Analysis of Business Structures Regarding the Level of Digital Maturity Using Data Mining Methods

Iryna Strutynska[1,†], Halyna Kozbur[2,†], Olena Sorokivska[2,*,†], Lesia Dmytrotsa[2,†]and Ihor Kozbur[2,†]

[1] *The Netherlands Loughborough University London, 3 Lesney Avenue, Queen Elizabeth Olympic Park, London, E20 3BS, UK*

[2] *Ternopil Ivan Puluj National Technical University, Ruska 56 46001 Ternopil, Ukraine*

Abstract

Cluster analysis is proposed as an unsupervised machine learning method to divide small and medium-sized businesses in Ukraine into groups based on their level and types of digital maturity. The input data used is a dataset formed by expert assessments of the state of digital technology usage in regional small and medium-sized businesses. The Digital Transformation Index "HIT" is used to numerically measure the level of digital maturity of domestic enterprises. Various approaches to building clustering models are implemented using built-in methods in the scikit-learn library for Data Mining problems. The quality of the constructed models is evaluated using three indicators. Groups of companies are identified based on similarity in understanding digital development, and a comparative analysis is performed. Performing clustering for a representative sample of domestic small and medium-sized businesses will allow understanding the current state of their use of digital technologies and developing a well-reasoned system of actions to effectively digitize entrepreneurship in Ukraine.

Keywords

Data Mining Algorithms, Digital Transformation, ICT for Data Analysis, Scikit-learn, Clustering

1. Introduction

Digital transformation of small and medium-sized enterprises (SMEs) is a top priority for the development of the Organization for Economic Cooperation and Development (OECD). OECD policy tools, such as the "Digital Policy Framework" and the approved national program "Digitalization for Recovery in Ukraine", envisage that in the long-term perspective (2026-2032) Ukraine can focus on creating a sound data infrastructure for measuring the digital economy [1].

The processes of digital transformation in domestic SMEs – the transformation of their business strategies, models, operations, goals, marketing approaches, etc. towards increased use of digital technologies and improved efficiency, – are slow and underdeveloped. One of

[1]*BAIT'2024: The 1st International Workshop on "Bioinformatics and applied information technologies", October 02-04, 2024, Zboriv, Ukraine*

* Corresponding author.

† These authors contributed equally.

soroka220996@gmail.com (O. Sorokivska); strutynskairy@gmail.com (I. Strutynska); kozbur.galina@gmail.com (H. Kozbur); dmytrotsa.lesya@gmail.com (L. Dmytrotsa); kozbur.igor@gmail.com (I. Kozbur)

0000-0001-8549-2910 (O. Sorokivska); 0000-0001-5667-6569 (I. Strutynska); 0000-0003-3297-0776 (H. Kozbur); 0000-0003-2583-3271 (L. Dmytrotsa); 0000-0002-3113-0014 (I. Kozbur)

the problems is the lack of necessary knowledge among entrepreneurs regarding the application of innovative digital technologies, as well as the insufficient number of tools (platforms, services, or applications) that would allow them to assess the current level of digital maturity of individual enterprises and at the same time provide a roadmap of digital opportunities for business transformation. Clustering SMEs by the level and types of digital maturity will allow to understand the current state of digitalization, identify problem groups of enterprises and bottlenecks in the process of digital transformation, as well as recommend a reasoned systemic program of actions for effective digital growth.

2. Related works

The process of digitalization of business and the use of digital technologies in activities is the subject of many scientific studies. Thus, in the work of J. Cenamor, V. Parida, and J. Vincent, the relationship between the use of digital platforms and small business performance indicators is analyzed [2]. Features of the use of digital business models are highlighted in the works of N. Ivanchenko, Zh. Kudrytska, K. Rekachynska [3], N. Kraus, O. Holoborodka, K. Kraus [4]. Digital transformation is proposed to be considered as "processes that aim to improve an economic entity by triggering significant changes in its properties through a combination of information, computing, communications and connectivity" [5]. Digital transformation affects business processes, operational procedures, and organizational capabilities [6], requiring enterprises to update workforce skills, achieve a certain level of digital maturity, and improve productivity and efficiency.

R. Ochoa in [7] summarizes and forms the semantic core of literature reviews of various scientists regarding the definition of the digital maturity models. Domestic scientists pay attention to factors specific to Ukraine (in particular, the low level of digital literacy of society and cyber security, insufficient regulatory and legal regulation of digitalization), which reduce the interest of small businesses in the digitalization of business processes [8, p. 231; 9, p. 58]. In connection with this, an important direction of scientific research in the field of digitization is the study of the peculiarities of the formation of the digital space in Ukraine, as well as the participation of the state in the institutional and legal regulation of this process (O. Pishchulina [8], H. Zhekalo [9], H. Karcheva, D. Ohorodnia, and V. Open'ko [10]).

Investigating the use of digital tools by business organizations [11, 12], the authors developed methodologies for applying mathematical and computer modeling methods to measure the level of digital transformations [13, 14]. The main methodological tool of this study is cluster analysis. General problems of clustering are fully covered in the sources [15, 16]. Authors of scientific studies use diversified methods of cluster analysis, depending on the problem to be solved. Thus, in the scientific works of C. Iyigun, M. Türkeş, I. Batmaz, C. Yozgatligil, V. Purutçuoğlu, E. Kartal, M. Öztürk [17] and K. Sablin, E. Kagan, E. Chernova [18] use hierarchical clustering methods, K. Gorbatiuk, O. Mantalyuk, O. Proskurovych, O. Valkov in [19] study fuzzy clustering methods. Cluster analysis is often used in scientific works by both domestic and foreign authors to perform macro analysis, namely the differentiation of socio-economic development of regions. Works [20, 23-25] are devoted to various directions of building clusters among the regions of Ukraine. As for tasks at the micro level, many scientific works are focused on the study of financial transactions in banking institutions and trade organizations. The work of foreign authors, M. R. Pinto, P. K. Salume, M. W. Barbosa, P.

R. de Sousa [26], is quite interesting and informative, in which the clustering of retail trade enterprises in relation to the levels of digital maturity according to five dimensions – strategy, market, operations, culture and technology. It is proposed to consider culture as a driver of digital transformation.

The importance of digital education, awareness, and skills for entrepreneurship, as well as the use of data analysis techniques in digital business transformation processes, has been discussed in the works of domestic and foreign scientists [27-31]. However, the question of clustering business structures by the level of digital maturity in order to develop practical recommendations for digital transformation currently requires further study.

3. Methodology for assessing the level of digital maturity of Ukrainian enterprises

Many countries have their own methodologies, frameworks, and tools for measuring digital maturity and digital transformation of business structures. For example, the UK uses diverse tools (Digital Acceleration Index (DAI) (Boston Consulting Group (BCG) and Google), The Digital Scorecard (Lloyds Bank), Digital Maturity Assessment (Department for Digital, Culture, Media & Sport (DCMS)), Digital Capability Assessment Tool (Department for Business, Energy and Industrial Strategy (BEIS)), Digital Business Academy Assessment (Tech Nation, a UK-based network for entrepreneurs)) based on different methodologies to understand the situation of digital business development. Collecting and processing relevant data provides an understanding of the development and implementation of various digital technologies and enables the formation of digital transformation "roadmaps".

The current state of digital technologies in domestic businesses sharply differs from the world. The use of international methodologies to determine the level of digital maturity in business using relevant indicators is not acceptable for domestic realities due to the low overall level of the use of digital technologies in the economic space. The low level of awareness of small and medium-sized enterprises about the opportunities for integrating technologies into their business processes hinders the development of companies and creates difficulties in the entry of domestic businesses into the international arena. Therefore, research on the development of digital transformation indicators for businesses, regular assessments of digital development, and the implementation of regular, systematic statistical observations [11, 12] deserve special attention.

It is necessary to develop our own methodology for determining the digital transformation index of businesses with corresponding indicators that reflect the current state of affairs, provide a deep analysis of the digital maturity indicators of business structures and take into account their dynamics, while remaining flexible to quickly respond to new economic processes and phenomena and ensure further alignment with international methodologies for comparing Ukraine with the most developed countries in the world.

A methodology for determining the Digital Transformation Index “HIT” of domestic SMEs was proposed in [14]. It allows not only to evaluate the level of digital maturity of a business structure but also obtain a vector of digital development strategy. The main indicators of the HIT index are:

- Humans (H): digital literacy (competence) of human capital, which is defined as the ability of an employee to perform complex tasks and requirements that involve both professional and personal digital skills.
- Instruments (I): use of digital tools, which includes components such as social media management, website functioning and search engine optimization, work with specialized business process automation systems, etc.
- Technologies (T): use of digital technologies, that is, the level of enterprise infrastructure provision with necessary equipment (personal computers, laptops, smartphones) and broadband Internet.

The value of the Digital Transformation Index is calculated as a weighted sum of the values of the three corresponding indicators:

$$HIT = \omega_H \cdot \sum ¿_H + \omega_I \cdot \sum ¿_I + \omega_T \cdot \sum ¿_T, ¿¿¿ \quad HIT \in [0;1]; \tag{1}$$

where $\sum ¿_H ¿$ – the aggregated indicator of the digital literacy level of the organization's human capital; $\sum ¿_I ¿$ – the aggregated indicator of the functioning of digital tools integrated into the organization's business processes; $\sum ¿_T ¿$ – the aggregated indicator of the functioning of the organization's digital infrastructure; ω_H, ω_I, ω_T – the respective weight factors of the indicators, where $\omega_H + \omega_I + \omega_T = 1$.

The weight factors were obtained by expert evaluation: $\omega_H = 0.3$, $\omega_I = 0.5$, $\omega_T = 0.2$.

The aggregated indicators $\sum ¿_X ¿$ for each of the indicators H, I, T are calculated using formula:

$$\sum ¿_X = \sum_{i=1}^{m_x} n_i^{(X)} \cdot k_i^{(X)}, ¿ \tag{2}$$

where $\sum ¿_X ¿$ – the aggregated value of indicator X (H, I, or T);

m_X – the number of components of indicator X;

$n_i^{(X)}$ – the functioning level of the i^{th} component of indicator X;

$k_i^{(X)}$ – the weight factor of the i^{th} component of indicator X.

Depending on the obtained value of the HIT index, such gradations for the levels of digital maturity of domestic SMEs were determined: [0; 0.2) is considered very low; [0.2; 0.4) – low; [0.4; 0.6) – medium; [0.6; 0.8) – high; and [0.8; 1] – very high.

4. Dataset description

The dataset represents the results of a survey conducted through Google Forms among Ukrainian entrepreneurs. Thirty four representatives of various small and medium-sized businesses registered in the Ternopil region participated in the survey. Participants were asked to answer 29 questions related to the level of digitization of business activity based on

the components of the HIT index. The set of responses was defined as an experimental dataset.

The answers of N respondents to M questions formed a matrix of dimension $(N \times M)$. It is assumed that each participant $\vec{u}_i$ answered each of the questions q_k. Thus, each surveyed participant is represented in the form of the vector: $\vec{u}_i = \{u_{i1}, u_{i2}, \ldots, u_{ik}, \ldots, u_{\mathfrak{Z}}\}$, where u_{ik} is the answer of the ith participant to the kth question. Each specific vector below in the work is considered a point.

	Questions						
Respondents	q_1	q_2	q_3	...	q_k	...	q_M
$\overrightarrow{u_1} =$	u_{11}	u_{12}	u_{13}	...	u_{1k}	...	u_{1M}
$\overrightarrow{u_2} =$	u_{21}	u_{22}	u_{23}	...	u_{2k}	...	u_{2M}
...	...	...	...	...	...	...	...
$\overrightarrow{u_N} =$	u_{N1}	u_{N2}	u_{N3}	...	u_{Nk}	...	u_{NM}

Figure 1: Matrix of Answers.

Encoding was used to transform categorical data into numeric data (Figure 2).

ID	Name	organization_type	import_export	business_model	computer_equipment	mobile_devices	internet_connection	ict_spec_expernal	ict_spec_internal	web_site	basket_chain	seo_optimization	smm_fb_inst	smm_other	smm_effectiveness	fb_ads	google_ads
1	Respondent1	1	0	0	2	2	0	0	0	0	0	1	0	0	0	0	0
2	Respondent2	1	0	0	2	2	1	1	0	0	0	0	1	0	2	0	0
3	Respondent3	1	0	0	2	2	1	1	2	0	0	0	2	0	2	0	0
4	Respondent4	1	0	0	2	2	2	1	2	1	1	2	2	0	3	3	0
5	Respondent5	0	0	0	0	0	2	1	2	0	0	0	0	0	0	0	0

Figure 2: The table portion of the input dataset with encoded values.

All procedures related to data processing were performed in a specially developed software application using Python. Python libraries used at various stages of the research:

- scikit-learn – for using clustering algorithms and computing quality metrics;
- scipy – for computing distance matrices based on a dataset;
- matplotlib – for visualizing obtained data in the form of graphs;
- pandas – for storing and manipulating a dataset in a special structure, a dataframe.

5. Choice of Clustering Specifications

After obtaining the values of the three components of the HIT index for each SME, the data set consisted of 34 items with 3 numerical attributes. Clustering of preprocessed data using the defined method and distance measure was performed sequentially using the number of clusters from 2 to 8. For each obtained clustering model, quality metrics (Silhouette, Calinski-

Harabasz, and Davies-Bouldin indices) were calculated. Based on visual analysis of the dependencies, the optimal number of clusters was selected. The Figure 3 shows the quality index dependence plots on the number of clusters obtained for agglomerative clustering using cosine distance and Ward linkage.

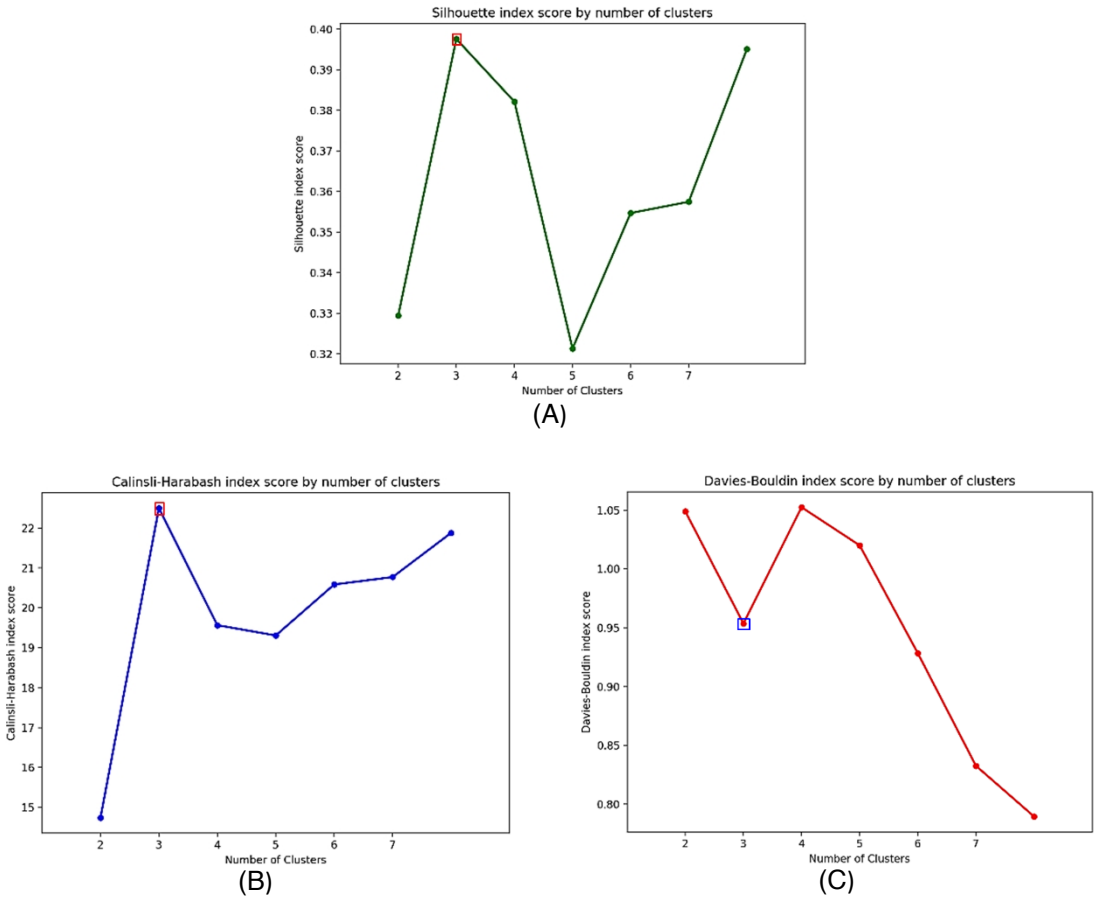

Figure 3: Choosing the optimal number of clusters by: (A) – Silhouette Coefficient, (B) – Calinski-Harabasz Index, (C) – Davies-Bouldin Index.

As it is shown in the Figure 3, local maxima of the Silhouette index and Calinski-Harabasz index are achieved at 3 and 8 clusters. At the same points, local minima are observed for the Davies-Bouldin index. Considering the features of the given problem, the value of 8 clusters seemed too large for the dataset with 34 points, so 3 clusters were chosen.

Since the concept of distance metric is used only for two clustering methods: agglomerative and OPTICS, the selection of criteria set: distance, number of clusters, neighbors was carried out only for them. For each distance metric, the optimal number of clusters was determined. Then, among all the used distance metrics, the one that showed the best results for the current method was selected. The tabular result of such comparison for the agglomerative method is shown in Table 2.

Table 2
An example of choosing the optimal metric and number of clusters

Metric used for intracluster distance	Number of Clusters	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harabasz Index
Euclidean	3	0.34	1	18
Cosine	**3**	**0.65**	**1.4**	**11**
Manhattan	7	0.36	0.9	18
Chebyshev	4	0.36	1	17
Hamming	7	0.13	3	3.5

A similar evaluation was conducted for each used method and distance measure. For each of the methods used, a summary analytical table was compiled with the main characteristics of the formed clusters (Tables 3-7). The figures also show a scatter plot of the dependence of the HIT index on the level of use of digital instruments (on the left) and a bar chart of clusters by HIT index value (on the right). The elements that belong to one cluster are highlighted in the same color.

1. The dataset was divided into 3 clusters using the K-means clustering algorithm. As seen in the scatter plot in the Figure 4, the clusters almost do not intersect with each other and contain sufficiently similar elements inside. Cluster #2 (blue dots) is clearly highlighted and is located at the bottom of the graph in terms of the value of the HIT index to the use of digital tools. Cluster #1 contains most of the points that are located within the intervals of both the HIT index value and the use of digital tools. Cluster #3 is characterized by the highest index values.

Members of Cluster #1 are partially effective in using social networks but do not use their own websites, advertising or analytics tools, while having sufficient technical equipment. The literacy of the human capital is at an elementary level (Table 3).

Cluster #2 shows similar indicators to Cluster #1, except that they do not use social networks or use them inefficiently, and the companies lack sufficient technical equipment. In contrast, Cluster #3 includes respondents who more effectively use the necessary digital tools: websites, social networks, advertising, and have sufficient human capital literacy.

2. Using the agglomerative method, the Euclidean distance measure and Ward linkage allowed for a fairly good result in dividing into 3 clusters (Figure 5). It can be noted that there is a fairly good separation of Cluster #2 (blue dots), which contains respondents with the lowest HIT index values. Additionally, Clusters #1 and #3 are fairly spread out in space, although they do overlap in a few points. Comparison of the main characteristics of the formed clusters is presented in the Table 4.

HIT index value by indicator "I" **HIT index value by participants**

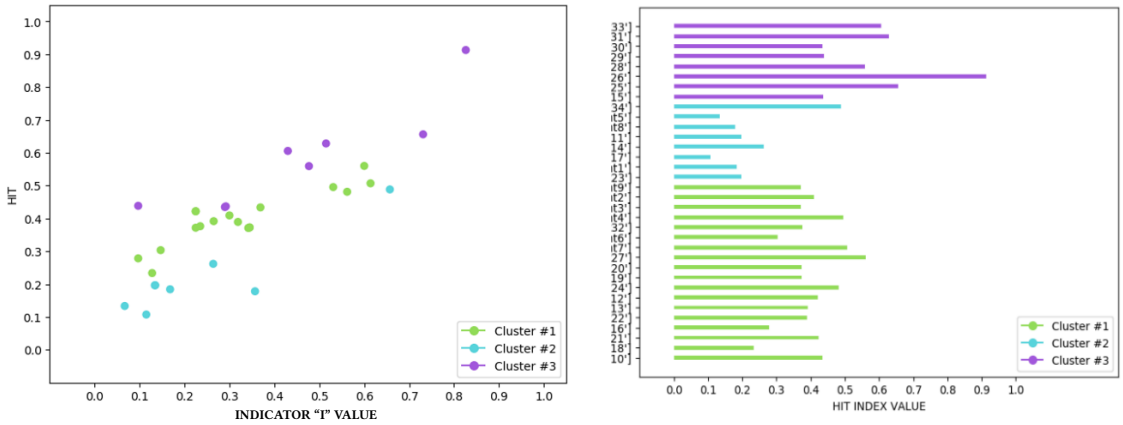

Figure 4: Results of clustering using the K-means method with Euclidean distance.

Table 3
Main characteristics of the clusters formed by the K-means method with Euclidean distance

K-means	Cluster # 1 (18) Ranges of Indicator values: H ϵ [0; 0,364] I ϵ [0,128; 0,614] T ϵ [0,7; 1] **Weighted Sum (HIT) ϵ [0,234; 0,56]**		Cluster # 2 (8) Ranges of Indicator values: H ϵ [0; 0,364] I ϵ [0,067; 0,657] T ϵ [0; 0,5] **Weighted Sum (HIT) ϵ [0,11; 0,488]**		Cluster # 3 (8) Ranges of Indicator values: H ϵ [0,636; 1] I ϵ [0,29; 0,826] T ϵ [0,5; 1] **Weighted Sum (HIT) ϵ [0,44; 0,91]**	
	Status	Percentage of cases	Status	Percentage of cases	Status	Percentage of cases
Website availability, optimization and effectiveness	Not optimized	61.1%	Not optimized	70.0%	Optimized	70.0%
Social media availability and effectiveness	Effectively	50.0%	Not effectively	70.8%	Effectively	70.0%
Use of online advertising and analytics	Not used	74.1%	Not used	91.6%	Used	58.3%
Use of specialized management systems	Not used	80.2%	Not used	73.2%	Not used	71.4%
Use of specialized technical systems	Not used	96.4%	Not used	79.2%	Not used	87.5%
Level of technical support	Satisfactory	98.1%	Not satisfactory	62.5%	Satisfactory	83.3%
Level of Digital Literacy	Basic	50.0%	Basic	62.5%	Intermediate or above intermediate	87.5%
Communication channels	With the use of ICT	74.7%	With the use of ICT	83.3%	With the use of ICT	75.0%

Silhouette Coefficient	**0.411**
Calinski-Harabasz Index	**24.105**
Davies-Bouldin Index	**0.889**

Cluster #1 members, who belong to the area with the highest indicator values, effectively use the website and social media, and also have a level of digital literacy that is at or above the average for most respondents. In contrast, Cluster #2 is characterized by ineffective use of digital tools for most members, as well as low digital literacy and unsatisfactory technical equipment for more than half of the surveyed. Cluster #3 has a certain intensity of social media use, but low indicators in other areas, such as elementary level of digital literacy among employees.

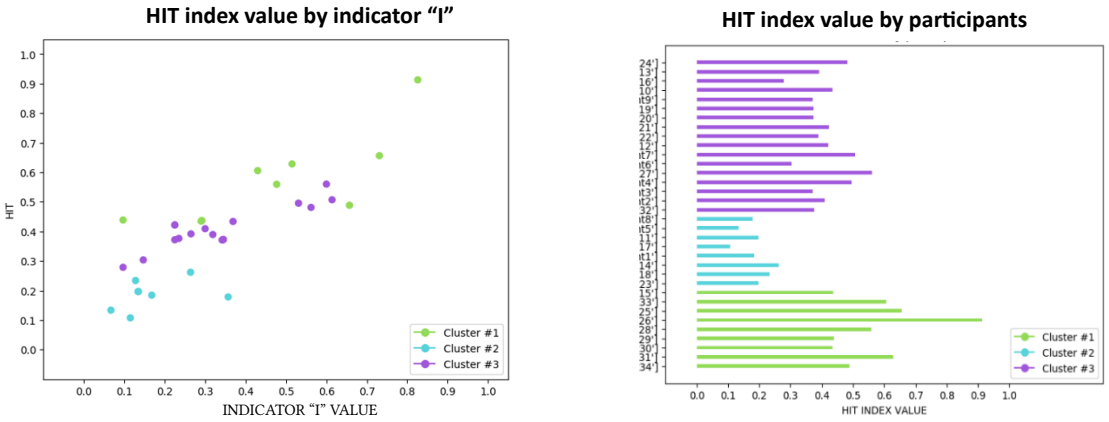

Figure 5: Results of clustering using the Agglomerative method with Ward linkage.

Table 4
Main characteristics of the clusters formed by the Agglomerative method with Ward linkage

	Cluster # 1 (9)		Cluster # 2 (8)		Cluster # 3 (17)	
Agglomerative clustering	Ranges of Indicator values: H ϵ [0,2; 1] I ϵ [0,097; 0,826] T ϵ [0,25; 1] **Weighted Sum (HIT) ϵ [0,43; 0,91]**		Ranges of Indicator values: H ϵ [0; 0,364] I ϵ [0,067; 0,357] T ϵ [0; 0,7] **Weighted Sum (HIT) ϵ [0,11; 0,26]**		Ranges of Indicator values: H ϵ [0; 0,364] I ϵ [0,097; 0,614] T ϵ [0,75; 1] **Weighted Sum (HIT) ϵ [0,28; 0,56]**	
	Status	Percentage of cases	Status r	Percentage of cases	Status	Percentage of cases
Website availability, optimization and effectiveness	Optimized	68.9%	Not optimized	80.0%	Not optimized	60.0%
Social media availability and effectiveness	Effectively	70.3%	Not effectively	75.0%	Effectively	51.0%
Use of online advertising and	Not used	55.6%	Not used	100.0%	Not used	72.5%

analytics						
Use of specialized management systems	Not used	65.1%	Not used	82.1%	Not used	79.8%
Use of specialized technical systems	Not used	88.9%	Not used	79.2%	Not used	98.0%
Level of technical support	Satisfactory	77.8%	Not satisfactory	58.3%	Satisfactory	100.0%
Level of Digital Literacy	Intermediate or above intermediate	83.3%	Basic	75.0%	Basic	70.6%
Communication channels	With the use of ICT	77.8%	With the use of ICT	75.0%	With the use of ICT	76.5%
Silhouette Coefficient			**0.398**			
Calinski-Harabasz Index			**22.497**			
Davies-Bouldin Index			**0.954**			

3. Using OPTICS with Chebyshev distance metric and a minimum of 7 points for cluster formation. Despite obtaining an optimal value for quality metrics, the clustering itself was not successful from a practical standpoint. As can be seen in the visualization in the Figure 6, the clusters contain almost the same number of members. Additionally, the clusters were distributed as internal and external, making it impossible to establish fundamental differences between them, as seen in the analytical Table 5.

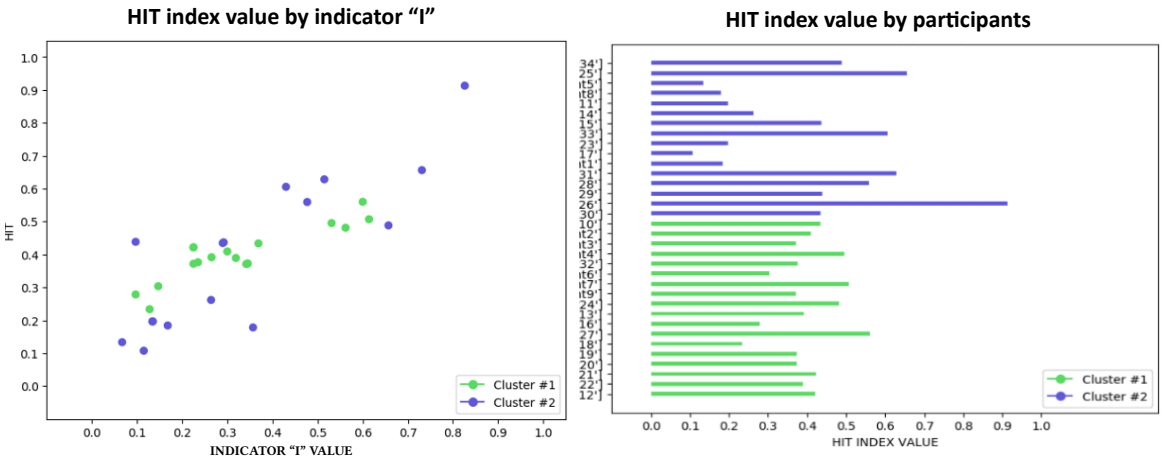

Figure 6: Results of clustering using the OPTICS method with Chebyshev distance and 7 neighbors.

The reason for this result is that OPTICS belongs to density-based algorithms, and the basic data set does not contain dense areas. Therefore, the internal cluster (green) turned out

to be an artificial area with dense values, while the external one was marked as outliers, meaning values that do not carry any value.

4. The Affinity Propagation method doesn't depend on the number of clusters and distance measures, so its results represent the inherent data structure without any user influence. As seen in the Figure 7 and Table 6, the data was divided into 6 clusters. Some of the clusters (such as #1, #5 and #6) are quite distinct from the others. At the same time, clusters #2, #3 and #4 overlap somewhat with other clusters. The distribution of respondents based on the value of the HIT index clearly highlights the cluster leader (#5), as well as the clusters with the lowest values (#2 and #4). Clusters #1, #3 and #6 consist of respondents with average and above-average values of the index.

Clusters #1, #3 and #5 are quite similar to each other, as can be seen from the table. However, it is interesting that about 2/3 of the participants in cluster #1 are successfully using the website and social media, although they rate the level of human capital literacy as elementary.

Table 5

Main characteristics of the clusters formed by the OPTICS method with Chebyshev distance and 7 neighbors

OPTICS	Cluster # 1 (18) Ranges of Indicator values: H ϵ [0; 0,364] I ϵ [0,128; 0,614] T ϵ [0,7; 1] **Weighted Sum (HIT) ϵ [0,23; 0,56]**		Cluster # 2 (16) Ranges of Indicator values: H ϵ [0; 1] I ϵ [0,067; 0,826] T ϵ [0; 1] **Weighted Sum (HIT) ϵ [0,13; 0,91]**	
	Status	Percentage of cases	Status	Percentage of cases
Website availability, optimization and effectiveness	Not optimized	61.1%	Not optimized	51.3%
Social media availability and effectiveness	Effectively	50.0%	Effectively	50.0%
Use of online advertising and analytics	Not used	74.1%	Not used	75.0%
Use of specialized management systems	Not used	80.2%	Not used	72.3%
Use of specialized technical systems	Not used	96.3%	Not used	85.4%
Level of technical support	Satisfactory	100.0%	Satisfactory	60.4%
Level of Digital Literacy	Basic	69.4%	Intermediate or above intermediate	56.3%

Communication channels	With the use of ICT	74.1%	With the use of ICT	79.2%
Silhouette Coefficient		**0.327**		
Calinski-Harabasz Index		**13.554**		
Davies-Bouldin Index		**1.408**		

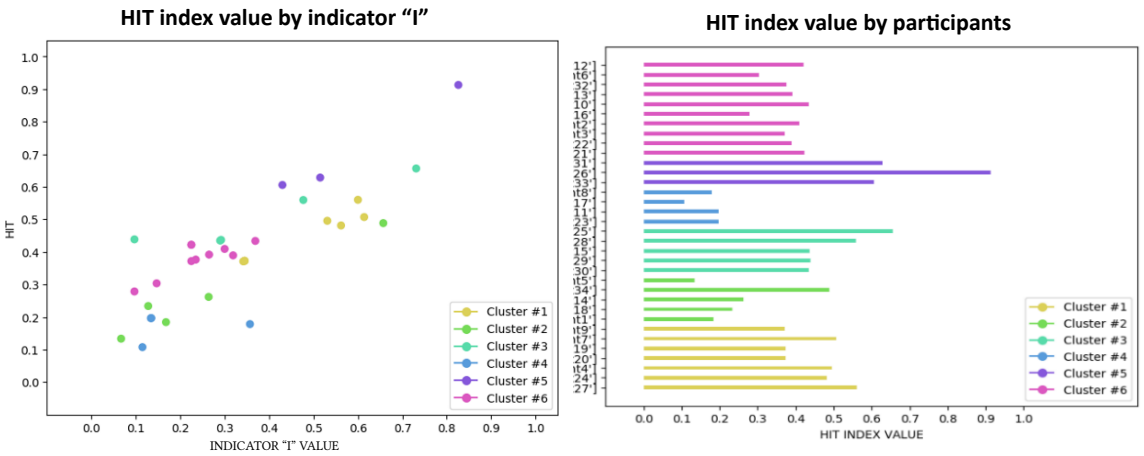

Figure 7: Results of clustering using the Affinity Propagation method

In contrast, cluster #4 has a high value of digital literacy, but only slightly more than half of the participants are successfully using digital technologies (given the size of the cluster, this may be within the margin of error). Cluster #5 is the smallest, but consists of respondents with the highest level of digital tool usage and transformation index value. Clusters #2 and #4 are characterized by inefficient use of digital resources. The difference between them lies in the value of the digital literacy indicator. Cluster #6 is also interesting, as it showed the effectiveness of social media use at low levels of other indicators.

5. The Gaussian Mixture Expectation-Maximization soft clustering algorithm divided the dataset into 3 clusters; visualization is shown in the Figure 8. Cluster #2 (blue dots) is dense, with its HIT index values falling in the interval with the mean values, indicating the use of digital tools. Slightly higher values can be observed in cluster #3, which is also well grouped.

In contrast, the largest cluster #1 is very dispersed and contains points with both the lowest and highest values of HIT index components. The points in this cluster, shown in green, are located around the perimeter of the scatter plot. Such dividing is likely due to the initial dataset being far from a normal distribution.

In Cluster #1, half of the respondents do not use digital tools, although almost 70% of those surveyed claim to have an average or high level of digital literacy. In Cluster #2, the majority do not use modern capabilities, despite that all respondents have a basic level of technical means.

Table 6

Main characteristics of the clusters formed by the Affinity Propagation method

Affinity Propagation	Cluster # 1 (7) Ranges of Indicator values: H ϵ [0; 0,2] I ϵ [0,34; 0,61] T ϵ {1} **Weighted Sum (HIT) ϵ [0,37; 0,56]**		Cluster # 2 (5) Ranges of Indicator values: H ϵ [0; 0,2] I ϵ [0,067; 0,657] T ϵ [0,5; 0,7] **Weighted Sum (HIT) ϵ [0,13; 0,488]**		Cluster # 3 (5) Ranges of Indicator values: H ϵ [0,636; 0,8] I ϵ [0,097; 0,73] T ϵ [0,25; 0,75] **Weighted Sum (HIT) ϵ [0,43; 0,66]**	
	Status	Percentage of cases	Status	Percentage of cases	Status	Percentage of cases
Website availability, optimization and effectiveness	Optimized	60.0%	Not optimized	64.0%	Optimized	52.0%
Social media availability and effectiveness	Effectively	66.7%	Not effectively	66.6%	Effectively	53.3%
Use of online advertising and analytics	Not used	52.4%	Not used	86.7%	Not used	60.0%
Use of specialized management systems	Not used	81.6%	Not used	71.4%	Not used	74.2%
Use of specialized technical systems	Not used	100.0%	Not used	66.6%	Not used	93.3%
Level of technical support	Satisfactory	100.0%	Satisfactory	53.3%	Satisfactory	86.7%
Level of Digital Literacy	Basic	85.7%	Basic	70.0%	Intermediate or above intermediate	80.0%
Communication channels	With the use of ICT	85.7%	With the use of ICT	66.6%	With the use of ICT	73.3%

	Cluster # 4 (4) Ranges of Indicator values: H ϵ [0; 0,36] I ϵ [0,12; 0,36] T ϵ [0; 0,25] **Weighted Sum (HIT) ϵ [0,11; 0,20]**		Cluster # 5 (3) Ranges of Indicator values: H ϵ [0,636; 1] I ϵ [0,43; 0,83] T ϵ [0,9; 1] **Weighted Sum (HIT) ϵ [0,61; 0,91]**		Cluster # 6 (10) Ranges of Indicator values: H ϵ [0,1; 0,36] I ϵ [0,097; 0,369] T ϵ [0,75; 1] **Weighted Sum (HIT) ϵ [0,28; 0,43]**	
	Status	Percentage of cases	Status r	Percentage of cases	Status	Percentage of cases
Website availability, optimization and effectiveness	Not optimized	85.0%	Optimized	93.3%	Not optimized	72.0%
Social media availability and effectiveness	Not effectively	75.0%	Effectively	100.0%	Effectively	60.0%
Use of online advertising and	Not used	100.0%	Not used	55.5%	Not used	86.7%

analytics						
Use of specialized management systems	Not used	78.6%	Not used	66.6%	Not used	78.6%
Use of specialized technical systems	Not used	91.6%	Not used	88.9%	Not used	96.7%
Level of technical support	Not satisfactory	75.0%	Satisfactory	100.0%	Satisfactory	100.0%
Level of Digital Literacy	Basic	75.0%	Intermediate or above intermediate	100.0%	Basic	60.0%
Communication channels	With the use of ICT	91.7%	With the use of ICT	77.8%	With the use of ICT	70.0%
Silhouette Coefficient			**0.351**			
Calinski-Harabasz Index			**21.607**			
Davies-Bouldin Index			**0.931**			

The Cluster #3 shows moderate success in using simple tools, such as a website and social networks, provided that 80% of respondents consider the digital competencies of their employees to be basic. Another observation is that half of the respondents use, for example, analytics and half do not, making it impossible to identify precise distinguishing features between the clusters.

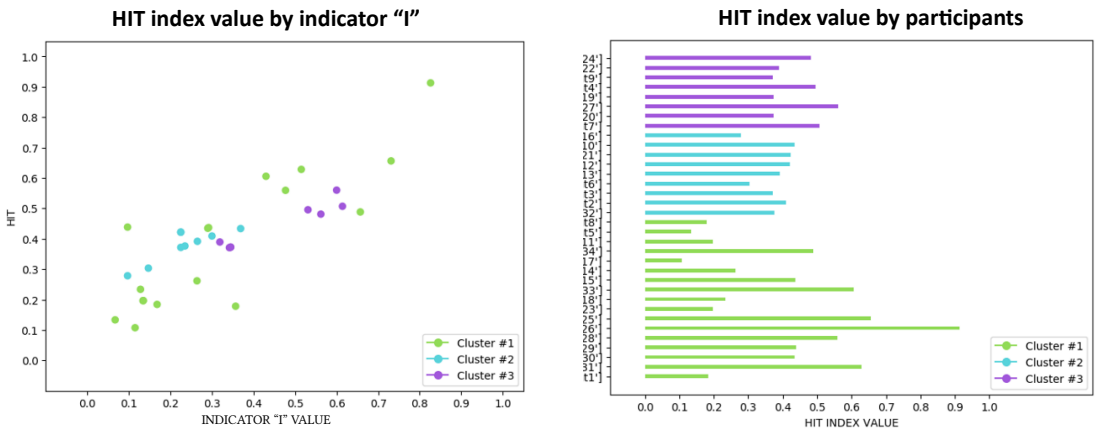

Figure 8: Results of clustering using the Gaussian Mixture (EM-method)

Analytical data with the main characteristics of the formed clusters are presented in the Table 7.

Table 7

Main characteristics of the formed clusters by the Gaussian Mixture (EM-method)

Gaussian Mixture (EM)	Cluster # 1 (17) Ranges of Indicator values: H ϵ [0; 1] I ϵ [0,067; 0,826] T ϵ [0; 1] **Weighted Sum (HIT) ϵ [0,13; 0,91]**		Cluster # 2 (9) Ranges of Indicator values: H ϵ [0,1; 0,364] I ϵ [0,097; 0,369] T ϵ [0,75; 1] **Weighted Sum (HIT) ϵ [0,28; 0,43]**		Cluster # 3 (8) Ranges of Indicator values: H ϵ [0; 0,2] I ϵ [0,319; 0,614] T ϵ {1} **Weighted Sum (HIT) ϵ [0,37; 0,56]**	
	Status	Percentage of cases	Status	Percentage of cases	Status	Percentage of cases
Website availability, optimization and effectiveness	Not optimized	61.1%	Not optimized	70.0%	Optimized	70.0%
Social media availability and effectiveness	Effectively	50.0%	Not effectively	70.8%	Effectively	70.0%
Use of online advertising and analytics	Not used	74.1%	Not used	91.6%	Not used	58.3%
Use of specialized management systems	Not used	80.2%	Not used	73.2%	Not used	71.4%
Use of specialized technical systems	Not used	96.4%	Not used	79.2%	Not used	87.5%
Level of technical support	Satisfactory	98.1%	Satisfactory	62.5%	Satisfactory	83.3%
Level of Digital Literacy	Intermediate or above intermediate	50.0%	Basic	62.5%	Basic	87.5%
Communication channels	With the use of ICT	74.7%	With the use of ICT	83.3%	With the use of ICT	75.0%
Silhouette Coefficient			**0.192**			
Calinski-Harabasz Index			**8.578**			
Davies-Bouldin Index			**1.352**			

It is worth noting that the level of digital literacy of employees has a significant impact on the overall state of digitalization of the enterprise. If the level of digital literacy of employees is defined as elementary, then such an enterprise lacks websites, social networks and other used tools. As the digital literacy of employees increases, the percentage of use of tools and technologies increases, so investing in people is seen as an important contribution to the success of digitalization. It is interesting that the level of technical equipment does not have a significant impact on the overall digital level of the enterprises.

6. Conclusions

The paper presents 5 data clustering models for understanding the current state of digitalization of business processes among small and medium-sized enterprises in the Ternopil region of Ukraine. The Digital Transformation Index "HIT" was used for numerical measurement of the current level of digital maturity of domestic enterprises. Clustering of enterprises was based on numerical values of three indicators – components of the Digital Transformation Index. A special software application was developed in Python programming language for solving the task. Various approaches to clustering model construction were implemented using built-in methods of the scikit-learn library for Data Mining problems. Four hard clustering methods (K-Means, Affinity Propagation, Hierarchical clustering, OPTICS) and one soft clustering method using the EM algorithm (Gaussian Mixture) were used. The Silhouette Index was used as the main quality metric. From the perspective of similarity between elements within groups and differences between different clusters, the best results on the dataset were demonstrated by Affinity Propagation, Ward's hierarchical clustering with 3 clusters, and K-Means with a division into 3 clusters. Analysis of the constructed models showed that high values of quality metrics do not always indicate an optimal and effective division into groups that can be successfully interpreted. New valuable ideas were obtained regarding the importance of individual components of the Digital Transformation Index. Common features of the obtained groups of enterprises, their strengths and weaknesses in the use of digital tools and digital literacy of human capital were identified. In the future, stable formed clusters can be used for classifying new surveyed enterprises and identifying significant attributes with the greatest impact on the value of digital maturity of the subject or for developing a methodology for providing recommendations to improve the level of digital maturity of the enterprise.

7. References

[1] OECD Policy Responses on the Impacts of the War in Ukraine “Digitalisation for recovery in Ukraine“, 2022. URL: https://www.oecd.org/ukraine-hub/policy-responses/digitalisation-for-recovery-in-ukraine-c5477864/.

[2] J. Cenamor, V. Parida and J. Wincent, How entrepreneurial SMEs compete through digital platforms: The roles of digital platform capability, network capability and ambidexterity, Journal of Business Research, Julay, vol. 100, (2019), pp. 196–216.

[3] N. Ivanchenko, Zh. Kudryts'ka. and K. Rekachyns'ka, Business models in the conditions of digital transformations, Vcheni zapysky TNU imeni V. I. Vernads'koho, Seriia: Ekonomika i upravlinnia, vol. 3, no. 31 (2020), pp. 185–190.

[4] N. M. Kraus, O. P. Holoborod'ko, and K. M. Kraus, Digital economy: trends and perspectives of the abangard change of development, Efektyvna ekonomika, vol. 1, 2018.

[5] A. Annarelli et al. Literature review on digitalization capabilities: Co-citation analysis of antecedents, conceptualization and consequences Technol. Forecast. Soc. Change, 2021.

[6] J. Mero et al. An effectual approach to executing dynamic capabilities under unexpected uncertainty Ind. Market. Manag, 2022.

[7] Digital Maturity Models: A Systematic Literature Review May 2021. doi:10.1007/978-3-030-69380-0_5 URL:

https://www.researchgate.net/publication/351975241_Digital_Maturity_Models_A_Systematic_Literature_Review.
[8] O. Pischulina, Digital economy: trends, risks and social determinants: report, Tsentr Razumkova, 2020, 271 p. URL: https://razumkov.org.ua/uploads/article/2020_digitalization.pdf.
[9] H. Zhekalo, Digital economy of Ukraine: problems and prospects of development, Naukovyj visnyk Uzhhorods'koho natsional'noho universytetu, Seriia: Mizhnarodni ekonomichni vidnosyny ta svitove hospodarstvo, vol. 26, no. 1, (2019): 56–60.
[10] H. Karcheva, D. Ohorodnia, and V. Open'ko, Digital economy and its impact on the development of national and international economy, Finansovyj prostir, vol. 3, no. 1, (2017): 13–21.
[11] I. Strutynska, L. Dmytrotsa, H. Kozbur, O. Hlado, P. Dudkin and O. Dudkina, Development of Digital Platform to Identify and Monitor the Digital Business Transformation Index, in: Proceedings of the 15th International Conference on Computer Sciences and Information Technologies (CSIT), Zbarazh, Ukraine, September 23, 2020, pp. 171-175, doi: 10.1109/CSIT49958.2020.9322016.
[12] I. Strutynska, L. Dmytrotsa, H. Kozbur, L. Melnyk, O. Hlado. Developing Practical Recommendations for Increasing the Level of Digital Business Transformation Index, in: Proceedings of the 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer, volume II: Workshops of ICTERI, Part III: 8th International Workshop Information Technology in Economic Research (ITER 2020), Kharkiv, Ukraine, October 06-10, 2020, pp. 351-362. URL: https://ceur-ws.org/Vol-2732/20200351.pdf.
[13] I. Strutynska, L. Dmytrotsa, H. Kozbur, L. Melnyk, System-Integrated Methodological Approach Development to Calculating the Digital Transformation Index of Businesses, in: Proceedings of the 16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer, volume I: Main Conference (ICTERI 2020), Kharkiv, Ukraine, October 06-10, 2020, pp. 373-379. URL: http://ceur-ws.org/Vol-2740/20200373.pdf.
[14] I. Strutynska, L. Dmytrotsa, H. Kozbur, L. Melnyk, The Digital Business Transformation Index Determining and Monitoring: Development of a National Online Platform, in: Proceedings of the 1st International Workshop on Information Technologies: Theoretical and Applied Problems, ITTAP 2021, Ternopil, Ukraine, 2021, pp. 327-334.
[15] H. Cuesta, S. Kumar, Practical Data Analysis. Birmingham, Packt Publishing Ltd, 2016.
[16] Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services. Indianapolis, John Wiley & Sons, Inc, 2015.
[17] C. Iyigun, M. Türkeş, I. Batmaz, C. Yozgatligil, V. Purutçuoğlu, E. Kartal, M. Öztürk, Clustering current climate regions of Turkey by using a multivariate statistical method. Theoretical and Applied Climatology, 114 (2013): 95-106.
[18] K. Sablyn, E. Kahan, E. Chernova, Clustering of coal mining regions of Russia: investment and innovation activity. Journal of New Economy, 21 (1) (2020): 89-106.
[19] K. Gorbatiuk, O. Mantalyuk, O. Proskurovych, O.Valkov, Application of Fuzzy Clustering to Shaping Regional Development Strategies in Ukraine, Proceedings of the 6th International Conference on Strategies, Models and Technologies of Economic Systems Management (SMTESM 2019), 2019, pp. 271-276.
[20] T. Paianok, Y. Vazhaliuk, Cluster analysis of labor potential of Ukraine. Economy and State, 12 (2019): 109-114.
[21] S. Behun, Application of cluster analysis to study the demographic situation in the region. Economic Journal of Lesya Ukrainka East European National University, 2 (2016): 122-128.

[22] S. Synytsia, O. Vakun, Clustering of regions by level of economic potential. Economy and society Mukachevo State University, 12 (2017): 776-784.
[23] L. Zomchak, Y. Dobrotii, Clustering of regions of Ukraine by competitiveness. Proceedings of the International scientific-practical conference Administrative-territorial vs economic spatial borders of regions, KNEU, 2020, pp. 328-332.
[24] V. Aulin, O. Lyashuk, O. Pavlenko, D. Velykodnyi, A. Hrynkiv, S. Lysenko, et al., "Realization of the Logistic Approach in the International Cargo Delivery System", COMMUNICATIONS, vol. 21, no. 2, pp. 3-12, 2019.
[25] Petraška, A.; Čižiuniene, K.; Jarašuniene, A.; Maruschak, P.; Prentkovskis, O. Algorithm for the assessment of heavyweight and oversize cargo transportation routes. J. Bus. Econ. Manag. 2017, 18, 1098–1114
[26] The path to digital maturity: A cluster analysis of the retail industry in an emerging economy Marcelo Rezende Pinto, Paula Karina Salume, Marcelo Werneck Barbosa, Paulo Renato de Sousa https://doi.org/10.1016/j.techsoc.2022.102191.
[27] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Clustering algorithms and validity measures, in: Proceedings of the Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM, Fairfax, VA, USA SourceIEEE Xplore, July 18, 2001. doi:10.1109/SSDM.2001.938534.
[28] Clustering, 2022. URL: https://scikit-learn.org/stable/modules/clustering.
[29] E. Zuccarelli, Performance Metrics in Machine Learning — Part 3: Clustering, 2021. URL: https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6.
[30] N. Bolshakova, F. Azuaje, Cluster validation techniques for genome expression data, volume 83 of Signal Processing, Issue 4, April 2003, pp. 825-833. doi: 10.1016/S0165-1684(02)00475-9.
[31] Yavorskyi, A.V.; Karpash, M.O.; Zhovtulia, L.Y.; Poberezhny, L.Y.; Maruschak, P.O. Safe operation of engineering structures in the oil and gas industry. J. Nat. Gas Sci. Eng. 2017, 46, 289–295.