INTELLIGENT ENVIRONMENTAL MONITORING PLATFORM

A.V. Uzhinskiy

Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia

E-mail: auzhinskiy@jinr.ru

Air pollution has a significant impact on human and environmental health. The aim of the UNECE International Cooperative Program (ICP) Vegetation within 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 develop an understanding of long-range transboundary pollution. The program is implemented in 43 countries of Europe and Asia. Mosses are collected at thousands of sites. The development of the data management system (DMS) for the ICP Vegetation program was initiated in 2016 at the Meshcheryakov Laboratory of Information Technologies of the Joint Institute for Nuclear Research. The DMS offers good options to simplify and automate the environmental monitoring process. We use several powerful technologies to provide a new level of service for ICP Vegetation participants. The platform has some worthwhile analytics, classification and prediction abilities. The current architecture, workflow, and principles of data processing and analysis will be presented.

Keywords: environmental monitoring, air pollution, data management, intelligent platform, intellectual data processing, machine learning, neural networks.

Alexander Uzhinskiy

1. Introduction

Air pollution is recognized as the fifth largest threat to human health [1]. According to the World Health Organization, 7 million people die from polluted air every year. Since the late 1960s in the USA and since 1970 in Europe, regulatory conventions and acts to protect health and the environment from air pollution have been signed. At present, there are many regional, local and global programs focused on the determination of air quality. In most cases, they use monitoring station networks or mobile measurement stations to provide information on regulatory air pollutants such as gaseous pollutants and particulate matter (PM). More detailed information on contamination can be obtained using more advanced techniques. During the last several decades, biomonitoring methods have been used to get data on heavy metals (such as antimony, mercury, lead, etc.), organic pollutants (benzo[a]pyrene), and radionuclides [2]. One of the projects aimed to identify the main polluted areas in Europe and produce regional maps is the UNECE International Cooperative Program (ICP) Vegetation. The program functions within the United Nations Convention on Long-Range Transboundary Air Pollution (CLRTAP). Transboundary pollution is a real threat clearly demonstrated by the sand from the Sahara desert lying on European streets. The UNECE ICP Vegetation program brings together researchers from 43 countries of Europe and Asia. Samples are collected at thousands of sites. In 2016, a Data Management System (DMS) of the UNECE ICP Vegetation was developed at the Meshcheryakov Laboratory of Information Technologies of the Joint Institute for Nuclear Research [3]. The DMS is designed to provide the ICP Vegetation community with a unified system for gathering, storing, analyzing, processing, and sharing biological monitoring data. The DMS can now considered to be an intelligent environmental monitoring platform. Since the first version of the platform presentation, a mobile application has been developed to simplify the process of collecting and verifying data. Deep learning models for image classification have been created. Statistical and neural models for pollution prediction based on remote sensing data have been elaborated. The analytical capabilities of the platform have been expanded.

2. Intelligent platforms

The level of automation and adoption of modern information technologies in environmental monitoring programs is constantly increasing, although it lags far behind areas where the use of advanced technology can lead to a rapid economic impact. Nevertheless, over the past decade, various powerful technologies have been used in environmental pollution control projects, which makes it possible to provide a new level of service, as well as the quality and speed of researches. Now we can talk about intelligent platforms capable of generating new knowledge based on incoming and available data and, in some cases, making decisions that previously required the competence of an expert. Here is a short list of such technologies: The Internet of things (IoT) specifies the principles of connecting and exchanging data between physical objects that are embedded with sensors and other objects, programs, and systems. Many platforms use IoT technologies to organize sensor networks and process environmental monitoring data. It allows one to minimize the number of errors, automate routine processes and speed up data-gathering routines. Big Data is a field that treats ways to analyze, systematically extract information, or otherwise, deal with datasets that are too large or complex to handle using traditional data-processing application software. In the case of environmental monitoring, the data we have to work with can be both large if we deal with a huge sensor network and complex if we deal with sampling sites' metadata. Artificial intelligence (AI) is a wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. In environmental monitoring, there are always operations demanding an expert opinion. AI technologies can execute primary analysis and save the expert’s time. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Both classification and prediction tasks of ML are highly useful for environmental monitoring. Many other technologies such as robotics, remote sensing, drones, etc. can also be mentioned. The general idea is that the use of one or more of these technologies can enhance the abilities of the digital platform and allows it to perform intelligent functions.

3. Data & Workflow

ICP Vegetation participants collect moss samples at thousands of sites. They should record the metainformation about sampling sites required by the UNECE ICP Vegetation manual. This data is used in the interpretation stage of research. We had to develop a mobile application to simplify metadata management. Now latitude, longitude, and altitude are set automatically at sampling sites, and most of the required parameters are implemented as lists. The mobile application allows getting visual information, for example, pictures of samples and of sampling sites. Such images can be used to verify the correctness of the input data. The given approach dramatically reduces the number of errors in metadata. Once collected, the samples are processed using different techniques, such as neutron activation analysis, to determine the concentrations of heavy metals, persistent organic pollutants, nitrogen, or radionuclides. Each sampling site has a unique ID that is used to import information on concentrations. Participants can manipulate data, create different kinds of maps, run prediction tasks, and get analytical reports on the platform. With simple statistical reports and geo indexes, it is also possible to carry out cluster or principal component analysis. Participants can build historical trends and compare the data with data from other countries or regions. For example, to better understand the global situation, the median values of heavy metal pollution with bordering countries and regions can be shown in one diagram. Coordinators have access to all tools of ordinary participants; in addition, they can perform group operations with data, receive summary reports and build global maps of pollution.

4. Architecture

To achieve the necessary scalability at the resource level, the platform is built on a cloud infrastructure based on Open Nebula. The amount of data coming to the platform from participants is rather small, but it has a complex structure. We have to manage collections of sampling sites, persistent organic pollutants, intercomparison data, etc. Each object can have tens to hundreds of fields, as well as geospatial data. Data automatically collected for forecasting is estimated at millions of records. In such conditions, it is preferable to use NoSQL solutions. In our case, this is MongoDB, which allows one to work with geospatial data and achieve high performance with correctly specified indexes. As a web server, we use Nginx, the performance of which is sufficient for our tasks.

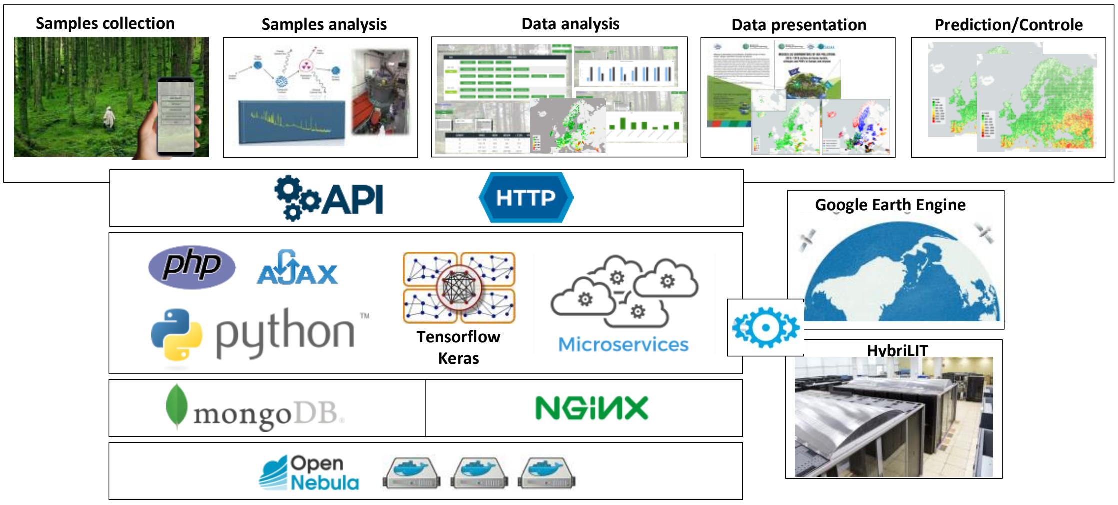

Figure 1. Architecture of the environmental monitoring platform

The server part of the platform provides client and program interfaces and also organizes the work of many services. Some tasks take time to complete, for example, collecting additional data for models or selecting their optimal parameters. To implement them, a microservice architecture is used. It allows one to scale the solution and make changes only to those processes where changes are taking place, without affecting other parts of the platform. For tasks related to machine learning and neural networks, the JINR HybriLIT heterogeneous computing platform is utilized. Machine learning models are implemented in Python using the Keras and TensorFlow libraries. To get satellite imagery and geospatial datasets, we use the Python interface of the Google Earth Engine platform.

5. Intelligent functions

The identification of moss species is important for the quality of analysis. We want to be able to classify moss images in our mobile application. Several deep learning models have been tested to solve recognition tasks on a limited training dataset. We have only 599 images of the five most demanded moss species. The current implementation uses a model of the Siamese neural network with a triplet loss function, the average accuracy of which amounts to 97.7% [4]. Forecasting is an important stage of environmental monitoring to fill data gaps. A forecasting mechanism based on the use of machine learning together with remote sensing data has been implemented within the platform. The approach is not universal, but some chemical elements, such as aluminum, copper, antimony, arsenic, chromium, nickel, iron, and vanadium, have shown good results. Images of various satellite programs are used to obtain so-called indices, which act as additional data when training the model and as basic data when conducting the forecast. The index includes the name of the satellite program, the data of which is used, the size of the analyzed area, the identifier of the spectral channel (band) in which the image is made, and the mathematical function applied to the digital matrix of the obtained image [5]. We have used the Google Earth Engine platform to automatically calculate the indices. This platform has advanced mechanisms for searching, processing, and analyzing satellite data. The data presented at GEE has already been preprocessed. There are more than 100 satellite programs and modeled datasets. Some programs have image resolution up to 15 - 30 meters. Our platform microservices are used to collect indexes, build global and local models, select optimal parameters, and predict contamination.

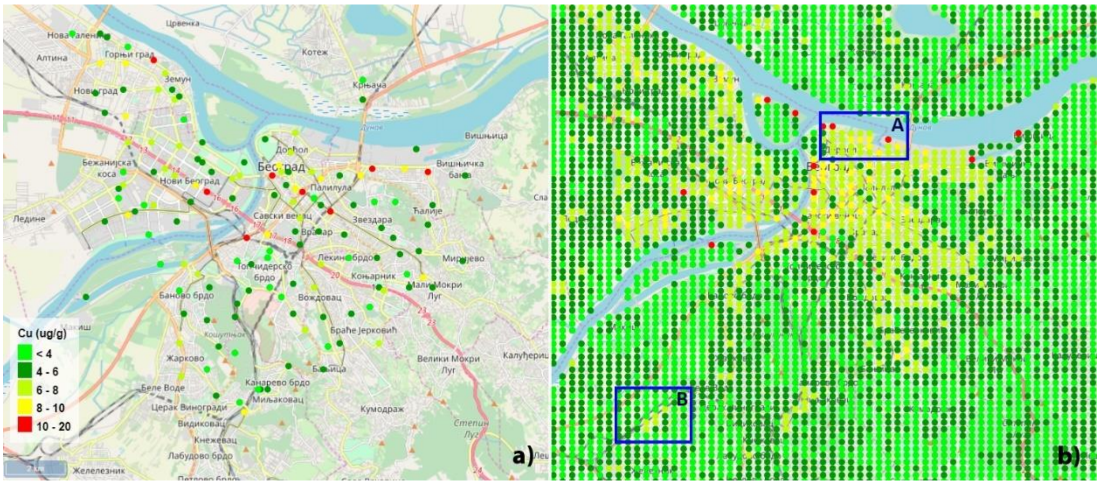

Figure 2. Concentration of Cu in summer in Belgrade: a) real measurements, and b) prediction values; area A represents the central part of Old Belgrade with a permanently high traffic flow; area B represents a large railway terminal

In the current implementation, statistical models or neural networks are used depending on the amount of training data. We focus on regression and classification tasks, but classification is prioritized for several reasons. Firstly, model data is mainly needed to build maps in which the gradation of pollution levels is already well known. Secondly, the determination of the accuracy metrics is clearer and more reliable for classification tasks. Thirdly, there are fewer sampling points that have a high level of contamination than points with a normal level, and better results can be achieved using balancing techniques for training datasets. At present, for local and regional maps of some elements, the accuracy of models reaches 90-95%.

6. Conclusion

The intelligent environmental monitoring platform provides a new level of service for UNECE ICP Vegetation participants. The platform has some worthwhile analytics, classification and prediction abilities based on modern technologies. The obtained results motivate several potential projects to consider our platform as a solution to their tasks. The platform will not only enhance the current functionality, but also provide new opportunities. One of the uppermost tasks is the automation of the environmental monitoring process based on modeling. Gathering satellite indexes is a much faster process than collecting and processing moss samples. We can gather new data and make predictions several times a year. If the contamination level in some regions exceeds certain limits, the platform will send notifications to the corresponding persons. The integration of data on PM and gaseous pollutants from air quality monitoring stations to the platform is considered. The mapping of PM concentrations with heavy metal concentrations can widen the analytical abilities of the platform. We are working on mechanisms of collecting and importing data on citizens' health to the platform. It will enable the comparison of contamination levels and human diseases in some areas. We are pursuing the possibility of obtaining information on diseases both from the Russian Compulsory Health Insurance Fund and from social networks using Big Data technologies.

References

[1] Cohen A., BrauerM., Burnett R., Anderson H.R., Frostad J., Estep K., Balakrishnan K., Brunekreef B., Dandona L., Dandona R., Feigin V., Freedman G., Hubbell B., Jobling A., Kan H., Knibbs L., Liu Y., Martin R., Morawska L., Pope III C.A., Shin H., Straif K., Shaddick G., Thomas M., van Dingenen R., van Donkelaar A., Vos T., Murray C.J.L. & Forouzanfar M.H., Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015 // Lancet, 389 (2017), 1907–1918.

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

[3] Frontasyeva M., Kutovskiy N., Nechaevskiy A., Ososkov G., Uzhinskiy A. Cloud platform for data management of the environmental monitoring network: UNECE ICP Vegetation case // CEUR Workshop Proceedings, 2016, 1787, pp. 224–229

[4] Uzhinskiy AV, Ososkov GA, Goncharov PV, Nechaevskiy AV, Smetanin AA. One shot learning with triplet loss for vegetation classification tasks // Computer Optics 2021; 45(4): 608-614. DOI: 10.18287/2412-6179-CO-856.

[5] A. Uzhinskiy, M. Aničić Urošević, M. Frontasyeva. Prediction of air pollution by potentially toxic elements over urban area by combining satellite imagery // Moss Biomonitoring Data and Machine Learning. Ciencia e Tecnica Vitivinicola Journal, ISSN:2416-3953, Vol. 35, No. 12, 2020.