Identification of Sister Cities using Pollution data in India

B.Tech Project June 2023-July 2024

Balaji Nagappan (CH20B022), Prof Ragunathan Rengasamy

Department of Chemical Engineering, IIT Madras

HIGHLIGHTS

In this project, the primary objective is to identify the sister cities, i.e. cities exhibiting similar pollution patterns. Sister cities offer valuable insights, enabling the interpolation of missing data points and facilitating the adoption of pertinent technologies.

Researchers have previously explored the clustering of air quality monitoring stations based on sister cities, acknowledging the potential benefits of such analyses. In this paper we show the importance of filtering out stations based on data quality score in order to mitigate the risk of misclassifications.

Results demonstrate tangible need for scoring mechanism and hierarchical clustering gave superior results as compared to K-Means clustering in terms of speed and accuracy.

INTRODUCTION

With urbanization and industrialization, air pollution in many countries and cities has become increasingly serious. Previous studies have shown that long-term inhalation of air pollutants causes adverse effects on human health. Thus monitoring air quality is at most necessary

CPCB provides high resolution data (15 min), but it suffers from high variabilities, inconsistencies and missing values. Poor-quality data can result in wasted resources, increased costs, unreliable analytics and bad business decisions.

Thus, a mechanism to score the quality of data is at most important. In this paper we have mentioned 5 key attributes for scoring any continuous dataset. Clustering algorithms like K-Means and hierarchical clustering were used to identify sister cities using DTW as distance metric to compare similarity between time series.

STATUS OF LITERATURE

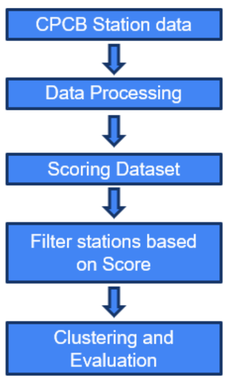
Scoring Mechanism

- Literature on methodology to score data quality is relatively extensive. The overall score gives us an idea about reliability of the data. Most papers use 5 attributes to measure data quality: accuracy, validity, consistency, uniqueness and completeness.
- For anomaly detection of timeseries, researchers have used Median Absolute Deviation (MAD) moving window approach and autoencoders. Both algorithms were explored in this paper.

Clustering

- There are existing studies on clustering air quality parameters. Researchers have used various clustering algorithms like K-Means, Hierarchical, K-Medoid, PAM etc.. Some papers performed PCA before time series clustering.
- Most papers used DTW (Dynamic Time Warping) as distance metric to measure similarity of timeseries.
- Studies show that multi-hour multi-site forecasting results are better when spatial clustering (i.e. identification of sister cities) is done.

METHODOLOGY



Data : CPCB station data (513 stations) **Resolution :** 15 min

Parameters: $PM_{2.5}$, PM_{10}, NO_2, SO_2, NH_3, O_3

Duration: Sept 2019 – Sept 2023

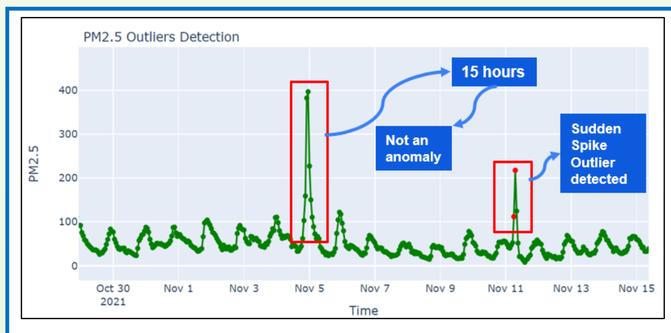
Data Quality Metrics

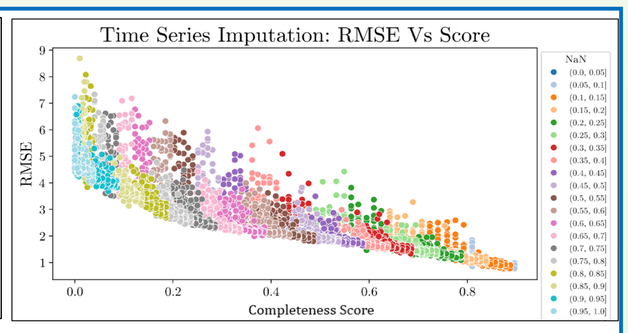
1. **Accuracy:** Detecting abnormalities in data using moving window MAD and autoencoders.
2. **Consistency:** Checks if data satisfies all the integrity constraints**.**

3. Uniqueness: Metric tracks duplicate records. Counting any data twice will unduly weigh the results.

4. Validity: Checks if records in data conforms to data type as required.

5. Completeness: Tracks number of missing values in data. We use custom completeness score that utilizes the mean missing block length for calculating completeness score.



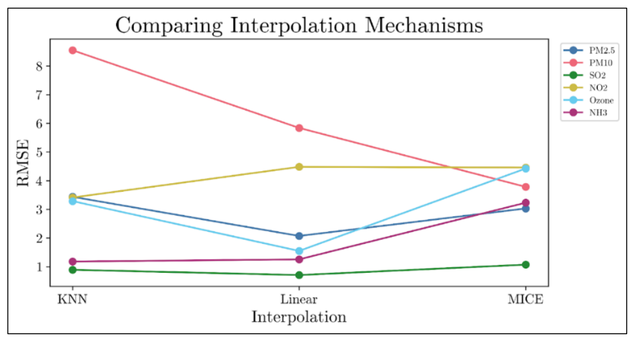


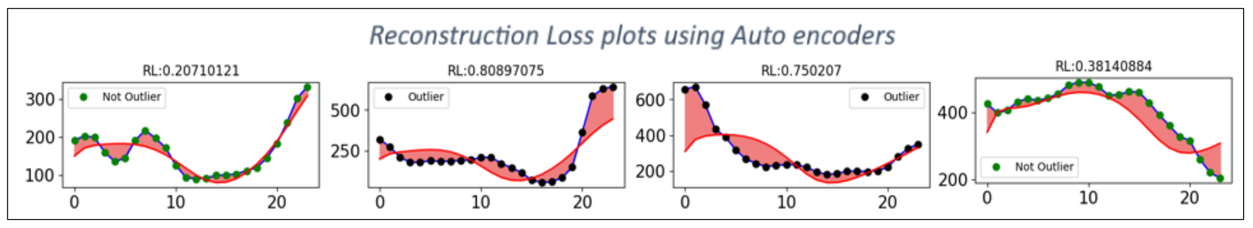
$$Completeness = (1 - \%\ NaN's) * (1 - \frac{Mean\ (Missing\ block\ Length)}{len(data)})$$

$$Overall\ Score = Completeness\ Score\ x\ Accuracy\ Score\ x\ Consistency\ Score\ x\ Validity\ Score$$

Interpolation: MICE for $PM_{2.5}$ and PM_{10}, linear interpolation for other parameters is best mechanism for interpolation.

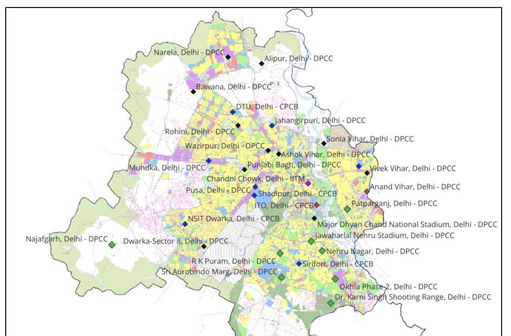
Clustering: K-Means and hierarchical clustering were implemented using DTW as the dissimilarity metric.

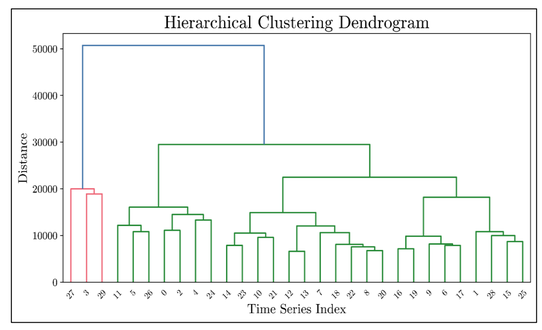


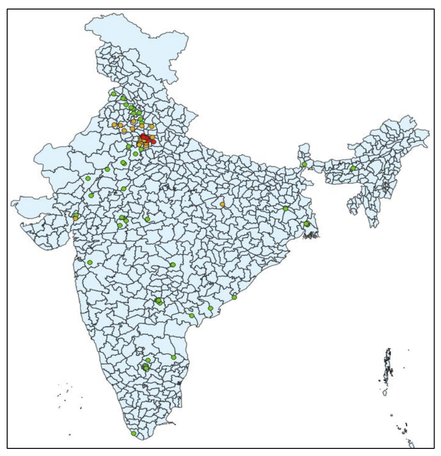


RESULTS AND DISCUSSION

- 30 Stations in Delhi clustered using Hierarchical clustering
- Hourly Resolution data was used for analysis due to computational reasons (DTW – time complexity)
- Land use pattern of Delhi is used as ground truth for clustering
- Chandini Chowk and NSIT Dwarka have low data quality score and have been misclassified into wrong cluster.
- Stations with lower data quality scores are susceptible to misclassification, likely because the data fails to capture the station's underlying pattern







- Hierarchical clustering proved to be both time-efficient and yielded superior results in comparison to k-means clustering using Rand-Index as metric for comparison.
- Clustering of CPCB stations in India was done after applying threshold on overall data quality score to avoid possible misclassifications. (102 stations were clustered in 3 groups)

CONCLUSION

The research article introduces a reliable scoring mechanism to assess the data quality of continuous datasets. Our analysis revealed that stations with lower data quality scores are susceptible to misclassification, likely because the data fails to capture the station's underlying pattern. This suggest the importance of filtering out stations before clustering. Hierarchical clustering gave us superior results as compared to k-means in terms of time and accuracy.

REFERENCES

[1] Firas Bayram, Bestoun S. Ahmed, Erik Hallin, Anton Engman, June 2023, Oulu, Finland
DQSOps: Data Quality Scoring Operations Framework for Data-Driven Applications
[2] Xu, Z.; Liu, Z.; Tian, J.; Liu, Y.; Pan, H.; Liu, S.; Yang, B.; Yin, L.; Zheng, W. Classification of Urban Pollution Levels Based on Clustering and Spatial Statistics. Atmosphere 2022, 13, 494.
[3] Suris, F.N.A.; Bakar, M.A.A.; Ariff, N.M.; Mohd Nadzir, M.S.; Ibrahim, K. Malaysia PM10 Air Quality Time Series Clustering Based on Dynamic Time Warping. Atmosphere 2022, 13, 503.
[4] Kulanuwat, L.; Chantrapornchai, C.; Maleewong, M.; Wongchaisuwat, P.; Wimala, S.; Sarinnapakorn, K.; Boonya-aroonnet, S. Anomaly Detection Using a Sliding Window Technique and Data Imputation with Machine Learning for Hydrological Time Series. Water 2021, 13, 1862.