International Journal on Engineering Technology and Infrastructure Development (InJET-InDev)
www.lec.edu.np/journal
Volume 2, Issue no. 2, Page 8-17
Received Date: 10/31/2025
Revised Date: 5/11/2026
Accepted Date: 5/16/2026
ISSN: 3021-999X (Print)
https://doi.org/10.3126/injet-indev.v2i2.95693

K-Means Clustering and Prophet-Based AQI Pattern Mining and Short-Term Forecasting for Asian Countries (2022–2025)

Ankit Neupane[1*], Ashlim Tamang[2], Rajad Shakya[3]

[1]*Department of Electronics and Computer Engineering, IOE Thapathali Campus, Thapathali, Nepal, ankit.net60@gmail.com*
[2]*Department of Electronics and Computer Engineering, IOE Thapathali Campus, Thapathali, Nepal, ashlimtamang1@gmail.com*
[3]*Department of Electronics and Computer Engineering, IOE Thapathali Campus, Thapathali, Nepal, shakyarajad1@gmail.com*

Abstract

This paper uses regional trends in air pollution in Asia as a case to explore the limitation of existing methods of analyzing pollution, which fail to effectively temporal pollution patterns. This is achieved by using a dataset of 19,949 records of weekly pollution in 40 countries in Asia obtained from Kaggle covering from 2022 to July 2025. In this paper, the K-Means clustering approach and Prophet algorithm have been used to analyze air quality data from the countries in Asia. Five clusters of countries with similar normalized AQI values have been identified, and each cluster has had four months of pollution forecasting performed at the cluster level. The results show varying levels of effectiveness of this approach to forecasting pollution with R^2 scores between 0.454 and 0.921 and mean absolute error of between 4.48 and 48.92. Large and stable clusters have relatively higher accuracy than smaller clusters with the smallest clusters (Cluster 4 with one country -India) being highly volatile. From this, it can be concluded that despite temporally sparse data, some valuable insights like monsoonal dip in pollution in July and August can be derived.

Keywords: AQI, K-Means Clustering, Prophet, Air Quality Forecasting, Time Series Analysis

1. Introduction

Air pollution is among the largest public health and environmental issues globally. Air pollution has increased rapidly in the majority of Asian countries as a result of rapid urbanization, industrial expansion, and the increase in motor vehicle emissions, worsening the quality of air dramatically. The Air Quality Index (AQI) is a standardized index to measure pollution and evaluate the impact of various air pollutants on human health. However, AQI trends exhibit spatiotemporal variability and therefore short-term analysis is insufficient to solely rely upon.

One of the biggest difficulties in such analyses is the spotty availability of long-term AQI information for a number of countries. With improved data sets now available, it is possible to apply advanced data mining and time series techniques to identify actionable intelligence. In this context, clustering algorithms may be helpful for nations with agglomerative pollution patterns, while forecasting models give room for making future air quality trend predictions.

This study analyzes AQI trends across Asian countries using K-Means clustering and the Prophet forecasting model. K-Means groups nations by their normalized AQI time series to identify shared pollution trends, and Prophet generates short-term AQI forecasts for each group. The system shows how even temporally sparse data can yield reasonable insights, facilitating data-informed policymaking and regional public health responses.

2. Literature Review

Maltare and Vahora (2023) also proposed an AI-based air quality prediction system for India, comparing SARIMA, SVM, and LSTM models to predict pollutant concentrations in Ahmedabad. Their study emphasizes the importance of utilizing both statistical and deep learning methods toward accurate, location-specific air quality prediction. Their findings showed great predictive performance, especially from the LSTM model, which demonstrated rapid convergence and low error rates. The use of freely available environmental data and the focus

* *Corresponding author*

on Indian cities is highly relevant to Asian regional forecasting needs, supporting the adoption of flexible, data-driven forecasting models across Asia.

Bishoi, Prakash, and Jain (2009) compared Air Quality Index (AQI) methodologies in an urban environment in India and introduced a new AQI using factor analysis (NAQI). Their research reveals shortcomings in the standard US-EPA AQI such as the lack of inclusion of synergistic pollutant effects and shows that NAQI does not differ from EPA methods in the long run but offers a more nuanced season- and location-specific ranking. These improvements in methodology uncover the necessity for authentic index development in cross-seasonal and cross-national comparisons of air quality patterns (Bishoi et al., 2009).

Tsai and Lin (2021) conducted a comprehensive trend analysis of the Air Quality Index (AQI) and greenhouse gas (GHG) emissions in Taiwan from official statistics over a period of years. They pointed out a reduction in the proportion of days with bad air quality (AQI > 100) from 18.1 % in 2017 to 10.1 % in 2020, which is primarily attributed to changes induced by COVID 19 lockdowns and enhanced pollution control measures. Their research also connects AQI trends with energy-related GHG emissions and shows the effects of regulatory policy and sustainable development initiatives on emission levels and air quality (Tsai & Lin, 2021).

3. Theoretical Background

3.1. Prophet Time Series Model

Prophet is a time series forecasting model developed by Facebook designed to handle time series data with seasonal effects and historical trends (Taylor and Letham, 2018). It is particularly robust to missing data, outliers and trend shifts.

Prophet decomposes a time series $y(t)$ into three main components.

$$y(t) = g(t) + s(t) + h(t) \quad \text{(Equation 1)}$$

Where,

$g(t)$ is trend component,

$s(t)$ is seasonal component,

$h(t)$ is holiday component.

3.2. K-Means Clustering

K-Means is an unsupervised machine learning algorithm most commonly used to split data sets into groups, or clusters, based on similarity (MacQueen, 1967). K-Means initializes a set of centroids (k) and assigns each data point to the nearest centroid before updating centroids iteratively to minimize within-cluster variance.

Mathematically, K-Means aims to minimize the sum of squared distances between data points and their respective cluster centroids. This makes the algorithm effective for discovering underlying structures in high-dimensional data where explicit labels are not available.

In the context of Air Quality Index (AQI) analysis, K-Means is particularly useful because AQI data is continuous and influenced by multiple correlated pollutants over time. The algorithm can group regions exhibiting similar pollution behavior patterns, even when such groupings are not explicitly defined. This enables the extraction of meaningful structures from large-scale environmental datasets, supporting comparative analysis across different geographical regions.

K-Means is also computationally efficient and interpretable, making it suitable for exploratory analysis of environmental time-series data where identifying general patterns is more important than precise classification boundaries.

4. Methodology

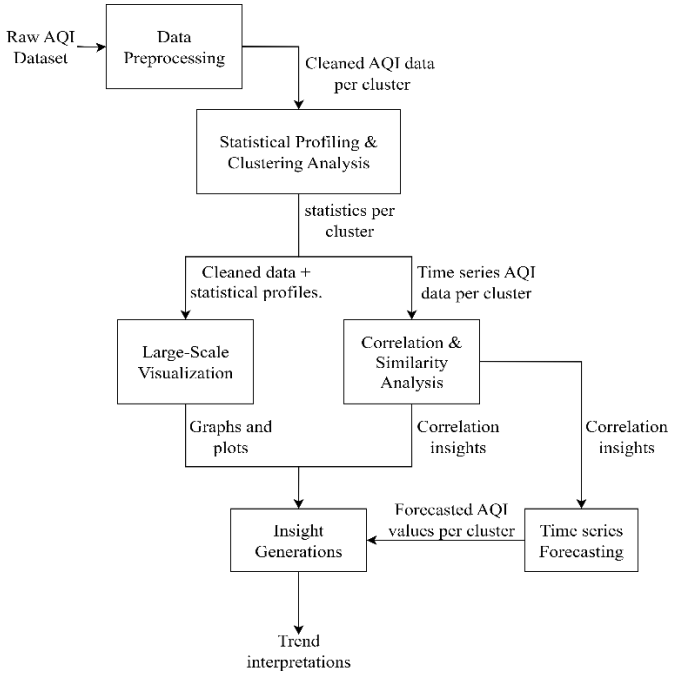


Figure 1. System Block Diagram

4.1 Dataset Description

The dataset for this project is compiled from Kaggle. It consists of four columns:

- Date
- Country
- AQI Value

There is a total of 19949 entries of data. The data is organized in a way that there is a record of average AQI of countries recorded in a weekly basis from July 2022 to July 2025. Each record consists of date, the country name, and the corresponding AQI value. The dataset is sourced from Kaggle and is structured such that each row corresponds to a specific country's AQI value for a given week. No additional features apart from AQI are present in the raw data.

Although the Kaggle dataset is updated daily, the records used in this study were aggregated into weekly averages for consistency and noise reduction.

4.2. Data Preprocessing

The original dataset contained AQI data for countries across the globe. The dataset was filtered to retain only the data corresponding to Asian countries, allowing a focused analysis on regional air quality trends and patterns within Asia. Upon doing so, the data of 40 Asian countries were collected. Some countries lacked data due to certain geopolitical reasons

4.2.1. Normalization

The AQI values were normalized using a Standard Scaler, which adjusted each value x to z as follows.

$$z = \frac{x - \mu}{\sigma} \qquad \text{(Equation 2)}$$

Where,

μ is mean of the data,

σ is the standard deviation of the data,

x is the AQI value of a particular country at a particular week.

The normalized data was utilized for clustering and forecasting models to ensure consistency in scale. However, for visualization purposes the original (unnormalized) AQI values were used.

4.2.2. Interpolation

To address missing values in the AQI time series data, linear interpolation was applied along the time axis. This method estimates a missing value by assuming a straight-line relationship between the known data points immediately before and after the missing entry. For a missing value x_t between two known points x_{t+1} and x_{t-1}, the interpolated value is given as follows.

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} \qquad \text{(Equation 3)}$$

Where,

y is the estimated AQI to be interpolated,

x is the time where data is missing,

x_0, y_0 are the coordinates of the closest known data point immediately preceding the missing gap,

x_0, y_0 are the coordinates of the closest known data point immediately following the missing gap.

This ensures a smooth and gradual transition in the AQI trend, avoiding any sudden jumps. The interpolation was performed independently for each country's AQI time series to preserve local temporal patterns.

5. Results

The number of clusters was set to k = 5. While typical data-driven approaches utilize the Elbow method or Silhouette scores, this study adopts k = 5 to align the resulting clusters with the established qualitative tiers of the AQI, which traditionally categorizes air quality into five levels (i.e., Good, Moderate, Unhealthy, Very Unhealthy, Unhealthy for Sensitive Groups). Although formal methods such as the Elbow Method and Silhouette Score are commonly used, this study prioritizes interpretability and alignment with standard AQI categories for policy relevance. This ensures that the mathematical clusters correspond to meaningful regulatory and public health categories.

The clustering along with the mean of each cluster is given below.

Table 1. Clustering Results

Cluster	Average AQI
Cluster 0	76.48
Cluster 1	117.00
Cluster 2	148.07
Cluster 3	29.95
Cluster 4	219.50

From the above table, we can conclude that the clusters represent the air quality as follows.

Table 2. Separation of Clustering Classes

Cluster	AQI Class
Cluster 0	Moderate
Cluster 1	Unhealthy for Sensitive Groups
Cluster 2	Unhealthy
Cluster 3	Good
Cluster 4	Very Unhealthy

To analyze air quality trends across Asian nations, a line graph of the average AQI values over time for each cluster was created. The graph distinguished between cluster patterns.

Clustering results divided the nations by AQI trends. The number of nations in each cluster are as follows.

Table 3. Countries per cluster

Cluster	Number of countries
Cluster 0	27
Cluster 1	3
Cluster 2	5
Cluster 3	4
Cluster 4	1

This analysis indicates that the majority of Asian countries possess moderate levels of air pollution with comparatively few countries exhibiting AQI above this range, representing localized areas of high pollution.

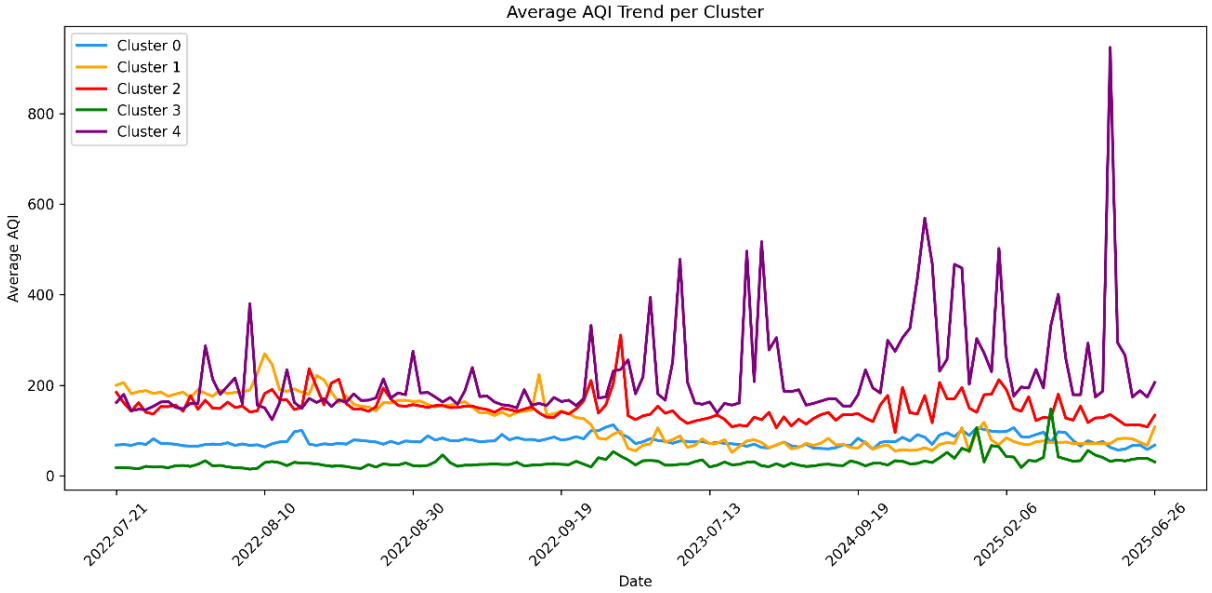


Figure 2. Trend across clusters

In addition, the AQI trend of Nepal was specifically analyzed to analyze its air quality pattern in relation to other Asian countries. Nepal was found to belong to Cluster 0, which consists of countries with moderate air quality. A simple line graph was plotted to illustrate Nepal's AQI variations over time, providing additional insight of its temporal air quality behavior.

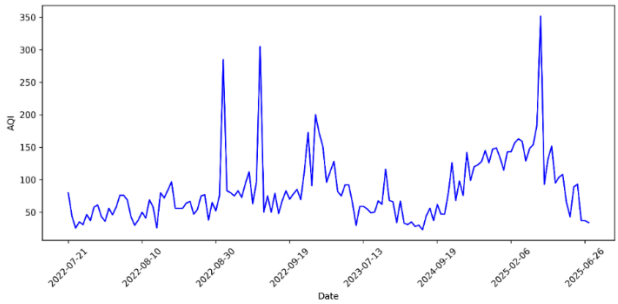
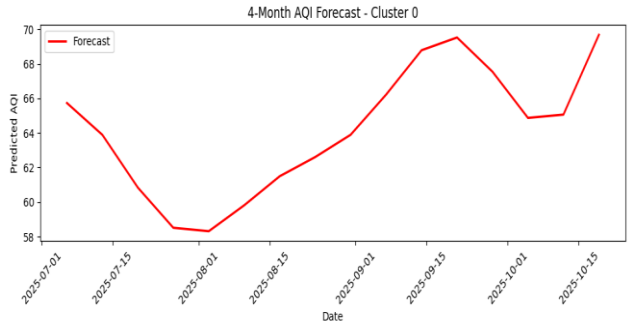


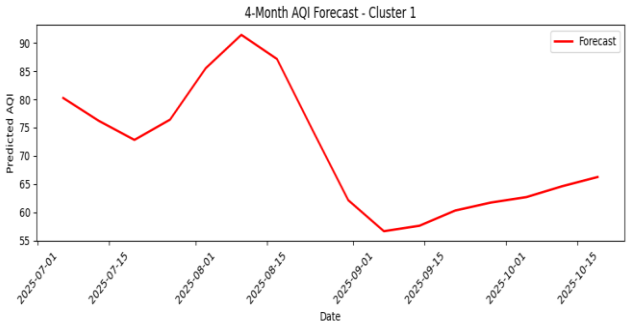
Figure 3. Trend of AQI in Nepal

The Prophet model was configured with “seasonality_mode='additive'”, which is appropriate for AQI data where seasonal fluctuations are roughly constant over time. The “changepoint_prior_scale” was set to the default 0.05 to maintain a balance between flexibility and overfitting. The model was trained to produce a forecast horizon of

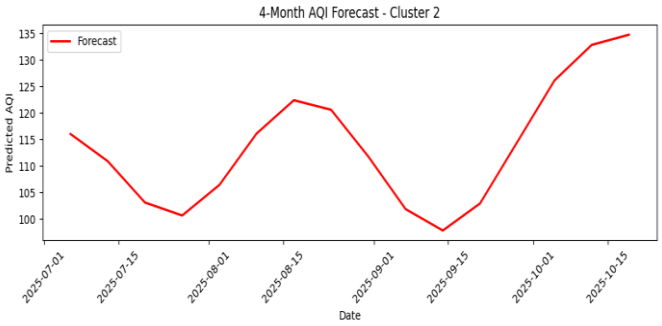
16 weeks (approximately 4 months) to provide actionable short-term insights. Using the Prophet model to forecast AQI for the next four months yielded the following results.



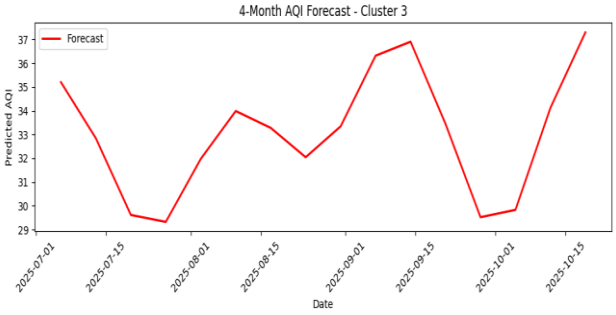
4 (a)



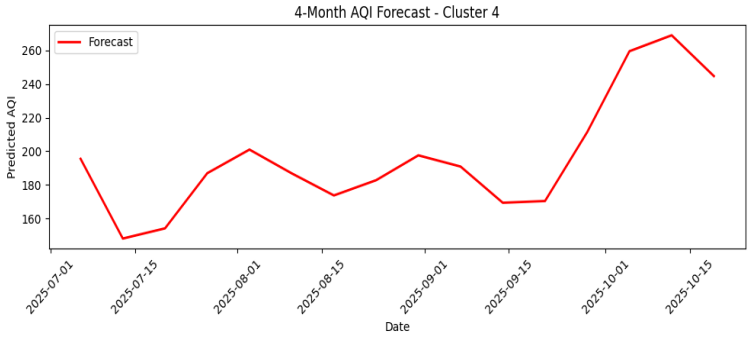
4 (b)



4 (c)



4 (d)



4 (e)

Figure 4. Four-month AQI forecasts generated by the Prophet model for each identified cluster: (a) Cluster 0 (Moderate), (b) Cluster 1 (Unhealthy for Sensitive Groups), (c) Cluster 2 (Unhealthy), (d) Cluster 3 (Good), and (e) Cluster 4 (Very Unhealthy).

5.1. Evaluation Metrics

Table 4. Metrics Table

Cluster	MAE	RMSE	R^2
Cluster 0	4.48	5.94	0.734
Cluster 1	10.31	15.01	0.921
Cluster 2	14.34	20.64	0.48
Cluster 3	5.76	10.11	0.557
Cluster 4	48.92	79.67	0.454

To evaluate the model's predictive performance, the time-series data was divided using a temporal hold-out validation approach. The initial 80% of the weekly records (from July 2022 to approximately October 2024) were

used as the training set to fit the Prophet model, while the remaining 20% were reserved as a test set to calculate the evaluation metrics (MAE, RMSE, and R^2) reported in Table 4.

6. Discussion

The integration of K-Means clustering and the Prophet forecasting model provides a unique lens through which to view the spatiotemporal dynamics of air quality across 40 Asian countries. By grouping nations into five distinct clusters, this study moved beyond individual country analysis to identify broader regional pollution profiles.

6.1. Comparison with Existing Literature

Our findings build upon recent predictive modeling efforts, such as the work of Maltare and Vahora (2023), who demonstrated the high efficacy of deep learning models like LSTM for localized AQI prediction. While LSTM models excel at capturing complex non-linearities in continuous, city-specific datasets, our application of the Prophet model proved more advantageous for forecasting on a multi-national scale. This is primarily due to Prophet's inherent robustness to the missing data and irregular temporal gaps frequently encountered in the aggregated Kaggle dataset. Furthermore, unlike the "black-box" nature of neural networks, Prophet yields highly interpretable seasonal components that directly align with observed regional environmental phenomena, such as the monsoon cycle.

Additionally, our K-Means clustering approach addresses the methodological gaps identified by Bishoi *et al.* (2009), who emphasized the necessity of season- and location-specific air quality rankings over generalized indices. By grouping nations with synchronous pollution behaviors, our framework provides the nuanced, cross-national comparability that traditional methods lack. Finally, consistent with the macro-level trend analyses conducted by Tsai and Lin (2021), the regional patterns extracted from our clusters underscore how broad temporal trends, whether driven by natural climatic cycles or coordinated pollution control measures can be effectively tracked and forecasted to aid cross-border environmental policy.

6.2. The Monsoon Effect and Seasonal Dynamics

A significant finding in this study is the consistent dip in AQI values observed across most clusters (specifically Figures 4(a) and 4(c)) during July and August. This pattern statistically correlates with the South Asian monsoon system. Increased precipitation during these months acts as a natural "washout" mechanism for atmospheric pollutants like PM2.5 and PM10, leading to the observed seasonal improvement in air quality. This suggests that regional environmental policies must account for these natural fluctuations when setting short-term pollution reduction targets.

Pearson correlation coefficients were computed between the average AQI time series of each cluster to analyze temporal synchronization patterns.

The correlation analysis between clusters (Figure 5) provides quantitative evidence of regional synchronization in air quality trends. High positive correlation coefficients between Clusters 0, 1, and 2 indicate that these groups, despite having different average AQI levels, experience nearly identical temporal fluctuations.

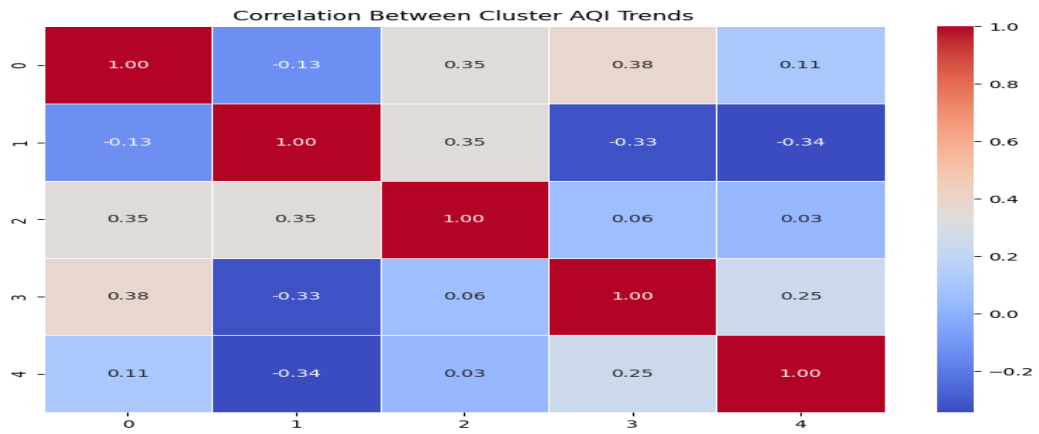


Figure 5. Correlation between clusters

This high degree of synchronicity is primarily driven by the South Asian monsoon system. The "monsoon-related dip" in July and August is not an isolated event but a regional phenomenon that "washes out" pollutants across multiple clusters simultaneously. The correlation matrix validates this; the shared downward trend during the rainy season creates the high correlation values observed. This suggests that the atmospheric "cleansing" effect of the monsoon is a dominant driver of air quality across the continent, over-riding local emission variations during those months.

Conversely, lower correlation coefficients between Cluster 4, which only contains India and the other groups highlight a divergence in pollution dynamics. While the broader region follows a synchronized seasonal cycle, Cluster 4 exhibits unique temporal peaks that do not align with the rest of Asia. This divergence is likely due to intense localized events, such as post-harvest crop residue burning in Northern India or specific industrial cycles that occur outside the standard regional peaks. This statistical divergence reinforces the decision to treat these regions as separate clusters; they are not only "dirtier" on average, but their pollution rises and falls at different times than their neighbors.

The high correlation between most clusters suggests that regional collaborative policy is essential. Because these clusters are synchronized, an environmental policy success in one region (e.g., reducing industrial emissions) is likely to have a 'multiplier effect' across synchronized clusters due to shared transboundary air masses. Conversely, the divergent clusters (like Cluster 4) require independent, targeted interventions that focus on their specific local emission schedules rather than general regional trends.

6.3. Correlation between Cluster Size and Predictive Accuracy

A key observation of this study is the positive correlation between the number of countries within a cluster and the resulting model accuracy. Clusters 0, 1, and 2 which aggregate data from multiple nations exhibited the highest R^2 values and lowest relative errors. This suggests that at a regional scale, the aggregation of multiple time series acts as a "natural filter," smoothing out localized noise and idiosyncratic pollution events (such as specific local festivals or short-term industrial disruptions) that might otherwise skew a single-country model.

In contrast, the performance significantly degrades in smaller or single-nation clusters. Cluster 4, which consists solely of India, yielded an R^2 of 0.454, the lowest in the study. This "small cluster penalty" occurs because the Prophet model, when applied to a single high-variance nation, cannot benefit from the shared regional trends that stabilize the larger groups. For nations like India, which experience extreme seasonal volatility due to agricultural burning and high-density industrial activity, a generalized cluster-level approach may be insufficient. These findings imply that while regional clustering is highly effective for broad policy-making across similar nations, "outlier" countries with high pollution volatility require individual, hyper-localized modeling parameters rather than being grouped into a general framework.

Accordingly, the evaluation metrics must be qualified in terms of cluster size: more general and accurate prediction results occur from larger clusters, but noise and reduced forecasting accuracy can plague smaller clusters. The insight underscores the importance of cluster composition when interpreting model performance findings in multi-country air quality research.

6.4. Policy Implications

The clustering results provide actionable insights for regional environmental governance. Nations within the same cluster (e.g., Cluster 3 "Good" vs. Cluster 2 "Unhealthy") can utilize these shared profiles to develop collaborative cross-border air quality management strategies. For instance, countries in "Unhealthy" clusters may benefit from synchronized regulations on agricultural burning or industrial emissions during high-risk seasons identified by the Prophet forecasts.

6.5. Limitations and Representativeness

Despite the framework's capability to extract insights from limited data, several limitations persist. The study lacks comprehensive meteorological variables, such as wind speed and humidity, which significantly influence AQI. Furthermore, the absence of data from specific nations (e.g. North Korea and Myanmar) due to geopolitical constraints may affect the absolute representativeness of the identified clusters. Future iterations of this work

should aim to integrate multi-source environmental data to improve the robustness of the forecasting in highly volatile regions.

7. Conclusion and Future Work

This study demonstrates the efficacy of combining K-Means clustering with Prophet time-series forecasting to analyze and predict air quality dynamics across 40 Asian countries. Rather than relying solely on individual national models, grouping these nations into five distinct clusters based on normalized AQI behavior revealed significant regional synchronization, most notably a consistent monsoon-related dip in pollution during July and August.

The quantitative evaluation indicates that predictive accuracy is closely tied to cluster size and stability. Larger, multi-nation clusters achieved robust forecasting accuracy, with R^2 values reaching up to 0.921, as localized noise was smoothed out across the group. Conversely, smaller, highly volatile clusters such as the single-nation cluster representing India exhibited lower predictability (R^2 = 0.454, MAE = 48.92). These findings carry important policy implications: while regional clustering provides a reliable framework for synchronized, cross-border environmental strategies, nations with extreme pollution volatility require hyper-localized, independent modeling. Ultimately, the proposed framework successfully extracts actionable short-term forecasts from temporally sparse data, providing a foundation for data-informed public health interventions.

7.1. Limitations and Future Enhancements

While the current framework is effective, it is limited by its reliance on historical AQI data alone, which hinders model accuracy in highly volatile clusters and may miss sudden, structural changes in emissions. Future research should focus on expanding the dataset to span longer time series and integrating exogenous environmental variables such as meteorological data (wind speed, precipitation) and emission source inventories to enhance prediction robustness. Additionally, implementing hybrid deep learning methodologies (e.g., combining Prophet with LSTM) could improve performance for outlier clusters. Finally, the integration of real-time data streams and dynamic clustering algorithms would enable adaptive tracking of evolving air quality trends.

7.2. Dataset Availability Statement

The dataset analyzed during the current study, "AQI - Air Quality Index," is a third-party dataset publicly available in the Kaggle repository. It can be accessed at: https://www.kaggle.com/datasets/azminetoushikwasi/aqi-air-quality-index-scheduled-daily-update. The date of access of this dataset for this study was July 06, 2025.

This research relies exclusively on publicly available, aggregate environmental data. The study does not involve human participants, animal subjects, or personally identifiable information; therefore, no formal ethical approval or participant consent was required for this manuscript.

Acknowledgements

The authors would like to express their gratitude to the Department of Electronics and Computer Engineering at IOE Thapathali Campus for providing the academic environment and resources necessary to conduct this research. We also thank the dataset creators on Kaggle for making the historical AQI records publicly accessible for analysis.

References

Azminetoushikwasi, 2024. *AQI - Air Quality Index*. [dataset] Kaggle. Available at: https://www.kaggle.com/datasets/azminetoushikwasi/aqi-air-quality-index-scheduled-daily-update?resource=download [Accessed 6 July 2025].

Bishoi, B., Prakash, A. and Jain, V.K., 2009. A comparative study of air quality index based on factor analysis and US-EPA methods for an urban environment. *Aerosol and Air Quality Research*, 9(1), pp. 1-17. Available at: https://doi.org/10.4209/aaqr.2008.02.0007.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley: University of California Press, pp. 281-297.

Maltare, N. N., and Vahora, S., 2023. Air Quality Index prediction using machine learning for Ahmedabad city. *Digital Chemical Engineering*, 7, p.100093. Available at: https://doi.org/10.1016/j.diche.2023.100093.

Taylor, S.J. and Letham, B., 2018. Forecasting at scale. *The American Statistician*, 72(1), pp. 37-45. Available at: https://doi.org/10.1080/00031305.2017.1380080.

Tsai, W.-T. and Lin, Y.-Q., 2021. Trend analysis of Air Quality Index (AQI) and greenhouse gas emissions in Taiwan and their regulatory countermeasures. *Environments*, 8(4), p. 29. Available at: https://doi.org/10.3390/environments8040029.