DYNAMIC APACHE SPARK CLUSTER FOR ECONOMIC MODELING

Iu. Gavrilenko [1, a], M. Sharma [2, b], M. Litmaath [2, c], T. Tikhomirova[1, d]

[1] *Plekhanov Russian University of Economics, 36 Stremyanny per., Moscow, 117997, Russia*

[2] *Department of Information Technologies CERN, Geneva 23 CH-1211 Switzerland*

E-mail: [a] yulya952@rambler.ru, [b] mayank.sharma@cern.ch, [c] maarten.litmaath@cern.ch, [d] t_tikhomirova@mail.ru

Modern econometric modeling of macroeconomic processes usually meets certain challenges due to the incompleteness and heterogeneity of the initial information, as well as huge data volumes involved. In the work, on the example of modeling the level of employment in the regions of the Russian Federation was shown the effectiveness of joint using Big Data technologies and automated deployment of a dynamic virtual computing cluster for solving such problems. There were constructed several models of the regional labor market, taking into account such basic macroeconomic indicators as per capita income, the volume of paid services to the population per capita, the industrial production index and others. The classification of the subjects of the Russian Federation according to the level of employment was obtained, it is stable against different methods (single linkage, complete linkage, Ward's method). For the analysis, it was used a dynamic Apache Spark cluster deployed by the means of the SIMPLE environment developed at CERN.

Keywords: SIMPLE, Apache Spark, Hadoop, economic modeling, labour market, classification.

Iuliia Gavrilenko, Mayank Sharma, Maarten Litmaath, Tatyana Tikhomirova

1. Introduction

Nowadays, there are great difficulties in econometric modeling of macroeconomic processes due to incompleteness, heterogeneity of the initial information, and large amounts of data. And also, there is a diversity of methods for analyzing economic data that should be applied while complementing each other.

Today the standard of living of people, the cost to society of training and professional development of employees, their employment, as well as the investment attractiveness of individual regions and their well-being, in general, depending on the employment of the population. One of the most pressing problems for the country's economy today is the underutilization of human resources and, as a result, low labor productivity and rising unemployment. That is why it is necessary to study the formation of the labor market, including on the basis of factors that have a significant impact on the level of employment.

In this paper a statistical analysis of the main macroeconomic factors affecting the employment rate in the regions of the Russian Federation has been conducted, quantitative estimates of the relationship between employment and socio-economic characteristics have been obtained, and a sustainable classification of the regions of the Russian Federation has been developed on the basis of similarity of their socio-economic characteristics.

2. Multidimensional labour market analysis

The data of the Russian Federal State Statistics Service of the Russian Federation for the period from 2000 to 2016 were taken for analysis. Information was collected on the level of employment in the regions of the Russian Federation, as well as on the following indicators: average per capita income of the population (x_1, rubles), the volume of paid services to the population per capita (x_2, rubles), the cost of a fixed set of consumer goods and services (x_3, rubles), the coefficient of migration growth per population (x_4, people), the average size of pensions per capita (x_5, rubles), the turnover of retail trade per capita (x_6, rubles), the actual final consumption of household owners per capita (x_7, rubles), industrial production index (x_8, %), demand for workers declared by employers to the employment services (x_9, people), investments in fixed assets (x_{10}, rubles) [1].

Then a correlation analysis was applied to form a system of labor resources and employment management. On the basis of the calculated correlation coefficients the correlation matrix was constructed, the results of which are presented in table 1.

Table 1. Matrix of paired correlation coefficients

	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
y	1,000	0,556	0,545	0,146	0,465	-0,054	0,800	0,683	0,009	0,961	0,121
x1	0,556	1,000	0,870	0,821	-0,151	0,635	0,856	0,939	0,224	0,624	0,635
x2	0,545	0,870	1,000	0,727	-0,074	0,568	0,785	0,856	0,114	0,628	0,508
x3	0,146	0,821	0,727	1,000	-0,395	0,806	0,518	0,703	0,281	0,252	0,602
x4	0,465	-0,151	-0,074	-0,395	1,000	-0,548	0,177	0,006	-0,044	0,414	-0,173
x5	-0,054	0,635	0,568	0,806	-0,548	1,000	0,258	0,489	0,148	0,001	0,605
x6	0,800	0,856	0,785	0,518	0,177	0,258	1,000	0,924	0,098	0,826	0,410
x7	0,683	0,939	0,856	0,703	0,006	0,489	0,924	1,000	0,175	0,746	0,541
x8	0,009	0,224	0,114	0,281	-0,044	0,148	0,098	0,175	1,000	0,025	0,251
x9	0,961	0,624	0,628	0,252	0,414	0,001	0,826	0,746	0,025	1,000	0,155
x10	0,121	0,635	0,508	0,602	-0,173	0,605	0,410	0,541	0,251	0,155	1,000

Different factors depend on each other to different degrees. There are both direct and inverse linear dependencies. Factors x_1 and x_2 (average per capita cash income of the population and volume of paid services to the population per capita), x_1 and x_3 (average per capita cash income of the population and the cost of a fixed set of consumer goods and services), x_1 and x_6 (average per capita cash income of the population and turnover of retail trade per capita), x_1 and x_7 (average per capita cash income of the population and actual final consumption of households per capita) x_2 and x_7 (volume of paid services to population per capita and actual final consumption of households per

capita), x_3 and x_5 (cost of a fixed set of consumer goods and services) and the average size of pensions per capita), x_6 and x_7 (turnover of retail trade per capita and actual final consumption of households per capita), x_6 and x_9 (turnover of retail trade per capita and the demand for workers declared by employers to the employment services) are connected most of all. The modular values of their correlation coefficients are higher than 0.8.

In order to identify the structure of the multivariate population under study, a cluster analysis was carried out. The main task of this method is to obtain homogeneous groups of objects, signs, and factors. The selected macroeconomic factors were analyzed for informativeness. For this purpose, we will conduct a cluster analysis of the factor space using the "Near Neighbor" method. First of all, we will standardize the data in order to eliminate economies of scale. Figure 1 displays the final dendrogram showing the degree of proximity of various factors. Thus, the group of factors x_1 (average per capita income of the population), x_6 (turnover of retail trade per capita) and x_7 (actual final consumption of households per capita) and a set of factors x_3 (cost of a fixed set of consumer goods and services), x_5 (average size of pensions per capita) and x_9 (demand for workers, declared by employers in the employment services) are most closely related.

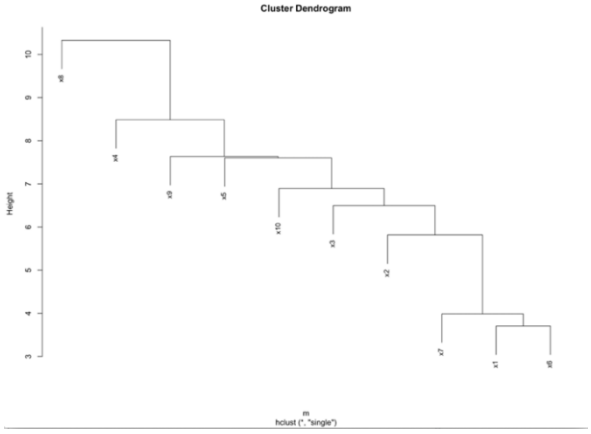

Figure 1. Dendrogram of the factor space

After studying the correlation vector that measures the degree of tightness of the linear relationship between the parameters, let's exclude successively the factors x_1, x_7, x_3, x_5 from the initial factor space and re-build clusters on the obtained data. After all non-informative factors have been excluded it is possible to continue the analysis of multidimensional data.

Then the following analysis of the multivariate emission data was carried out. For this purpose, a cluster analysis of the object space using the "Near Neighbor" method, using standardized data, was used. Figure 2 shows a dendrogram, on which atypical regions are clearly distinguished, namely: Moscow, Chukotka Autonomous Okrug, Tyumen Region, Sakhalin Region, Kamchatka Territory, Magadan Region. We will exclude the identified atypical regions from the common object space and continue to form a stable cluster structure.

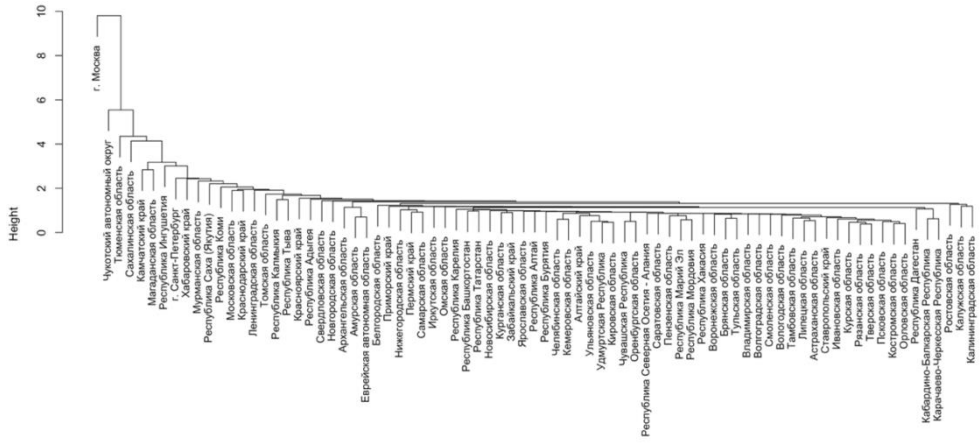

Figure 2. Dendrogram of the object space by the "Near Neighbor" method

For the remaining regions, a cluster analysis was conducted using Ward, Far Neighbour and k-means methods [3]. Based on these methods, regions were sustainably classified (fig. 3). The first cluster is characterized by the negative behavior of the regions according to the chosen characteristics: paid services to the population are provided here in a smaller volume, the negative coefficient of migration growth - people of working age leave these constituent territories of the Russian Federation (possibly because of the tense social situation), the average size of pensions is lower than the median value of the indicator throughout Russia, low indices of industrial production and investment in fixed capital, small need for workers. The second cluster is characterized by the average behavior of the regions according to the selected characteristics since the difference between the centroids of the groups and the median values of indicators throughout Russia are close to zero. The third cluster is

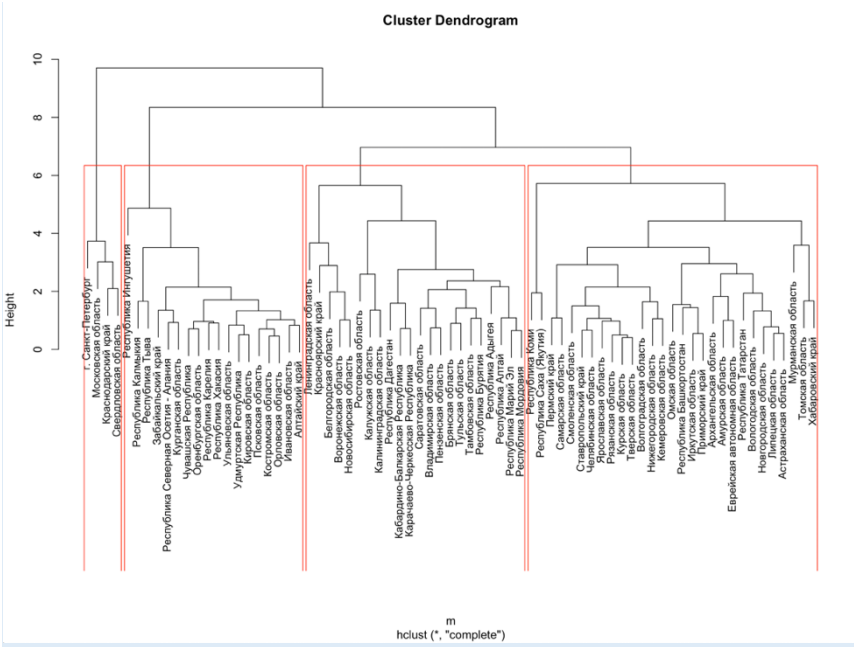

Figure 3 Stable classification of the regions of the Russian Federation

characterized by positive behavior of the regions in terms of the selected indicators, as the difference between the cluster centroids and median values for all major macroeconomic indicators in Russia is already above zero.

.

3. Distributed computing technologies for economic data analysis

For analysis performing, a dynamic Apache Spark cluster deployed by the means of the SIMPLE framework [3, 4] developed at CERN was used. The SIMPLE Grid Project is a Solution for Installation, Management, and Provisioning of Lightweight Elements. SIMPLE Grid Framework, the core component of the project, offers a common interface to a site admin for configuring such grid services using the infrastructure management tools they prefer to use or already use at their site. The framework does not expect a site admin to be familiar with the intricate architecture and workflows involved with the grid services they need to work with. Instead, it focuses on abstracting low-level details of implementing a working WLCG site based on the grid services and middleware packages specified by site admins, which they may or may not be familiar with. The framework ensures that at the end of the day, a site has installed and configured all the essential packages and is running the necessary services that are expected for processing WLCG workloads. At the same time, the framework also ensures that the process of setting up these packages and services is simplified from the site admin's perspective [3].

Let's consider the main stages of building the Hadoop Spark cluster. The first step is to create Docker containers to initialize the master and slave nodes. With the help of the *docker-compose.yaml* file, where the names of the containers are indicated, all the necessary environment variables are written, ports are opened for connecting the nodes with each other and for the possibility of editing the work execution in the Web-page mode, and the received containers are launched. Then there is local

testing - the launch of a simple test task, the correctness of the network operation is checked. Then all containers are placed together in a cluster environment in which the Swarm network of the Docker Swarm orchestra application becomes the master. After that, a separate repository component is created for nodes performing different tasks (in this case, two repositories have been created for the master and slave nodes) [5]. The components of the repositories are special files: *meta-info.yaml*, which declares such parameters as the node name, its type, the way of launching the container, the necessary ports; *config-schema.yaml*, where all the required configuration variables to be provided by the user to SIMPLE framework are listed; *default-data.yaml* contains default values to avoid a task stop situation if the user does not provide information about any variable; *requirements.txt* with all required installation files and packages*; pre_config.py* generates output files, including *augmented_site_level_config_file.yaml* - it is from this application that all the information for launching the SIMPLE framework containers is taken; the main configuration file is *init.sh* where all the commands for launching containers are specified. To test the obtained components, two files *simple_spark_hadoop_master.sh* and *simple_spark_hadoop_worker.sh* are created, which simulate the life cycle of the container. And finally a *site_level_config_file.yaml* configuration file is created, where the IP-addresses and names of virtual machines - the main and working nodes - are written, as well as links to the corresponding repositories are specified and the main parameters of the environment required by the user are set. Then the cluster is launched, the Web-interface of which is shown in Figure 4.

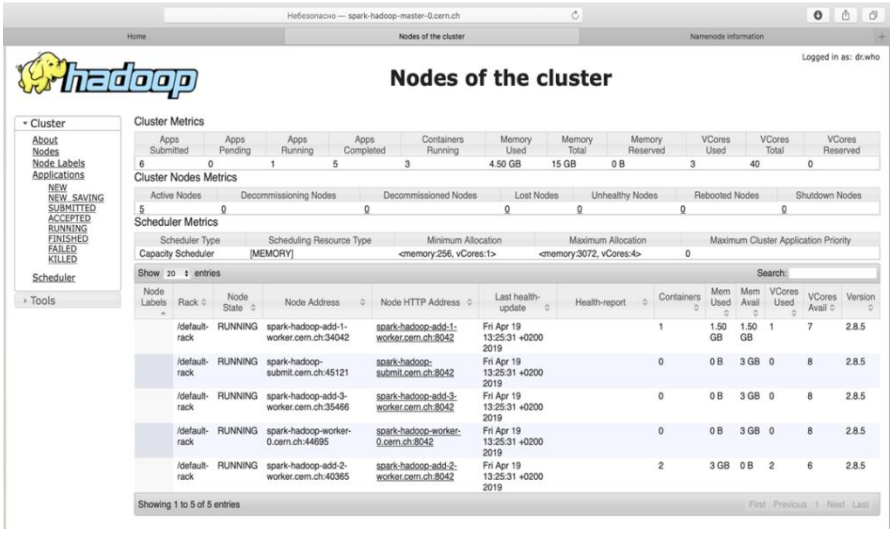

Figure 4. Hadoop Spark Cluster Web Interface

5. Conclusions

Summarizing, we may say that a modern approach to data analysis often requires high operating costs. Due to the large volume of information, it is necessary to use special programs even for simple modeling of the economic process. Created on the basis of distributed technologies cluster Hadoop Spark allows you to run the code written in Python for integrated statistical analysis of data, conducting a cluster analysis of data by hierarchical and iterative methods and obtaining a sustainable classification of the regions of the Russian Federation.

Acknowledgments

The study on the creation of dynamic scalable infrastructure solutions for economic modeling was carried out at the expense of the Russian Science Foundation grant (project No. 19-71-30008).

References

[1] Database at the Russian Federal State Statistics Service. Available at: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/demography/ (accessed: 17.05.2019).

[2] Steinley D. K-means clustering: a half-century synthesis. British Journal of Mathematical and Statistical Psychology, 2010, https://doi.org/10.1348/000711005X48266

[3] M. Sharma, M. Litmaath, E. Silva Junior, R. Santana – Lightweight WLCG Sites The SIMPLE Grid Framework, proceedings of CHEP'2018 conference, 9-13 July 2018, Sofia, Bulgaria, https://doi.org/10.1051/epjconf/201921407019

[4] M. Sharma et al. The SIMPLE Grid project. Available at: https://wlcg-lightweight-sites.github.io. (accessed: 20.11.2019).

[5] Iuliia Gavrilenko. Master node repository. Available at: https://github.com/JuliaGavrilenko/simple_spark_cluster_master (accessed: 20.11.2019).