Regression analysis as a tool for identifying patterns in atmospheric air monitoring data

Dmytro V. Shevchenko[1], Bella L. Holub[1]

[1]*National University of Life and Environmental Sciences of Ukraine Ukraine, Kyiv, Heroyiv Oborony st., 15, 03041*

Abstract

This study investigates the application of regression analysis as a tool for identifying patterns in atmospheric air quality monitoring data collected from IoT-based monitoring systems. Given the increasing importance of air pollution control and environmental safety, precise analytical methods are required to assess the relationships between pollutant concentrations and external factors such as meteorological conditions. The research focuses on employing the ordinary least squares (OLS) regression method to analyze the influence of key atmospheric parameters—including temperature, humidity, wind speed, and radiation—on air quality indices, specifically the Air Quality Index (AQI) and the Common Air Quality Index (CAQI). The study is based on data obtained from a network of IoT-enabled sensors deployed across various monitoring stations, which continuously measure air pollutants and meteorological parameters in real-time. A custom-built analytical module was developed to facilitate flexible data selection, enabling comparative assessments across different time frames, monitoring stations, and measurement parameters. Through statistical modeling, it was determined that radiation significantly influences both AQI and CAQI, while wind speed has a more pronounced effect during daytime hours. Furthermore, the study revealed that hourly aggregation of pollutant data is optimal for CAQI calculations, whereas daily averages better align with AQI assessments. The findings highlight the advantages of integrating IoT technology with regression-based analysis in modern air quality monitoring systems. By leveraging Python-based statistical tools such as StatsModels, this approach enables the identification of critical environmental factors affecting air pollution levels, thereby supporting more effective predictive modeling and decision-making in environmental policy and urban planning. Future research will explore the integration of nonlinear regression models and time series forecasting to further refine air quality assessments.

Keywords

IoT devices, edge computing, atmospheric air, air quality monitoring, pollutants, data analysis, statistical modeling, Python, StatsModels, linear regression, ordinary least squares (OLS), Common Air Quality Index (CAQI), Air Quality Index (AQI)

1. Introduction

1.1. Relevance

Atmospheric air is a critically important component of the ecosystem, providing oxygen to all living organisms and sustaining essential metabolic processes. Air pollution has a significant negative impact on human health, causing respiratory and cardiovascular diseases, weakening immune responses, and contributing to the development of chronic illnesses. At the same time, air pollution adversely affects animals and plants, compromising their health and vitality, which in turn negatively impacts ecosystems.

The state of atmospheric air quality is a determining factor for environmental safety and public health. In Ukraine, air quality monitoring involves the systematic collection, analysis, and processing of data on pollutant concentrations, serving as a foundation for developing state policies in environmental protection. Traditionally, monitoring has relied on centralized stations with limited spatial coverage. However, the advancement of IoT-based monitoring systems has enabled the deployment of low-cost, real-time air quality sensors, significantly enhancing data granularity and accessibility.

doors-2025: 5th Edge Computing Workshop, April 4, 2025, Zhytomyr, Ukraine

dimashevchenko10021999@gmail.com (D. V. Shevchenko); bellalg@nubip.edu.ua (B. L. Holub)
https://nubip.edu.ua/en (D. V. Shevchenko); https://nubip.edu.ua/en (B. L. Holub)
0009-0001-7736-8263 (D. V. Shevchenko); 0000-0002-1256-6138 (B. L. Holub)

Modern air quality monitoring systems utilize IoT-enabled devices (or edge devices) to measure pollutant levels and meteorological parameters continuously. These IoT sensors form a distributed network, transmitting data to centralized platforms for processing and analysis. This approach allows for real-time air quality assessment, early warning systems, and better urban environmental management.

Recognizing these challenges, the development of effective strategies for air quality monitoring and analysis becomes particularly important. Such approaches enable timely responses to threats, forecasting the impacts of pollution, and implementing measures to mitigate its adverse effects. The integration of IoT technology with statistical modeling, such as regression analysis, offers new opportunities for identifying hidden dependencies and trends in air pollution data.

1.2. Related work

In recent studies, significant attention has been given to integrating IoT technologies and machine learning methods for monitoring and predicting air quality. Ángel Martín-Baos et al. [1] developed a system combining IoT sensors for collecting pollution and traffic data in urban environments. By employing linear regression (LR), Gaussian process regression (GPR), and random forests (RF), the authors constructed models to assess the Air Quality Index (AQI) based on a limited set of parameters, allowing real-time insights into urban air quality and traffic conditions.

Similarly, Banciu et al. [2] proposed an air quality monitoring system that utilizes IoT devices to collect data on temperature, humidity, and particulate matter (PM10, PM2.5). The collected data is transmitted to the ThingSpeak cloud platform for storage and preliminary analysis. A regression model based on TensorFlow is applied for AQI prediction, providing timely alerts and recommendations for preventive measures.

Bobulski et al. [3] introduced an IoT-based pollution monitoring system with secure data transmission. This system enables localized real-time air quality monitoring and sends data to users. A key feature of the system is implementing a secure data transmission protocol, ensuring protection against cyber-attacks and data interception.

Ravindra et al. [4] focused on improving the accuracy of air quality sensors using machine learning methods. The study demonstrated that applying machine learning models for PM2.5 sensor calibration significantly enhances the accuracy of air quality monitoring systems.

Additionally, Dhanalakshmi and Radha [5] proposed an air pollution forecasting method based on discretized linear regression and multi-class support vector machines. The proposed IoT system enables monitoring and controlling air quality within a cloud computing environment, ensuring high prediction accuracy and reduced data processing time.

These studies highlight the effectiveness of combining IoT technologies and regression analysis for air quality monitoring and prediction. The integration of modern machine learning methods and secure data transmission improves the accuracy and reliability of monitoring systems, which is crucial for making informed decisions in environmental protection and public health. While global studies focus on improving monitoring accuracy and data transmission security, national-level reports highlight the limitations of existing monitoring infrastructure, such as in Kyiv, Ukraine.

The 2023 report by the Kyiv City State Administration [6] presents the results of atmospheric air quality monitoring based on data collected from 46 indicative and 7 reference automatic stations. Despite a high level of the Common Air Quality Index (CAQI), numerous exceedances of the maximum allowable concentrations (MAC) for nitrogen oxides were recorded, as well as significant exceedances of daily average MAC levels for ground-level ozone at nearly all reference stations. A comparison of these findings with the results from the Borys Sreznevskyi Central Geophysical Observatory revealed the technical obsolescence of the national air monitoring system, which fails to provide comprehensive and up-to-date information about air quality. This situation highlights the need to modernize the monitoring system and ensure regular access to data for informed decision-making aimed at improving air quality in Kyiv.

1.3. Objective of the study

The objective of the study is to utilize regression analysis to identify patterns in atmospheric air quality monitoring data collected from IoT-based environmental monitoring systems. The research aims to develop a methodology for analyzing dependencies that will support effective decision-making to improve the state of atmospheric air.

2. System architecture

The air quality monitoring system is based on IoT-enabled devices that continuously collect real-time data on key air parameters, including pollutant concentrations, temperature, humidity, and other environmental indicators.

The spatial distribution of these IoT sensors plays a crucial role in data reliability. Urban regions with high traffic density require a denser network of sensors compared to suburban or rural areas, where air circulation patterns differ significantly. In Kyiv, for example, the placement of sensors near industrial zones and major highways ensures a more accurate representation of pollution sources.

These IoT devices serve as edge computing units, transmitting collected data to cloud-based or on-premise servers for further analysis. The architecture of the monitoring system is illustrated in figure 1.

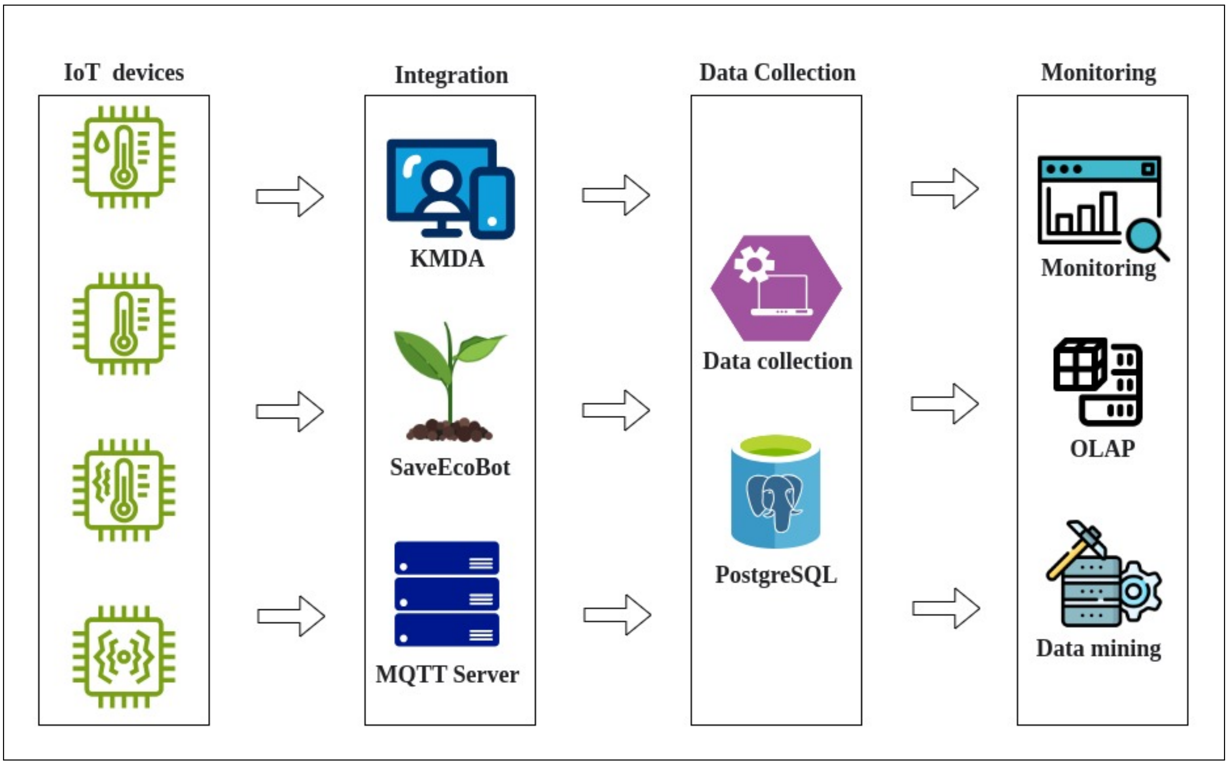


Figure 1: Monitoring system architecture.

The collected data is integrated with other sources, such as municipal monitoring platforms (e.g., Kyiv City State Administration), community initiatives (SaveEcoBot), and MQTT servers. This multi-source system enables a comprehensive view of air quality and ensures data relevance. The data is transmitted to a centralized database built on PostgreSQL, providing reliable storage and accessibility for further processing.

This integration allows for cross-validation of sensor readings with officially recognized monitoring stations, reducing the impact of sensor biases and ensuring higher data reliability. By leveraging municipal platforms, real-time data can be incorporated into existing environmental management systems, facilitating timely responses to pollution spikes.

An essential component of such systems is the organized storage and processing of datasets. In our architecture, these tasks are managed using the PostgreSQL database management system (DBMS). This open-source relational DBMS is well-suited for efficiently managing large datasets and executing advanced queries. The database's logical structure is depicted in figure 1.

A vital element of the system is a set of tables designed to store data related to monitoring stations and their recorded measurements. The station table holds detailed information about each monitoring station, including unique identifiers, geographical coordinates, type of data source (e.g., stationary or mobile), and operational status.

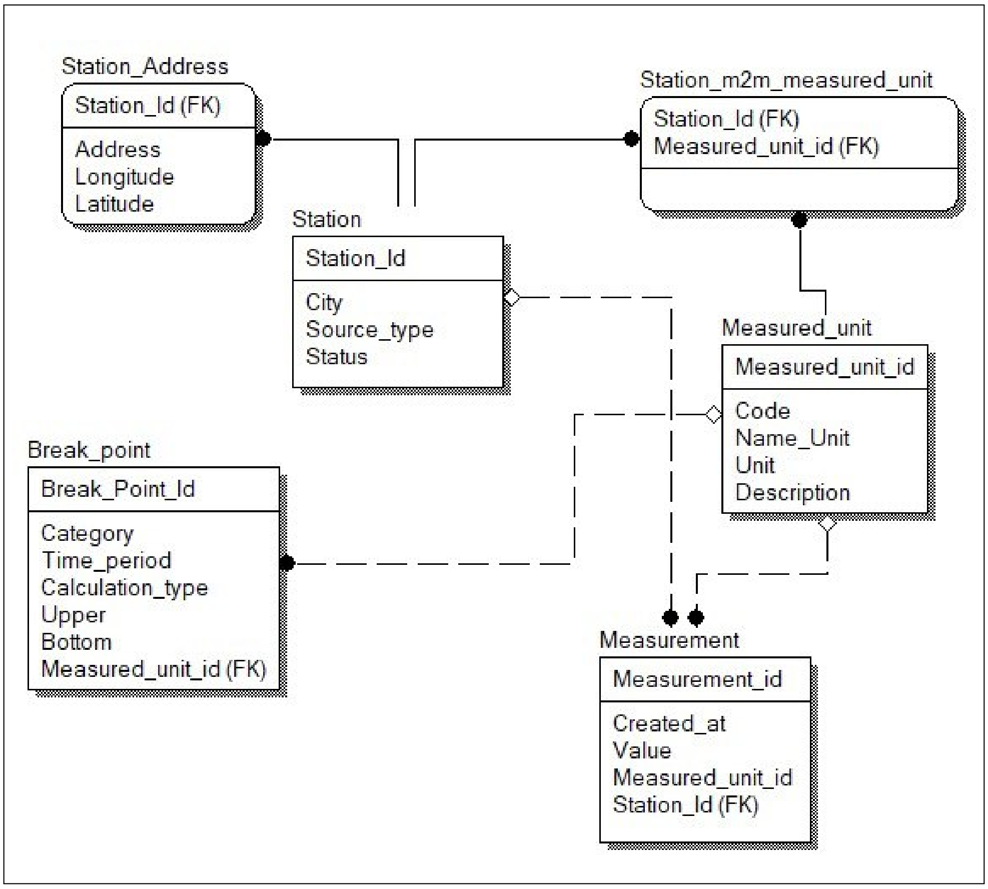


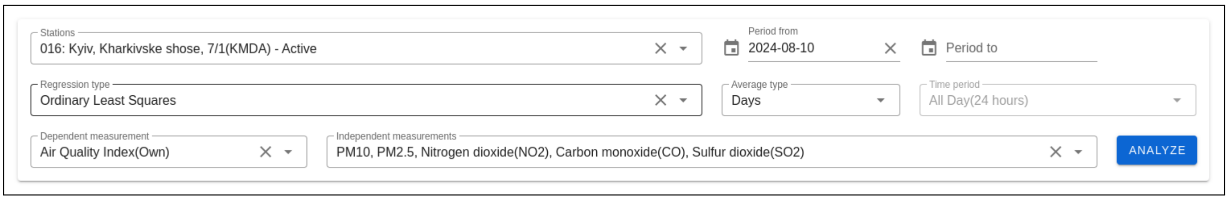
Figure 2: Logical model of the database.

Based on the accumulated data, two approaches are implemented for assessing air quality: the Air Quality Index (AQI, USA) and the Common Air Quality Index (CAQI, Europe). Both indices serve as key indicators of air pollution levels and reflect potential health risks, though there are notable differences between them:

- AQI (Air Quality Index): The AQI is calculated based on the concentrations of primary air pollutants, such as fine particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), and carbon monoxide (CO). It is widely used to provide a clear assessment of health risks depending on pollution levels.
- CAQI (Common Air Quality Index): The CAQI is a standard adopted in European cities, also based on pollutant concentrations, including PM2.5, PM10, NO_2, O_3, and SO_2. This index allows for a unified evaluation of air quality, which is crucial for regional comparisons.

3. Results

As part of the air quality monitoring system, a module was developed to enable flexible configuration of parameters for data analysis using the OLS regression model, implemented via the Statsmodels library [7], a powerful tool for statistical modeling in Python. This module facilitates the analysis of dependencies across selected stations, measurement parameters, and various time intervals. The tool allows for personalized queries through filtering options, as illustrated in figure 3.



Stations: 016: Kyiv, Kharkivske shose, 7/1(KMDA) - Active
Period from: 2024-08-10
Period to
Regression type: Ordinary Least Squares
Average type: Days
Time period: All Day(24 hours)
Dependent measurement: Air Quality Index(Own)
Independent measurements: PM10, PM2.5, Nitrogen dioxide(NO2), Carbon monoxide(CO), Sulfur dioxide(SO2)

Figure 3: Configuration and parameter settings for analysis.

The user can specify any time range for analysis, including both daily and hourly intervals. This flexibility enables targeted investigations, such as focusing on peak activity periods or unusual meteorological conditions. Additionally, the system allows for selecting either a single station or multiple stations, which is particularly useful for comparative analysis across different areas or locations.

The regression analysis results yielded several key components that characterize the model's quality and the relationships between variables, as shown in figure 4. The primary model metrics include the proportion of variance in the dependent variable explained by the independent variables (R-squared). The F-statistic was also calculated to evaluate the overall significance of the model, along with its p-value, confirming statistical relevance.

Furthermore, residual analysis provides deeper insights into the model's fit to the data. For each independent variable, a regression coefficient table (β) was generated, containing standard errors, t-values, and p-values. The importance of variables is assessed based on their statistical significance, helping to identify the most impactful factors.

The visualization of results includes a prediction plot that illustrates the relationship between the actual values of the dependent variable and those predicted by the model. This plot demonstrates the model's accuracy, providing a visual assessment of its fit to the real data and its predictive performance. The graph highlights the degree to which the predicted values align with the actual observations, showcasing the effectiveness of the regression model in capturing key patterns and dependencies within the data.

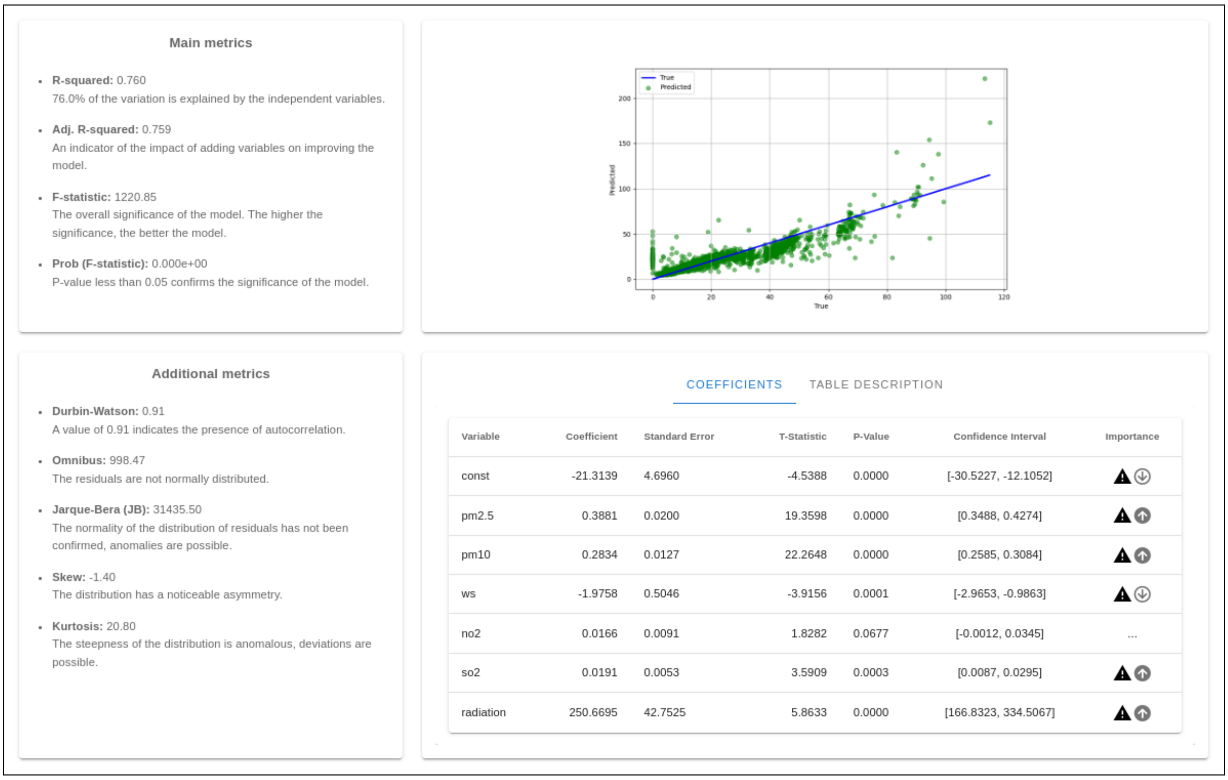
4. Discussion

The analysis of results obtained through the developed module revealed important patterns in the formation of air quality indices (AQI and CAQI) and the nuances of their application. The radiation factor emerged as a significant variable, substantially influencing the calculation of both indices.

When comparing the approaches to index calculation, it was found that hourly averages are more effective for CAQI, while daily averages are more appropriate for AQI (figure 5). This difference reflects the distinct algorithms used for each index and their impact on the final results. Such findings allow for more precise adaptation of analysis to specific tasks and regional characteristics.

An intriguing observation concerns the influence of wind on CAQI. During nighttime, wind has minimal impact on the air quality index, while during daytime or over an entire day, its influence becomes noticeable. This suggests a reduction in wind intensity at night, diminishing its capacity to disperse pollutants and, consequently, affecting the modeling outcomes.

These findings highlight the importance of tailoring air quality analyses to account for specific temporal and environmental factors, ensuring more accurate and actionable insights. Despite the advantages of using IoT-based monitoring systems and regression analysis, several limitations exist. IoT



Main metrics

- **R-squared:** 0.760
  76.0% of the variation is explained by the independent variables.
- **Adj. R-squared:** 0.759
  An indicator of the impact of adding variables on improving the model.
- **F-statistic:** 1220.85
  The overall significance of the model. The higher the significance, the better the model.
- **Prob (F-statistic):** 0.000e+00
  P-value less than 0.05 confirms the significance of the model.

Additional metrics

- **Durbin-Watson:** 0.91
  A value of 0.91 indicates the presence of autocorrelation.
- **Omnibus:** 998.47
  The residuals are not normally distributed.
- **Jarque-Bera (JB):** 31435.50
  The normality of the distribution of residuals has not been confirmed, anomalies are possible.
- **Skew:** -1.40
  The distribution has a noticeable asymmetry.
- **Kurtosis:** 20.80
  The steepness of the distribution is anomalous, deviations are possible.

Variable	Coefficient	Standard Error	T-Statistic	P-Value	Confidence Interval	Importance
const	-21.3139	4.6960	-4.5388	0.0000	[-30.5227, -12.1052]	⚠ ↓
pm2.5	0.3881	0.0200	19.3598	0.0000	[0.3488, 0.4274]	⚠ ↑
pm10	0.2834	0.0127	22.2648	0.0000	[0.2585, 0.3084]	⚠ ↑
ws	-1.9758	0.5046	-3.9156	0.0001	[-2.9653, -0.9863]	⚠ ↓
no2	0.0166	0.0091	1.8282	0.0677	[-0.0012, 0.0345]	...
so2	0.0191	0.0053	3.5909	0.0003	[0.0087, 0.0295]	⚠ ↑
radiation	250.6695	42.7525	5.8633	0.0000	[166.8323, 334.5067]	⚠ ↑

Figure 4: Results of regression analysis for CAQI.

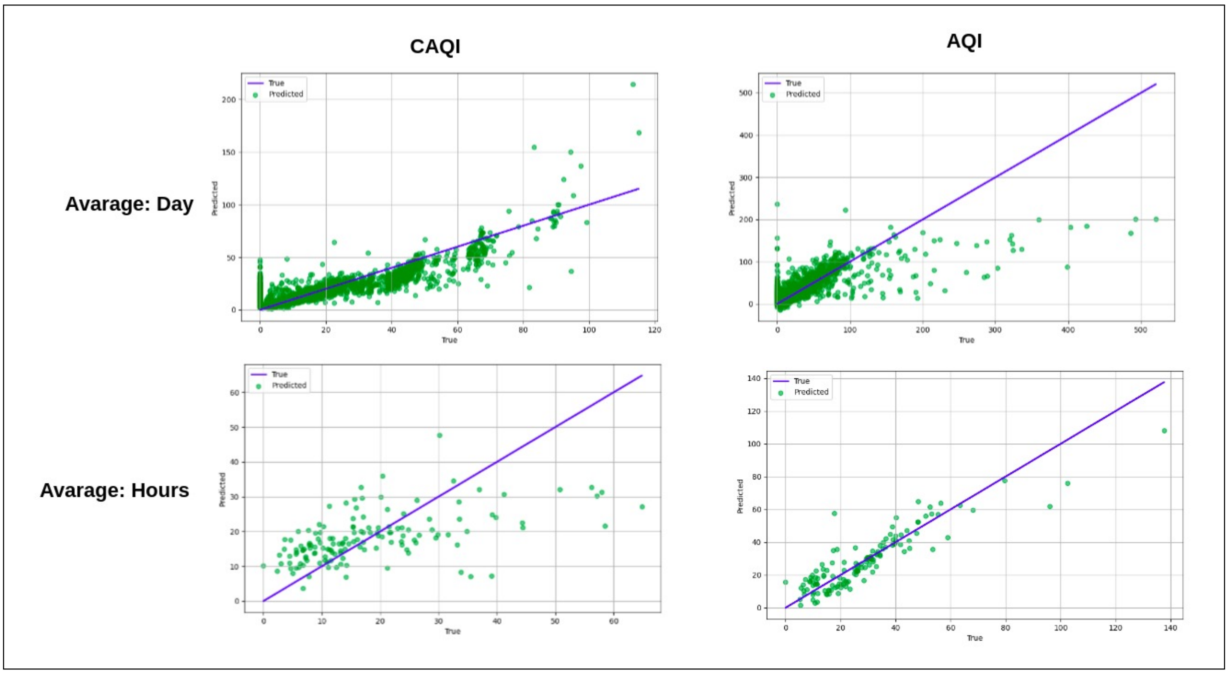


Figure 5: Comparing AQI and CAQI with different average type.

sensors are prone to measurement drift over time, requiring regular calibration to maintain accuracy. Additionally, regression models assume linear relationships between variables, which may not fully capture complex atmospheric interactions. Incorporating nonlinear models or hybrid approaches with machine learning could help address these challenges.

5. Conclusions

This study confirms the effectiveness of applying regression analysis to identify patterns in atmospheric air quality monitoring data, particularly in combination with IoT-based monitoring systems. The integration of real-time sensor networks allows for more accurate and dynamic assessment of pollutant levels and meteorological factors, improving the overall reliability of air quality analysis. Unlike traditional methods that rely on static datasets from reference stations, this research introduces a flexible approach based on Ordinary Least Squares (OLS) regression, which enables the identification of statistically significant dependencies. One of the key findings is the impact of radiation levels on air quality indices, a factor often overlooked in previous studies. The results also highlight fundamental differences in data aggregation strategies, demonstrating that hourly averaging is more suitable for CAQI, whereas daily averaging better aligns with AQI calculations. These insights allow for a more precise selection of analytical techniques depending on the index used, enhancing predictive accuracy in air quality assessments.

A significant contribution of this work lies in the development of a custom analytical module that facilitates flexible configuration of regression parameters, allowing users to explore pollutant dependencies across different periods, stations, and environmental conditions. This tool supports decision-making processes in environmental management by enabling targeted analysis of pollution dynamics. The use of Python-based statistical libraries, particularly StatsModels, proves to be a cost-effective alternative to more complex machine learning models while maintaining high interpretability.

The study not only advances the methodological approach to air quality assessment but also sets the foundation for future research. Expanding the analytical framework with nonlinear regression models and time series forecasting could improve predictive capabilities, while further integration of data mining techniques may reveal hidden dependencies within air pollution trends. The findings underscore the importance of real-time monitoring systems and data-driven approaches for environmental policy-making, providing a technological basis for more adaptive and responsive air quality management strategies.

The developed analytical framework can be utilized by policymakers to design data-driven air pollution control strategies. By identifying critical pollution sources and peak contamination periods, urban planners can implement targeted interventions such as emission restrictions in high-risk areas, traffic rerouting, or increased green zones to mitigate pollution levels.

Author Contributions: Conceptualization, Dmytro V. Shevchenko and Bella L. Holub; methodology, Dmytro V. Shevchenko; software, Dmytro V. Shevchenko; writing – original draft, Dmytro V. Shevchenko; writing – review and editing, Bella L. Holub. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analysed during this study. Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration on Generative AI: During the preparation of this work, the authors used GPT-4o in order to: Drafting content, Generate literature review, Paraphrase and reword, Improve writing style, Abstract drafting, Grammar and spelling check, Content enhancement. After using this service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] J. Ángel Martín-Baos, L. Rodriguez-Benitez, R. García-Ródenas, J. Liu, IoT based monitoring of air quality and traffic using regression analysis, Applied Soft Computing 115 (2022) 108282. doi:10.1016/j.asoc.2021.108282.

[2] C. Banciu, A. Florea, R. Bogdan, Monitoring and Predicting Air Quality with IoT Devices, Processes 12 (2024) 1961. doi:10.3390/pr12091961.

[3] J. Bobulski, S. Szymoniak, K. Pasternak, An IoT System for Air Pollution Monitoring with Safe Data Transmission, Sensors 24 (2024) 445. doi:10.3390/s24020445.

[4] K. Ravindra, S. Kumar, A. Kumar, S. Mor, Enhancing accuracy of air quality sensors with machine learning to augment large-scale monitoring networks, npj Climate and Atmospheric Science 7 (2024) 326. doi:10.1038/s41612-024-00833-9.

[5] M. Dhanalakshmi, V. Radha, Discretized Linear Regression and Multiclass Support Vector Based Air Pollution Forecasting Technique, International Journal of Engineering Trends and Technology 70 (2022) 315–323. doi:10.14445/22315381/ijett-v70i11p234.

[6] Kyiv City State Administration, Discussion of the Kyiv City State Administration Report on Atmospheric Air Quality Monitoring Results in Kyiv for 2023, 2023. URL: https://nubip.edu.ua/node/141976.

[7] StatsModels Developers, StatsModels Ordinary Least Squares Documentation, 2023. URL: https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html.