Building a model for predicting the water level in a river using remote sensing data from open sources

Alexey Kolesnikov[1,*], *Pavel Kikin*[2] , and *Anastasia Nungesser*[1]

[1] Siberian State University of Geosystems and Technologies, Department of Cartography and Geoinformatics, 630108 Plakhotnogo street 10, Novosibirsk, Russia
[2] Peter the Great St. Petersburg Polytechnic University, Higher School of Theoretical Mechanics, 195251 Polytechnicheskaya Street 29, Saint Petersburg, Russia

Abstract. The article discusses the possibility of predicting the water surface area of a river (and based on these values, the calculation of the water level) based on only open data of remote sensing. The area and depth of snow cover, the intensity of precipitation according to Landsat and Sentinel data and the monitoring indicators MODIS, Copernicus, REMSS are used as initial parameters. For the selected parameters, the degree of influence on the final forecast was assessed.

1 Introduction

Summer-autumn floods occurring on rivers can cause serious damage to industrial facilities, settlements, agricultural land. Therefore, their forecasting is a complex multi-level task, the relevance of which is due to the current state of economic systems. This primarily applies to urbanized and industrial regions. [1-3]. Since now a large amount of various remote sensing data is presented in the open access on a time interval of more than 10 years (with a fairly high frequency), it becomes possible to study almost any territory in terms of hydrographic objects and their relationships with climatic factors [4,5]. When building a forecast, it is very important and relevant to choose the appropriate research method. In hydrology, various methods of forecasting are widespread: graphic-analytical, water-balance, mathematical modeling, etc. There are a number of approaches for solving this problem, both deterministic and stochastic. In this study, machine learning methods were chosen to construct a mathematical model of the river flood, so it was necessary to carefully prepare the data set in order to "train" the relevant algorithms.

Today's remote sensing open data sources make it possible to collect dozens of environmental parameters every day, which, on the one hand, makes it possible to build the most comprehensive models, on the other hand, it can lead to unnecessary complication of the model and it's overfitting. [6,7]. Therefore, it was important to analyze the algorithms for assessing the importance (impact for the target feature prediction) of each of the

* Corresponding author: alexeykw@mail.ru

parameters to the final result and use them to analyze the initial data in the described problem.

2 Materials and Methods

In the very beginning of the study the sufficiency of the most logically grounded data was checked in order to construct a forecast of the water level and assess its accuracy. The territory of Irkutsk was chosen as the object of research. One of the most difficult, but often determining factors in the formation of runoff is the absorption of incoming water by the soil and precipitation. The latter are predicted only with a lead time of 2-3 days, or are replaced by their norm over the lead time period. The amount of runoff during the flood is determined by three main factors: the amount of water reserves in the snow cover accumulated by the time of the forecast, the amount of precipitation from the moment of making the forecast until the expiration of its lead time, the water-absorbing capacity of the basin, which depends on the depth of soil freezing, its humidity in autumn and during the winter thaws, negative air temperatures at the Earth's surface and the snow reserves themselves. These factors determine the moisture absorption capacity of different parts of the basin in different ways, depending on the composition of the underlying surface and the characteristics of a particular year. As the initial data, we used data on snow cover in areas of mountain ranges from which the flow of Lake Baikal and the Angara River is formed (MODIS), information on precipitation on the territory of catchment areas (Copernicus, REMSS), as well as water surface parameters based on the interpretation of satellite images (Landsat, Sentinel). The collected data from open sources were transformed in such a way that it was possible to form a general table and, on the basis of which, build a mathematical forecasting model and evaluate its quality (the execution scheme is shown in Figure 1).

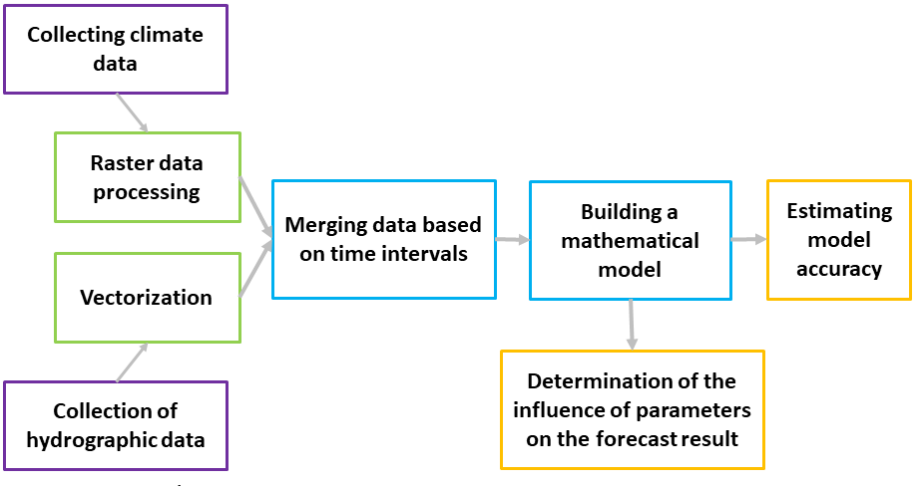

Fig. 1. Study execution pipeline

At the stages of setting a machine learning problem and generating data, it is not always clear which features are important for constructing an optimal algorithm, therefore, data often contains a lot of redundant (noise) information. The appearance of noise signs degrades the quality of the algorithm and slows down its work. Therefore, in most cases, before solving the problem of classification, regression or forecasting, it is necessary to select those features that are most informative. Choosing the right features can be a more significant task than reducing data processing time or improving classification accuracy.

Therefore, a block for determining the importance of individual features was added to the study. In addition to demonstrating the importance of features directly, one can evaluate the general concept of the model, its interpretability. Interpreting methods for machine learning models are divided into local and global. Global methods are designed to show which factors in general have the greatest influence on the structure of the model and on its predictions. Local methods try to explain how a given prediction was made. Often local methods can be used as a basis for a more global interpretation, for example, through averaging or visualization. Local methods LIME and SHAP were used for the assessment.

Data were collected for certain areas (Figure 2), the location of which was selected based on the hydrological description of the Angara River and Lake Baikal, for which the main source of water is the Selenga River. The time interval from 2014-2019 was taken for the analysis.

For remote study of the areas of the water surface of lakes according to space data, the following methods of processing satellite images are widely used - determination of multichannel spectral indices (water indices), thematic classification with training, linear separation, single-channel classification using the separation threshold. The paper considers a method for decoding water surfaces using multi-channel spectral indices MNDWI - Modified Normalized Difference Water Index. MNDWI is determined using 3 and 6 channels of Landsat 8 OLI and is currently the most common index for identifying surface water objects in satellite imagery. The boundary of the definition of water surfaces passes through zero, the values of the indices greater than zero correspond to water surfaces.

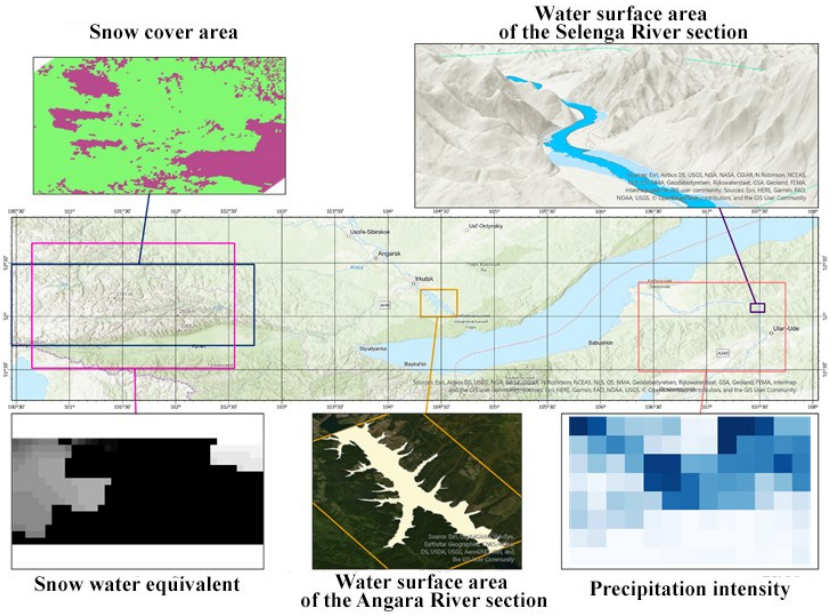

Fig. 2. Data and collection sites

Since most of the training sample is presented in raster form, it was necessary to choose a way to represent each raster in the form of one row of the table (for comparison with the corresponding value of the water cover area). The simplest option is to representing each band of a raster image as a flat array. But it was not suitable due to the significant increase in the dataset. In this regard, it was decided to calculate summary quantitative indicators for each raster: maximum, minimum, average, median, standard deviation. An important step when using multispectral images is to eliminate the redundancy of such images while preserving important spectral information. In this regard, considerable attention of

researchers was paid to both methods of controlled selection of features and uncontrolled methods of reducing the dimension of data. While both of these directions inevitably lead to some loss of spectral information, successful results in terms of the quality of subsequent classification have been demonstrated in both directions. For a number of objective reasons, such as high computational costs, less resistance to scene changes, the need for a classified set, feature selection methods are inferior to dimensionality reduction methods. And, although when using data dimensionality reduction methods, the original interpretation of the data (spectral signatures) is lost, the latter are more often used when working with multispectral images [10-12]. In order to at least partially preserve the information about the spatial arrangement of pixels of the original rasters, the PCA algorithm was used for each of the rasters. The first 5 components for each of the image channels were extracted into a pivot table.

The ArcGIS Model Builder tool was used as an automation tool. For each raster, cropping was performed according to the data collection area and statistical indicators were calculated, which were saved to a CSV file. This approach made it possible to reduce the time for processing each raster separately, but a fairly large part of the operations (forming datasets for loading, placing them on local disks, combining the results into one table) still had to be done manually. As a result, for further experiments, it is planned to either extend the ArcGIS scripts by writing additional functions in Python, or to transfer most of the GIS operations to some kind of spatial database, since their functionality now allows this. It should also be noted that Model Builder does not always use RAM efficiently, and on datasets containing hundreds of elements, the processing speed at each iteration gradually decreases.

3 Results

The importance of individual parameters was assessed based on the calculation of the correlation and the rating of the contribution of each parameter to the final forecast using the LIME [13] and SHAP [14] algorithm, as well as the Feature Importance mechanism built into the random forest and decision tree model of scikit learn" library [15]. According to the assessment results, the area of snow cover received the highest impact rating.

To construct mathematical models, linear regression, decision tree and random forest algorithms were selected. The choice is due to the fact that these algorithms are computationally simple and their results are well interpreted. The root mean square error error was chosen as the criterion for assessing the accuracy.

As a result, the random forest algorithm showed the highest accuracy, which for the Angara River section was 0.31 sq. km. And for the section of the Selenga River 0.3 sq. km., which is about 10% of the deviation from the true value.

4 Discussion

To improve the accuracy of the forecast, it is planned to increase the number and areas of data collection regions, which will allow analyzing the effectiveness of the choice of locations and time intervals for collecting data on snow cover and formalizing the selection principles based on an open digital elevation model and basic hydrological analysis. Also, the quality can be improved by creating ensembles of standard machine learning algorithms together with time series forecasting algorithms (such as SARIMAX or LSTM).

Also, since the resulting value is only a digit, you need to choose an algorithm for forming the polygon based on the relief data and the calculated area value.

References

1. M. Norwaziah, P. Nur, I. Izleen, M. Umi, J. Siti, O. Nurul, Proceedings of the Second International Conference on the Future of ASEAN (ICoFA) *Forecasting River Water Level Using Artificial Neural Network*, **2** (2018), 10.1007/978-981-10-8471-3_41.
2. V. Hung, L. Xuan Hien, H. Tuan, XVII World Water Congress of International Water Resources Association. *Using Machine Learning to forecast river water levels in tidal areas for operating sluice gates in Hai Duong province, Vietnam*. (2020)
3. H. Dinna, N. Hardini, L. Dinda, International Conferences on Information System and Technology. *Forecasting River Water Quality using Autoregressive Integrated Moving Average (ARIMA),* 158-163. (2019) 10.5220/0009907201580163.
4. L. Plastinin, V. Stupin, B. Olzoev, Digital geography: proceedings of the All-Russian scientific-practical conference with international participation. *Integration of the regional cartographic and space monitoring system into the geo-information environment,* **1,** (2020)
5. D.V. Andreev, The successes of modern natural science. *Application of gis technologies to determine flooding in the republic of Sakha (Yakutia).* **11**, 43-47, (2019)
6. W.E. Chapman, A.C. Subramanian, L. Monache, S.P. Xie, F.M. Ralph, Geophysical Research Letters. *Improving Atmospheric River Forecasts With Machine Learning.* **46(1),** (2019) 10.1029/2019GL083662.
7. C. Tarau, K. Lee, W. Anderson, C. Morrison, T. Hendrick, International Conference on Environmental Systems (ICES 2020), *Thermal Concept for Planetary Ice Melting Probe*, (2020)
8. M.A. Golyatina, K.A. Kurganovich, Materials of Intern. Conf. Geosystems and their components in Northeast Asia: evolution and dynamics of natural, natural-resource and socio-economic relations, *The study of the change in the number of lakes in the Transbaikal Territory with the use of MNDWI*, 174-179, (2016)
9. G.L. Feyisa, H. Meilby, R. Fensholt, S.R. Proud, Remote Sensing of Environment, *Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery*, **140**, 23-35, (2014)
10. J. Wang, C.-I. Chang, IEEE Trans. Geosci. Remote Sens. *Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis,* **44 (6)**, 1586-1600, (2006)
11. Z. Zhang, Z. Hongyuan, SIAM Journal on Scientific Computing, *Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment*, **26(1)**, 313-338, (2005)
12. C.M. Bachmann, T.L. Ainsworth, R.A. Fusina, IEEE Trans. Geosci. Remote Sens., *Improved Manifold Coordinate Representations of Large-Scale Hyperspectral Scenes*, **44(10)**, 2786-2803, (2006)
13. R. Marco Tulio, S. Singh, C. Guestrin. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, *Why should I trust you?: Explaining the predictions of any classifier,* (2016)
14. S. Lundberg, S. Lee, Advances in Neural Information Processing Systems (NIPS 2017), *A Unified Approach to Interpreting Model Predictions,* **30,** 4765—4774, (2017)
15. J. Demsar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Starc, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik,

B. Zupan, Journal of Machine Learning Research. *Orange: Data Mining Toolbox in Python*. 2349−2353, (2013)