Data Modelling to Analyze How the Cities in the Volga Region Correspondent to the Digital State Format

I N Khaimovich[1,2]**, V M Ramzaev**[1] **and V G Chumak**[1]

[1]Samara University of Public Administration International Market Institute, G.S. Aksakova Street 21, Samara, Russia, 443030
[2]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086

Abstract. The article suggests the methodology for assessing the readiness of Volga region municipal entities to introduce digital state. The authors worked out a model of statistic tests based on multiple probability theoretic and statistical modelling of parameter values of technological elements of digital economy. This model will allow the Volga region cities to define their possibility to participate in the State Program of Digital Economy, to choose cities most suitable for introduction of modern technologies, to identify the main shortcomings hindering their integration into the program. This research may be of interest to experts in the field of digital economy and Big Data management.

1. Introduction

At present, Russia is carried out a transition to digital development of the state and the digital economy. In the context of fierce competition among municipal entities, the issue of introducing information technologies becomes especially urgent. Municipal entities (ME) - the city districts are key elements of the territorial organization of the country's economy, where new business, financial and cultural centers are formed, which stimulate change. The implementation of modern forms and tools of the digital economy in the ME is currently constrained by the following problems: restricting access to the digital systems in municipalities, the depressed state of the ME economies, the lack of social involvement of local government in the management process, and the lack of bases for building the digital economy.

At the same time, further development of the ME is impossible without attracting new intellectual services, one of which is the technology of the digital state. The following questions arises:

- Are MEs ready to introduce digital state technologies, i.e. to the elimination of intermediaries in services, the implementation of direct transactions in the economy, etc?
- What is the methodology which is used to check the readiness.

When managing the development of the regional economy in the modern digital format, it is necessary to solve the following tasks:

- consider the main ratings of the conformity assessment of the ME to the principles of the digital state;
- develop a methodology for applying, through analysis of compliance data, to the principles of the "digital state";
- analyze the cities of the Volga region using this method.

2. Analysis of existing ratings in the concept of “digital state”

The concept of "digital state" includes the following concepts:

- "smart economy", which includes a high level of indicators in the areas of innovation, employment, trade, productivity, physical infrastructure;
- "smart environment" that links air quality index, water supply, noise level, environmental quality, biodiversity, power economy;
- "smart society and culture", which includes education, health service, security, housing, culture, social involvement.

All these indicators should interact only through information and communication technologies (ICT).To analyze the readiness and attract investment in the ME, the "Smart City" rating is developed by the United Nations [1] and along with that the Russian rating for the assessment of sustainable urban development was developed [2]. Analysis results are given in table 1.

Table 1. Ratings of “digital” state in Russia and abroad.

Ratings	Indicators
Rating of sustainable development of cities in RF	Economy, municipal infrastructure, social sphere, environmental situation
Smart city system of indicators	Economy (physical infrastructure, innovations, employment, trade, productivity, information and communication technologies). Environment (air quality, water supply, noice level, environmental quality, biodiversity, power economy). Society and culture (education, health service, security, housing, culture, social involvement)

Let's consider in detail the rating of sustainable development of Russian cities, proposed by SGM Agency. This rating includes more than thirty indicators that characterize ME: economy, municipal infrastructure, social sphere, ecology. The respondents of the selection are the administrative centers of Russia. The advantage of this rating is the balance of the indicators under consideration, since unbalancing adversely affects the sustainable development of the ME. The disadvantages include the fact that high performance does not always determine the leading position of the ME in the country. Every year the leading cities are the same in this ranking: Moscow, St. Petersburg and Ufa. The administrative centers of the Volga region are annually referred to the outsiders. The shortcomings in the context of the topic under consideration include the isolated character of the indicators of this rating to the topic "digital state". For example, the indicator "demography", consisting of the criteria "natural growth rate", "migration growth rate", "demographic burden" only indirectly affects the use of ICT in all areas of the city.

Next, consider the smart cities metrics system developed by the United Nations Economic Commission for Europe. The degree of readiness to introduce new ICTs is assessed through the city's innovation indicators to improve the living standards of the population. This system assesses the effectiveness of activities and services to meet future generations in various aspects of activities.

This system consists of the following three blocks:

Block 1. Economics: ICT infrastructure, innovation, employment, trade - e-commerce, trade - export / import, productivity, physical infrastructure - water supply, electricity, health service, transport, buildings.

Block 2. Environment: air quality, water supply, noise level, environmental quality, biodiversity, power economy.

Block 3. Society and culture: education, health, safety – consequence management, security - emergencies, security - ICT, housing, culture, social involvement. The last indicator includes calculation of levels of public participation, gender equality of incomes, the ability of people with special needs, attractiveness for qualified personnel, the Gini coefficient.

The advantage of this system is the detailed calculation of all indicators. The disadvantage is that this assessment can be applied only to European countries.

After a detailed analysis of European and Russian ratings and evaluation systems, a methodology was developed for assessing cities to introduce technologies of the "digital state", taking into account all the features of the Volga region [3-6].

3. Involvement model of ME of Volga region in “digital economy” technology

Despite a large number of studies in this field [7-11], there is no unified methodology for assessing the involvement of ME in “digital state” technology. International ratings and systems are only being tested and they answer specific questions: is the economy of the city stable; does it have elements of a "smart" city. They do not assess any indicators of intellectualization of the urban environment.

Further, a model of statistical tests will be considered based on repeated probability-theoretical and statistical modeling of the parameter values of the "digital state" concept. This model will be associated with the analysis of a large amount of data, which will require the automation of evaluation indicators, and it can be based on the Big Data technology [12,13].

The method of using database mining technique in order to support business objectives is as follows:

1. Formation of a big data set in the hadoop from the twitter using the filter “Samara region”, revealing the number of calls;
2. Division of the formed set into various filters associated with the performance measures of the involvement of the ME in the "digital state" technology;
3. Monitoring of the stream analysis of unstructured data sources using filters;
4. Development of a program in Scala Programming Language for working with filtering in the field of Big Data;
5. Debugging and program testing with a set of practical data;
6. Analysis of calculation results.

The social network “Twitter” is used in order to receive data, since it is an “open” product, its application does not require additional investment, and 50% of Internet users have profiles in this program. Twitter is the second most popular network among users worldwide, second only to Facebook. However, unlike Facebook, which does not provide open access to its data, Twitter provides such access; there are no restrictions on access to the server's data sets. Users of this social network exchange mainly textual information, which is an undoubted advantage in processing. Twitter is not an object network and most widely reflects public opinion on many issues of interest, so data processing from this social network was best possible for the formation of small business zones in the region.

To work with BIG DATA in social networks it is necessary to use methods of data collection, processing and analysis. The data is collected in real time, within a certain geolocation, or within the entire network, according to certain patterns. Information of interest for analysis is: location, date and time, content, “author” of the content (user), communication between users. Data collection in social networks can be performed using the following tools: Apache Hadoop, Biglnsights (IBM), Cloudera, Hortonworks, and Storm. Hortonworks was chosen to carry out research on ME involvement in the "digital economy". Twitter Application (apps.twitter.com) was used, in which the following key parameters were defined and refined: API key, API secret, Access token, Access token secret.

To collect data using Hortonworks, Twitter App, the Flume service configuration file was used in the Hortonworks Sandbox virtual machine. After installing the virtual machine Hortonworks_ Sandbox version 2.3 and the Flume service settings, the system is ready to download data from Twitter. To view and download files, go to the HDFS folder where the data is process. The types of the HDFS file structure in the Hortonworks virtual machine when solving the task of ME involvement is shown in figure 1.

The collected data must be structured (i.e. processed) in accordance with the MapReduce paradigm. MapReduce is a framework for performing distributed tasks using a large number of computers that form a cluster.

Using MapReduce helped to structure the data stream from social networks by the criteria: fonts, text size, color, link to user profile, location, time and so on.

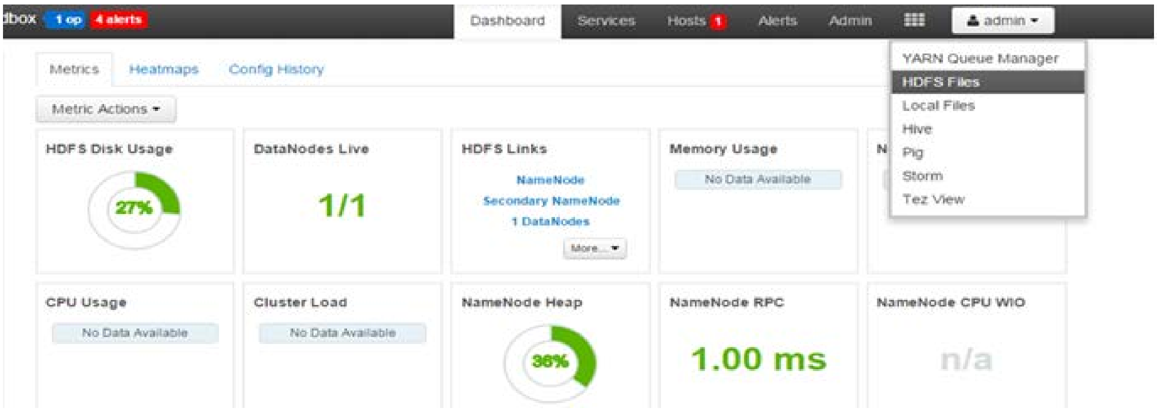

Figure 1. Visualization of HDFS in Hortonworks when downloading files to solve the problem of ME involvement in the “digital economy”.

To determine the data for ME analysis, for our study it is necessary to collect the data of the following types: placement, text, language and time. In order to extract only this information, you can use the MapReduce technology built into the Hortonworks Sandbox tool. For data processing we use DBMS Hive in Hadoop environment, which allows performing operations on data and their analysis by SQL-sequel like. To do this, it is necessary to create a file for processing and creating the necessary hivedll.sql. tables.

Run this file using the following command: Hive_f hiveddl.sql. Structured data will be placed in table 2.

Table 2. Type of headings for the analysis of structured data in tasks for the estimation of ME indicators.

A	B	C	D	E	F
Data/Time	Time/Zona	language	Text	location	Sentiments

This table was obtained from social network “twitter” data.

Thanks to the BIG DATA technology, it is possible to store and update data in the file system "hadoop" for the filter "Samara region" (filter1 = {Samara region}). Then it is necessary to filter this area according to the basic parameters for estimating the ME, by setting, for example, the following filters: Filter2 (economy) = {roads, goods}; Filter3 (environment) = {forest, air}; Filter4 (society and culture) = {nightclub, concert, session, hangout}.

It is possible to obtain graphs of the number of users accessing filters (i.e., the value of the ME internetization indicator) from the data collection time.

The time of data collection from the Internet in BIG DATA technology is unlimited.

As a result, we receive a dynamic change in the information in real time from the Internet, which allows us to monitor the stream analysis of unstructured information (the technology of In-Memory Data Processing and Stream) by filters with minimal investment. To implement this method, a program was written in Scala Programming Language:

```
val file = spark.textFile(“hdfs://… “)
val errors=file.filter(line=>line.contains(“Samara region“))
//count all the data
errors.count()
//count data mentioning Filter
errors.filter(line=>line. contains(“concert“)).count()
//Fetch the filter as an array of string
errors.filter(line=>line. contains(“doctor consultation“)).collect()
```

After the work of the program we obtain a dynamic change of parameters in the BIG DATA environment, which allow us to determine some of the indices of ME involvement in the “digital economy” taking into account unstructured information. This method of collecting information for estimating parameters can also be used for other social systems and sites, and also statistics of official sources posted on the Internet can be used to collect information

Thus, a tool is proposed for data collection in the ME system of indicators in the "digital economy".

The system of indicators of ME involvement in the introduction of "digital state" technologies is shown in figure 3.

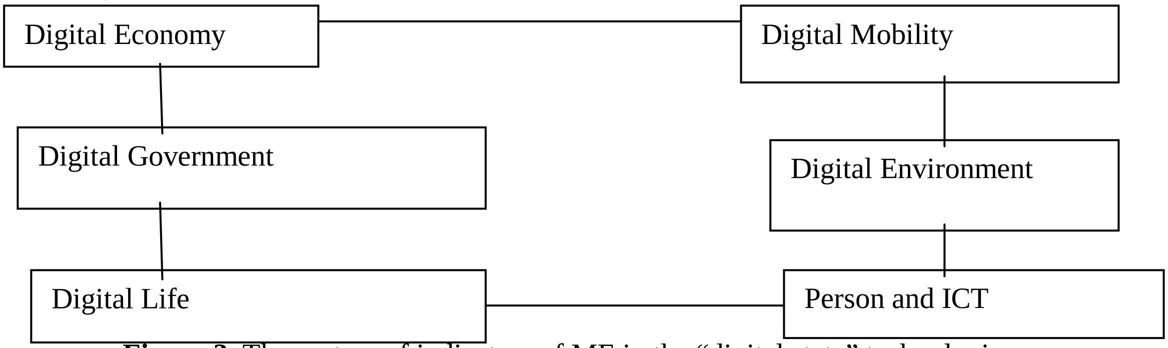

Figure 3. The system of indicators of ME in the "digital state" technologies.

The indicators in the above mentioned indicator system include the following:

- Digital Economy: the indicator for innovation, entrepreneurship, the city's competitiveness, the indicator for producibility, the labor market, the indicator for financial independence;
- Digital Mobility: local transport system, (inter-) national accessibility, ICT infrastructure, transport system stability;
- Person and ICT: an indicator for intelligence, lifelong learning, an ethnic variety;
- Digital Life: cultural and entertainment facilities, health status, individual security, housing quality, educational institutions, tourist attraction, social cohesion;
- Digital Government: political awareness, public and social services, effective and transparent administration;
- Digital Environment: air quality (without pollution); environmental awareness, sustainable resource management.

These indicators have expert values. To compare different indicators, it is necessary to standardize the values from the samples of several cities. The study uses the standardization method of z-transformation using the following formula:

$$z_i = \frac{x_i - \bar{x}}{S},$$

where $\bar{x}$ is the average value in the sample, S is the standard variation in the sample. This method converts all values of the indicators into standardized values with an average of 0 and a standard deviation of 1. The method has the advantages of considering heterogeneity within groups and maintaining its metric information. In addition, high sensitivity to changes is achieved.

In order to obtain results by indicator level, indicators and final result for each city, it is necessary to summarize the values at the level of the indicator. To aggregate the corresponding group of indicators by domain, we also take into account the coverage factor of each indicator. A certain result from the indicator covering all cities weighs more than the indicator covering only, for example, 6 cities. In addition to this slight correction, the results were aggregated at all levels without any weighting. Aggregation was added, but divided by the number of values added. This allows us to include cities that do not cover all indicators. Their results are calculated from the available values. However, it is necessary to ensure good coverage of all cities in order to obtain reasonable results.

Some indicators can be not only expert, but also calculated, they include the indicator of manufacturability of production, the index of innovation, the indicator of Internetization, the index of intellectualization, the indicator of financial independence, the index of energy efficiency, the indicator of the introduction of creative technologies. These groupings can allow us to obtain quick management decisions depending on the average values of the indicators. If the average is in the range of 3.7 and above, then the ME is ready for the introduction of digital economy technologies. If the average value is in the range from 2, 5 to 3.7, then the ME has an average level of readiness. If the average value is in the range from 1.95 to 2.5, then the ME has a satisfactory level of readiness. If the average is below 1.95, then the ME is not ready to introduce digital economy technologies [14].

4. Results and discussion

According to the methodology developed above, there were carried out some calculations of some key figures and indicators of readiness of cities in the Volga region to implement the technologies of the "digital state".

Table 3 shows the calculations of the relative and absolute values of the ME in the Volga region, taking into account the z - transformation.

Table 3. Indicators of municipal entities of Volga region taking into account the z-transformations.

Digital Economy	ME (abs)	variation	ME (rel)
Indicator for innovation	3	0,547722558	0,912871
Entrepreneurship	3		0,912871
Competitiveness	2		-0,91287
Indicator for producibility	3		0,912871
Labour market	2		-0,91287
International integration (indicator for financial independence)	2		-0,91287
Total	15		
Person and ICT			
Indicator for intelligence	1	1,707825128	-1,0247
Lifelong learning	2		-0,43916
Ethnic variety	3		0,146385
Openness	5		1,317465
Total	11		
Digital Mobility			
Local transport system	2	2,217355783	-0,56373
(Inter-) national accessibility	1		-1,01472
ICT- infrastructure	4		0,338241
Transport system stability	6		1,240216
Total	13		
Digital Life			
Cultural and entertainment establishments	6	0,975900073	1,610235
Health conditions	5		0,58554
Personal safety	3		-1,46385
Energy index	4		-0,43916
Educational institutions	4		-0,43916
Tourist attractiveness	5		0,58554
Social cohesion	4		-0,43916
Total	31		
Digital Government			
Political awareness	3	0,577350269	-0,57735
Public and social services	3		-0,57735
Effective and transparent administration	4		1,154701
Total	10		
Digital Environment			
Air quality (without pollution)	4	1,154700538	0,57735
Ecological awareness	4		
Sustainable resource management	2		
Total	10		

Further the figures 4-9 show the histograms of digital city indicators for the cities of Samara and Ulyanovsk.

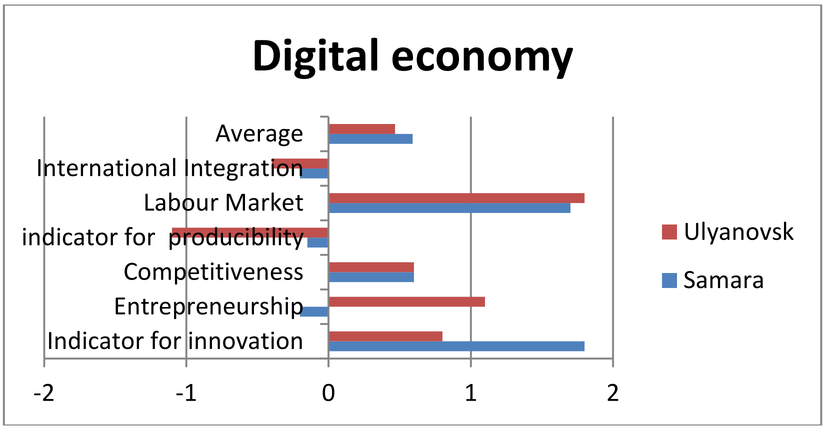

Figure 4. Indicators of "Digital Economy" for Samara and Ulyanovsk cities.

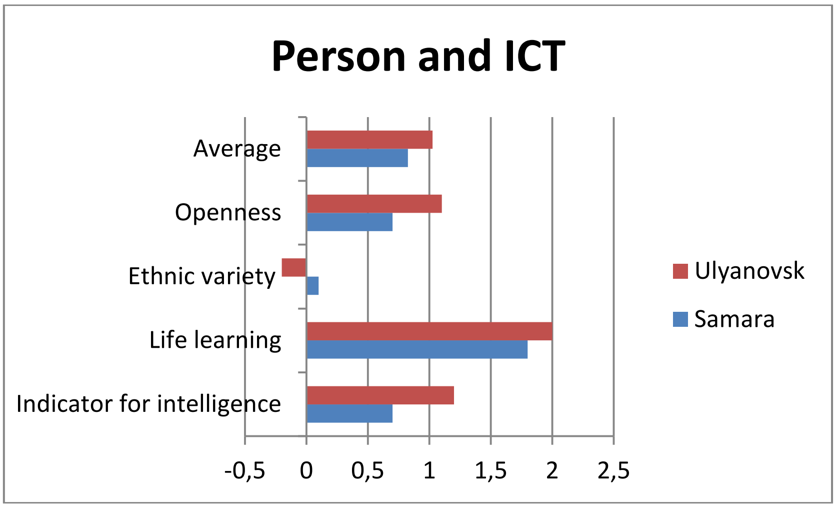

Figure 5. Indicators of "Person and ICT" for Samara and Ulyanovsk cities.

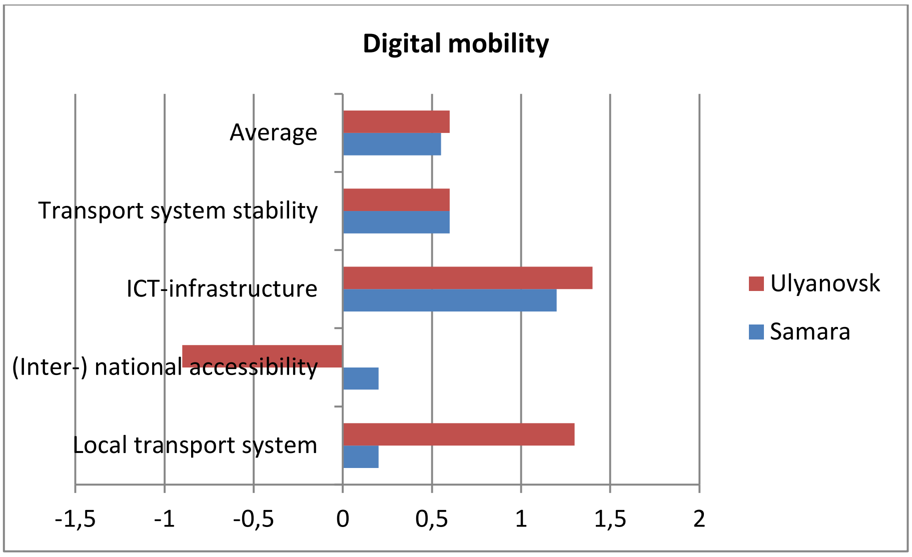

Figure 6. Indicators of "Digital Mobility" for Samara and Ulyanovsk cities.

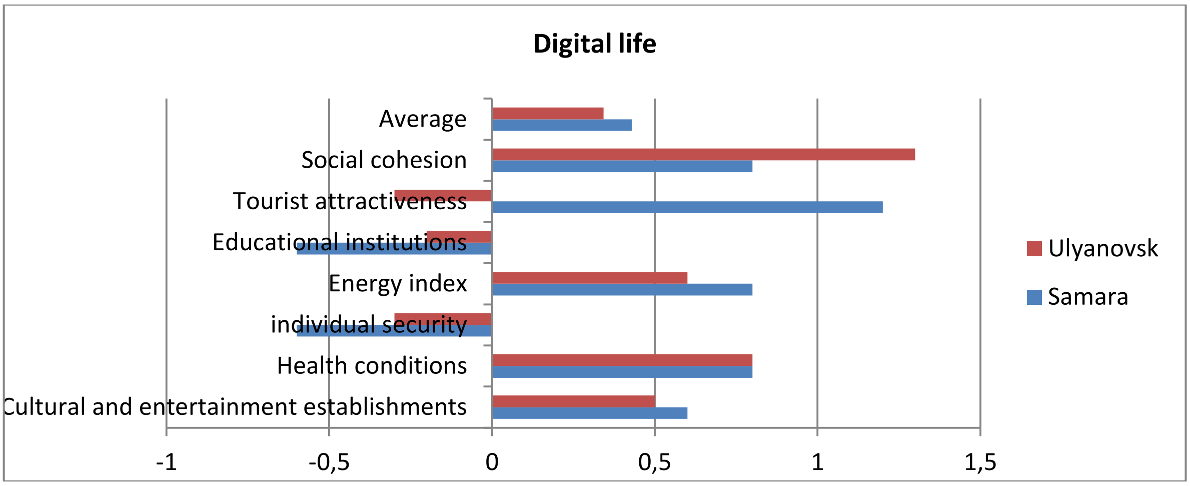

Figure 7. Indicators of “Digital Life” for Samara and Ulyanovsk cities.

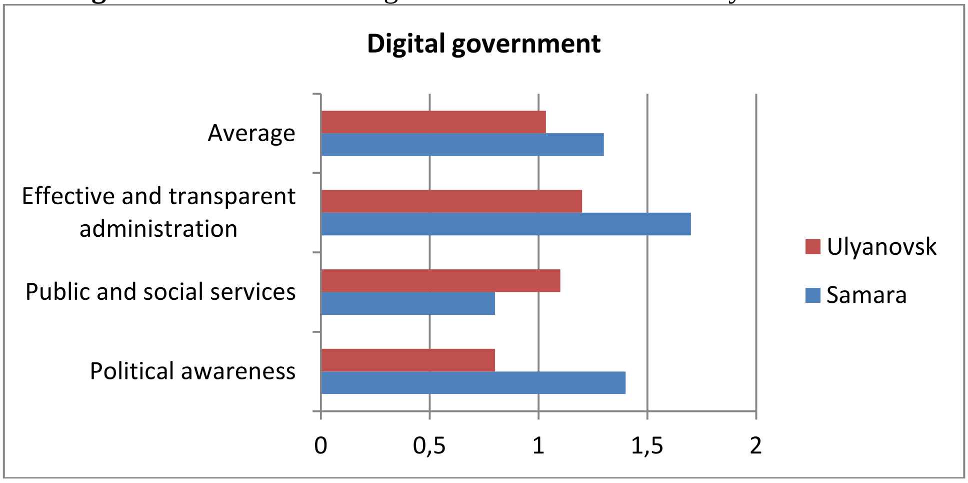

Figure 8. Indicators of “Digital Government” for Samara and Ulyanovsk cities.

Then applying a similar method it is possible to obtain level of indicators for the cities of Ulyanovsk and Samara (figure 10). From the histogram data, it can be seen that Samara is ahead of Ulyanovsk in the indicators of “digital production”, “digital economy”, “digital life”, but the indicators of "people and ICT", "digital environment", "digital life" are better for the Ulyanovsk municipal entity. In general, the willingness to introduce ICT in both cities is the same.

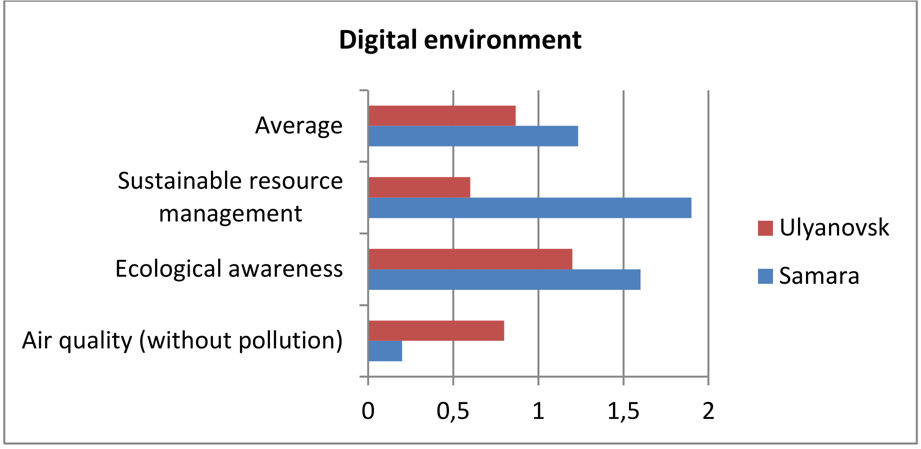

Figure 9. Indicators of “Digital Environment” for Samara and Ulyanovsk cities.

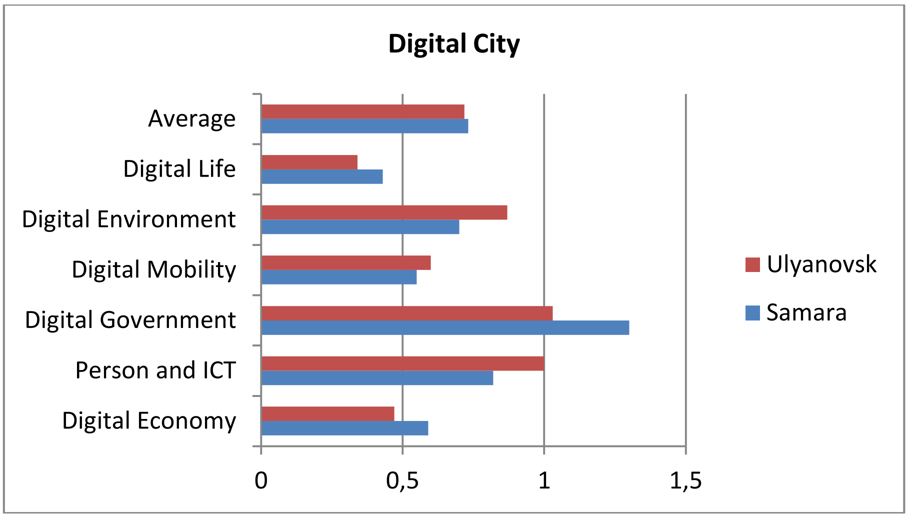

Figure 10. Indicators of “Digital City” for Samara and Ulyanovsk cities.

Let us carry out a comparative analysis of indicators of the "digital city" for municipalities with a lower level of readiness, i.e. located below the zero level, i.e. requiring significant investment in the introduction of ICT. A comparative graph of the study results is shown in figure 11.

As a result, it is possible to identify the main trends for investing in the ME of Volga region on the basis of a comparative analysis of the "digital city" indicators. Figure 11 shows that investing in a region with indicators below the zero level is not profitable. Many cities of the Volga region belong to this zone, for example, Zhigulevsk. It is better to invest in the ME with a level of readiness above the zero level, for example, Ulyanovsk. These cities correspond more to the concept of the "digital city", they are almost ready to introduce the technologies of the "digital state".

Thus, the assessment model will allow to determine the level of development of municipal entities, which are ready to implement the digital state, to identify shortcomings in the group "which is not ready for implementation", will improve the performance of the ME on the basis of a detailed analysis of data of all major cities of the Volga region.

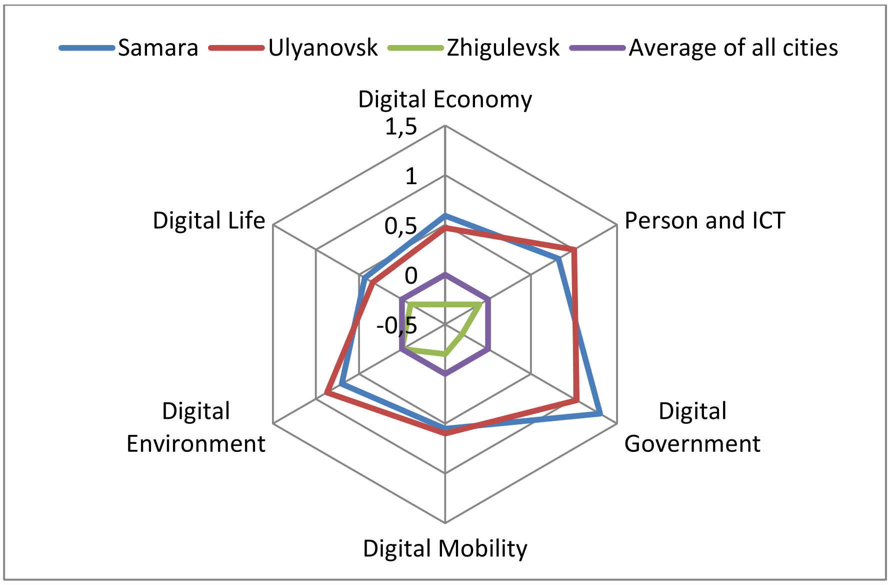

Figure 11. Graph of indicator level of Digital City for Volga region cities.

5. References

[1] Access mode: http://www.unece.org/fileadmin/DAM/hlm/documents/2015/ECE_HBP_2015_4.ru.pdf (9. 11.2017).

[2] Access mode: http://agencysgm.com/projects/Рейтинг%20устойчивого%20развития-2015.pdf (9.11. 2017)

[3] Terekhin E A 2017 *Computer Optics* **41(5)** 719-725 DOI: 10.18287/2412-6179-2017-41-5-719-725

[4] Afanasyev A A and Zamyatin A V 2017 *Computer Optics* **41(3)** 431-440 DOI: 10.18287/2412-6179-2017-41-3-431-440

[5] Vorobiova N S,Sergeyev V V and Chernov A V 2016 *Computer Optics* **40(6)** 929-938 DOI: 10.18287/2412-6179-2016-40-6-929-938

[6] Boori M S, Kuznetsov A V, Choudhary K K and Kupriyanov A V 2015 *Computer Optics* **39(5)** 818-822 DOI: 10.18287/0134-2452-2015-39-5-818-822

[7] Akaslan D and Taskln S 2016 *4th Int. Istambul Smart Grid Congress and Fair* (New York: IEEE Press)

[8] De Domenico M, Lima A A and Gonzalez M C 2015 *EPJ Data Science* **1** 1-11

[9] Glebova I S, Yasnitskaya Y S and Maklakova N V 2014 *Mediterranean J. of Social Sciences* **12** 129-133

[10] Ishkineeva G, Ishkineeva F and Akhmetova S 2015 *Asian Social Science* **5** 70-73

[11] Khatoun R and Zeadally S 2016 *Communications of the ACM* **8** 46-57

[12] Khaimovich I N, Ramzaev V M and Chumak V G 2016 *CEUR Workshop Proceedings* **1638** 864-872

[13] Khaimovich I N, Ramzaev V M and Chumak V G 2015 *CEUR Workshop Proceedings* **1490** 327-337

[14] Komarevtseva O O 2017 *R-Economy* **3** 32-39