International Journal on Emerging Technologies **10**(2a): 225-231(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Cluster Analysis of the Russian Regions by the Human Capital Development and Digital Resources Factors

***Gabdullin Nail Maratovich*[1]*, Kamaev Bulat Nailevich*[2] *and Kirshin Igor Alexandrovich*[3]**
[1]Associate Professor, Head University/Institute of Management, Economics and Finance,
Kazan Federal University, Russia.
[2]Design engineer, KFU/REC for the study of the problems of development of market relations in
the context of globalization of the world economy (main employee), Kazan Federal University, Russia.
[3]Professor, KFU/Higher School of Business, Kazan Federal University, Russia.

(Corresponding author: Gabdullin Nail Maratovich)
(Received 05 May 2019, Revised 16 July 2019 Accepted 25 July 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The article performs the analysis of the structure of multidimensional data on the factor of development of human capital and digital resources of Russia. The objects under observation were the subjects of the Russian Federation, the factors of each object under observation were the average annual indicators of the corresponded factors of the monitoring of the information society development in the Russian Federation for the period from 2010 to 2017. The article substantiates an increasing role of digital resources in improving the quality of human capital of clusters of the subjects of the Russian Federation which differ in the level of development of digital infrastructure, social conditions and the peculiarities of rendering electronic services. The theses of this study can serve as methodological instructions for the formation of regional cluster networks aimed at improving the competitiveness of the subjects of the Russian Federation that form clusters. To determine the optimal number of clusters, several models were applied with the subsequent selection of the best number by means of the BIC (Bayesian Information Criterion) assessment. The structuring of data through cluster analysis was performed using the EM method (Expectation Maximization) and the Ward Hierarchical Clustering method. These methods were implemented in the R studio integrated development environment. As a result of applying the methods of clusterization, clusters of the regions of the RF that differ in the level of development of resource base, processes and results of using digital technologies for the development of constituent entities of the Russian Federation have been singled out. A stable cluster of the subjects of the RF differing from the majority of the subjects of the RF by special characteristics of human capital development and electronic technologies has been identified.

Keywords: cluster analysis, digital economics, E-education, E-Health, E-culture.

I. INTRODUCTION

In modern hyper-competitive global economy the regions are as much innovative and competitive as they effectively capitalize human capital and apply digital technologies. The subjects of the Russian Federation are experiencing a shortage not only in STEM-workers (Science, Technology, Engineering and Mathematics) but also in specialists who master the skills to implement innovations that create competitive advantages.

Ensuring sustainable growth of regional economies requires balancing the requirements of employees skills at different levels of the hierarchy. The role of human capital of managers of organizations that the owners tend to attract as strategic business partners is becoming stronger. This is a complex task, and to accomplish it, the heads of the HR department of firms use digital technologies that accelerate the accumulation of human capital through the digitalization of jobs and, above all, the working conditions of talented employees. This specifies the growing importance of digital self-sufficiency as a modern HR management, as well as working and potential employees. The more valuable is today the ability to manage the activities using online services that provide a higher level of job satisfaction and productivity. Human capital development digitization transforms human resource management via supporting electronic technologies. They comprehend:

– E-education,
– E-Health,
– E-culture.

This paper is concerned with a consistent clustering of the subjects of the Russian Federation on the factors of human capital development and the use of electronic technologies in the subjects of the Russian Federation. The subjects of the RF were chosen as the objects of observation, and the average annual values of the relevant factors for the period considered in monitoring were chosen as the factors of each object of observation [1].

II. METHODS

The following analytical procedures were used within the framework of the cluster analysis of multidimensional data:

– Prestarting procedures for cluster analysis:

Step 1. Standardization procedure. Standardization of the initial matrix of factors.

Step 2. Carrying out the procedure for reducing the dimension of the standardized matrix of factors by the PCA method.

Step 3. Determination of the optimal set of clusters in a group of 30 indices.

Cluster analysis of multidimensional data (research results to identify homogeneous clusters that integrate the characteristics of groups of initial factors):

- the EM method -Expectation Maximization,
- the Hierarchical Clustering method (Ward.D2) comparative analysis of multidimensional data in the identified clusters (data analysis of individual clusters).

To determine the optimal number of clusters, several models were applied with the subsequent selection of the best number by means of the BIC (Bayesian Information Criterion) assessment. The structuring of data through cluster analysis was performed using the EM method (Expectation Maximization) and the Ward Hierarchical Clustering method.

These methods were implemented in the R studio integrated development environment, using the library m-clust - for statistical data processing and working with graphics for clustering, classification and density estimation on the ground of the models based on the final modeling of Gaussian mixture of distributions (GMM - Gaussian mixture models).It provides a set of tools and functions for estimating parameters using *the* EM algorithm for ordinary mixture models with different covariance structures and functions for simulation modeling of these models. The EM algorithm is used to find the estimates of maximum likelihood for the parameters of probabilistic models, in the case when the model depends on some hidden variables. Each iteration of the algorithm involves two steps. At the E-step (expectation), the expected value of the likelihood function is calculated, at the same time the hidden variables are treated as observables. At the M-step (maximization), the maximum likelihood estimate is calculated, thus increasing the expected likelihood being calculated at the E-step. This value is then used for the E-step at the next iteration. The algorithm runs until convergence [2]. When using the Ward method inside clusters, the minimum dispersion is optimized, as a result, the clusters of approximately equal sizes are created [3].

To carry out a complex procedure of cluster analysis, the problem of optimization in order to reduce their dimension by the method of Principal Components (Principal Component Analysis, PCA)was preliminarily solved for each multidimensional matrix of factors [4]. The data transformation matrix to the main components consists of the vectors of the main components that are arranged in decreasing order of eigen values. Most of the data variation will be concentrated in the first coordinates, which makes it possible to move to a space of a smaller dimension. The resulting first two main components are used to create a dot data chart, where Dim1 is the first main component, Dim-2 is the second. The percentages in the dot chart indicate the percentage of information disclosure for the entire sample.

III. RESULTS AND DISCUSSION

First, we consider the results of the cluster analysis of the factors of dataset" Human capital". Multiple factors that characterize human capital include: the rate of employment population aged 25-64 years old, having a higher educational attainment to the ratio of population of relevant age group; student population enrolled in educational programs of higher education –Bachelor's programs, Specialist's programs, Master's programs, per 10,000 population; adult literacy rate; the proportion of students of educational institutions in the total population; the proportion of students enrolled in the training program for skilled workers and employees to general population; the proportion of students enrolled in the training programs for mid-level specialists to general population; the proportion of organizations that conducted remedial ICT education and training of the staff to the overall number of organizations surveyed. Following the results of the evaluation of the BIC criterion, we will model a scattering diagram with 4 clusters localized in it.

The scattering diagram with the localization of three clusters is shown in Fig. 1.

The set of clusters of the subjects of the RF is given numerically in Table 1.

Then, using the Ward Hierarchical Clustering method, a dendrogram was constructed and 4 clusters were selected (Fig. 2).

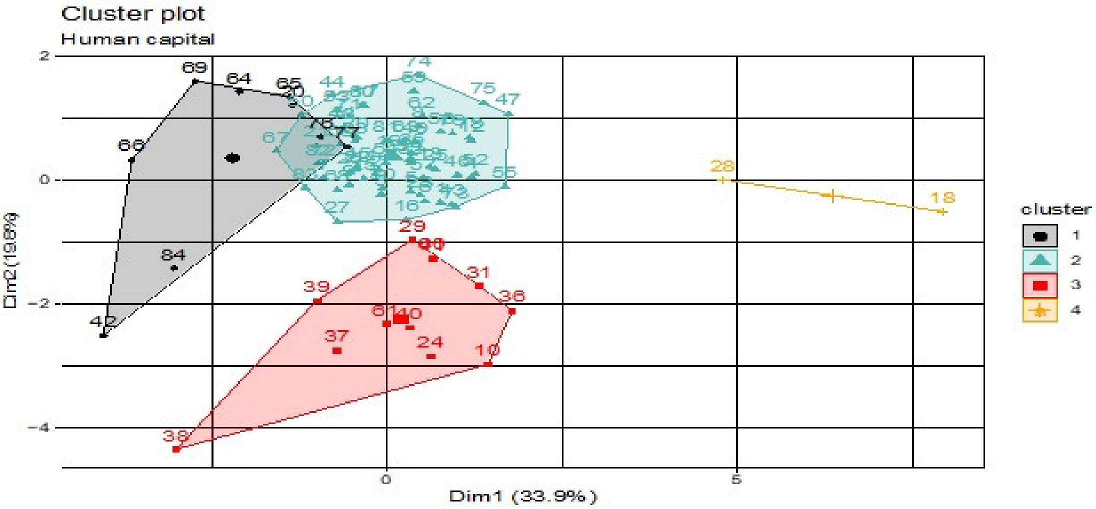

Fig. 1. Scattering Diagram (Visualization of the Results of Cluster Analysis using the EM method).

Maratovich et al., *International Journal on Emerging Technogies* **10(2a): 225-231(2019) 226**

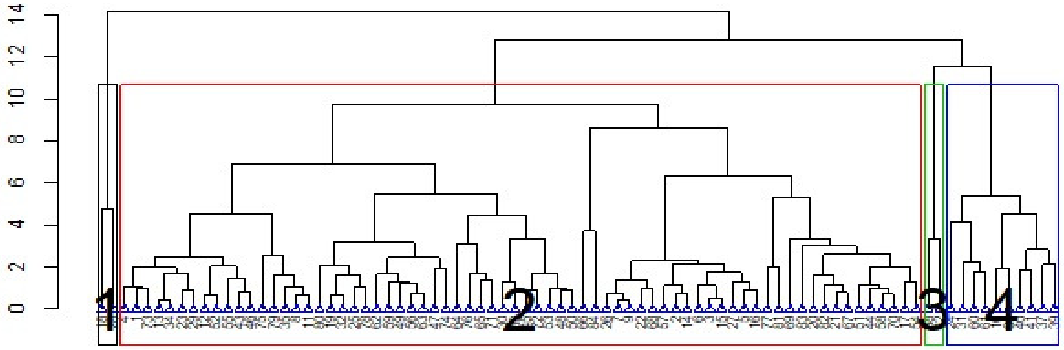

Fig. 2. Dendrogram of Hierarchical Cluster Analysis (Visualization of Results by the Ward method).

Table 1:

Cluster number	Cluster's Composition											
1	18	28										
2	1	2	3	4	5	6	7	8	9	11	12	13
	14	15	16	17	19	20	21	22	23	25	26	27
	29	30	32	33	34	35	43	44	45	46	47	48
	49	50	51	52	53	54	55	56	57	58	59	62
	63	64	65	66	67	68	69	70	71	72	73	74
	75	76	77	78	79	80	81	82	83	84		
3	38	42										
4	24	31	36	37	39	40	41	60	61			

Let us pass on to the results of clustering the subjects of the Russian Federation based on the factors of E-education. These factors include: the number of personal computers used for educational purposes per 100 students of state and municipal educational institutions; the number of personal computers used for educational purposes that are comprehended in Local Area Networks (LAN) per 100 students at educational institutions of Secondary Professional Education (SPE); the number of personal computers used for educational purposes that are comprehended in local area networks (LAN) per 100 students at educational institutions of Higher Professional Education (HPE); the share of educational institutions of higher education that are connected to the Internet in the total institutions of higher education surveyed at a speed of 256 Kbps and above; the number of personal computers used for educational purposes, with access to the Internet, per 100 students (students) at educational institutions of secondary vocational education; the number of personal computers used for educational purposes with access to the Internet per 100 students at educational institutions of higher education; the share of educational institutions that have a website on the Internet in the total independent educational institutions of secondary professional education [5, 6].

Following the results of the evaluation of the BIC criterion, we will model a scattering diagram with localization of five clusters (Fig. 3).

Hierarchical clustering by Ward's Hierarchical Clustering method also generates five clusters (Fig. 4).

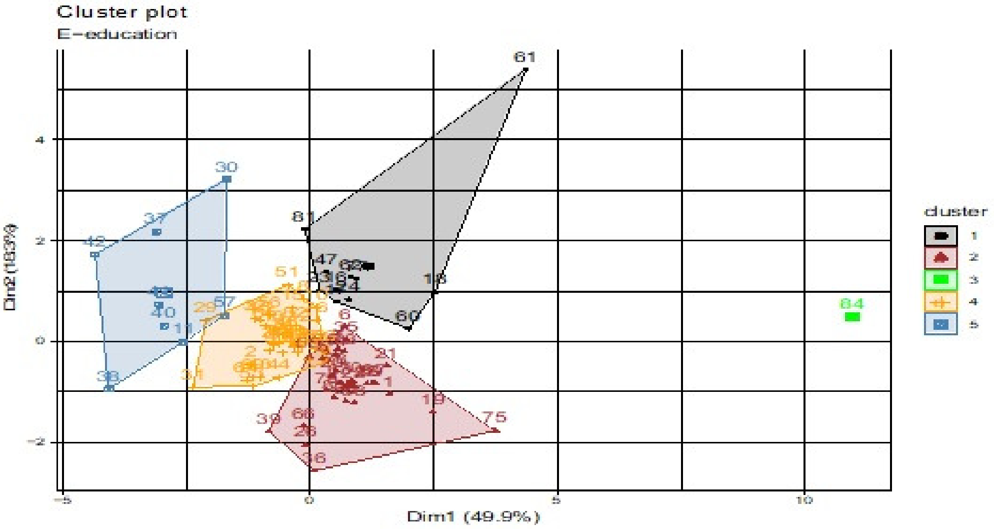

Fig. 3. Scattering Diagram (Visualization of the Results of Cluster Analysis by the EM Method)

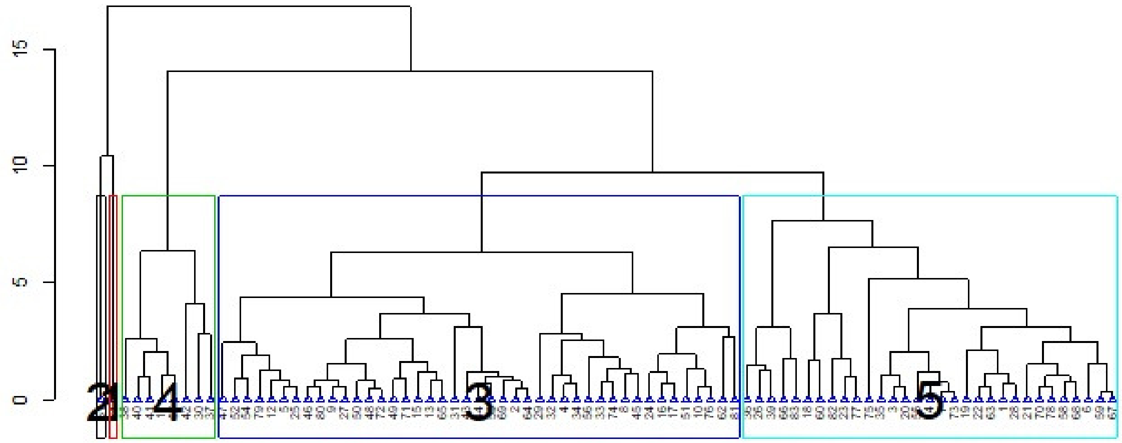

Fig. 4. Dendrogram of Hierarchical Cluster Analysis (Visualization of the Results by the Ward Method).

Table 2: The Results of Hierarchical Cluster Analysis.

Cluster number	Cluster's Composition											
1	61											
2	84											
3	2	4	5	8	9	10	12	13	15	16	17	24
	25	27	29	31	32	33	34	43	44	45	46	47
	48	49	50	51	52	53	54	56	62	64	65	69
	71	72	74	76	79	80	81					
4	11	30	37	38	40	41	42	57				
5	1	3	6	7	14	18	19	20	21	22	23	26
	28	35	36	39	55	58	59	60	63	66	67	68
	70	73	75	77	78	82	83					

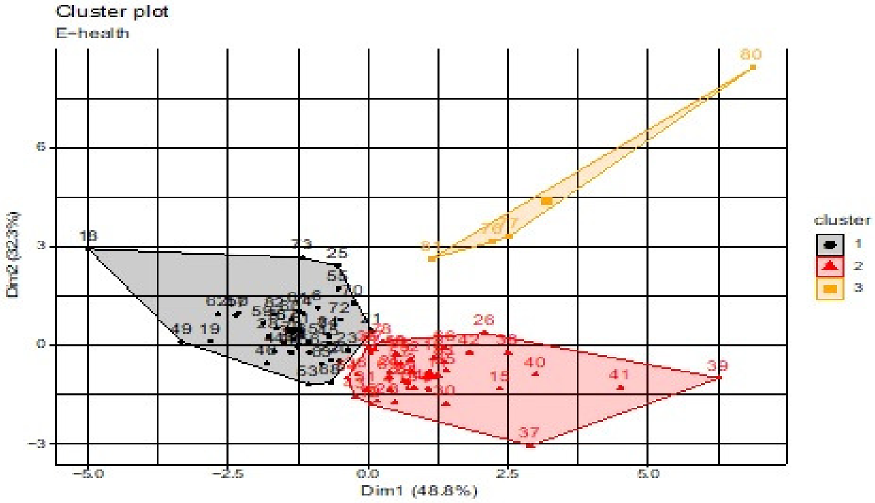

Fig. 5. Distribution of the Subjects of the RF in Three Clusters (Visualization of the Results of Cluster Analysis by the EM Method).

Table 2 demonstrates the results of hierarchical cluster analysis (classification by the subjects of the RF)

In the next step, we perform cluster analysis of the subjects of the Russian Federation according to the dataset of the "E-health" factors. The dataset includes: the share of healthcare providing institutions that use personal computers in the total examined healthcare institutions; the proportion of health facilities with local computer networks in the total health facilities surveyed; the share of health facilities using the Internet in the total

Analyzing the clustering results for 3 clusters, we can note the possibility of further decomposing the first cluster in order to optimize the resulting emissions.

number of health facilities; the number of personal computers per 100 workers in medical care units; the number of personal computers with access to global information networks per 100 employees in health care institutions; the number of personal computers connected to the Internet per 100 healthcare workers; the proportion of health facilities with a website in the total health facilities surveyed. According to the majority rule, the best number of clusters is 3 (Fig. 5).

In view of the considerable distance of the Amur Region (serial number 80) from the centroids of all the clusters built, it is impossible to conduct a normal clustering procedure. Consider a variant of building five clusters (based on the BIC indices to determine the optimal set).

The new distribution (Fig. 6) places the Amur region in a single cluster, which allows for more informative clustering. Further, pass on to hierarchical clustering by the Ward Hierarchical Clustering method, which also results in the five clusters (Fig. 7).

Table 3 presents the results of hierarchical cluster analysis (classification by subjects of the RF).

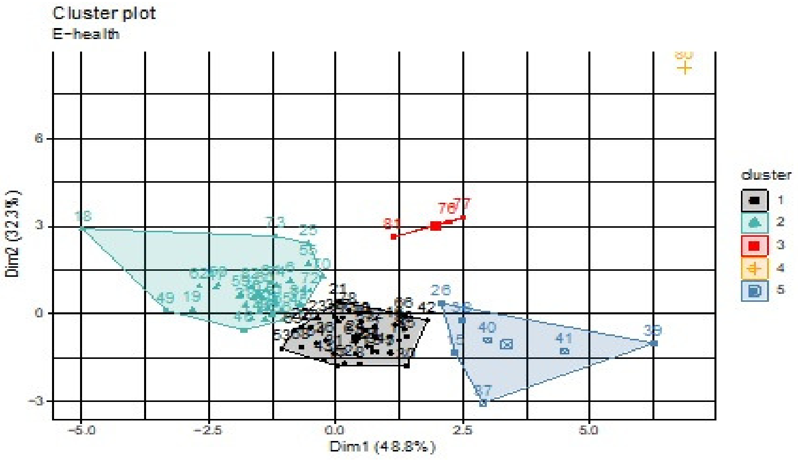

Fig. 6. The Distribution of the Subjects of the Russian Federation in Five Clusters (Visualization of the Results of Cluster Analysis Using the EM Method)

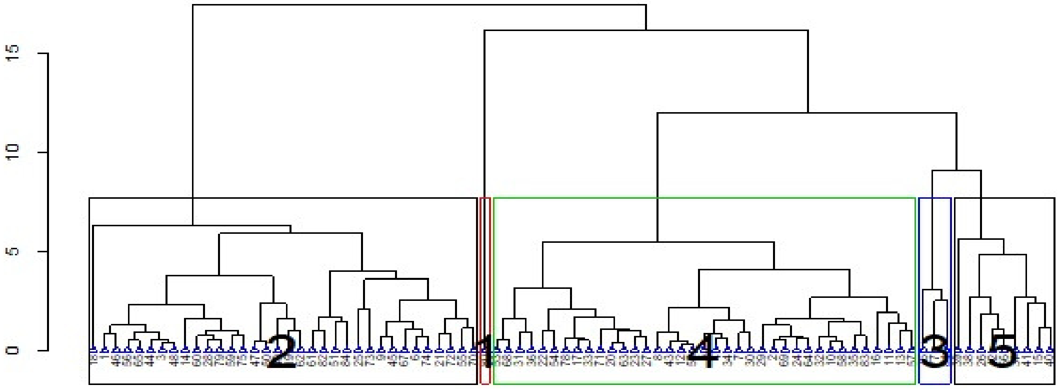

Fig. 7. Dendrogram of Hierarchical Cluster Analysis (Visualization of the Results by Ward's Method)

Table 3: The Results of Hierarchical Cluster Analysis.

Cluster number	Cluster's Composition										
1	80										
2	1	3	6	9	14	18	19	28	44	45	46	47
	48	49	50	51	56	59	60	61	62	65	67	71
	72	74	75	79	82	84						
3	2	4	5	7	8	10	11	12	13	16	17	20
	21	22	23	24	27	29	30	31	32	33	34	35
	36	42	43	52	53	54	57	58	63	64	66	68
	69	78	83									
4	15	26	37	38	39	40	41					
5	25	55	70	73	76	77	81					

At the final stage of cluster analysis, we conduct clusterization of the subjects of the Russian Federation based on the factors that identify E-culture [7-9]. These factors comprehend: the proportion of cultural institutions that had a website in the total cultural institutions surveyed; the share of electronic documents on removable media in the total volume of the library fund; the share of museum items included in the electronic catalog in the total volume of the general museum fund; the volume of electronic library catalog available on the Internet; the number of library documents converted into electronic form; the number of museum items listed in the electronic catalog; the number of museum items available on the Internet, listed in the electronic catalog and having digital images per 10,000 items of the general museum fund [10-14].

We will carry out cluster analysis with localization of four clusters using the EM method (Fig. 8).

We will complete the cluster analysis using Ward's Hierarchical Clustering method with four clusters (Fig. 9).

Table 4 presents the results of hierarchical cluster analysis (classification by subjects of the RF).

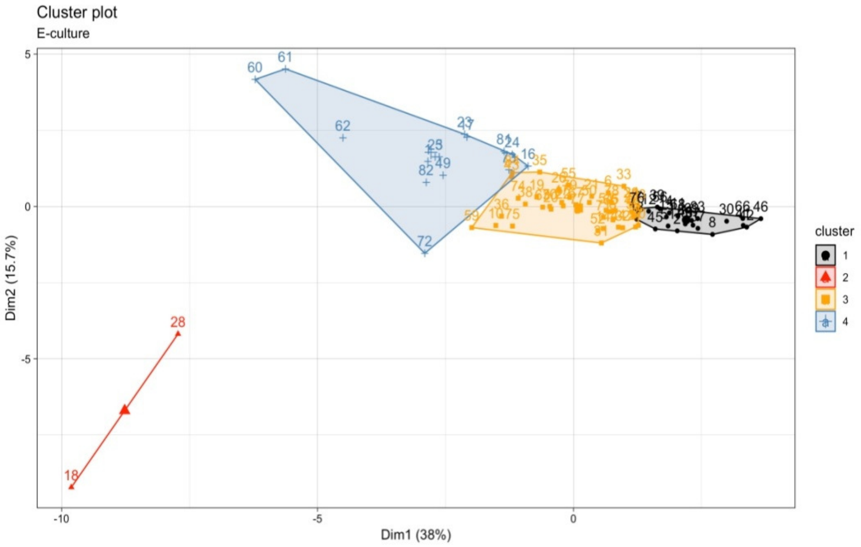

Fig. 8. Distribution of the Subjects of the Russian Federation in Four Clusters (Visualization of the Results of Cluster Analysis by the EM Method).

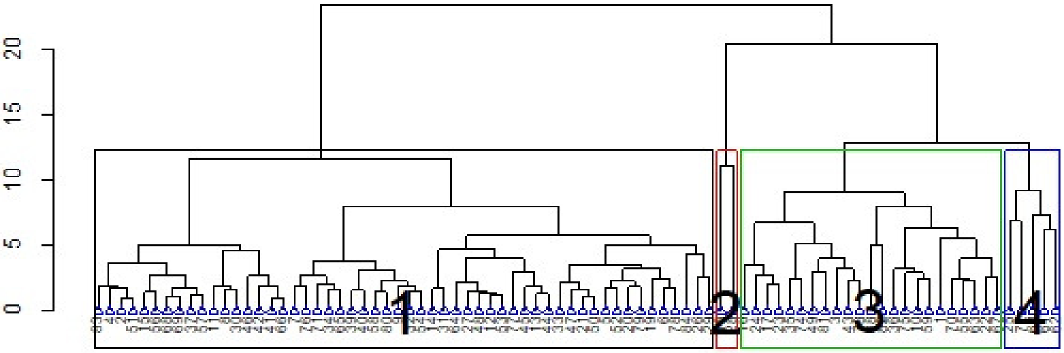

Fig. 9. Dendrogram of Hierarchical Cluster Analysis (Visualization of the Results by Ward's Method).

Table 4: The Results of Hierarchical Cluster Analysis.

Cluster number	Cluster's Composition											
1	2	4	5	6	7	8	9	11	12	13	14	15
	19	20	21	26	27	29	30	31	32	33	34	37
	39	40	41	42	44	45	46	47	48	50	51	52
	53	54	56	57	58	64	65	66	68	69	71	76
	77	78	79	80	83	84						
2	18	28										
3	1	3	10	16	17	22	23	24	35	36	38	43
	49	55	59	63	67	70	73	74	75	81		
4	25	60	61	62	72							

IV. SUMMARY

As a result of cluster analysis of data, the clusters of the Russian regions differing in the level of development of the resource base, processes and results of digitization of the economy of the subjects of the Russian Federation have been formed.

Based on the data analysis, the existence of a stable set of the subjects of the Russian Federation, differing from the majority of the subjects of the Russian Federation by specific characteristics of human capital and electronic technologies, has been substantiated. These include the following subjects of the Russian Federation: Moscow Region (10), Moscow (18), St. Petersburg (28), Republic of Crimea (31), Chechen Republic (42), Republic of Tatarstan (47), Khanty-Mansi Autonomous Okrug–Yugra (60), Yamalo-Nenets Autonomous Okrug (61), Tyva Republic (66), Magadan Region (81), Chukotka Autonomous Okrug (84). This is quite predictable since it is these subjects of the Russian Federation that are distinguished by extremes of the values of human capital factors and electronic technologies. The results obtained are consistent with the rating data of the Council for Regional Informatization of the Government Commission on the Use of Information Technologies for Improving the Quality of Life and the Conditions for Business.So, from the identified list of separate objects, the city of Moscow (18), the Republic of Tatarstan (47), the Autonomous

Okrug of Khanty-Mansiysk–Yugra (60) are in the top five of this rating. Among the ratings that close the list are the Republic of Crimea (31), the Chechen Republic (42), the Republic of Tyva (66), the Region of Magadan (81) and the Autonomous Region of Chukotka (84).

V. CONCLUSIONS

The results of the cluster analysis of the subjects of the RF substantiate the effectiveness of the use of Principal Component Analysis, EM (Expectation Maximization) and the Ward Hierarchical Clustering for clustering multidimensional human capital data and digital transformation of the economy of the subjects of the RF. At the same time, the number of selected clusters and their composition are almost identical when using the EM methodand applying hierarchical cluster analysis, Ward's method.

ACKNOWLEDGEMENTS

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

[1]. Monitoring of Information Society Development in the Russian Federation (data October 03, 2018). – URL: http://www.gks.ru/free_doc/new_site/figure/anketa1-4.html, free

[2]. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques Morgan Kaufmann, - 703 P. - Isbn: 0123814790 (3rd Edition)

[3]. Ivanenko, L.V., Karaseva, E.A., & Solodova, E.P. (2020).Clusters, Digital Economy and Smart City. In: Ashmarina S., Mesquita A., Vochozka M. (Eds) Digital Transformation of the Economy: Challenges, Trends and New Opportunities. *Advances In Intelligent Systems and Computing*, Vol. *908*, Springer, Cham

[4]. Weichhart, G., & Rosemann, M. (2016). Guest Editorial: Cooperative Information Systems In the Digital Age. *International Journal Of Cooperative Information Systems*. October 2015.

[5]. Sundberg, L. (2019). Electronic Government: Towards E-Democracy or Democracy At Risk?.*Safety Science,118*: 22-32

[6]. Altilio, R., Di Lorenzo, P., &Panella, M. (2019). Distributed Data Clustering Over Networks Pattern Recognition. *93*: 603-620.

[7]. Kiseleva, T.V., &Boiko, Y.A. (2015). Application of Mobile Technologies in E-education.*The World of Science, Culture, Education*, *3*.

[8]. Kiseleva, T. V., & Khudoverdova, S.A. (2017). Formation of the Information and Education Environment of Higher Establishment Based on Portal Technology. *Informatics and Education*, *5*: 49-59.

[9]. Saner, H. (2016). E-Health and Telemedicine: Current Situation and Future Challenges. *Cardiology and Cardiovascular Surgery*, *9*(1): 8-12

[10]. Krivosheev, V. V. (2013). Electronic Culture: The Pre-Requisite of Interdisciplinary Approach to Learning. *Herald of Immanuel Kant Baltic Federal University*, *6*: 76–81.

[11]. Lyakh, S. S. (2014). Electronic Culture as a Constituent Part of Spiritual Growth of Modern Russian Society.*Highlights of Social Sciences: Sociology, Politology, Philosophy, History: Collected Papers of the 33 th International Science and Practice Conference*,*1*(33), Novosibirsk: SibAK,.

[12]. Urnyshev, R. On Monitoring of Information Society Development in the Subjects of the Russian Federation (data date May 03, 2019). – URL: http://tomedu.ru/wp-content/uploads/2015/02/Vopros_8_Rejting.pdf, free

[13]. Mendonça, C. M. C. D., & Andrade, A. M. V. D. (2018). Dynamic Capabilities and Their Relations with Elements of Digital Transformation in Portugal. *Journal of Information Systems Engineering & Management*, *3*(3), 23.

[14]. Ranjbari, M. H., Shaheri, A., Dalili, R., & Soroush, R. (2015). Optimal allocation of distributed generation using an analytical method with consideration of technical and economic parameters. *UCT Journal of Research in Science, Engineering and Technology*, *3*(1), 9-17.