**Supplementary Notebook (RTEP - Brazilian academic journal, ISSN 2316-1493)**

# SPATIAL GEOMARKETING POWERED BY BIG DATA

S.V. Shaytura [1]
V.M. Feoktistova [2]
A.M. Minitaeva [3]
L.A. Olenev [4]
V.O. Chulkov [5]
Y.P. Kozhaev [6]

[1] *Russian University of Transport, Russia*
[2] *Russian State University of Tourism and Service, Russia*
[3] *Bauman Moscow State Technical University, Russia*
[4] *Plekhanov Russian University of Economics, Russia*
[5] *Moscow State University of Civil Engineering (National Research University), Russia*
[6] *Russian State University of Physical Education, Sport, Youth and Tourism, Russia*

***Abstract:*** *New approaches to processing and visualization are required due to the accumulation of a large amount of geospatial data. One of these approaches is creating a geomarketing system with a fundamentally new toolkit based on data clustering. Examples of the capabilities of such a system are assessing the cost of housing, locating a new shopping center, a bank branch, and railway workers examination points. The article describes the technology used for assessing the cost of housing. This technology implements a fuzzy logic assessment based on expert rules. A site for a bank has been searched for with the use of geospatial analysis of rectangular clusters. A site for an outpatient clinic has been searched for by going through the clusters located along railways.*

***Keywords****: spatial geomarketing, geomarketing system, spatial data, geomarketing, big data, clusters, geomarketing technologies, information technologies.*

## INTRODUCTION

The digitalization of the national economy and the rapid development of the global information networks have resulted in the emergence of the big data concept (Tsvetkov, Shaytura, Sultaeva, 2020; Buravtsev, Tsvetkov, 2019; Dyshlsko, 2015). Most of the data that companies work with are georeferenced. For example, sales to customers are determined by the geographic location of the target groups. All clients have several addresses (residence, registration, work, study, etc.) and move quite logically between them. Revealing these patterns allows "learning" and scaling the algorithms. Geodata

improve the quality of the decisions made and the effectiveness of the actions related to the current and potential markets, customers, media planning, customer attraction, and sales growth.

If the data are related to spacetime, one speaks of spatially distributed or geospatial data. The source of big geospatial data is the Earth remote sensing and space monitoring data, the Internet, and nondigital information of various kinds (Shaytura et al., 2018a; Markelov, Tsvetkov, 2015; Chumachenko et al., 2017; Shaytura, Vaskina, 2019; 2020). The big data concept includes not only large amounts of digital data but also the means of working with these data. The means of processing big geospatial data are, first of all, the geographic information system and the methods of geospatial analysis: crowdsourcing, classification, aggregation, and integration of data of various kinds, predictive analytics, machine learning, pattern recognition, and analytical data visualization (Bayandurova, Rosenberg, Shaytura, 2016; Rosenberg, Shaytura, 2016; Kureichik, Kazharov, 2013; Shaytura et al., 2016a; Petrov et al., 2019; Kulagin, Tsvetkov, 2013).

There are many tasks for geospatial analysis. These tasks include searching for the locations of various places, addresses, and coordinates, and vice versa, the objects with certain coordinates on site (Shaytura et al., 2018b; Shaytura et al., 2016b; 2019; Tymchenko, 2014; Mayorov, 2014; Sharifyanov, 2017). Spatial geoportals such as Yandex, Google, Bing, and others are engaged in these tasks. There are also many specialized geoportals: global, regional, and municipal ones, which solve the problems of geospatial analysis in certain territories. With that, the crowdsourcing method is often used for clarifying and enriching the digital content of the maps by involving large numbers of users for creating and generalizing the content. Thus, Solaris, a kind of computer brain is created, which contains huge amounts of data. Using these data, one can solve many geomarketing tasks, for example, locating a new shopping center, a bank branch, an outpatient clinic, etc. (Gavrilova, Shaytura, 2012; Kitova, Shaytura, 2016; Maratkanova, 2018; Vorobyova, Degteva, 2018).

## METHODS

The studies described in the article were based on cluster analysis, geoinformation systems and technologies, and geomarketing analysis.

### Geomarketing studies

Geomarketing information systems (GMISs) appeared due to the integration of geographic information systems and marketing information systems (Petrov et al., 2019; Gerasimenko, Tkhorikov, Naplekova, 2020; Tsvetkov et al., 2020). The use of geomarketing systems and technologies is advisable where there is a need to process spatially localized data, or where it is necessary to use thematic maps with business graphics for the substantiation of decision-making. Spatial localization may be coarse or precise. GMISs allow applying visual methods of statistical data representation and processing for the substantiation of decision-making. These should include the possibility of the generalization of various qualitative phenomena and characteristics. Visual data analysis is two to three orders of magnitude faster than tabular data analysis, especially

in monitoring critical or anomalous situations. The listed additional processing capabilities determine the effectiveness of geomarketing as a market information technology, especially in analyzing spatially distributed market characteristics, such as political situation, demographic situation, economic situation, transport networks, tourist routes, etc.

Geomarketing is a marketing research technique for making strategic, conceptual, and managerial decisions based on the methods of geographical analysis of various spatially distributed objects and phenomena. Such studies make it possible to determine the target audience in certain territorial units, making competitive analyses, determining the best location for a new facility, predicting the commercial real estate turnover, developing concepts for the existing or planned facilities, finding the best use for a land plot, and much more. Geomarketing studies allow analyzing external and internal geospatial (georeferenced) indicators of companies, various aspects of their past, current, and future activities, including their infrastructure and competitive environments. Since in geomarketing studies, the customer has to provide various company data (for example, the revenue for several years, the employee turnover, etc.), the customer and the contractor are in close contact during the entire process of the studies. This is an undoubted advantage of geomarketing since all sorts of specialists and experts that businessmen turn to for help do their job and simply pass the working mechanism over to the customer.

**Coordinate referencing and data clustering**

Since the data are coordinate-referenced (and this is the main thing they have in common), it is advisable to use the coordinates as the main axes of a multidimensional data cube. At the same time, in the qualitative analysis, the accuracy of referencing often varies, since averaged data are referenced to a certain territory, rather than to a certain point. In the tasks where precise data georeferencing is not required, models of clustered raster data are used. This method allows processing the data with predetermined precision, thus avoiding the unnecessary use of computing capacities. For example, a Landsat satellite image of the Earth has 74,000,000 raster cells. Processing this amount of data is a very laborious task. Clustering (or cluster analysis) is a machine-based data processing method aimed at separating the data into several groups according to certain characteristics. The purpose of cluster data analysis is to separate the groups with similar characteristics and to distribute them across clusters. Currently, the relevance of such a process is determined by the large flow of the data to be properly structured for further processing.

To clearly show the differences between the types of clustering, let us consider the example of store X. Let us assume that the director of store X is interested in analyzing the customer preferences for expanding the business. It is impossible to know each customer's needs and develop a unique business strategy for each customer. Using statistical analysis, it is possible to combine all customers, for example, into 10 groups, based on their purchasing habits. Next, individual strategies may be developed for the clients in these 10 groups. Two types of clustering are identified:

*Hard clustering.* In hard clustering, each data point is either fully owned by the cluster, or not owned. In the example above, each customer is allocated to one of 10 groups.

*Soft clustering.* Soft clustering determines the likelihood that the data point is present in these clusters, instead of placing each data point in a separate cluster. For

example, in the above scenario, each customer is assigned the probability of being present in any of the 10 clusters of store X.

There are many tools for solving clustering goals. Each method has a different set of rules for determining the data points "similarity". Currently, more than 100 clustering algorithms are known. Hierarchical clustering is an algorithm that builds a hierarchy of clusters. This algorithm starts with all the data points assigned to their cluster. Next, the two nearest clusters are merged. The algorithm comes to a logical conclusion when there is only one cluster left. The results of hierarchical clustering may be shown in a tree diagram (in the form of a tree built on the proximity matrix). The tree diagram (Fig. 1) may be illustrated as follows (Fig. 1):

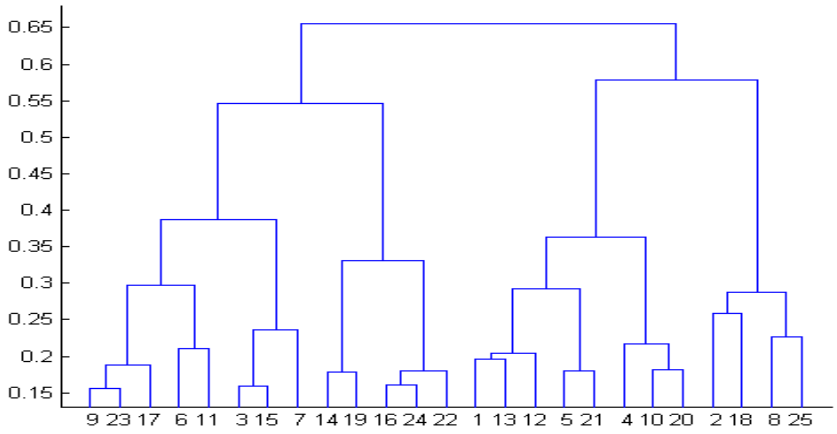


**Figure 1.** The tree diagram of hierarchical clustering

In the quadrotomic representation, the entire surface of the Earth is represented in the form of a tree-like structure with each area divided into four squares. This data representation was used for making the standard protocol for map fragments preliminarily created and processed in real-time, and georeferenced via the Internet (Fig. 2).

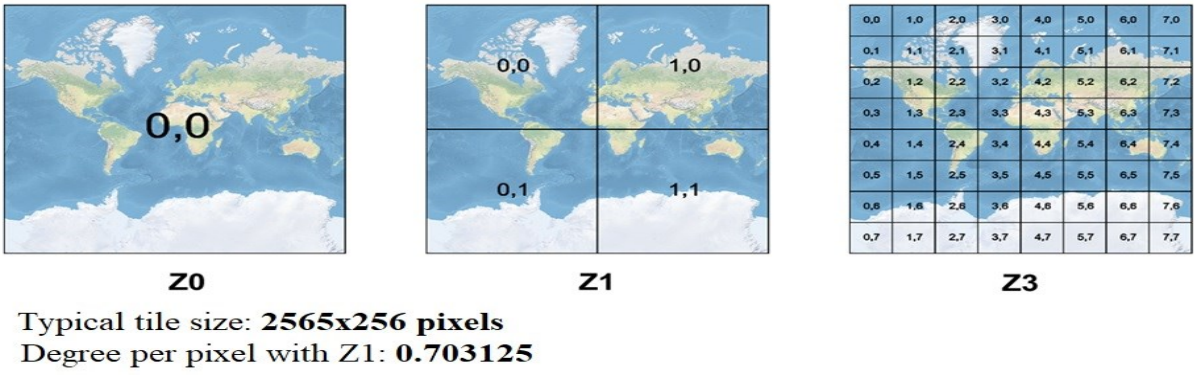


**Figure 2.** Tile forming

The cluster approach allows using the geospatial data with varying degrees of the accuracy of the spatial coordinates.

## RESULTS

### Assessing the residence attractiveness based on the geomarketing systems

In assessing the residence attractiveness, several factors are to be considered (Bayandurova, Rosenberg, Shaytura, 2016; Rosenberg, Shaytura, 2016; Gavrilova, Shaytura, 2012). The general rule is obvious: the consumer properties of the real estate are consistently lost with time, which decreases its usability, i.e., decreases its value. Separating the region into clusters allows forming an intelligent evaluation model based on the neural network methods (Bayandurova, Rosenberg, Shaytura, 2016). This model implements the process of assessing the clusters in the form of a multilayer perceptron. Based on the initial data, a neural network model is formed, which generates Kohonen maps. The previously obtained results were visualized using the capabilities of the geographic information system (GIS). The concept of assessing the clusters included the following stages: collecting the initial data, creating a mathematical model for real estate valuation, assessing the model parameters based on the nonparametric statistical methods, formulating the problem for the neural network basis, and analyzing the parameters of the mathematical model for real estate valuation based on the results of the neural network learning. The proposed approach to assessing urban clusters makes it possible to expand the range of assessment tasks to be solved. In this case, subjectivity is considered in the mathematical model, thereby the calculation accuracy increases, and the cost of adjustment with time reduces.

### Predicting the financial performance of premium clothes stores

One of the first tasks in processing geospatial data is geocoding. Coordinates are added to the target variable. Next, one has to bind the data to each square (cluster). With the help of machine learning, one can predict the profit in each square, i.e., the profit of a store placed in a certain square. Next, one has to draw the generalized portrait of potential buyers with the income above average. It is necessary to understand where they live; for this purpose, one has to open the geographic information system and enable the required layer. Rich people live in locations with the most expensive real estate. Next, one has to understand whether they will or will not go to the planned store. For this purpose, the Huff model is used.

Consider a single store and a single point. What is the likelihood of a person traveling from this point (square) to the other end of the city? The number of expensive clothing stores in this square is taken for the weight. The people from this square are distributed across the points around the store. The result is that with the probability of 10 % he will go to store D, with the probability of 8% he will go to store C, and for all the stores likewise. In the square, we have the percent of potential buyers, which may bring profit in the future. Take the most expensive 20 % of the real estate and instead of stating the number of people allocated to the store, write the number of people multiplied by the cost of a square meter. The result is some kind of the people-and-money equivalent. A rich person who comes to the store will bring more profit than two poor persons.

As a result of machine learning, a model is formed that is used for this task. The customer presents the initial information in the form of a table with columns: address, area, and sales by year. Next, the data are geocoded, i.e., latitude and longitude are obtained. The next step is defining the goal function. From the sales volume over the

previous period, one can find the amount the store can ideally sell. For example, store number 0 has earned a maximum of 964.886 Euros over a year. However, stores have different areas, and it is incorrect to compare a 30-square-meter store to a 300-square-meter store. By dividing the profit by the area, one can get the profit per square meter. Let us calculate the median value for the sample. After that, the goal function is the amount of money per square meter. A regression model is built. It is a model that should receive the number of certain objects for any point as the input value. The profit for all the squares in Moscow is determined. The profit table is obtained. The results are placed on the map (Fig. 3).


**Figure 3.** A map with predicted financial performance for premium clothing stores.

## Choosing the optimal solution for placing a bank branch

Base layers are created with the initial value for each variable displayed for each cell (exactly in this 100x100 meter cell). For example, the Population cell contains a digit that corresponds to the number of people living in this square (Fig. 4). These layers include external data: welfare, population, traffic, bus stops, routes, and competitors of the bank. After loading, the display of each grid may be customized, color added, the gradient changed, text added, and the changes saved.

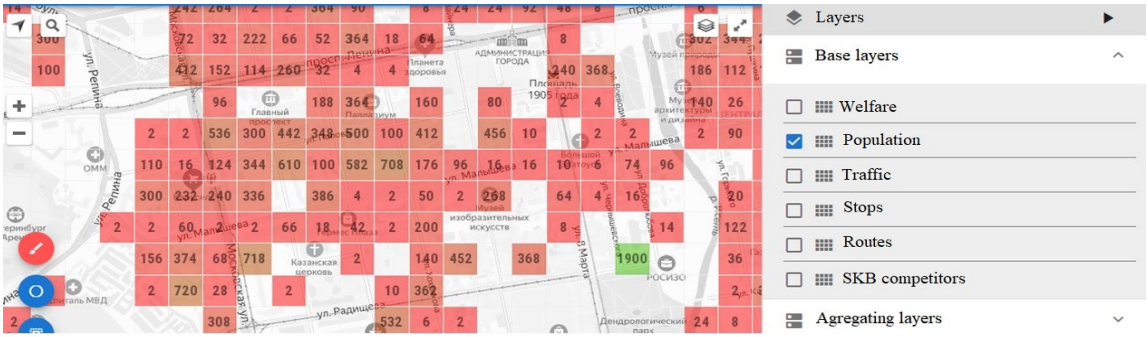


**Figure 4.** The base layer — Population.

Aggregate layers are created, in which one cell contains a digit equal to the sum of the cells of the base layer in the corresponding radius. For example, Population 1,000 — the cell contains the digit that is equal to the number of people living within 1,000 meters from this square (Fig. 5). For each layer, the distance is set: Population 1,000; Welfare 0; Traffic 300; Stops 500; Routes 500, and Competitors of the bank 500.

**Figure 5.** The Population 1,000 aggregate layer

Binary relative layers are created, in which the score is indicated following the adopted expert rule. An Expert Score layer is created that shows the sum of the binary layers reduced to a positive range (Fig. 6). In the case considered, it is a number from 0 to 7 (the minimum amount is -2, the maximum amount is 5, recalculated to the 0 – 7 score scale).


**Figure 6.** The Expert Score layer

A Normalized Expert Score layer is created (with values from 0 to 50), which is a variable that aggregates the information for the six expert rules (selected by the contractor based on their experience) embedded in the model (Fig. 7).

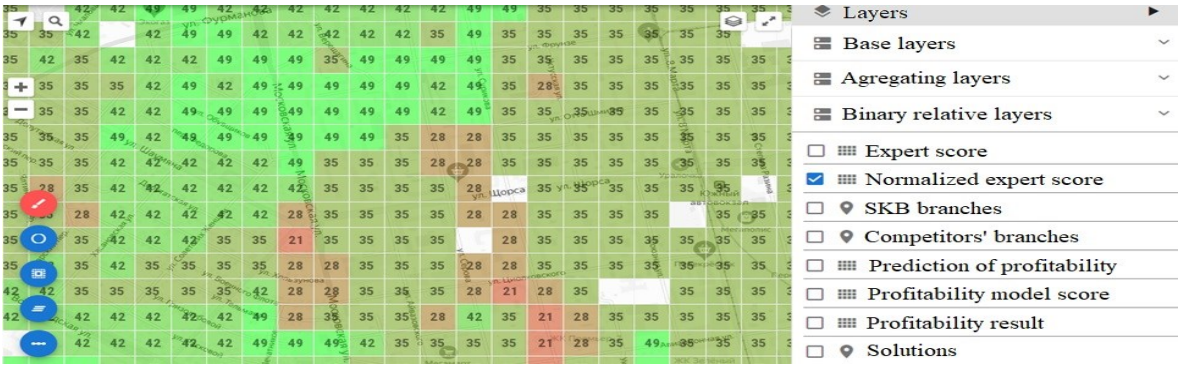


**Figure 7.** The Normalized Expert Score layer

A Bank Branches label layer is created, in which the data are displayed on the map as the characteristics of each label (Fig. 8). Each label has mandatory attributes: the

address and geocoordinates, the icon (to be chosen from Google Material Design Icons), and the color (blue by default).


**Figure 8.** The Bank Branches label layer

Similarly, a layer for competitor bank branches is created. A Profitability Prediction Layer is created for each cell, the office profitability value predicted by the model (machine learning) is displayed if it is present at this point (similar to the financial performance before localization) (Fig. 9).

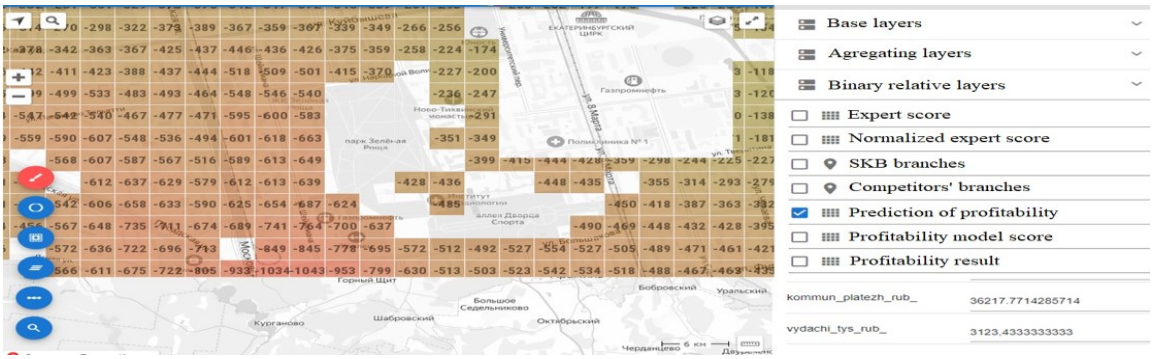


**Figure 9.** A Profitability Prediction Layer

A Profitability Model Score layer is created, which is the result of normalization (reduction to a comparable scale from 0 to 50) of the Profitability Prediction variable, which is obtained by machine modeling of the profitability indicator for each office based on the external factors (welfare, population, shopping centers within a radius of 2,000 m, parking lots within a radius of 2,000 m) (Fig. 10). Based on the analysis of a large data array – about 850 variables (traffic, routes, points of attraction, population, competitors, etc.), 13 factors have been selected that allow predicting the profitability of an office in a particular location. Assessment of profitability for the model is the calculated analog of the financial performance indicator before localization. Next, for comparison, the indicator is converted (normalized) into a scale of 0 – 50 scores. For example, for a bank branch, a prediction of the profitability of -828 has been obtained based on modeling, or 24 scores after normalization.

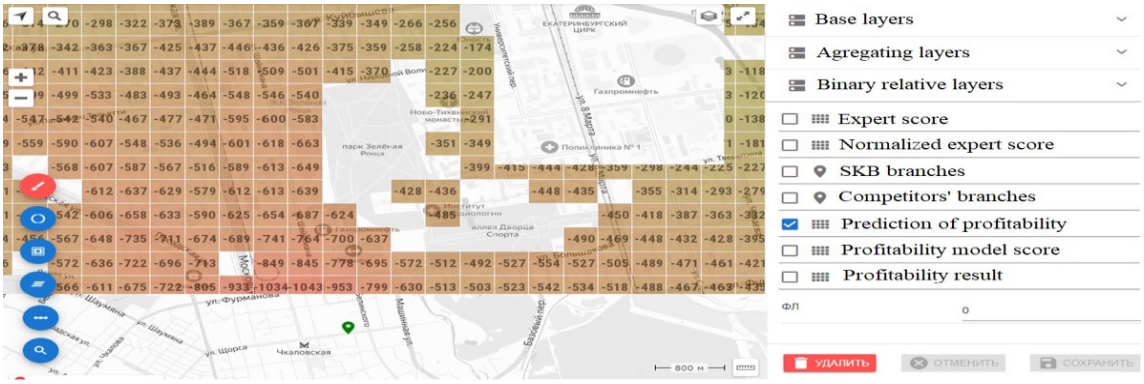


**Figure 10.** The Profitability Model Score layer

A Profitability Result layer is created (with the values from 0 to 100), which is the sum of scores for the Normalized Expert Score and the Profitability Model Score variables. It is the variable that equally considers the result of expert assessments and machine modeling. The success of a location is assessed from the standpoint of the probability of achieving high profitability (Fig. 11).


**Figure 11.** The Profitability Result layer

The Solution layer is the result of analyzing the office network of the bank (the assessment of the potential of each office, given the prerequisites for the location). It is displayed on the map as a colored marker according to the following rules. Using the geographic information system, potential locations for the bank branches are displayed on the map (Fig. 12).

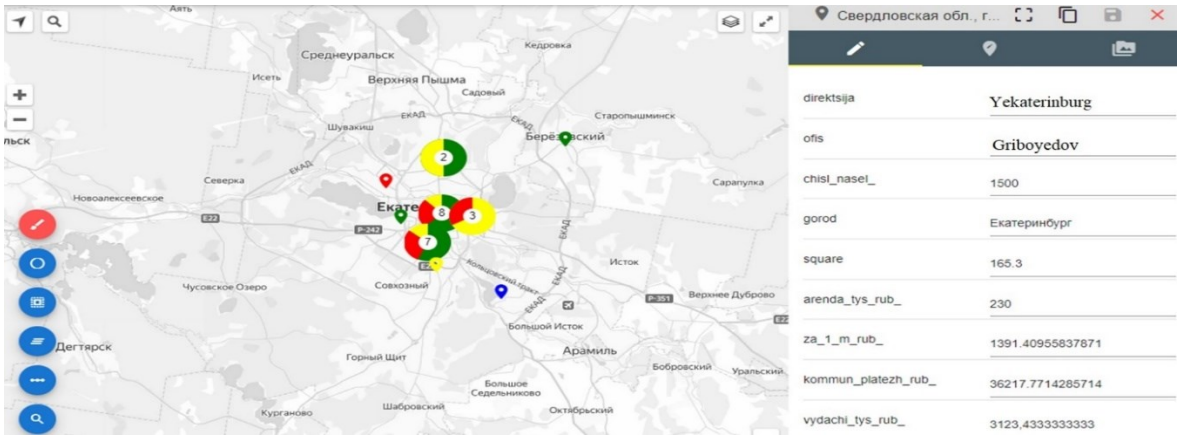


**Figure 12.** Potential locations for the bank branches

As a result, combined methods of geomarketing have been tested, which allow assessing a bank branch location and solving the problem of placing new branches. Promising areas for bank branches are the territories with high population densities, both living and working in the territory, with many organizations and sales-and-entertainment facilities. The use of this technique contributes to the competent organization of a banking network.

**Solving the logistical problems in medical care for the employees**

The healthcare system is a complex interweaving of practicing specialists, hospitals, clinics, employees, companies engaged in medical care, diagnostics, and treatment. Planning the development of such a system is not an easy task, since one of the main problems is a large amount of unstructured data. For optimizing the business processes of passing a medical examination, the GMIS has been used that was developed specifically for the employees of the Gorky Railway. With the help of GMIS, such tasks as affordable passing medical examination in time, increasing the safety of railway traffic, and careful monitoring of the health of the employees whose work is associated with harmful factors are solved. Using an atlas, users can quickly see the deadlines for the medical examination and the nearest medical institution for the delivery of medical tests.

The project provides a variety of opportunities for studying the data with the use of dynamic tables, charts, and algorithms. This system contains a large set of medical institutions for passing the medical examination and the information about the employees of the Gorky Railway (place of residence; position). The website logically organizes about 40 thousand employees and provides information about them. To improve the efficiency of passing a medical examination, the employees are distributed across the clusters and assigned to a specific clinic. The purpose of hierarchical clustering in the GMIS project is to bring together the employees who are located near each other based on the principle of passing the medical examination at the same time. All employees are separated into nine groups; new clusters are created within each of them according to the time of examination for choosing the optimal date for a particular group to pass the medical examination. In each group, groups of people living not far from each other are found with the use of agglomeration clustering.

## DISCUSSION

The article discusses several examples of using a geomarketing system for solving various problems in economics. Geospatial data clustering allows efficiently making calculations given the accuracy of spatial localization. Geospatial analysis may be applied in many business aspects. It will help analyze the outflow of customers to the competitors, optimize the number of hall managers for effective customer service, determine the workload of transport hubs, determine the need for the construction of a skyscraper or a park, etc.

## CONCLUSION

The uniqueness of the method in the base of geospatial analysis is that it allows, through cluster analysis of indicators, to obtain the information about the routes of people and the movement of material and information flows without using and processing personal data. Geospatial analysis has several advantages over the other data processing methods: the ability to provide continuous coverage of the population residing in the territory; the ability to consider the uneven distribution of the population across the territories of residence; providing data for any time intervals; prompt collection and provision of up-to-date data about the population and the dynamics of its movements across the territories, and economic efficiency. Geospatial analysis allows commercial enterprises and government agencies to assess the efficiency of doing business, the potential of the territories from the standpoint of the existence of social infrastructure (hospitals, schools, kindergartens, cultural institutions), traffic (commuting, traffic conditions, public transport), and payback rate (relevant for hotels, cafes, restaurants, and places of recreation).

## REFERENCES

1 Bayandurova, A.A., Rosenberg, I.N., Shaytura, S.V. (2016). Complex analysis of the Crimean tourist destinations. Scientific notes of the Crimean Federal University named after V.I. Vernadsky. Economics and Management, 2(68), 3-10.

2 Buravtsev, A.V., Tsvetkov, V.Ya. (2019). Cloud computing for large geospatial data. Information and Space, 3, 110-115.

3 Chumachenko, S.I., Knyazeva, M.D., Mitrofanov, E.M., Shaytura, S.V. (2017). Space Monitoring: study guide. Burgas.

4 Dyshlsko, S.G. (2015). Big Data in Earth Sciences. Slavic Forum, 3(10), 88-96.

5 Gavrilova, V.V., Shaytura, S.V. (2012). Intellectual processing of information in the field of real estate valuation. Slavyanskiy forum, 1(1), 164 - 171.

6 Gerasimenko, O.A., Tkhorikov, B.A., Naplekova, Yu.A. (2020). Essential representation, role, evolutionary stages and approaches of geomarketing. Bulletin of Belgorod University of Cooperation, Economics and Law, 3(82), 248-259.

7 Kitova, O.V., Shaytura, S.V. (2016). Information Marketing: study guide. Burgas.

8 Kulagin, V.P., Tsvetkov, V.Ya. (2013). Geology: representation and linguistic aspects. Information technologies, 12, 2-9.

9 Kureichik, V.M., Kazharov, A.A. (2013). Analysis and state of the problem of routing vehicles. Bulletin of the Rostov State University of Communications, 4(52), 73-77.

10 Maratkanova, O.E. (2018). Geomarketing approach to the placement of logistic infrastructure facilities. Socio-economic management: theory and practice, 1(32), 34-35.

11 Markelov, V.M., Tsvetkov, V.Ya. (2015). Geomonitoring. Slavic Forum, 2(8), 177-184.

12 Mayorov, A.A. (2014). Geomarketing research. Educational resources and technologies, 5(8), 43-48.

13 Petrov, Ya.A., Stepanov, S.Yu., Sidorenko, A.Yu., Martyn, I.A., Petrov, A.D. (2019). Geomarketing research as a tool for analyzing the target audience when choosing the location of a retail outlet. Information technologies and systems: management, economics, transport, law, 4(36), 44-48.

14 Petrov, Ya.A., Stepanov, S.Yu., Sidorenko, A.Yu., Martyn, I.A., Petrov, A.D. (2019). Geomarketing research as a tool for analyzing the target audience when choosing the location of a retail outlet. Information technologies and systems: management, economics, transport, law, 4(36), 44-48.

15 Rosenberg, I.N., Shaytura, S.V. (2016). Cluster analysis of tourist destinations on the Crimean peninsula. In the collection: Organizational and economic mechanism for managing the advanced development of regions.

16 Sharifyanov, T.F. (2017). Planning of social information and communication infrastructure of the region based on geomarketing methods. Practical marketing, 12(250), 29-34.

17 Shaytura, S.V., Kozhaev, Yu.P., Ordov, K.V., Vintova, T.A., Minitaeva, A.M., Feoktistova, V.M. (2018b). Geoinformation services in a spatial economy. International Journal of Civil Engineering and Technology, 9(2), 829-841.

18 Shaytura, S.V., Minitaeva, A.M, Ordov, K.V., Shaparenko, V.V. (2019). Virtual enterprises in a spatial economy. International Journal of Recent Technology and Engineering (IJRTE), 7(6), 719 - 724.

19 Shaytura, S.V., Ordov, K.V., Lesnichaya, I.G., Romanova, Yu.D., Khachaturova, S.S. (2018a). Services and mechanisms of competitive intelligence on the internet. Espacios, 39(45), 24.

20 Shaytura, S.V., Stepanova, M.G., Shaytura, A.S., Ordov, K.V., Galkin, N.A. (2016a). Application of information-analytical systems in management. Journal of Theoretical and Applied Information Technology, 90(2), 10-22.

21 Shaytura, S.V., Tsvetkov, V.Ya., Shaytura, A.S., Kozhaev, Yu.P., Kharitonov, S.V., Stepanenko, N.V. (2016b). Theory and Practice of Geomarketing: study guide. Burgas.

22 Shaytura, S.V., Vaskina, M.Yu. (2019). Integrated digital model of monitoring the area. Ecology of urbanized territories, 4, 71-76.

23 Shaytura, S.V., Vaskina, M.Yu. (2020). Monitoring of lands in the regions of the Far East. Land management, cadastre and monitoring of lands, 1, 28 - 33.

24 Tsvetkov, V.Ya., Shaytura, S.V., Minitaeva, A.M., Feoktistova, V.M., Kozhaev, Yu.P., Belyu, L.P. (2020). Metamodelling in the information field. Amazonia Investiga, 9(25), 395-402.

25 Tsvetkov, V.Ya., Shaytura, S.V., Sultaeva, N.L. (2020). Digital Enterprise Management in Cyberspace. Proceedings of the 2nd International Scientific and Practical Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth" (MTDE 2020). Yekaterinburg.

26 Tymchenko, E.V. (2014). Organization of data in geomarketing. Prospects for Science and Education, 6(12), 160-165.

27 Vorobyova, D.E., Degteva, E.V. (2018). Geomarketing as a tool for studying the competitive advantages of hotels. Russian regions: a look into the future, 5(4), 47-56.