А. В. Ужинский

Современные информационные технологии в экологическом мониторинге

Проблемы загрязнения окружающей среды и экологической безопасности всегда актуальны. Особое внимание уделяется контролю загрязнения воздуха. Большинство программ в данной области направлены на определение мелкодисперсных частиц (particulate matter) и некоторых химических соединений, например CO_2. Для получения подробной информации о составе загрязнения используются методы мониторинга, основанные на отборе проб. В рамках Конвенции ООН по дальнему трансграничному переносу воздушных загрязнений (CLRTAP) в 1980-е гг. была создана программа UNECE ICP Vegetation, участники которой собирают образцы мха и используют различные аналитические методы, в том числе и нейтронный активационный анализ на реакторе ИБР-2 ЛНФ ОИЯИ, чтобы получить данные по содержанию в воздухе тяжелых металлов, азота, стойких органических соединений и радионуклидов.

Проект объединяет исследователей из 43 стран и регионов Европы и Азии. С 2014 г. за координацию программы отвечает Лаборатория нейтронной физики им. И. М. Франка (координатор М. В. Фронтасьева). Несмотря на несомненную важность проекта UNECE ICP Vegetation, уровень применения современных технологий и автоматизации в нем был достаточно низок, что серьезно ограничивало эффективность сбора данных и их статистической обработки. В 2016 г. в Лаборатории информационных технологий (ныне — Лаборатория информационных технологий им. М. Г. Мещерякова) началась разработка системы управления данными проекта UNECE ICP Vegetation. Изначально планировалось, что система упростит и частично автоматизирует типовые операции с данными, а также позволит оперативно создавать карты загрязнения. Со временем система эволюционировала, вбирая в себя все новые и новые технологии и подходы, и

A. V. Uzhinskiy

Modern Information Technologies in Environmental Monitoring

The problems of environmental pollution and environmental safety are always relevant. Special attention is paid to air pollution control. Most of the programmes in this area focus on identifying particulate matter and some chemical compounds such as carbon dioxide. To obtain detailed information on the composition of contamination, monitoring methods based on sampling are used. In the 1980s, the UNECE ICP Vegetation project was created within the UN Convention on Long-Range Transboundary Air Pollution (CLRTAP). Its participants collect moss samples and use different analytical methods, including neutron activation analysis at the IBR-2 reactor of FLNP JINR, to determine the concentrations of heavy metals, nitrogen, persistent organic pollutants and radionuclides in the air.

The project brings together researchers from 43 countries and regions of Europe and Asia. Since 2014, the Frank Laboratory of Neutron Physics has been coordinating the programme (coordinator M. Frontasyeva). Despite the undoubted importance of the UNECE ICP Vegetation project, its level of automation and adoption of modern information technologies was quite low, which seriously limited the efficiency of data acquisition and statistical processing. In 2016, the Laboratory of Information Technologies (now the Meshcheryakov Laboratory of Information Technologies) started developing a data management system of the UNECE ICP Vegetation project. It was initially planned that the system would simplify and partially automate typical operations with data, as well as enable the fast creation of pollution maps. Over time, the system has evolved, incorporating more and more novel technologies and approaches, and now it can be considered an intelligent environmental monitoring platform [1].

Studies within the UNECE ICP Vegetation project are based on the analysis of mosses as biomonitors. Participants collect samples every five years, recording various information about sampling sites. Errors, which negatively affect the results, are possible in the process

в настоящее время может быть причислена к интеллектуальным платформам экологического мониторинга [1].

Исследования в рамках программы UNECE ICP Vegetation базируются на анализе мхов-биомониторов. Участники раз в пять лет собирают образцы, фиксируя различную информацию о местах сбора. Естественно, что в процессе записи и передачи метаинформации возможны ошибки, которые отрицательно сказываются на полученных результатах. Для их минимизации было разработано мобильное приложение, которое позволяет вносить большинство обязательных параметров вручную, а часть данных, например широту, долготу и высоту над уровнем моря, — автоматически. В приложении есть возможность фотографировать места сбора и образцы и отправлять их в платформу для распознавания. Это позволило значительно упростить процесс определения типа мха, что является важной частью сбора метаинформации и в некоторых случаях вызывает трудности даже у экспертов. На базе платформы было апробировано несколько моделей глубокого обучения для решения задач распознавания на ограниченной выборке. В настоящее время используется модель сиамской нейронной сети с трехчленной функцией потерь. Сиамская сеть состоит из нескольких сетей-близнецов, соединенных между собой слоем подобия (рис. 1, *a*).

Веса близнецов одинаковы, поэтому результат является инвариантным и гарантирует, что похожие изображения не могут находиться в разных местах в многомерном пространстве свойств. При использовании трехчленной функции потерь на вход близнецам подаются два изображения одного класса и одно изображение другого класса. В результате это позволяет лучше подобрать веса, чтобы векторные представления схожих изображений находились ближе друг к другу, а изображения другого класса — дальше от них (рис. 1, *c* и *d*). После обучения один из близнецов используется в связке с многомерным персептроном, выступающим в качестве классификатора (рис. 1, *b*). Подобная архитектура сети позволяет классифицировать пять наиболее распространенных разновидностей мха с точностью порядка 97,6 % [2].

Рис. 1. Архитектура сиамской сети (*a*); один из близнецов и MLP-классификатор (*b*); представление векторов изображений в двумерном пространстве до обучения (*c*) и после обучения (*d*)

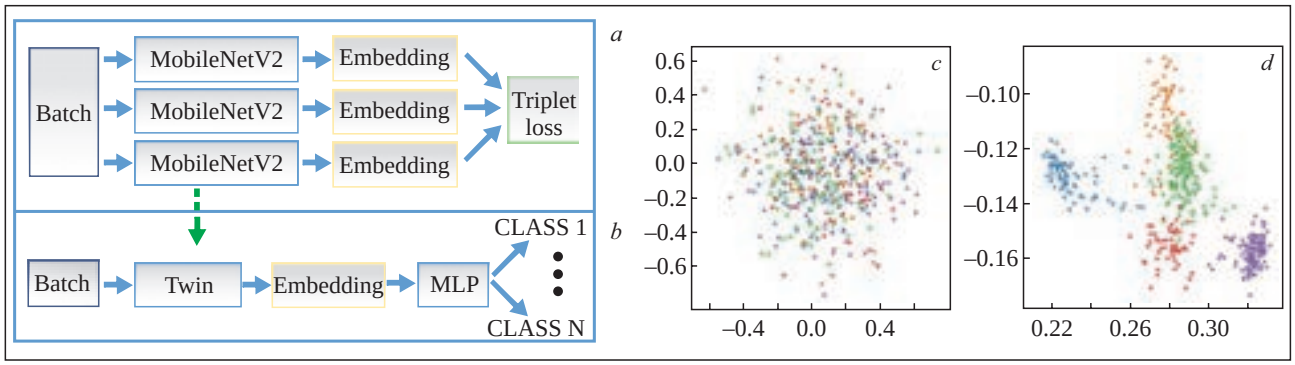

Fig. 1. Architecture of the Siamese network (*a*); one of the twins and MLP classifier (*b*); 2D image vector representation before (*c*) and after (*d*) training

of recording and transferring metainformation. To reduce them, a mobile application was developed. It allows one to fill in most of the required parameters manually, while some data, such as latitude, longitude and altitude, are set automatically. Using the mobile application, one can take pictures of sampling sites and samples and send them to the platform for recognition. This made it possible to significantly simplify the process of defining the moss type, which is crucial for collecting metainformation and in some cases creates difficulties even for experts. On the basis of the platform, several deep learning models were tested to solve recognition tasks on a limited training dataset. The current implementation uses the model of a Siamese neural network with a triplet loss function. The Siamese network comprises several twin networks joined by the similarity layer (Fig. 1, *a*).

The weights of the twins are the same, so the result is invariant and ensures that similar images cannot be in different locations of the multidimensional feature space. When using the triplet loss function, the input consists of three images, two of which belong to the same class, and the third one belongs to another class. After evaluation, the weights are chosen so that the vector representations of similar images are closer to each other, and images of another class are farther from them (Fig. 1, *c* and *d*). After training, one of the twins is used in conjunction with a multilayer perceptron as a classifier (Fig. 1, *b*). Such a network architecture enables the classification of five most common moss species with an accuracy of about 97.6% [2].

In the process of setting metainformation, each sampling site has a unique ID, which is used to import data on the concentrations of elements and compounds after

В процессе внесения метаинформации точкам отбора проб присваиваются уникальные идентификаторы, которые используются после проведения анализа образцов для импорта данных по концентрациям элементов и соединений. В рамках платформы проводятся поиск статистических аномалий, проверка полноты и корректности данных. Полная автоматизация данного процесса невозможна, поскольку аномалии могут иметь естественный характер и для принятия решения об их включении или исключении требуется согласованное мнение участника и координатора. Одной из основных задач при реализации платформы было сведение к минимуму необходимости использования сторонних систем. В настоящее время участники проекта могут производить манипуляции с данными, строить локальные и региональные карты загрязнений, запу-

Рис. 2. Примеры снимков программ в платформе Google Earth Engine

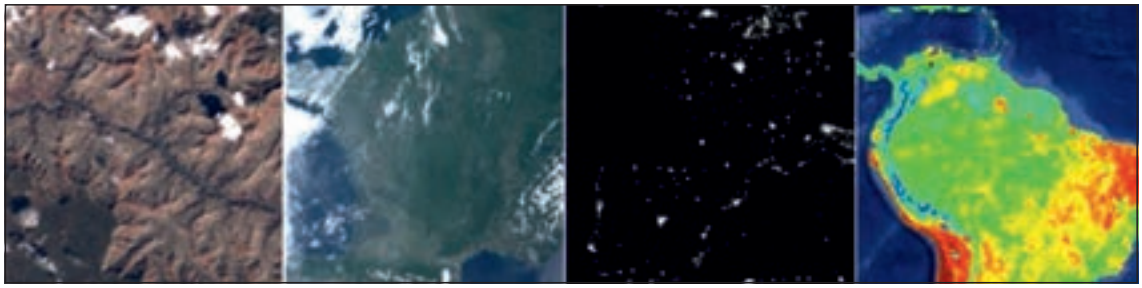

Fig. 2. Examples of Google Earth Engine program images

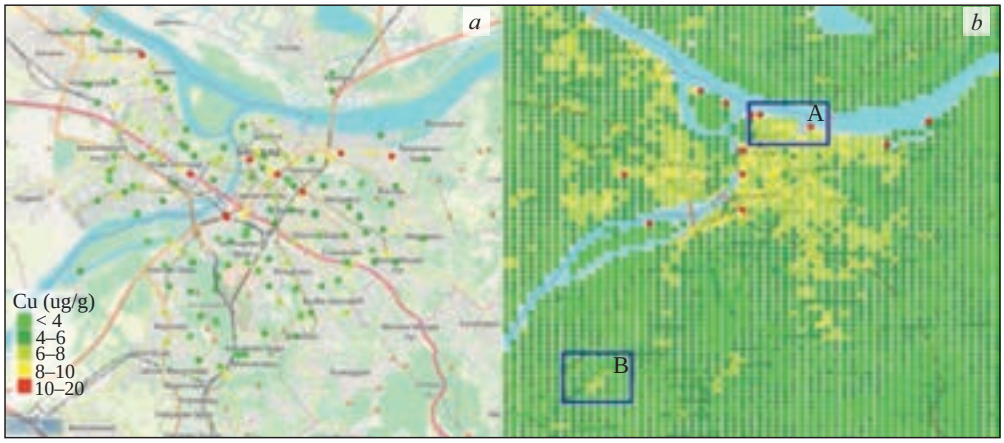

Рис. 3. Прогноз концентрации меди на примере Белграда: *a*) данные мониторинга; *b*) прогноз модели. Область А — центральная часть города с высоким трафиком; В — железнодорожный терминал

Fig. 3. Examples of predicting the concentration of Cu in Belgrade: *a*) monitoring data; *b*) prediction values. Area A represents the central part of the city with high traffic; area B — a railway terminal

analyzing the samples. The platform allows searching for statistical anomalies and checking the completeness and correctness of data. Full automation of the given process is impossible since anomalies can be of a natural kind and the agreed opinion of the participant and the coordinator is required for making a decision on their inclusion or exclusion. One of the uppermost tasks in the implementation of the platform was to minimize the need to use third-party systems. At present, project participants can manipulate data, create local and regional pollution maps, run prediction tasks and get analytical reports directly on the platform. In addition to simple statistics and geo-indexes, tools of a higher level, such as cluster analysis or principal component analysis, are available. Users can build historical trends and make comparisons with the data of other participants with the appropriate permission on their part. 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 create global pollution maps.

Forecasting is an essential stage in environmental monitoring to fill data gaps. A forecasting mechanism based on machine learning and remote sensing data is implemented within the platform.

Images of various satellite programs are utilized to obtain so-called indexes, which act as additional data when training the model and as basic data when conducting the

скать задачи прогнозирования и получать различные аналитические отчеты непосредственно в платформе. Кроме простых статистических выкладок и геоиндексов доступны инструменты более высокого уровня, например кластерный анализ или метод главных компонент. У пользователей есть возможность анализировать временные тренды и проводить сравнение с данными других участников при наличии соответствующих разрешений с их стороны. Так, для лучшего понимания глобальной ситуации можно показать на одной диаграмме медианные значения загрязнения тяжелыми металлами с граничащими странами и регионами.

Координаторам доступны все инструменты рядовых участников, кроме того, они могут осуществлять групповые операции с данными, получать сводные отчеты и строить глобальные карты загрязнений.

Прогнозирование — важный этап экологического мониторинга, позволяющий заполнять пробелы в данных. В рамках платформы реализован механизм прогнозирования, основанный на применении машинного обучения совместно с данными дистанционного зондирования земли.

Снимки разных спутниковых программ используются для получения так называемых индексов, которые являются дополнительными данными при обучении модели и основными — при построении прогноза. Для вычисления индексов используется платформа Google Earth Engine, содержащая данные десятков различных программ и продуктов (рис. 2).

Отдельные микросервисы платформы используются для сбора индексов, построения глобальных и локальных моделей, подбора оптимальных параметров и прогнозирования. В текущей реализации в зависимости от количества исходных данных используются статистические модели машинного обучения либо глубокие нейронные сети. Решаются задачи регрессии и классификации, но последние более приоритетны, так как появляется возможность использования методов балансировки обучающей выборки, и изначально при построении карт используется градация уровней загрязнения. При построении локальных и региональных карт некоторых элементов точность моделей достигает 90–95 % [3] (рис. 3).

В планах развития платформы предусматривается не только улучшение существующего функционала, но и предоставление новых возможностей. Например, большой интерес вызывает задача сбора и предоставления данных по заболеваемости населения, что позволило бы в рамках платформы проводить сравнение уровней загрязнения и количества определенных заболеваний в различных регионах.

forecast. The Google Earth Engine platform, containing data from dozens of different programs and products, is used to calculate the indexes (Fig. 2).

Platform microservices are used to collect indexes, build global and local models, select optimal parameters and predict contamination. In the current implementation, statistical machine learning models or deep neural networks are used depending on the amount of training data. We are focused on regression and classification tasks, however, classification is prioritized since it becomes possible to apply balancing techniques for training datasets, and a gradation of pollution levels is initially used when building maps. For local and regional maps of some elements, the model accuracy reaches 90–95% [3] (Fig. 3).

To develop the platform, it is planned to enhance the existing functionality, as well as to provide new opportunities. For example, the task of collecting and importing data on the morbidity of the population to the platform arouses great interest, which would enable the comparison of contamination levels and the number of certain diseases in different areas within the platform.

Список литературы / References

1. *Ужинский А.* Интеллектуальная платформа экологического мониторинга // Открытые системы. СУБД. 2021. № 2. С. 21–23.

Uzhinskiy A. Intelligent Environmental Monitoring Platform // Open Systems. DBMS. 2021. No. 2. P. 21–23 (in Russian).

2. *Uzhinskiy A., Ososkov G., Goncharov P., Nechaevskiy A., Smetanin A.* One-Shot Learning with Triplet Loss for Vegetation Classification Tasks // Comp. Opt. 2021. V. 45, No. 4. P. 608–614; doi: 10.18287/2412-6179-CO-856.

3. *Uzhinskiy A., Aničić Urošević M., Frontasyeva M.* Prediction of Air Pollution by Potentially Toxic Elements over Urban Area by Combining Satellite Imagery, Moss Biomonitoring Data and Machine Learning // Ciência e Técnica Vitivinícola J. 2020. V. 35, No. 12.