A Big Data Framework for Automated Environmental Risk Stratification Using a Composite Hazard Index and Unsupervised Clustering

Kyrylo Vadurin[1], Olena Kortsova[2], Kateryna Vasiutynska[3], Andrii Perekrest[1] and Volodymyr Bakharev[4]

[1]*Department of Computer Engineering and Electronics, Kremenchuk Mykhailo Ostrohradskyi National University, Universytetska Str. 20, 39600 Kremenchuk, Ukraine*

[2]*Department of Ecology and Biotechnologies, Kremenchuk Mykhailo Ostrohradskyi National University, Universytetska Str. 20, 39600 Kremenchuk, Ukraine*

[3]*Department of Environmental Safety and Hydraulics, National University Odesa Polytechnic, Shevchenko Avenue 1, 65044 Odesa, Ukraine*

[4]*Educational and Scientific Institute of Mechanical Engineering, Transport and Natural Sciences, Kremenchuk Mykhailo Ostrohradskyi National University, Universytetska Str. 20, 39600 Kremenchuk, Ukraine*

kir3337@gmail.com, equivalent.eco@gmail.com, e.a.vasutinskaya@op.edu.ua, pksg13@gmail.com, v.s.baharev@gmail.com

Keywords: Big Data, Environmental Risk Assessment, Composite Hazard Index, Unsupervised Clustering, Decision Support System.

Abstract: The proliferation of IoT-based environmental monitoring networks has led to an exponential increase in air quality data, creating a significant challenge for effective and timely risk assessment. Traditional analysis methods often struggle to convert this vast volume of raw data into actionable insights for decision-makers. This paper introduces a novel big data framework designed to automate the process of environmental risk assessment through a two-stage analytical approach. First, we propose a Composite Hazard Index (CHI), a quantitative metric that normalizes and aggregates data from disparate pollutants into a single, interpretable risk score for each monitoring location, incorporating expert-defined weights to reflect the relative toxicity of different substances. Second, this index, along with normalized pollutant profiles, is used as input for an unsupervised clustering algorithm (K-Means) to automatically stratify locations into distinct risk tiers (e.g., Low, Medium, High). A key novelty of our approach is a descriptive analysis method that identifies the primary pollutants, or "key risk drivers," responsible for the elevated hazard levels in high-risk clusters. The framework was validated using a real-world dataset from a public air quality monitoring network. The results demonstrate the system's ability to successfully identify high-risk "hotspots," quantify their danger level, and pinpoint the specific pollutants contributing to the threat, thereby providing a powerful, data-driven tool for proactive environmental management.

1 INTRODUCTION

The global challenge of air pollution, with its well-documented adverse effects on public health and ecosystems, has entered a new era defined by data abundance [1]. The widespread deployment of low-cost Internet of Things (IoT) sensors and citizen science initiatives has generated unprecedented volumes of high-frequency environmental data [2]. This "data deluge" presents a dual opportunity and challenge. On one hand, it offers the potential for granular, real-time analysis of environmental dynamics. On the other, the sheer volume, velocity, and heterogeneity of this data overwhelm traditional analytical approaches, creating a critical gap between data availability and the ability to derive actionable intelligence for environmental governance [3].

1.1 Related Work

A review of current scientific literature reveals several distinct but often fragmented approaches to this problem. A significant body of research focuses on predictive modeling, employing machine learning techniques like Long Short-Term Memory (LSTM) networks to forecast pollutant concentrations [1, 2]. While valuable for short-term warnings, these models often operate on raw concentration values and do not inherently provide an integrated assessment of the overall environmental risk, which is often a product of multiple interacting pollutants. Furthermore, their accuracy is fundamentally dependent on the quality of input data, a major issue given the known reliability concerns of low-cost sensors [3].

Another prominent research direction involves the use of exploratory data analysis, particularly unsupervised clustering algorithms, to identify spatial patterns and pollution "hotspots" [4, 5]. For instance, Hu et al. (2022) applied clustering and spatial statistics to classify urban pollution levels, successfully identifying distinct geographical patterns [5]. More advanced techniques like co-clustering have been used to simultaneously group spatial and temporal objects, revealing complex interactions [6]. However, a common limitation of these methods is their descriptive nature. They effectively answer the question "Where are the problem areas?" but often fall short of quantifying the severity of the risk or identifying its primary drivers, leaving interpretation to human experts.

To bridge this gap, hybrid methodologies have emerged, combining clustering with physical-chemical receptor models like Positive Matrix Factorization (PMF) to link clusters with specific emission sources (e.g., traffic, industry) [7, 8]. This approach provides a powerful diagnostic tool, offering insights into the "Why?" behind pollution patterns. Nevertheless, it typically requires detailed and costly chemical speciation data, which is not available from standard monitoring stations that measure bulk pollutants like PM2.5 or CO.

1.2 Problem Formulation

Based on the identified gaps in the existing literature, this research addresses the challenge of transforming high-volume, multi-pollutant monitoring data into a prioritized and interpretable risk assessment. The problem can be formally stated as follows:

GIVEN:

1) A raw dataset of time-series air quality measurements from a set of m monitoring locations, where each record contains a location identifier, timestamp, pollutant type, and concentration value.
2) A set of regulatory standards (e.g., Maximum Allowable Concentrations) for each of the n monitored pollutants.
3) A set of expert-defined raw weights representing the relative health impact or importance of each pollutant.

FIND:

1) A single, quantitative Composite Hazard Index (CHI) for each of the m locations that represents the integrated environmental risk.
2) An automated stratification of the m locations into a predefined number of K discrete risk tiers (e.g., "Low," "Medium," "High") based on their pollutant profiles and overall CHI.
3) For the highest risk tier, an identification of the key pollutants (risk drivers) that are the primary contributors to the elevated risk level, quantified by a comparative metric.

2 MAIN PART

2.1 Research Methodology

2.1.1 Data Pre-Processing and Profile Generation

The initial stage transforms raw, heterogeneous time-series data into a structured, analysis-ready format. This is a prerequisite for any meaningful analysis of environmental indicators [9]. The process involves several steps:

1) Data from disparate sources and timeframes are consolidated into a unified structure, with each record containing a location identifier, timestamp, pollutant type, and concentration value.
2) Absolute concentration values are converted into a dimensionless, universally comparable metric of risk. This is achieved by dividing each measurement by its corresponding regulatory standard, such as the Maximum Allowable Concentration (MAC). A value of 2.5 after this step is unequivocally interpreted as "the concentration is 2.5 times the established safe limit," regardless of the pollutant.

3) To handle missing data and irregular measurement intervals, time-series for each unique "location-pollutant" pair are resampled onto a regular time grid (e.g., daily). Gaps are filled using linear interpolation, a reasonable assumption for environmental processes over short intervals.
4) The processed time-series data is aggregated over a defined analytical period (e.g., one year) to create a static "environmental profile" for each monitoring location. This profile is represented as a feature vector where each element is the mean MAC-normalized value for a specific pollutant. This results in a feature matrix X of size $m \times n$, where m is the number of locations and n is the number of pollutants.

2.1.2 Composite Hazard Index Calculation

To move from a multi-pollutant profile to a single, rankable measure of overall danger, we introduce the CHI. The calculation is a three-step process:

1) The feature matrix X (containing mean MAC-normalized values) is further normalized to a scale of $[0,1]$ to ensure all risk factors contribute equally to the index calculation, regardless of their original range. For each pollutant j, the normalized value for location i is calculated as:

$$x_{norm,ij} = \frac{x_{ij} - \min(X_j)}{\max(X_j) - \min(X_j)}, \quad (1)$$

where $\min(X_j)$ and $\max(X_j)$ are the minimum and maximum mean values for pollutant j across all locations.

2) The framework allows for the incorporation of expert knowledge by assigning weights to each pollutant based on its perceived toxicity or local importance. A user-defined vector of raw weights $W_{\text{raw}} = [w_{\text{raw},1}, \ldots, w_{\text{raw},n}]$ is normalized to sum to 1:

$$W_{norm,j} = \frac{w_{raw,j}}{\sum_{k=1}^{n} w_{raw,k}}. \quad (2)$$

The use of expert-defined weights, rather than a machine learning approach, is a deliberate decision that prioritizes interpretability, control, and practicality.

Experts assign these transparent weights based on established public health knowledge, pollutant toxicity, and regulatory standards, such as reference concentrations (RfCs). This ensures the resulting risk score is scientifically justified and aligned with health priorities, for example, by giving a higher weight to PM2.5 due to its significant cardiorespiratory effects. This method also allows authorities to adjust weights based on local concerns or new toxicological findings.

From a practical standpoint, a machine learning model would require a large, labeled dataset linking pollution mixes to verified health outcomes, which is exceptionally rare. Therefore, the expert-driven system provides a more robust and operationally feasible solution for decision support, while still allowing for future refinement through machine learning as suitable data becomes available.

3) The CHI for each location i is computed as the weighted sum (scalar product) of its normalized pollutant vector and the normalized weight vector:

$$CHI_i = \sum_{j=1}^{n} x_{norm,ij} \cdot w_{norm,j}. \quad (3)$$

The result is a vector $R = [\text{CHI}_1, \ldots, \text{CHI}_m]$, where each CHI_i is a value in $[0,1]$ representing the integrated environmental hazard at that location.

2.1.3 Unsupervised Risk Stratification (Auto-Clustering)

This stage uses the generated profiles and indices to automatically group locations into meaningful risk tiers. We employ the K-Means clustering algorithm, but with a novel feature engineering approach to guide the process.

1) Instead of clustering on the pollutant profiles alone, we create an augmented feature vector Z_i^* for each location i. This vector concatenates the normalized pollutant profile with its calculated CHI:

$$Z_i^* = \left[x_{norm,j1}, x_{norm,j2}, \cdots, x_{norm,jn}, CHI_i \right]. \quad (4)$$

This augmentation forces the clustering algorithm to consider not only the similarity in the mix of pollutants but also the overall level of hazard as captured by the CHI.

2) The K-Means algorithm is applied to the set of augmented feature vectors {Z1*, … Zm*}. For

risk stratification, the number of clusters K is fixed (typically K =3) to represent “Low Risk,” “Medium Risk,” and “High Risk” tiers. The algorithm partitions the locations into K clusters C={C1, …Ck} by minimizing the within-cluster sum of squares (inertia):

$$\arg\min_{C} \sum_{k=1}^{K} \sum_{Z_i^* \in C_K} \left\| Z_i^* - \mu_k^* \right\|^2, \qquad (5)$$

where μ_k^* is the centroid of cluster C_k.

3) After partitioning, the clusters are ordered based on the average CHI of their constituent locations, ensuring that Cluster 0 corresponds to the lowest risk and Cluster K–1 to the highest.

2.1.4 Key Risk Driver Analysis

The final stage provides an interpretable explanation for why a cluster is classified as high-risk. Instead of relying on complex statistical tests that may be unreliable with small sample sizes, we use a straightforward descriptive metric.

For each pollutant j, we calculate its mean value (from the original, un-normalized feature matrix X) within the “High Risk” cluster (C_{High}) and the “Low Risk” cluster (C_{Low}). The Risk Multiplier (M_j) is then calculated as the ratio of these means:

$$M_j = \frac{mean(X_j \ in \ C_{High})}{mean(X_j \ in \ C_{Low})}. \qquad (6)$$

Such a mathematical model is easily integrated and can be implemented both in public monitoring information systems [10] and used for municipal monitoring.

2.2 Results and Discussion

The proposed framework was validated using the "Air Quality Monitoring from EcoCity" dataset [11], a comprehensive, real-world collection of public monitoring data from the city of Vinnytsia, Ukraine. For this validation, higher expert weights were assigned to particulate matter (PM2.5), reflecting its significant impact on respiratory health, as detailed in the "Risk Factor Importance" charts within the generated reports. The analysis was conducted for two periods: a historical review from January 1, 2021, to September 9, 2023, and a forecasted period from September 9, 2023, to October 9, 2023, to assess both retrospective patterns and future trends.

2.2.1 Risk Index Calculation and Stratification

After pre-processing the historical data, the framework calculated a CHI for each monitoring station. The results, visualized in the "Comprehensive Risk Index" chart (Fig. 1), revealed a stark distribution of risk, with CHI scores ranging from approximately 0.1 to a severe outlier of 0.786 [12].

The subsequent risk stratification, using K-Means with K=3, produced a clear and actionable grouping of the stations into three distinct risk tiers. The results for the historical period are summarized in Table 1.

The clustering process unequivocally isolated a single station, s-1612, as the sole member of the "High Risk" cluster. This result immediately focuses analytical and regulatory attention on a specific geographical point of major concern. The quantitative nature of the CHI provides critical context for this classification; its index of 0.786 is nearly double the index of the "Medium Risk" station and four times higher than the "Low Risk" average [13].

This is further confirmed by the spatial distribution map (Fig. 1), where s-1612 is represented by a significantly larger circle, visually identifying it as a risk "hotspot".

2.2.2 Identification of Key Risk Drivers

To understand the underlying cause of this anomaly, the Key Risk Driver Analysis was performed. This analysis compares the average, real-world pollutant concentrations (in mg/m³) between the "High Risk" cluster (containing only s-1612) and the "Low Risk" cluster. The results for the historical period, detailed in the "In-Depth Analytical Report" (Fig. 2), were striking, as shown in Table 2.

The analysis provides an unambiguous conclusion: the extreme risk associated with station s-1612 is overwhelmingly driven by particulate matter. The average concentrations of PM10 and PM2.5 at this location are over 400-500 times higher than in the safest zones. In contrast, the levels of gaseous pollutants like CO and NO_2 are nearly identical across all clusters [14]. This finding effectively rules out generalized urban pollution (e.g., from traffic) as the primary cause and strongly suggests the presence of a powerful, localized source of particulate emissions in the immediate vicinity of station s-1612. The "Anatomy of Risk" chart (Fig. 2) visually corroborates this, showing that the risk index for s-1612 is almost entirely composed of the contribution from PM2.5.

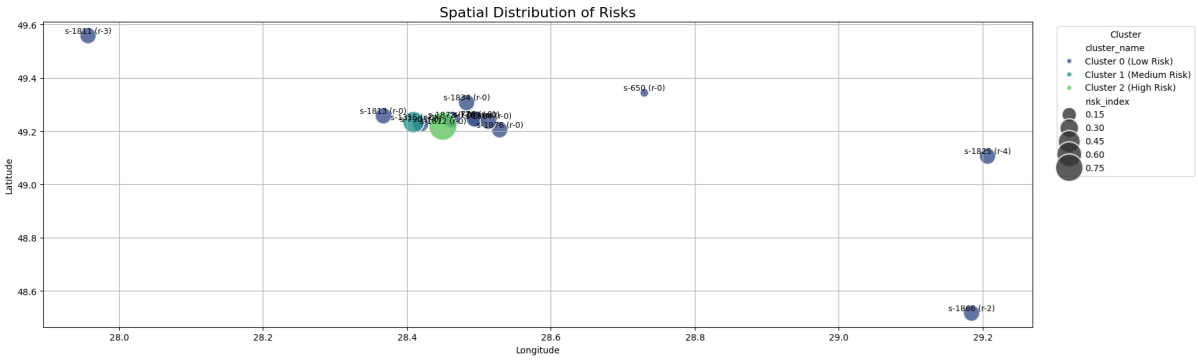

Figure 1: Spatial distribution of the CHI across monitoring stations for the historical period (2021-01-01 to 2023-09-09).

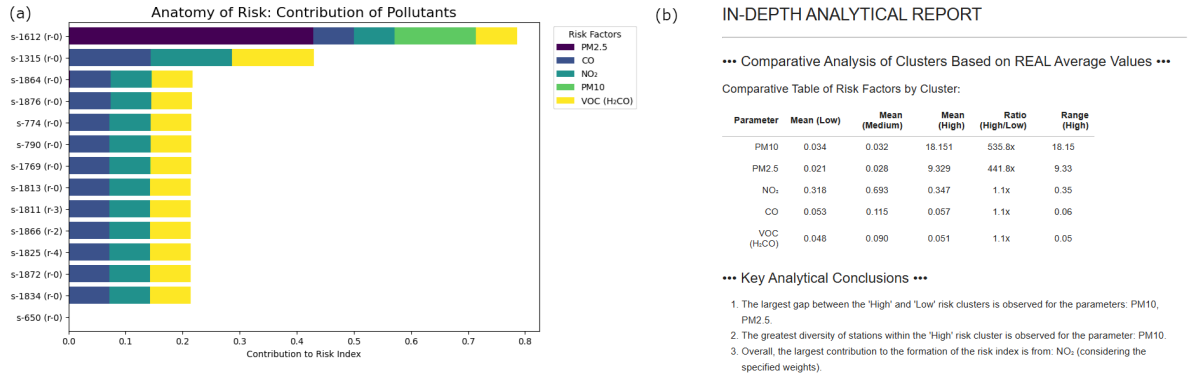

Figure 2: Analytical dashboard for the historical period (2021-01-01 to 2023-09-09): a) pollutant contribution to the CHI for each station and b) summary of the Key Risk Driver Analysis.

Table 1: Results of Risk Stratification on Historical Data (2021-01-01 to 2023-09-09).

Risk Tier	Low Risk	Medium Risk	High Risk
Cluster ID	0	1	2
Number of Stations	12	1	1
Average CHI	0.197	0.430	0.786
Key Stations Identified	s-650, s-1834, s-774, etc.	s-1315	s-1612

Table 2: Key Risk Driver Analysis for Historical Data.

Pollutant	Mean (Low Risk Cluster)	Mean (High Risk Cluster)	Risk Multiplier
PM10	0.034	18.151	535.8x
PM2.5	0.021	9.329	441.8x
NO_2	0.318	0.347	1.1x
CO	0.053	0.057	1.1x

2.2.3 Spatio-Temporal Dynamics and Regional Impact

The framework's ability to analyze forecasted data provides crucial insights into the stability of these risk patterns. The analysis of the forecasted period shows that the risk structure at the station level remains remarkably stable. Station s-1612 persists as the single occupant of the "High Risk" cluster, and while the forecasted absolute concentrations are lower, the risk multipliers for PM10 (30.6x) and PM2.5 (24.5x) remain an order of magnitude higher than for any other pollutant. This indicates that the identified pollution source is likely constant and its impact will persist, making it a priority for long-term strategic intervention.

Aggregating the results to a regional level further clarifies the impact of this local hotspot. The regional analyses for both the historical (Fig. 4) and forecasted (Fig. 3) periods consistently identify region r-0 – the region to which station s-1612 belongs – as the area with the highest environmental risk. The "Structure of Risk" charts for both periods demonstrate that the elevated risk in region r-0 is driven by a disproportionately high contribution from particulate matter, a direct consequence of the anomaly at s-1612. This highlights the system's ability to trace the impact of a micro-level issue on macro-level regional assessments.

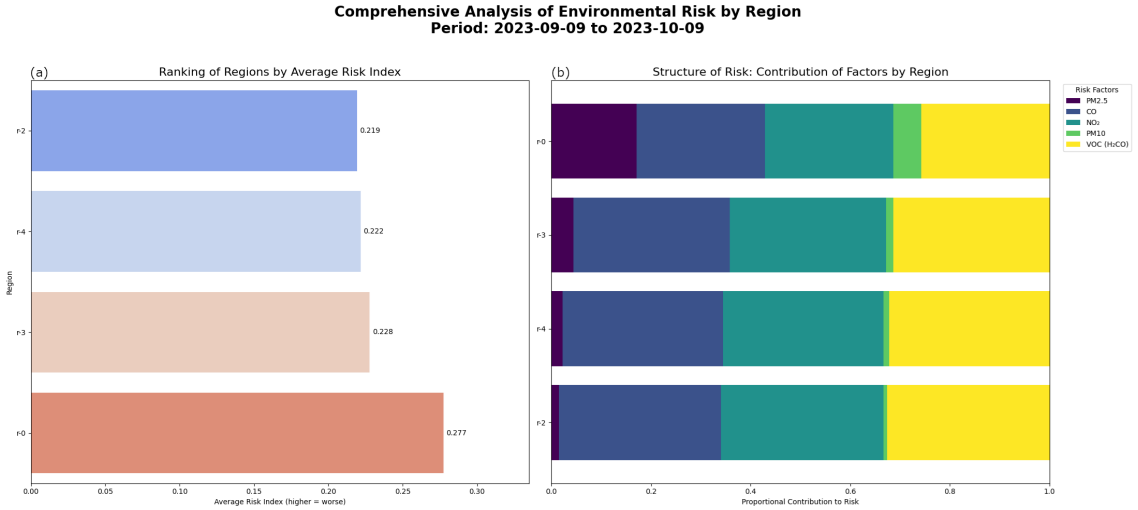

Figure 3: Regional environmental risk analysis for the forecasted period (2023-09-09 to 2023-10-09): a) Ranking of regions by average CHI and b) proportional contribution of each pollutant to regional risk.

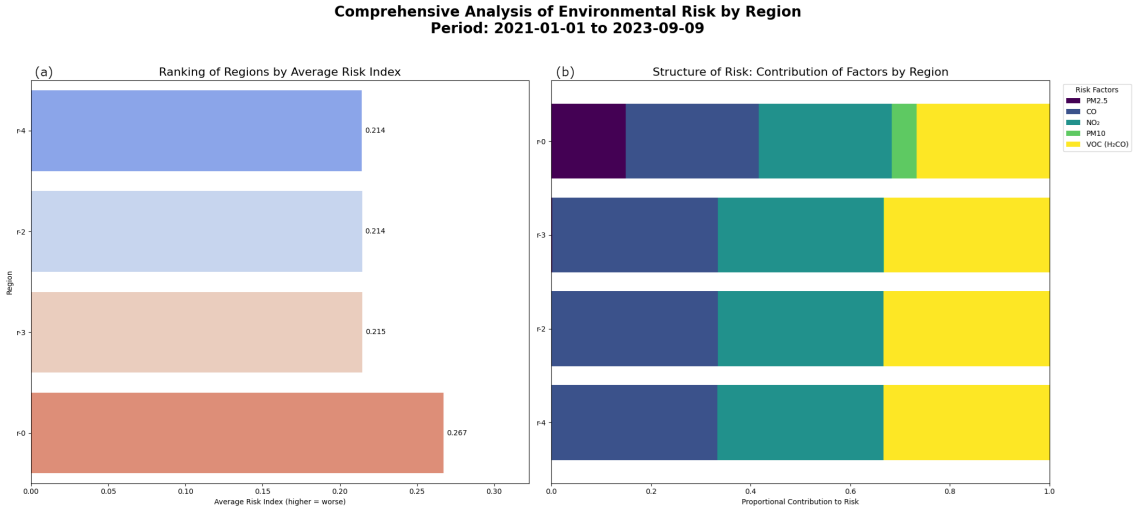

Figure 4: Regional environmental risk analysis for the historical period (2021-01-01 to 2023-09-09): a) ranking of regions by average CHI and b) proportional contribution of each pollutant to regional risk.

3 CONCLUSIONS

This paper presented a novel big data framework for the automated analysis and stratification of environmental risk. In response to the challenge of processing vast and heterogeneous data from modern monitoring networks, our methodology provides a structured, multi-stage pipeline that transforms raw measurements into actionable intelligence.

The core contributions of this work are twofold. First, the development of a CHI offers a method to synthesize complex, multi-pollutant data into a single, quantitative, and rankable score of environmental danger. Second, we demonstrated a guided clustering approach that uses these comprehensive risk profiles to automatically stratify monitoring locations into clear risk tiers, effectively identifying "hotspots." The framework's ability to then pinpoint the key risk drivers for these high-risk zones provides a crucial diagnostic capability.

The scientific novelty of this work is multi-faceted and lies not only in the individual components but, crucially, in their synergistic integration into a single analytical pipeline that bridges the gap from raw data to decision support. Specifically, this novelty is defined by:

1) A methodological shift in data pre-processing, which moves beyond standard cleaning to include a risk-based normalization step. By converting absolute concentrations into a dimensionless metric of multiplicity relative to MAC, the framework transforms the data into an inherently risk-aware format before any analysis is performed.
2) An improved approach to exploratory analysis, where the Composite Hazard Index is not merely an output but a critical input for the clustering algorithm. This "guided clustering" technique forces the stratification to be based not just on the similarity of pollution profiles, but on the overall level of danger, thereby moving beyond descriptive grouping to perform prescriptive risk stratification.

Validation on a real-world dataset confirmed the framework's effectiveness. It successfully isolated a single location of extreme risk, quantified its hazard level relative to other areas, and unambiguously identified particulate matter (PM10 and PM2.5) as the primary cause. This result exemplifies the system's potential to provide clear, evidence-based, and targeted insights for environmental agencies, enabling a shift from reactive responses to proactive, data-driven management. While acknowledging limitations such as the reliance on expert weights, the proposed framework represents a significant step towards building more intelligent, interpretable, and effective decision support systems for environmental security.

ACKNOWLEDGMENTS

The authors would like to thank the creators and developers of the EcoCity air quality monitoring dataset for their support in providing public access to this valuable resource.

REFERENCES

[1] M. N. A. Ramadan, M. A. H. Ali, S. Y. Khoo, M. Alkhedher, and M. Alherbawi, "Real-time IoT-powered AI system for monitoring and forecasting of air pollution in industrial environment," Ecotoxicol. Environ. Saf., vol. 283, p. 116856, 2024. https://doi.org/10.1016/j.ecoenv.2024.116856.

[2] Q. A. Tran, Q. H. Dang, T. Le, H. T. Nguyen, and T. D. Le, "Air quality monitoring and forecasting system using IoT and machine learning techniques," in Proc. 6th Int. Conf. Green Technol. Sustain. Dev. (GTSD), Danang, Vietnam, Jul. 2022. https://doi.org/10.1109/GTSD54989.2022.9988756.

[3] R. Kozłowski, M. Szwed, A. Kozłowska, J. Przybylska, and T. Mach, "Quality management system in air quality measurements for sustainable development," Sustainability, vol. 16, no. 17, Art. no. 7537, 2024. https://doi.org/10.3390/su16177537.

[4] A. Rahman and M. T. Khatun, "Multivariate analysis of urban air pollution: Clustering and patterns across major Asian cities," in Proc. IEEE Int. Conf. Future Mach. Learn. Data Sci. (FMLDS), Nov. 2024. https://doi.org/10.1109/FMLDS63805.2024.00087.

[5] Z. Hu, Z.Liu, J.Tian, Y.Liu, H.Pan, Sh.Liu, B.Yang, L.Yin, and W.Zheng "Classification of urban pollution levels based on clustering and spatial statistics," Atmosphere, vol. 13, no. 3, p. 494, 2022. https://doi.org/10.3390/atmos13030494.

[6] C. Bouveyron, J. Jacques, A. Schmutz, F. Simões, and S. Bottini, "Co-clustering of multivariate functional data for the analysis of air pollution in the South of France," Ann. Appl. Stat., vol. 15, no. 4, 2020.: https://doi.org/10.1214/21-AOAS1547

[7] H. Jorquera and A. M. Villalobos, "Combining cluster analysis of air pollution and meteorological data with receptor model results for ambient PM2.5 and PM10," Int. J. Environ. Res. Public Health, vol. 17, no. 22, p. 8455, 2020. [Online]. Available: https://doi.org/10.3390/ijerph17228455

[8] I. Savchenko, A. Shapoval, and I. Kuziev, "Modeling of high module power sources systems safety processes," Mater. Sci. Forum, vol. 1052, pp. 399–404, 2022. https://doi.org/10.4028/p-24y9ae

[9] K. Vadurin, A. Perekrest, V. Bakharev, V. Shendryk, Y. Parfenenko, and S. Shendryk, "Towards Digitalization for Air Pollution Detection: Forecasting Information System of the Environmental Monitoring," Sustainability, vol. 17, no. 9, Art. no. 3760, 2025. [Online]. Available: https://doi.org/10.3390/su17093760.

[10] K. Akshara, K. Shamita, G. Prashant, K. Neelesh, N. Dev. (2022). "SmartAirQ: A Big Data Governance Framework for Urban Air Quality Management in Smart Cities," Frontiers in Environmental Science, 10, Article no. 7537. https://doi.org/10.3389/fenvs.2022.785129.

[11] V. Mokin, "Air Quality Monitoring from EcoCity," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/datasets/vbmokin/air-quality-monitoring-from-ecocity

[12] Z. Allam and Z. A. Dhunny, "On Big Data, Artificial Intelligence and Smart Cities," Cities, vol. 89, pp. 80–91, 2019. https://doi.org/10.1016/j.cities.2019.01.032

[13] O. Alvear, C. T. Calafate, J.-C. Cano, and P. Manzoni, "Crowdsensing in Smart Cities: Overview, Platforms, and Environment Sensing Issues," Sensors 18, no. 2, p. 460, 2018. https://doi.org/10.3390/s18020460

[14] M. Asgari, M. Farnaghi, and Z. Ghaemi, "Predictive Mapping of Urban Air Pollution Using Apache Spark on a Hadoop Cluster," in Proceedings of the 2017 International Conference on Cloud and Big Data Computing (London, UK: Association for Computing Machinery), pp. 89–93, 2017. https://doi.org/10.1145/3141128.3141131