

Elaboration of directions for sustainable socio-economic development using Big DATA

Ferova I.S.
Siberian Federal University
Krasnoyarsk, Russia
iferova@yandex.ru

Makarova S.N.
Siberian Federal University
Krasnoyarsk, Russia
msn2004@list.ru

Kozlova S.A.
Siberian Federal University
Krasnoyarsk, Russia
s.a.kozlova@inbox.ru

Abstract — At the present stage, one of the most important areas of the economic policy development agenda is the course towards sustainable development, which is aimed to integrate the solution for economic, environmental and social problems. The relevance of the research topic outlined in this article is due to the fact that in the current circumstances effective solutions to the existing problems in the social and economic sphere are required, which demands significant investments aimed at improving the quality and safety of the population's life. At the same time, the adoption of effective decisions should be based on the results of an objective assessment of the level of sustainable development of territories. The receipt of this assessment is impossible without the use of reliable and promptly obtained data.

This study examines approaches to using Big data in managing socio-economic systems and suggests elements of a methodology for diagnosing sustainable development of territories in accordance with the principles and content of the paradigm of a sustainable economy.

The authors proposed a methodology for assessing the sustainability of the regional economy, which allow to identify the "growth points" of the socio-economic development of the region. The authors also mark that it is necessary to use the scenario approach in combination with the concept of cluster development. In the proposed methodology, the formation of scenarios is carried out using the methodology of experiment planning, which allows the formation of numerous development scenarios with different combinations of variables. This can be realized on the basis of using the theory of Big data. The core of each experiment is to make a cluster.

Based on the results of the calculation, it will be possible to draw a conclusion on the sustainability of the development of basic clusters of the region, their impact on the socio-economic development of the territory as a whole, and to form management decisions to support them.

Keywords — *clusters, development scenarios, big data, management, sustainable development, regional economy, growth points*

I. INTRODUCTION

The problem of improving the efficiency of the Russian economy is one of the key at the present stage, one of the ways to solve it is the transition to an innovative development path. Lately, it has become obvious that systematization of various approaches to the creation of an innovative economy, an

assessment of their adequacy to Russian realities, is necessary. Nowadays, regulatory legal and by-laws of a strategic nature were adopted and are prepared, in which various development scenarios and corresponding GDP growth rates are considered. But it should be noted that Russia's lag in the creation of innovative economy is so great that even when fulfilling the targets of program documents, there is no reason for a breakthrough in the medium term. For example, it's highlighted by a comparative analysis of labor productivity, dynamics of investments and the capital-labor ratio of the Russian Federation and the United States. Moreover, if modern industrial policy is preserved, they will not appear in the long term. The situation can be reversed only by an enhanced investment process, which, as known, is closely related to the innovation one. Currently, sustainable development goals (hereinafter - SDGs) declared by the UN, are formulated, and they can be the basis for regional development strategies formation. Regarding this, an innovation and investment breakthrough is needed, which makes it possible to balance the goals of sustainable development in terms of economic, social and environmental sustainability components.

A model that would allow such a breakthrough is a cluster development model that combines all the components of sustainable development goals. At the same time, due to the effects of cluster overflows, there is an impetus for the development of the regional economy.

Thus, the task is to identify base clusters or "growth points" for a region, to form scenarios for their development and, basing on results of the scenario, to assess the sustainability of the cluster and formulate tasks for their development.

II. RESEARCH METHODOLOGY

In order to determine the role of Big data in assessing the sustainability of development of territories, identify gaps in existing research, suggest areas for further research activities, determine further research trajectories, the study conducted content analysis of publications presenting problems of applying big data and analysis techniques big data existing in modern scientific research literature.

A. Defining base clusters

There are various approaches to identify clusters: localization coefficients, size and "focus" of the cluster group, profile. In this study, emphasis was made on the investment "breakthrough." Therefore, the central point should be the reimbursement of invested capital.

In this regard, it is proposed to identify growth points with maximum return and put all efforts to obtain maximum effect, including cluster one, using the advantages of cluster influence on neighboring industries and territories.

As a base rate for determining growth points, we suggest using the investment return coefficient K_i (1), along with the already known localization and concentration factors:

$$K_i = (I_j q / I_q) / (I_j / I), \quad (1)$$

where "j" is branch;

"q" is region

"I" is the coefficient of return on changes in investments in fixed capital of industry for the total volume of industrial production, rub./rub.

The choice is made in favor of maximum performance, cluster core, elements and structure of the cluster and directions of state policy on its development and effective functioning are determined.

B. Cluster development scenarios

When solving the problem we will use mathematical models of research. By a mathematical model, we mean an equation relating an optimization parameter to factors. Each factor can take one of several values in experience. These values are called levels. To facilitate the construction of a black box and experiment, the factor should have a certain number of discrete levels. A fixed set of levels of factors determines one of the possible states of the black box. At the same time, this is a condition for conducting one of the possible experiments. If we enumerate all possible sets of states, we get many different black box states. Simultaneously, there will be the number of possible different experiments. The number of possible experiments is determined by the expression. Each factor can take one of several values in the experiment. These values are called levels. To facilitate the construction of a black box and experiment, the factor should have a certain number of discrete levels. A fixed set of levels of factors determines one of the possible states of the black box. At the same time, this is a condition for conducting one of the possible experiments. If we enumerate all possible sets of states, we get many different black box states. Simultaneously, there will be the number of possible different experiments. The number of possible experiments is determined by the expression (2):

$$N = p^i \quad (2)$$

where "N" is a number of

experiments; "p" is a number of levels;

"i" is a number of factors

Real objects usually have tremendous complexity. So, at first glance, a simple system with five factors on five levels has 3125 states, and for ten factors on four levels they are over a million. In these cases, the implementation of all experiments is

almost impossible. So, there is a question: How many and what experiments should be included in the experiment to solve the problem? Here we need to apply experiment planning in the Big Data area.

The study proposes to use 3 levels of each parameter: optimistic, pessimistic and average. As parameters of influence, it is proposed to use indicators of sustainable development goals with their classification by economic, social and environmental unit. The response function may be proposed by the GRP or human development index

C. Cluster Stability Assessment

The sustainability assessment is carried out on the basis of statistical parameters: coefficient of variation, the criterion of Student and Fisher.

D. Evaluation of the impact of the cluster on the development of the region

One of the effective tools for analyzing quantitative changes in individual sustainability criteria and assessing structural changes in the territorial-production system of a region under the influence of changes in economic relations, conjuncture processes, interregional and inter-industry relations is an indicator of structural differences - the V. Ryabtsev index. [1]. The index characterizes the differences in the structure of the indicator at two time intervals and makes it possible to understand the sustainability of the development of the economic system. The index (3) is calculated as the ratio of the actual measure of the differences in the values of the components of the two structures with their maximum possible value:

$$\Sigma \frac{(1 - \theta)^2}{\dots} \quad (3)$$

where is the index of structural differences of the components of sustainability of the i -territory;

¹ - estimated values of the j -stability criterion of the i -territory achieved under the influence of changes in economic relations, conjuncture processes, inter-regional and inter-sectoral relations;

⁰ - estimated values of the j -stability criterion of the i -territory, found at the initial stage of the study (before assessing the impact of the interaction of territories within the cluster).

The following intervals of values of the criterion of structural differences are introduced: [0; 0.03] - the identity of the structures; [0.031; 0.07] - a very low level of differences; [0.071; 0.15] - low level of differences; [0.151; 0.3] - a significant level of differences; [0.301; 0.5] - a significant level of difference; [0.501; 0.7] - a very significant level of difference; [0.701; 0.9] - the opposite type of structures; [0.901; 1] - the exact opposite of structures.

III. RESULTS OF THE RESEARCH

Currently, the amount of data in the world is constantly growing: according to UN data, up to 90% of all data accumulated in the world were created in the last two years,

and, according to forecasts, the volume of data will increase annually by 40% [2].

A significant proportion of such data is passively collected data obtained from daily digital operations: for example, mobile communications, credit cards operations, and social media data. This amount of digital data is known as big data (hereinafter - Big data, BD).

Advances in computing and computer science now allow to process and analyze big data in real time. New items derived from the processing of such data can serve as an important complement to official survey data and statistical information, thereby helping to analyze the behavior of people and their experience. The use of these new data in conjunction with the data obtained in the traditional way allows to conduct the timely analysis of information at a qualitatively higher level.

Data is an essential element of the decision-making process and the basis of the accounting and control process. Big data can be used to obtain information about people's well-being in real time and provide timely assistance to vulnerable groups. New data sources, new technologies and new analytical approaches, if applied responsibly, can allow a better tracking of progress towards achieving the SDGs.

In 2015, the world began to formulate a new development agenda, supported by sustainable development goals (SDGs). Achieving these goals requires comprehensive actions to address social, environmental and economic problems, provided that comprehensive, inclusive development is ensured.

At the same time, the problem of the lack of Big data needed for policy development at the global, regional and national levels remains relevant. Many governments still do not have access to full-fledged data on the population of their countries.

The first UN World Data Forum, held in January 2017, gathered more than 1,400 data users and representatives of public and private sector enterprises, politicians, scientists and representatives of civil society to explore ways to use data for sustainable development. The Cape Town Global Sustainable Development Data Plan was launched at the forum.

In order to conduct a study of the use of Big Data in the development of areas of sustainable development of territories, a content analysis of scientific publications of the Scopus database was conducted.

According to the results of the analysis the following problems were revealed:

- problems associated with Big data characteristics: the authors distinguish three to six characteristics: three characteristics of the "3V" data are highlighted by the authors Shah T., Rabhi F., Ray P. (2015) - volume, speed and diversity [3], four characteristics of the "4V" data are defined in Liao Z., Yin Q. Huang Y., Sheng L (2014) - volume, speed, diversity and variability [4], six key characteristics of "6V" are indicated in Gandomi A., Haider M. (2015) - volume, speed, diversity, reliability, variability and value [5];

- problems associated with large scale and amount of data: for example, the heterogeneity, ubiquity and dynamic nature of various resources and devices for data generation, as well as a huge amount of data themselves (Barnaghi P., Sheth A., Henson C., 2013) [6];

- the problem of lack of unification: a huge amount of data is not consistent and does not fit a specific pattern or format — it is recorded in various forms and different sources (Chen J. et al., 2013) [7], the various forms and quality of data clearly indicate that heterogeneity is a natural BD property, and it is a difficult task to understand and manage such data (Labrinidis A., Jagadish H.V., 2012) [8];

- inconsistency in large data sets: in this case, it is not about the quality of the data, but about their understanding and interpretation, since there are inaccuracies in almost all the collected data, for example, by truthfulness they mean overcoming bias, doubt, inaccuracies, speculation, confusion and inappropriate evidence in the data (Akerkar R., 2014) [9], the need to deal with inaccurate and ambiguous data is another aspect of BD, which is solved using tools and analytics developed for managing and analyzing unreliable data;

- the problem of speed: is associated with the need to control the high flow rate of heterogeneous data, which leads either to the creation of new data, or updating the existing data, this problem is common for data sets that are generated through large complex networks, including data generated by the proliferation of digital devices that are located everywhere, which leads to the need for real-time analytics and evidence-based planning [10];

- the problem of data collection and storage: BD complexity and exponentially growing demands create unprecedented problems in BD development, such as collecting and storing data [11], one of the main barriers to analyzing BD arises from the lack of data origin, knowledge, and scale inconsistencies inherent in data collection and processing. This limits the speed and resolution at which data can be received and stored. As a result, this affects the ability to extract useful information from the data;

- data cleaning problem: in order to use BD effectively, it is necessary to develop an extraction method that allows us to get the necessary information from an unstructured BD and formulate it in a standard and structured form that is easy to understand;

- confidentiality problem: BD poses big confidentiality issues as privacy in the digital age is paramount, in a smart city environment where touch devices collect data on citizens' actions that can be accessed, several government and security agencies pose serious privacy concerns. There is also a problem of protecting privacy - the inability to protect the privacy of citizens is illegal and open to the relevant state supervisors bodies [12];

- the problem of operating costs: constantly growing data in all different forms led to a growing demand for BD processing in complex data centers, significant resources were allocated to support operations with large amounts of data (for example, data collection, warehousing, mining and cleaning, aggregation and integration, processing and interpretation) - all this leads to the high cost of storing and processing data [13]. Researchers say cost minimization is an emerging problem (Irani Z., 2010) [14], data processing costs and other data center operating costs are a problem that can also impact on how organizations implement technology solutions [15].

Thus, the carried out analysis suggests that it is possible to use Big data to develop directions for sustainable socio-economic development of territories at the present stage, but it

is necessary to ensure their proper ordering, storage and systematic collection, and also to improve the quality of Big data.

We proposed a methodology for assessing the directions of sustainable socio-economic development, taking into account the basic clusters. Approbation was carried out on Krasnoyarsk Region.

1. Defining base clusters

The object of the study was the Krasnoyarsk Region, which is characterized, on the one hand, by the richest natural resources and production and innovation potential, and on the other hand, low investment returns. A comparative analysis of the sensitivity of industrial production to investment in foreign economic activity showed the following (Table 1):

TABLE 1. THE SENSITIVITY OF INDUSTRIAL PRODUCTION TO INVESTMENT IN FOREIGN ECONOMIC ACTIVITY

Region	Coefficient of return on changes in investments in fixed capital of processing industry for the total volume of industrial production, rub./rub.	The coefficient of return from changes in investment in fixed capital for the production of the total industrial production, rub./rub.
Krasnoyarsk Region	0.21	0.27
Novosibirsk Region	16.34	6.47
Irkutsk Region	11.53	2.47

Calculations carried out according to the data on the Krasnoyarsk Region allowed to obtain the following results:

- base clusters identified: oil and gas and timber industry;
- potential clusters identified: cluster of building materials, food production and aluminum cluster.

The most stable are recognized food cluster, IPM and timber industry.

The experts of the Stakeholder Forum performed an analysis of the balance of the SDGs, which aims to determine the proportion of the three components of each goal, based on the corresponding tasks [16]. At the same time, experts note that individual goals are not well balanced within themselves: some goals have priority as economic, others as social, and others as environmental (Table 2).

TABLE 2. BALANCE OF SDGs BY SUSTAINABLE DEVELOPMENT COMPONENTS

SDG	Social area	Ecological area	Economical area
Poverty eradication	60%	13%	27%
Hunger eradication	60%	27%	13%
Health and well-being	96%	-	4%
Quality education	81%	5%	14%
Gender equality	100%	-	-
Clean water and sanitary	44%	44%	11%
Low-cost clean energy	22%	44%	33%
Decent work and economic growth	37%	10%	53%

Industrialization, Innovation and Infrastructure	13%	20%	27%
Lowering of inequality	67%	-	33%
Sustainable cities and towns	62%	33%	5%
Responsible consumption and production	21%	58%	21%
Fighting the climate change	56%	-	44%
Preservation of marine ecosystems	22%	67%	11%
Preservation of earth ecosystems	22%	67%	11%
Peace, justice, and effective Institutions	93%	-	7%

The analysis of statistical data on morbidity, mortality and life expectancy of the population of the Krasnoyarsk Region was carried out in order to assess the achievement of the SDGs in terms of social and environmental parameters. Moreover, among the limitations of the study, it is worth noting that the existing environmental monitoring system requires improvement aimed at completeness of data collection.

Thus, the development of directions for the sustainable socio-economic development of territories should be implemented taking into account the SDGs for the economic, social and environmental components. In order to assess implementation of the SDGs by component to be possible, a centralized organization of Big data collection is needed to carry out the calculations.

According to the results of the analysis conducted for the period from 2012 to 2016, the fall in the contribution of the Krasnoyarsk Region to the aggregate GRP was 2.2%.

In order to assess the achievement of the SDGs on the social development parameters, the Human Development Index (hereinafter - HDI) was assessed for the period from 2012 to 2016. According to the results of the assessment of the HDI in 2016 compared to 2012, there is an increase of 6%.

IV. DISCUSSION OF RESULTS

The analysis of using Big data to develop areas of sustainable socio-economic development of territories revealed a number of problems, the presence of which at the present stage does not allow full use of big data to assess the achievement of sustainable development goals in the formation of economic, social and environmental policies.

As a result of the study, it was found that the food cluster, IPM and timber industry are recognized as the most stable in the Krasnoyarsk Region. At the same time, it is necessary to create an infrastructure in the region that will allow competing successfully for investments and labor resources. The implementation of the cluster approach will contribute to the regional localization of social, economic and environmental effects of the activities of cluster members.

Thus, at the present stage, it is necessary to ensure the formulation of regional development objectives oriented towards the SDGs, while the objectives must be consistent with the regional development priorities, be real and achievable.

V. CONCLUSIONS (INFERENCE)

The authors of this article gave an assessment of the possibilities of using Big Data to develop areas of sustainable socio-economic development of territories, taking into account the implementation of the cluster approach. Based on the results of existing studies published in the Scopus science-metric database, the work analyzes, summarizes and presents the characteristics of Big data needed to support management decision making based on the use of Big data.

The work proposed a methodology for assessing the sustainability of the development of basic clusters in the region, consisting of four stages. The methodology is based on the use of the scenario approach (optimistic, pessimistic and average), as well as on the assessment of the impact of parameters based on indicators of sustainable development goals with their classification by economic, social and environmental unit. GRP or human development index is proposed as a response function.

In our opinion, ensuring the use of Big data for the development of directions for sustainable socio-economic development of territories should be provided by the following activities:

- to ensure the comparability of data from national statistical systems with international statistical systems in the field of SDGs;
- provide data for a specific indicator of the achievement of the SDGs with only one specific responsible agency;
- create an electronic data platform for reporting on SDGs.

The study was carried out within the framework of the competition of scientific projects of interdisciplinary fundamental research conducted by RFFI together with the Government of Krasnoyarsk Krai, in 2018 Project No. 18-410-242005 r _ mk "Development of a methodology for assessing the sustainable development of the territories of Krasnoyarsk Krai."

References

- [1] M. Polikarpova. Statistical analysis of the diversification of integration activity in the Russian economy // Young Scientist. 2013, vol. 10 (57). Pp. 377-379.
- [2] Innovations in data for development: UN materials. URL: <http://www.un.org/ru/sections/issues-depth/big-data-sustainable-development/index.html>.
- [3] Shah T., Rabhi F., Ray P. Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions. Cluster Computing, 18 (1) (2015), pp. 351-367.
- [4] Liao Z., Yin Q., Huang Y., Sheng L. Management and application of mobile big data. International Journal of Embedded Systems, 7 (1) (2014), pp. 63-70.
- [5] Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35 (2) (2015), pp. 137-144.
- [6] Barnaghi P., Sheth A., Henson C. From data to actionable knowledge: big data challenges in the web of things. IEEE Intelligent Systems, 28 (6) (2013), pp. 6-11.
- [7] Chen J., Chen Y., Du X., Li C., Lu J., Zhao S., Zhou X. Big data challenge: a data management perspective. Frontiers of Computer Science, 7 (2) (2013), pp. 157-164.
- [8] Labrinidis A., Jagadish H.V. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5 (12) (2012), pp. 2032-2033.
- [9] Akerkar R. Big data computing. CRC Press, Taylor & Francis Group, Florida, USA (2014).
- [10] Lu R., Zhu H., Liu X., Liu J.K., Shao J. Toward efficient and privacy-preserving computing in big data era. IEEE Network, 28 (4) (2014), pp. 46-50.
- [11] Wang Y., Wiebe V.J. Big Data Analytics on the characteristic equilibrium of collective opinions in social networks. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 8 (3) (2014), pp. 29-44.
- [12] Machanavajjhala A., Reiter J.P. Big privacy: protecting confidentiality in big data. XRDS: Crossroads. The ACM Magazine for Students, 19 (1) (2012), pp. 20-23.
- [13] Raghavendra R., Ranganathan P., Talwar V., Wang Z., Zhu X. No power struggles: coordinated multi-level power management for the data center. In ACM SIGARCH Computer Architecture News, 36 (1) (2008), pp. 48-59.
- [14] Irani Z. Investment evaluation within project management: an information systems perspective. Journal of the Operational Research Society, 61 (6) (2010), pp. 917-928.
- [15] Al Nuaimi E., Al Neyadi H., Mohamed N., Al-Jaroodi J. Applications of big data to smart cities. Journal of Internet Services and Applications, 6 (1) (2015), pp. 1-15.
- [16] "Sustainable development goals and integration: Achieving a better balance between the economic, social and environmental dimensions". URL: <http://www.stakeholderforum.org/index.php/our-publications-sp-1224407103/reports-in-our-publications/616-sdgs-integration-and-balance-of-three-dimensions>.