# Data analysis using OLAP and data mining technologies in the study of atmospheric air quality

Nikolay Kiktev[1,†], Bella Golub[1,†], Maryna Lendiel[1,†], Taras Lendiel[1,†], Vitalii Larin[2,*,†] and Danylo Hradoboiev[3, †]

[1] National University of Life and Environmental Sciences of Ukraine, Heroiv Oborony str. 15, 03041, Kyiv, Ukraine
[2] National Aviation University, Liubomyra Huzara Ave., 1, Kyiv, 03058, Ukraine
[3] Beetroot LLC, Hollandargatan 20, 11160, Stockholm, Sweden

## Abstract

The article presents the results of statistical analysis of atmospheric air quality. Open data on the air quality index (AQI) and its components: levels of ozone ($O_3$), nitrogen dioxide ($NO_2$) and fine particulate matter (PM2.5) were used for the study. Based on the analysis, a multivariate regression model was built, which made it possible to assess the significance of each of the pollutants. Using the elasticity index made it possible to determine the relative impact of each factor on the air quality index. According to the results of the study, PM2.5 and ozone levels had the greatest influence on the AQI index, while the influence of $NO_2$ was insignificant. When developing a decision support system, the use of statistical methods together with OLAP and Data Mining technology can be useful, as statistical methods can confirm or refute the hypotheses that arise in the process of using OLAP and Data Mining, which helps to analyze the data more deeply and implement the decision-making process more effectively solutions.

## Keywords

neural network, agricultural land, image recognition, blast craters, training, dataset

## 1. Introduction

Atmospheric air is one of the key factors that affects all life on the Earth. Weather and wind have influence into the entire ecosystem, including biotic components. The current development of humanity is clearly reflected in the world of nature. Through various spheres of human activity, the global economy is hampered by waste products from industry, exhaust gases from transport, solid particles, freons from waste, which causes the greenhouse effect and, consequently, a change in climate mat.

Therefore, monitoring the intensity of the atmospheric air is an important task in order to ensure the health of the population and protect the environment. The study of peer obstacles helps to identify the obstacles and develop effective strategies for their change. From these data and formed hypotheses that would help to reduce the intensity of the pollution of the atmospheric air. The main purpose of the study is to improving the state of atmospheric air by analyzing and processing data on the state of atmospheric air.

## 2. Materials and methods

Remote sensing techniques of the Earth and atmosphere provides reach feature to study atmospheric state. artificial intelligence and machine learning helps to analyse measured data.

*ADP'24: International Workshop on Algorithms of Data Processing, November 5, 2024, Kyiv, Ukraine*
* Corresponding author.
† These authors contributed equally.
nkiktev@ukr.net (N. Kiktev); bella.golub55@gmail.com (B. Golub); marynalendel@gmail.com (M. Lendiel); taraslendel@gmail.com (T. Lendiel); vjlarin@gmail.com (V. Larin); gradoboev2607@gmail.com (D. Hradoboiev)
0000-0001-7682-280X (N. Kiktev); 0000-0002-1256-6138 (B. Golub); 0009-0008-0042-7705 (M. Lendiel); 0000-0002-6356-1230 (T. Lendiel); 0000-0002-5042-2426 (V. Larin); 0009-0007-9636-1095 (D. Hradoboiev)

Decision support systems play an important role in the process of air quality regulation. They provide an opportunity to analyze a large amount of data, which allows us to make informed decisions to reduce the level of air pollution. One of the key components of decision support systems are On-line Analytical Processing (OLAP) and Data Mining technologies.

OLAP is a system of analytical data processing. It is designed to prepare reports, build forecast scenarios and perform statistical calculations based on large information arrays with a complex structure.

The overall goal of OLAP is to make it easier to manage large amounts of data without requiring the data to be stored in a specific way or in a specific location. This means you can specify it in your data repository and index the data where it lives, abstracting away the complexity of managing large, distributed data sets.

Data Mining is a process of finding previously unknown, non-trivial, practically useful and easily interpretable knowledge in arrays of raw data, which is necessary for decision-making in various areas of human activity. The essence and purpose of Data Mining technology: it is a technology that is designed to search for non-obvious, objective and practically useful patterns (knowledge) in large volumes of data.

Both technologies complement each other, and while DM finds patterns based on known knowledge, OLAP analyzes data in real-time and can confirm or refute hypotheses generated by the intelligent analysis process. The use of statistical methods, such as linear regression and correlation analysis, in combination with OLAP and Data Mining allows for a detailed analysis of the impact of individual pollutants on the overall quality of atmospheric air and a more accurate assessment of the significance of the identified hypotheses.

To solve some problems, not only linear regression models can be used. The authors [1, 2] in their study show the best minimization of the standard deviation and the best predictive properties when using a segmented regression model to solve the problem of increasing the assessment of the degree of degradation of aviation equipment. The authors of the study [3] also confirm the effectiveness of using segmented linear regression for estimating time series in financial analysis.

To identify a hypothesis, Bayesian methods for evaluating statistical data are often used. In paper [4], a maximum posterior probability method was developed to minimize errors in determining the position of an aircraft. The Bayesian approach is also successfully used in machine learning.

An open dataset on the Air Quality Index (AQI) and its components: levels of ozone ($O_3$), nitrogen dioxide ($NO_2$) and fine particulate matter (PM2.5) [5] was used for the analysis. To implement the technology of processing data on the degree of air pollution using the OLAP method in real time, it is necessary to have sensors or sensor systems that measure the required parameters. Determination of the concentration of fine particulate matter in the air mass can be realized using aviation weather radars, for which it is necessary to develop particle identification algorithms, as shown in article [6], which describes the functioning of a new algorithm for determining and classifying turbulence. The structure of aviation weather radar is represented in Figure 1.

Radar transmitter generates radio testing signal at a frequency band of 9345 ± 15 MHz. Signal through antenna switch is supplied to the antenna system. Such the signal scans a defined area of air space before an aircraft. Weather radar signals are reflected from air objects return in the opposite direction to the antenna system of Weather Radar. At some revision of weather radar software we will able to recognize domains polluted by fine particulate matter.

The need to measure pollution parameters in local areas of the atmosphere will require solving the problems of determining the coordinates of polluted areas and their possible dynamics, taking into account changes in their configuration, including altitude changes [7–9].

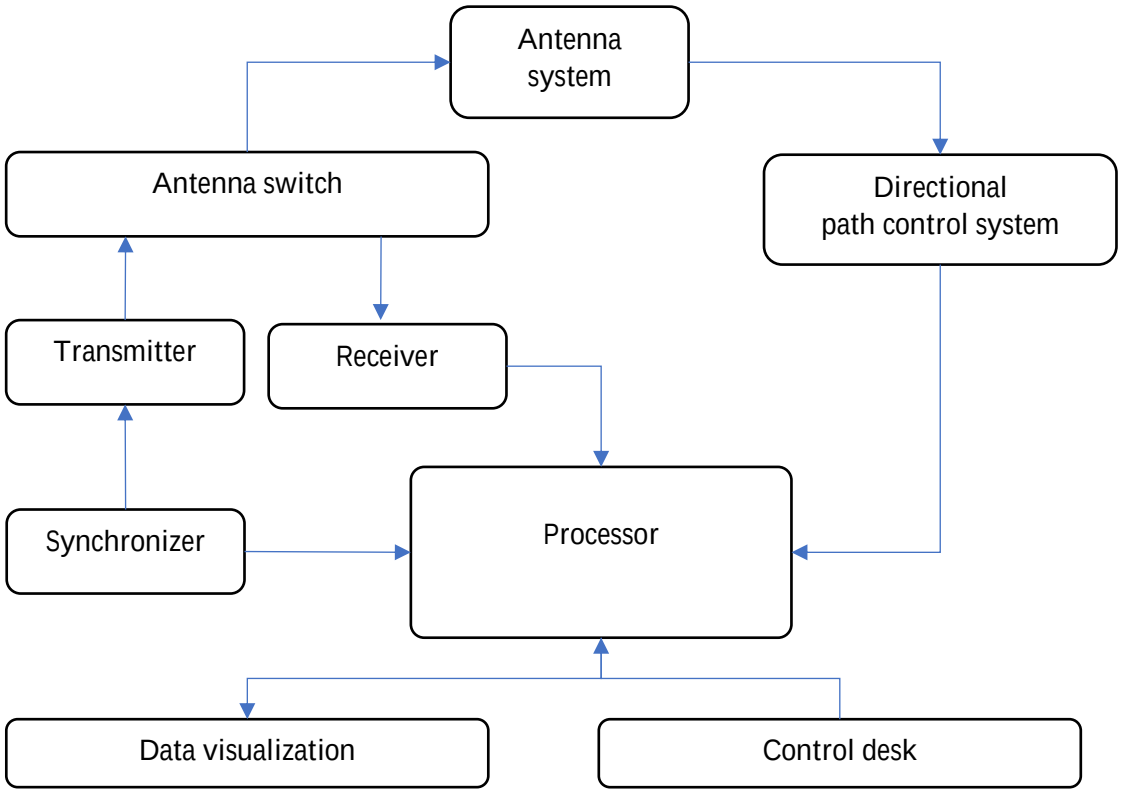


Figure 1: Common structure scheme of aviation weather radar.

Determining the concentration of ozone and nitrogen dioxide ($NO_2$) will require the use of different types of measuring instruments. The most common type of devices for measuring nitrogen dioxide are gas analyzers, the operating principle of which is based on the effect of chemiluminescence. This effect is observed in the process of converting chemical energy into quantum energy, accompanied by the emission of a certain number of photons.

In the gas analyzer, the resulting photon flow is passed through special light filters, after which the flow enters the chamber of the photoelectron multiplier, where it is converted into an electrical signal convenient for processing. Modern advances in minimizing electronic circuits significantly reduce the size and weight of modern gas analyzers. In modern devices a response times have been increased to 30 seconds. The structure of a chemiluminescence sensor is represented in Figure 2. Chemiluminescence analyzers due only to nitrogen dioxide measurement have a narrow spectrum of application.

One more method to define both nitrogen dioxide and nitrogen monoxide is Non-Dispersion InfraRed Spectroscopy (NDIR). The gas contains various atoms that absorb light with a characteristic wavelength in the infra-red region of the spectrum. For measurements, the total absorption of the molecule at the maximum frequency or wavelength is used. One beam passes through the measuring camera, the other through a comparison camera containing a gas that does not absorb infrared radiation, usually nitrogen. If the sample contains a substance being determined, some of the infra-red energy is absorbed and the fraction of the infrared energy reaching the detector will be proportional to the amount of such the substance in the sample. Detector has sensitivity to emitting of long wave, characterizing of researched substance. The structure of a NDIR-sensor is represented in Figure 3.

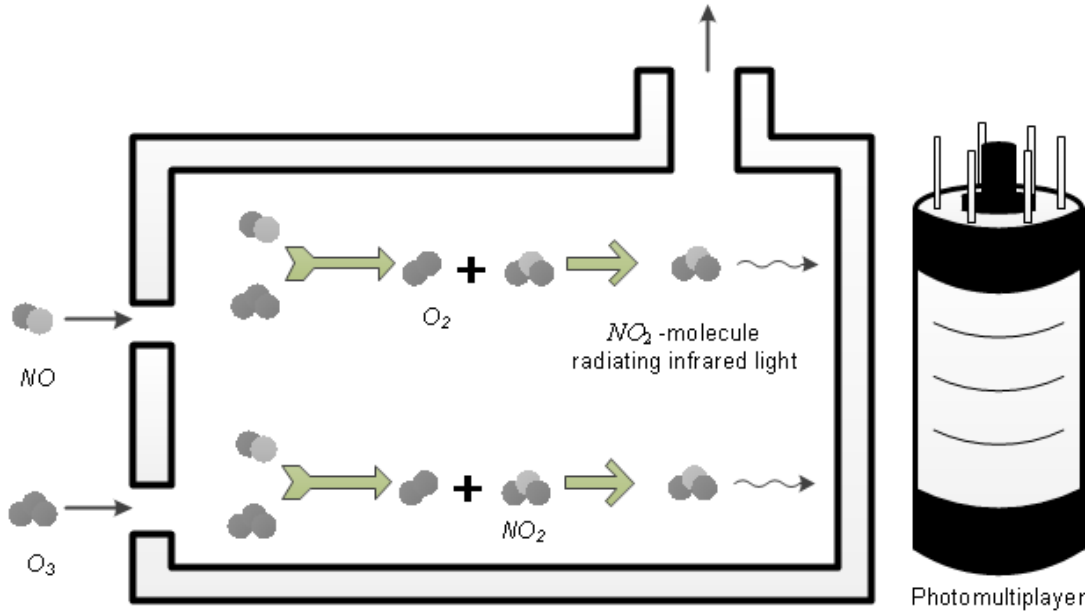


Figure 2: Chemiluminescence sensor structure.

The noted improvement makes it possible to develop a scheme for using the gas analyzer as a payload and install it on an unmanned aerial vehicle. Such an assembling can easily measure concentrations of nitrogen dioxide and ozone not only in air samples taken under normal conditions a meter or two from the earth's surface, but also at significantly higher altitudes. The maximum measurement altitude will be limited by the maximum flight altitude of the unmanned aerial vehicle. Offered measuring devices must be joined into measuring system like Wireless Sensor Network. Wireless Sensor Network advantage is availability of ready to use special devices to manage such the sensors system - Network Coordinator. Network Coordinator able to integrate of wireless sensors into the network, to maintain network performance, define the state of individual sensor and the network as a whole, detect and eliminate extraordinary situations [10, 11].

Measured and processed data should be indexed and joined with navigation coordinates and collects in Database. When using OLAP and Data Mining methods some researches offer to use a structured Databases - Data Warehouse [12].

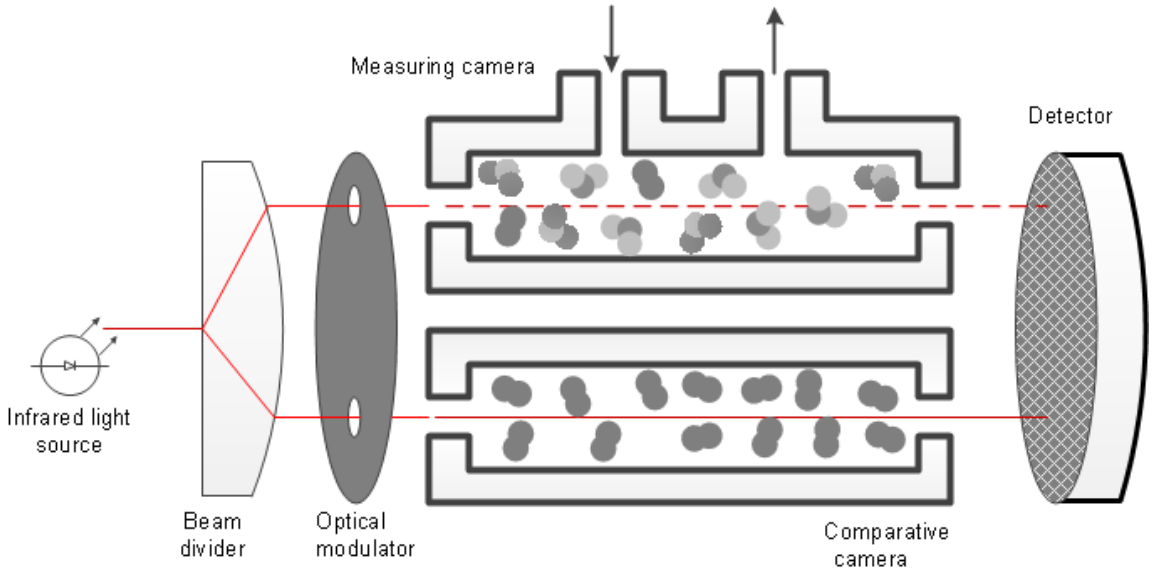


Figure 3: NDIR sensor structure.

The Data Warehouse concept is not a complete DSS architectural solution and certainly not a finished software product.

The purpose of the Data Warehouse concept is to define the requirements for data placed in the Data Warehouse, the general principles and stages of building a Data Warehouse, the main sources of data, and to provide recommendations for solving potential problems that arise during their unloading, cleaning, reconciliation, transportation, and loading.

The Data Warehouse concept defines only the most general principles for constructing an analytical system and is primarily focused on the properties and requirements of data, but not on the methods of organizing and presenting it in the target Data Base and the modes of its use. Data Warehouse is a concept for building an analytical system, but not a concept for using it.

Today, Data Warehouse construction technologies are the basis for creating full-fledged intelligent data analysis systems, focused on solving poorly structured decision-making problems, because they contain data with the following properties:

- Integrity and internal interconnection. Although the data is loaded from different sources, but they are united by uniform naming laws, methods of measuring attributes, etc. This is of great importance for corporate organizations, in which computer systems with different architectures, representing the same data in different ways, can be operated at the same time.
- Subject orientation. Local databases contain megabytes of information not needed for analysis. Such information is not entered into the repository, which limits the range of data considered when making a decision to a minimum.
- Lack of time reference. Operating systems cover a small time interval, which is achieved due to periodic data archiving. Data warehouses, on the contrary, contain historical data accumulated over a long period of time (years, decades).
- Read-only availability. Data modification is not carried out, as it may lead to a violation of the integrity of the Data warehouse. Since there is no need to minimize the immersion time, the Data warehouse structure can be optimized for processing certain requests, which is achieved by denormalizing the relational schema, prior aggregation and building the most relevant indexes.
- Integration. This means that the data satisfies the requirements of the entire project, not a single design procedure. This way Data Warehouse ensures that the same reports generated for different analysts will contain the same results.
- Immutability means that, once in the Data Warehouse, the data is stored there and does not change. Data in the Data Warehouse can only be added.

The software underlying OLAP is a Python library that allows the user to index data (add entries to the index), receive data (optimize indexed data for performance), query data (return data in a standard data format). and a wide range of other functions related to data management.

It is necessary to determine the connection between the air quality index and its components. For this, the Seaborn library in the Python programming language was used, which allows you to construct a pair graph.

Such a plot helps reveal correlations and other dependencies between numerical variables in a data set.

Result of analysis connection between different air quality indicators is shown in Figure 4. Obtained results indicates clear correlation between the air quality index and PM2.5.

Also, to determine the effect of each pollutant on the overall value of the air quality index, elasticity was calculated, which shows how much the dependent variable (AQI) changes by 1% when the independent variable (the level of a particular pollutant) changes by 1% can be calculated as follows [11]:

$$E = \beta_i \frac{x_i}{y}, \quad (1)$$

where $E$ is elasticity; $\beta_i$is the regression coefficient for the $i^{th}$ pollutant; $x_i$ is the average value of the $i^{th}$ pollutant; $y$ is the average value of the air quality index.

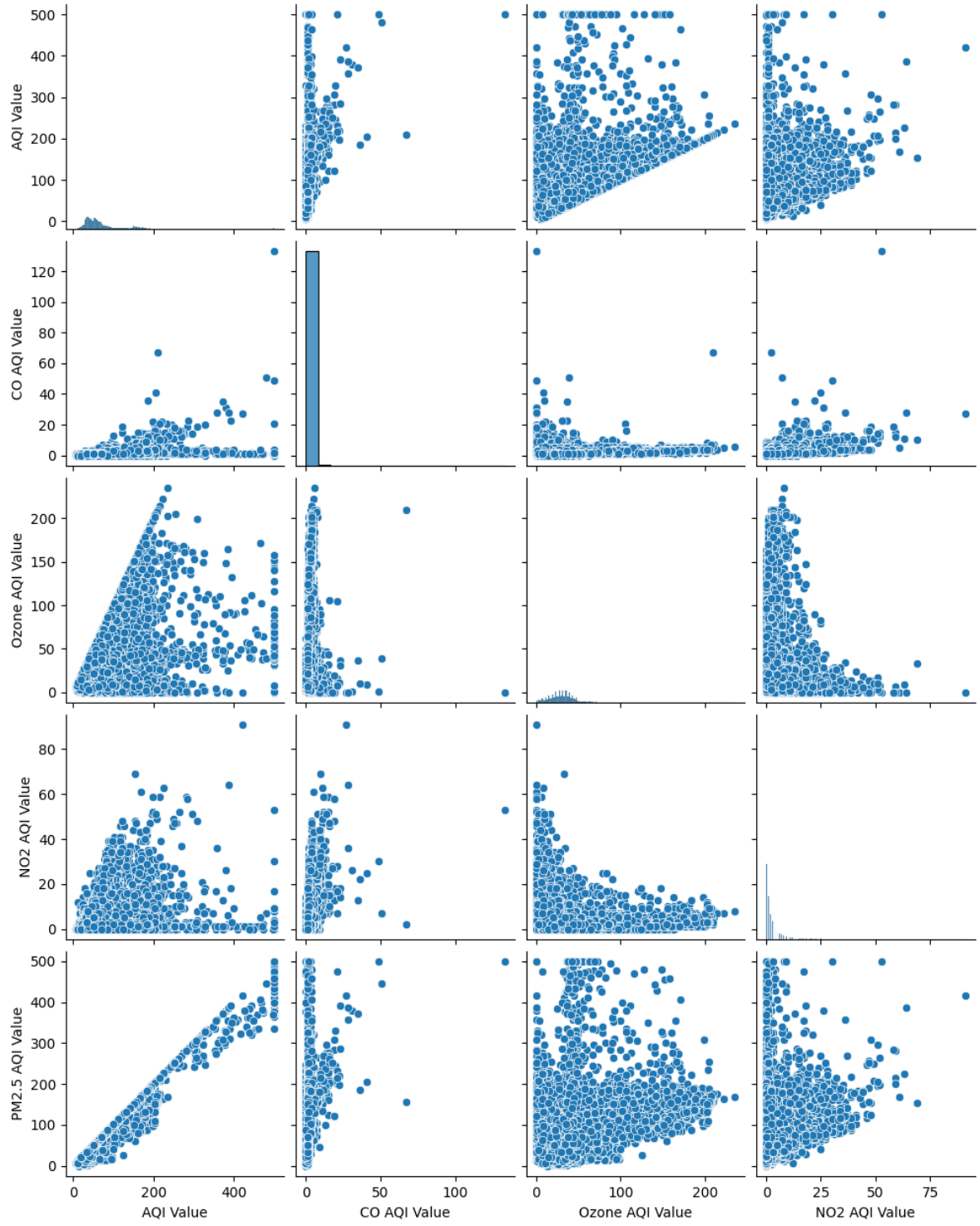


Figure 4: Visualization of connection between the air quality index and its components.

The obtained results demonstrate the following:

- ozone value (Ozone AQI Value): an elasticity of 0.075802 means that a 1% increase in Ozone AQI Value is associated with an increase in AQI Value of about 0.0758%. This indicates a relatively low but positive sensitivity of AQI to changes in the ozone level.

- value of the $NO_2$ indicator ($NO_2$ AQI Value): elasticity -0.001569 means that a 1% increase in $NO_2$ AQI Value is associated with a decrease in AQI Value of about 0.0016%. This suggests a very small negative relationship between $NO_2$ levels and overall AQI, i.e. as $NO_2$ increases, overall AQI decreases slightly. However, this effect is very minimal.
- PM2.5 indicator (PM2.5 AQI Value): an elasticity of 0.934452 means that a 1% increase in PM2.5 AQI Value is associated with an increase in AQI Value of approximately 0.9345%. This indicates a very high sensitivity of AQI to changes in PM2.5 levels, showing that PM2.5 has a significant impact on the overall air quality index.

Looking at the results obtained, among the pollutants measured, PM2.5 has the most significant effect on the air quality index, followed by ozone, while *$NO_2$* has a negligible effect.

When applying intelligent analysis technologies, elasticity will help to assess the impact of each pollutant on the overall air quality, which will help to assess the accuracy of the identified hypotheses. In the future, the obtained results will help determine a strategy for improving air quality.

Based on the obtained elasticity results, the level of linear relationship between the pollutant PM2.5 and the overall index of air quality was estimated. Figure 5 shows the block diagram of the algorithm based on which a linear regression model is created [13, 14].

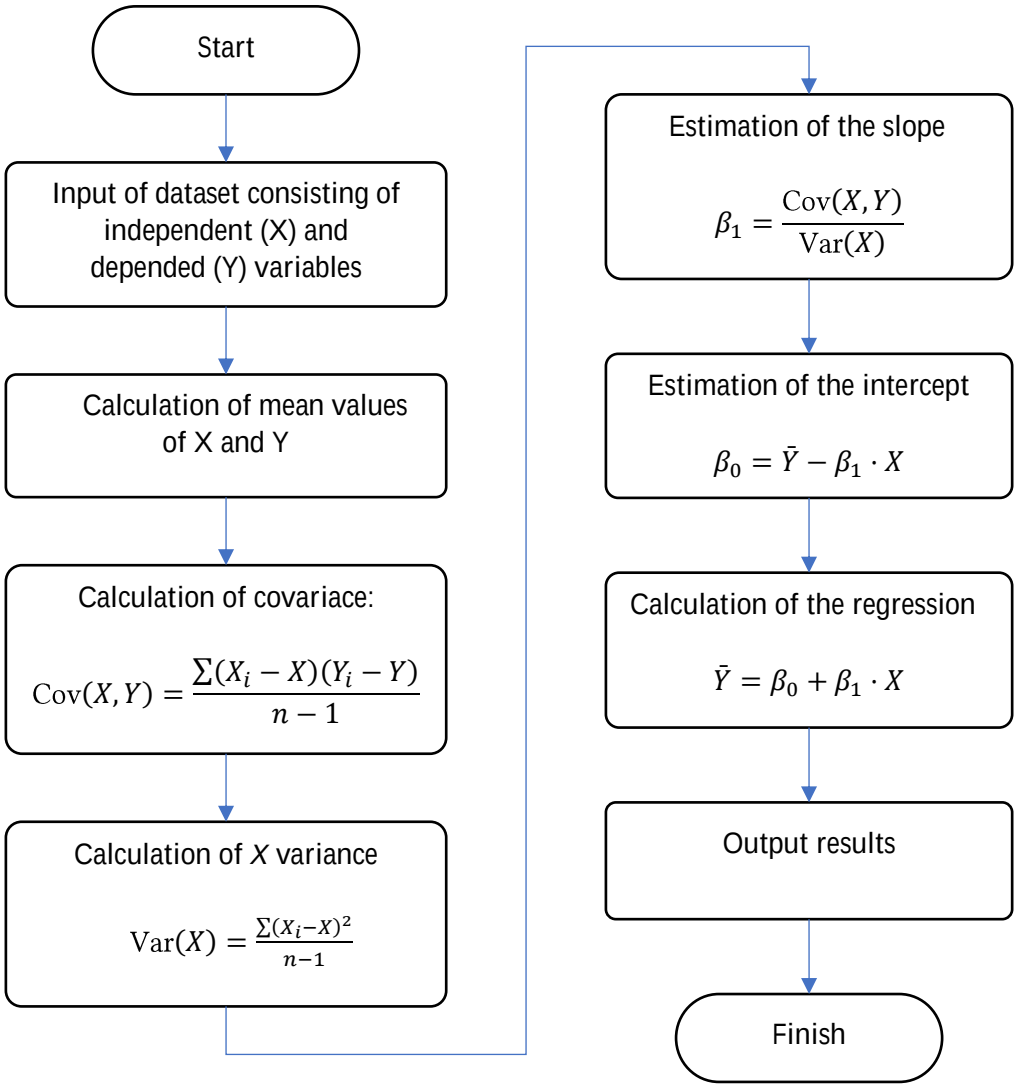


Figure 5: Flow chart for linear regression.

The mean square error, coefficient of determination (R2) and correlation between PM2.5 AQI and AQI were also calculated [15-17]. Its use can help reveal connections between different indicators of pollution, which allows for a better understanding of their impact on the overall ecological situation. The results of estimation of a linear regression model: mean squared error is 103.37, R2: 0.96897, and correlation coefficient is 0.9845.

Coefficient of determination ($R^2$) reached 0.969, which demonstrates that approximately 96.9% of the AQI (Air Quality Index) value can be explained by the AQI PM2.5 value. The correlation coefficient with a value of 0.985 demonstrates a strong linear relationship between the variables PM2.5 AQI Value and AQI Value (Figure 6).

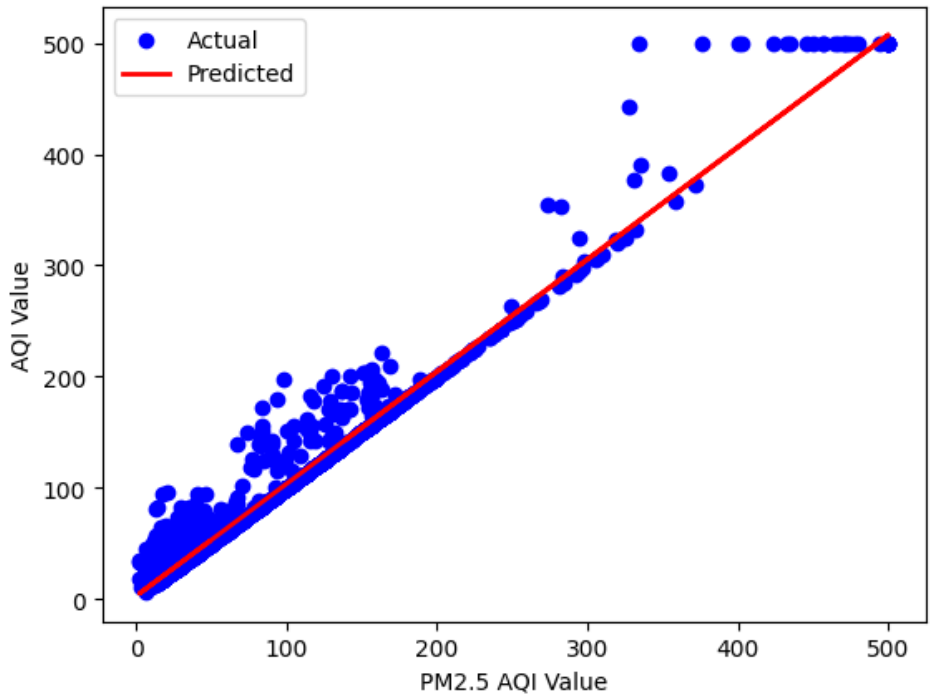


Figure 6: Scatter plot of actual air quality index values (blue dots) and line of predicted values (red line) based on PM2.5.

## 3. Conclusions

Paper presents results of statistical analysis of atmospheric air quality by using the methods of multivariate regression and elasticity. Application of these methods made it possible to determine the relative impact of pollutants on the overall air quality index. In particular, the above results showed that the levels of fine particulate matter (PM2.5) and ozone ($O_3$) had the greatest impact on the air quality index, while the impact of nitrogen dioxide ($NO_2$) was negligible.

The used methods can be effective tools for evaluating the impact of various factors on the overall air quality index. They can be adapted and applied to other data sets to determine the impact of different pollutants in other locations or over different time periods, which will contribute to a deeper understanding of the problem and increase the effectiveness of actions focused on improving air quality.

To obtain data on the concentration of the substances in the air, an unmanned aircraft with a payload in the form of a fluorescent or infrared spectroscope to determine the concentration of nitrogen dioxide and ozone would be an appropriate research tool, and the determination of the concentration of small particles is proposed to be performed using an aviation weather radar with appropriate software.

The use of statistical methods can be a component in the development of a decision support system, as they allow you to clean and prepare data for further analysis using OLAP and Data Mining technologies.

Further, in the process of forming hypotheses that arise during data analysis using OLAP and Data Mining, they can be confirmed or refuted using statistical methods. For example, with the help of elasticity, it is possible to quantitatively assess the relationships between various parameters that affect the quality of atmospheric air. On the basis of these estimates, it is possible to analyze the formed hypothesis and more accurately and effectively shape the decision-making process. Therefore, the use of statistical methods in combination with OLAP and Data Mining creates a fairly powerful tool for decision-making, which allows you to reduce risks and more effectively make decisions that will improve the quality of atmospheric air.

## References


[1] M. Zaliskyi, O. Solomentsev, V. Larin, Y. Averyanova, N. Kuzmenko, Model Building for Diagnostic Variables during Aviation Equipment Maintenance, in: proceedings of the IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2022, pp. 160–164. doi: 10.1109/CSIT56902.2022.10000556.

[2] O. Solomentsev, M. Zaliskyi, O. Sushchenko, Y. Bezkorovainyi, Y. Averyanova, Data Processing through the Lifecycle of Aviation Radio Equipment, in: proceedings of the IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 2022, pp. 146-151, doi: 10.1109/CSIT56902.2022.10000844.

[3] C.J. Wu, W.S. Zeng, J.M. Ho, Optimal Segmented Linear Regression for Financial Time Series Segmentation, in: proceedings of the International Conference on Data Mining Workshops (ICDMW), 2021, pp. 623–630.

[4] N. Kuzmenko, I. Ostroumov, Y. Bezkorovainyi, O. Sushchenko, Airplane Flight Phase Identification Using Maximum Posterior Probability Method, in: Proceedings of IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC), Kyiv, Ukraine, 2022, pp. 1–5, doi: 10.1109/SAIC57818.2022.9922913.

[5] Global Air Pollution Dataset. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset.

[6] Y. Averyanova, V. Larin, N. Kuzmenko, M. Zaliskyi, O. Solomentsev, Turbulence Detection and Classification Algorithm Using Data from AWR, in: Proceedings of IEEE 2nd Ukrainian Microwave Week (UkrMW), Ukraine, 2022, pp. 518–522. doi: 10.1109/UkrMW58013.2022.10037172.

[7] I. Ostroumov, et al., Relative navigation for vehicle formation movement, in: Proceedings of IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek), IEEE, Kharkiv, Ukraine, 2022, pp. 1–4, doi: 10.1109/KhPIWeek57572.2022.9916414.

[8] O. Sushchenko, Y. Bezkorovainyi, V. Golitsyn, N. Kuzmenko, Y. Averyanova, M. Zaliskyi, Integration of MEMS Inertial and Magnetic Field Sensors for Tracking Power Lines, in: Proceedings of IEEE XVIII International Conference on the Perspective Technologies and Methods in MEMS Design (MEMSTECH), Polyana, Ukraine, 2022, pp. 33–36, doi: 10.1109/MEMSTECH55132.2022.10002907.

[9] O. Sushchenko, et al., Airborne sensor for measuring components of terrestrial magnetic field, in: Proceedings of IEEE 41st International Conference on Electronics and Nanotechnology (ELNANO), IEEE, Kyiv, Ukraine, 2022, pp. 687–691. doi: 10.1109/ELNANO54667.2022.9926760.

[10] B. Golub, A. Hudz, A. Dudnyk, A. Bushma, Production of Biotechnological Objects using Business Intelligence, in: Proceedings of the 9th International Conference on Advanced Computer Information Technologies (ACIT), Ceske Budejovice, Czech Republic, 2019, pp. 200–204, doi: 10.1109/ACITT.2019.8780061.

[11] V. Romanov, O. Palagin, I. Galelyuka, O. Voronenko, Wireless Sensor Network for Precision Agriculture and Ecological Monitoring, Computer means, networks and systems 13 (2014) 53–62.
[12] D.Y. Yashchuk, B.L. Golub, Research on the Use of OLAP Technologies in Management Tasks, in: Z. Hu, S. Petoukhov, I. Dychka, M. He (Eds.), Advances in Computer Science for Engineering and Education (ICCSEEA 2018). Advances in Intelligent Systems and Computing, Springer, Cham, 2019. vol. 754. pp. 683–691. doi: 10.1007/978-3-319-91008-6_67.
[13] D. Liu, A. K. Mishra, D. K. Ray, Sensitivity of global major crop yields to climate variables: A non-parametric elasticity analysis. Science of the Total Environment 748 (2020) 141431.
[14] D. Maulud, A. M. Abdulazeez, A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends 1(2) (2020) 140–147.
[15] G. James, D.Witten, T. Hastie, R. Tibshirani, J. Taylor, Linear regression. In An introduction to statistical learning: With applications in python (2023) 69–134.
[16] A. Ali, K. Al-Hameed, Spearman's correlation coefficient in statistical analysis. International Journal of Nonlinear Analysis and Applications 13(1) (2022) 3249–3255.
[17] M. Berggren, Coefficients of determination measured on the same scale as the outcome: Alternatives to R2 that use standard deviations instead of explained variance. Psychological Methods (2024) 1–58.