Tariere Timitimi

Professor Christina Schweikert

CUS 610 Project Paper

20 December 2024

Spatial Clustering of Air Pollution in New York City: A Comparative Analysis of K-Means and Hierarchical Clustering

Abstract

This paper explores the spatial clustering of PM 2.5 and NO2 levels in New York City (NYC) using K-means and hierarchical clustering algorithms. By analyzing annual averages of pollutant data, it identifies distinct clusters, evaluates areas of pollutant dominance, and compares the interpretability and informativeness of clustering methods. The findings reveal three primary clusters and highlights areas with disproportionately high pollutant concentrations. This paper provides insights into pollution management and the efficacy of clustering techniques in environmental studies.

Air pollution remains a pressing concern, particularly in urban environments like NYC, where PM 2.5 and NO2 significantly impact public health. Prior studies have emphasized the importance of spatial and temporal analyses to understand pollutant distributions. For this project, I examine the clustering of air pollution data to identify patterns and inform policy interventions. I selected NO2 and PM2.5 as my measures primarily because these pollutants are key indicators of urban air quality; NO2 is primarily associated with traffic emissions and industrial activities, whilePM2.5 comes from activities such as burning fuel and chemical reactions in the air, which helps us understand where pollution comes from and how it spreads."

Related Work

Numerous studies have utilized clustering and spatial analyses to study air pollution. For instance, Jerrett et al. (2005) conducted a spatial analysis of PM 2.5 in Los Angeles, revealing intra-urban variations in mortality risk due to pollution gradients (Jerrett, 2005). Similarly, Kibria et al. (2002) applied Bayesian spatial prediction methods to map PM 2.5 exposure, addressing gaps in monitoring network coverage (Kibria, 2002). In NYC, Shukla et al. (2022) developed the ZAPPA tool for ZIP code-level air pollution analysis, enabling policy evaluation at localized scales (Shukla, 2022). These articles highlight the utility of spatial clustering and predictive modeling in urban air quality research. My approach builds on these methods by comparing K-means and hierarchical clustering to assess their applicability in identifying pollution patterns and dominance.

Data

The dataset used for this project contains information from New York City air quality surveillance data, sourced from the Data.gov Air Quality dataset (https://catalog.data.gov/dataset/air-quality). This dataset provides a comprehensive view of air quality indicators across NYC neighborhoods, including annual averages of PM 2.5 and NO2 levels. Air pollution is a critical environmental threat in urban areas, contributing to respiratory and cardiovascular diseases, cancers, and premature deaths. The data reflects variations in pollutant emissions, exposure levels, and population vulnerability across NYC's neighborhoods, offering valuable insights for public health and environmental policies. More details on this dataset can be explored online at the Environment and Health Data Portal (http://nyc.gov/health/environmentdata). Preprocessing steps included filtering data for annual averages, removing irrelevant attributes, and standardizing features. A new feature, "Pollution Trend," was created to capture year-over-year percentage changes in pollutant levels, providing a more detailed perspective on pollution patterns.

Methodology

Preprocessing

Data cleaning involved removing rows with missing values and irrelevant attributes (e.g., 'Message'). Additionally, standardization was applied using the StandardScaler to ensure comparability across features.

Clustering Algorithms

I applied K-means and hierarchical clustering to scaled data comprising "Data Value" and "Pollution Trend" and optimal cluster numbers were determined using the elbow method for K-means and dendrogram analysis for hierarchical clustering (3).

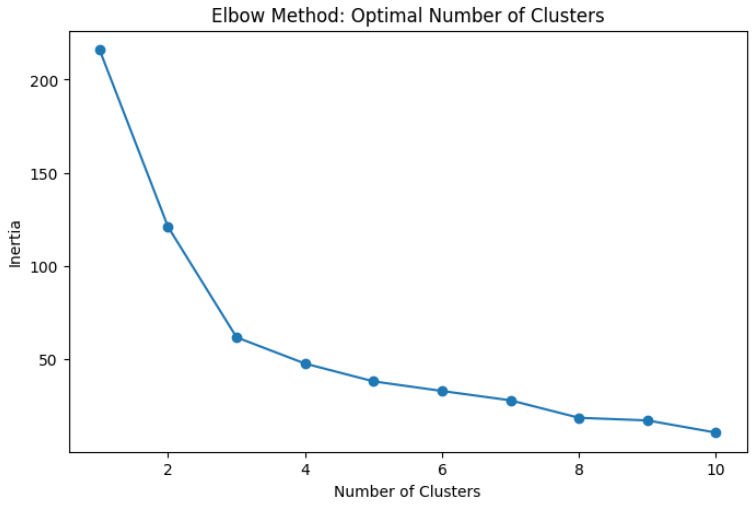

Figure 1: Elbow Method for Determining Optimal Clusters in K-Means Clustering

Evaluation Metrics

The clusters were then evaluated based on intra-cluster variance, inter-cluster separability, and interpretability. Visualizations included scatter plots and dendrograms.

Results

Both algorithms identified three distinct clusters. K-means produced well-separated, compact clusters, whereas hierarchical clustering highlighted nested relationships.

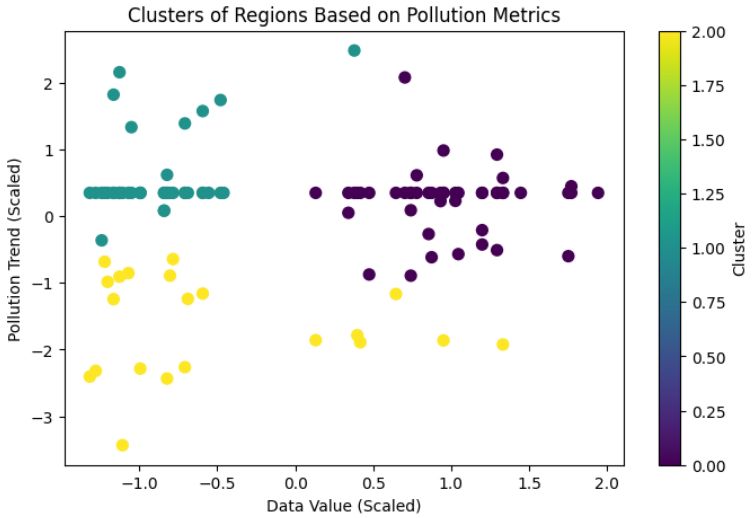

Figure 2 : K-Means Clustering of Regions Based on PM 2.5 and NO2 Levels

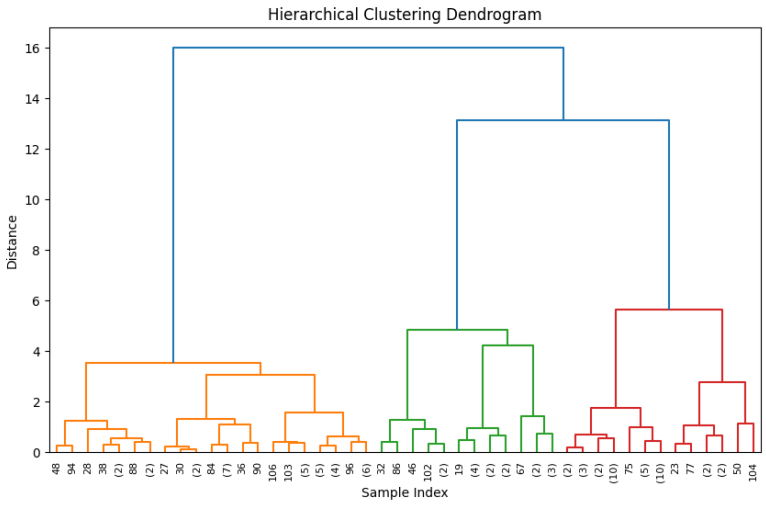

Figure 3: Dendrogram for Hierarchical Clustering of Regions

Analysis also revealed regions where NO2 concentrations exceeded PM 2.5 levels, particularly in traffic-dense areas. Conversely, residential zones exhibited higher PM 2.5 dominance. Regions with extreme pollution characteristics include Bensonhurst - Bay Ridge, which demonstrated a high PM2.5 level, which is likely influenced by residential heating and proximity to traffic

corridors. Southeast Queens also exhibits elevated NO2 levels, possibly due to traffic congestion and nearby industrial zones. Additionally, Midtown Manhattan shows higher concentrations of NO2, likely due to dense urban infrastructure and vehicle emissions.

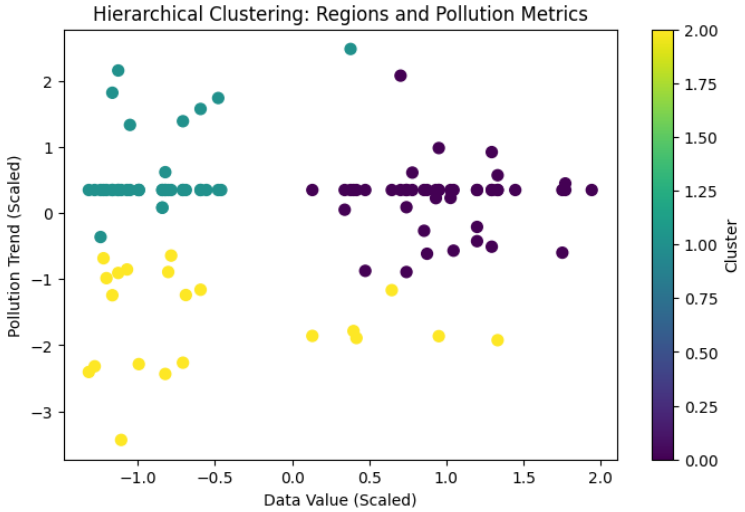

Figure 4: Hierarchical Clustering of Regions Based on Pollution Metrics

K-means clustering proved more interpretable for practical applications, offering clear centroids for pollution levels. Hierarchical clustering provided richer insights into the hierarchical structure of relationships but required more effort for interpretation. These findings underscore the importance of clustering techniques in environmental studies. K-means' simplicity and clarity makes it suitable for policy-oriented tasks, while hierarchical clustering's granularity aids exploratory research. Future studies could integrate additional pollutants and employ ensemble clustering approaches to enhance robustness.

This project demonstrates the utility of clustering techniques in analyzing air pollution data. By comparing K-means and hierarchical clustering, distinct spatial patterns and pollutant dominance areas were identified, providing insights for urban pollution management. Both methods offer unique strengths, emphasizing the value of methodological pluralism in environmental data analysis.

References

1. Shukla, K., Seppanen, C., Naess, B., et al. (2022). ZIP Code-Level Estimation of Air Quality and Health Risk Due to Particulate Matter Pollution in New York City. *Environmental Science & Technology*. (Shukla, 2022)

2. Jerrett, M., Burnett, R. T., Ma, R., et al. (2005). Spatial Analysis of Air Pollution and Mortality in Los Angeles. *Epidemiology*. (Jerrett, 2005)

3. Kibria, B. M. G., Sun, L., Zidek, J. V., & Le, N. D. (2002). Bayesian Spatial Prediction of Random Space-Time Fields with Application to Mapping PM2.5 Exposure. *Journal of the American Statistical Association*. (Kibria, 2002)

4. Abbasi, M. T., Alesheikh, A. A., & Jafari, A. (2024). Spatial and Temporal Patterns of Urban Air Pollution in Tehran. *Scientific Reports*. (Abbasi, 2024)

5. Todd, J. J. (2005). Urban Air Quality. *Environment Design Guide*. (Todd, 2005)

6. Hassan, N. A., Hashim, Z., & Hashim, J. H. (2016). Impact of Climate Change on Air Quality and Public Health in Urban Areas. *Asia Pacific Journal of Public Health*. (Hassan, 2016)

7. Khan, S., Bajwa, S., & Brahmbhatt, D. (2021). Multi-Level Socioenvironmental Contributors to Childhood Asthma in New York City. *Journal of Urban Health*. (Khan, 2021)

8. Scientists Walk NYC Neighborhoods to Map Air Quality Differences. *Spheres*. (2024). (Spheres, 2004)

9. Air Quality Dataset. (n.d.). Retrieved from https://catalog.data.gov/dataset/air-quality.