Central Russia heavy metal contamination model based on satellite imagery and machine learning

A. Uzhinskiy [1], *K. Vergel* [1]
[1] *Joint Institute for Nuclear Research, 141980, Dubna, Moscow region, Russia, 6 Joliot-Curie*

Abstract

Atmospheric heavy metal contamination is a real threat to human health. In this work, we examined several models trained on in situ data and indices got from satellite images. During 2018-2019, 281 samples of naturally growing mosses were collected in the Vladimir, Yaroslavl, and Moscow regions in Russia. The samples were analyzed using Neutron Activation Analysis to get the contamination levels of 18 heavy metals. The Google Earth Engine platform was used to calculate indices from satellite images that represent summarized information about sampling sites. Statistical and neural models were trained on in situ data and the indices. We focused on the classification task with 8 levels of contamination and used balancing techniques to extend the training data. Three approaches were tested: variations of gradient boosting, multilayer perceptron, and Siamese networks. All these approaches produced results with minute differences, making it difficult to judge which one is better in terms of accuracy and graphical outputs. Promising results were shown for 9 heavy metals with an overall accuracy exceeding 89 %. Al, Fe, and Sb contamination was predicted for 3,000 and 12,100 grid nodes on a 500 km^2 area in the Central Russia region for 2019 and 2020. The results, methods, and perspectives of the adopted approach of using satellite data together with machine learning for HM contamination prediction are presented.

Keywords: heavy metal contamination, modeling, air pollution, biomonitors, prediction, satellite imagery, machine learning, neural architectures, Siamese neural networks.

Citation: Uzhinskiy A, Vergel K. Central Russia heavy metal contamination model based on satellite imagery and machine learning. Computer Optics 2023; 47(1): 137-151. DOI: 10.18287/2412-6179-CO-1149.

Introduction

This research complements the author's earlier paper [1] related to the platform-based intelligent environmental monitoring system of the UNECE ICP Vegetation program. The United Nations Convention on Long-Range Transboundary Air Pollution (CLRTAP) obligates participants to annually collect samples of naturally growing mosses and analyze them to generate information on the contamination of air with pollutants, particularly heavy metals (HM), such as Antimony, Mercury, Lead, etc., in addition to organic pollutants like benzo[a]pyrene, and radioactive constituents like radionuclides. These data are then presented in the form of an atlas with statistical metrics and pollution maps [2]. Atmospheric heavy metal contamination is a real threat to human health, and the severity of the hazard depends on the inherent characteristics of metal contaminants and their concentration in the exposed surroundings [3]. The toxic/carcinogenic potencies of these pollutants are compound-specific and depend on their dose exposure. Understanding compound-specific contamination levels on local, regional, and global scales is important for assessing the severity of the adverse consequences of contamination on humans. Biomonitoring can provide much more information than stationary and mobile air quality monitoring stations focused on determining the levels of airborne pollutants, including particulate matter (PM) [4, 5]. Unfortunately, biomonitoring surveys are limited both spatially and temporally. In such a situation, one can choose modeling, which has an enhanced analysis potential.

We have some experience in predicting contamination on regional and urban levels using machine learning and satellite imagery techniques [6, 7]. There is still a lot of direction for improving and verifying our methods. We have access to the Vladimir, Yaroslavl, and Moscow region data reported in 2018 and 2019. This motivates us to create a heavy metal contamination model for Central Russia. We have information on HM concentrations at 281 sampling sites. The data are obtained using a moss biomonitoring technique and Neutron Activation Analysis. Mosses are regarded as one of the main bioindicators of air pollution since they have a superficial root system [8]. We use these data to train the model and create a prediction for 3.000 and 12.100 grid nodes on a 500 km^2 area in the Central Russia region. The keystone of our research is additional data for models and advanced statistical and neural network architectures. We focus on satellite imagery as it is an easily accessible source of data. Google Earth Engine (GEE) is a cloud-based geospatial analysis platform that enables users to visualize and analyze satellite images of our planet. We use the GEE Python interface to calculate the so-called indices representing sampling areas. The indices are then used together with in situ data on heavy metal contamination to train the model. After that, indices are calculated for a specific area with a high spatial resolution and then used as the input to the model for a prediction. This ap-

proach works only if there is a meaningful, informative, or correlative connection between the indices and in situ measurements. Since the late 1990s, there have been published many research reports that mention the use of spectroscopy and satellite imagery to determine the concentration of heavy metals in soil. [9, 10, 11, 12]. The studies focus on the estimation of heavy metal contamination near mines and earth breaks. The determination of different PMs based on satellite imagery is a highly popular research work [13, 14, 15]. At present, there are specialized satellite programs, such as Sentinel-5, useful for assessing air quality, including concentrations of ozone, methane, formaldehyde, aerosol, carbon monoxide, nitrogen oxide, and sulfur dioxide. In our case, we try to explain the connection between data from satellite images and the concentration of heavy metals in mosses. Many researchers report that heavy metals in biomass directly influence satellite imagery data. Muradyan et al. [16] estimate the content of Mo, Cu, Ni, Cd (heavy metals) in potatoes and bean leaves using multispectral satellite imagery. Meiling et al. [17] show the possibility of applying multi-temporal satellite images to detect heavy metal-induced stress (i.e., Cd stress) in rice crops. Amer et al. [18] report that hyperspectral vegetation indices have a potential for monitoring Zn and Cu concentrations in wheat plants and grains. Zhou et al. [19] examine the relationship between the leaf reflectance of different seasons and the concentration of heavy metal elements, such as Co, Cu, Mo, and Ni, in leaves in a post-mining area. Yu et al. [20] show that elevated Cd and Pb concentrations induce contrasting spectral changes in the red-edge (690 – 740 nm) region for Tilia tomentosa trees.

The forest and tree leaves in particular can be a good source of contamination data. Bjerke et al. [21] show that birch leaves can be used to determine Cu and Ni concentrations. Khosropour et al. [22] report that Platanus orientalis leaves have a potential for monitoring Cd, Pb, Ni, and Cr. Alahabadi et al. [23] examine different tree species in terms of their capability to accumulate airborne and soilborne HMs and report that a number of species can be used for the phytoextraction of HM pollution. Terekhina et al. [24] examine the accumulation of chemical elements by the leaves of trees and shrubs in urban environments. Lyanguzova et al. [25] present the results of long-term monitoring of the state of boreal forest ecosystems (the Kola Peninsula, the European North of Russia) that experience industrial pollution from the Norilsk Nickel Mining and Metallurgical Company. Lassalle et al. [26] demonstrate the potential of hyperspectral imaging to assess metal uptake by plants.

Many types of studies focus on the determination of HMs in soil [27, 28, 29, 30]. They report the possibility of As, Pb, Cr, Cu, Cd, Fe, Mn, Zn, Sb, Hg concentration determination using reflectance spectroscopy. For some HMs, there is also an indirect relation between anthropogenic emission and sample contamination, which can be detected by satellite observations. Some researchers report that a high level of Antimony (Sb) can be associated with a higher level of temperature and/or luminosity, which characterize areas with heavy traffic, such as urban [31, 32, 33]. Thus, it is recognized that heavy metals in aerosol, biomass, and soil influence reflection, absorption, or transmission spectral characteristics. The above mentioned studies on the estimation of HMs in soil use advanced machine learning approaches, and their methods are close to ours. They mostly use soil samples and one or a few satellite programs as a source of additional data. Fang et al. [34] use a multi-layer perceptron to determine Cu contamination. Xu et al. [35] estimate Hg, Cr, Cu, and other heavy metal pollution areas in agricultural soils with the help of a generalized regression neural network. In a study by Pyo et al. [36], a convolutional neural network is adopted to estimate Arsenic (As), Copper (Cu), and Lead (Pb) concentrations using measured soil reflectance. There are also several interesting investigations, but we could not find any work related to the estimation of atmospheric heavy metal contamination using mosses as a source of in situ data or Siamese networks as a basic neural architecture.

In our previous studies, we used classical machine learning approaches (gradient boosting, learning trees, etc.) to train a regression model [7]. We have several experiments using neural networks, however, statistical methods over-perform them, probably due to the limited training dataset. We are currently focused on the classification task for several reasons. Firstly, maps are the main object of interest, and contamination levels are already well known or can be easily determined. In most cases, it is 5 to 8 levels of contamination. Secondly, the determination of accuracy metrics is clearer in the classification task. We can use accuracy, loss, F-score, and other metrics, while for the regression task it is in a general variation of the mean square error and R squared. Thirdly, there are fewer sampling points with a high level of contamination than points with a normal level, and better results can be achieved by using training dataset balancing techniques. For example, after balancing the dataset for Al, we have 1.176 samples instead of 281 (147 members in each of the 8 classes). In this research, we examine three approaches: Gradient Boosting, Multilayer perceptron, and Siamese network. Gradient Boosting is a well-established algorithm that showed the best results in our previous research. A multilayer perceptron is also a well-known solution for solving classification tasks. Expectations are raised on a neural architecture based on a Siamese network. Siamese networks are a well-known solution in image classification and facial recognition [37, 38, 39], but they also show good results in different areas such as object tracking [40], brain imaging modality recognition [41], bioacoustics classification [42], and remote sensing scene classification [43]. The benefit of Siamese networks is training in pairs of examples. Their main task is to distinguish one object from another. Here, the training dataset is extended because of combinatorial

functions. For Al, we have 8.234 pairs for training. Thus, in our study, we use both classical methods and Siamese networks to compare the results.

1. Materials and methods

1.1. Sampling

The moss biomonitoring technique was developed in the late 1960s by Scandinavian scientists [44]. Since then, this method has become widespread. Mosses absorb and accumulate chemical compounds and substances from the air since they have a superficial root system [45], and for this reason, they are widely used as biomonitors. In the late ’80s, an international research program was founded to investigate the impacts of air pollutants on crops and semi-natural vegetation (ICP Vegetation). As a part of this program, the Atlas of atmospheric deposition of heavy metals is published every 5 years. This Atlas summarizes data available from associated scientific groups in Europe and Asia. The Department of Neutron Activation Analysis (NAA) of the Joint Institute for Nuclear Research has been taking part in ICP Vegetation since its first survey in 1995.

Since then, it has performed investigations in many Russian regions (Moscow, Tula, Tver, Ivanovo, Udmurtia, Yaroslavl, Vladimir, Ryazan, Saint-Petersburg, etc.). NAA is a multi-element analysis that enables the analysis of up to 45 elements from samples [46]. The ongoing research uses sampling data gathered in the Moscow region during the summer of 2019 and in the Yaroslavl and Vladimir regions during the summer period in 2018 following a special protocol of the ICP Program [47]. The protocol sets requirements for the choice of sampling sites and samples, for example, moss spices, locations of nearby trees, areas to be avoided, preferable places for moss collection (ground or surface of decaying stumps), and other aspects.

The regions located in the central part of Russia are densely populated and have diverse industrial establishments, which determines the choice of these regions for the investigation. We use the information on 73, 53, and 156 samples from the Vladimir, Yaroslavl, and Moscow regions. The sampling map is presented in Fig. 1.

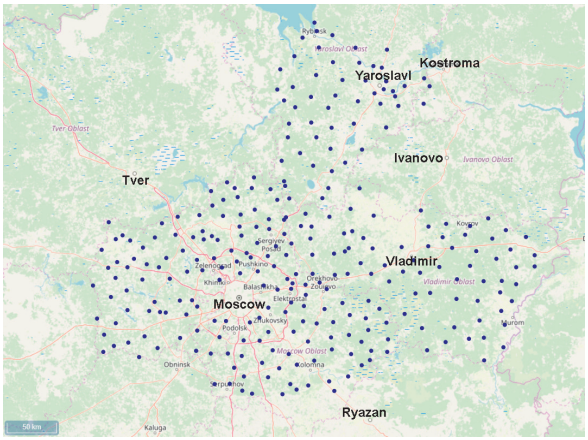

Fig. 1. Sampling map of 281 sampling sites in Central Russia

The Moscow region is more contaminated than the others. However, there are also some hot spots in the Vladimir and Yaroslavl regions. The combined dataset reveals the diversity and extent of contamination observed in the Central Russia region.

1.2. Satellite data

Hyperspectral images are a unique source for obtaining many kinds of information about the Earth's surface. Modern platforms support users to perform complex analyses with a collection of images without using any specialized software. Google Earth Engine (GEE) is a planetary-scale platform for Earth science data & analysis. Atmospheric, radiometric, and geometric corrections have been made to a number of image collections at GEE. There are over 100 satellite image collections and modeled datasets. Some collections have a spatial resolution of up to 15 meters. With just a few commands, GEE enables to get a median image by specifying the collection name, date frame, and area of interest. We use GEE to auto-calculate indices for model training and basic data for prediction. The index includes the name of the satellite image collection, the data retrieved and used, the size of the analyzed area, the identifier of the spectral channel (band), and the mathematical function applied to the digital matrix of the obtained image. For example, we can use data of monthly average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/Night Band (DNB). This collection has only one band, “avg rad”, average DNB radiance values. We can then get the median image of the collection for 01.01.2018 – 31.12.2018, by specifying the area of interest (10 km^2). Finally, we use mathematical functions, such as sum, max, and median, to reduce the volume of data and get some numerical values. Fig. 2 shows a general description of the index calculation to transform spectral information into textual information.

In this research, indices for over 40 collections using 5 reducer functions, i.e., max, min, sum, median, and mean, are computed. While working with row data, we use the built-in GEE function to filter data and create composites to get the cloud score threshold and percentile. We also use a variation of the SIAC module for atmospheric corrections [48].

We calculate Spearman and Pearson correlations between the indices and HM concentrations. While the Pearson correlation assesses linear relationships, the Spearman correlation evaluates monotonic relationships. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. Based on all the data, we select collections and reducer functions with the best correlation. It turns out that we can exclude some collections and reducer functions as they never appear on the best index list.

As a result, we keep working with only 13 collections (see App. 1 for a full list) and the median reducer function. We are going to predict contamination not only in 2019, but

also in 2020. Therefore, we assume that the median reducer is preferable since it can eliminate outliers. We experiment with the mapping unit size and end up working with 4 km^2 for 10 collections and 1 km2 for 3 Sentinel V5 collections. This combination shows the best results in terms of performance and correlation. GEE has some limits, and if we increase the analyzed area, the computation time also increases. At the final stage, the calculation of indices for 3.000 grid nodes takes over 2 days. The calculation of indices for 12.100 grid nodes takes approximately 1 week.

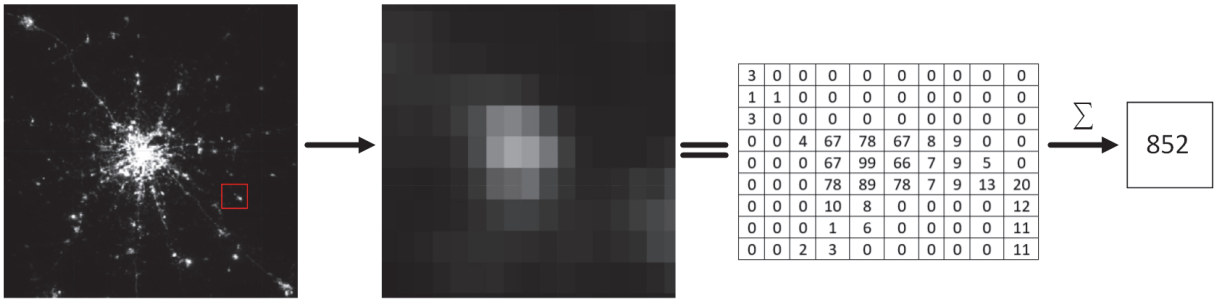

Fig. 2. Index calculation algorithm

1.3. Machine learning and neural networks

In our previous research, the best results were obtained using statistical techniques, in particular, variations of Gradient Boosting. We have some experiments with neural networks, but due to the lack of training data for regression tasks, the accuracy is lower, and it is not worth pursuing. Most sampling sites in the regions studied have a low concentration of HMs, and there are only a few sites with a high concentration of HMs. In such a situation, one should prefer neural networks to work with a balanced training dataset. In this study, we focus on the classification task, thus it is possible to use dataset balancing techniques. We use the imblearn Python package, in particular, its over-sampling method. The idea is to balance minority classes to reach the majority class. The technique requires a careful attention to ambiguous regions with class overlapping in the data. We analyze dot plots and frequently observe overlapping in close classes. However, there are parameters in close classes in which the intersections are single (fig. 3).

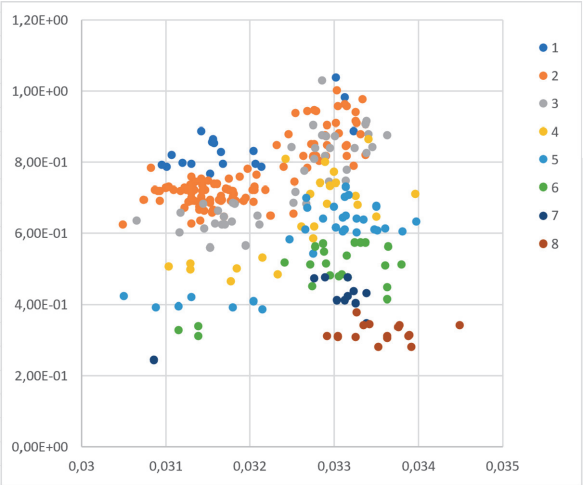

Fig. 3. Scatter plot between the Sentinel-2 and Sentinel-5P Indexes for 8 contamination classes

A typical sample consists of 10 or more parameters. Thus, we assume that the classes will not intersect strongly in a multidimensional feature space, and the application of the technique is acceptable. Below we use RandomOverSampler from the imblearn package, the Minority Oversampling Technique (SMOTE) [49] and ADASYN [50] to create new samples with fairly similar results.

For example, Al, Sb, and Fe minority classes have 4, 16, and 5 elements, and majority classes have 147, 117, 133 elements. After balancing, the training dataset increases over 2.5 times. Instead of 281 samples, we have 1,176 samples for Al (147 members in each of the 8 classes).

As part of the training data, we have some indices and HM element contamination classes. Two approaches are tested. First, we use only noncollinear indices with Spearman or Pearson correlations with HMs that are meaningful for our task (>0.35 and <-0.35). We have 10 to 14 of such indices. For different elements, the indices may vary. Second, we use all the indices we have, despite correlation and collinearity. It is assumed that neural models can use only principal features for prediction. In the second approach, the feature vector has 88 elements. For all models, we use the 80/20 train /test split. For classification, we test three approaches: variations of gradient boosting, multilayer perceptron, and Siamese networks.

Gradient boosting (GB) [51] relies on the judgment that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set target outcomes for this next model to minimize the error. In gradient boosting, the target outcomes for each case are set on the basis of the gradient of the error for the prediction. Each new model takes a step towards minimizing the prediction error in the space of predictions for each training case. We try XGBoost [52] and gradient boosting from the scikit-learn library with the parameters found by the grid search procedure. In most cases, we use several estimators (nearly 100) and a learning rate of 0.075. XGBoost is a unique and universal instrument as all fine-tuning procedures are done automatically; therefore, we use it to create final models.

A multi-level perceptron (MLP) [53] is a supplement of a feed-forward neural network. It has three types of

layers: input, output, and hidden. The input layer receives an input signal for processing. The classification task is performed by the output layer. Hidden layers, which are placed between the input and output layers, are the true computational engine of the MLP. Neurons in the MLP are trained with a back-propagation learning algorithm. MLPs are designed to approximate any continuous function and can solve problems that are not linearly separable. Different architectures are tested; however, the best results are obtained with a multi-level perceptron of three hidden layers. A grid-search procedure applies to find the best hyperparameters. We use the Relu activation function [54] on each layer with binary cross-entropy as a loss function. The learning rate is set to 0.01. We try different batch sizes and training approaches, but cannot get any better accuracy metrics than that of GB.

The Siamese neural network (SNN) [37] architecture comprises two or more twin networks with tied weights joined by the similarity layer with the energy function at the top. Parameter updating is mirrored across both twin networks. It is used to find the similarity of the inputs by comparing their feature vectors. When we pass an object to the network input, we extract some features of the object in the output, the so-called encoding. Similar objects cannot be in very different locations of the feature space since each of the twins computes the same function due to weight sharing. The architecture of the Siamese network is illustrated in Fig. 4.

We use the MLP with three hidden layers as the basic twin architecture. On the input, we pass a feature vector of 11–13 parameters for the selected indices or a vector with 88 parameters for all indices. Our Siamese network unites the twins within the L1 distance layer, followed by sigmoid activation to train the network with a cross-entropy objective. The dimension of the feature vector extracted from the embedding model is 60.

The training dataset for the Siamese network comprises positive and negative objects of different classes. Because of the different combinations of pairs, the Al dataset for training increases from 1.176 to 10.290 elements. As a result, we have a network that can distinguish object classes with high accuracy. The rest looks similar to the transfer learning approach. After training, one twin is used as a feature extractor for the one-layer MLP, which acts as a classifier. While the training weights of the feature extractor are frozen, the resulting network is trained with Adam's optimizer [55] and the categorical cross-entropy loss function. For most elements, this approach achieves an accuracy equal to or better than the accuracy of GB.

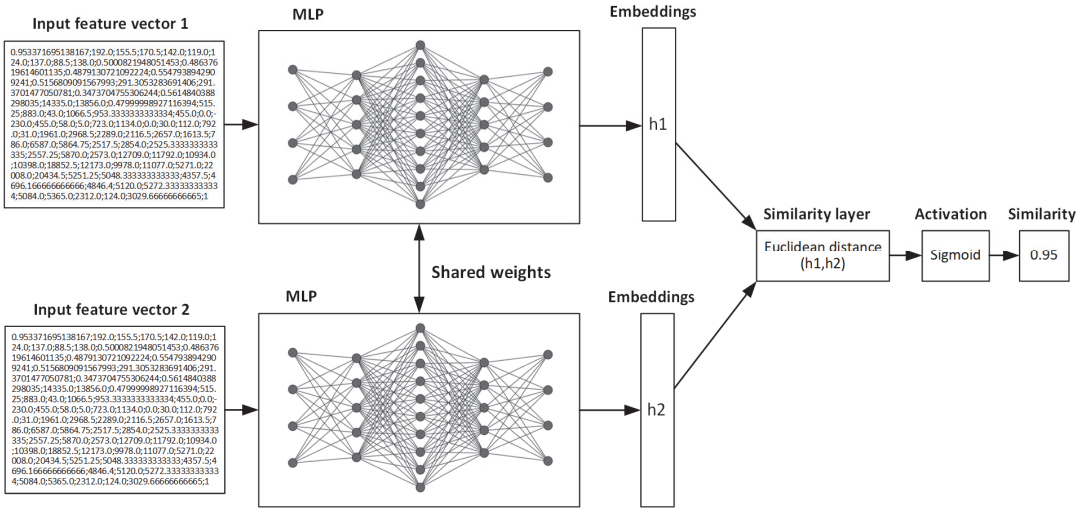

Fig. 4. Siamese network architecture

1.4. General pipeline

Our approach to heavy metal contamination modeling can be presented in a series of steps. The first is data preparation. It is necessary to calculate indices for sampling sites where in situ measurements are carried out. Then, through all the indices, those with a good connection with HMs should be selected. It is better to use only noncollinear indices with sufficient correlation with real measurements. We heuristically determine 0.35 as a correlation level for our task. After that, indices for modeling grid nodes with the required spatial resolution should be calculated. We calculate indices for different periods (2019 and 2020). At the end of this phase, we have the training data and data necessary for prediction. The second step is the search for the best model. We have some scripts that automatically run tests. All we have to provide to them is the training data, and as a result, we have all accuracy metrics for each model. In our research, we only fine-tune the models at the beginning. Fine-tuning the model for each element can have a positive influence on the accuracy, but to get the unified solution, we do it only once and then use the established parameters. The third step is the evaluation of the model and the interpretation of the prediction. Even for models with high accuracy, the prediction is not always invariant and may differ in some areas. We use our best judgment to select plausible models through all candidate models. Finally, if the

results are not satisfactory, we can repeat all the previous steps with needed corrections. The model is accepted if the results are satisfactory. In our case, we use the models to get a prediction for the area with a high spatial resolution for 2019 and 2020.

2. Results and discussion

The indices are gathered for data from 13 collections of 281 sampling sites. Their linkage with the concentration of 18 heavy metals is verified. We test two approaches to creating a training dataset. In the first one, we pass a feature vector with only pre-selected noncollinear indices with sufficient Spearman or Pierson correlations. In the second approach, we pass all indices. In our previous research, the best results were obtained when we used 8 or more noncollinear indices along with concentration as training data. This is not an obligatory requirement as there may be indirect connections, but the presence of such indices can be an indicator of successful modeling. Currently, all HMs have indices with sufficient Spearman or Pierson correlations, but only 9 of them have 8 or more noncollinear indices: Al, Fe, Sb, Na, Sc, Sm, Tb, Th, and U. We focus on these elements as they seem to be highly promising for modeling.

The selected indices are prepared for Al, Fe, and Sb. They are of particular interest for experts in air pollution monitoring. The experts in our case are three specialists from the Frank Laboratory of Neutron Physics (JINR), with great experience in air pollution monitoring in Russia and Europe. The selected indices and their correlation for Al, Sb and Fe are presented in App. 2.

Twelve training datasets (3 with the selected indices, and 9 with all indices) are compiled. After that, we prepare indices for a prediction of 3.000 grid nodes on an area of 500 km^2 in the Central Russia region for 2019 and 2020. Indices for 12.100 grid nodes in the same area are also computed to provide detailed information about zones of interest. GB, MLP, and SNN models are then trained on the data. For all models, we use the 80/20 train /test split. Tab. 1 gives the mean accuracies for 10 runs of Al, Sb, and Fe models, trained on the selected indices and all indices, respectively, except for those of Na, Sc, Sm, Tb, W, Th, and U models, trained on all indices.

Tab. 1. Mean accuracy of the models. GB is gradient boosting. MLP is the multilayer perceptron. SNN is the Siamese neural network. Acc Si is the accuracy on the selected indices. Acc Al is the accuracy on all indices

	Al		Fe		Sb		Na	Sc	Sm	Tb	Th	U
	Acc si	Acc ai	Acc si	Acc ai	Acc si	Acc ai	Acc si	Acc si	Acc si	Acc si	Acc si	Acc si
GB	0.91	0.92	0.92	0.93	0.94	0.94	0.94	0.93	0.92	0.93	0.93	0.92
MLP	0.89	0.91	0.92	0.92	0.89	0.92	0.92	0.92	0.92	0.92	0.91	0.90
SNN	0.92	0.93	0.93	0.93	0.93	0.94	0.93	0.94	0.93	0.93	0.93	0.93

SNN training has two parts. First, we train the Siamese network on pairs of samples. The mean accuracy of similarity determination for Al, Sb, and Fe on the selected indices amounts to 0.86, 0.84, and 0.85, respectively. It shows that the network is good at comparing objects of different classes. Then we use one of the trained twins with the frozen weights as a feature extractor and the one-layer perceptron as a classifier. The analysis of the confusion matrix shows that in most cases the models flounder on low-level contamination classes. Training on all indices shows slightly better accuracy and simplifies the process. It is good for universality; we do not need to specifically search for noncollinear indices with a sufficient connection with contamination. After expert verification, it turns out that the prediction of such a model is less plausible than the prediction of the model trained on the selected indices. Fig. 5 illustrates an example of the output of SNN models for Sb trained on the selected (B) and all (C) indices.

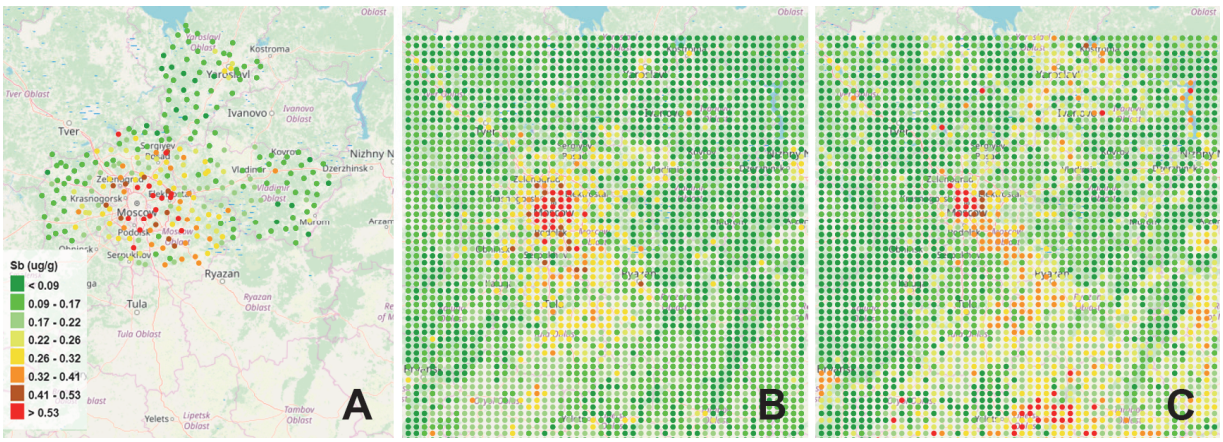

Fig. 5. Sb contamination in the Central Russia region in 2019. A shows the in situ data. B shows the output of the SNN model trained on the selected indices. C shows the output of the SNN model trained on all indices

In Fig. 5C, one can see clusters of yellow and red points at the bottom and in the top right corner of the map. It is known that the contamination level is not high over these areas, while the contamination level near the border of the Moscow region is moderate.

The models trained on all indices are likely to be more imbalanced since some collinear indices dominate the others. Such a situation is observed for GB, but also affects MLP and SNN models. The accuracy of the models trained on all data for Na, Sc, Sm, Tb, Th, and U leads to a reasonable expectation of good results, as obtained for them with the selected indices. The process of index selection can be automated for ease of working. Since we have resource limitations, we focus only on the models trained on the selected indices for Al, Sb, and Fe.

All the models show good accuracy, and we have to identify some way to find the best one. We cannot rely only on statistical metrics, as the prediction results can vary from run to run because of random selection in train-test splits. GB is less vulnerable from this point of view than the MLP and SNN due to fewer stochastic processes in training procedures. Models equal to statistical parameters can show different results in some regions. Here, while general tendencies will be the same, details can vary.

As the next level of model verification, a comparison of some hot spots is proposed. The experts specify six-ten points for each element on the map where contamination cannot be higher or lower than certain specified levels. For example, we have one of 3.000 points in the southern part of Yaroslavl. There is a working oil refinery here, and therefore, we can infer that the contamination level at this point cannot be too low. The same principle applies to huge transport nodes. Conversely, contamination in national parks or forest reserves cannot be too high. After determining the hot-spot list, a cycle procedure of model retraining is run. On each iteration, we make a prediction and check if the predicted values in the hot spots are consistent with the expectations. Several validated models are selected for further analysis. A blind test is used; therefore, the experts do not know about the model used to make a prediction. Twelve Sb contamination maps are passed to the experts; 4 maps for each approach, namely, GB, MLP, SNN. Based on their experience, the experts rank the models using their judgment. Unfortunately, we cannot determine the best modeling approach, as representatives of each of them appear at the top of the list. We can only confirm that MLP models appear at the top of the list less often, neural model prediction is less positive than GB, and it takes fewer iterations to get a prediction that fits the hot-spot check with neural models.

The best model can only be identified by collecting and examining samples on modeled grid points. Unfortunately, we cannot conduct such a kind of research. We try to solve the task by using a model trained on the Vladimir and Yaroslavl data to predict contamination in the Moscow region. This idea also does not work as GB, the MLP and SNN show a fairly comparable accuracy ranging from 54 to 58 %. We believe it is a good result, given the huge difference in contamination levels in the regions. Since we cannot identify the best model, we choose to present the outputs of all the three models and in situ data in Fig. 6.

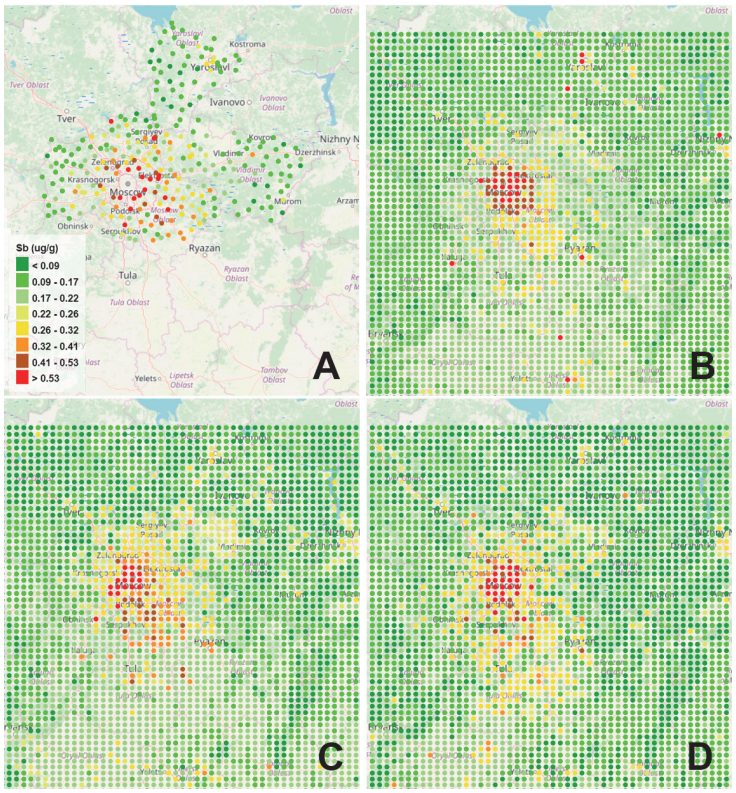

Fig. 6. Sb contamination in the Central Russia region in 2019. A shows the in situ data, B shows the GB prediction, C shows the MLP prediction, D shows the SNN prediction

Considering that we have limitations in the volume of the research, we focus only on the SNN model, as we truly believe in the potential of this approach. To demonstrate the abilities of our method, we use the best model to get a prediction on the indices calculated for 12.100 grid nodes to obtain detailed information on the Central Russia region (Fig. 7).

Moscow is densely populated, and its population grows fast. Published information reveals, there are about 12.5 million habitants in Moscow. Therefore, the Sb contamination level there is bound to be very high. The map also displays clusters of hot spots in large cities, such as Tula, Kaluga, Vladimir, Tver, Nizhny Novgorod, Yaroslavl, etc. It is also seen that from Sergiyev Posad to the north direction, the contamination level is rather low, except for Yaroslavl, where the already mentioned working oil refinery is located.

The Tula region stands out on the map. There is a lot of industry located in the region, i.e., chemical, metallurgical, and machine-building, in addition to several large thermal power plants. In terms of the concentration of such enterprises per unit area, the Tula region is second only to the Moscow region. Huge transport nodes and federal freeways are seen, rather clearly, on the map.

As an experiment, we use the SNN model to get a prediction for Sb contamination of 2020 and compare it with the 2019 results (Fig. 8).

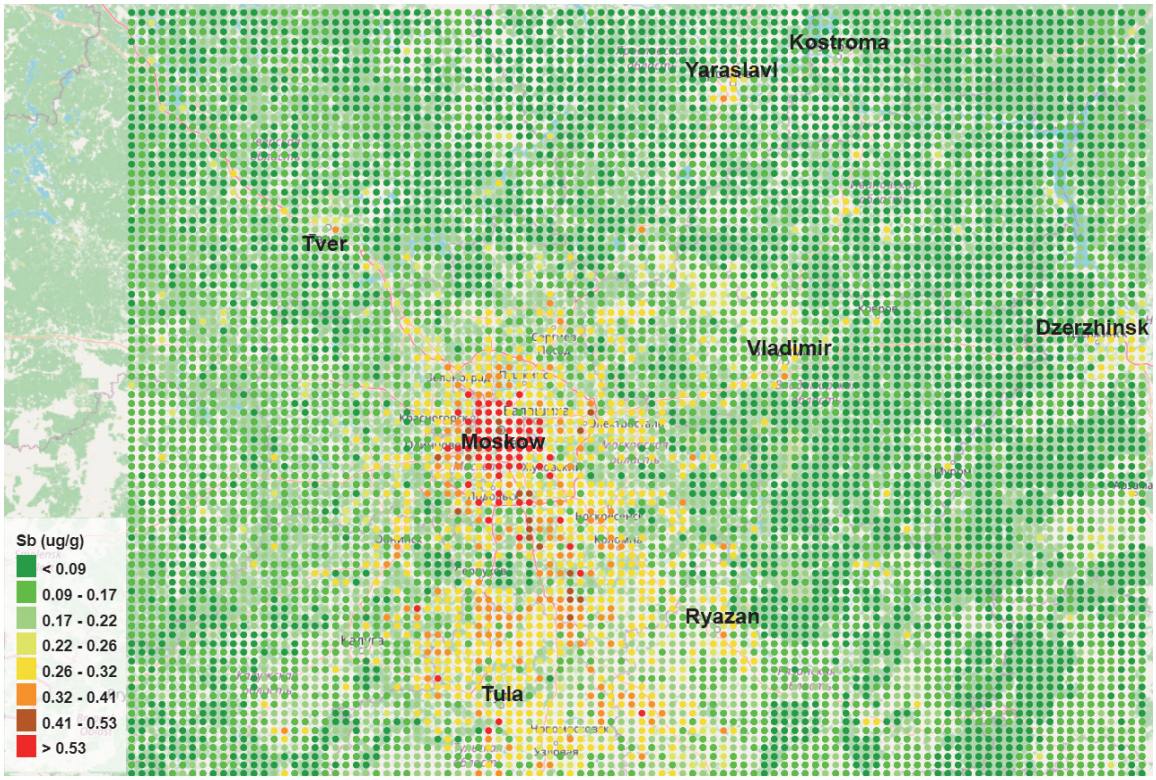

Fig. 7. High spatial resolution of the SNN model prediction of Sb contamination

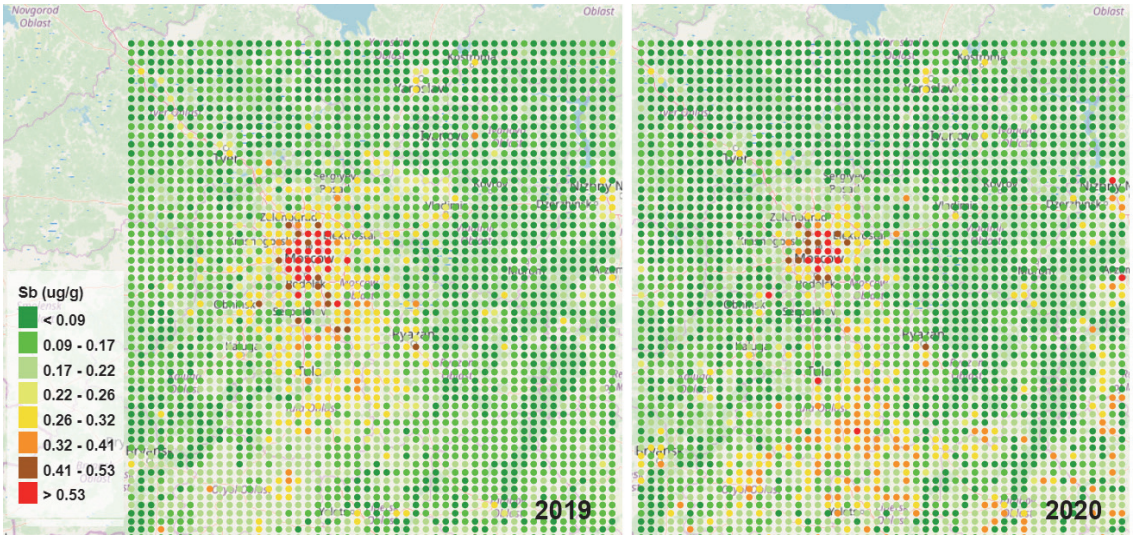

Fig. 8. Sb contamination prediction of the SNN model for 2019 (left) and 2020 (right)

The lockdown in Russia, which lasted approximately 1.5 months, imposed different limitations. Most of the limitations restricted the movement activities of the population. According to the official statistics, by the end of 2020, industrial production in Russia decreased by 2.9 % from the past. While there is a general decrease in the contamination level, high levels persist in areas with active production or high population density. The exceptions to the general decrease in contamination are areas near Nizhny Novgorod and Arzamas. A plausible explanation for this is that the Nizhny Novgorod region is a well-known place for attracting tourists. In 2020, this region appeared at the top of the list of destinations attracting tourists from different regions. Arzamas – Diveevo – Sarov is a well-known tourist pilgrimage route.

Contrary to the general decrease in contamination, there is an increase in contamination in the Lipetsk region. We have no reasonable explanation and cannot find any clue in the official records. Probably it is a model error, but we cannot ignore a real increase in contamination. Proceeding further, it would be appropriate to analyze changes in the correlation of indices with HMs from year to year to select the best indices. Unfortunately, ICP Vegetation surveys are held only once in five years. We only have information on the Moscow and Tver regions for 2014 – 2015, however, as some satellite programs had not started by that time, the information available is incomplete. We believe that the best practice for future research is annual sampling and the determination of satellite indices, which continue to have a good connection with contamination over the years. As a next step, to show the applicability of the suggested approach, the SNN model is trained to get a prediction for 2019 and 2020 for Al (Fig. 9) and Fe.

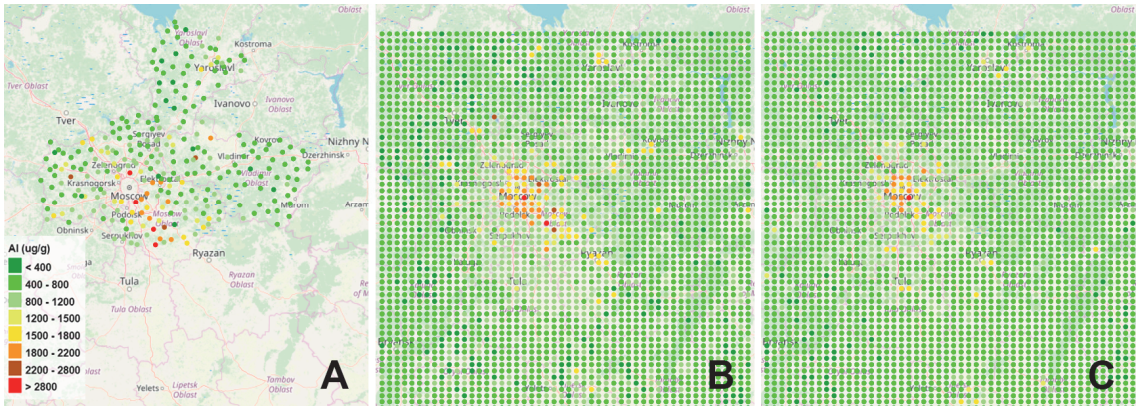

Fig. 9. Al contamination in the Central Russia region. A shows the in situ data. B shows the output of the SNN model trained on the selected indices for 2019. C shows the output of the SNN model for 2020

In developed societies, aluminum is the most widely used metal after steel and its derivatives. After steel, aluminum is the most produced metal, and the most produced non-ferrous metal [56]. Al contamination is mostly connected with production and less related to traffic [57]. There is some decrease in contamination in the model prediction for 2020. However, in most hot spots close to Al-production factories, the contamination level stays high.

Iron is an essential element in the blood pigment that helps transport oxygen to all parts of the body; however, its excessive intake poses a risk to human health [58]. Fe contamination is associated with production or natural sources [59, 60]. Thus, like Al, Fe is a good candidate for research.

There is no significant decrease in contamination in the model prediction for 2020 (Fig. 10). One can even see some concentration of contamination around Moscow. The lockdown did not affect Fe-industrial processes as significantly as the traffic activity represented by Sb.

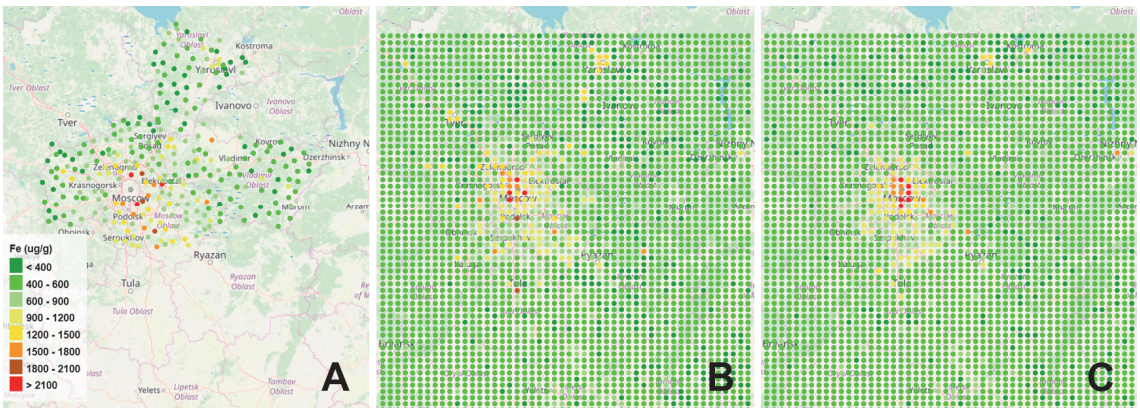

Fig. 10. Fe contamination in the Central Russia region in 2019. A shows the in situ data. B shows the output of the SNN model trained on the selected indices for 2019. C shows the output of the SNN model for 2020

3. Discussion

Obviously, it is better to have a sample grid covering all areas of interest, to execute sampling every year, and have enough samples to split data for training, testing, and validation. However, even with the limits in training data, we try to show the advantages and prospects of using satellite data together with machine learning for HM contamination prediction. With the help of the method, we can monitor and evaluate the situation when needed, get detailed information about areas of interest, check the situation in the areas where sampling is forbidden, and partly automate the environment control process.

Unfortunately, we cannot determine which architecture (GB, MLP, or SNN) is better for this. We believe that the Siamese network is more versatile as there is a lot of direction for evaluation. In our future research, we are going to examine other loss functions and training procedures. We see the direction of future research in the improvement of calculation techniques of satellite indices, the examination of new modeling approaches, and the extension and automation of our pipeline.

Modern satellite programs such as Sentinel-5 provide a great deal of data. A high-resolution spectrometer onboard Sentinel-5 operates in the ultraviolet to shortwave infrared range with 7 different spectral bands ranging from 270 to 2,385 nm. We believe that data of such programs, together with an advanced neural architecture, can broaden the horizon of environmental monitoring and contribute to improving the environmental performance in the world.

Nevertheless, satellite programs are not the only possible source of additional data. Meteorological and topological data can also be used for modeling. We believe that many interesting results can be obtained using Big Data technologies, and the further identification of potential contamination sources can seriously improve the accuracy of models.

Conclusion

The models for air contamination by heavy metals were designed using in situ data and satellite imagery. The concentrations of elements from 281 samples of naturally growing mosses, chosen in the Vladimir, Yaroslavl, and Moscow regions, were used as training data. Indices from satellite images gathered by the Google Earth Engine platform, which represent summarized information about sampling sites, were used as additional data for training. We focused on the classification task with 8 levels of contamination and used balancing techniques to extend the training data. Gradient boosting, Multi-layer perceptron, and Siamese neural network approaches were examined. Any of the approaches could be called better than the other two. The median accuracy of the models for 9 heavy metals (Al, Fe, Sb, Na, Sc, Sm, Tb, Th, U) exceeded 89 %. Al, Fe, and Sb contamination of 3,000 and 12,100 grid nodes on a 500 km^2 area in the Central Russia region for 2019 and 2020 was modeled. The output model was accepted by scientists involved in ecology monitoring over Central Russia. The potential of using satellite data together with machine learning to predict contamination was demonstrated.

References

[1] Uzhinskiy A. Intelligent environmental monitoring platform. CEUR Workshop Proc 2019; 2267: 351-358.

[2] Harmens H, Norris DA, Steinnes E, Kubin E, Piispanen J, Alber R, Aleksiayenak Y, Blum O, Coskun M, Dam M, De Temmerman L, Fernández JA, Frolova M, Frontasyeva M, González-Miqueo L, Grodzinska K, Jeran Z, Korzekwa S, Krmar M, Kvietkus K, Leblond S, Liiv S, Magnússon SH, Mankovská B, Pesch R, Rühling Å, Santamaria JM, Schröder W, Spiric Z, Suchara I, Thöni L, Urumov V, Yurukova L, Zechmeister HG. Mosses as biomonitors of atmospheric heavy metal deposition: spatial patterns and temporal trends in Europe. Environ Pollut 2010; 158: 3144-3156.

[3] Yuan Y, Wu Y, Ge X, Nie D, Wang M, Zhou H, Chen M. In vitro toxicity evaluation of heavy metals in urban air particulate matter on human lung epithelial cells. Sci Total Environ 2019; 678: 301-308.

[4] Jakubowski M. Biological monitoring versus air monitoring strategies in assessing environmental-occupational exposure. J Environ Monit 2012; 14(2): 348-352. DOI: 10.1039/c1em10706b.

[5] Holt EA, Miller SW. Bioindicators: Using organisms to measure environmental impacts. Nature Education Knowledge 2010; 3(10): 8.

[6] Uzhinskiy A, Urošević MA, Frontasyeva M. Prediction of air pollution by potentially toxic elements over urban area by combining satellite imagery, moss biomonitoring data and machine learning. Cienc e Tec Vitivinic J 2020; 35(12): 34-46.

[7] Uzhinskiy A, Ososkov G, Goncharov P, Frontsyeva M. Combining satellite imagery and machine learning to predict atmospheric heavy metal contamination. CEUR Workshop Proc 2018; 2267: 351-358.

[8] Salo H, Mäkinen J. Magnetic biomonitoring by moss bags for industry-derived air pollution in SW Finland. Atmos Environ 2014: 97: 19-27. DOI: 10.1016/j.atmosenv.2014.08.003.

[9] Ben-Dor E, Irons JR, Epema GF. Soil reflectance. In Book: Rencz AN, Ryerson RA, eds. 3rd ed, Vol 3. Manual of remote sensing: Remote sensing for the earth sciences. New York: John Wiley & Sons Inc; 1999: 111-189.

[10] Kemper T, Sommer S. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. Environ Sci Technol 2002; 36(12): 2742-2747.

[11] Choe E, van der Meer F, van Ruitenbeek F, van der Werff H, de Smeth B, Kim KW. Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: a case study of the Rodalquilar mining area, SE Spain. Remote Sens Environ 2008; 112(7): 3222-3233.

[12] Ren H-Y, Zhuang D-F, Singh AN, Pan J-J, Qui D-S, Shi R-H. Estimation of As and Cu contamination in agricultural soils around a mining area by reflectance spectroscopy: a case study. Pedosphere 2009; 19(6): 719-726.

[13] Beloconi A, Chrysoulakis N, Lyapustin A, Utzinger J, Vounatsou P., Bayesian geostatistical modelling of PM10

and PM2.5 surface level concentrations in Europe using high-resolution satellite-derived products. Environ Int 2018; 121(1): 57-70. DOI: 10.1016/j.envint.2018.08.041.

[14] Alvarez-Mendoza CI, Teodoro AC, Torres N, Vivanco V. Assessment of remote sensing data to model PM10 estimation in cities with a low number of air quality stations: A case of study in Quito, Ecuador. Environments 2019; 6: 85. DOI: 10.3390/environments6070085.

[15] Zheng T, Bergin MH, Hu S, Miller J, Carlson DE. Estimating ground-level PM2.5 using micro-satellite images by a convolutional neural network and random forest approach. Atmospheric Environ 2020; 230: 117451. DOI: 10.1016/j.atmosenv.2020.117451.

[16] Muradyan V, Tepanosyan G, Asmaryan S, Maghakyan N, Sahakyan L, Saghatelyan A. Estimating Mo, Cu, Ni, Cd contents in the crop leaves growing on small land plots using satellite data. Commun Soil Sci Plant Anal 2020; 51(11): 1457-1468. DOI: 10.1080/00103624.2020.1784922.

[17] Liu M, Wang T, Skidmore AK, Liu X. Heavy metal-induced stress in rice crops detected using multi-temporal Sentinel-2 satellite images. Sci Total Environ 2018; 637-638: 18-29. DOI: 10.1016/j.scitotenv.2018.04.415.

[18] Amer M, Tyler A, Foudat T, Hunter P, Elmetwalli A, Wilson C, Vallejo-Marin M. Spectral characteristics for estimation heavy metals accumulation in wheat plants and grain. J Product Dev 2017; 22(3): 409-428.

[19] Zhou C, Chen S, Zhang Y, Zhao J, Song D, Liu D. Evaluating metal effects on the reflectance spectra of plant leaves during different seasons in post-mining areas, China. Remote Sens 2018; 10(8): 1211. DOI: 10.3390/rs10081211.

[20] Yu K, Van Geel M, Ceulemans T, Geerts W, Ramos MM, Serafim C, Sousa N, Castro PML, Kastendeuch P, Najjar G, Ameglio T, Ngao J, Saudreau M, Honnay O, Somers B. Vegetation reflectance spectroscopy for biomonitoring of heavy metal pollution in urban soils. Environ Pollut 2018; 243(B): 1912-1922. DOI: 10.1016/j.envpol.2018.09.053.

[21] Bjerke JW, Tømmervik H, Finne TE, Jensen H, Lukina N, Bakkestuen V. Epiphytic lichen distribution and plant leaf heavy metal concentrations in the Russian–Norwegian boreal forests influenced by air pollution from nickel–copper smelters. Boreal Env Res 2006; 11: 441-450.

[22] Khosropour E, Attarod P, Shirvany A, et al. Response of Platanusorientalis leaves to urban pollution by heavy metals. J For Res 2019; 30: 1437-1445. DOI: 10.1007/s11676-018-0692-8.

[23] Alahabadi A, Ehrampoush MH, Miri M, Ebrahimi Aval H, Yousefzadeh S, Ghaffari HR, Ahmadi E, Talebi P, AbaszadehFathabadi Z, Babai F, Nikoonahad A, SharafiK, Hosseini-Bandegharaei A. A comparative study on capability of different tree species in accumulating heavy metals from soil and ambient air. Chemosphere 2017; 172: 459-467. DOI: 10.1016/j.chemosphere.2017.01.045.

[24] Terekhina NV, Ufimtseva MD. Leaves of trees and shrubs as bioindicators of air pollution by particulate matter in Saint Petersburg. Geogr Environ Sustain 2020; 13(1): 224-232. DOI: 10.24057/2071-9388-2019-65.

[25] Lyanguzova I, Yarmishko V, Gorshkov V, Stavrova NN, Bakkal I. Impact of heavy metals on forest ecosystems of the European North of Russia. In: Saleh HEM, Aglan RF, eds. Heavy metals. London: IntechOpen; 2018. DOI: 10.5772/intechopen.73323.

[26] Lassalle G, Fabre S, Credoz A, et al. Mapping leaf metal content over industrial brownfields using airborne hyperspectral imaging and optimized vegetation indices. Sci Rep 2021; 11: 2. DOI: 10.1038/s41598-020-79439-z.

[27] Liu Z, Lu Y, Peng Y, Zhao L, Wang G, Hu Y. Estimation of Soil heavy metal content using hyperspectral data. Remote Sens 2019; 11(12): 1464.

[28] Gholizadeh A, Saberioon M, Ben-Dor E, Borůvka L. Monitoring of selected soil contaminants using proximal and remote sensing techniques: Background, state-of-the-art and future perspectives. Crit Rev Environ Sci Technol 2018; 48(3): 243-278.

[29] Gholizadeh A, Coblinski JA, Saberioon M, Ben-Dor E, Drábek O, Demattê JAM, Borůvka L, Němeček K, Chabrillat S, Dajčl J. vis-NIR and XRF data fusion and feature selection to estimate potentially toxic elements in soil. Sensors 2021; 21(7): 2386. DOI: 10.3390/s21072386.

[30] Ahado SK, Nwaogu C, Sarkodie VYO, Borůvka L. Modeling and assessing the spatial and vertical distributions of potentially toxic elements in soil and how the concentrations differ. Toxics 2021; 9(8): 181. DOI: 10.3390/toxics9080181.

[31] Michaelides S, Paronis D, Retalis A, Tymvios F. Monitoring and forecasting air pollution levels by exploiting satellite, groundbased, and synoptic data, elaborated with regression models. Adv Meteorol 2017; 2017: 2954010.

[32] Foldi C, Sauermann S, Dohrmann R, Mansfeldt T. Traffic-related distribution of antimony in roadside soils. Environ Pollut 2018; 237: 704-712.

[33] Goddard SL, Williams KR, Robins C, et al. Determination of antimony and barium in UK air quality samples as indicators of nonexhaust traffic emissions. Environ Monit Assess 2019; 191(11): 641.

[34] Fang Y, Xu L, Peng J, Wang H, Wong A, Clausi DA. Retrieval and mapping of heavy metal concentration in soil using time series Landsat 8 imagery. ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2018; XLII-3: 335-340.

[35] Xu X, Chen S, Ren L, Han C, Lv D, Zhang Y, Ai F. Estimation of heavy metals in agricultural soils using vis-NIR spectroscopy with fractional-order derivative and generalized regression neural network. Remote Sens 2021; 13: 2718. DOI: 10.3390/rs13142718.

[36] Pyo JC, Hong SM, Kwon YS, Kim MS, Cho KH. Estimation of heavy metals using deep neural network with visible and infrared spectroscopy of soil. Sci Total Environ 2020; 741: 140162. DOI: 10.1016/j.scitotenv.2020.140162.

[37] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. arXiv Preprint. 2015. Source: ⟨https://arxiv.org/abs/1503.03832⟩.

[38] Cheng D, Gong Y, Zhou S, Wang J, Zheng N. Person re-identification by multi-channelparts-based CNN with improved triplet loss function. 2015 IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2015: 1335-1344.

[39] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person reidentification. arXiv Preprint. 2017. Source: ⟨https://arxiv.org/abs/1703.07737⟩.

[40] Dong X, Shen J. Triplet loss in Siamese network for object tracking. In Book: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. Computer Vision – ECCV 2018. Cham: Springer Nature Switzerland AG; 2018: 472-488.

[41] Puch S, Sánchez I, Rowe M. Few-shot learning with deep triplet networks for brain imaging modality recognition. In Book: Wang Q, Milletari F, Nguyen HV, Albarqouni S, Cardoso MJ, Rieke N, Xu Z, Kamnitsas K, Patel V, Roysam B, Jiang S, Zhou K, Luu K, Le N, eds. Domain adaptation and representation transfer and medical image learning with less labels and imperfect data. Cham: Springer Nature Switzerland AG; 2019: 181-189.

[42] Anshul T, Daksh T, Padmanabhan R, Aditya N. Deep metric learning for bioacoustic classification: Overcoming

training data scarcity using dynamic triplet loss. J Acoust Soc Am 2019; 146: 534-547.
[43] Zhang J, Lu C, Wang J, Yue X, Lim S, Al-Makhadmeh Z, Tolba A. Training convolutional neural networks with multi-size images and triplet loss for remote sensing scene classification. Sensors 2020; 20(4): 1188.
[44] Rühling A, Tyler G. An ecological approach to the lead problem. Botaniska Notiser 1968; 121: 321-342.
[45] Markert BA, Breure AM, Zechmeister HG. Definitions, strategies and principles for bioindications/biomonitoring of the environment, In Book: Markert BA, Breure AM, Zechmeister HG, eds. Bioindicators & biomonitors: Principles, concepts and applications. Vol 6. Pergamon; 2003: 3-39.
[46] Frontasyeva MV. Neutron activation analysis in the life sciences. Phys Part Nucl 2011; 42: 332-378. DOI: 10.1134/S1063779611020043.
[47] CLRTAP. Manual on methodologies and criteria for modeling and mapping critical loads and levels and air pollution effects, risks and trends. UNECE Convention on Long-range Transboundary Air Pollution; 2015. Source: ⟨https://icpvegetation.ceh.ac.uk⟩.
[48] Yin F, Lewis PE, Gomez-Dans J, Wu Q. A sensor-invariant atmospheric correction method: application to Sentinel-2/MSI and Landsat 8/OLI. EarthArXiv Preprint. 2019. Source: ⟨https://eartharxiv.org/repository/view/1034/⟩. DOI: 10.31223/osf.io/ps957.
[49] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001; 29(5): 1189-1232.
[50] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE Int Joint Conf on Neural Networks 2008: 1322-1328. DOI: 10.1109/IJCNN.2008.4633969.
[51] Chen C, Liaw A, Breiman L, et al. Using random forest to learn imbalanced data. Berkeley: University of California; 2004.
[52] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. KDD '16: Proc 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining 2016: 785-794. DOI: 10.1145/2939672.2939785.
[53] Hertz J, Krogh A, Palmer RG. Introduction to the theory of neural computation. 1st ed. CRC Press; 1991. DOI: 10.1201/9780429499661.
[54] Nair V, Hinton G. Rectified linear units improve restricted boltzmann machines. Proc 27th Int Conf on Machine Learning 2010: 807-814.
[55] Kingma DP, Ba J. Adam: A method for stochastic optimization. 3rd Int Conf on Learning Representations (ICLR). 2015. Source: ⟨https://arxiv.org/abs/1412.6980⟩.
[56] Abdollahi J, Emrani N, Chahkandi B, at al. Environmental impact assessment of aluminium production using the life cycleassessment tool and multi-criteria analysis. Ann Environ Sci Toxicol 2021; 5(1): 059-066. DOI: 10.17352/aest.000038.
[57] Paraskevas D, Kellens K, Van de Voorde A, Dewulf W, Duflou JR. Environmental impact analysis of primary aluminium production at country level. Procedia CIRP 2016; 40: 209-213.
[58] Nkansah MA, Agorsor PI, Opoku F. Heavy metal contamination and health risk assessment of mechanically milled delicacy called fufu. Int J Food Contam 2021; 8: 6. DOI: 10.1186/s40550-021-00085-y.
[59] Zhang X, Gao S, Fu Q, Han D, Chen X, Fu S, Huang X, Cheng J. Impact of VOCs emission from iron and steel industry on regional O3 and PM2.5 pollutions. Environ Sci Pollut Res 2020; 27: 28853-28866. DOI: 10.1007/s11356-020-09218-w.
[60] Wang R, Balkanski Y, Boucher O, Bopp L, Chappell A, Ciais P, Hauglustaine D, Peñuelas J, Tao S. Sources, transport and deposition of iron in the global atmosphere. Atmospheric Chem Phys 2015; 15: 6247-6270. DOI: 10.5194/acp-15-6247-2015.

Appendix 1

Tab. 1. List of the used collections

1	Name: USGS Landsat 7 Collection 1 Tier 1 Raw Scenes Description: Landsat 7 Collection 1 Tier 1 DN values, representing scaled, calibrated at-sensor radiance. Spatial resolution: 15 – 60 m Number of meaningful bands (channels): 9
2	Name: USGS Landsat 8 Collection 1 Tier 1 Raw Scenes Description: Landsat 8 Collection 1 Tier 1 DN values, representing scaled, calibrated at-sensor radiance. Spatial resolution: 15 – 30 m Number of meaningful bands: 11
3	Name: MOD11A2.006 Terra Land Surface Temperature and Emissivity 8-Day Global 1km Description: The MOD11A2 V6 product provides an average 8-day land surface temperature (LST) in a 1200 x 1200 kilometer grid. Each pixel value in MOD11A2 is a simple average of all the corresponding MOD11A1 LST pixels collected within this 8-day period. Spatial resolution: 1000 m Number of meaningful bands: 2
4	Name: VIIRS Stray Light Corrected Nighttime Day/Night Band Composites Version 1 Description: Monthly average radiance composite images using nighttime data from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day/ Night Band (DNB). Spatial resolution: 463 m Number of meaningful bands: 1
5	Name: PROBA-V C1 Top Of Canopy Daily Synthesis 333m Description: Proba-V is a satellite mission tasked to map land cover and vegetation growth. It was designed to provide continuity for the VGT optical instrument from the SPOT-4 and SPOT-5 missions. Spatial resolution: 333 m Number of meaningful bands: 5

Tab. 1. List of the used collections (continuation)

6	Name: TerraClimate: Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces, University of Idaho Description: TerraClimate is a dataset of monthly climate and climatic water balance for global terrestrial surfaces. It uses climatically aided interpolation, combining high-spatial resolution climatological normals from the WorldClim dataset, with coarser spatial resolution, but time-varying data from CRU Ts4.0 and the Japanese 55-year Reanalysis (JRA55). Spatial resolution: 4638 m Number of meaningful bands: 12
7	Name: MOD09A1.006 Terra Surface Reflectance 8-Day Global 500m Description: The MOD09A1 V6 product provides an estimate of the surface spectral reflectance of Terra MODIS bands 1-7 at 500m resolution, which is corrected for atmospheric conditions such as gasses, aerosols, and Rayleigh scattering. Spatial resolution: 500 m Number of meaningful bands: 7
8	Name: NOAA CDR AVHRR: Surface Reflectance, Version 5 Description: The NOAA Climate Data Record (CDR) of AVHRR Surface Reflectance contains gridded daily surface reflectance and brightness temperatures derived from Advanced Very High Resolution Radiometer (AVHRR) sensors onboard seven NOAA polar orbiting satellites. The data are gridded at a resolution of 0.05° and computed globally over land surfaces. Spatial resolution: 5566 m Number of meaningful bands: 6
9	Name: MOD13A1.006 Terra Vegetation Indices 16-Day Global 500m Description: The MOD13A1 V6 product provides the Vegetation Index (VI) value at a per pixel basis. Spatial resolution: 500 m Number of meaningful bands: 2
10	Name: Sentinel-2 MSI: MultiSpectral Instrument, Level-1C Description: Sentinel-2 is a wide-swath, high-resolution, multi-spectral imaging mission supporting Copernicus Land Monitoring studies, including the monitoring of vegetation, soil and water cover, as well as the observation of inland waterways and coastal areas. Spatial resolution: 10 – 60 m Number of meaningful bands: 13
11	Name: Sentinel-5P OFFL CO: Offline Carbon Monoxide Description: This dataset provides offline high-resolution imagery of CO concentrations. Spatial resolution: 1113 m Number of meaningful bands: 2
12	Name: Sentinel-5P OFFL NO2: Offline Nitrogen Dioxide Description: This dataset provides offline high-resolution imagery of NO2 concentrations. Spatial resolution: 1113 m Number of meaningful bands: 6
13	Name: Sentinel-5P OFFL AER AI: Offline UV Aerosol Index Description: This dataset provides offline high-resolution imagery of the UV Aerosol Index (UVAI), also called the Absorbing Aerosol Index (AAI). Spatial resolution: 1113 m Number of meaningful bands: 1

Appendix 2

Tab. 1. Indices selected for contamination and their connection with AL concentrations at sampling sites (rs – Spearman correlation coefficient, rp – Pearson correlation coefficient)

Collection	Band	Area	All regions		Moscow r.		Yaroslavl r.		Vladimir r.	
			rs	rp	rs	rp	rs	rp	rs	rp
USGS Landsat 7 Collection 1 Tier 1 Raw	B7 (2.08 – 2.35 μm)	4 km^2	0.39	0.33	0.2	0.21	0.3	0.26	0.26	0.16
VIIRS Stray Light Corrected Nighttime Day/Night Band Comp.	avg_rad	4 km^2	0.48	0.32	0.33	0.24	0.48	0.44	0.18	0.06
PROBA-V C1 Top Of Canopy Daily Synthesis	SWIR (1610 nm, FWHM: 89 nm)	4 km^2	0.4	0.33	0.3	0.15	0.21	0.15	0.32	0.28
TerraClimate: Monthly Climate and Climatic Water Balance	srad	4 km^2	0.36	0.36	0.34	0.34	0.15	0.08	– 0.14	– 0.26
MOD09A1.006 Terra Surface Reflectance 8-Day Global 500m	sur_refl_b03 (459 – 479 nm)	4 km^2	0.42	0.32	0.24	0.16	0.18	0.3	0.43	0.37

Tab. 1. Indices selected for contamination and their connection with AL concentrations at sampling sites (rs – Spearman correlation coefficient, rp – Pearson correlation coefficient) (continuation)

Collection	Band	Area	All regions		Moscow r.		Yaroslavl r.		Vladimir r.	
			rs	rp	rs	rp	rs	rp	rs	rp
NOAA CDR AVHRR: Surface Reflectance V. 5	SREFL_CH2 (860 nm)	4 km^2	0.47	0.43	0.25	0.26	0.29	0.28	0.18	0.46
MOD13A1.006 Terra Vegetation Indices 16-Day Global 500m	NDVI	4 km^2	–0.43	–0.41	–0.26	–0.29	–0.25	–0.26	–0.19	–0.29
Sentinel-2 MSI: Multi-Spectral Inst., L 1C	B10 (1373.5 – 1376.9 nm)	4 km^2	0.39	0.34	0.31	0.18	0.12	0.17	0.14	0.16
Sentinel-5P OFFL CO: Offline Carbon Monox.	CO_column number_density	1 km^2	0.44	0.37	0.15	0.17	0.35	0.39	–0.2	–0.2
Sentinel-5P OFFL NO2: Offline Nitrogen Diox.	NO2_column number_density	1 km^2	0.52	0.54	0.46	0.49	0.34	0.39	0.23	0.12
Sentinel-5P OFFL NO2: Offline Nitrogen Diox.	Absorbing aerosol index	1 km^2	–0.4	–0.31	–0.28	–0.11	0.19	0.13	0.14	0.25

Tab. 2. Indices selected for contamination and their connection with Sb concentrations at sampling sites (rs – Spearman correlation coefficient, rp – Pearson correlation coefficient)

Collection	Band	Area	All regions		Moscow r.		Yaroslavl r.		Vladimir r.	
			rs	rp	rs	rp	rs	rp	rs	rp
USGS Landsat 7 Collection 1 Tier 1 TOA Reflect	B6_VCID_2 (10.40 – 12.50 μm)	4 km^2	0.36	0.4	0.38	0.38	0.33	0.36	–0.13	0.14
VIIRS Stray Light Corrected Nighttime Day/Night Band Comp.	avg_rad	4 km^2	0.6	0.45	0.57	0.41	0.55	0.45	0.38	0.17
MOD11A2.006 Terra Land Surface Temperature and Emissivity	LST_Night_1 km	4 km^2	0.41	0.39	0.46	0.38	0.2	0.25	0.19	0.13
TerraClimate: Monthly Climate and Climatic Water Balance	srad	4 km^2	0.39	0.37	0.5	0.41	–0.25	–0.26	–0.32	–0.38
MOD09A1.006 Terra Surface Reflectance 8-Day Global 500m	sur_refl_b03 (459 – 479 nm)	4 km^2	0.38	0.32	0.25	0.29	0.29	0.39	0.52	0.44
NOAA CDR AVHRR: Surface Reflectance V5	SREFL_CH2 (860 nm)	4 km^2	0.43	0.45	0.27	0.34	0.26	0.29	0.32	0.51
MOD13A1.006 Terra Vegetation Indices 16-Day Global 500m	NDVI	4 km^2	–0.41	–0.43	–0.33	–0.36	–0.14	–0.23	–0.25	–0.31
Sentinel-5P OFFL CO: Offline Carbon Monox.	SO2_column number density_amf	1 km^2	0.48	0.46	–0.48	–0.46	0.25	0.19	0.11	0.15
Sentinel-5P OFFL NO2: Offline Nitrogen Diox.	tropospheric_NO2 column_number density	1 km^2	0.63	0.71	0.74	0.73	0.4	0.41	0.14	0.19
Sentinel-5P OFFL NO2: Offline Nitrogen Diox.	Absorbing aerosol index	1 km^2	–0.37	–0.34	–0.26	–0.17	0.15	0.03	0.3	0.31

Tab. 3. Indices selected for contamination and their connection with Fe concentrations at sampling sites (rs – Spearman correlation coefficient, rp – Pearson correlation coefficient)

Collection	Band	Area	All regions		Moscow r.		Yaroslavl r.		Vladimir r.	
			rs	rp	rs	rp	rs	rp	rs	rp
USGS Landsat 7 Collection 1 Tier 1 TOA Reflectance	B6_VCID_2 (10.40 – 12.50 μm)	4 km^2	0.36	0.4	0.38	0.38	0.33	0.36	–0.13	0.14
VIIRS Stray Light Corrected Nighttime Day/Night Band Comp	avg_rad	4 km^2	0.6	0.45	0.57	0.41	0.55	0.45	0.38	0.17

Tab. 3. Indices selected for contamination and their connection with Fe concentrations at sampling sites (rs – Spearman correlation coefficient, rp – Pearson correlation coefficient) (continuation)

Collection	Band	Area	All regions		Moscow r.		Yaroslavl r.		Vladimir r.	
			rs	rp	rs	rp	rs	rp	rs	rp
MOD11A2.006 Terra Land Surface Temperature and Emissivity	LST_Night_1 km	4 km^2	0.41	0.39	0.46	0.38	0.2	0.25	0.19	0.13
TerraClimate: Monthly Climate and Climatic Water Balance	srad	4 km^2	0.39	0.37	0.5	0.41	– 0.25	– 0.26	– 0.32	– 0.38
MOD09A1.006 Terra Surface Reflectance 8-Day Global 500m	sur_refl_b03 (459 – 479 nm)	4 km^2	0.38	0.32	0.25	0.29	0.29	0.39	0.52	0.44
NOAA CDR AVHRR: Surface Reflectance V5	SREFL_CH2 (860 nm)	4 km^2	0.43	0.45	0.27	0.34	0.26	0.29	0.32	0.51
MOD13A1.006 Terra Vegetation Indices 16-Day Global 500m	NDVI	4 km^2	– 0.41	– 0.43	– 0.33	– 0.36	– 0.14	– 0.23	– 0.25	– 0.31
Sentinel-5P OFFL CO: Offline Carbon Monoxide	SO2_column number density_amf	1 km^2	0.48	0.46	– 0.48	– 0.46	0.25	0.19	0.11	0.15
Sentinel-5P OFFL NO2: Offline Nitrogen Dioxide	tropospheric_NO2 column_number density	1 km^2	0.63	0.71	0.74	0.73	0.4	0.41	0.14	0.19
Sentinel-5P OFFL NO2: Offline Nitrogen Dioxide	Absorbing aerosol index	1 km^2	– 0.37	– 0.34	– 0.26	– 0.17	0.15	0.03	0.3	0.31

Authors' information

Alexander Vladimirovich Uzhinskiy, (b. 1983) Ph.D. in Technology, graduated from Dubna International University for Nature, Society, and Man in 2006. Works as a leading programmer at the Joint Institute for Nuclear Research. Research interests: neural networks, computer vision, software engineering, remote sensing.
E-mail: *auzhinskiy@jinr.ru* .

Konstantin Nikolaevich Vergel, (b. 1981) graduated from Dubna International University for Nature, Society, and Man in 2004. Works as a researcher at the Joint Institute for Nuclear Research. Research interests: biomonitoring, atmospheric pollution, environmental monitoring, metal bioaccumulation. E-mail: *verkn@mail.ru* .

Code of State Categories Scientific and Technical Information (in Russian – GRNTI)): 28.23.37, 87.15.03
Received – April 20, 2022. Final version – August 10, 2022.