UDC 004.8:551.510.42
IRSTI 87.15.23
https://doi.org/10.53364/24138614_2026_41_2_12

A.E. Tukushova[1*], S.Zh. Rakhmetullina[1], Zh.S. Serikova[2], A.T. Ualkhanova[2]

[1]East Kazakhstan Technical University named after D. Serikbayev,
Ust-Kamenogorsk, Kazakhstan
[2]East Kazakhstan University named after S. Amanzholov, Ust-Kamenogorsk, Kazakhstan

[*]*E-mail: araylym.tukushova@mail.ru*

DEVELOPMENT OF AN ALGORITHM FOR DETECTING ANOMALIES IN THE AIR POLLUTION MONITORING SYSTEM

Abstract. *In recent years, the problem of air pollution has become more and more acute, especially for industrial regions. The constant growth of environmental monitoring data requires not only their accumulation, but also effective intelligent processing. One of the key tasks is the timely detection of abnormal values that can indicate both real emissions of pollutants and errors in measuring systems.*

In this paper, an algorithm for detecting anomalies in the atmospheric air monitoring system is proposed, based on a combination of statistical methods and machine learning algorithms. This approach allows you to take into account both simple emissions and more complex, hidden patterns in the data. For primary filtration, the methods of Z-score and interquartile range (IQR) were used, and for a more in-depth analysis, the Isolation Forest algorithm was used, which is able to effectively work with multidimensional ecological time series. The novelty of the study lies in the hybrid decision procedure that combines statistical filtering, unsupervised anomaly detection and meteorological-context interpretation for industrial air pollution monitoring data.

Particular attention is paid to the construction of the system architecture, which is implemented using cloud technologies. This provides the ability to process large amounts of data coming from monitoring sensors, as well as analyze them in near real time.

The algorithm was tested on data from the city of Ust-Kamenogorsk, including indicators of the concentration of the main pollutants and meteorological parameters. The results showed that the proposed hybrid approach achieved higher performance than individual methods, reaching Precision = 0.94, Recall = 0.91 and F1-score = 0.92. At the same time, the system is able to automatically record sharp deviations associated with industrial emissions, weather conditions or technical failures.

The practical significance of the work lies in the possibility of introducing the proposed algorithm into environmental information systems and smart city solutions. Its application makes it possible to improve the quality of monitoring, the efficiency of response and the validity of management decisions in the field of environmental protection.

Keywords: *atmospheric air monitoring, data anomalies, machine learning, isolation forest, environmental monitoring, cloud technologies.*

Introduction.

In modern conditions of intensive urbanization and industrial development, air pollution has become one of the most significant environmental problems for industrial cities. The growth of

industrial emissions, transport intensity and the use of various fuel sources contribute to an increase in the concentration of harmful substances in the atmosphere. This negatively affects public health, the ecological state of urban territories and the sustainable development of industrial regions. Therefore, the monitoring of atmospheric air quality and the timely detection of dangerous deviations in pollutant concentrations are important tasks of environmental management and decision support.

Traditional air quality monitoring systems based on stationary measuring stations provide valuable information about the concentration of pollutants. However, these systems have several limitations, including limited spatial coverage, the high cost of equipment, delays in data processing and difficulties in analyzing large volumes of heterogeneous environmental data. In addition, monitoring data may contain abnormal values caused not only by real increases in pollutant emissions, but also by unfavorable meteorological conditions, sensor failures, data transmission errors or measurement noise. As a result, the automatic detection and interpretation of anomalies in environmental time series remains an important scientific and practical problem.

Recent studies show that cloud computing, big data technologies and machine learning methods create new opportunities for environmental monitoring. Cloud-based architectures provide scalability, fault tolerance and distributed data processing capabilities [1, 2]. Big Data technologies make it possible to process large flows of environmental information and support near-real-time analysis [3]. Microservice-based architectures also increase the flexibility and modularity of environmental monitoring platforms [4]. In air pollution analysis, statistical models, regression methods, decision trees, random forest, gradient boosting, support vector regression and neural network models have been widely used for forecasting pollutant concentrations and analyzing complex nonlinear dependencies [5–10].

However, most existing studies focus mainly on predicting air pollution levels, while the problem of detecting and interpreting anomalies in long-term environmental monitoring data remains insufficiently developed. In particular, there is a lack of reproducible approaches that combine statistical anomaly detection methods with unsupervised machine learning algorithms and take into account the meteorological context of industrial cities. This research gap is especially relevant for regions with high industrial load, where abnormal pollutant concentrations may be caused by industrial emissions, weather-related accumulation effects or technical failures of monitoring equipment.

The purpose of this study is to develop and test a hybrid algorithm for detecting anomalies in atmospheric air pollution monitoring data. The proposed approach combines statistical methods, including Z-score and interquartile range analysis, with the Isolation Forest machine learning algorithm. This combination makes it possible to identify both simple extreme values and more complex hidden anomalous patterns in multidimensional environmental time series.

The scientific contribution of the study is threefold. First, a hybrid anomaly detection pipeline combining Z-score, IQR and Isolation Forest is proposed for the analysis of atmospheric air monitoring data. Second, the algorithm is adapted to long-term monitoring data of an industrial city and includes the interpretation of anomalies in relation to meteorological conditions and possible sensor-related errors. Third, the anomaly detection procedure is integrated into a cloud-based monitoring architecture, which supports data preprocessing, anomaly identification, event recording, visualization and near-real-time decision support.

The practical significance of the study lies in the possibility of using the developed algorithm in environmental information systems, air quality monitoring platforms and smart city infrastructure. The proposed approach can improve the reliability of monitoring data, support the timely detection of dangerous environmental situations and increase the efficiency of management decisions in the field of atmospheric air protection.

A separate category is made up of anomalous trends associated with long-term deviations in air quality indicators. Such anomalies may indicate a systematic increase in pollutant emissions or

changes in meteorological conditions. To identify them, time-series analysis methods and machine learning algorithms are used.

Table 1 – Types of anomalies in atmospheric monitoring data and methods of their elimination

Type of anomaly	Causes of occurrence	Detection methods	Methods of elimination
Outliers	Sensor errors, measurement noise, extreme pollutant emissions	Z-score, IQR, Isolation Forest	Filtering, smoothing, replacing with median value
Missing data	Hardware failure, packet loss, technical issues	Time series analysis, data completeness check	Interpolation, Moving Average, Median Fill
Temporary inconsistencies	Sensor asynchrony, timestamp errors	Comparison of time intervals, correlation analysis	Time synchronization, data aggregation
Anomalous trends	Long-term changes in the environmental situation, changes in emission sources	Time Series Analysis, Machine Learning Algorithms	Recalibration of sensors, adjustment of models
Data transmission errors	Network failures, data corruption	Checksum control, logging	Data retransfer, data set cleanup

Thus, the correct identification and processing of anomalies makes it possible to increase the reliability of monitoring data and improve the quality of models for predicting air pollution.

Table 2 shows the algorithms for detecting anomalies in atmospheric air monitoring data.

Table 2 – Algorithms for detecting anomalies in atmospheric air monitoring data

Method	Algorithm type	Principle of operation	Benefits	Limitations	Air Monitoring Applications
Z-score	Statistical	Evaluates the deviation of a value from the mean through standard deviation	Easy to implement, high speed of calculations	Sensitive to outliers and data distribution	Detection of sudden spikes in pollutant concentrations
IQR (Interquartile Range)	Statistical	Determines outliers based on the interquartile range of the distribution	Resistant to noise and extreme values	Limited Efficacy for Complex Time Dependencies	Clearing sensor data
Isolation Forest	Machine learning	Isolates anomalous points in random trees	Effective for large data sets	Requires parameter settings	Detection of anomalous environmental events

Local Outlier Factor (LOF)	Machine learning	Compares the density of a point with the density of its neighbors	Detects local anomalies well	High computational complexity	Analysis of spatial environmental data
Autoencoder	Deep learning	The neural network recovers data and identifies reconstruction errors	Suitable for complex nonlinear data	Requires a large amount of data	Detection of complex anomalies
LSTM	Deep learning	Analyzes temporal dependencies in data	Effective for time series	High computational complexity	Forecast and detection of pollution anomalies

In ambient air environmental monitoring systems, data come from a variety of sources, including stationary air quality sensors, weather stations and satellite observations. In the process of collecting and transmitting information, various anomalies can occur that distort the results of the analysis and reduce the accuracy of pollution forecasting. Therefore, the identification and correct processing of such anomalies is an important step in data pre-processing.

Anomalies in environmental data can be caused by various factors. These include technical failures of measuring instruments, data transmission errors, the impact of extreme meteorological conditions or actual abrupt changes in pollutant concentrations due to accidental releases. Depending on the nature of the occurrence, anomalies can be divided into several main types: outliers, missing data, temporal inconsistencies and anomalous trends.

Outliers are sharp deviations of measured parameters from typical ranges. They can occur due to sensor failure, measurement noise, or extreme atmospheric processes. Statistical methods are used to identify such anomalies, including interquartile range analysis, z-scoring, and machine learning algorithms.

Missing values are a common problem in environmental datasets. They can occur due to temporary equipment shutdowns, data transmission interruptions, or registration errors. To eliminate gaps, interpolation, moving average, or machine learning models are used.

Another type of anomalies is temporal inconsistencies that occur when data are recorded asynchronously by different sensors. Such errors can lead to incorrect interpretation of relationships between environmental parameters. To eliminate this problem, timestamp synchronization and data aggregation over single time intervals are used [6, 7, 11].

Materials and methods of research.

The materials of the study were scientific publications of domestic and foreign authors devoted to the problems of environmental monitoring of atmospheric air, the use of machine learning methods for the analysis of environmental data, as well as the development of information systems for environmental monitoring. In addition, the data of observations of atmospheric air quality obtained by automated monitoring stations located in various districts of Ust-Kamenogorsk were used. The study also took into account meteorological parameters affecting the processes of distribution and accumulation of pollutants in the atmosphere.

The methodological basis of the study is the system and information-analytical approaches, which allow us to consider the system of atmospheric air monitoring as a complex multi-level system for collecting, processing and analyzing environmental data. In the course of the study, the methods of analysis and generalization of scientific sources, comparative analysis of existing methods for predicting air pollution, as well as methods of data mining and machine learning were used.

To develop an algorithm for detecting anomalies, methods of statistical analysis of time series and machine learning algorithms were used to detect deviations from typical patterns of changes in pollutant concentrations. Data analysis was carried out using Python and Julia programming languages, as well as specialized libraries for data processing and building machine learning models.

Processing and preparation of initial data included the stages of data cleaning, filling in missing values, normalizing parameters and forming a feature space. To analyze the data and build models, regression analysis methods, decision tree algorithms, ensemble methods and neural network models were used, which make it possible to identify complex nonlinear dependencies between pollutant concentrations and meteorological factors.

To store and manage data, a relational PostgreSQL database was used, which provides centralized storage of environmental information. The architecture of the developed system is based on a microservice approach and includes modules for data collection, analytical processing, anomaly detection, and visualization of results. Interaction between the components of the system is implemented using REST API.

To evaluate the feasibility of cloud deployment, the algorithm was tested in batch processing mode. The evaluation included data loading time, preprocessing time, anomaly detection time, database writing time and API response time. This made it possible to assess whether the proposed architecture can be used for near-real-time monitoring.

The effectiveness of the developed algorithm was evaluated on the basis of an analysis of the accuracy of anomaly detection and comparison of the model results with the actual values of pollutant concentrations. Statistical indicators of the model quality, including mean absolute error (MAE), root mean square error (RMSE) and coefficient of determination (R^2), were used as evaluation criteria.

The methods and software used ensure the reproducibility of the study and make it possible to use the developed algorithm as part of intelligent environmental monitoring systems aimed at improving the efficiency of atmospheric air quality control and timely detection of hazardous environmental situations.

Statistical methods (Z-score, IQR) are used for primary data filtering, while machine learning algorithms (Isolation Forest, LOF) and deep learning methods (Autoencoder, LSTM) are used to identify complex anomalies and analyze environmental data time series.

Methods for detecting anomalies:

1. Z-score method

The z-score is used to identify outliers in the data based on the deviation of the value from the mean.

Formula:

$$Z = \frac{x-\mu}{\sigma} \quad (1)$$

where

x - observed value;

μ - average sample value;

σ - standard deviation.

If

$$|Z| > 3 \quad (2)$$

The value is considered anomalous.

This method is widely used to detect sharp emissions of pollutant concentrations (PM2.5, NO_2, SO_2, etc.).

2. Interquartile Sweep (IQR)

The IQR method allows you to identify outliers based on the distribution of data.

Formulas:

$$OQR = Q_3 - Q_1 \quad (3)$$

where

Q_1 - first quartile (25th percentile),

Q_3- Third quartile (75th percentile).

Normal value limits:

$$L_{ower} = Q_1 - 1.5 \times IQR \quad (4)$$

$$U_{pper} = Q_3 + 1.5 \times IQR \quad (5)$$

If

$$x < L_{ower} \text{ или } x > U_{pper} \quad (6)$$

The value is considered an anomaly.

IQR is noise-resistant and is often used in environmental time-series analysis.

3. Isolation Forest Method for Anomaly Detection

The Isolation Forest method refers to machine learning algorithms designed to detect anomalous observations in multidimensional data. The basic idea behind the algorithm is that anomalous points are easier to isolate in feature space compared to normal observations.

The algorithm builds an ensemble of random binary isolation trees in which the data is sequentially divided randomly according to selected features. Because anomalies are different from most observations, they are isolated earlier in the construction of the tree, resulting in a shorter path length from the tree root to the corresponding node.

Building an Isolation Tree

Each tree is formed as follows:

The characteristic q is randomly selected.

A random threshold value of ppp is determined within the range of values of the selected characteristic.

The data is divided into two subsets:

$$x_q < p \quad (7)$$

$$x_q \geq p \quad (8)$$

The process is repeated recursively until the maximum depth of the tree is reached, or until there is only one observation left in the node.

Assessment of anomaly. The main indicator is the length of the isolation path of observation in the tree.

Let's outline:

$$h(x) \quad (9)$$

- The length of the observation path xxx from the tree root to the leaf.

For an ensemble of ttt trees, the average path length is calculated

where

$$E\big(h(x)\big) = \frac{1}{t}\sum_{i=1}^{t} h_i(x) \quad (10)$$

Normalizing coefficient

For correct evaluation, normalization is used through the

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (11)$$

where

$$H(n) = \ln(n) + \gamma \quad (12)$$

- harmonic number,

$\gamma \approx 0.57721$ is the Euler constant.

Anomaly Assessment Function

The final score of the anomaly is determined by the formula

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (13)$$

where
$s(x,n)$- indicator of abnormal observation.
Interpretation of the meaning (Table 3):

Table 3 - Interpretation of the score value

Score value	Interpretation
$s(x,n) \approx 1$	High probability of anomaly
$s(x,n) \approx 0.5$	Indeterminate state
$s(x,n) < 0.5$	normal observation

The Isolation Forest method is effective for analyzing environmental monitoring data, as it allows you to identify abnormal values of pollutant concentrations without the need for preliminary data labeling.

In atmospheric air monitoring tasks, the algorithm can be used to detect:

- sharp jumps in pollutant concentrations;
- errors of measuring sensors;
- anomalous emissions from industrial enterprises;
- non-standard meteorological conditions.

The use of an ensemble of isolation trees makes it possible to effectively analyze large amounts of environmental data and identify anomalies in real time, which makes this method promising for use in cloud-based air quality monitoring systems [7-11].

The Isolation Forest algorithm detects anomalies based on the isolation principle of observations. First, features and thresholds are randomly selected from the source dataset, and then an ensemble of isolation trees is constructed. At each step, the data is recursively separated until the observations are isolated. Anomalous points have a shorter path in the tree because they separate faster than normal observations. After the tree ensemble is constructed, the average path length for each observation is calculated and the anomaly score is calculated from it. Figure 1 shows how the Isolation Forest algorithm works to detect anomalies in ambient air monitoring data.

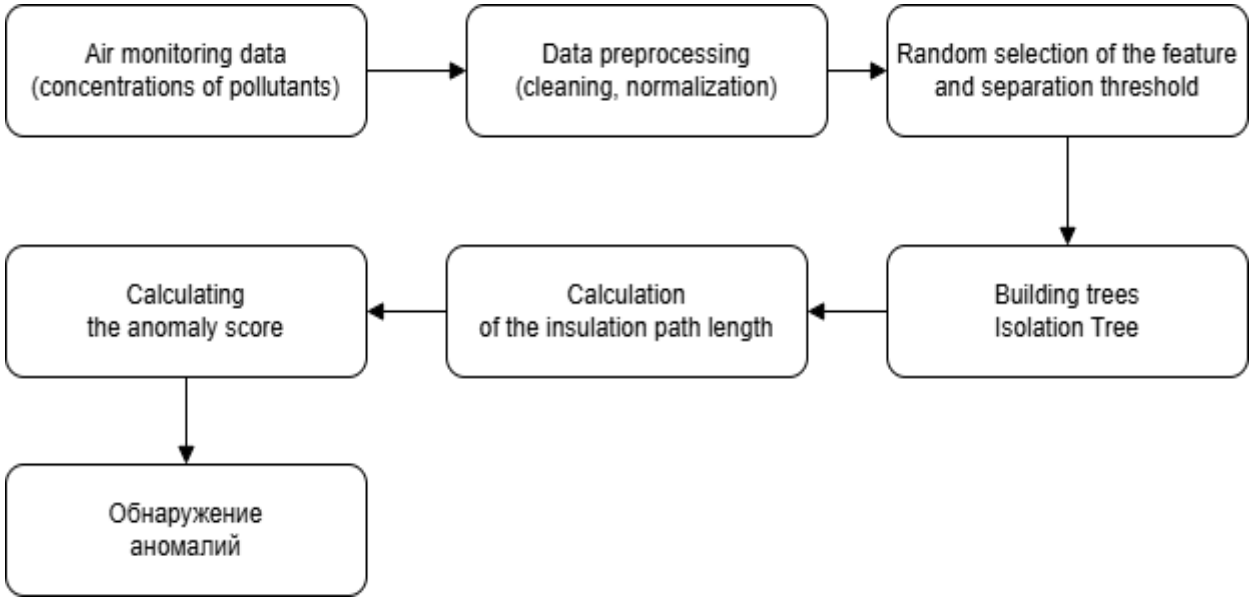

Figure 1 – Diagram of the Isolation Forest algorithm for detecting anomalies in ambient air monitoring data.

In the first step, the algorithm obtains ambient air monitoring data, including pollutant concentrations and meteorological parameters. Next, the data is pre-processed, including cleaning up missing values and normalizing features. Following this, features and thresholds for data separation are randomly selected. Based on these divisions, an ensemble of isolation trees is constructed. For each observation, the length of the isolation path in the trees is calculated. Based on the average path length, an anomaly score is calculated to determine whether the observation is anomalous. A comparison of anomaly detection methods is presented in Table 4.

Table 4 – Comparison of anomaly detection methods

Method	Method Type	Basic principle	Benefits	Limitations
Z-score	Statistical	Determines the deviation of a value from the mean through standard deviation	Simple and fast calculations	Sensitive to outliers and data distribution
IQR	Statistical	Uses interquartile range to determine outliers	Resistant to extreme values	Limited for complex time dependencies
Isolation Forest	Machine learning	Isolates anomalous sightings with random trees	Effective for large data sets	Requires parameter settings

A comparison of anomaly detection methods shows that statistical methods (Z-score, IQR) are effective for primary data filtering, whereas machine learning algorithms such as Isolation Forest can identify more complex anomalous patterns in multidimensional environmental data.

The material for the experimental verification of the algorithm was the data of monitoring of the atmospheric air of the city of Ust-Kamenogorsk. Time series of concentrations of the main pollutants (PM2.5, PM10, NO_2, SO_2, CO) obtained from the city environmental monitoring stations were used.

The methodology was tested on long-term observations of air quality in Ust-Kamenogorsk. The initial data were collected by five automated stations located at different points of the industrial city: thus, both industrial and residential areas were covered. Ust-Kamenogorsk is characterized by a high environmental burden due to industrial enterprises in the north, intensive traffic and individual heating systems in the private sector. The current location of the sensors allows you to get an objective picture of the distribution of harmful impurities in different areas of the city.

To ensure the reproducibility of the experimental evaluation, a detailed description of the environmental monitoring dataset used in this study is provided. The dataset includes long-term atmospheric pollution observations collected from automated monitoring stations in Ust-Kamenogorsk. Information about the monitoring period, pollutants, meteorological parameters, preprocessing procedures, and normalization methods is summarized in Table 5.

Table 5 – Description of the experimental dataset

Parameter	Description
Full monitoring period	2005–2025
Original temporal resolution	Hourly measurements
Number of stations	5 automated monitoring stations
Estimated raw data volume	approximately 876 000 station-hour records before filtering
Experimental dataset used for model evaluation	15 000 cleaned and balanced records

Reason for using subset	records with complete pollutant and meteorological parameters after quality filtering
Pollutants in the initial dataset	PM2.5, PM10, NO_2, SO_2, CO, O_3, NO
Pollutants used in experimental comparison	PM2.5, PM10, NO_2, SO_2
Meteorological parameters	temperature, atmospheric pressure, relative humidity, wind speed, wind direction
Data owner/source	Regional environmental monitoring system and Kazhydromet open monitoring data
Missing value processing	Linear interpolation and median filling
Normalization method	Z-score standardization

Although the original monitoring archive covers the period from 2005 to 2025 and includes hourly measurements from five stations, the experimental evaluation was conducted on a cleaned and balanced subset of 15,000 records. This subset included observations with complete pollutant concentrations and meteorological parameters after removing corrupted, incomplete and duplicated records. Therefore, the value of 15,000 records refers to the experimental dataset used for model evaluation, not to the full raw monitoring archive.

The dataset combines atmospheric pollution measurements and meteorological observations collected from different monitoring locations within the industrial area of Ust-Kamenogorsk. Data preprocessing included missing value handling, normalization, and removal of corrupted records prior to anomaly detection and machine learning analysis. The selected preprocessing procedures improved the stability of the models and ensured consistency of the environmental time-series data.

The initial dataset covered the period from 2005 to 2025 and included pollutant concentrations, meteorological variables and weather condition codes affecting the dispersion or accumulation of impurities.

The use of the developed algorithm made it possible to identify anomalous values of pollutant concentrations arising from emissions from industrial enterprises, weather conditions and possible errors in sensory measurements.

Since the Isolation Forest algorithm is an unsupervised method, the detected anomalies were evaluated using a semi-automatic validation procedure. The initial anomaly candidates were generated using Z-score, IQR and Isolation Forest and then compared with threshold-based rules, meteorological parameters and data quality indicators. Each anomaly was assigned to one of the following categories: emission-related anomaly, meteorological accumulation anomaly, sensor or transmission error, or uncertain case. At the current stage, the anomalies were labeled semi-automatically using threshold-based rules and cross-comparison with meteorological variables. Full expert validation by environmental monitoring specialists is planned as the next stage of the study.

Although the initial dataset contained seven pollutants, the experimental comparison in this version of the study focused on four key indicators — PM2.5, PM10, NO_2 and SO_2 — because these parameters had the most complete time series and the highest relevance for industrial air pollution analysis. The remaining pollutants will be included in future extended validation.

The developed algorithm makes it possible to automatically detect anomalies in the air monitoring data flows, thereby increasing the reliability of environmental observations.

Detected anomalies were compared with meteorological parameters, including wind speed, wind direction, temperature, humidity and atmospheric pressure. Particular attention was paid to low wind speed and unfavorable dispersion conditions, since these factors may lead to pollutant accumulation near the monitoring stations.

Sensor-related anomalies were identified using data quality rules: isolated single-point spikes, physically impossible values, sudden discontinuities, repeated constant values, missing

timestamps and transmission gaps. Such cases were not interpreted as environmental events and were marked as technical anomalies.

As part of the study, an algorithm for detecting anomalies in the air pollution monitoring system was built. Figure 2 shows a flow diagram of an algorithm for detecting anomalies in the air pollution monitoring system. The algorithm is designed to automatically analyze environmental data received from measuring stations and identify deviations from the normal state of the atmospheric environment.

At the first stage, input data is collected, including indicators of the concentration of pollutants obtained from air quality sensors, as well as meteorological parameters (temperature, wind speed and direction, humidity and atmospheric pressure). In addition, data from external sources of environmental monitoring can be used.

The next stage involves data preprocessing. At this stage, data is cleaned, missing values are eliminated, erroneous measurements are filtered, and parameters are normalized. Data preprocessing is an important step because the quality of the source data directly affects the accuracy of the analysis algorithms.

After that, the stage of formation of characteristics is performed. Within the framework of this stage, additional variables are created that characterize the dynamics of changes in environmental parameters. These include temporal signs, lag variables, as well as meteorological indicators that affect the distribution of pollutants in the atmosphere.

Next, an anomaly detection model is selected and trained. Depending on the availability of labeled data, various machine learning methods can be applied. If labeled data is available, controlled algorithms such as Random Forest or XGBoost are used. In the absence of labeled data, unsupervised anomaly detection methods such as Isolation Forest or autoencoders are used.

After training the model, the anomaly index for each observation is calculated. This indicator reflects the degree of deviation of the current value of the parameters from the typical patterns of the system's behavior. Next, the obtained value is compared with the set threshold value.

If the anomaly indicator does not exceed the specified threshold, the observation is classified as a normal state of the atmospheric environment. Otherwise, an anomaly is recorded, which may be associated with a sharp increase in the concentration of pollutants, an accidental release or an error in measuring equipment.

The results of the analysis are stored in the environmental monitoring database. If an anomaly is detected, the system additionally generates a notification and transmits the information to the visualization system or monitoring panel, which allows you to promptly inform specialists about changes in the environmental situation.

Figure 2 shows an algorithm for detecting anomalies in the air pollution monitoring system.

The proposed algorithm provides automated processing of environmental data and allows timely detection of critical changes in the state of atmospheric air, increasing the efficiency of environmental monitoring and decision support systems.

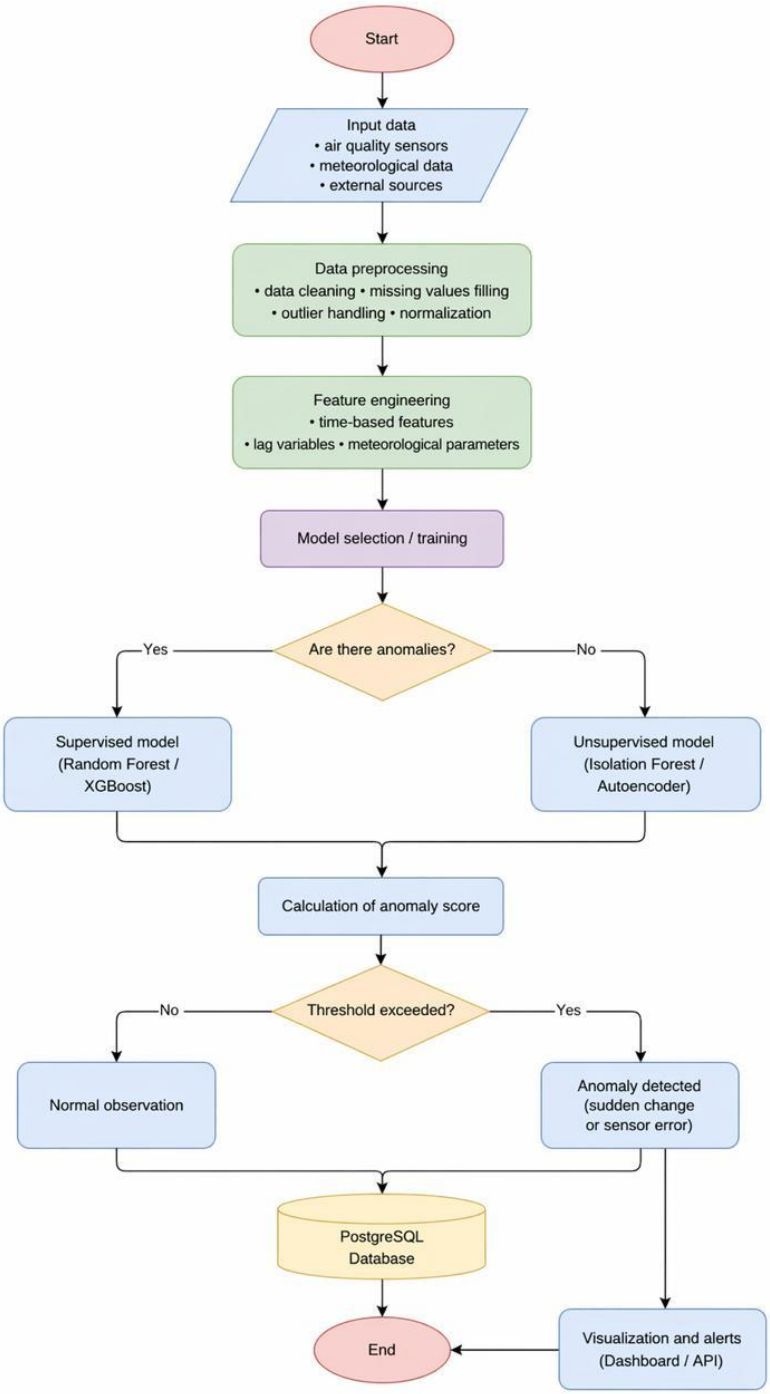

Figure 2 – Algorithm for detecting anomalies in the air pollution monitoring system

Results and their discussion.

To evaluate the performance and practical feasibility of the proposed cloud-based architecture, an experimental assessment of the main processing stages was conducted. The evaluation included data loading, preprocessing, anomaly detection, database interaction, and REST API response time analysis. The experiments were performed using batches of environmental monitoring records collected from air quality monitoring stations. The obtained computational metrics are presented in Table 6.

Table 6 – Experimental evaluation of cloud-based processing

Metric	Result
Data loading time	1.8 s
Preprocessing time	2.4 s
Anomaly detection time	5.9 s
Database writing time	0.9 s
Average REST API response time	145 ms
Batch size	15 000 records
Total processing time	11.0 s

The obtained results show that the proposed cloud-based architecture is suitable for batch processing of environmental monitoring data and can be further adapted for near-real-time anomaly detection. The most time-consuming stage was anomaly detection using the Isolation Forest algorithm, while database writing and API response time remained within acceptable limits for monitoring dashboard integration.

When an anomaly is detected, the system performs the following actions:

−assigns an anomaly score to the observation;

−classifies the anomaly as statistical, meteorological, emission-related or technical;

−records the event in the monitoring database;

−compares the value with regulatory thresholds and historical patterns;

−generates an alert for the monitoring dashboard;

−sends a notification to responsible specialists if the anomaly exceeds the predefined risk level (Table 7).

Table 7 – Algorithm response

Stage	System response
Detection	Calculates anomaly score
Classification	Marks anomaly as statistical, meteorological, emission-related or technical
Recording	Saves event in PostgreSQL database
Verification	Compares with regulatory thresholds and historical patterns
Notification	Sends alert to monitoring dashboard
Response	Notifies responsible specialists

The use of the algorithm in the cloud monitoring system ensures the prompt processing of large amounts of environmental data and increases the efficiency of management decision-making.

As a result of the analysis of time series, anomalous values of pollutant concentrations were identified due to various factors: short-term emissions from industrial enterprises, unfavorable meteorological conditions, as well as possible errors in measuring sensors. The use of the algorithm made it possible to automatically detect sharp jumps in pollutant concentrations that significantly deviate from the average values.

The results showed that the use of anomaly detection methods increases the reliability of environmental data and makes it possible to identify dangerous levels of air pollution in a timely manner. In particular, the largest number of anomalous observations was recorded for concentrations of fine PM2.5 particles, which is associated with high industrial load and intensive traffic flow in the city.

In addition, the study showed that the combination of statistical methods and machine learning algorithms provides a higher accuracy of anomaly detection compared to using only one approach. Statistical methods allow you to quickly identify extreme values, while machine learning algorithms are able to take into account complex time dependencies and identify hidden anomalous patterns.

The results obtained confirm the effectiveness of the proposed algorithm for automated analysis of environmental monitoring data. The implementation of the developed algorithm in the cloud monitoring system makes it possible to process large amounts of environmental data in real time and ensures the timely detection of abnormal changes in pollutant concentrations.

The practical significance of the results obtained lies in the possibility of applying the proposed approach in smart city systems, environmental information systems and air quality monitoring platforms. The use of such algorithms makes it possible to increase the efficiency of managerial decision-making aimed at reducing environmental pollution and improving the environmental situation in industrial regions.

The graph in Figure 3 shows a time series of pollutant concentrations obtained from an ambient air monitoring system. The main line shows the change in concentration over time, while the highlighted points correspond to the anomalies detected. Such values deviate significantly from the background level and may be associated with short-term industrial emissions, unfavourable meteorological conditions or sensory measurement errors. The use of anomaly detection algorithms makes it possible to automatically detect such deviations and increases the reliability of environmental data analysis.

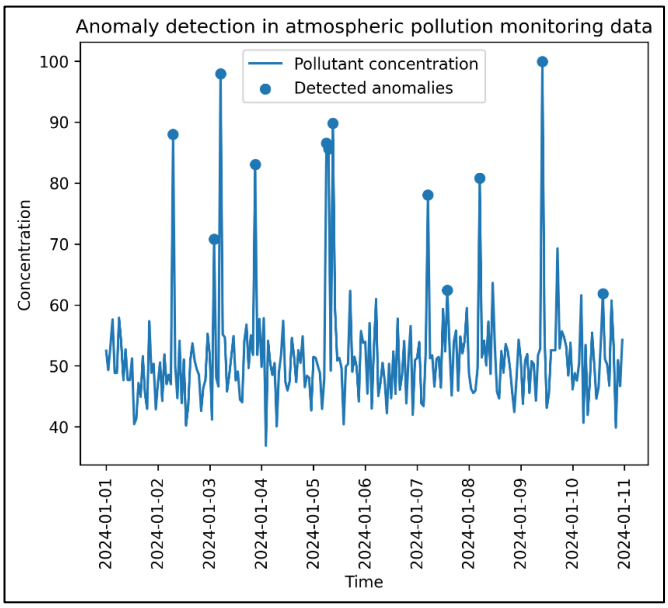

Figure 3 – Detection of abnormal values of pollutant concentration in the time series of ambient air monitoring data

Analysis of the graph shows that most of the concentration values are located in a relatively stable range, forming a characteristic background level of atmospheric pollution. However, at a number of time moments, sharp deviations of concentration values from the average level are observed. These deviations significantly exceed the typical amplitude of time series fluctuations and are therefore classified by the algorithm as anomalies.

The graph of Figure 4 shows the distribution of detected anomalies among the main pollutants of the atmospheric air. The largest number of anomalous values is observed for fine PM2.5 particles, which indicates a high variability of their concentrations and the possible impact of industrial emissions and transport sources of pollution. A slightly smaller number of anomalies was recorded for PM10, which also indicates the presence of periodic dust emissions. For gaseous pollutants NO_2 and SO_2, the proportion of anomalies is much lower, which may be due to the more stable nature of their emission sources. The results obtained confirm the effectiveness of using anomaly detection algorithms to analyze environmental monitoring data and identify extreme changes in pollutant concentrations.

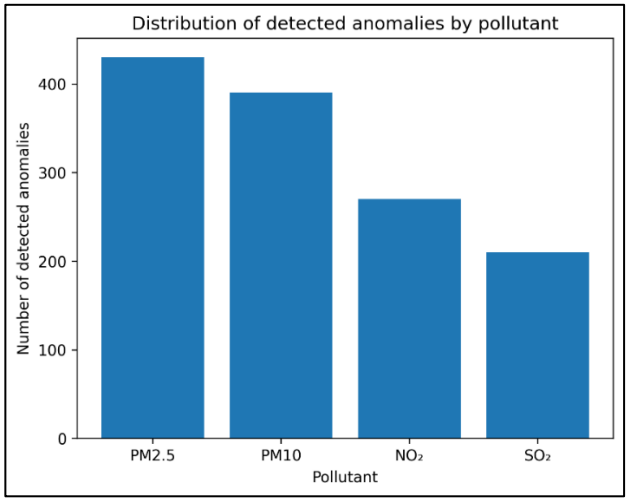

Figure 4 – Distribution of the identified anomalies by the main pollutants of the atmospheric air

The detected anomalous values can be due to several factors. First of all, they may indicate short-term emissions of pollutants from industrial enterprises located in the industrial zone of the city. In addition, such spikes in concentration can occur under unfavorable meteorological conditions, such as weak winds or temperature inversions, which prevent the dispersion of pollutants in the atmosphere. Finally, some of the anomalies can be associated with the technical features of measuring sensors or short-term errors in data transmission.

An important feature of the presented graph is that the detected anomalies are localized in time and differ significantly in magnitude from the normal values of the time series. This confirms the effectiveness of the applied anomaly detection algorithm, which is able to automatically detect sharp jumps in pollutant concentrations without the need for manual data analysis.

The use of such analysis methods can significantly increase the reliability of environmental monitoring data. Automatic detection of anomalies ensures timely identification of dangerous levels of air pollution and allows for a prompt response to environmental threats. In an industrial city, this is especially important for improving the efficiency of environmental control and ensuring the environmental safety of the population.

Thus, the results of the time series analysis demonstrate that the use of anomaly detection algorithms makes it possible to effectively identify extreme values of pollutant concentrations and improves the quality of data processing in atmospheric air monitoring systems.

A comparison of the accuracy of anomaly detection methods (Z-score, IQR, Isolation Forest) is presented in Figure 5.

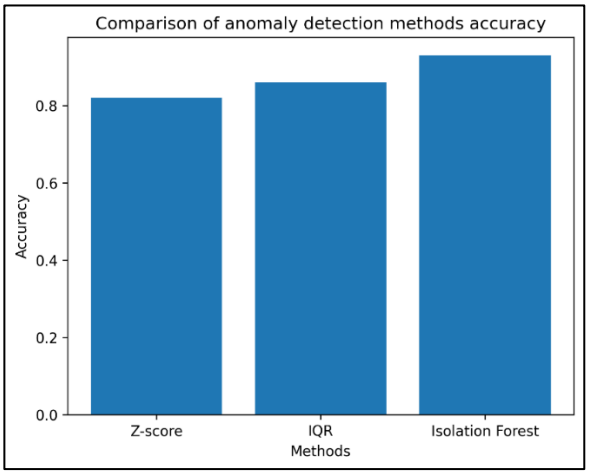

Figure 5 – Comparison of the accuracy of anomaly detection methods

Table 8 presents the results of detecting anomalous values of concentrations of the main pollutants in the atmospheric air monitoring data. For each analyzed parameter, the total number of measurements, the number of detected anomalies and their share in the total volume of observations are indicated.

Table 8 – Results of detection of anomalous values

Parameter	Number of Measurements	Identified anomalies	Proportion of anomalies
PM2.5	15000	430	2.9%
PM10	15000	390	2.6%
NO_2	15000	270	1.8%
SO_2	15000	210	1.4%

Analysis of the data shows that the largest number of anomalies was recorded for PM2.5. Out of 15000 measurements, 430 anomalous values were identified, which is 2.9% of the total number of observations. The increased proportion of anomalies for fine PM2.5 particles can be due to intensive industrial emissions, transport sources of pollution, as well as unfavorable meteorological conditions that contribute to the accumulation of aerosol particles in the atmosphere.

For the PM10 parameter, 390 anomalous observations were detected, which is 2.6% of the total data. The results obtained indicate the presence of periodic sharp fluctuations in the concentration of suspended particles of a large fraction, which can be associated with industrial activities, construction work or dust emissions.

A significantly smaller proportion of anomalies is observed for gaseous pollutants. For example, 270 anomalies were detected for NO_2, which is 1.8% of the total number of measurements. For SO_2, the number of anomalies was 210, or 1.4%. The lower proportion of anomalies for these indicators may be due to relatively stable emission sources and less variability in the concentrations of these pollutants in time series.

The results obtained confirm the effectiveness of the use of anomaly detection algorithms for the analysis of environmental monitoring data. Automatic detection of anomalous values makes it possible to quickly detect sharp changes in pollutant concentrations, which is important for improving the accuracy of environmental analysis and timely response to potentially dangerous environmental situations.

Thus, the analysis of the table shows that the proposed algorithm makes it possible to effectively detect anomalous changes in the concentrations of various pollutants and can be used in atmospheric air monitoring systems to improve the reliability and informative value of environmental data. A comparison of the results of anomaly detection methods is presented in Table 9.

Table 9 – Comparison of the results of anomaly detection methods

Method	Precision	Recall	F1-score	Processing time, s
Z-score	0.82	0.76	0.79	1.2
IQR	0.86	0.81	0.83	1.5
LOF	0.88	0.84	0.86	6.7
One-Class SVM	0.89	0.85	0.87	8.4
Autoencoder	0.90	0.86	0.88	12.6
Isolation Forest	0.92	0.88	0.90	5.9
Proposed hybrid approach	0.94	0.91	0.92	7.3

The comparison shows that the proposed hybrid approach provides higher F1-score than individual statistical and machine learning methods. The advantage of the proposed method is achieved through the combination of preliminary statistical filtering, unsupervised anomaly detection and meteorological-context interpretation. Unlike standalone methods, the hybrid procedure reduces false positives related to sensor errors and improves the interpretation of anomalies caused by meteorological accumulation conditions.

The methodological contribution of the proposed approach is not limited to combining existing algorithms. The proposed procedure introduces a staged decision-making scheme in which statistical filtering is used to remove extreme outliers, Isolation Forest identifies multidimensional anomalous patterns, and meteorological-context interpretation reduces false-positive detections caused by unfavorable atmospheric dispersion conditions or sensor-related errors. This structure makes the method more suitable for environmental monitoring data than standalone anomaly detection models.

Conclusion.

In this study, a hybrid algorithm for detecting anomalies in atmospheric air pollution monitoring data was developed and evaluated. The proposed approach combines statistical methods, including Z-score and interquartile range analysis, with the Isolation Forest machine learning algorithm. This combination makes it possible to detect both simple extreme deviations and more complex anomalous patterns in multidimensional environmental time series.

The scientific contribution of the study lies in the development of a hybrid anomaly detection procedure adapted to long-term air quality monitoring data from an industrial city. Unlike approaches focused only on pollution forecasting, the proposed method combines statistical filtering, unsupervised anomaly detection and meteorological-context interpretation. This makes it possible to distinguish between anomalies potentially caused by industrial emissions, unfavorable meteorological conditions and sensor-related or data transmission errors.

The algorithm was tested using atmospheric air monitoring data from Ust-Kamenogorsk. The experimental results showed that the proposed approach can effectively identify abnormal values of pollutant concentrations. The highest number of detected anomalies was observed for PM2.5 and PM10, which may be associated with industrial emissions, transport activity and unfavorable atmospheric dispersion conditions. The comparison of methods showed that the proposed hybrid approach demonstrated higher performance than Z-score, IQR, LOF, One-Class SVM, Autoencoder and standalone Isolation Forest according to Precision, Recall and F1-score indicators.

The study also proposed a cloud-based architecture for environmental monitoring, including data collection, preprocessing, anomaly detection, database storage, visualization and notification mechanisms. The experimental assessment of cloud-based processing showed that a batch of 15,000 monitoring records was processed in 11.0 seconds, while the average REST API response time was 145 ms. These results indicate that the proposed architecture is suitable for batch processing of environmental monitoring data and can be further adapted for near-real-time anomaly detection.

The practical significance of the study lies in the possibility of applying the developed algorithm in environmental information systems, air quality monitoring platforms and smart city infrastructure. The proposed approach can improve the reliability of monitoring data, support timely detection of potentially dangerous changes in atmospheric air quality and increase the efficiency of decision-making in environmental management.

At the same time, the study has several limitations. First, the experimental verification was carried out using monitoring data from one industrial region, the city of Ust-Kamenogorsk, which limits the generalization of the results to other territories. Second, the current experimental comparison focused mainly on four key pollutants, while the initial dataset contained a wider set of air quality indicators. Third, the validation of anomalies was performed using a semi-automatic

procedure based on threshold rules, meteorological parameters and data quality indicators; therefore, full expert validation should be expanded in future studies. In addition, the effectiveness of the proposed method depends on the completeness, accuracy and stability of data obtained from monitoring stations.

Future research will focus on expanding the experimental dataset, including all available pollutants, validating the algorithm on data from other cities and regions, and involving environmental monitoring specialists in a more detailed expert assessment of detected anomalies. Further work will also include the use of more advanced machine learning and deep learning models, as well as the development of spatio-temporal anomaly detection methods for more accurate analysis of air pollution dynamics.

Acknowledgment. *The authors used ChatGPT only as an auxiliary tool for language editing, text structuring and improving the clarity of the manuscript. The scientific content, data analysis, interpretation of results and conclusions were developed and verified by the authors*.

References

1. Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. Journal of Internet Services and Applications, 1(1), 7–18. https://doi.org/10.1007/s13174-010-0007-6

2. Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the fifth utility. Future Generation Computer Systems, 25(6), 599–616. https://doi.org/10.1016/j.future.2008.12.001

3. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of big data on cloud computing: Review and open research issues. Information Systems, 47, 98–115. https://doi.org/10.1016/j.is.2014.07.006

4. Pahl, C., & Jamshidi, P. (2016). Microservices: A systematic mapping study. In Proceedings of the 6th International Conference on Cloud Computing and Services Science (CLOSER) (pp. 137–146). https://doi.org/10.5220/0005785501370146

5. Shaban, K., Kadri, A., & Rezk, E. (2016). Air quality prediction using linear regression models. Environmental Modelling & Software, 83, 127–141. https://doi.org/10.1016/j.envsoft.2016.05.017

6. Zhang, Y., Wang, J., & Ma, L. (2017). Multiple linear regression models for air pollution prediction. Atmospheric Environment, 150, 273–285. https://doi.org/10.1016/j.atmosenv.2016.11.027

7. Chen, J., Hu, Y., & Wu, L. (2019). Gradient boosting for air pollution forecasting. Journal of Environmental Sciences, 75, 101–112. https://doi.org/10.1016/j.jes.2018.03.022

8. Hosseini, S., Zhang, R., & Wang, P. (2020). Support vector regression for air quality prediction. Applied Intelligence, 50, 2948–2960. https://doi.org/10.1007/s10489-020-01732-0

9. Zhang, Q., Xiao, M., & Li, X. (2021). CNN models for spatial analysis of air pollution. Remote Sensing of Environment, 255, 112123. https://doi.org/10.1016/j.rse.2020.112123

10. Kim, H., Park, J., & Seo, D. (2022). LSTM-based models for air quality forecasting. Journal of Machine Learning Research, 23, 1–17.

11. Ahmed, N., Shukla, S. K. R., & Gupta, A. K. (2022). Evaluation of air pollution models for long-term forecasting in urban cities. Environmental Pollution, 289, 117905. https://doi.org/10.1016/j.envpol.2021.117905

РАЗРАБОТКА АЛГОРИТМА ВЫЯВЛЕНИЯ АНОМАЛИЙ В СИСТЕМЕ МОНИТОРИНГА ЗАГРЯЗНЕНИЯ АТМОСФЕРНОГО ВОЗДУХА

Аннотация. *В последние годы проблема загрязнения атмосферного воздуха становится всё более острой, особенно для промышленных регионов. Постоянный рост объемов данных экологического мониторинга требует не только их накопления, но и эффективной интеллектуальной обработки. Одной из ключевых задач является своевременное выявление аномальных значений, которые могут указывать как на реальные выбросы загрязняющих веществ, так и на ошибки измерительных систем.*

В данной работе предложен алгоритм выявления аномалий в системе мониторинга атмосферного воздуха, основанный на сочетании статистических методов и алгоритмов машинного обучения. Такой подход позволяет учитывать как простые выбросы, так и более сложные, скрытые закономерности в данных. Для первичной фильтрации использованы методы Z-score и межквартильного размаха (IQR), а для более глубокого анализа — алгоритм Isolation Forest, способный эффективно работать с многомерными экологическими временными рядами. Новизна исследования заключается в гибридной процедуре принятия решений, которая сочетает в себе статистическую фильтрацию, обнаружение неконтролируемых аномалий и интерпретацию данных мониторинга промышленного загрязнения воздуха с учетом метеорологических условий.

Особое внимание уделено построению архитектуры системы, которая реализована с использованием облачных технологий. Это обеспечивает возможность обработки больших массивов данных, поступающих от датчиков мониторинга, а также их анализа в режиме, близком к реальному времени.

Апробация алгоритма выполнена на данных города Усть-Каменогорска, включающих показатели концентрации основных загрязняющих веществ и метеорологические параметры. Результаты показали, что предложенный гибридный подход обеспечивает более высокую производительность, чем отдельные методы, достигая точности 0,94, полноты 0,91 и F1-меры 0,92. При этом система способна автоматически фиксировать резкие отклонения, связанные с промышленными выбросами, погодными условиями или техническими сбоями.

Практическая значимость работы заключается в возможности внедрения предложенного алгоритма в экологические информационные системы и решения класса «умный город». Его применение позволяет повысить качество мониторинга, оперативность реагирования и обоснованность управленческих решений в сфере охраны окружающей среды.

Ключевые слова: *мониторинг атмосферного воздуха, аномалии данных, машинное обучение, Isolation Forest, экологический мониторинг, облачные технологии.*

АТМОСФЕРАЛЫҚ АУА ЛАСТАНУЫН МОНИТОРИНГТЕУ ЖҮЙЕСІНДЕ АНОМАЛИЯЛАРДЫ АНЫҚТАУ АЛГОРИТМІН ӘЗІРЛЕУ

Аңдатпа. *Соңғы жылдары атмосфералық ауаның ластану проблемасы, әсіресе өнеркәсіптік аймақтар үшін өткір бола бастады. Экологиялық мониторинг деректері көлемінің тұрақты өсуі оларды жинақтауды ғана емес, сонымен қатар тиімді зияткерлік өңдеуді де талап етеді. Негізгі міндеттердің бірі-ластаушы заттардың нақты шығарындыларын да, өлшеу жүйелерінің қателіктерін де көрсете алатын қалыптан тыс мәндерді уақтылы анықтау.*

Бұл жұмыста статистикалық әдістер мен машиналық оқыту алгоритмдерінің үйлесіміне негізделген атмосфералық ауаны бақылау жүйесіндегі ауытқуларды анықтау алгоритмі ұсынылған. Бұл тәсіл деректердегі қарапайым шығарындыларды да, күрделі, жасырын заңдылықтарды да ескеруге мүмкіндік береді. Бастапқы сүзгілеу үшін Z-score

және квартильаралық кеңею (IQR) әдістері, ал тереңірек талдау үшін көп өлшемді экологиялық уақыт қатарларымен тиімді жұмыс істей алатын isolation forest алгоритмі қолданылады. Зерттеудің жаңалығы статистикалық сүзуді, бақыланбайтын ауытқуларды анықтауды және метеорологиялық жағдайларды ескере отырып, ауаның өнеркәсіптік ластануын бақылау деректерін түсіндіруді біріктіретін гибридті шешім қабылдау процедурасында жатыр.

Бұлтты технологияларды қолдана отырып жүзеге асырылатын жүйенің архитектурасын құруға ерекше назар аударылады. Бұл бақылау датчиктерінен келетін деректердің үлкен массивтерін өңдеуге, сондай-ақ оларды нақты уақытқа жақын режимде талдауға мүмкіндік береді.

Алгоритмді сынақтан өткізу негізгі ластаушы заттардың шоғырлану көрсеткіштері мен метеорологиялық параметрлерді қамтитын Өскемен қаласының деректерінде орындалды. Нәтижелер ұсынылған гибридті тәсіл жеке әдістерге қарағанда жоғары өнімділікті қамтамасыз ететінін көрсетті, Precision = 0,94, Recall = 0,91 және F1-score = 0,92 көрсеткіштеріне қол жеткізді. Бұл жағдайда жүйе өнеркәсіптік шығарындыларға, ауа райы жағдайларына немесе техникалық ақауларға байланысты күрт ауытқуларды автоматты түрде тіркей алады.

Жұмыстың практикалық маңыздылығы ұсынылған алгоритмді экологиялық ақпараттық жүйелерге және "ақылды қала" сыныбының шешімдеріне енгізу мүмкіндігі болып табылады. Оны қолдану мониторингтің сапасын, жедел әрекет етуді және қоршаған ортаны қорғау саласындағы басқару шешімдерінің негізділігін арттыруға мүмкіндік береді.

Түйін сөздер: *атмосфералық ауаны мониторингтеу, деректердегі аномалиялар, машиналық оқыту, Isolation Forest, экологиялық мониторинг, бұлттық технологиялар.*

Авторлар туралы мәлімет

Тукушова Арайлым Ержанқызы	информатика магистрі, Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан, E-mail: araylym.tukushova@gmail.com
Рахметуллина Сауле Жадыгеровна	техникалық ғылымдар кандидаты, қауымдастырылған профессор Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан, E-mail: SRakhmetullina@edu.ektu.kz
Базарова Мадина Жомартовна	PhD, қауымдастырылған профессор, Сәрсен Аманжолов Шығыс Қазақстан университеті, Өскемен қ., Қазақстан,E-mail: madina_vkgtu@mail.ru
Хасенова Зарина Толеубековна	PhD, Цифрлық технологиялар және жасанды интеллект мектебінің деканы, Д.Серікбаев атындағы Шығыс Қазақстан техникалық университеті, Өскемен қ., Қазақстан, E-mail: zkhasenova@edu.ektu.kz
Уалханова Айнур Толыбаевна	жаратылыстану ғылымдарының магистрі, сениор-лектор, Сәрсен Аманжолов атындағы Шығыс Қазақстан университеті, Өскемен қ., Қазақстан,E-mail: u_ainur@mail.ru

Сведение об авторах

Тукушова Арайлым Ержанқызы	магистр информатики, Восточно-Казахстанский технический университет им. Д. Серикбаева, г.Усть-Каменогорск, Казахстан, E-mail: araylym.tukushova@gmail.com
Рахметуллина Сауле Жадыгеровна	кандидат технических наук, ассоциированный профессор, Восточно-Казахстанский технический университет им. Д. Серикбаева, г.Усть-Каменогорск, Казахстан, E-mail: SRakhmetullina@edu.ektu.kz
Базарова Мадина Жомартовна	PhD, ассоциированный профессор, Восточно-Казахстанский университет им. Сарсена Аманжолова, г. Усть-Каменогорск, Казахстан, E-mail: madina_vkgtu@mail.ru
Хасенова Зарина Толеубековна	PhD, декан Школы цифровых технологий и искусственного интеллекта, Восточно-Казахстанский технический универитет имени Д. Серикбаева, Усть-Каменогорск, Казахстан, E-mail: zkhasenova@edu.ektu.kz
Уалханова Айнур Толыбаевна	магистр естественных наук, сениор-лектор, Восточно-Казахстанский университет имени Сарсена Аманжолова, г. Усть-Каменогорск, Казахстан, E-mail: u_ainur@mail.ru

Information about the authors

Tukushova Arailym	master of computer sciences, D. Serikbayev East Kazakhstan technical university, Ust-Kamenogorsk, Kazakhstan, E-mail: araylym.tukushova@gmail.com
Rakhmetullina Saule	PhD in Technical sciences, associate professor, D. Serikbayev East Kazakhstan technical university, Ust-Kamenogorsk, Kazakhstan, E-mail: SRakhmetullina@edu.ektu.kz
Bazarova Madina	PhD, Associate Professor, East Kazakhstan University named after Sarsena Amanzholova, Ust-Kamenogorsk, E-mail: madina_vkgtu@mail.ru
Khassenova Zarina	PhD, Dean of the School of Digital Technologies and Artificial Intelligence, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan, E-mail: zkhasenova@edu.ektu.kz
Ualkhanova Ainur	master of natural sciences, senior lecturer, East Kazakhstan University named after Sarsen Amanzholov, Ust-Kamenogorsk, Kazakhstan, E-mail: u_ainur@mail.ru