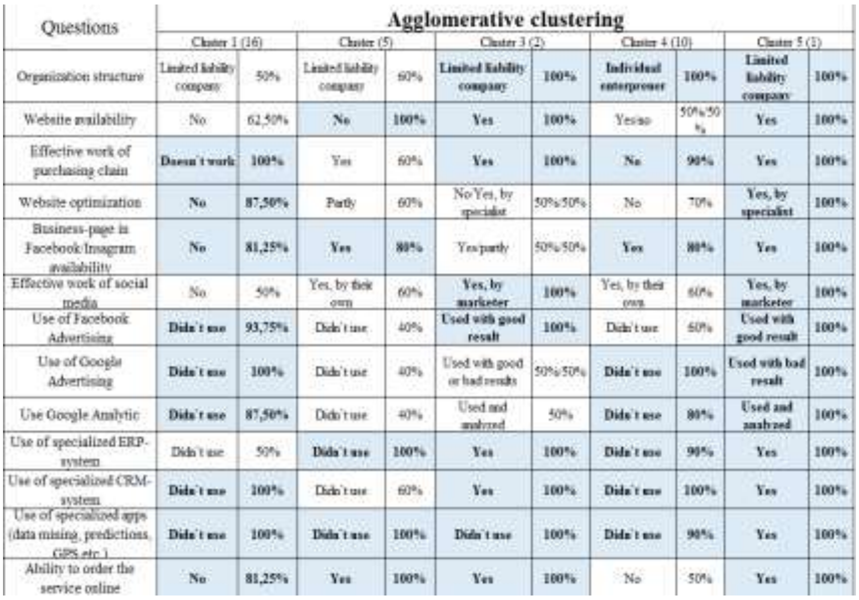
Table 1. Matrix of distances between questions



	Question numbers						
	1	2	3	4	5	…	34
1	0	0.34	0.26	0.46	0.36	…	0.57
2		0	0.15	0.33	0.36	…	0.49
3			0	0.24	0.27	…	0.52
4				0	0.37	…	0.41
5					0	…	0.47
…						0	…
34							0

As a result of the agglomerative clustering algorithm using Python sklearn, a stable distribution of 5 clusters was obtained, with a satisfactory value of the quality metric, the Silhouette index [22]: $sil \approx 0.16$. A comparative analysis of the clusters obtained by main characteristics is shown in Table 2.

Table 2. Comparative characteristics of clusters formed by agglomerative clustering with Python tools



Questions	Agglomerative clustering									
	Cluster 1 (16)		Cluster (5)		Cluster 3 (2)		Cluster 4 (10)		Cluster 5 (1)	
Organization structure	Limited liability company	50%	Limited liability company	60%	**Limited liability company**	**100%**	**Individual enterpreneur**	**100%**	**Limited liability company**	**100%**
Website availability	No	62,50%	**No**	**100%**	**Yes**	**100%**	Yes/no	50%/50 %	**Yes**	**100%**
Effective work of purchasing chain	**Doesn`t work**	**100%**	Yes	60%	**Yes**	**100%**	**No**	**90%**	**Yes**	**100%**
Website optimization	**No**	**87,50%**	Partly	60%	No/Yes, by specialist	50%/50%	No	70%	**Yes, by specialist**	**100%**
Business-page in Facebook/Imagram availability	**No**	**81,25%**	**Yes**	**80%**	Yes/partly	50%/50%	**Yes**	**80%**	**Yes**	**100%**
Effective work of social media	No	50%	Yes, by their own	60%	**Yes, by marketer**	**100%**	Yes, by their own	60%	**Yes, by marketer**	**100%**
Use of Facebook Advertising	**Didn`t use**	**93,75%**	Didn`t use	40%	**Used with good result**	**100%**	Didn`t use	60%	**Used with good result**	**100%**
Use of Google Advertising	**Didn`t use**	**100%**	Didn`t use	40%	Used with good or bad results	50%/50%	**Didn`t use**	**100%**	**Used with bad result**	**100%**
Use Google Analytic	**Didn`t use**	**87,50%**	Didn`t use	40%	Used and analyzed	50%	**Didn`t use**	**80%**	**Used and analyzed**	**100%**
Use of specialized ERP-system	Didn`t use	50%	**Didn`t use**	**100%**	**Yes**	**100%**	**Didn`t use**	**90%**	**Yes**	**100%**
Use of specialized CRM-system	**Didn`t use**	**100%**	Didn`t use	60%	**Yes**	**100%**	**Didn`t use**	**100%**	**Yes**	**100%**
Use of specialized apps (data mining, predictions, GPS etc.)	**Didn`t use**	**100%**	**Didn`t use**	**100%**	**Didn`t use**	**100%**	**Didn`t use**	**90%**	**Yes**	**100%**
Ability to order the service online	**No**	**81,25%**	**Yes**	**100%**	**Yes**	**100%**	No	50%	**Yes**	**100%**

The percentages indicate the proportion of respondents in each cluster who answered the same question the same way. The number of respondents who answered equally to the selected questions varied from 40 to 100%. To distinguish the characteristic features of the formed clusters, we leave the values greater than 80% and depict the comparison of the clusters.

According to the results, the largest number of respondents (16) was attributed to the first cluster, the main characteristics of which are:

- lack of experience with any digital tools;
- absence of companies in the Internet environment;
- site inefficiency, if any.

The second cluster was formed by 5 companies, the main characteristics of which are defined as follows:

- availability of companies in the Internet space;
- usage of simple tools to a limited extent.

2 more respondents formed the third cluster that is characterized by:

- effective functioning of the site and the purchase chain;
- use of most digital tools including advertising;
- the work of a marketer to promote a brand or product.

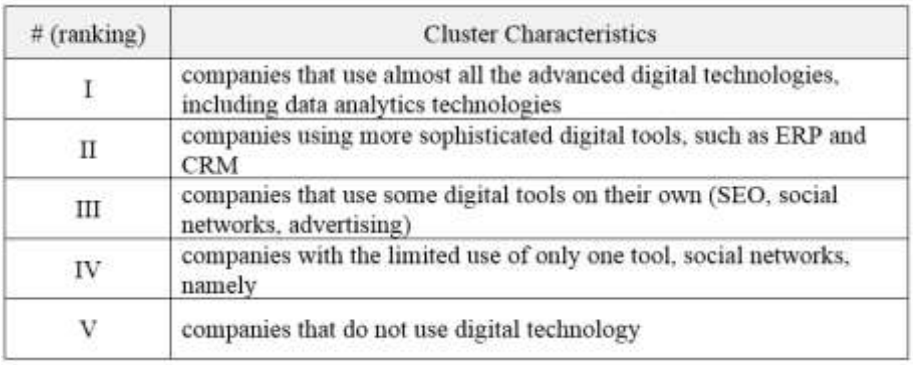
The fourth cluster consists of 10 respondents, and its characteristic features are:

- lack of functioning of the purchase chain on the site;
- non-use of sophisticated digital tools;
- availability of social networks only.

The fifth cluster consists of only one company that successfully uses virtually all digital tools with the help of specialists.

A sufficient degree of differences between clusters and a sufficient degree of similarity of elements within the cluster (80-100%) makes it possible to clearly identify the following groups and rank them by the level of use of digital technologies and tools in business activities. Table 3 shows the ranking of types of business structures by the decline in digital maturity.

Table 3. Ranking of business structure clusters by digital maturity level



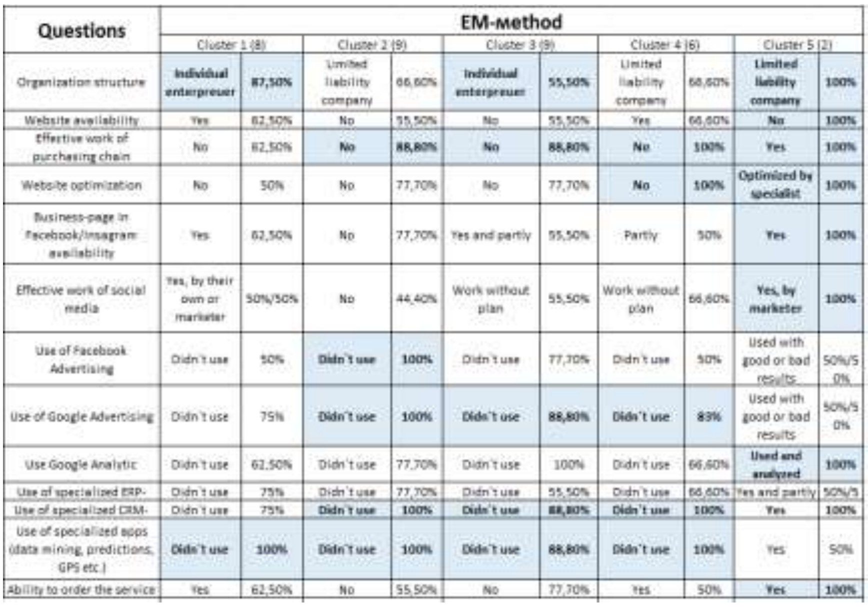
# (ranking)	Cluster Characteristics
I	companies that use almost all the advanced digital technologies, including data analytics technologies
II	companies using more sophisticated digital tools, such as ERP and CRM
III	companies that use some digital tools on their own (SEO, social networks, advertising)
IV	companies with the limited use of only one tool, social networks, namely
V	companies that do not use digital technology

6.2 The results of EM-clustering

The clustering performed by the EM method in MS Excel proved to be unstable. Because the EM algorithm is a group of iterative fuzzy clustering methods, this result is normal and suitable for use in a particular class of tasks. The comparative characteristics of the clusters obtained are shown in Table 4. In contrast to agglomerative clustering, the degree of uniformity of answers to questions within the clusters is much lower, and fluctuates on average within 60-70%. As we can see, the degree of difference between clusters is also low. Repeated application of the EM method did not improve the quality of the results.

Similar to the previous clustering method, a cluster was identified that included two business entities that are actively using digital technology in their businesses. The differences between the other clusters are small, the differences in the percentage of answers to the questions are minimal, so it is impossible to distinguish the characteristics of each subset. The inability to distinguish clusters with distinct features does not meet the objective of the study.

Table 4. Comparative characteristics of clusters formed by using an iterative approach



Questions	EM-method									
	Cluster 1 (8)		Cluster 2 (9)		Cluster 3 (9)		Cluster 4 (6)		Cluster 5 (2)	
Organization structure	**Individual enterpreuer**	**87,50%**	Limited liability company	66,60%	**Individual enterpreuer**	**55,50%**	Limited liability company	66,60%	**Limited liability company**	**100%**
Website availability	Yes	62,50%	No	55,50%	No	55,50%	Yes	66,60%	**No**	**100%**
Effective work of purchasing chain	No	62,50%	**No**	**88,80%**	**No**	**88,80%**	**No**	**100%**	**Yes**	**100%**
Website optimization	No	50%	No	77,70%	No	77,70%	**No**	**100%**	**Optimized by specialist**	**100%**
Business-page in Facebook/Insagram availability	Yes	62,50%	No	77,70%	Yes and partly	55,50%	Partly	50%	**Yes**	**100%**
Effective work of social media	Yes, by their own or marketer	50%/50%	No	44,40%	Work without plan	55,50%	Work without plan	66,60%	**Yes, by marketer**	**100%**
Use of Facebook Advertising	Didn`t use	50%	**Didn`t use**	**100%**	Didn`t use	77,70%	Didn`t use	50%	Used with good or bad results	50%/50%
Use of Google Advertising	Didn`t use	75%	**Didn`t use**	**100%**	**Didn`t use**	**88,80%**	**Didn`t use**	**83%**	Used with good or bad results	50%/50%
Use Google Analytic	Didn`t use	62,50%	Didn`t use	77,70%	Didn`t use	100%	Didn`t use	66,60%	**Used and analyzed**	**100%**
Use of specialized ERP-	Didn`t use	75%	Didn`t use	77,70%	Didn`t use	55,50%	Didn`t use	66,60%	Yes and partly	50%/5
Use of specialized CRM-	Didn`t use	75%	**Didn`t use**	**100%**	**Didn`t use**	**88,80%**	**Didn`t use**	**100%**	**Yes**	**100%**
Use of specialized apps (data mining, predictions, GPS etc.)	**Didn`t use**	**100%**	**Didn`t use**	**100%**	**Didn`t use**	**88,80%**	**Didn`t use**	**100%**	Yes	50%
Ability to order the service	Yes	62,50%	No	55,50%	No	77,70%	Yes	50%	**Yes**	**100%**

7 Conclusions and future work

In this study, we conducted an experimental comparison of the use of two approaches to the clustering of respondents according to online survey results using the Google Forms service, hard and soft clustering, in particular. Hard clustering was implemented with the use of Python tools and the hierarchical agglomerative method, while soft clustering was viewed through the use of the Data Mining add-in MS Excel and the iterative EM method.

A comparative analysis of the results obtained by the two methods showed the following results:

- Using hierarchical agglomerative clustering, we obtained 5 clusters, sufficiently different from each other and with a high degree of similarity between the elements of the cluster (60-100% depending on the question). The cluster features are distinguished (use of social networks, advertising offices and services, analytical tools, search engine optimization of sites, etc.);
- the use of the EM method did not allow to obtain good clustering results and to achieve the goal of the task, the results of the EM method implementation changed with each run of the algorithm.

It has been experimentally established that the method of agglomerative hierarchical clustering is an effective method for solving the problem of clustering of mixed-type data obtained from the survey of respondents.

In addition to improving the parameters of the algorithm, the tasks for the further studies are: elimination of mechanical errors when entering answers and the presence of empty values; diversity of data, which causes complexity of their unification and proper ordering; selection of mathematical metrics used as arguments for clustering functions and calculating their quality.

References

[1] Cherezov D., Tyuakchev N. Obzor osnovnykh metodov klassifikatsii I klasterizatsii dannykh [Overview main methods of data classification and clustering]. Vestnik VGU, Seria: Sistemnyi analiz i informachionnye tekhnologii, no 2. pp.25-29, 2009. Avaliable at: http://www.vestnik.vsu.ru/pdf/analiz/2009/02/2009-02-05.pdf (Accessed 14 August 2019). (In Russian).

[2] I. Katakis, N. Tsapatsoulis, C. Tziouvas and F. Mendes. (2012). Clustering Online Poll Data: Towards a Voting Assistance System, 2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization, http://www.katakis.eu/wp-content/uploads/2014/11/katakissmap12.pdf.doi.org/10.1109/SMAP.2012.19

[3] J. McCaffrey, Machine Learning Using C#. Syncfusion, 2014.

[4] Chae, S. S., Kim, J.-M. & Yang, W. Y., (2006). Cluster analysis with balancing weights on mixed-type data. The Korean communications in statistics, 13(3) doi.org/10.5351/CKSS.2006.13.3.719

[5] Gower JC (1967) A comparison of some methods of cluster analysis. Biometrics 23:623–637 doi.org/10.2307/2528417

[6] William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association. Vol. 66, No. 336 (Dec., 1971), pp. 846-850 doi.org/10.2307/2284239

[7] Gower, J. C., (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4), p. 859. doi.org/10.2307/2528823

[8] I.Strutynska, H.Kozbur, L.Dmytrotsa, I.Bodnarchuk and O.Hlado. (2019) Small and Medium Business Structures Clustering Method Based on Their Digital Maturity. 2019 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology, pp.278-282 (in print)

[9] Klasterizatsia [Clustering] *Wikiconspekts, ITMO University* [Online]. 2019. Avaliable: http://neerc.ifmo.ru/wiki/index.php?title=Кластеризация (in Russian)

[10] Cluster analysis, *Wikipedia* [Online]. Avaliable: https://en.wikipedia.org/ wiki/Cluster_analysis

[11] Zatserkovnyi V. I., Burachek V. H., Zhelezniak O. O., Tereshchenko A. O. Heoinformatsiini systemy i bazy danykh [Geoinformation systems and databases]. Nizhyn, Ukraine: NDU im. M. Hoholia, 2017. pp. 77-95. Avaliable at: https://studfiles.net/preview/6440954 (In Ukrainian).

[12] Google Forms, https://www.google.com/intl/uk_ua/forms

[13] Microsoft (2017 Dec). Cluster Wizard (Data Mining Add-ins for Excel), Microsoft Docs [On-line]. Avaliable: https://docs.microsoft.com/en-us/sql/analysis-services/cluster-wizard-data-mining-add-ins-for-excel?view=sql-server-2014

[14] Python, https://www.python.org.

[15] Scikit. Clustering documentation. *Scikit learn* [Online]. Avaliable: https://scikit-learn.org/stable/modules/clustering.html.

[16] Expectation-maximization algorithm, *Wikipedia* [Online]. Avaliable: https://en.wikipedia.org/wiki/Expectation-maximization_algorithm

[17] Yair Weiss. Bayesian motion estimation and segmentation. PhD thesis, Massachusetts Institute of Technology, May 1998.

[18] Zinchenko D. EM-algoritm klasterizatsii [EM-clustering algorithm], *AlgoWiki* [Online], 2016. Avaliable: https://algowiki-project.org/ru/Участник:Noite/EM-алгоритм_кластеризации

[19] Oreshkov V. EM – masshtabiruyemyy algoritm klasterizatsii [EM – scalable clustering algorithm], *BaseGroup Labs* [Online], 2013. Avaliable: https://basegroup.ru/community/articles/em

[20] Lotfi, E. (2018). [image] Available at: https://www.researchgate.net/profile/Elaachak_Lotfi/publication/322641344/figure/fig3/AS:585607373922305@1516631086753/flowchart-of-EM-algorithm.png [Accessed 15 Oct. 2019].

[21] Scikit. Agglomerative Clustering. *Scikit learn* [Online]. Avaliable: https://scikitlearn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[22] Scikit. Silhourtte Score. *Scikit learn* [Online]. Avaliable: https://scikit-learn.org/stable/modules/ generated/sklearn.metrics.silhouette_score.html