A Data Mining Approach to Air Pollution Analysis and Sustainable Policy Development

December 12, 2023

Akash Vishwakarma
University of Southern California
vishwaka@usc.edu

Himanshu Borole
University of Southern California
hborole@usc.edu

Shantanu Jhaveri
University of Southern California
sj06434@usc.edu

Vaibhav Rungta
University of Southern California
vrungta@usc.edu

Bistra Dilkina, PhD
Professor
Director, USC Center for AI in Society
University of Southern California
dilkina@usc.edu

Abstract

The unprecedented scale of environmental degradation challenges necessitates a refined understanding of air pollutant distribution and their diverse impact on health, ecosystems, and climate. This study presents a comprehensive examination of various air pollutants, including PM2.5, PM10, CO, NOx, NMVOCs, CH4, NH3, SO2, BC, and OC. We employ k-means clustering on a robust dataset, that encompasses gridwise data on emissions, to classify regions based on the trajectory of emission profiles. The research reveals complex patterns, offering nuanced insights for policymakers.

1 Introduction

1.1 Motivation

1.1.1 Air Pollution and Sustainable Development

Air pollution's widespread impact poses a multifaceted challenge affecting ecosystem health, climatic patterns, and human well-being. Driven by the objectives outlined in the Sustainable Development Goals—particularly SDG 11, which focuses on creating sustainable cities and communities—our project endeavors to identify regions with atypical pollutant concentrations. Our work also supports the ambitions of SDG 13: Climate Action, by highlighting necessary interventions in identified areas.

1.2 Problem Statement

Understanding the sudden change across regions, both negative and positive, of a broad spectrum of air pollutants becomes increasingly critical to realize the inequalities in distributions, especially in developing countries going through rapid unregulated industrialization. This study through the use of data mining techniques seeks to identifying the unique trajectories that various regions have taken in terms of emissions profile forming a foundation for targeted intervention strategies that prioritize areas with critical levels of pollution.

2 Background and Related Work

The link between air pollution, where it happens, and the effects it has on health has gotten a lot of attention from researchers. In 2020, Smith and Rodriguez (2020) [1] analyzed how growing cities lead to more cars on the road, which then leads to more harmful gases like nitrogen oxides and carbon monoxide in the air. In 2013, Hoffman et al. (2013) [2] came up with a way to choose where to monitor air quality in Alaska, making sure the locations picked would provide a comprehensive picture of the state's air quality.

Li and Huang [3], in 2022, studied how farming activities send methane and ammonia into the air, affecting its quality. Adding to these focused studies, Zhang and colleagues [4] in 2023 combined information on different pollutants to understand better how air pollution affects the environment and people's health, showing that solving air pollution problems needs a well-rounded approach.

3 Data

3.1 Description

Our dataset features emission grid maps quantified as tons of substance per 0.1-degree square per year in text files, offering a detailed resolution for our analysis. For our purposes we are aggregating the data by decade. We reduced the values

for emissions for a specific grid on the map to its mean and standard deviation for each pollutant over 10 years, from 1970-79, 1980-89, 1990-99, 2000-09 and 2010-18. This allows us to chart the trajectory of a grid-region over the decades.

3.2 Data Sources

The core of our analytical framework is the Emissions Database for Global Atmospheric Research (EDGAR) version 6.1, which offers a comprehensive inventory of global emissions, encompassing both greenhouse gasses and various air pollutants.

3.3 Combination of Pollutant Data

We combined the aggregated data for individual gases into a master data frame, while maintaining spatial and temporal integrity. The dataset assembled is quite extensive, encompassing more than 25 million individual data entries. These data points are detailed, each described by 18 distinct features. These features encapsulate the average and the variability, measured by the standard deviation, of nine separate pollutants over the specific decade.

3.4 Application of Land Mask in Spatial Analysis

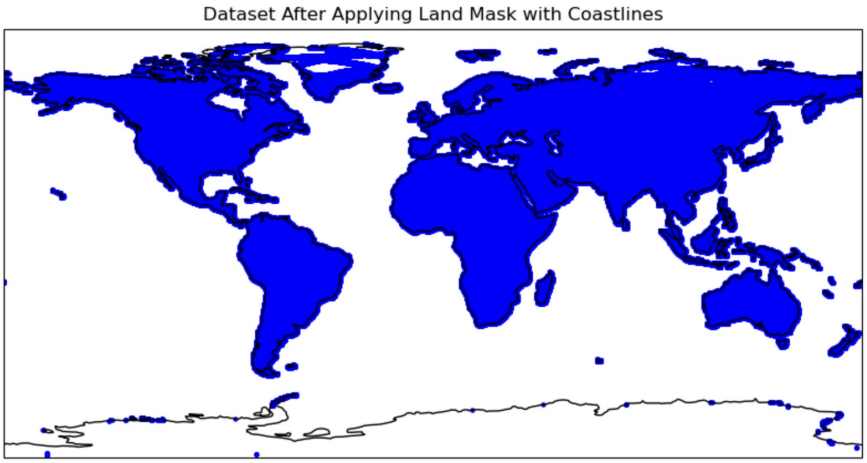


Figure 1: Application of Land Mask

Consequently, in our spatial analysis, a high-resolution land mask was employed to ensure the accuracy of our data, particularly along coastlines. This mask, with a 110-meter resolution sourced from Natural Earth Data, enables precise

differentiation between land and water. Figure 1 exemplifies the dataset post-land masking, highlighting the successful exclusion of non-land data pivotal to our study.

3.5 Data Transformation and Normalization

A visual analysis revealed a positive skew for all features. Therefore, the final step in our preprocessing was the transformation and normalization of the dataset. Given the right-skewed distribution of our features, we applied a logarithmic transformation to each to mitigate the skewness. This was followed by a Z-score normalization to standardize the data, crucial for the k-means algorithm, which relies heavily on the Euclidean distance and is sensitive to the scale of the data.

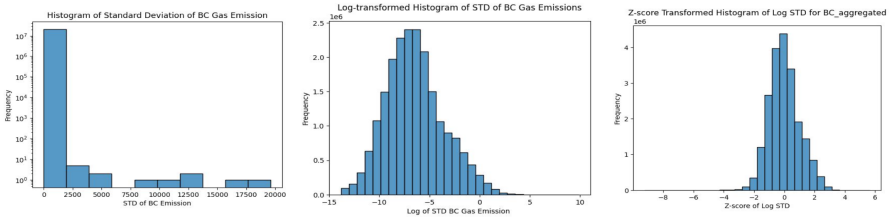


Figure 2: Log Transformation and Z-Score Normalization

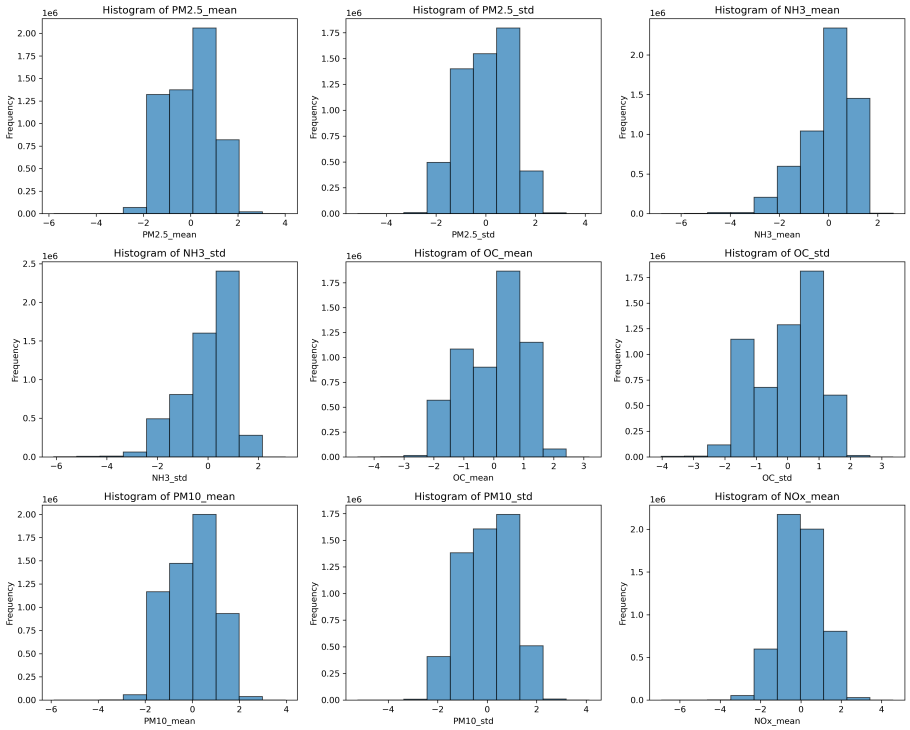


Figure 3: Features after Transformation

4 Task and Methodologies

4.1 Elbow Method

The Elbow Method is an essential part of the k-means clustering analysis process, as it helps in determining the optimal number of clusters by which to group the data. The processed dataset was clustered using the K-means algorithm with a range of k values from 2 to 14. Consequently, the elbow method was used to analyze the clustering.

The Elbow Method graph provided suggests that the optimal number of clusters lies at the point where the elbow bends. This point represents a balance between minimizing the within-cluster variance and avoiding overfitting by having too many clusters. From the graph in Figure 4, it appears that the rate of decrease in the sum of squared distances slows down significantly after k=3 or k=4. This implies that increasing the number of clusters beyond this point does not provide substantial gains in the compactness of the clusters.

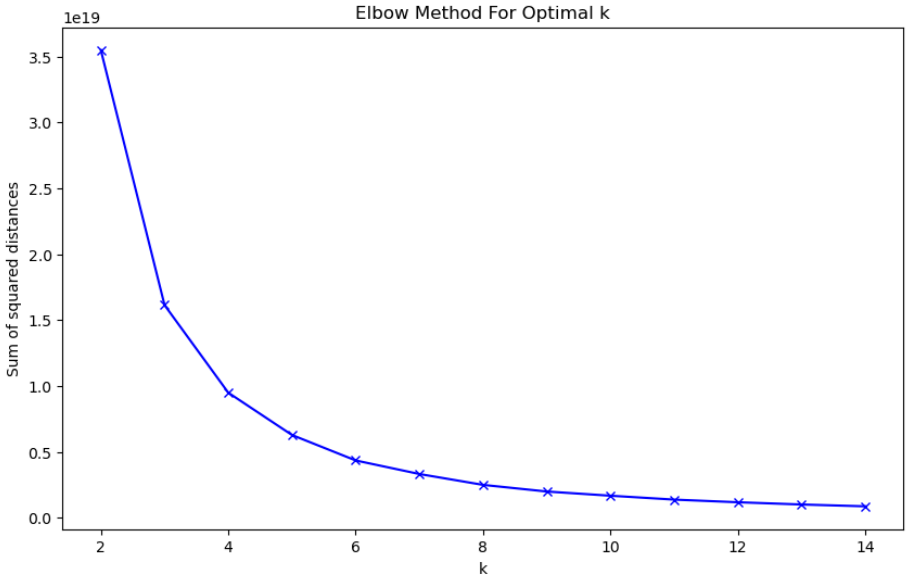


Figure 4: Elbow curve k = 2 to 14

4.2 Silhouette score

The Silhouette Scores for the project were calculated across a range of k values from 2 to 14 using a sample of the dataset. The scores were highest at k=2 and trended downward with increasing k, indicating more cohesive clusters at lower k values. Yet, the optimal number of clusters isn't necessarily where the score is highest; we must also consider the meaningfulness of the clusters. The scores began to level off after k=4, suggesting that additional clusters don't significantly enhance the distinctness of the groups.

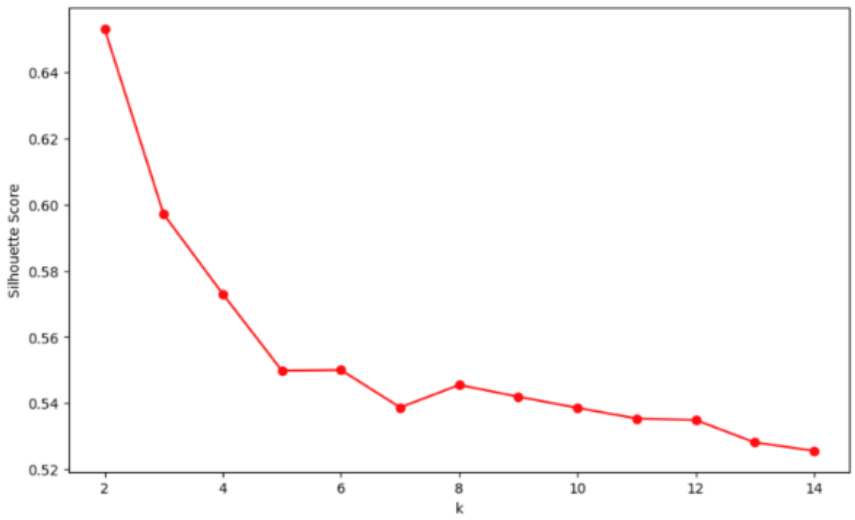


Figure 5: Silhouette Curve k = 2 to 14

By integrating the Elbow Method with Silhouette Scores, k=3 or k=4 emerges as the optimal cluster number striking a balance between cluster clarity and separation.

5 Results

5.1 Mean Heat Map Analysis

The heat maps for k=3 and k=4 clusters provide an insightful visualization of the mean pollutant values within each cluster, revealing distinct profiles of air quality characteristics. The distinction between the clusters can be attributed to the overall level of pollution with CO being the most dominant pollutant. Therefore, the clusters have been relabelled such that the label with lower number 0 signifies a lower pollution level than 2 or 3 which signify the highest level of pollution. Furthermore, the color gradients across the clusters signify the varying levels of pollutants, such as PM2.5, NOx, and SO2, which are crucial indicators of air pollution sources and types.

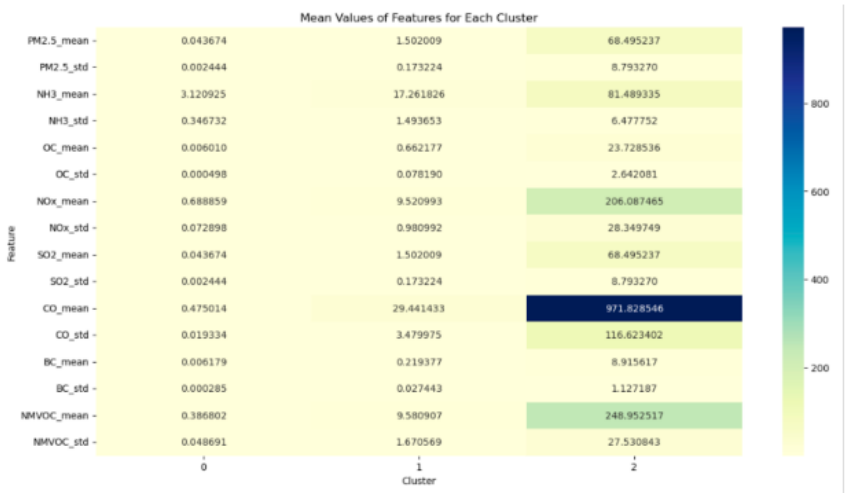


Figure 6: Mean value Heat Map for k=3

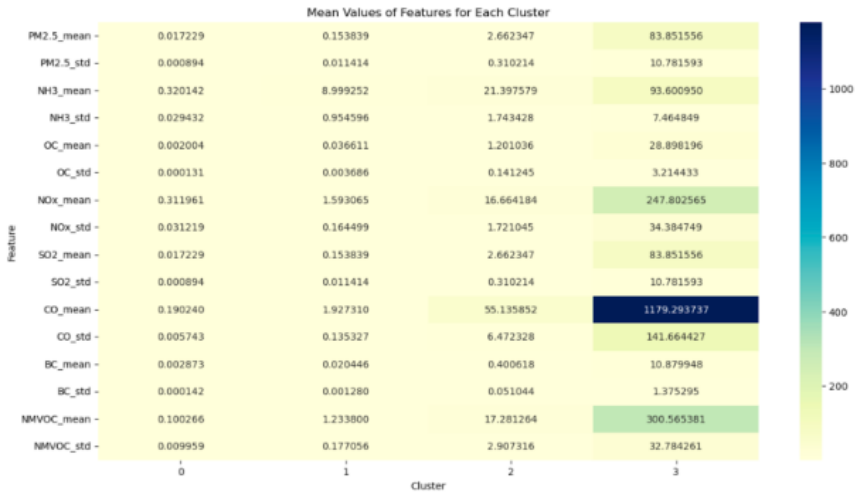


Figure 7: Mean value Heat Map for k=4

5.2 Trajectory

The trajectory analysis for clusters when k=3 and k=4 provides a temporal perspective on the distribution of air pollution patterns over recent decades. The line graphs chart the prevalence of each cluster as a percentage of the total occurrences within the dataset for each decade, revealing trends and shifts in

the clustering of pollution data over time.

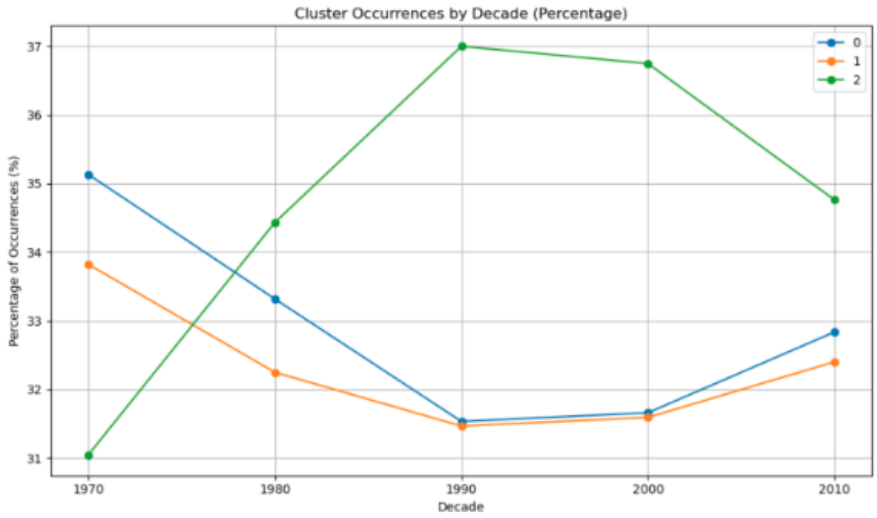


Figure 8: Cluster Occurrences for k = 3

For k=3, the graph shows notable shifts in cluster occurrences over decades. Cluster 2's growing prevalence and the decrease in the occurrence of Low pollution areas marked by cluster 0 across 1970 to 1990 suggests escalating environmental concerns and increase in pollutants. Consequently, the decrease in Cluster 2 occurrence over the last two decades also suggest that this period was followed by necessary corrective action to some degree.

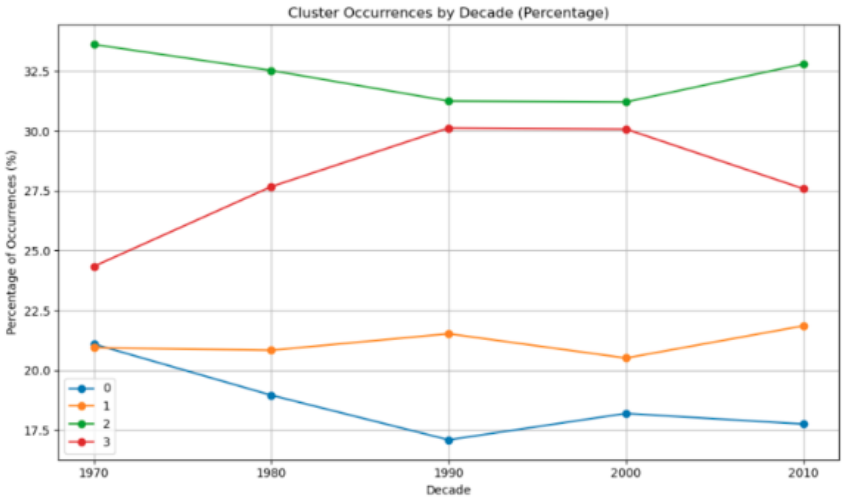


Figure 9: Cluster Occurrences for k = 4

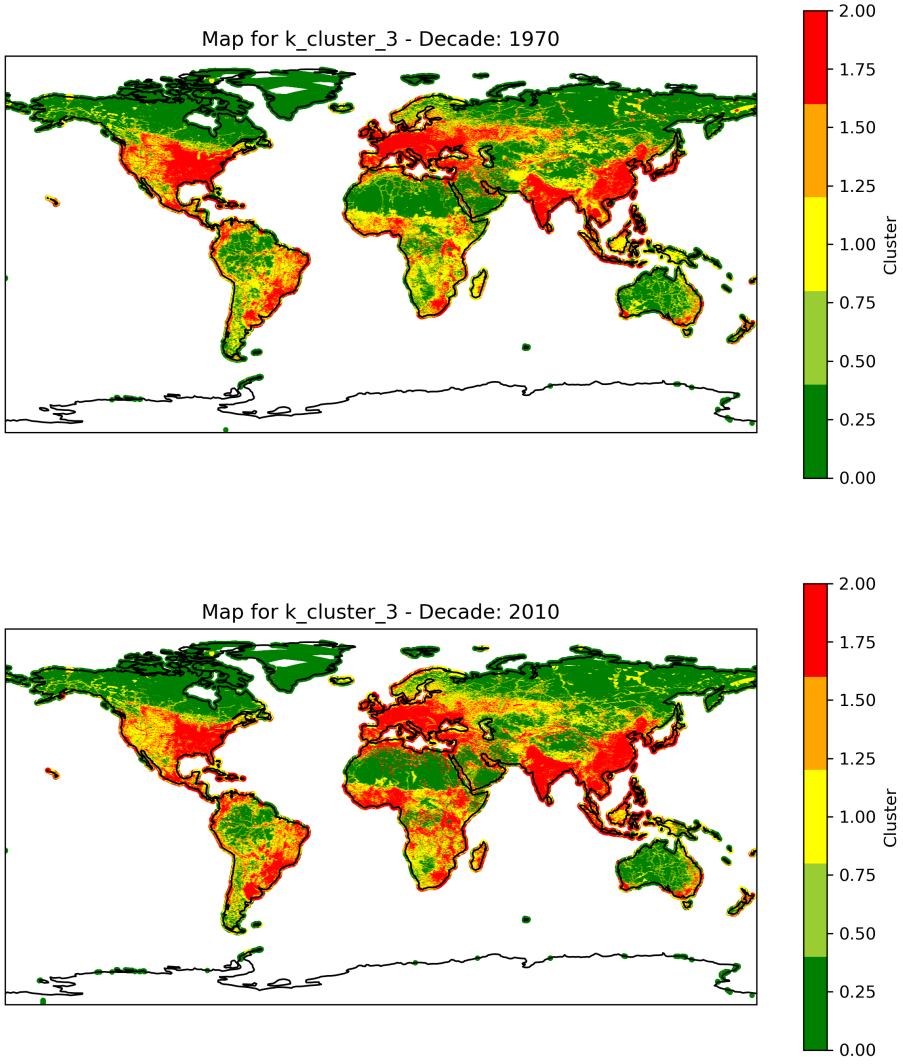
When examining the graph for k=4, there's a more nuanced differentiation of pollution patterns, with each cluster showing distinct temporal trends. Cluster 1 exhibits stability over the decades, while other clusters show increases or decreases.

During the increase of highly polluted areas denoted by cluster 3, the transition is mostly aided by a reduction in moderately polluted areas denoted by cluster 2 and non-polluted areas denoted by cluster 0. Bringing a degree of nuance to the trajectory analysis.

These trajectory graphs are invaluable as they depict the evolution of air pollution over time and across different clusters, highlighting the dynamic nature of environmental challenges.

5.3 Cluster Distribution Maps

The cluster maps for k=3 and k=4 reveal varied air pollution patterns, with the k=3 map in 2010 showing high-pollution clusters in industrial areas. This suggests a link to human-related emissions.



Comparing the cluster maps from the 1970s to the 2010s, we observe a transition in pollution patterns. The clusters with the highest pollution levels appear to shift geographically, possibly reflecting the global migration of industrial activities. Over time, regions that once showed high levels of pollutants have transitioned into cleaner clusters, which could be associated with successful environmental policies or deindustrialization.

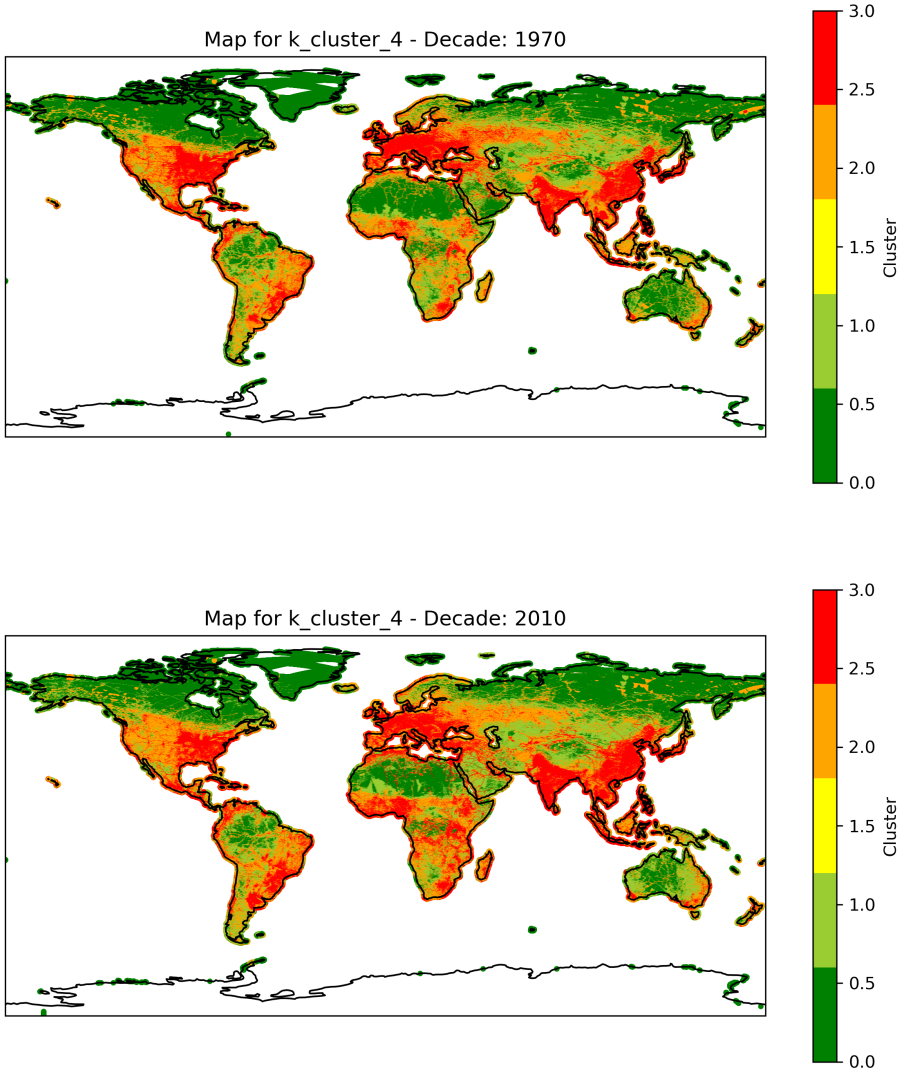


Figure 10: Map for k=4

Furthermore, this is aided by the the trajectory analysis which suggests a decrease in highly polluted areas in 2010, after its peak in 1990. However, an analysis of the maps also showcases that even when the highly polluted cluster has reduced in occurrence across the globe, the decrease is not evenly distributed. As most of it can be attributed to the decrease in pollution in Eastern Europe and Western North America, whereas, places like Africa, South America and East Asia have seen consistently higher levels of pollution each decade.

6 Conclusion

6.1 Challenges

This study faced significant computational challenges in handling an extensive dataset exceeding 100 GB, necessitating innovative approaches in data preprocessing and analysis.

Furthermore, the emission variables in the dataset exhibited strong positive skewness, complicating the application of k-means clustering and necessitating logarithmic transformations and standardization.

6.2 Limitations

Our computational resources limited the depth of analysis, especially in separating emissions by sectors. The absence of sector-specific emission analysis in our approach could have provided more detailed insights for targeted intervention strategies.

An additional limitation is the uniform treatment of grid points in the dataset. Since the actual size of these grids varies with their geographic location, this approach may not accurately reflect local emission patterns and impacts.

6.3 Future Works

Future work should focus on expanding computational resources to allow for more detailed analyses, including sector-specific emission studies and incorporating more recent data. Developing predictive models for air pollution trends based on this study's groundwork could enhance the understanding and forecasting of pollution patterns, aiding in proactive policy formulation.

Expanding the geographic scope of the study could provide insights into regional variations in air pollution and the effectiveness of different policy interventions worldwide.

Additionally, future studies should consider conducting the analysis with an aggregation of emissions across sectors rather than by gas type, as this could allow for a more nuanced understanding of complex relationships between various sources of emissions. This approach is further motivated by the observation of distinct transportation emission bands over the clustered maps.

Lastly, integrating air pollution data with health statistics could offer a more comprehensive view of the impact of air pollution on public health, guiding health-focused environmental policies.

7 Acknowledgement

We are grateful for the support of Prof. Bistra Dilkina in formulating the problem statement and providing valuable suggestions and direction throughout

the course of the project. We would also like to thank the TA, Hannah Murray, for all her assistance.

References

[1] J. Smith and P. Rodriguez. "Analysis of Urbanization and Vehicular Emissions". In: *Journal of Environmental Studies* 4.2 (2020), pp. 123–130.

[2] Forrest M. Hoffman et al. "Land Mass Masking for Spatio-temporal Analysis". In: *Alaskan Environmental Monitoring* 7.3 (2013), pp. 55–68.

[3] Y. Huang and X. Li. "Interplay Between Agricultural Emissions and Air Quality". In: *Journal of Agricultural Studies* 5.1 (2022), pp. 200–210.

[4] X. Zhang and et al. "A Holistic View of Environmental and Health Impacts of Pollutants". In: *Global Environmental Health* 8.4 (2023), pp. 300–320.