Advances in Intelligent Systems Research, volume 174
Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

Developing Risk Assessment Model for Altering Conditions of Forest Reserves in an Oil-Production Region

Alexander Yakimchuk*
Yugra State University
Khanty-Mansiysk, Russia
YakimchukAV@uriit.ru

Vladimir Burlutskiy
Ugra Research Institute of Information Technologies
Khanty-Mansiysk, Russia
BurlutskyVV@uriit.ru

Alexander Tsaregorodtsev
Ugra Research Institute of Information Technologies
Khanty-Mansiysk, Russia
TsaregorodtsevAL@uriit.ru

Andrey Melnikov
Ugra Research Institute of Information Technologies
Khanty-Mansiysk, Russia
andmelnikov1956@yandex.ru

***Abstract*—The scientific problem, the solution of which is aimed at this work, is the implementation of a systematic method of assessing and predicting the influence of anthropogenic impact on the natural environment of the oil-producing region. The article considers the process of manipulating specific heterogeneous data and presents an implemented neural network model for predicting areas with the most likely risk of oil spill. A special feature of the proposed approach is the use of hybrid methods of machine learning in conjugation with the geo information analysis on the basis of the history of incidents occurred on sections of license areas. Some statistical estimates of the influence of the assessed factors of risk formation on an emergence probability of an incident have been obtained. Within the framework of this investigation, a vector-based description of signs of incidents was formulated and validated followed by a forecast based on methods of machine learning.**

Keywords—data analysis, machine learning, neural networks, spatial analysis, geographic information systems, risk-based approach

I. INTRODUCTION

According to the Federal Law as of July 13, 2015, No. 246-FZ, bodies of state control are to apply the risk-oriented approach while organizing separate types of state control as from January 1, 2018. By now, some criteria have been developed (approved by the Government of the Russian Federation's Resolution as of from 22.11.2017, No. 1410) of referring production facilities which are used by legal entities and individual entrepreneurs and which make a negative impact on the natural environment to a certain category of a risk as viewed by a regional state ecological supervision, and some particulars of the implementation of the said supervision have been described.

However, there are no criteria – which could assess risks of non-compliance with requirements of forest legislation – for making decisions on carrying out control-and-supervision activities with regards to controlled subjects (legal entities, individual entrepreneurs and citizens), as well as objects (sites on the forest territory in the range of the KMAO). Identifying such risks and establishing criteria of their assessment, as well as creating a model of the risk assessment of the condition of the forest reserves is the main target of the present paper.

The specificity of the KMAO from the viewpoint of the risk assessment of environmental management consists in its leading role among oil producing and electricity generating regions. The region also ranks second in the Russian Federation in volumes of industrial output and natural gas production. As a result more than 93% of the territory of the KMAO is occupied by forest lands, which sustain a constant negative impact from enterprises of the oil and gas complex and other users of natural resources. Control-and-supervision authorities of the region detect a large number of various violations in environmental management, which make a significant negative effect on the condition of the forest reserves. Taking all of the above into consideration, the goal for monitoring and risk assessment of the conditions of forest reserves on the territory of KMAO as well as for increased efficiency of control-and-supervised activities are of the extreme importance to the region.

Setting goals

The purpose of the present paper consists in assessing modern risks of negative ecological impact within the territory of the KMAO and forecasting their dynamics based on the spatial analysis of zones subject to a negative impact of natural and anthropogenic factors. Based on the results of modeling, a map of zoning of various levels of risk occurrence is plotted, which permits to solve an important task of increasing the efficiency of monitoring and patrolling of forest territory by control-and-supervision authorities.

In order to achieve the set objectives, it is required to define a list of significant pollution risks of the of territories, rank them and then develop a model which will allow to assess an integrated value of a risk for any site on the territory of the autonomous area. [3]

When determining pollution risks of forest reserves, diverse data is employed that of the actual state of forest reserves and its dynamic over the last year, which is obtained by means of analysis of temporal data of the remote sensing. Also, data on the availability of infrastructural objects, including those of the oil-and-gas complex, data on various ecological violations, incidents that were revealed earlier are considered when determining pollution risks. In order to sort various data into risk groups, the territory of the KMAO is covered with a grid, each cell of which represents a minimum structural quad of the model thus permitting to classify initial set of data on the basis of geographical coordinates.

The result of this solution will be a visualized model of risk assessment using geo-information systems.

The problem of forecasting the pollution of forest reserves from the point of statistical theory of decision-making can be considered as a task of classification of elementary sites of pollution, based on information about these territories over the past years, including that about emergencies occurred on the sites of oil production.

A formal setting of this task consists in the following. Let there be a set of basic elements, Zi, i =1,...,n, each of which is characterized by a p-dimensional vector of signs Xi = (x i1, …, X ip)T influencing impacting pollutions. The appurtenance of each site Yi to belonging of each site (1) to one of two classes of risks is also known:

$$Y = \begin{cases} y = 1 - \text{there is a risk of the pollution of a site} \\ y = 0 - \text{there is no risk of the pollution of a site} \end{cases} \quad (1)$$

According to the Government of Russian Federation's Resolution as of September 28, 2015, No 1029, "On the approval of criteria of referring objects making a negative impact on the environment to those of the I, II, III and IV categories", the carrying-out of economic and/or other activities connected with oil production refers a pollution object to the I category. However, such a binary classification is not sufficient for a subsequent analysis. Therefore, it is proposed that not only a category of risk of a negative impact, but also posteriori distribution of a negative impact should be used as an output information.

The subsequent distribution points – for each category of a risk – to a probability of the affiliation of a site to this category of a risk. For example, in case of two categories of risks with distributions 90%/10% and 55%/45%, a site has a risk of pollution in both situations with an apparent difference though.

Literature review

In order to eliminate an excess quantity of administrative barriers and various inspections imposed by the state, the legislation of the Russian Federation has recently established a principle of a risk-oriented approach while organizing state control [4].

According to A.V. Martynov [5], "The world-wide experience of using the risk-oriented approach testifies to the fact that its application has enabled to reduce the total number of inspections from 30 to 90 percent and entirely exempt separate categories of business from the need of undergoing planned inspections". N.E. Yegorova [6], in her turn, claims that the introduction of the risk-oriented approach in the Russian Federation "will permit reducing the number of inspections by 40 percent, and the number of inspectors by 30 percent – that will enable to save 20 percent of the funds of the budget".

They distinguish two types of model within the framework of the risk-oriented approach – static and dynamic. In a statistical model, results of an field checks are not considered in the classifying objects for the purpose of changing the measure of a risk. In a dynamic approach, results of an outdoor inspections and other factors make a direct impact on the category of an object of supervision and, as a result, on the intensity of inspections which are carried-out in relation to it [7].

By now, a large number of various classification algorithms have been developed, each of them allowing to single out their advantages and shortcomings. While establishing systems of risk management, the following methods of machine learning are used on the regular basis: method K of the closest neighbours (Stone, 1977), logistic regression (Walker, 1967), the randomised forest (Leo, 2001), the stochastic gradient descent (Robbins, 1971), the method of support vectors (Cortes, Vapnik, 1995), all of which have a sufficiently high precision of recognition. Frequently, they use combinations of classifiers (assemblies) [8,11]. Methods of machine learning are successfully applied when forecasting natural disasters. For example, method of support vectors is used to predict floods with 96% accuracy. [8] Research described in papers [9,10] show an efficiency of application of various methods of machine learning (SVM, DT, ANFIS) in forecasting landslides.

II. THE PROCESS OF GATHERING AND STRUCTURING OF VARIOUS DATA

In order to develop a model for forecasting pollution risks, it is proposed to use [9] data of the following information entities:

- Emergencies and incidents;
- Technogenic objects:
  - Objects of waste emplacement;
  - Inhabited localities;
  - Transport infrastructure.
- Forest users;
- Mineral developers.

Data describing information entities of a model have been obtained from the following sources:

- Data on violations and emergencies connected with environmental pollution and on mineral developers was obtained from the state information system «Ekonadzor» – the Service of Control-and-Supervision in the Sphere of Environmental Protection, Objects of Wildlife and Forest Relations of the KMAO (hereafter referred to as the Prirodnadzor of KMAO).
- Data on technogenic infrastructure, on forest users and on the subdivision of forests resources was retrieved from the archive of satellite images of the Centre of Space Services of the Yugra Research Institute of Information Technologies (hereafter referred to as URIIT).

- Additionally, specifying data on inhabited localities and transport infrastructure have been received from open sources of public cartographical systems.

The uniform five-kilometer grid is a binding structural component for all information entities. An individual site of this grid is a binding entity for all the information entities described above (hereafter referred to as an elementary site).

The process of integration of information entities on the base of a grid consisted of unification of geographical coordinates of various objects, accumulation of general qualifiers, removal of duplicating information and reduction of text signs to the uniform view.

As a result, the primary array of data on environmental pollution on the territory of the autonomous area was received based on the Prirodnadzor of KMAO data for the period of 2012-2019. The amount of this informative selection considering risk-determining factors made 29087 lines. The total of elementary sites amounted to 22054. It permitted to admit a possibility of the correct application of machine learning methods.

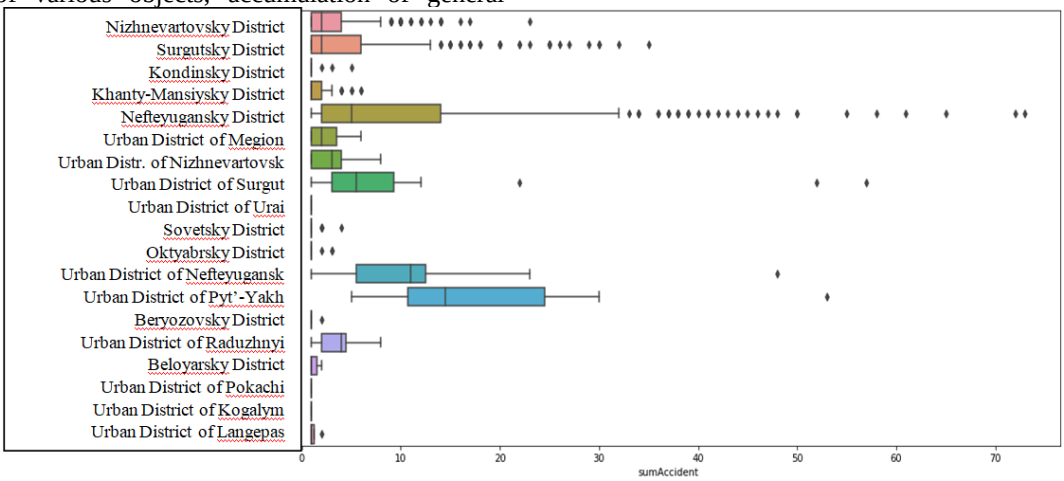

Fig.1. An area-wise distribution of emergences

III. PRELIMINARY PROCESSING OF GATHERED INFORMATION FOR THE PURPOSE OF ORGANIZATION AND RECOVERY OF MISSING DATA. DEFINING CRITERIA FOR EVALUATION OF RISKS FACTOR

Before initiation of models development, it is necessary to prepare initial data for analysis in order to be able to face difficulties when analyzing real statistical data.

For each symbol or element of a data set a measuring scale was defined taking into consideration the method of processing:

- Text elements were replaced with defined numerical ones;
- Continuous values were normalized based on their maximum value.

Coordinates are a key feature of a data set as they allow to unambiguously establish the appurtenance association of an object to an elementary site of a grid. In the obtained data, no more than 50% of objects (11946) have this feature. Therefore, it is necessary to affix the remaining objects if at all possible.

The following features were used to fit the objects on a grid: a site of a forest, a forest compartment, a forest sub-compartment or a cluster (if available). The remaining objects were affixed based on the data received by the Centre of Space Services on the sectionary of forest territory into sites, forest compartments, forest sub-compartments and clusters. This information enabled to bind 10721 objects to elementary sites which increased the selection by 41%.

Thus, no more than 10% of the collected data remained unbound. During data analysis, it was established that the unbound objects belonged mostly to the year of 2012. Therefore, it is beneficial to exclude this years' data while testing the model.

In order to verify correctness of data allocation on a grid, histograms and cartographical analysis were used as a few methods of visual analysis.

A box-plot is used for a visual analysis (Figure 1) which compactly represents an area-wise distribution of emergences. This way a diagram conveniently demonstrate a average, lower and top quartiles, the minimum and maximum value of a selection and projections.

One can see on the diagram that in some areas projections relate to the number of emergencies.

IV. RISK FACTORS

Before initiation of models development, it is necessary to prepare initial data for analysis in order to be able to face difficulties when analyzing real statistical data.

For each symbol or element of a data set a measuring scale was defined taking into consideration the method of processing:

- Text elements were replaced with defined numerical ones;
- Continuous values were normalized based on their maximum value.

Coordinates are a key feature of a data set as they allow to unambiguously establish the appurtenance association of an object to an elementary site of a grid. In the obtained data, no more than 50% of objects (11946) have this feature. Therefore, it is necessary to affix the remaining objects if at all possible.

The following features were used to fit the objects on a grid: a site of a forest, a forest compartment, a forest sub-compartment or a cluster (if available). The remaining objects were affixed based on the data received by the Centre of Space Services on the sectionary of forest territory into sites, forest compartments, forest sub-compartments and clusters. This information enabled to bind 10721 objects to elementary sites which increased the selection by 41%.

Thus, no more than 10% of the collected data remained unbound. During data analysis, it was established that the unbound objects belonged mostly to the year of 2012. Therefore, it is beneficial to exclude this years' data while testing the model.

In order to verify correctness of data allocation on a grid, histograms and cartographical analysis were used as a few methods of visual analysis.

A box-plot is used for a visual analysis (Figure 1) which compactly represents an area-wise distribution of emergences. This way a diagram conveniently demonstrate a average, lower and top quartiles, the minimum and maximum value of a selection and projections.

One can see on the diagram that in some areas projections relate to the number of emergencies.

Some objects allocated to elementary quad are presented in Figure 2. Sites with pollution emergencies occurring during the period from 2013 to 2018 are highlighted. Allocation together with the diagram, enabled one to make sure that emergencies accompanied with pollution most commonly happen in a part of areas.

Next, the dataset was analyzed for the presence of non-informative elements using the method of principal components.

As a result of the carried-out analysis and pre-processing of data, the initial data set of 22054 was obtained. The main elements are distances to technogenic objects; distance to a solid waste landfill; class and kind of a forest, of an area, and of a company which holds a license forest occurrence of emergences during the period from 2013 to 2018.

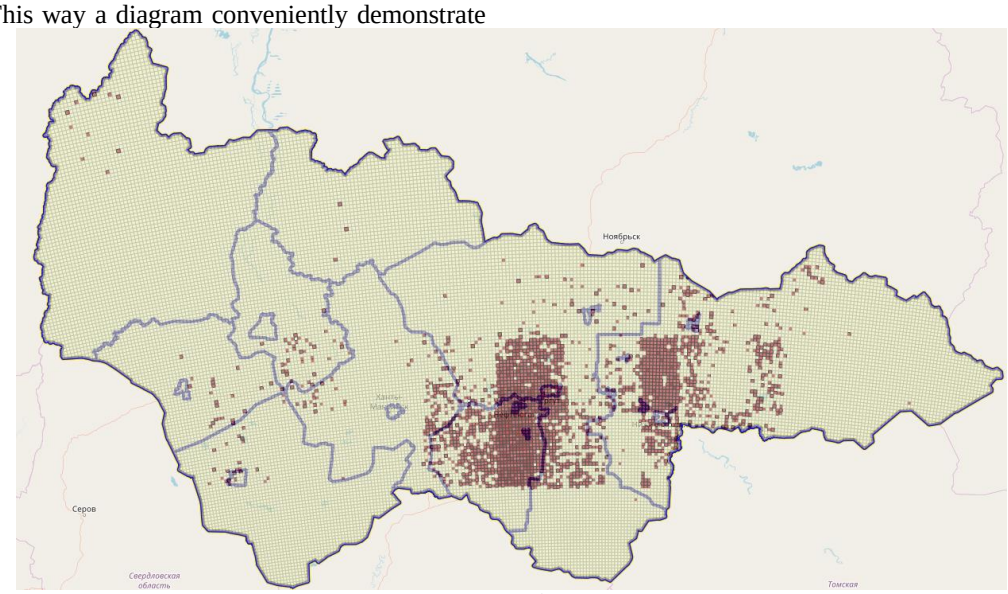

Fig. 2. A visual model of an initial data set

V. Description of the Developed Model of Risk Assessment

As a result of the carried-out analysis, a neural network model of the assessment of risks was elaborated. The incidents that occurred in 2018 were used as a label.

Objects in which emergences occurred in 2018 are a data set for forecasting. Their number amounted to 2248 objects out of 22054. The same quantity of incident-free objects was chosen. A data set for building up a model made 4496 objects.

A multilayered neural network with 3 layers was used as a forecasting model.

The following signs were used as data for an entrance layer: a distance to technogenic objects, a type and variety of a forest, an area, licenses forests, and occurrence of emergencies in the period from 2013 to 2017. It should be noted that the elements of a type and variety of a forest, a license and areas were vectorised, thus, an amount of input neurons was 83.

An amount of neurons in the 2-nd layer was 166, in the 3-rd layer - 83, in the 4-th layer – 40, and the output layer contained 2 neurons, one replied for an emergency occurred, the second one replied for no emergency occurred. The greatest output value of these two neurons was used as a reply.

VI. Description of the Obtained Results

Some K-fold cross-validations [12] were used as a method of model assessment. In this case, an initial data set was subdivided into 10 parts of identical size. One out of ten parts was left for testing a model and the remaining 9 were used as a training set. The process was repeated 10 times, and each one of the ten parts was used once as a testing set. Eventually, 10 results came out, one for each part. They are averaged or combined using some other method and yield one assessment. Results are presented in Table 1.

Table 1. An assessment using a cross-validation method

Number of a model	1	2	3	4	5	6	7	8	9	10
Assessment obtained	91.5	90.3	90.8	91.2	89.9	90.6	91.4	89.5	91.2	90.4

On average assessment makes 90.68.

An advantage of such a method over random subsampling consists in that all observations are used for both training and validation a model, and each observation is used for testing only once.

The resulting map with forecasted emergencies is presented on Figure 3.

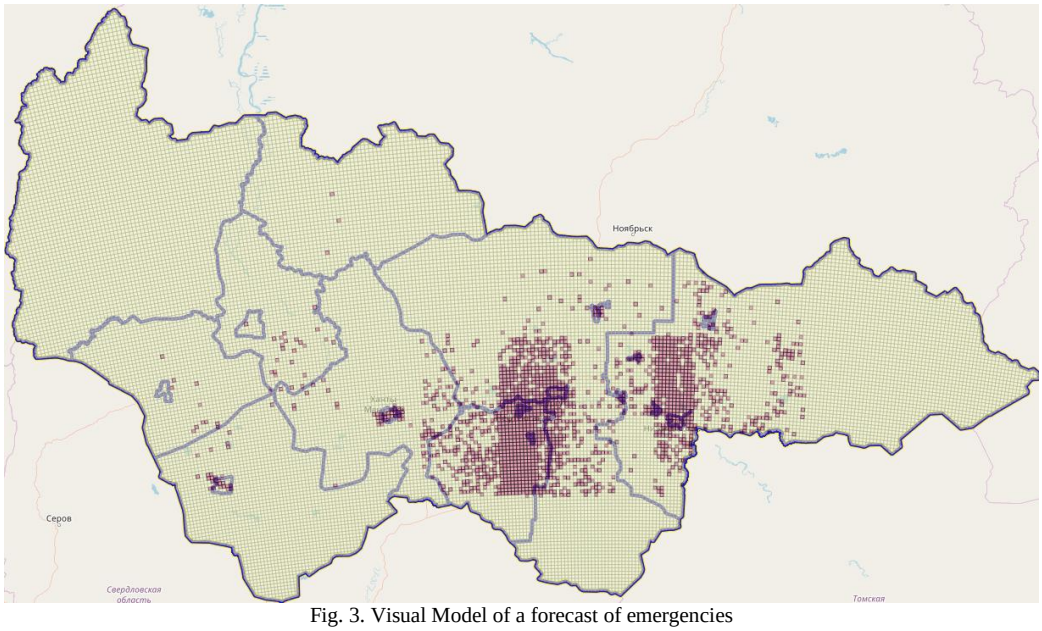

Fig. 3. Visual Model of a forecast of emergencies

VII. CONCLUSION

The solution of set objectives has permitted to identify the most significant risks for the forest resources in the region and development of a model risk assessment. The predicted data were visualized on the geoportal of the Centre of Space Services. Analysis of the obtained data will enable to lower a load upon the enterprises of the region arising from planned inspections of control-and-supervision authorities, and will increase the efficiency of the detection of violations in zones of elevated risks.

ACKNOWLEDGMENT

The reported research was funded by Russian Foundation for Basic Research grant No № 18-45-860003.

REFERENCES

[1] Gumenjuk V. I., Karmishin A. M., Kireev V. A. O kolichestvennyh pokazatelyah opasnosti tekhnogennyh avarij [Quantitative risk of industrial accidents] // Nauchno-tekhnicheskie vedomosti SPbPU. Estestvennye i inzhenernye nauki[St. Petersburg Polytechnic University Journal of Engineering Science and Technology] №2 (171):281-288. 2013.

[2] Seth D.Guikema. Natural disaster risk analysis for critical infrastructure systems: An approach based on statistical learning theory // Reliability Engineering & System Safety.94(4). 2009.

[3] Shokin Y., Moskvichev V., Nicheporchuk V. Metodika ocenki antropogennyh riskov territorij i postroeniya kartogramm riskov s ispol'zovaniem geoinformacionnyh system [Technique for estimation of anthropogenous risks for territories and construction of risks cartograms using geoinformation systems] //Vychislitel'nye tekhnologii [Computing technology]. 15(1):120-131. 2010

[4] Fomin A. I., Besperstov D. A., Saibel S. Yu. Fire risks and their influence on the risk-oriented approach in the organisation and implementation of federal state fire supervision // Bulletin of Scientific №3. 2017.

[5] Martynov, A. V. The application of the risk-oriented approach in the implementation of state control and supervision as a pre-requisite for reducing pressure on business // “Lawyer”, No. 18. Pages 22–27. 2016.

[6] Yegorova N.E. Three years without inspections // Information Bulletin “Express-Bookkeeping”, No 29. 2016.

[7] Black, J., & Baldwin, R. Really responsive risk-based regulation. Law & policy, 32(2), 181-213. 2010.

[8] Tehrany, Mahyat Shafapour, Biswajeet Pradhan, and Mustafa Neamah Jebur. "Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS." Journal of hydrology 512: 332-343. 2014.

[9] Hong H., Pradhan B., Jebur M.N. Spatial prediction of landslide hazard at the Luxi area (China) using support vector machines. Environmental Earth Sciences. 75(1): 40. 2016.

[10] Pradhan, Biswajeet. "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS." Computers & Geosciences 51: 350-365. 2013.

[11] Kussul, Nataliya N., et al. "Disaster risk assessment based on heterogeneous geospatial information." Journal of Automation and Information Sciences 42.12. 2010.

[12] Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. 2 (12): 1137–1143.