Machine learning approach for simulation of heavy metal concentration in river water: the Crimean peninsula case study

Evgeniy Malygin*, and *Mikhail* Lychagin

Lomonosov Moscow State University, GSP-1, Leninskie Gory, 119991, Moscow, Russia

Abstract. This study proposes an approach for simulation of heavy metal concentration in river waters using machine learning techniques. A regression model was built and it captured the relationship between the concentration of heavy metal and metalloids (HMM) and several characteristics of studied catchment. Machine learning techniques allowed to simulate the annual concentration variability of HMM. This approach allows exploring the impact of different factors on studied processes.

1 Introduction

Water scarcity increasing and water quality problems require a complex study of the processes, which control water and hydrochemical regime of rivers. The same problems are also relevant for the Crimean peninsula, which has important economic, agricultural and recreational significance. Currently, there is water scarcity on the Crimean peninsula and this problem needs to develop new water supply solutions. The Crimean Rivers are the main source of fresh water. About 60% of fresh water is withdrawn from rivers. This study aims to find effective methods for water quality assessment. The main aim is to simulate water quality parameters such as heavy metal based on machine learning techniques.

2 Materials and methods

2.1 Data collection

This research uses the results of hydrological and geochemical studies on the Crimean rivers Salgir, Alma, Kacha, and Belbek, collected by Faculty of Geography of Moscow State University in different hydrological conditions in 2015 and 2016 (Fig. 1). In this work, data for three seasons was used: winter high-water period (February 2015), summer flood (June 2015), and summer low water (August 2016). The studies included the river discharge measurements and physicochemical water parameters. The total concentration of heavy metal was measured with inductively coupled plasma mass spectrometry.

* Corresponding author: malygin.ev@gmail.com

All information was included in geodatabase based on satellite imagery, field, and cartographic materials. This database also contains information on soil, land cover, land use, urban area and population.

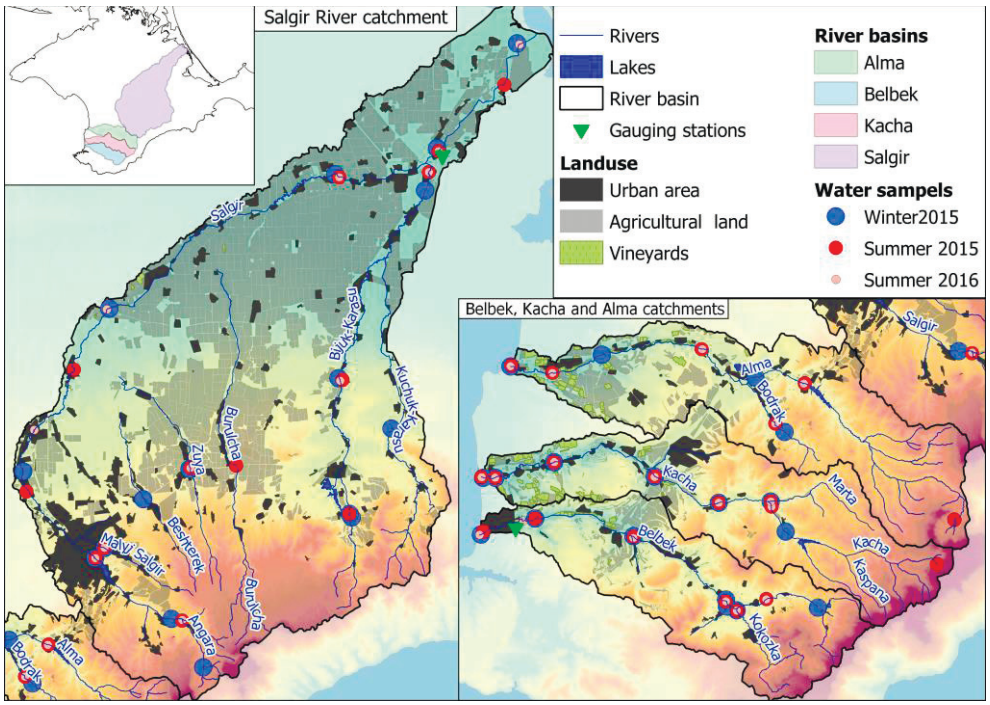

Fig. 1. Water sampling sites.

2.2 Model implementation

To simulate annual variability of heavy metal concentrations Random Forest regression model was used. Random forest is a machine learning algorithm using the composition (ensemble) of decision trees [1,2]. A single decision tree gives a low-quality model. However, when using Random Forest, a good result is achieved by averaging the responses of a large number of single decision trees.

To simulate heavy metals, a regression model was built. This model captured a relationship between the concentration of heavy metal and several characteristics of studied catchment. The model is written in the Python 3 programming language using the scikit-learn machine learning library [3]. The Random Forest is used to build a regression model. Input data for the model are the following: anthropogenic characteristics of studied catchments (population density and shares of the catchment area attributable to urban area, agricultural land, and vineyards), water discharge, average daily air temperature, average catchment elevation, river basin ID (categorical feature). The target variable is the concentration of heavy metals in river waters. Leave-one-out cross-validation was used to evaluate the results. This approach can be used to calculate feature importance. Feature importance gives a score for each feature of the input dataset, the higher score more important is the feature towards the output variable.

3 Results and discussion

The spatio-temporal distribution of heavy metals is given based on the field measurement results. As an example, the distribution of concentrations of copper and molybdenum during the summer low water period in 2016 is presented at Fig. 1.

It can be possible to simulate the annual variability of the concentrations of Cu, Mo, As, Sr using machine learning techniques. Fig. 2 shows the distribution of the concentration of heavy metals. The simulations were carried out for gauging stations Salgir – Listvennoe and Belbek – Fruktovoe (Fig. 3). Also, the feature importance was calculated (Fig. 4).

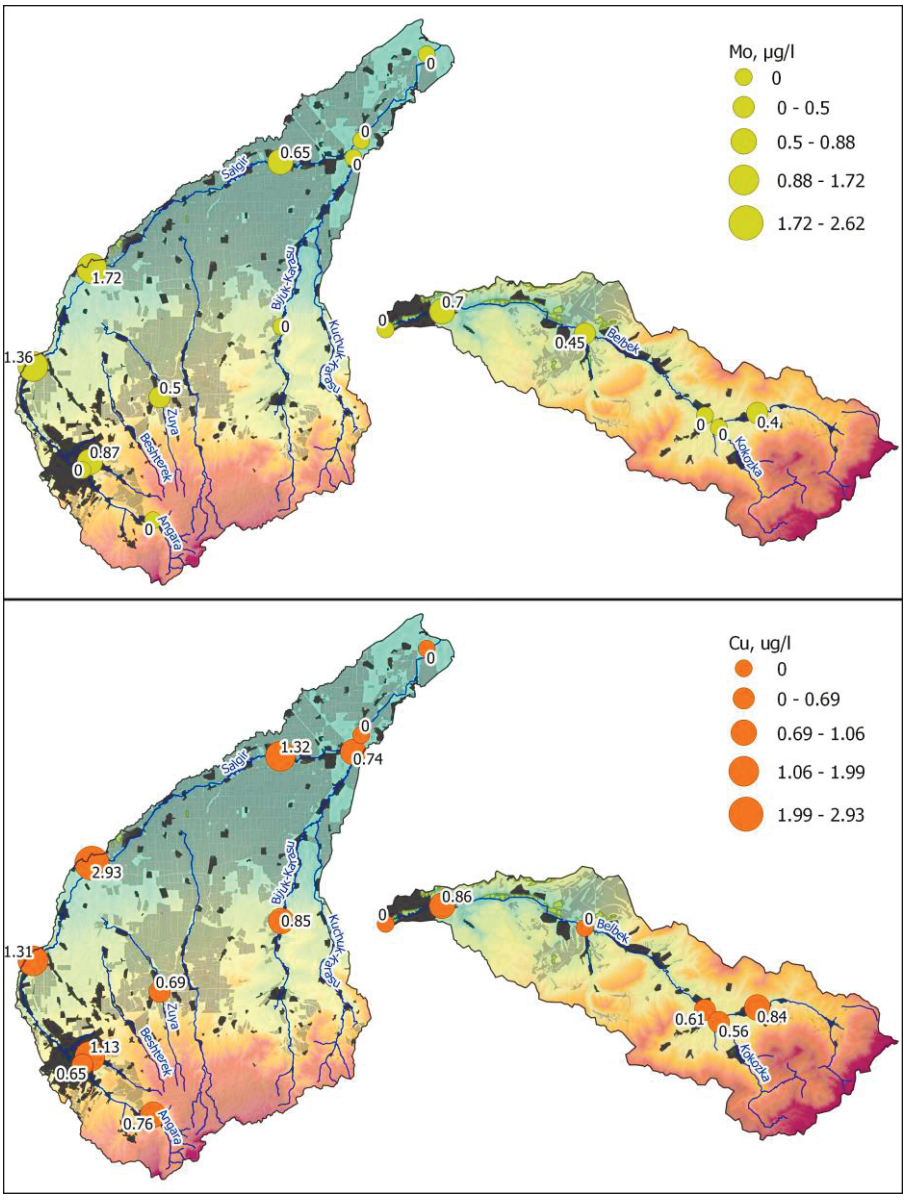

Fig. 2. Distribution of the Mo, and Cu concentration, µg/l.

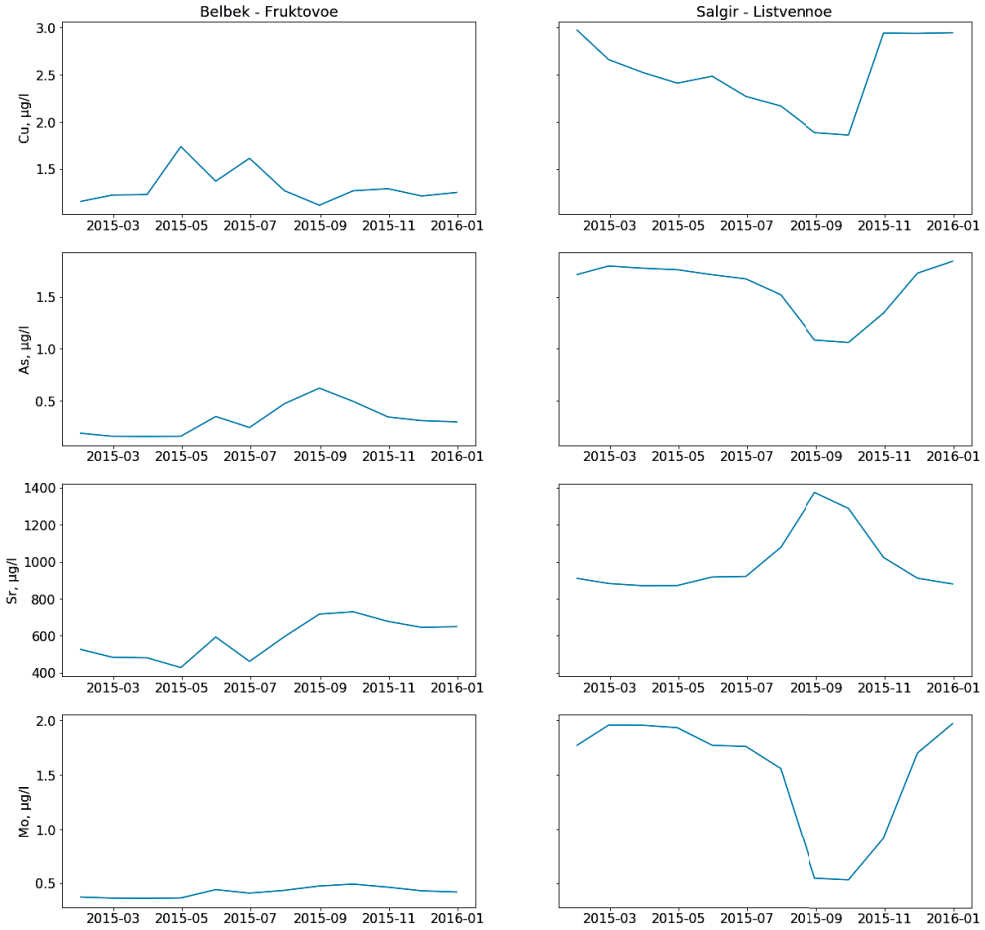

Fig. 3. Simulated concentrations of heavy metal and metalloid, µg/l.

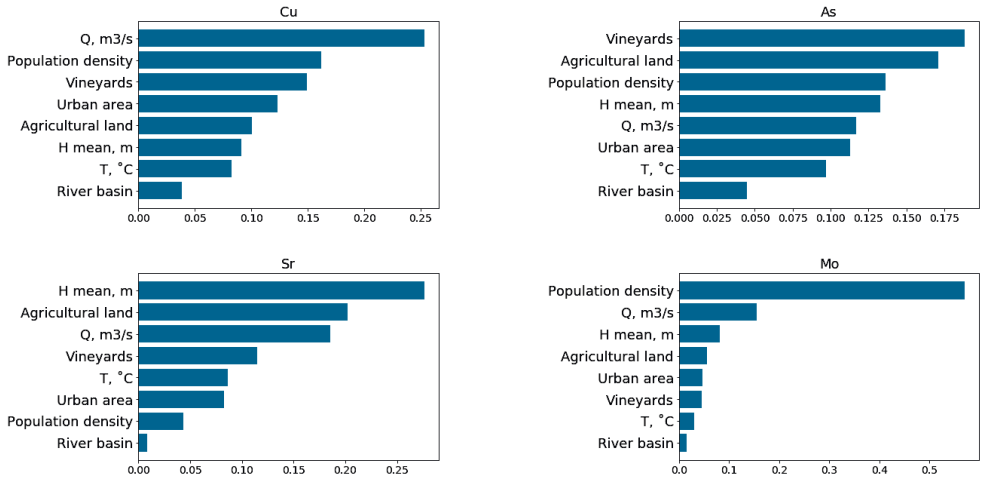

Fig. 4. Feature importance.

The maximum content of Cu in the Belbek River water was observed during the summer flood 2015. It may be associated with the intensive leaching of substances from the catchment area. The minimum values are characteristic for the low-water period. In the Salgir River water two peaks of Cu concentration were found: at summer flood and winter high-water periods. The greatest influence on the concentration of copper is exerted by the river water runoff and area of vineyards within the catchment. To combat harmful insects, plantations are treated with pesticides, which contain copper [4, 5]. Runoff increasing during a high-water period leads to an increase in copper inflow from the catchment.

The monthly average As concentration in the Salgir River 2-3 times greater than in the Belbek River. This may be associated with the presence of arsenic in the pesticides and phosphate fertilizers, which are widely used in agriculture. Arsenic migratory ability increases in alkaline environment which is typical for the semi-arid landscapes. High As concentration in the Salgir River is due to the intensive inflow of this element from agricultural land during high-water periods. During the summer low water period As content is the lowest. The maximum concentration of arsenic in the Belbek River is observed during the summer low water period, which may be associated with the intake of this element from the saline groundwaters [6]. The most important features, in this case, are agricultural land, vineyards, and population density.

The monthly average Mo concentration in the Salgir River is 3-4 times greater than in the Belbek River. This is due to its high migratory ability in alkaline steppe soils. The most important feature, in this case, is a population density, which indicates the mountain and the steppe regions of the Crimean Peninsula [7].

The most important factor for Sr concentration is the average catchment elevation, in this case, which may reflect the influence of calcareous bedrocks. The peak of Sr concentration occurs in the summer low water. It is associated with the ground feeding of rivers prevailing during this period. The concentration of Sr in the Salgir River is several times higher than in the Belbek River water. This may be due to the intake of Sr from agricultural land from phosphate fertilizers, which contain a large amount of Sr.

Thereby, our approach allowed simulating the annual concentration variability of several pollutants and estimating their flows from river catchments. The simulation results show that the model reproduces well the results of field studies at low absolute error value.

4 Conclusion

Machine learning techniques allowed simulating the annual concentration variability of several heavy metals and metalloids. This approach allows for exploring the impact of different features on the modelling result.

The modelling results based on a machine learning approach can be used to explore complex interconnections in environmental systems with high anthropogenic influence for the improvement of environmental management.

The study was carried out in frames of the Crimean Expedition of the Russian Geographical Society and research project of Russian Science Foundation 16-05-01037\16.

References

1. L. Breiman. Mach. Learn. **45**, 5–32 (2001)
2. L. Breiman. Mach. Learn. **26**, 123–140 (1996).

3. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. J. Mach. Learn. Res. **12,** 2825–2830 (2011)
4. A. Desiree, J. Viers, M. Guiresse, A. Probst, D. Aubert, J. Caparros, F. Charles, K. Guizien, J. Probst. Sci. Total Environ. **463–464**, 91-101 (2013)
5. S. Blotevogel, P. Oliva, S.E. Sobanska, J. Viers, H. Vezin, S.e Audry, J. Prunier, J. Darrozes, L. Orgogozo, P. Courjault-Rade, E. Schreck. Chem. Geol. **477**, 35–46 (2018)
6. R. Bowell, C. Alpers, H. Jamieson, D. Nordstrom, J. Majzlan. Rev. Mineral. Geochem. **79**, 1–16 (2014)
7. P. Smedley, D. Kinniburgh. Appl. Geochemistry. **84**, 387–432 (2017)