Data Management of the Environmental Monitoring Network: UNECE ICP Vegetation Case

G. Ososkov[1], M. Frontasyeva[2], A. Uzhinskiy[1], N. Kutovskiy[1], B. Rumyantsev[1,2,]
A. Nechaevsky[1], S. Mitsyn[1], K. Vergel[2]
[1]Laboratory of Information Technologies and [2]Frank Laboratory of Neutron Physics
Joint Institute for Nuclear Research, Dubna, Moscow Region, Russia
ososkov@jinr.ru marina@nf.jinr.ru

Abstract

A new data management cloud platform is presented. The platform is to be applied for global air pollution monitoring purposes to assess the pathway of pollutants in the atmosphere. For this purpose a set of inter-connected services and tools will be developed and hosted in the JINR cloud.

1 Introduction

Air pollution has a significant negative impact on the various components of ecosystems, human health, and ultimately cause significant economic damage. That is why air pollution is a main concern of the Doctrines of the environmental safety all over the world. Increased ratification of the Protocols of the Convention on Long-range Transboundary Air Pollution (LRTAP) is identified as a high priority in the new long-term strategy of the Convention. Full implementation of air pollution abatement policies is particularly desirable for countries of Eastern Europe, the Caucasus and Central Asia (EECCA) and South-Eastern Europe (SEE). 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 [1]. Due to intense activity of the Joint Institute for Nuclear Research (JINR), as a coordinator of the moss surveys since 2014, Azerbaijan, Belarus, Georgia, Kazakhstan, Moldova, Turkey and Ukraine participated in the moss survey for 2015/2016. Nowadays the UNECE ICP Vegetation programme [2] is realized in 36 countries of Europe and Asia. Mosses are collected at thousands of sites across Europe and their heavy metal (since 1990), nitrogen (since 2005), POPs (pilot study in 2010) and radionuclides (since 2015) concentrations are determined. The goal of this study program is to identify the main polluted areas, produce regional maps and further develop the understanding of long-range transboundary pollution [3].

Proceedings of the XVIII International Conference «Data Analytics and Management in Data Intensive Domains» (DAMDID/RCDL'2016), Ershovo, Russia, October 11 - 14, 2016

2 Experiment and data interpretation

Sampling is carried out in compliance with the internationally accepted guidelines [4]. Such analytical techniques as AAS, AFS, CVAAS, CVAFS, ETAAS, FAAS, GFAAS, ICP-ES, ICP-MS, as well an INAA are used for elemental determination. A total of 13 elements are reported to the Atlas (As, Cd, Cr, Cu, Fe, Hg, Ni, Pb, V, Zn, Al, Sb, and N). Nowadays POPs (whichever determined) and radionuclides (namely, 210Pb and 137Cs) are accepted for air pollution characterization. The results are reported as number of sampling sites, minimum, maximum and median concentrations in mg/kg. The data interpretation is based on Multivariate statistical analysis (factor analysis), description of sampling sites (MossMet information package) and distribution maps for each element produced using ArcMAP, part of ArcGIS, an integrated geographical information system (GIS) [5]. Examples of GIS maps are presented in Fig. 1.

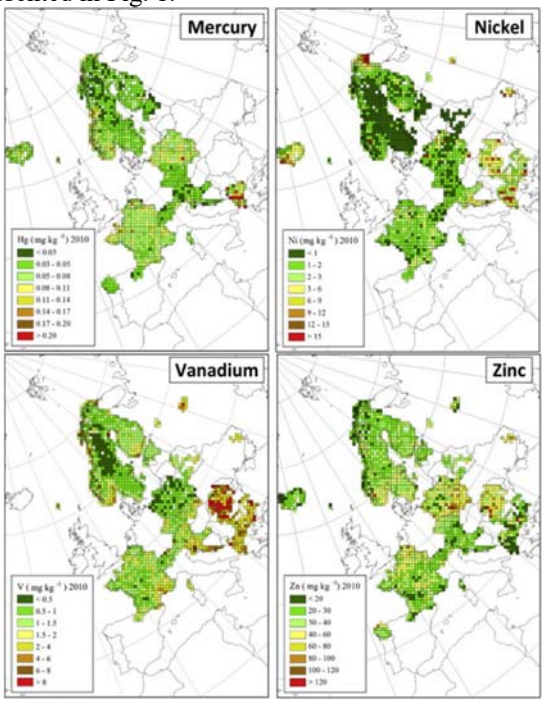

Figure 1 Examples of distribution maps [3]

Analytical results and information on the sampling sites (MossMet set) reported to JINR include confidential

acceptance of the data from individual contributors, the storage of large data arrays, their initial multivariate statistical possessing followed by applying GIS technology, and the use of artificial neural networks for predicting concentrations of chemical elements in various environments.

As an example of the importance of this study, the tendency of average median metal concentrations in moss (± one standard deviation) since 1990 to 2010 are presented in Fig. 2.

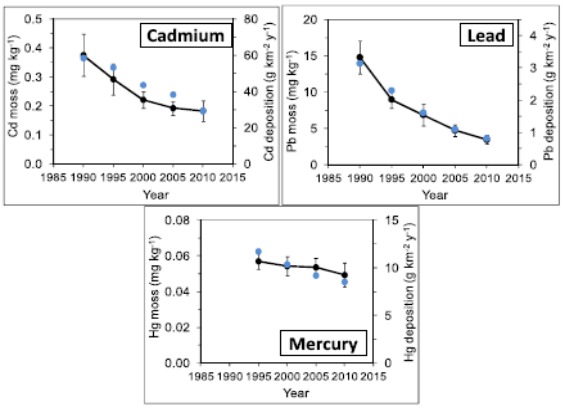

Figure 2 Change in atmospheric deposition of elements in time. The black dots in the graphs show the decline in deposition across Europe and blue dots as modeled by the Environmental Monitoring and Assessment Program

3 Motivation and aim

As discussed above, the ICP Vegetation programme is very important project, but it has a serious weakness related to its weak adoption of modern informational technologies. There are dozens of respondents in existing monitoring network and their number is increasing, but information on collecting and processing of samples is carried out manually or with minimum automation. Data mostly stored in xls files and aggregated manually by the coordinator. Files from respondents are usually passed to the coordinator by email or by ordinary mail. There are no common standards in data transfer, storing and processing software. Such situation does not meet the modern standards for quality, effectiveness and speed of research. Lack of a single web-platform that provides comprehensive solution of biological monitoring and forecasting tasks is a major problem for research.

Therefore the aim of the project is to create a cloud platform using modern analytical, statistical, programmatic and organizational methods to provide the scientific community with unified system of gathering, storing, analyzing, processing, sharing and collective usage of biological monitoring data.

The platform elements are to facilitate IT-aspects of all biological monitoring stages starting from a choice of collection places and parameters of samples description and finishing with generation of pollution maps of a particular area or state-of-environment forecast in the long term. Mechanisms and tools for association of participants of heterogeneous networks of biological monitoring are to be provided in the platform. That enables verifying obtained results and optimizes research. The open part of the platform can be used for informing public authorities, local governments, legal entities and individuals about state-of-environment changes.

One more important aspect of ecological researches relates to various statistical methods applied to process collected data. Modern approaches to explore air pollutions provided by heavy metals, nitrogen, POPs and radionuclides include as a mandatory part multivariable statistical and intellectual data processing. Latest tendencies in data processing include extension of a set of georeferenced data that is integrated in data processing of surveyed data. So it is not limited by geographical, topographical or geological information, what is traditional in such cases, but also includes, for example, satellite imagery and their products, topographic high-precision data derived from aerial photography, etc. These new data classes, contrary to the traditional ones, are characterized by a high resolution and dynamic nature – for example, satellite images represents a reflection of solar radiation, which depends on the time of day, season, cloud cover, etc. This in turn greatly increases the amount of data to be processed. The task of integration of different types of data is tied to the problem of the development of new models and algorithms – such as neural networks [6], self-organizing maps [7], etc. – during the study of dynamic properties of ecological processes among other things.

So, one more aim of our project is to develop modern software tools for multivariable statistical and intellectual data processing oriented on the GIS-technology.

4 ICP vegetation data and required resources

The moss data are to be collected for about 50 countries in Europe, Asia and Central America. Each country has more than 100 monitored points, and several hundred parameters must be taken into account for each of them. A bulky archive is needed to perform comparative studies and to estimate dynamics of explored air pollution processes. Keeping in mind the intensive data exchange and non-relational and poor-structured character of data we can assess the size of our database on the level of terabytes.

Thus it is necessary for scientists to manage large amounts of data, and it leads to many non-trivial problems in IT field. It seems natural that a solution should be centralized and outsourced to a cloud.

From a cloud point of view, the amount of data and computing leads to data intensive processing.

5 Data management on the unified cloud platform

To optimize the whole procedure of data management, it is proposed to build a unified platform consisting of a set of interconnected services and tools to be developed, deployed and hosted in the JINR cloud [6].

The JINR cloud currently has 400 CPU cores, 1000 TB of RAM and about 30 TB of total local disk space on cloud worker nodes for virtual machines and containers deployment. Hosting services in the cloud allows scaling up and down cloud resources assign to the services depending on theirs load. When some component will require more resources cloud can provide it without affecting other components. This increases the efficiency of hardware utilization as well as the reliability and availability of the service itself for the end-users. Such auto-scaling behavior will be achieved by using the OneFlow component of OpenNebula platform [6], which the JINR cloud is based on.

We define requirements for the platform and specify its components. The general architecture of the platform and technologies used are depicted in Fig. 3.

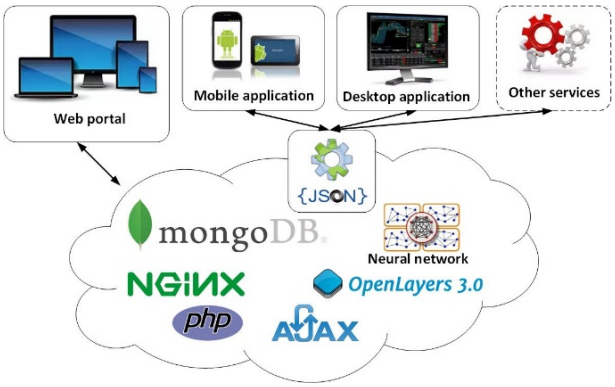

Figure 3 General architecture of the platform and technologies used

We analyze data that comes from the contributors. The data samples can have 10 to 40 metrics depending on the collecting area. Most of the metrics are optional, so traditional relation databases will be ineffective. We also want to have a possibility to change structure of the data sample object without hard code modification to easily integrate new projects and experiments into the platform. We have a positive experience with MongoDB (open-source, document database designed for ease of development and scaling [9]) at our previous projects where more than 5 million data records from 200+ contributors are processed so it was decided to use the data base to store sampling results.

The portal back-end will be built on Nginx (an open source reverse proxy web server for HTTP protocol [10]) and developed with PHP (widely-used open source general-purpose scripting language that is especially suited for web development). That should provide necessary performance and scalability. Web-portal with responsive design that adjusts to different screen sizes is the main interface of the platform. The portal allows multilevel access to the data and has advanced data processing and reporting mechanisms. Currently basic functionality of the portal has been implemented and authorized users can manage their project/regions, import data samples and generate regional maps. At top of Fig. 4 one can see the interface for project management where contributor can add, delete, edit or copy the datasets. At bottom of the figure the map with the indication of pollution distributions and basic instruments to configure the map are presented.

We have tried QGIS (Open Source Geographic Information System [11]) and OpenLayers (opensource javascript library to load, display and render maps from multiple sources) for regional and global maps representation. But QGIS and its web plugin is too hard to maintain and develop. Now we are using OpenLayers [12] and some of its specific layers that allows to do basic interpolation to create concentration maps.

Another interface to the platform is RESTful service [13] that we are going to provide to the mobile and desktop application and also for third-party services that can be interested in the environmental monitoring data.

Data import and export mechanism will be available for the platform, so users can process data online or upload it and use their local processing application. Intelligent multi-level statistical data processing is one of the platform important parts. We have tried several solutions but statistical and analytical packages are still under discussion. A very promising direction is the use of artificial neural network applications for predicting concentrations of chemical elements in various environments. We have done some research in this field but do not yet have the finished solution.

6 Prediction and GIS-oriented data processing

Prediction is an important step of data analysis of any ecological survey. Application of prediction methods enables mapping of estimate values. Maps in their turn provide visualization of spatial variability of data and can be used for visual analysis so that ecological hazards can be identified [14].

Kriging is a widely-used interpolation technique used for prediction, e.g. concentration of heavy elements in moss [15], soil contamination [6]. Recently more and more research is made towards integration of different data sources like aerial and satellite photography together with incorporation of new methods like artificial neural networks.

Mathematically, given a discrete function $f(x_i, y_i)$ (response variable defined by measurements over Cartesian coordinates) on an irregular grid of a set of points $V = \{(x_i, y_i)\}$, an interpolation procedure finds $\tilde{f}(x, y)$ for f such that $\tilde{f}(x, y)$ is prediction for f for $\forall (x, y) \in \mathbb{R}^2$. Integration of data is done in such a way that helps an interpolation procedure, like artificial neural network frameworks, to make a better predictor (interpolator). Such an approach is based on a conjecture that neural networks are capable to employ hidden non-linear correlations that exist and hidden in the data.

Formally, if compared to "classic" interpolation where predictor variables are limited by Cartesian coordinates, in this case a set of predictors is to be expanded with other predictor variables $g_1(x, y)$, $g_2(x, y), \ldots, g_n(x, y)$. These can be topographical features, elevation, products of aerial photography and satellite imagery and many more different surface properties.

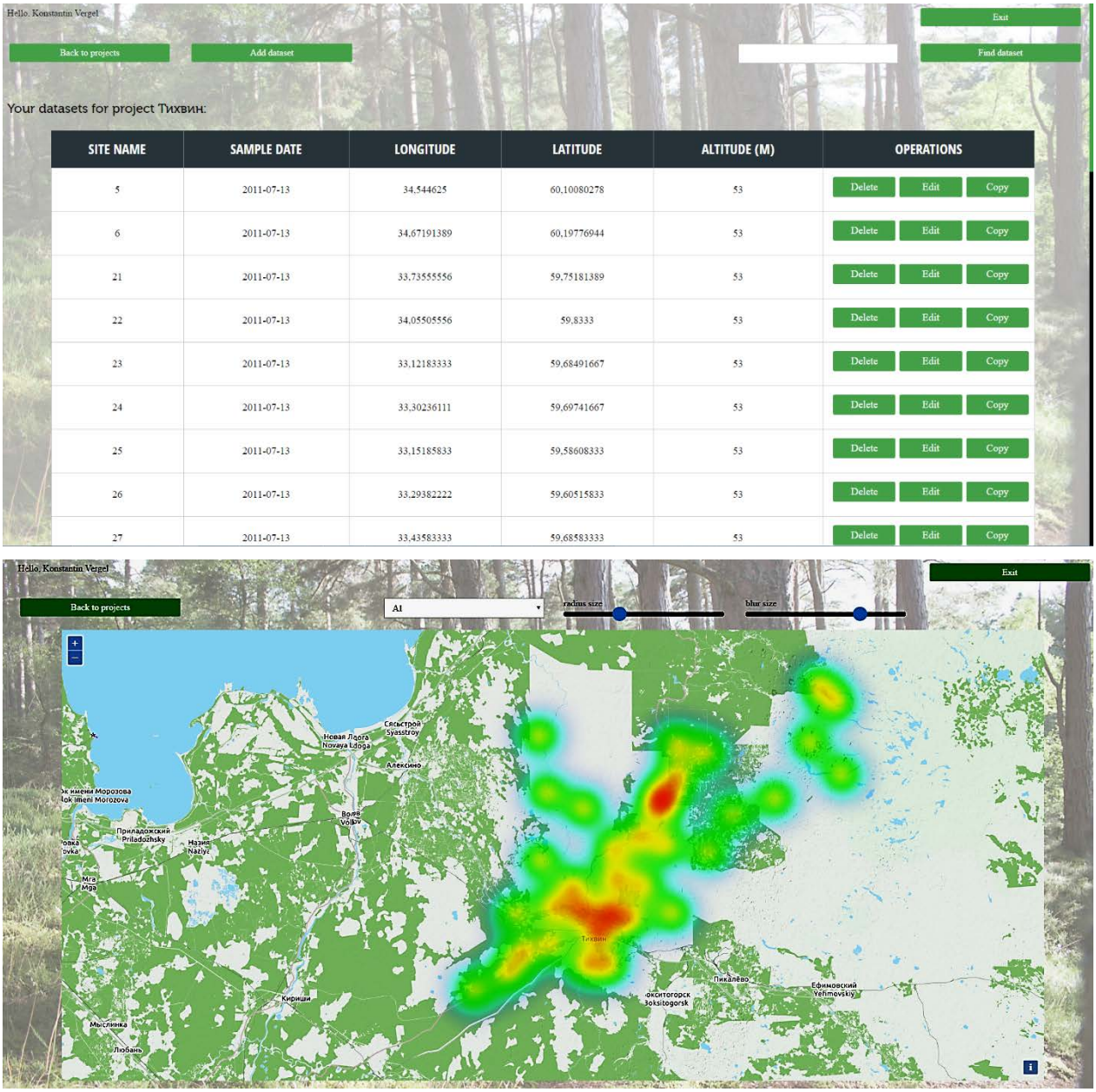

Figure 4 Web-portal interfaces

Thus the interpolation method is replaced in this way: given

$f(x_i, y_i, g_1(x_i, y_i), g_2(x_i, y_i), \dots, g_n(x_i, y_i)) = f(x_i, y_i)\ \forall (x_i, y_i) \in V$,

build a predictor

$\tilde{f}(x, y, g_1(x, y), g_2(x, y), \dots, g_n(x, y))$

and establish an equality

$\tilde{f}(x, y) = \tilde{f}(x, y, g_1(x, y), g_2(x, y), \dots, g_n(x, y))$. While extended form for $\tilde{f}(x, y)$ seems more complicated, it simply allows an interpolation method to fuse in possible non-linear correlations and make "better" predictions, while other formal parts of the method stay the same.

A modification of kriging called cokriging has been proposed [16]. It is oriented on using these secondary variables, as aerial and satellite protography, but has some problems with applying them to real-world data. Kriging is oriented on data that is normally distributed. While it is somewhat true for lags of spatial coordinates that are utilized in semivariogram and covariance function, it is not always true for secondary predictors g_i. Also, it is problematic to construct a covariance function as it may naturally be anisotropic and even non-symmetric. An artificial neural network, on the other hand, automatically adapts to nonlinearities and non-normally distributed variables.

This approach also constitutes some problems which are inherent to artificial neural networks. As each concrete neural network is a product of a learning

procedure, some form of predictor evaluation has to be incorporated. Usually, several different learning procedures and network topologies are evaluated and results of interpolation procedure are analyzed for deficiencies like overfitting. Such superprocedure effectively increases computational costs and may be sped up with parallel computing. Other problems are caused by data specifics, so some approaches of regularization should be employed, like learning with Gaussian noise.

Different types of satellite imagery are currently employed in data processing, like LandSat [17] and MODIS (the Moderate-resolution imaging spectroradiometer [18]). The latter project incorporates two satellites with spectroradiometers (hence the name) that is able to take satellite imagery with high spectral resolution of 36 spectral bands. Whilst, if compared to LandSat, it has moderate resolution (hence the name), it allows deeper and more thorough analyses of Earth surface, thus enabling interesting possibilities for research towards correlation and causality (e.g. contamination spreading catalysts and accelerants).

Using raw spectral radiation bands of spectral imagery is confronted with obvious interfering factors such as sun azimuth, time of day and season, surface altitude and slope and other.

Thus, in addition to the running standard statistical procedures which calculated descriptive statistics and factor analysis, neural network data processing is considered to be used in the given project, together with various MODIS products, as surface reflectance, land surface temperature, land cover, vegetation indices, land use, etc.

7 Conclusion

The study of migration and accumulation of highly toxic pollutants, which include heavy metals, persistent organic pollutants and radionuclides, the influence of pollutants on the various components of the natural and urban ecosystems is the key problem of modern biogeochemistry and ecology. The aim of the given project is to create cloud platform using modern analytical, statistical, programmatic and organizational methods to provide the scientific community with unified system of collecting, analyzing and processing of biological monitoring data.

Parts of the project have already been implemented. The rest is going to be implemented in the next two years.

References

[1] United Nations Economic Commission for Europe International Cooperative Programme on Effects of Air Pollution on Natural Vegetation and Crops (http://icpvegetation.ceh.ac.uk/

[2] Harmens H. and Mills G. (Eds.) Air Pollution: Deposition to and impacts on vegetation in (South-East Europe, Caucasus, Central Asia (EECCA/SEE) and South-East Asia. Report prepared by ICP Vegetation, March 2014. ICP Vegetation Programme Coordination Centre, Centre for Ecology and Hydrology, Bangor, UK. ISBN: 978-1-906698-48-5, 2014, 72p.

[3] Harmens H., Norris D.A., Sharps K., Mills G. … Frontasyeva M., et al. Heavy metal and nitrogen concentrations in mosses are declining across Europe whilst some “hotspots” remain in 2010. Environmental Pollution. 2015, 200:p. 93-104. http://dx.doi.org/10.1016/j.envpol.2015.01.036

[4] HEAVY METALS, NITROGEN AND POPs IN EUROPEAN MOSSES: 2015 SURVEY http://icpvegetation.ceh.ac.uk/publications/documents/MossmonitoringMANUAL-2015-17.07.14.pdf

[5] Buse A. et al. (2003). Heavy metals in European mosses: 2000/2001 survey. UNECE ICP Vegetation Coordination Centre, Centre for Ecology and Hydrology, Bangor, UK. http://icpvegetation.ceh.ac.uk.

[6] J. Alijagić, 2013. Application of multivariate statistical methods and artificial neural network for separation natural background and influence of mining and metallurgy activities on distribution of chemical elements in the Stavnja valley (Bosnia and Herzegovina) : PhD thesis. University of Nova Gorica.

[7] Žibret, G., Šajn, R., 2010. Hunting for Geochemical Associations of Elements: Factor Analysis and Self-Organising Maps. *Mathematical Geosciences*, 42(6): 681–703, doi:10.1007/s11004-010-9288-3. http://link.springer.com/article/10.1007/s11004-010-9288-3

[8] Kutovskiy N., Korenkov V., Balashov N., Baranov A., Semenov R. JINR cloud infrastructure. Procedia Computer Science, ISSN: 1877-0509, Publisher: Elsevier. 2015, 66, p. 574-583.

[9] MongoDB site and description, URL: https://www.mongodb.com/

[10] Nginx for Windows URL: http://nginx.org/ru/docs/windows.html

[11] QGIS description URL: http://www.qgis.org/en/site/

[12] OpenLayers description URL: http://docs.openlayers.org/

[13] RESTful Web services: The basics URL: https://www.ibm.com/developerworks/library/ws-restful/

[14] Goodchild M.F., Parks B.O., Steyaret L.T. 1993. Environmental modelling with GIS, Oxford University Press, New York, 488 p.

[15] S. Nickel et al. / Atmospheric Environment 99 (2014) 85e93

[16] H. Wackernagel Cokriging versus kriging in regionalized multivariate data analysis. Geoderma, 62 (1994) 83-92

[17] LandSat program description URL: http://yceo.yale.edu/what-landsat-program

[18] MODIS spectrometer on TERRA satellite URL: http://modis.gsfc.nasa.gov/about/