Prediction of Future AQI for Cities in India using Geospatial Analysis of Air Pollution data

Mini Project for Spatial Modeling

Submitted by:
Rutuja Deshpande: 23070243009
Vibhoor Bhatnagar: 23070243058
Smruti Dighe: 23070243013
Sharon Futardo: 23070243015

Under the Guidance of Dr. Sandipan Das

ABSTRACT

This project aims to forecast the Air Quality Index (AQI) for various cities in India by leveraging machine learning models and geospatial analysis. The dataset, sourced from the Government of India's data portal, includes historical pollutant levels for multiple cities. Using a Long Short-Term Memory (LSTM) model and XGBoost, we developed predictive models to forecast AQI values, achieving a Root Mean Square Error (RMSE) of 0.90 for both techniques. LSTM was chosen for its ability to capture temporal dependencies within time-series data, while XGBoost was selected for its performance in structured data and capacity to analyze feature interactions.

To further enhance the spatial analysis, we applied geospatial techniques using ArcGIS Pro, including Kriging and Inverse Distance Weighting (IDW), to interpolate AQI levels both for historical data and predicted values across different urban regions. This approach allows for a more comprehensive understanding of AQI distribution and spatial variations across cities. Our findings underscore the utility of combining machine learning models with geospatial analysis to provide accurate AQI predictions, which are vital for urban planning and public health policy. The project highlights key pollutant contributions and spatial trends, supporting efforts to mitigate air pollution in urban environments across India.

INTRODUCTION

Air quality has become a critical concern in urban and industrial areas due to frequent exceedances of safe pollutant levels, which pose significant risks to public health. Accurate predictions of the Air Quality Index (AQI) are essential to enable timely interventions that mitigate the negative health impacts of pollution. Various researchers have applied machine learning and statistical methods to enhance AQI prediction by analyzing both spatial and temporal patterns of pollutants across different regions.

For instance, **Wang et al. (2023)** employed a high-resolution spatiotemporal model to estimate daily AQI levels in Shanghai and to identify key pollutants contributing to AQI variations across urban regions. Their approach, which incorporated partial least squares (PLS) and universal kriging within a geostatistical modeling framework, enabled accurate mapping of pollutant concentrations and highlighted the impact of ozone (O3), PM2.5, and NO2 on unhealthy AQI levels. Similarly, **Sarmadi et al. (2021)** investigated AQI changes in 87 cities worldwide before, during, and after the COVID-19 pandemic. Using a range of statistical tests, they observed significant reductions in pollutants such as PM2.5 and NO2 during lockdown periods, along with an increase in ozone levels that revealed regional variations.

In Jakarta, **Handhayani (2023)** analyzed causal relationships between pollutants and meteorological conditions using both LSTM and GRU models, finding that LSTM offered superior forecasting accuracy. **Liu et al. (2019)** compared the performance of Support Vector Regression (SVR) and Random Forest Regression (RFR) for AQI predictions in Beijing and an Italian city. Their findings indicated that SVR achieved high accuracy for AQI prediction, while RFR was more effective for nitrogen oxides (NOx) forecasting. These studies underscore the importance of employing advanced models to handle complex datasets and capture pollutant dynamics across diverse environments.

In an Indian context, **Janarthanan et al. (2021)** combined SVR and LSTM models to forecast AQI in Chennai, achieving high accuracy and highlighting the potential of hybrid modeling approaches to inform local environmental policies. **He and Luo (2020)** further explored the predictive efficacy of LSTM in Chengdu, showing its advantages over traditional neural networks for short-term AQI forecasting when considering meteorological variables.

LITERATURE REVIEW

S. No	Author & Year	Objective	Technique	Dataset	Results
1	Wang et al., 2023	Estimate daily AQI levels at 100-meter resolution in a metropolitan city (Shanghai, 2019), understand spatial patterns of AQI, and identify predominant pollutant contributions.	Spatiotemporal model using a geostatistical framework; includes partial least squares (PLS), spline smoothing, universal kriging; data imputation using Random Forest.	Hourly air quality data from 58 monitoring sites in Shanghai (CNEMC & EMA); MAIAC AOD, OMI NO2, TROPOMI NO2 satellite data; Geographic covariates (land use, NDVI, elevation, road networks, POIs, population density, nighttime lights).	Good model performance (CV R2 = 0.86, RMSE = 10.05 for AQI); substantial spatial variation in AQI; Ozone (O3) dominated unhealthy air quality (AQI > 100) in a large area; PM2.5 and NO2 were also significant contributors to AQI, particularly in specific areas.
2	Sarmadi et al. (2021)	Investigate changes in Air Quality Index (AQI) in various cities before, during, and after the COVID-19 pandemic.	Bivariate correlation analysis, Paired-sample t-test, Wilcoxon signed-rank test, Multivariable linear regression	AQI data from the World Air Quality Index (WAQI) for 87 capital, industrial, and polluted cities worldwide (2018-2021, Jan-Apr). Meteorological data (temperature, wind speed, relative humidity) also included.	Significant improvement in AQI for PM2.5, PM10, and NO2 in 2020 compared to 2019, reversed in 2021. Temperature and relative humidity inversely correlated with AQI-PM2.5, AQI-PM10, and

					AQI-NO2. Significant differences in AQI across developed and developing countries, and high vs. lower-middle income countries. Ozone levels increased in 2020 compared to 2019 in many cities.
3	Handhayani, 2023	Analyze causal relationships between air pollutants and meteorological conditions in Jakarta; forecast AQI and meteorological conditions	PC algorithm, LSTM, GRU	Integrated dataset of Jakarta AQI and meteorological data (2010-2021)	Causal relationships identified; LSTM performs better than GRU for forecasting using integrated data; Covid-19 outbreak impacted forecasting accuracy.
4	Liu et al., 2019	Predict Air Quality Index (AQI) in Beijing and nitrogen oxides (NOx) concentration in an Italian city.	Support Vector Regression (SVR), Random Forest Regression (RFR)	Beijing Air Quality Dataset (Dec 2013-Aug 2018), Italian city air quality data (Mar 2004-Feb 2005)	SVR performed better for AQI prediction (RMSE = 7.666, R^2 = 0.9776, r = 0.9887). RFR performed better for NOx prediction (RMSE =

					83.6716, R^2 = 0.8401, r = 0.9180).
5	Janarthanan et al., 2021	Predict Air Quality Index (AQI) in Chennai, India.	Data preprocessing (missing value replacement, redundant data removal); Feature extraction (Grey Level Co-occurrence Matrix - GLCM); Feature optimization (Modified Fruit Fly Optimization - MFOA); Classification (Support Vector Regression - SVR with Long Short-Term Memory - LSTM)	Data from 3 Central Pollution Control Board (CPCB) monitoring stations in Chennai (Manali, Velachery, Alandur). Data included relative humidity (RH), PM2.5, atmospheric pressure (BP), wind speed (WS), wind degree (WD), NO2, SO2, CO, and Ozone. Data collected at 15-minute intervals from May 1, 2019 to April 30, 2020.	Accurate AQI prediction for specific locations within Chennai. Improved prediction accuracy compared to existing techniques. RMSE and R^2 values reported for PM2.5, NO2, SO2, CO, and Ozone. The model showed good fit for PM2.5 prediction (R^2 of 0.632 for training and 0.570 for testing). Other pollutants showed lower R^2 values.
6	. He & Luo (2020)	To predict Air Quality Index (AQI) accurately, considering pollution sources, meteorological conditions, and time series, and to compare the prediction accuracy of LSTM with	LSTM (Long Short-Term Memory) neural network; Transfer entropy for meteorological factor selection; BP neural network; GRU (Gated	Hourly air quality index data from Chengdu, China (January 1, 2018 - September 15, 2019); Corresponding meteorological data (from Leshan City meteorological center).	LSTM showed better prediction accuracy and robustness than BP neural network and GRU, particularly for short-term (0-24h) and

		traditional BP neural network and GRU	Recurrent Unit).		very short-term (0-12h) forecasts. GRU showed no significant advantage over LSTM. RMSE used as evaluation metric.

DATASET REVIEW

This project utilizes a publicly available air quality dataset from data.gov.in](https://data.gov.in), which contains air quality data for multiple cities across India. The dataset records daily AQI values along with concentrations of key pollutants that contribute to air quality. It spans from January 2015 to recent years, providing a robust basis for time-series and spatial analysis. Here, we detail the dataset's key attributes, structure, and any preprocessing steps necessary to prepare the data for modeling.

The dataset includes 29,531 records and 16 columns, each representing a specific feature or data attribute. The following list describes each column:

- City: Name of the city where the data was recorded. This categorical feature allows for spatial segmentation in our analysis.
- Date: The date on which measurements were taken. This is essential for time-series analysis, enabling us to capture seasonal and temporal variations in AQI and pollutant levels.
- AQI (Air Quality Index): A computed index representing overall air quality based on the concentration of multiple pollutants. This is the primary target variable for our predictive models.

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{HI} - BP_{Lo}}\left(C_p - BP_{Lo}\right) + I_{L0}$$

Where ,
I_p = the index for pollutant p
I_{Hi} = the AQI value corresponding to BP_{HI}
I_{Lo} = the AQI value corresponding to BP_{Lo}
BP_{HI} = the concentration breakpoint that is greater than or equal to C_p
BP_{Lo} = the concentration breakpoint that is lesser than or equal to C_p
C_p = the truncated concentration of pollutant p

- AQI_Bucket: A categorical representation of AQI, indicating air quality categories such as "Good," "Moderate," and "Poor." This is derived based on AQI ranges and provides an additional perspective on the severity of pollution.

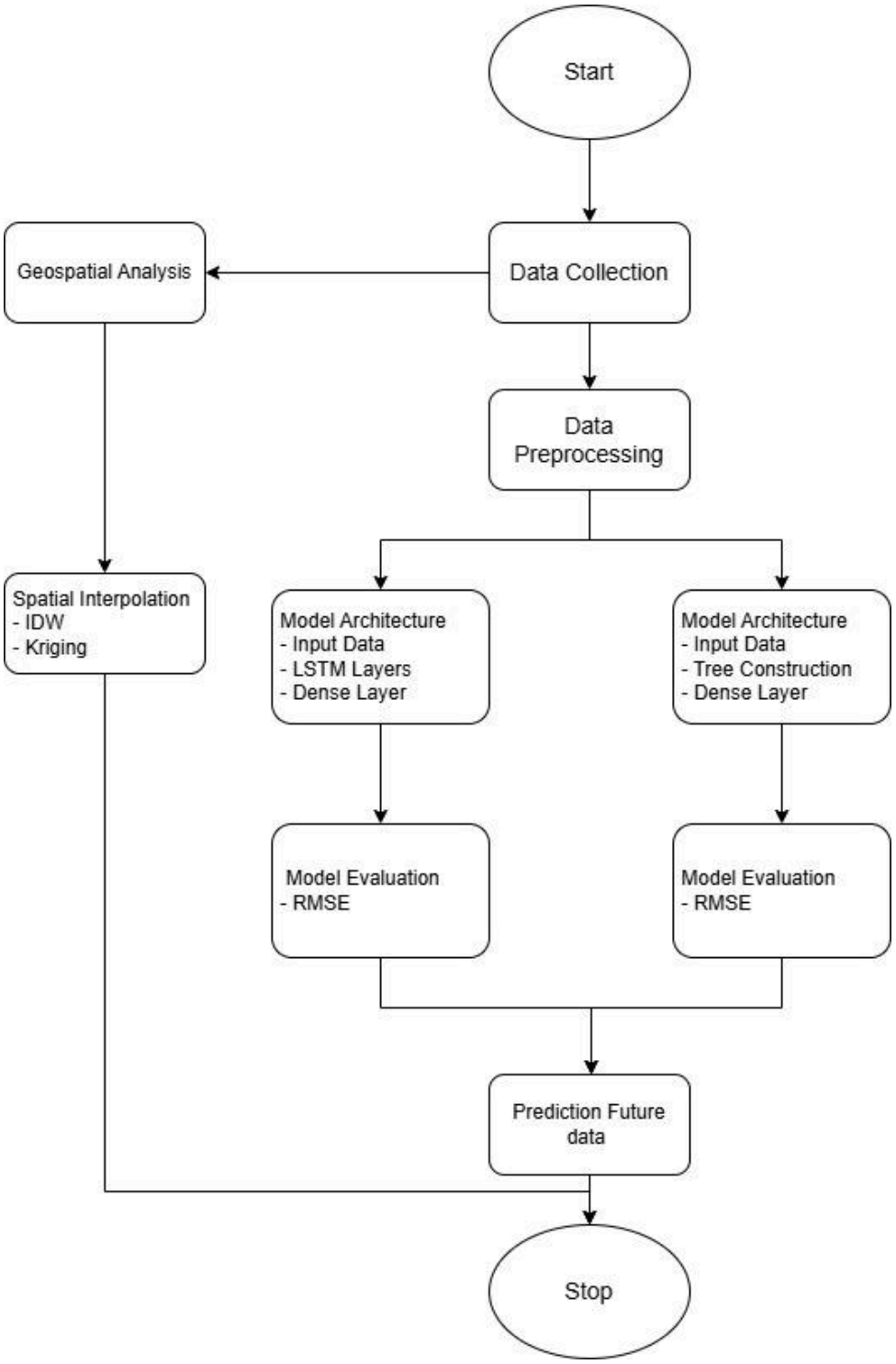
The remaining columns represent specific air pollutants, each of which contributes uniquely to air quality. Below is a description of each pollutant:

- PM2.5 (Particulate Matter 2.5): Particulate matter with a diameter of less than 2.5 micrometers. PM2.5 is particularly harmful as it can penetrate deep into the lungs and bloodstream, leading to severe health issues.
- PM10 (Particulate Matter 10): Particulate matter with a diameter of less than 10 micrometers. PM10 can affect the respiratory system but is less penetrative than PM2.5.

- NO (Nitric Oxide): A gas emitted primarily from vehicular emissions and industrial activities. It contributes to the formation of ground-level ozone and can affect respiratory health.
- NO2 (Nitrogen Dioxide): Formed by the oxidation of NO, NO2 is associated with various respiratory problems and is a key contributor to smog.
- NOx (Nitrogen Oxides): A collective measure of NO and NO2, commonly used in air quality monitoring to assess overall nitrogen oxide levels.
- NH3 (Ammonia): Primarily emitted from agricultural activities, NH3 can contribute to the formation of secondary pollutants, such as particulate matter.
- CO (Carbon Monoxide): A colorless, odorless gas produced by incomplete combustion of fossil fuels. High levels of CO can be life-threatening, as it reduces the blood's oxygen-carrying capacity.
- SO2 (Sulfur Dioxide): Produced mainly from industrial activities, SO2 can irritate the respiratory system and contribute to the formation of particulate matter.
- O3 (Ozone): Ground-level ozone is a secondary pollutant formed through the reaction of sunlight with NOx and volatile organic compounds. O3 can lead to respiratory issues and is a major component of smog.
- Benzene: An organic compound and carcinogen, commonly found in industrial emissions and vehicular exhaust.
- Toluene: Another volatile organic compound with harmful health effects, primarily emitted from industrial sources and vehicles.
- Xylene: Similar to Benzene and Toluene, Xylene is a toxic compound released from industrial activities and vehicles.

METHODOLOGY

DATA FLOW CHART



Data Preprocessing

In this project, data preprocessing involved several steps to clean, format, and prepare the data for accurate AQI prediction. Given the critical importance of data quality in machine learning models, the following steps were undertaken:

1. Date Formatting :

- The `Date` column was converted to a standard datetime format. This conversion enabled time-series sorting and ensured that each entry could be accurately associated with a specific day, essential for temporal analysis and predictions.

2. Sorting and Grouping :

- The dataset was sorted by `City` and `Date` to allow for accurate chronological analysis within each city. Sorting ensures that temporal dependencies, crucial for LSTM modeling, are preserved in the data structure.

3. Handling Missing Values :

- Forward and Backward Filling :

Missing values in the pollutant columns (e.g., PM2.5, PM10, NO, NO2, etc.) were addressed using forward-filling and backward-filling within each city. This method helped to fill missing pollutant values by referencing adjacent days, preserving temporal continuity without introducing synthetic trends.

- Mean Imputation :

Numerical columns without a clear temporal order were filled using the mean values of each column, reducing the bias introduced by missing data.

- Mode Imputation for `AQI_Bucket`:

The `AQI_Bucket` column, which categorizes AQI, was filled with the mode (most common value) to maintain consistency without losing categorical information.

4. Feature Selection :

A selection of relevant pollutants (e.g., PM2.5, PM10, NO, NO2, etc.) was defined as the feature set for predicting AQI. This selection focused on pollutants with the highest impact on air quality to enhance model relevance and interpretability.

- Target Variable :

AQI and AQI_Bucket were identified as primary target variables. AQI serves as the continuous target for regression-based approaches, while AQI_Bucket can be used in classification models to categorize air quality levels.

Algorithms Used

The main algorithms used were Long Short-Term Memory (LSTM) for time-series analysis and XGBoost for structured data modeling.

1. **Long Short-Term Memory (LSTM) Network :**

   - Purpose:

   LSTM networks are particularly well-suited for time-series forecasting, as they can capture long-term dependencies within sequential data. In this project, the LSTM model was used to predict future AQI values by analyzing historical pollutant levels for each city.

   - Model Structure :

   The LSTM model was configured to take historical values of selected pollutants (PM2.5, PM10, NO, etc.) as inputs, allowing it to learn patterns over time. It was trained to output AQI values, thereby providing a continuous prediction of air quality levels.

   - Training and Validation :

   The data was split into training and validation sets, ensuring the model's ability to generalize by evaluating performance on unseen data. Standard performance metrics like RMSE (Root Mean Square Error) were used to assess the model's accuracy in predicting AQI.

2. **XGBoost :**

   - Purpose :

   XGBoost is a gradient-boosted decision tree algorithm valued for its strong predictive performance and interpretability on structured datasets. In this project, XGBoost was chosen to predict AQI values while capturing complex feature interactions, aiding in understanding each pollutant's contribution to overall air quality.

   - Feature Importance Analysis :

   Through XGBoost, we assessed the importance of each pollutant in influencing AQI, providing insights into which pollutants have the most substantial impact on air quality. This analysis highlighted key contributors, allowing for a more targeted understanding of pollutant effects.

   - Model Configuration:

   The XGBoost model was trained on the pollutant data to predict AQI values. To optimize performance, the model underwent hyperparameter tuning, which enhanced prediction accuracy and minimized the risk of overfitting by carefully selecting parameters suited to the dataset.

- **Training and Evaluation** :

To ensure model generalizability, data was split into training and validation sets, and performance was evaluated using metrics like Mean Absolute Error (MAE) and R^2 (coefficient of determination). These metrics allowed us to measure prediction accuracy and the model's ability to explain variations in AQI.

3. Geospatial Analysis :

- **Objective** :

To create a comprehensive spatial representation of AQI across different regions, we applied **interpolation techniques** in ArcGIS Pro, allowing us to estimate AQI values at unmonitored locations based on values from existing monitoring stations.

- **Mapping and Visualization :**

Interpolation is used to create a continuous surface from discrete points, offering a complete view of pollutant concentrations across urban and regional areas. This technique is essential for understanding the spatial distribution of air quality, as AQI measurements are often limited to specific monitoring locations, leaving gaps in data coverage.

- **Inverse Distance Weighting (IDW)**: IDW is an interpolation method where closer points have a greater influence on the estimated values than those farther away. The rationale behind using IDW is its effectiveness in regions with densely located data points, as it assumes that points close together will have more similar values. For this project, IDW was used to highlight AQI hotspots around monitoring stations, providing a straightforward visual representation of pollution intensity based on proximity to measured values.

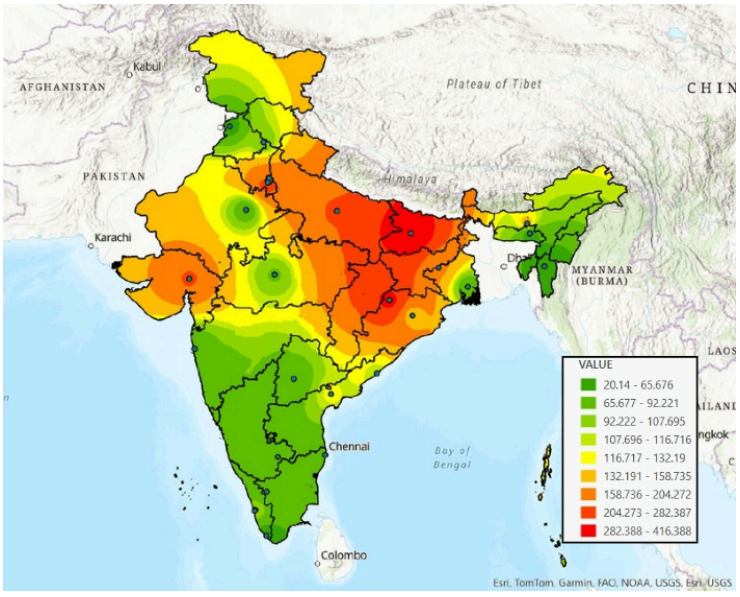


Fig: Interpolation of the past data using IDW

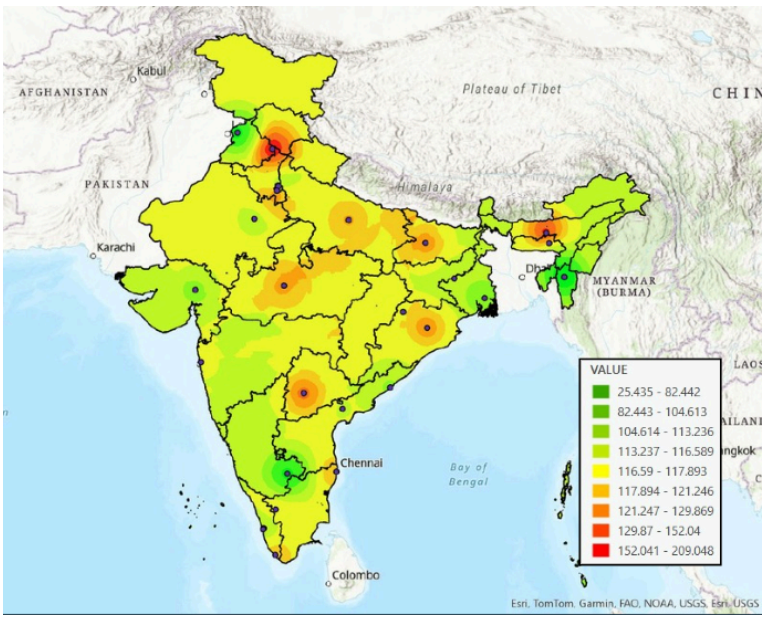


Fig: Interpolation of the predicted data using IDW

- **Kriging**: Kriging is a more advanced geostatistical interpolation technique that, unlike IDW, incorporates both the distance and the spatial arrangement of all sample points to predict values across a surface. It assumes spatial autocorrelation, meaning that points closer to each other are more likely to be similar than those further apart. Kriging was chosen for its statistical rigor and ability to model more complex spatial relationships, which is especially valuable for accurately predicting AQI in regions with variable data density. This method provides a smoother, more reliable spatial surface, revealing subtle patterns in pollution that may not be captured by IDW alone.

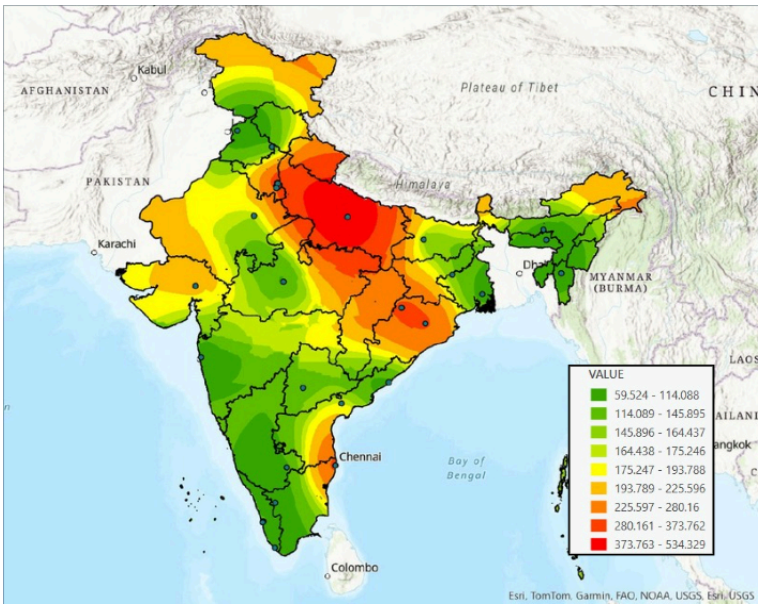


Fig: Interpolation of the past data using Kriging

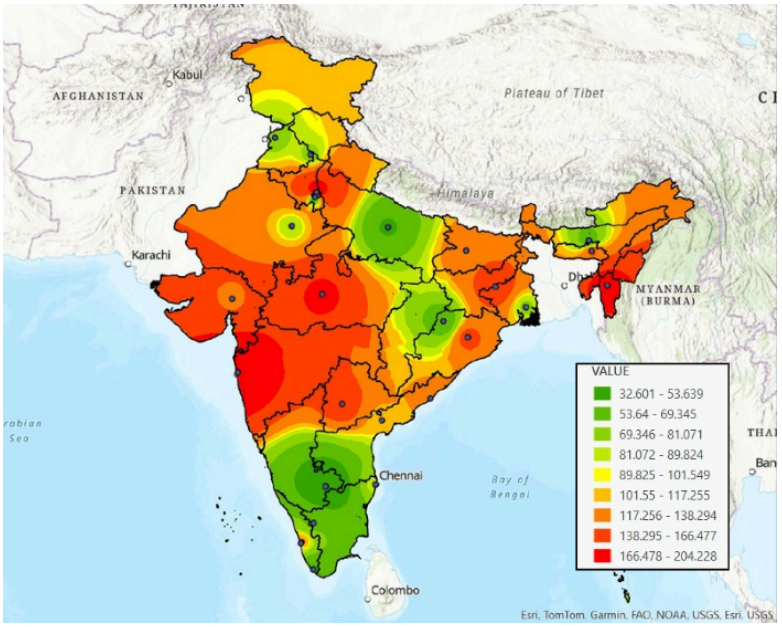


Fig: Interpolation of the predicted data using Kriging

These interpolated AQI surfaces, generated through IDW and Kriging, were visualized in ArcGIS Pro. The resulting maps show color gradients indicating AQI levels, with high-AQI areas clearly marked, often correlating with urbanized or industrial regions. This geospatial perspective enhances our temporal AQI predictions by identifying specific areas that may require environmental management, enabling policymakers to target interventions effectively based on both spatial and temporal insights.

4. Evaluation Metrics

To validate the model's performance, several metrics were used:

- **RMSE (Root Mean Square Error)**: Used for LSTM predictions, RMSE measures the average magnitude of errors, indicating how closely predictions match actual AQI values.
- **MAE (Mean Absolute Error)**: This metric provides a straightforward average error measurement that helps interpret XGBoost's prediction accuracy.

RESULTS

Using Root Mean Square Error (RMSE) as the major criterion for accuracy assessment, the performances of various models were compared. The models compared were LSTM, XGBoost, Kriging, and IDW. The results of the analysis are summarized below:

- **LSTM:**
  The LSTM model has shown good predictive capability with the measure of RMSE of 0.90490. This shows how LSTM can capture sequences and time series data it can easily identify the temporal dependencies common in the data set.

- **XGBoost:**
  As can be seen from the table, the best result is given by XGBoost, a gradient-boosting framework with the RMSE of 0.81445. Thus, using specific testing metrics, that is, preventing/co-efficient of variation/rating etc. RMSE can be seen where XGboost exhibited the lowest RMSE and therefore, the most accurate amongst the tested methods, attributable due to XGboost's mechanisms against non-linear relationship and over-fitting.

- **Kriging:**
  Thus, the Kriging, which is a geostatistical interpolation technique, provided an RMSE of 1.709579. Compared with machine learning algorithms, Kriging is good at modeling spatial dependencies but has a higher error due to the large variance of the data set.

- I**DW(Inverse Distance Weighting):**
  The lowest accuracy was recorded by the IDW interpolation technique that had an RMSE of 2.17137. This method can also lack the necessary sophistication in assessing the weights since distance-based weights do not capture interactions between variables sufficiently well.

Methods	**Root Mean Square Error (RMSE)**
LSTM	0.90490
XGBoost	0.81445
Kriging	1.709579
IDW(Inverse Distance Weighting)	2.17137

CONCLUSION

XGBoost was found to be the best method, which gave a lower value of RMSE among the four methods considered. Next to it, the LSTM model also showed reasonable accuracy based on the result analysis. At this point Kriging and IDW interpolations could be considered less efficient with much higher RMSE values demonstrating the potential of Machine learning for predictive modeling for the given dataset.

Based on the RMSE analysis, it is evident that the performance of the tested models varies significantly depending on their ability to handle the underlying data structure and relationships. Among the models, **XGBoost** stands out with the lowest RMSE of **0.81445**, showcasing its superior ability to model complex, non-linear relationships and mitigate overfitting. This makes XGBoost the most accurate method in this comparison.

The **LSTM model** also demonstrated strong predictive capabilities with an RMSE of **0.90490**, highlighting its proficiency in capturing temporal dependencies inherent in sequential or time-series data. While it did not outperform XGBoost, it remains a robust choice for tasks involving temporal patterns.

On the other hand, **Kriging**, with an RMSE of **1.709579**, showed higher error rates, reflecting its limitations in handling datasets with large variance despite its strength in modeling spatial dependencies. This suggests that while Kriging is effective in geostatistical interpolation, it may not be as suitable for datasets with complex relationships.

Finally, the **IDW (Inverse Distance Weighting)** method recorded the highest RMSE at **2.17137**, indicating the lowest accuracy among the models. Its reliance on distance-based weights without adequately accounting for variable interactions makes it less effective for this dataset.

In conclusion, **XGBoost** emerged as the most accurate model, followed by LSTM, while Kriging and IDW lagged due to their inability to address the dataset's complexity and variability effectively. These findings underscore the importance of selecting appropriate models tailored to the data's characteristics for achieving optimal accuracy.

REFERENCES

1 - Wang, Y., Huang, L., Huang, C., Hu, J., Wang, M. (2023). High-resolution modeling for criteria air pollutants and the associated air quality index in a metropolitan city. Environment International, 172, 107752.

2 - Sarmadi, M., Rahimi, S., Rezaei, M., Sanaei, D., & Dianatinasab, M. (2021). Air quality index variation before and after the onset of COVID-19 pandemic: a comprehensive study on 87 capital, industrial and polluted cities of the world. Environmental Sciences Europe **https://doi.org/10.1186/s12302-021-00575-y**

3 - Handhayani, T. (2023). An integrated analysis of air pollution and meteorological conditions in Jakarta. Scientific Reports, 13(5798).

https://doi.org/10.1038/s41598-023-32817-9

4 - Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. Applied Sciences, 9(19), 4069. https://doi.org/10.3390/app9194069

5 - Janarthanan, R., Partheeban, P., Somasundaram, K., & Elamparithi, P. N. (2021). A deep learning approach for prediction of air quality index in a metropolitan city. Sustainable Cities and Society, 67, 102720. https://doi.org/10.1016/j.scs.2021.102720

6 - He, H., & Luo, F. (2020). Study of LSTM air quality index prediction based on forecasting timeliness. IOP Conference Series: Earth and Environmental Science, 446(3), 032113. https://doi.org/10.1088/1755-1315/446/3/032113