COMBINING SATELLITE IMAGERY AND MACHINE LEARNING TO PREDICT ATMOSPHERIC HEAVY METAL CONTAMINATION

Alexander Uzhinskiy [1,a], Gennady Ososkov [1], Pavel Goncharov [2], Marina Frontsyeva [1]

[1] *Joint Institute for Nuclear Research*

[2] *Sukhoi State Technical University of Gomel, Belarus.*

E-mail: [a] auzhinskiy@jinr.ru

Air pollution has a significant impact on the European and Asian countries. More than nine out of 10 of the world's population – 92%, lives in places where the air pollution exceeds safe limits, according to the research of the World Health Organization. There are a lot of regional and international environment control programs. They use different techniques and tools but, as a result, they all intend to understand what is the current situation and how does it evolve. Generally, to get some indexes, researches take samples and analyze them. For the natural reasons, sampling is carried out rarely, and the dimension of the sampling grid can be very big. In such a situation, modeling can be a right choice. In our research, we have tried to predict atmospheric heavy metal contamination combining satellite images and machine learning. Data sources for model training were satellite images from Google Earth Engine platform and sampling data from Data Management System of UNECE International Cooperative Program (ICP) Vegetation. We obtained satisfactory results in the prediction of Sb for Norway and Mn for Serbia.

Keywords: prediction, ecological monitoring, satellite images, machine learning

1. Introduction

Air pollution is the fourth largest threat to human health, behind high blood pressure, dietary risks and smoking. The health risks of breathing dirty air include respiratory infections and cardiovascular diseases, stroke, chronic lung diseases, and lung cancer. In the study of the World Bank and the Institute for Health Metrics and Evaluation (IHME) the economic cost of air pollution was calculated. It was found that air pollution led to one of 10 deaths in 2013, which cost for the global economy about $225 billion in loss of labor income [1].

There are many regional and international environment control programs. They use different techniques and tools but as a result, they all intend to understand what is the current situation and how does it evolve. Generally, to get some indexes researches take samples and analyze them. For the objective reasons a sampling is carried out rarely, and the dimension of the sampling grid can be very big. In such a situation, modeling can be a right choice. Our idea is to use real-life information about heavy metals concentration and indexes taken from the satellite images to train the special statistical model. After that, the model together with the new satellite indexes can be used to predict contamination for any part of the study area at any time.

It is clear that getting the indexes from the satellite images is a much easier process than a field sampling. A sampling and analysis of an area like Moscow Region can take 4-6 month. Gathering of indexes for the same area can be done for a few days.

2. Data sources for model training

To obtain reliable results we should have real contamination data and additional parameters. The sources of the contamination data are environment control programs. The most perspective source of the additional parameters for the models is satellite images done in various spectra.

2.1. UNECE ICP Vegetation

The aim of the UNECE International Cooperative Program (ICP) Vegetation in the framework of the United Nations Convention on Long-Range Transboundary Air Pollution (CLRTAP) is to identify the main polluted areas of Europe, produce regional maps and further to develop the understanding of the long-range transboundary pollution [2] . Atmospheric deposition study of heavy metals, nitrogen, persistent organic compounds (POPs) and radionuclides is based on the analysis of naturally growing mosses through moss surveys carried out every 5 years [3]. Nowadays the UNECE International Cooperative Program (ICP) Vegetation is realized in 39 countries of Europe and Asia. Mosses are collected at thousands of sites across Europe, and their heavy metals (since 1990), nitrogen (since 2005), POPs (persistent organic compounds, pilot study in 2010) and radionuclides (since 2015) concentrations are determined. A total of 13 elements are reported for the Atlas (As, Cd, Cr, Cu, Fe, Hg, Ni, Pb, V, Zn, Al, Sb, and N). Results are reported as a number of sampling sites, minimum, maximum and median concentrations in mg/kg. Specialists of the Joint Institute for Nuclear Research (JINR) developed a cloud platform (ICP Vegetation Data Management System, DMS, dms.jinr.ru) consisting of a set of interconnected services. The platform provides ICP Vegetation participants with the modern unified system of collecting, analyzing and processing of biological monitoring data [4]. More than 6000 sampling sites from 47 regions of different countries are presented at the DMS now from the Moss Survay 2015-2016.

As the developers of the DMS, we have an agreement with ICP Vegetation participants. Due to this circumstance, we can use these data in our research.

2.2. Google Earth Engine

Satellite programs like LandSat, MODIS, Sentinel provides free access to their data. One can search their database and find necessary images. Special software such as ENVI or ERDAS can be used to process images after that. Such an approach is not too comfortable, because images are of gigabyte size, and we should have a few of them to cover the region. Some software exists to search

through the image archives, but the functionality of these programs is rather poor, and they often work with only one satellite data source.

We used Google Earth Engine (GEE) – a cloud-based platform for the planetary-scale environmental data analysis to get satellite image indexes. The purpose of the Earth Engine is to perform highly interactive algorithm development at global scale, push the edge of the envelope for Big Data in remote sensing, enable high-impact, data-driven science and make substantive progress on global challenges that involve large geospatial datasets [5]. There are more than 100 satellite programs and modeled datasets. GEE has the JavaScript online editor to create and verify code online and Python API to communicate with user's applications.

There are as common programs presented at GEE - Landsat, Modis, Sentinel, as pretty specific - the MOD11A2 V6 product that provides an average 8-day land surface temperature (LST) in a 1200 x 1200 kilometer grid, VIIRS Nighttime Day/Night - Monthly average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB) or the MOD13A2 V6 – product that provides two Vegetation Indices (VI): the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), see examples at fig 1.

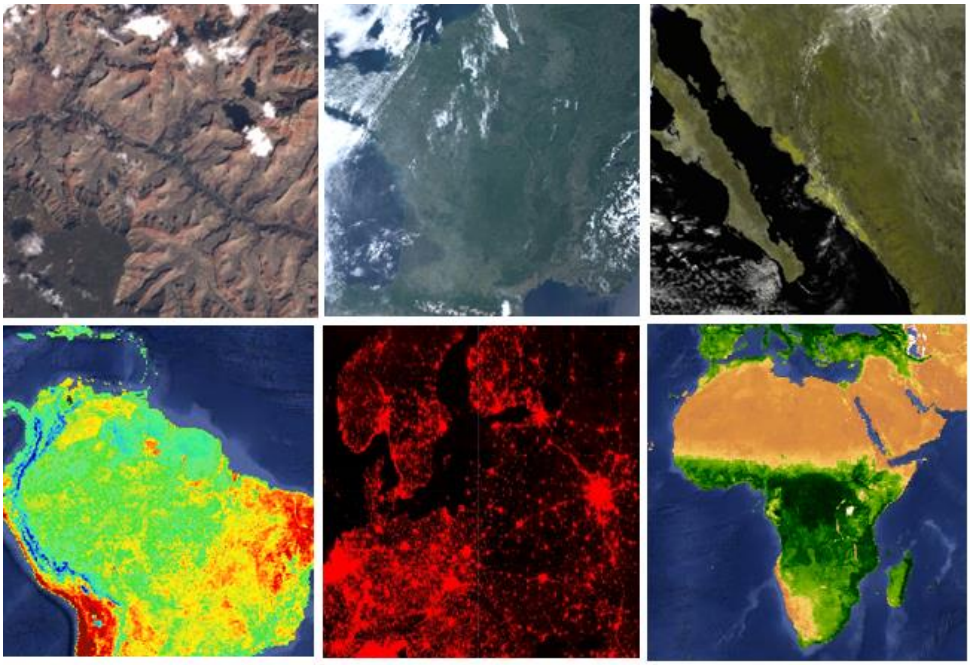

Figure 1. Examples of the satellite images at Google Earth Engine

To calculate the satellite image index with GEE, we undertake a few steps. We define a region, for example, 2 square kilometers with the center at the known sampling site point. Then we choose satellite program/product, for example, MODIS/006/MOD09A1 and define some filters like dates, weather, region, *etc*. After that, we get a collection of images and combine them with the median function. At the moment, we have an image of the region. For the image, we have a set of spectral channels, and GEE provides functionality to execute some mathematical functions (max, min, median, *etc*) for each channel. For example, we can calculate max(NDVI) for our 2 km^2 area based on MODIS satellite images.

3. Correlation of the contamination and satellite images

We have created a piece of software that takes coordinates of the sampling site from DMS and calculates indexes from different satellite program images. Then the correlation between the contamination in a region and indexes is defined. We had analyzed information for seven countries where the number of sampling sites were more than 200: Norway, France, Germany, Sweden, Rumania, Serbia, and Iceland. We used more than 20 satellite programs/products, and different types of indexes for them to find a correlation. We see that connection between elements and indexes varies

from region to region, and we can find several regions where few elements have correlations more than 0.5 with the number of indexes, see table 1.

Nature of such correlations can be an objective of the independent study. We consider the following possibilities: 1. Strait influence of aerosol heavy metal particles on the images in exact spectrum. 2. Indirect connection through technogenic and anthropogenic factors. 3. Indirect connection through landscape features and vegetation characters. 4. The mix of first three factors.

Table 1. Examples of the connection between contaminations of some elements and some satellite indexes

region	element	Proggram	index	area	correlation
France	Al	VITO/PROBAV/C1/S1_TOC_333M	max(SZA)	~ 0,9km2	0.554
Norway	Na	IDAHO_EPSCOR/TERRACLIMATE	min(aet)	~ 0,3km2	0.602
Norway	Sb	MODIS/006/MOD11A2	max(LSTDay_1km)	~ 2km2	0.609
Norway	Mn	MODIS/006/MOD17A2H	sum(PsnNet)	~ 0,9km2	0.566
Serbia	Mn	IDAHO_EPSCOR/TERRACLIMATE	sum(def)	~ 0,3km2	-0.599
Serbia	Na	IDAHO_EPSCOR/TERRACLIMATE	mean(srad)	~ 0,3km2	0.604
Romania	U	MODIS/006/MOD09A1	mean(surf_refl_b04)	~ 0,3km2	0.731
Iceland	Sb	NOAA/VIIRS/DNB/MONTHLY_V1/VCMSLCFG	max(avg_rad)	~ 0,9km2	-0.821

To use some statistical methods for prediction, we should find for an element at some region six or more indexes which have a satisfactory connection with element concentration, but a weak cross-correlation. This is a complicated task because different programs can make images in analogous or similar spectra, so the satellite image indexes of such programs can be strongly correlated. We manage to find eight or more indexes satisfactory correlated with Sb for Norway, Mn for Serbia and U at Romania.

4. Machine learning methods and algorithms

We used two types of statistical approaches: regression and classification. For each of them, we tried a tree-based model and artificial neural networks. Tree-based models include gradient boosting, decision trees, random forest and bagging [6]. The neural network approach assumes a usage of multi-layer perceptron with two hidden layers.

The goal of the regression task is to predict a single output, which represents the concentration. The common metric for regression tasks is the mean-squared error. The output layer of the neural network regressor is a single neuron with linear activation. The linear activation is used because concentration values are unbounded. We can reduce the regression problem to the classification task because a prediction of the exact concentration value in the particular point is not mandatory. The K-Nearest Neighbours (KNN) [7] algorithm was trained on the data of contamination for selection K data clusters, each of which corresponds to a particular level of contamination. Contamination values were replaced with the labels of corresponding clusters after the training of the KNN to pass them to the input of classifiers. The output layer of neural network classifier has K neurons equal to the number of clusters. Softmax activation on the output layer of the neural network classifier computes the distribution among classes to identify which of them can be chosen as model's response depending on the input data. We use the cross-entropy loss as the minimized metric during the training process of the neural network classifier. The predicted labels indicate the level of the contamination. We also tried to add the weights of the classes, but it did not have any significant impact on the prediction efficiency.

To find optimal parameters for the tree-based models we performed a special procedure named a grid-search cross-validation. This procedure consists of a total check of all possible combinations of the parameters and finding one, which allows obtaining a minimal cross-validation loss. Training data were permuted and splited into ten equal portions. Nine of them were used for training and the remaining one for the evaluation. We repeated the procedure ten times per each training epoch. The finding of the optimal parameters of the neural networks even with only two

hidden layers is a very time-consuming task. Thus, for the parameters selection of the neural network models, we used Tree-structured Parzen Estimator Approach (TPE). TPE allows doing a smaller number of iterations than the grid-search, while the results are better than random search [8]. The full list of explored parameters is available at [9].

In addition, we tried a few new models and improvements: we applied the MinMax normalization not only to the input values, but also to the concentrations – it allowed us to minimize the binary-cross entropy loss in the case of neural networks regressor instead of the mean-squared error method which works well only when dealing with the normally distributed data. Also, we tried to use the robust scaling of the input features. It performs a subtraction of the median value and scales the data according to the quantile range (defaults to IQR: InterQuantile Range). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). The value of the mean-squared error for the best regression algorithms measured on the test subset of the data was about 0.0035 (± 0.0015).

5. Modeling results

Having the models, we tried to predict the contamination for some regions. We focused on Sb for Norway and Mn for Serbia. One can see in fig. 2 a color-graduated map of Sb distribution in Norway and Mn distribution in Serbia. Various colors represent concentrations in mg/kg of the element in taken at that point sample. In table 2, one can also see some statistics about element distribution.

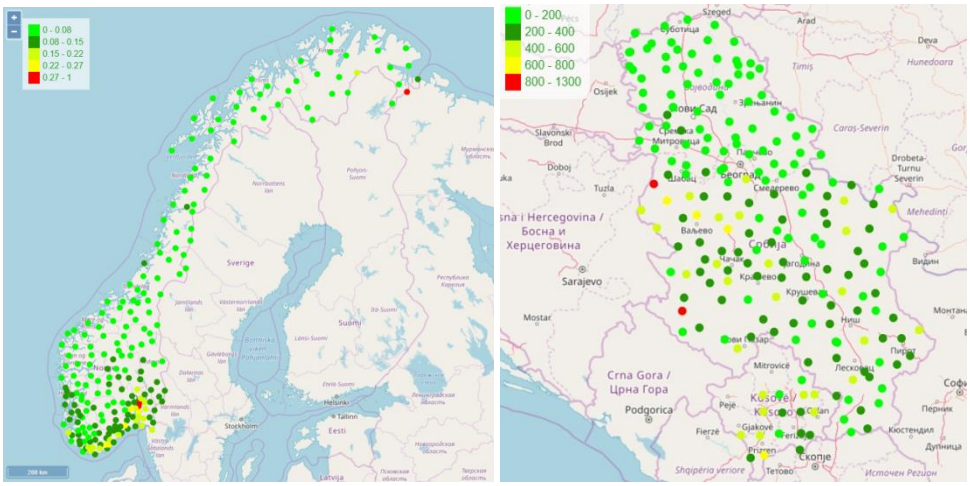

Figure 2. Color graduated map of Sb for Norway (left) and Mn for Serbia (right)

Table 2. Basic Mn and Sb distribution statistics

region	element	Sampling sites	Range	mean	median	± st.dev.
Norway	Sb	228	0.0065-0.376	0.08	0.067	0.06
Serbia	Mn	216	17.593 – 1136.108	226	188.835	190.31

Sb sources are mostly anthropogenic ones (traffic and industrial). We have a few programs in the list that represent temperature, radiance and other signs of the human population. 8 indexes are given in table 3, that we choose to train the model for Norway. Sources of Mn for Serbia could have a natural, agricultural, or industrial origin. As one can see, we used much smaller analyzed areas in this experiment than for Sb in Norway. In table 4, for 9 indexes are given that we choose to train our model for Serbia. We have removed two indexes with correlation ~0.52 from Serbia experiment to get better results.

Table 3. Indexes chosen to train model for Norway

Program	Index	Area	Correlation
PROBA-V C0 Level 3 Top Of Canopy Daily Synthesis at 100m resolution [PROBA-V 100m resolution]	sum(NDVI)	~ 36km2	0,636
MOD11A2.006 Land Surface Temperature and Emissivity 8-Day Global 1km [MOD11A2.006]	median(LST_Day_1km)	~ 16km2	0,628
Program	Index	Area	Correlation
PROBA-V C1 Level 3 Top Of Canopy Daily Synthesis at 333m resolution [PROBA-V 333m resolution]	median(SZA)	~ 6,25km2	-0,605
Sentinel-3 OLCI EFR: Ocean and Land Color Instrument Earth Observation Full Resolution [Sentinel-3 OLCI EFR]	max(Oa03_radiance)	~ 25km2	-0,57
VIIRS Nighttime Day/Night Band Composites Version 1 [VIIRS Nighttime]	max(avg_rad)	~ 16km2	0,587
USGS Landsat 7 Raw Scenes [Landsat 7]	max(B6_VCID_2)	~ 20,25km2	0,593
ASTER L1T Radiance [ASTER L1T Radiance]	max(B13)	~ 16km2	0,587
MODIS Nadir BRDF-Adjusted Reflectance, daily 500m [MCD43A4.006]	max(Nadir_Reflectance_Band5)	~ 49km2	-0,571

Table 4. Indexes chosen to train model for Serbia

Program	Index	Area	Correlation
Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces, University of Idaho [IDAHO_EPSCOR/TERRACLIMATE]	min(def) min(pr) max(soil)	~ 0,3 km2 ~ 1,5 km2 ~ 0,3 km2	-0.601 0.552 0.564
Sentinel-3 OLCI EFR: Ocean and Land Color Instrument Earth Observation Full Resolution [COPERNICUS/S3/OLCI]	mean(Oa21_radiance)	~ 0,9km2	0,556
NASA-USDA Global Soil Moisture Data [NASA_USDA/HSL/soil_moisture]	min(ssm)	~ 0,9km2	-0,54
MOD17A2H.006: Terra Gross Primary Productivity 8-Day Global 500m [MODIS/006/MOD17A2H]	sum(PsnNet)	~ 3km2	0,585
PROBA-V C1 Top Of Canopy Daily Synthesis 333m [VITO/PROBAV/C1/S1_TOC_333M]	max(SAA) max(VNIRVZA)	~ 0,9km2 ~ 0,9km2	-0,547 0,53
NOAA CDR AVHRR LAI FAPAR: Leaf Area Index and Fraction of Absorbed Photosynthetically Active Radiation[NOAA/CDR/AVHRR/LAI_FAPAR/V4]	max(FAPAR)	~ 3km2	0,563

We trained models on the real-life data and satellite image indexes. After that, we calculated the new indexes to get the prediction. We had 1198 points at the south part of Norway. We applied our models, and gradient boosting regressor gives the best results, which presented in fig 3.

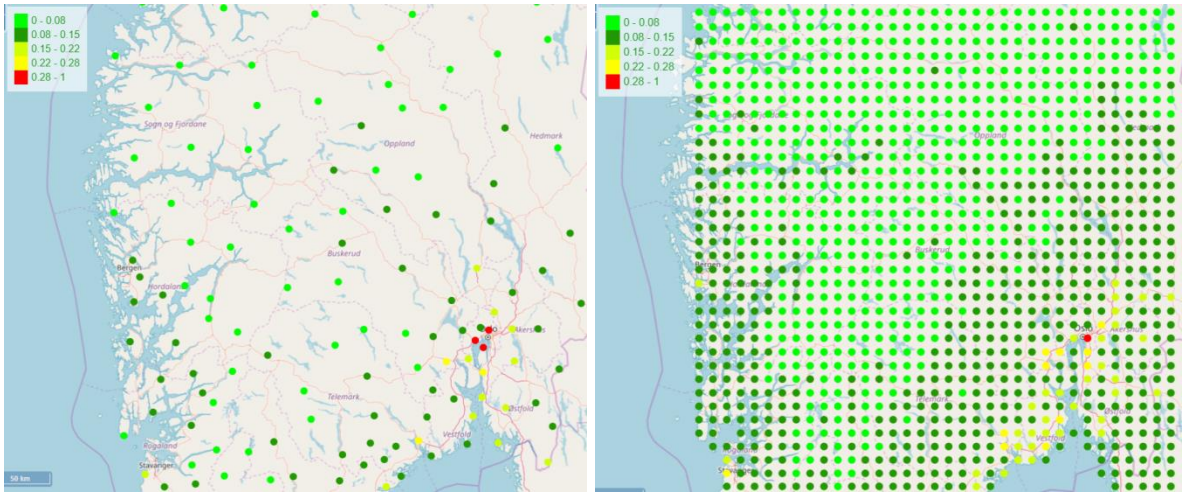

Figure 3. Sb distribution in Norway. Real (left) and modeled (right) data

One can see that the model shows tendencies very well. We tried the same model on the north part of Norway and the results were the same.

We have calculated indexes for 958 points covering the whole Serbia and some trans-border region. We used an automation to choose the best model. To the first set of 742 points, we added 216 points geographically close to the sampling sites points (their longitude was moved on 0.001 aside). Then we done 1000 iterations of training and prediction. On each step, we verified is the new model statistically closer to the real-life data. At the end, we chose the best model from different model types. Gradient boosting regressor gave the best result, see fig. 4.

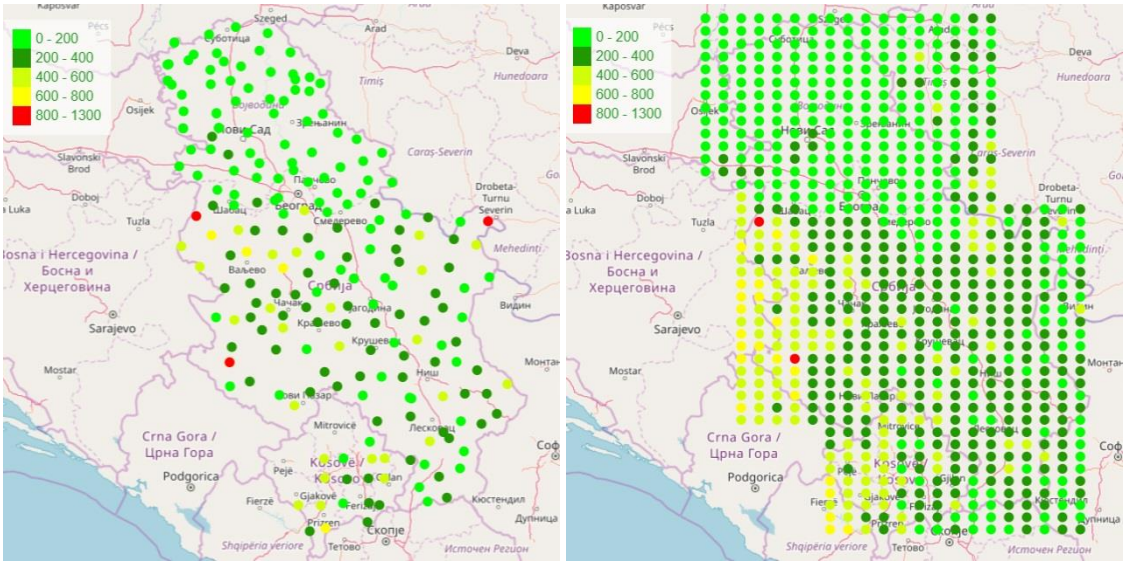

Figure 4. Mn distribution at Serbia. Real (left) and modeled (right) data

The model shows tendencies and peak values clearly. If we can find indexes with better correlation with real data, the accuracy of the model can be increased even more.

We made also some experiments with U for Romania but results were not satisfying, in our opinion. We are going further to investigate where the problem is in indexes, or it is in models.

4. Conclusion

Indexes from satellite images combined with the special statistical models can be used to predict the atmospheric contamination of some heavy metals in some regions. The connection between elements and indexes varies from region to region, so there is no unified solution or a unified model. Modeling can open new horizons for the contamination analysis. Researchers will be able to monitor the evaluation of the situation when it is needed, get detailed information about the areas of interests, check the situation in the cross-border areas, partly automate environment control process (automatically run the model and get the notification when the contamination level is higher than the critical level).

We keep on searching connection of the contamination and satellite indexes and testing new models and approaches.

References

[1] Rosamond Hutt, Keith Breene, 7 shocking facts about air pollution//World Economic Forum, 2016. Available at: https://www.weforum.org/agenda/2016/10/air-pollution-the-true-cost-in-numbers/

[2] United Nations Economic Commission for Europe (UNECE) Conventions and Protocols [Electronic resource]: http://www.unece.org/env/treaties/welcome.html. (Accessed 1.10.2018)

[3] UNECE International Cooperative Programme on Effects of Air Pollution on Natural Vegetation and Crops [Electronic resource]: http://icpvegetation.ceh.ac.uk/. (Accessed 1.10.2018)

[4] Ososkov G., Frontasyeva M., Uzhinskiy A., Kutovskiy N., Nechaevsky A., Cloud platform for data management of the environmental monitoring network: UNECE ICP Vegetation case // CEUR Workshop Proceedings, Vol. 1787, 2016, Pages 224-229.

[5] Google Earth Engine [Electronic resource]: https://earthengine.google.com/. (Accessed 1.10.2018)

[6] Hastie T. Trees Bagging Random Forests and Boosting // Stanford University — 2003.

[7] Alsabti K., Ranka S., Singh V. An efficient k-means clustering algorithm — 1997.

[8] Bergstra J. S. et al. Algorithms for hyper-parameter optimization // Advances in neural information processing systems. – 2011. – P. 2546-2554

[9] Uzhinskiy A., Ososkov G., Goncharov P., Frontasyeva M., Perspektivy ispol'zovaniya kosmosnimkov dlya prognozirovaniya zagryazneniya vozdukha tyazhelymi metallami [Perspectives of using a satellite imagery data for prediction of heavy metals contamination] // Computer Research and Modeling, 2018, vol. 10, no. 4, pp. 535-544, ISSN 2076-7633