

Н. М. ТЕМИРБЕКОВ^{1*}, Д. Р. БАЙГЕРЕЕВ^{1,2}, М. Н. МАДИЯРОВ^{1,2}

¹Национальная инженерная академия Республики Казахстан,
Алматы, Казахстан

²Восточно-Казахстанский университет им. С. Аманжолова,
Усть-Каменогорск, Казахстан

e-mail: temirbekov@rambler.ru, dbaigereyev@gmail.com, madiyarov_mur@mail.ru

РАЗРАБОТКА ПОДСИСТЕМ ИНТЕГРАЦИИ И АНАЛИЗА ДАНЫХ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ ДЛЯ ПОДДЕРЖКИ И ПРИНЯТИЯ РЕШЕНИЙ В ОБЛАСТИ РЕГИОНАЛЬНОЙ ЭКОЛОГИЧЕСКОЙ ПОЛИТИКИ

В статье представлены результаты исследований, проведенных в рамках проекта программно-целевого финансирования BR18574148 Министерства науки и высшего образования Республики Казахстан, направленных на разработку единой информационно-аналитической платформы esomodeling.kz для поддержки и принятия решений в области региональной экологической политики. В настоящей статье приводится описание трех подсистем данной платформы, осуществляющих интеграцию платформы с различными информационными источниками, анализ данных мониторинга атмосферы промышленных городов Казахстана, прогнозирование значений концентрации вредных веществ с применением моделей машинного обучения и отображение результатов. В статье описаны технические аспекты получения данных измерения из двух информационных источников, структура базы данных, разработанные инструментальные средства для анализа данных, используемые модели машинного обучения, а также пользовательский интерфейс веб-приложения.

Ключевые слова: информационно-аналитическая платформа, загрязнение атмосферы, анализ данных, машинное обучение, программный интерфейс (API), картографический веб-интерфейс.

Введение. Существует множество информационных платформ, предоставляющих информацию о качестве воздуха и прогноз загрязнения воздуха в режиме реального времени. Данные платформы направлены на улучшение экологической обстановки региона и общественного здоровья посредством комплексной оценки данных мониторинга воздуха, предоставления практических рекомендаций для охраны здоровья, а также своевременного уведомления граждан о неблагоприятной экологической ситуации с целью принятия превентивных мер.

Например, IQAir [1] является одной из крупнейших информационных платформ, предоставляющих комплекс решений, направленных на улучшение качества воздуха. Наряду с информированием о текущих значениях концентрации загрязняющих веществ и индекса качества воздуха в выбранном городе, платформа предоставляет информацию о главных источниках загрязнения, тенденцию загрязнения, результаты исследования влияния экологической ситуации в городе на здоровье людей, а также предлагает практические рекомендации по улучшению качества воздуха. Кроме того, платформа позволяет проследить динамику изменения индекса качества воздуха на

* E-mail корреспондирующего автора: temirbekov@rambler.ru

основе исторических данных, сделать его прогноз на недельный период и предлагает ранжированный список городов по данному показателю в разрезе выбранной страны или всего мира.

Проект «Всемирный индекс качества воздуха» [2, 3] имеет аналогичные цели и предоставляет неограниченный доступ к мировым историческим данным и текущим данным о качестве воздуха от 30000 станций из более чем 2000 крупных городов в 130 странах в режиме реального времени через сайты aqicn.org и waqi.info. Аналогичные функциональные возможности предоставляет платформа air-quality.com [4].

Проект airly.org [5] предлагает набор инструментов для анализа данных. Наряду с представлением отчетности о загрязнении воздуха, платформа предлагает прогноз загрязнения воздуха на 24 часа с точностью 95% на основе каскадной системы нейронных сетей. Используется предиктивный алгоритм с применением машинного обучения исторических данных и прогнозных метеорологических параметров.

Принимая во внимание актуальность проблемы, авторами статьи поставлена цель разработать единую цифровую информационно-аналитическую платформу для поддержки и принятия решений в области региональной экологической политики. Настоящая статья открывает цикл публикаций, посвященных описанию технических аспектов разрабатываемой платформы. В частности, в данной статье приводится описание трех ее подсистем, которые осуществляют интеграцию платформы с различными информационными источниками, позволяют проводить анализ данных мониторинга атмосферы промышленных городов Казахстана, прогнозирование значений концентрации вредных веществ с применением моделей машинного обучения, а также отображение результатов мониторинга. Разрабатываемая платформа имеет несколько ключевых отличий от созданных ранее аналогичных платформ. Во-первых, расчет прогнозных значений концентрации загрязняющих веществ осуществляется на основе комбинированного использования математического аппарата и моделей машинного обучения. Во-вторых, платформа предоставляет набор инструментов для проведения анализа данных и машинного обучения с использованием данных о загрязнении воздуха в промышленных городах Казахстана.

Методы и материалы. Разрабатываемая информационно-аналитическая платформа состоит из пяти подсистем. В настоящей статье приводится описание следующих трех ее подсистем:

- 1) подсистемы интеграции платформы с различными информационными источниками;
- 2) подсистемы анализа данных;
- 3) подсистемы отображения данных измерений и прогноза.

Первая подсистема предназначена для получения, обработки и хранения данных измерений концентрации загрязняющих веществ и метеорологических параметров из различных информационных источников. Основным поставщиком данных по концентрациям загрязняющих веществ в атмосфере является ТОО «Экосервис-С», который предоставляет доступ к данным посредством программного интерфейса (API). Указанные данные поступают от автоматизированных систем мониторинга эмиссий, которые предназначены для непрерывного измерения концентраций в газовых выбросах в режиме реального времени, а также расчета валового объема выбросов [6].

Программный интерфейс состоит из нескольких конечных точек, позволяющих получить перечень автоматизированных постов мониторинга с указанием географических координат, измеренные данные концентрации загрязняющих веществ с постов мониторинга, перечень загрязняющих веществ со значениями предельно допустимых концентраций, индекс качества воздуха, вычисленный на каждом посту мониторинга, и другую информацию. Получение указанных выше данных из ТОО «Экосервис-С» осуществляется с использованием программы, написанной на высокопроизводительном языке программирования Julia, которая запускается с интервалом в 20 минут.

Информация, получаемая из ТОО «Экосервис-С» содержит значения концентрации следующих веществ на каждом посту наблюдения:

- оксид азота (NO_2);
- диоксид серы (SO_2);
- оксид углерода (CO);
- взвешенные частицы $\text{PM}_{2.5}$ и PM_{10} ;

а также следующую метеорологическую информацию:

- температура воздуха;
- направление и скорость ветра;
- атмосферное давление;
- относительная влажность воздуха.

Вторым источником данных является РГП «Казгидромет», предоставивший исторические данные значений концентраций загрязняющих веществ в файлах формата CSV. Импорт полученных данных в базу данных осуществлен с помощью программы, написанной на языке Julia.

В результате база данных информационно-аналитической платформы содержит данные о загрязнении атмосферы с 2005 года. Данная подсистема хранит данные в следующих основных таблицах:

- таблица провайдеров данных;
- таблица городов;
- таблица автоматизированных постов наблюдения;
- таблица наблюдаемых параметров;
- таблица единиц измерений;
- таблица значений измерений

и других.

В работе используется свободная объектно-реляционная система управления базами данных PostgreSQL. Ее выбор обоснован гибкостью, высокопроизводительностью транзакций и репликации, возможностью индексирования геометрических объектов и наличием базирующегося на ней расширения PostGIS, поддержкой слабоструктурированных данных с возможностью индексирования, а также возможностью создания пользовательских типов.

Назначение второй подсистемы информационно-аналитической платформы состоит в предоставлении набора инструментов для проведения анализа данных. Данная подсистема имеет модульную структуру, каждый модуль которой направлен на решение определенной задачи.

Первый модуль содержит методы для выполнения общего анализа данных, составления матрицы корреляции между параметрами для выявления скрытых закономерностей. В частности, модуль вычисляет статистические характеристики набора данных, такие как медиана, стандартное отклонение, минимальное и максимальное значение и квартили. Результаты вычислений матрицы корреляции существенно используются при выборе признаков в моделях машинного обучения, поскольку выбор признаков является критическим этапом для эффективного обучения модели. Матрица корреляции составляется индивидуально для каждого поста наблюдения.

Суть второго модуля заключается в выявлении закономерности в изменении концентрации загрязнителей на основе анализа архивных данных, с целью ее использования при прогнозе будущих значений. Работа модуля осуществляется в несколько этапов. Выполнение первого этапа опирается на применение моделей машинного обучения. На текущий момент реализованы три ансамблевые методики обучения. Первая модель, XGBoost, представляет собой реализацию алгоритма стохастического повышения градиента. Это алгоритм дерева решений ансамбля, в котором новые деревья исправляют ошибки тех деревьев, которые уже являются частью модели. Деревья добавляются до тех пор, пока в модель нельзя будет вносить дальнейшие улучшения. Вторая модель, LightGBM представляет собой структуру повышения градиента, использующую алгоритмы обучения на основе дерева. Она предназначена для распределения и эффективности и нацелена на более высокую скорость обучения и более высокую эффективность, меньшее использование памяти, повышенную точность, поддержку параллельного, распределенного и графового обучения, возможности обработки больших данных. Согласно некоторым исследованиям, LightGBM в большинстве случаев обучается быстрее, чем XGBoost. Третья модель – повышение градиента на основе гистограммы (HistGradientBoosting), представляет собой реализацию деревьев повышения градиента. Как и LightGBM, данная модель также имеет встроенную поддержку отсутствующих значений, что позволяет избежать необходимости в импутере. В результате выполнения первого этапа вычисляются показатели R^2 , MAE, MSE и RMSE, которые характеризуют степень обученности модели.

Второй этап заключается в вычислении прогнозных значений концентрации на каждом автоматизированном посту наблюдения на основе обученной модели. Результаты данного этапа используются в подсистеме моделирования рассеивания загрязняющих веществ.

Третий модуль подсистемы заключается в выявлении возможной сезонности, влияющей на повышение концентрации того или иного вещества. Алгоритм реализован индивидуально для каждого поста мониторинга и стремится выявить зависимость изменения концентрации от четырех факторов:

- времени суток;
- дня недели;
- месяца;
- квартала года.

Четвертый модуль предназначен для проведения статистических расчетов на основе получаемых данных измерений.

Третья подсистема информационно-аналитической платформы, рассматриваемая в данной статье, предназначена для предоставления пользовательского веб-интерфейса ко всем функциям и модулям, входящих в состав информационно-аналитической платформы, а также для графического отображения данных измерений и прогнозных расчетов. Веб-приложение разработано с использованием веб-фреймворка Django REST Framework. Обмен данными между различными частями веб-приложения осуществляется посредством Web API.

Результаты и их обсуждение. Приведем описание пользовательского интерфейса разработанной информационно-аналитической платформы. Веб-приложение состоит из нескольких разделов, ссылки на которые расположены в верхней части страницы.

Отображение данных измерений загрязняющих веществ приводится в разделе «Карта». На карте маркерами обозначены автоматизированные пункты наблюдения (Рисунок 1). Цвет маркеров зависит от концентрации на соответствующем посту наблюдения: при превышении концентрации однократного и пятикратного допустимого показателя маркер закрашивается, соответственно, в желтый и красный цвета. В остальных случаях отображается зеленый маркер. Кроме того, окрестность маркеров покрывается градиентным рисунком, отображающим градиацию изменения значений концентрации.



Рисунок 1 – Отображение данных измерений концентрации

С помощью выпадающего списка, расположенного в правом верхнем углу, пользователь имеет возможность выбрать город. В настоящий момент платформа располагает данными мониторинга по городам Алматы, Шымкент, Темиртау, Караганда, Усть-Каменогорск и Жанатас. В левом верхнем углу расположен переключатель, с помощью которого выбирается один из пяти параметров наблюдения – $PM_{2.5}$, PM_{10} , SO_2 , NO и CO .

При нажатии на маркер отображается всплывающее окно с текущим значением концентрации на соответствующем посту наблюдения (Рисунок 2). При переходе по ссылке «Детали», расположенной в данном окне, появляется модальное окно, в котором можно проследить динамику изменения концентрации выбранного вещества на данном посту за одни сутки. Кроме того, в верхней части модального окна отображены суточные и разовые предельно допустимые значения концентрации.



Рисунок 2 – Динамика изменения концентрации на посту наблюдения

В нижней части окна расположена кнопка «Детали» для вывода таблицы значений концентрации на всех постах наблюдения выбранного города. Строки таблицы выделены зеленым, желтым или красным цветами в зависимости от отношения текущих значений к предельно допустимому значению.

Кроме того, в нижней части страницы расположены переключатели с временными отметками. При нажатии на кнопки данного переключателя пользователь имеет возможность просмотреть значения, соответствующие выбранной временной отметке. Имеется возможность запустить анимацию, при которой можно проследить динамику изменения состояния воздуха за последние 6 часов с интервалом в 20 минут.

Цель раздела приложения «Анализ» состоит в предоставлении пользователю набора инструментов для проведения анализа данных и машинного обучения на исторических данных. Данный раздел имеет следующие функциональные возможности:

а) Просмотр данных измерений в табличном виде с возможностью фильтра по городам, постам наблюдения, произвольному периоду наблюдения, а также экспорта полученных значений в файл формата CSV.

б) Статистика получаемых данных измерений в разрезе городов, дней, месяцев, а также статистика получаемых данных в рамках одного запроса.

в) Построение корреляционной матрицы, показывающей зависимость между веществами и метеорологическими данными по каждому посту наблюдения в заданном периоде наблюдения (Рисунок 3).

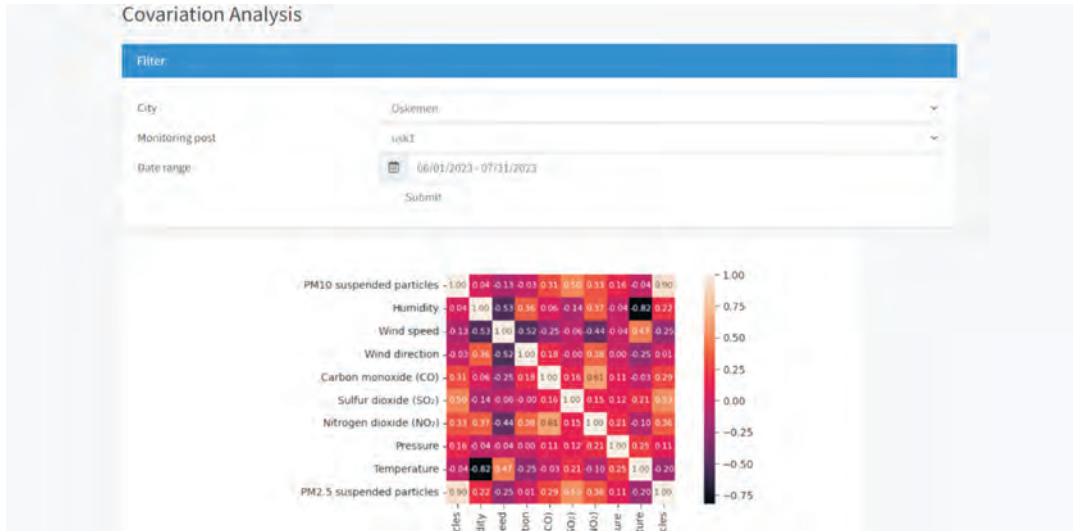


Рисунок 3 – Матрица корреляции

г) Выявление возможных сезонных закономерностей изменения концентрации в разрезе часов (Рисунок 4), дней недели (Рисунок 5), квартала, месяца. Входными параметрами алгоритма являются пост наблюдения, период наблюдения и параметр группировки. Результатом работы алгоритма является блочная диаграмма с ограничителями выбросов, компактно иллюстрирующая медиану, нижний и верхний квартили, минимальное и максимальное значение выборки, а также выбросы.



Рисунок 4 – Определение возможной сезонной закономерности

д) Обучение моделей машинного обучения XGBoost, LightGBM, HistGradient-Boosting на основе данных измерений и прогноз значений концентрации на каждом посту наблюдения на основе обученной модели. Входными параметрами алгоритма являются пост наблюдения, период наблюдения и параметры модели машинного обучения. Результатом вычислений являются коэффициент детерминации R^2 , среднеквадратическая ошибка, средняя абсолютная ошибка, результат валидации обученной модели на тестовых данных и прогноз концентрации.



Рисунок 5 – Определение возможной сезонной закономерности

Заключение. Таким образом, в разрабатываемой информационно-аналитической платформе реализован ряд инструментов для проведения анализа данных мониторинга атмосферы промышленных городов Казахстана, которые будут использованы в других подсистемах платформы. В следующем цикле статей будет приведено описание подсистемы моделирования рассеивания загрязняющих веществ на основе комбинированного использования математических методов и моделей машинного обучения, осуществляющей интеллектуальный анализ сценариев развития экологической обстановки в промышленных городах Казахстана.

Благодарности. Данное исследование выполнено при поддержке Министерства науки и высшего образования Республики Казахстан по программно-целевому финансированию ИРН BR18574148 «Развитие геоинформационных систем и мониторинга объектов окружающей среды».

ЛИТЕРАТУРА

- 1 IQAir – About IQAir. <https://iqair.com/about-iqair> (10.07.2023).
- 2 AQICN – aqicn.org (10.07.2023).
- 3 WAQI – waqi.info (10.07.2023).
- 4 Air Quality – air-quality.com (10.07.2023).
- 5 Airly – airly.org (10.07.2023).
- 6 Автоматизированные системы мониторинга – <https://ecoservice.kz/services/cat11/> (10.07.2023).

REFERENCES

- 1 IQAir – About IQAir. <https://iqair.com/about-iqair> (10.07.2023).
- 2 AQICN – aqicn.org (10.07.2023).
- 3 WAQI – waqi.info (10.07.2023).
- 4 Air Quality – air-quality.com (10.07.2023).
- 5 Airly – airly.org (10.07.2023).
- 6 Avtomatizirovannyye sistemy monitoringa – <https://ecoservice.kz/services/cat11/> (10.07.2023)

Н. М. ТЕМИРБЕКОВ¹, Д. Р. БАЙГЕРЕЕВ^{1,2}, М. Н. МАДИЯРОВ^{1,2}

¹Қазақстан Республикасының Ұлттық инженерлік академиясы, Алматы, Қазақстан
²С. Аманжолов атындағы Шығыс Қазақстан университеті, Өскемен, Қазақстан

Өңірлік экологиялық саясат саласында ҚОЛДАУ ЖӘНЕ ШЕШІМДЕР ҚАБЫЛДАУ ҮШІН АҚПАРАТТЫҚ-ТАЛДАУ ПЛАТФОРМАСЫНЫҢ ИНТЕГРАЦИЯЛАНДЫРУ ЖӘНЕ ДЕРЕКТЕРДІ ТАЛДАУ ІШКІ ЖҮЙЕЛЕРІН ӨЗІРЛЕУ

Мақалада Қазақстан Республикасы ғылым және жоғары білім министрлігінің BR18574148 бағдарламалық-мақсатты қаржыландыру жобасы аясында аймақтық экологиялық саясат саласында қолдау көрсету және шешімдер қабылдау үшін ecomodeling.kz бірыңғай ақпараттық-талдау платформасын әзірлеуге бағытталған зерттеулердің нәтижелері берілген. Бұл мақалада платформаны әртүрлі ақпарат көздерімен біріктіретін, Қазақстандағы өнеркәсіптік қалалар атмосферасының мониторингі деректерін талдайтын, машиналық оқыту үлгілерін пайдалана отырып, зиянды заттардың концентрациясының мәндерін болжайтын және нәтижелерін көрсететін үш ішкі жүйесі сипатталған. Мақалада екі ақпарат көзінен өлшеу деректерін алудың техникалық аспектілері, деректер қорының құрылымы, деректерді талдау үшін әзірленген құралдар, қолданылатын машиналық оқыту үлгілері және веб-қосымшаның пайдаланушы интерфейсі сипатталған.

Түйін сөздер: ақпараттық-аналитикалық платформа, ауаның ластануы, деректерді талдау, машиналық оқыту, бағдарламалау интерфейсі (API), веб-карталау интерфейсі.

N. M. TEMIRBEKOV¹, D. R. BAIGEREYEV^{1,2}, M. N. MADIYAROV^{1,2}

¹National Engineering Academy of the Republic of Kazakhstan, Almaty, Kazakhstan
²S. Amanzholov East Kazakhstan University, Ust-Kamenogorsk, Kazakhstan

DEVELOPMENT OF INTEGRATION AND DATA ANALYSIS SUBSYSTEMS OF INFORMATION AND ANALYTICAL PLATFORM FOR SUPPORT AND DECISION-MAKING IN THE FIELD OF REGIONAL ENVIRONMENTAL POLICY

The paper presents the results of research conducted within the framework of the program-targeted financing project BR18574148 of the Ministry of Science and Higher Education of the Republic of Kazakhstan, aimed at developing a unified information and analytical platform ecomodeling.kz to support and make decisions in the field of regional environmental policy. This paper describes three subsystems

of this platform that integrate the platform with various information sources, analyze the monitoring data of the atmosphere of industrial cities in Kazakhstan, predict the values of the concentration of harmful substances using machine learning models and display the results. The article describes the technical aspects of obtaining measurement data from two information sources, the structure of the database, the developed tools for data analysis, the machine learning models used, and the user interface of the web application.

Key words: *information and analytical platform, air pollution, data analysis, machine learning, programming interface (API), web mapping interface.*