Study of the socio-economic development of the federal subjects of Russia on basis of Data Science Technology

Andrey Shevandrin
Volgograd State University,
Institute of Management and Regional Economics
Volgograd, Russia
shevandrin@mail.ru

Petr Bondarenko
Volgograd State University,
Institute of Management and Regional Economics
Volgograd, Russia
bondarenko.volsu@gmail.com

***Abstract* — The paper shows the possibility of conducting a study of the regional economy based on alternative sources of official statistics. Official statistical resources in selected studies cannot be used or do not meet the requirements of relevance, reliability, and completeness. As new data resources, live data on the Internet should be considered. Two approaches to obtaining data on economic processes from the Internet can be distinguished: through a specially created program interface of websites (API) and as a result of scanning the contents of web pages (web-mining). Based on the open dataset published by PJSC "Sberbank", a number of features of social and economic processes in the regions related to the formation and use of household incomes have been established. Methods of time series analysis, variance analysis, multidimensional classification and factor analysis are applied. As a result, the seasonality and rhythm of the time series of salary and the amount of pension are established; the reduction of differentiation of regions according to the incomes of the population was revealed, with their insignificant growth; It is shown that the income indicators of the population are not determinative in financial activity; for relatively "poor" regions is characterized by a savings strategy of the households, "rich" - credit strategy.**

Keywords* — *data science, big data, data minig, neural network, Kohonen self-organizing maps

I. Introduction

The traditional cycle of research in the regional economy involves the formation of a hypothesis of research or a theoretical model, the choice of methods for analyzing its verification, the collection, and analysis of the data obtained, the final conclusion with regard to verification of the assumptions made [1, 2, 3]. In economic research, until recently, reliable analysis, diagnostics, and hypothesis verification could be based only on the final annual (for selected indicators of quarterly) data included in the published information resources of Rosstat. The data of official statistical bulletins in the context of accelerating the dynamics of economic processes are characterized by the following significant shortcomings: low level of relevance (data are published with a delay of 3 months to 2 years), distortion due to the quality of the primary statistical material (enterprises tend to underestimate their official reporting, statistics); conservative composition of statistical indicators (new economic phenomena and processes are not observed by official statistics); different degree of depth of detail and analytical slices of indicators (does not allow for comparative studies); fragmentary time series; refinement of previously published values (makes conclusions of researchers initially not reliable).

Ensuring reliable selective research is associated with high costs, which research teams, even taking into account grant support, objectively cannot bear.

The development of information technologies, the introduction of methods of artificial intelligence and machine learning, the need to search for heuristic solutions in the modern creative economy led to the spread of new approaches to obtaining a factual basis in economic research. Conventionally, we can distinguish two alternative official statistics and selective studies of the data extraction approach - obtaining "live" data from the Internet (Web Mining) and large data (Big Data).

Modern resources of the Internet contain huge arrays of poorly structured information. Since the advent of Web 2.0 technology, according to which users themselves create content on Internet sites, the worldwide network increasingly reflects interests, preferences, sensitivity to the properties of various goods and the characteristics of their users. Effective methods of Web Mining are:

- obtaining data from Internet sites through a specially created software interface (API). Such interfaces have the majority of social networks and data banks (citation systems, ad placements, registries, etc.);

- "parsing" sites, i.e. the analysis of texts of pages of Internet sites, special programs-robots with the purpose of extracting information according to pre-established rules. "Parsing sites", as a rule, apply in the absence of API-interfaces; the most interesting for parsing are sites of online stores, job banks, exchanges and organizations.

Obviously, the disadvantage of Web Mining is the need for the researcher to have the relevant competences in the field of software engineering or to involve the relevant specialist in the research.

Big Data technology involves processing a significant amount of data recorded by information systems and analog-to-digital converters, or obtained from processing unstructured data (images, video, audio data). Unlike the "parsing" of sites, the Big Data array is not formed for a specific query and is designed to search for new (not obvious) patterns in socio-economic phenomena and processes. In this regard, large owners of such data sets (corporations, national governments) are interested in providing researchers with access to their data banks or publish them in aggregate form. The most popular open data projects include Data.gov (US Government, http://data.gov), US Census Bureau (US Census Bureau portal, http://www.census.gov/data.html), European Union Open Data Porta (Open Data Portal of the European Union, http://open-data.europa.eu/en/data), Government of the United Kingdom Open Data Portal (http://data.gov.uk/), Amazon Web Services public datasets (Amazon Internet store datasets, http://aws.amazon.com/datasets), UnData (UN Open Data Portal, http://data.un.org/Default.aspx), DBPedia (Database Publishing Service , http://wiki.dbpedia.org/).

The most popular Russian public Internet resources include the Open Data Portal of the Russian Federation (http://data.gov.ru/), the open data portals of the subjects of the Russian Federation (for example, the official portal of the open data of the Volgograd Region http://opendata.volganet.ru), open data of the Savings Bank of the Russian Federation (http://www.sberbank.com/en/analytics/opendata).

The latter is of particular value for research in the regional economy, since it offers a set of aggregated monthly data for all regions of Russia.

The purpose of this study is to find trends and features that are significant for the economy of Russia's regions on the basis of aggregated data on the activities of the largest financial institution in the country - PJSC "Sberbank".

II. Materials and Methods (Model)

An array of data published by PJSC "Sberbank" and intended for free use as of January 2018 contains more than 45.5 thousand records for 83 regions of the Russian Federation for the period from January 2014 to October 2017, inclusive. The set includes the following indicators:

- The number of applications for consumer loans;
- The average amount of an application for a consumer loan;
- The number of applications for mortgage loans;
- The average amount of an application for a mortgage;
- The number of new deposits;
- The average amount of a new deposit;
- The average salary;
- The average pension;
- The average rub. on the current account per person;
- The average deposits in rubles. per person;
- The average spending on cards.

PJSC "Sberbank" is the largest financial institution of the country, according to the Frank Research Group as of 01.07.2017, the share of private lending is 38.8%, in the credit card market - 40.5%, in the current account and term deposits market - 44, 2%, mortgage lending - more than 50% [4], having offices in all constituent entities of the Russian Federation, a bank of these indicators of its activities can be qualified as representative for the study of the economy of Russian regions.

Based on the available data set, it seems appropriate to carry out the analysis in the following areas:

1. Having a monthly data array, it is relevant to analyze the seasonality of a number of income indicators of the population, since official statistics do not publish such data;

2. The data in the context of the regions of the Russian Federation will make it possible to draw a number of conclusions on the differentiation of subjects in terms of income indicators of the population.

3. It is of interest to analyze certain aspects of the social and economic well-being of the population, which, according to official statistics, is difficult to ascertain.

4. Expediency classification of regions according to a set of indicators that will allow to establish not only differences, but also to identify common groups of regions.

III. Results and Discussion

1. The dynamics of the average wage in the RF as a whole, separately for the Volgograd region and the ratio of salary and transferred to the accounts of PJSC "Sberbank" pensions are shown in Figure 1.

According to the diagram above, the observed seasonality and rhythmicity of time series by average wage should be noted: traditionally the largest payments fall in December (the seasonality ratio is 1.6), less significant growth is observed in July (seasonality ratio 1.1), and the minimum payments are made in the first months of the year (seasonality ratio ≈ 0.7).

In the ratio of salary and pensions, using the linear approximation, the regression coefficient will have a negative sign, i.e. on the observed period; there is an increase in the gap between the average pension received and salary.

2. The study of the differentiation of the regions of the Russian Federation in terms of these indicators was carried out by means of generally accepted statistical indicators for estimating the regional inequality: mean value, variation range, coefficient of variation, asymmetry and kurtosis coefficients. The values obtained are presented in Table I.

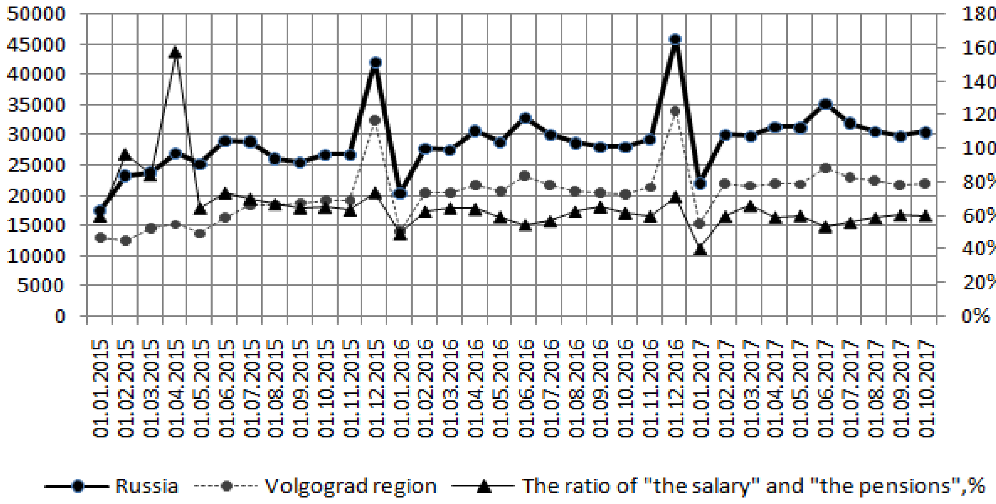

Fig. 1. Dynamics of the average salary transferred to the accounts of PJSC "Sberbank"

The greatest differentiation of the regions of the Russian Federation is observed in terms of the average volume of deposits per person, since 2015 the relative range of variation here has increased by 46.2%. It should be assumed that the population of the regions has different propensities to savings, which will be considered separately.

By the level of salary, pensions received and the average application for consumer credit, regions show a reduction in heterogeneity. In 2017, the smallest transferred average pension in the Tula region is 8587 rubles; the largest in the Kamchatka Territory is 22588 rubles. The lowest average wage is observed in the Republics of Karachaevo-Cherkessia and Kalmykia, Ivanovo and Kostroma regions. The highest salary are typical for the Chukotka Autonomous District (72022 rubles), the Magadan Region (67697 rubles), the Kamchatka Territory (60552 rubles) and Moscow (55661 rubles). For these indicators, if the kurtosis index is not high, asymmetry has a positive value, and therefore the distribution of clients' income of PJSC "Sberbank" is still shifted to the right. Thus, on this set of data, the hypothesis of increasing differentiation of regions according to the incomes of the population in the observed period is not observed.

3. Analysis of hidden relationships between the indicators considered is carried out through factor analysis (main components method).

As a result, 2 factors have been identified, which can explain 85% of the variation in the initial values. The first factor includes the variable salary and the size of pensions, i.e. this is the income factor of the population, it gives only 16% of the explained variance. The second factor consists of variable-value applications for consumer and mortgage loans, a new deposit and card expenses - 69% of the explained variance. Thus, salary (together with the value of the average pension) are not determinative in the financial activity of the population.

4. To analyze the data array and solve the problem of classification of regions by the set of indicators of PJSC "Sberbank", a multidimensional statistical grouping based on the application of the Kohonen neural network was applied. This method allows you to identify latent rules and patterns in the data set.

TABLE I. STATISTICAL DATA FOR THE EVALUATION OF DIFFERENTIATION FOR INDICATORS ON SERVICES RECEIVED IN PJSC "SBERBANK" ON JANUARY 01, 2015 AND OCTOBER 15, 2017

The coefficients	**The average salary**		**The amount of the pension**		**The average volume of deposits in rub. per person**		**The average amount of an application for a consumer loan in rub.**	
Period	01.2015	10.2017	01.2015	10.2017	01.2015	10.2017	01.2015	10.2017
Average value	18001,0	28221,1	12213,9	14508,5	335322,0	445107,4	192907,6	247657,2
Relative range of variation	311,3%	124,2%	253,6%	196,1%	409,2%	455,4%	166,0%	144,7%
Excess	7,8	3,8	4,4	4,5	21,1	23,5	4,3	3,3
Asymmetry	2,4	1,6	1,9	2,1	4,1	4,3	2,0	1,9

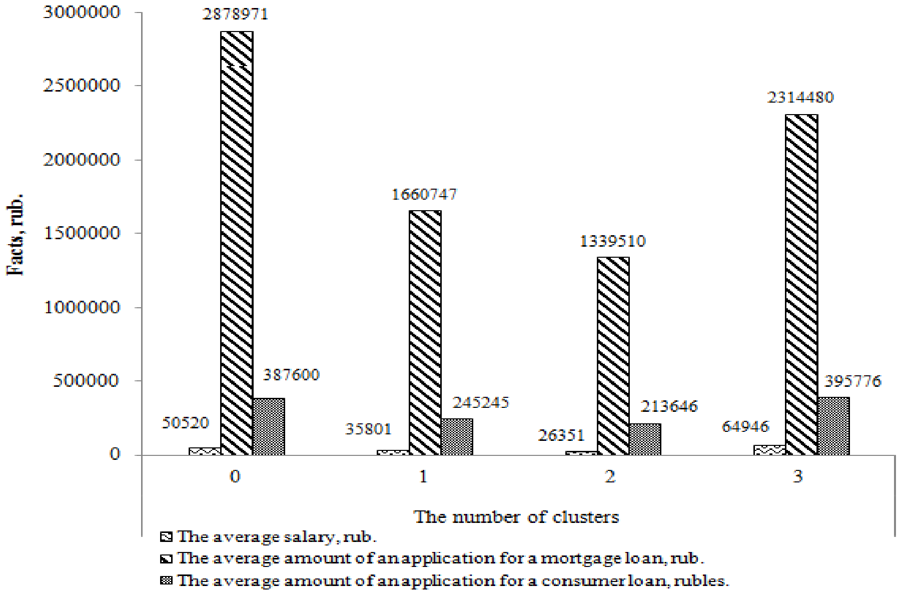

Fig. 2. The diagram of clusters in the context of average values of the population's wage and the loan amount.

TABLE II. THE CLUSTER ANALYSIS OF THE LEVEL OF REGISTRATION OF REGIONS OF RUSSIA

Cluster number	Level of crediting
2	Low (55 regions)
1	Below the average (16 regions)
3	Abow the average (9 regions)
0	High (3 regions)

As a result of processing the data array, Kohonen self-organizing maps were built, whereby the regions of Russia were combined into clusters, each of which characterizes the extent to which the savings, investment or credit strategy of the population is implemented (Table II).

The credit strategy is typical for the population of the cities of St. Petersburg, Moscow and the Moscow Region, which are classified in the zero cluster. The Volgograd Region is included in the second cluster with a low level of applications for consumer and mortgage loans, which indicates the prevalence of a savings strategy with low salary relative to other regions.

Considering the profiles of clusters in terms of average salary, the amount of the application for mortgage and consumer loans, (Fig. 2), it can be noted that the most "secured" regions form the main portfolio of mortgage loans, in regions with lower incomes, the share of consumer loans in the aggregate portfolio of applications.

Analyzing the profiles of clusters in the context of the average values of the turnover of funds on bank cards and the average size of the new deposit, it can be seen that the population of regions classified as a cluster with a propensity for credit behavior is actively using bank cards.

IV. Conclusion

Studies of regularities and features of the economy of the regions can now be performed not only on the basis of an array of official statistical information and sample studies, but also on the basis of data collected or published on the Internet. The need to search for heuristic solutions in management, as well as using the capabilities of artificial intelligence and machine learning leads to the formation of an open access to a new class of information resources obtained by information owners on the basis of aggregation of large data.

In this study, based on the open data set of PJSC "Sberbank", certain features of the socio-economic status of the regions have been identified, which it is difficult to identify on official open statistics, including: the seasonality and rhythm of time series of salary and pensions have been established; the reduction of differentiation of regions according to the incomes of the population was revealed, with their insignificant growth; it is shown that salary (together with the value of the average pension) are not decisive in the financial activity of the population; for the relatively "poor" regions, the savings strategy of the population is characteristic, the "rich" - the credit strategy; the higher the incomes of the population in the region, the more demanded mortgage lending.

Expansion of the indicators, which is announced on the website "Open data of Sberbank", will allow solving other analytical problems of the regional economy, and the addition of time series will make them suitable for forecasting, which will make such information resource for researchers as valuable as official ones statistical bulletins.

References

[1] A. Borisova, A. Kalinina, M. Buyanova “The mechanism for detecting and controlling regional socio-economic risks” 3rd International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2016, SGEM2016 Conference Proceedings, Book 2 Vol. 3, pp.1127-1134, 24 - 31 August 2016, DOI: 10.5593/SGEMSOCIAL2016/B23/S07.142

[2] Petrova E., Tarakanov V., Kalinina A. Methodological Toolset of Regional Authority Bodies’ Cluster Policy // Russia and the European Union. Development and Perspectives // Contributions to Economics. 2017. - pp. 55 – 63

[3] Petrova E., Kalinina A., Buyanova M. Efficiency of Public Administration and Economic Growth in Russia: Emperical Analysis // European Research Studies Journal. - 2015. - Vol. XVIII, Issue 4, Special Issue. – p. 73-86

[4] The biggest players in the retail banking market [Krupneishie igroki rynka roznichnykh bankovskikh uslug] [Electronic resource]. URL: : https://frankrg.com/index.php?new_div_id=145