Spatiotemporal Analysis and Prediction Modeling of Air Pollution in Dublin Based on Traffic and Meteorological Factors

Abstract

This report uses air pollution data, weather data (wind speed), and traffic flow data in Dublin, Ireland. The analysis includes data cleaning and integration, time series trend research, correlation analysis, pollutant regression predictive modeling analysis, and traffic flow clustering analysis, supplemented by PCA spatial pollution distribution. The results show that pollutant concentrations have obvious differences in different time periods; wind speed and traffic flow show a certain negative correlation with some pollutants; at the same time, the prediction model shows that historical pollution data is of great value for predicting current pollutant concentrations. In addition, cluster analysis found that there are abnormalities in the distribution of pollutant concentrations under different traffic flow levels, and the pollution concentration is lower when the traffic flow is higher.

Data Collection and Preprocessing

We used air pollution data, wind speed data, and SCATS traffic flow data for Dublin. During data processing, the time field of air pollution data was used as the main table, and wind speed and traffic flow data were uniformly aggregated to the hourly scale for integration. Negative values in pollutant data were uniformly replaced with 0, and time series interpolation was used to fill in missing values in pollution data, ultimately forming a complete data set.

Table 1

Data Example (partial)

datetime	O3_ugm3	CO_mgm3	PM25_ugm3	traffic_volume	wind_speed
2021-05-06 07:00:00	97.031800	0.518892	7.273711	471529	9.0
2021-05-06 12:00:00	55.077142	0.351417	4.626918	861725	11.0
2021-05-06 13:00:00	57.717606	0.335741	4.339602	913258	11.0

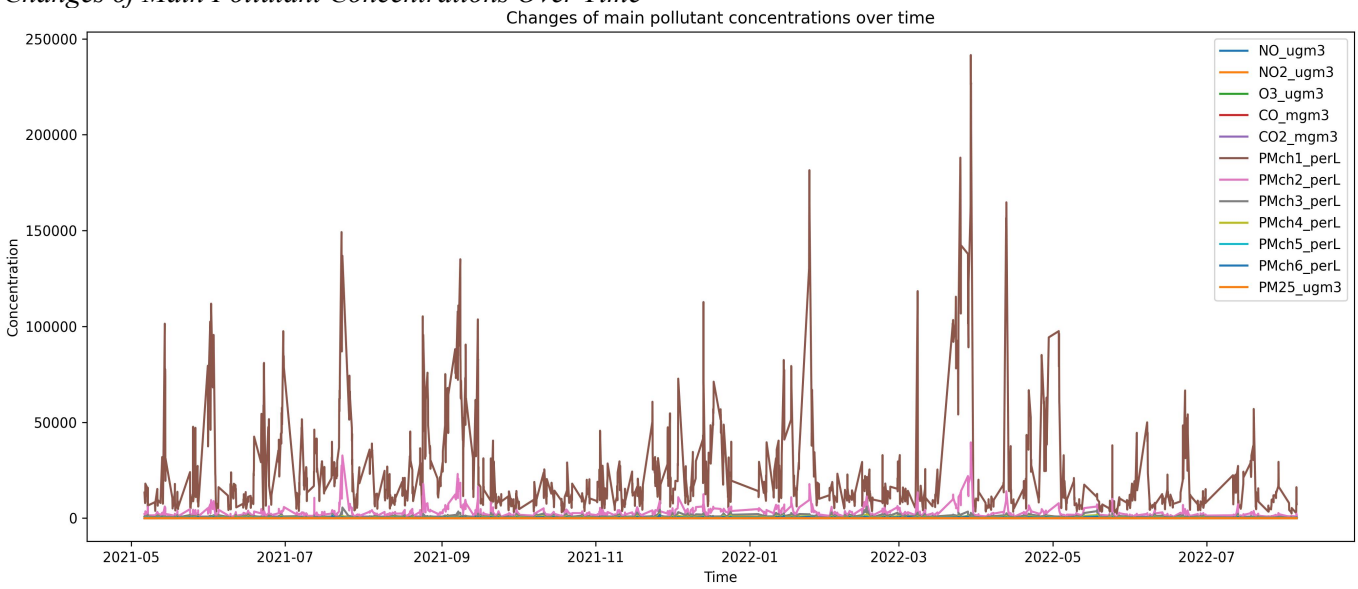
Exploratory Data Analysis

Time Series Analysis

In order to explore the changing trends of different pollutant concentrations over time and understand the potential periodic or seasonal patterns, we conducted a time series analysis of pollutants. With time as the horizontal axis and pollutant concentration as the vertical axis, a line graph of the changing trends of each pollutant from 2021 to 2022 was drawn.

Figure 1

Changes of Main Pollutant Concentrations Over Time



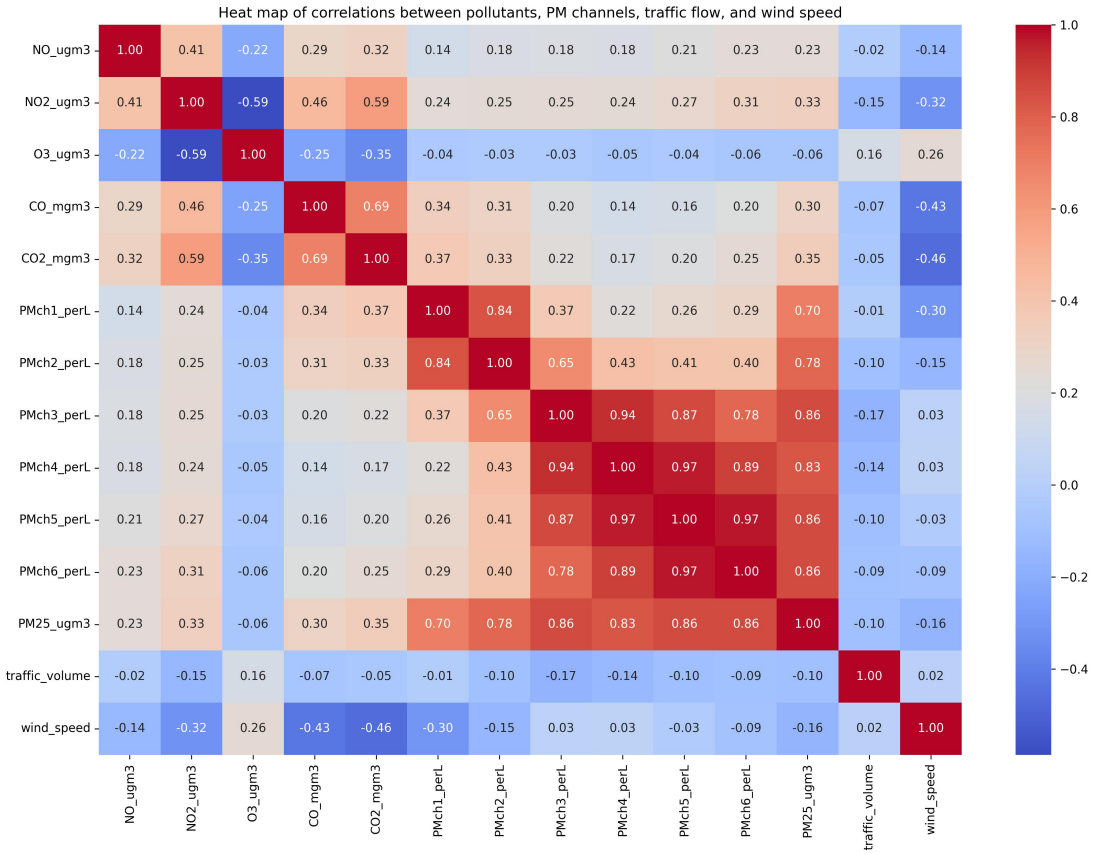
It can be clearly seen from the figure that PMch1_perL has drastic fluctuations, and its concentration is significantly higher than other pollutants, suggesting that it may be related to special emission events or seasonal factors. Other pollutants showed a relatively stable and low-amplitude fluctuation trend.

Correlation Analysis

Preliminary analysis shows that traffic flow and most pollution concentrations show an unexpected negative correlation. After data integrity checks and statistical verification, we believe that this is the result of the combined effect of data characteristics and sampling mechanisms. The following is a heat map of the correlation between pollutant concentrations and wind speed and traffic flow and its analysis:

Figure 2

Heat Map of the Correlation Between Pollutant Concentration, Wind Speed and Traffic Flow



In terms of traffic flow, O3 is slightly positively correlated with traffic flow (0.16). NO, NO2, CO, CO2 and PM channels are mostly weakly negatively correlated with traffic flow, especially PMch3 (-0.17).

In terms of wind speed, wind speed is generally negatively correlated with NO, NO2, CO, CO2 and PM particles, among which CO2 (-0.46) and CO (-0.43) are the most significant negative correlations. O3 is slightly positively correlated with wind speed (0.26).

Pollutant Prediction Modeling and Analysis

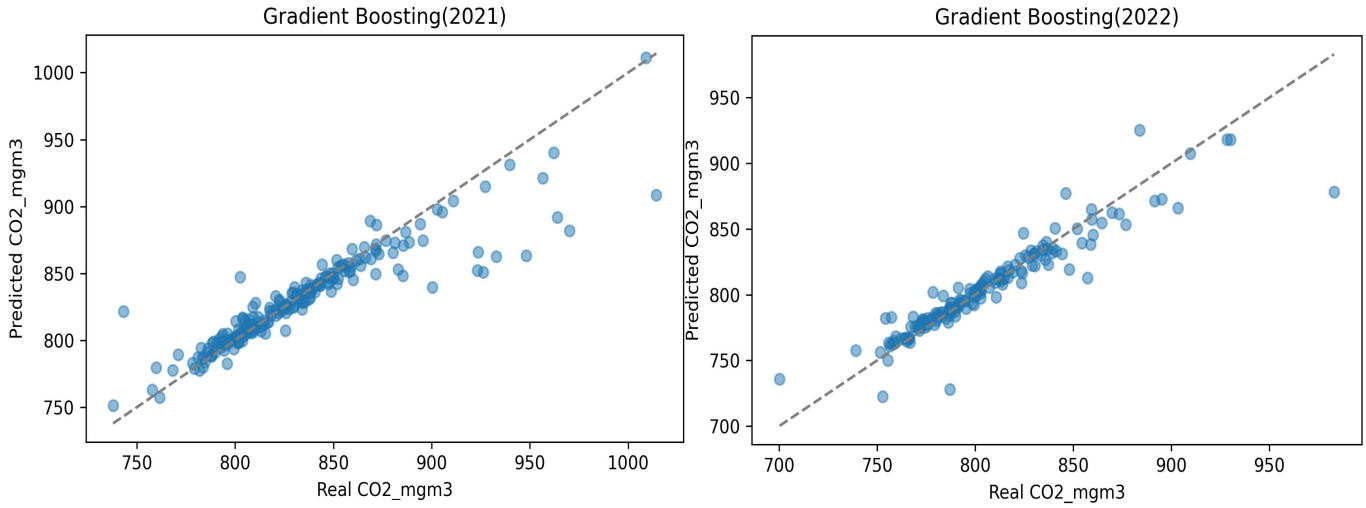
Pollutant Prediction Modeling

We compared the performance of 8 regression models (Linear Regression, Ridge, Lasso, Random Forest, Gradient Boosting, XGBoost, KNN, MLP) on 13 pollutant prediction tasks, and used R^2 as the core evaluation indicator. The Gradient Boosting Regressor model was initially selected to perform best on the pollutant CO2_mgm3, with an R^2 of 0.321, and the prediction trend was consistent with the true value.

To improve the performance of the model, we added lag features, rolling averages, and time dimension information. And the parameters were adjusted through RandomizedSearchCV. Independent modeling was performed on the subsets of 2021 and 2022 respectively. The results show that the respective models performed well in the two years. The 2021 model has R^2 = 0.819, RMSE = 19.00, MAE = 9.24, MSE = 360.90, and the 2022 model has R^2 = 0.870, RMSE = 14.18, MAE = 7.69, MSE = 201.17.

Figure 3

Gradient Boosting for 2021 and 2022



This phenomenon indicates that there is a distribution drift between pollutant concentrations and their influencing factors across years, and a single model is difficult to capture the complex trend of changes over many years at the same time. This result emphasizes that time sensitivity should be considered when modeling non-stationary data.

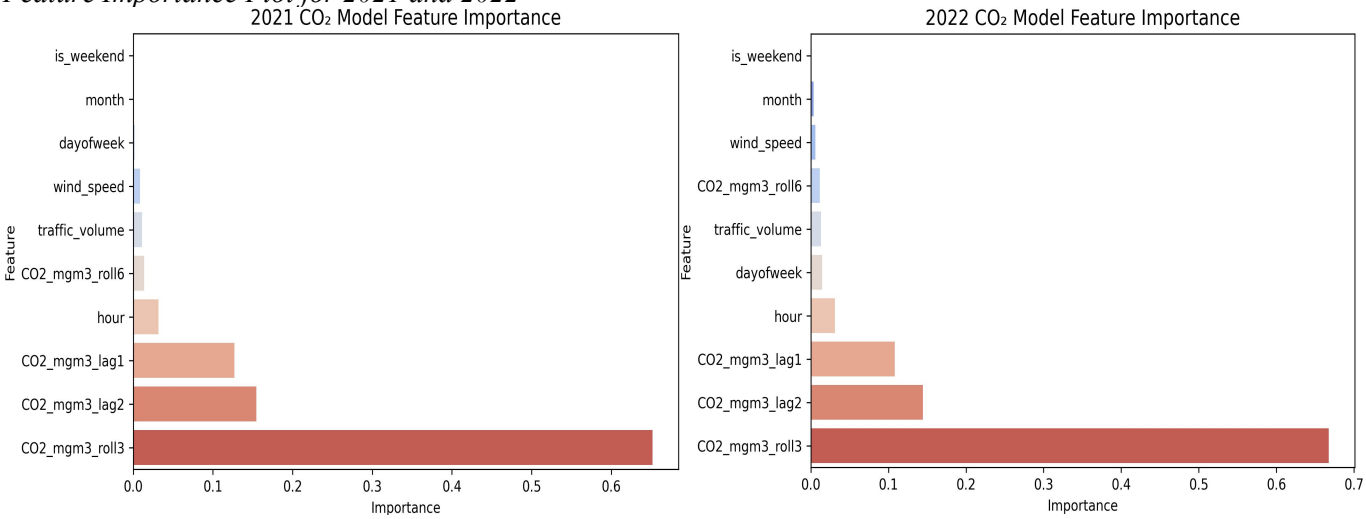
Feature Importance

In the feature importance analysis of the model, the rolling average feature CO2_mgm3_roll3 is the most important for pollutant concentration prediction. This shows that the pollutant concentration prediction is mainly affected by the fluctuation characteristics of the historical pollution concentration, indicating that it has strong temporal continuity.

Besides, the wind speed and traffic flow contribute less to the model prediction, and the contribution of traffic flow is slightly higher than wind speed. The low importance of traffic flow and wind speed in the model may be related to the location of the measurement equipment and the contribution of non-motor vehicles from CO_2 sources.

Figure 4

Feature Importance Plot for 2021 and 2022

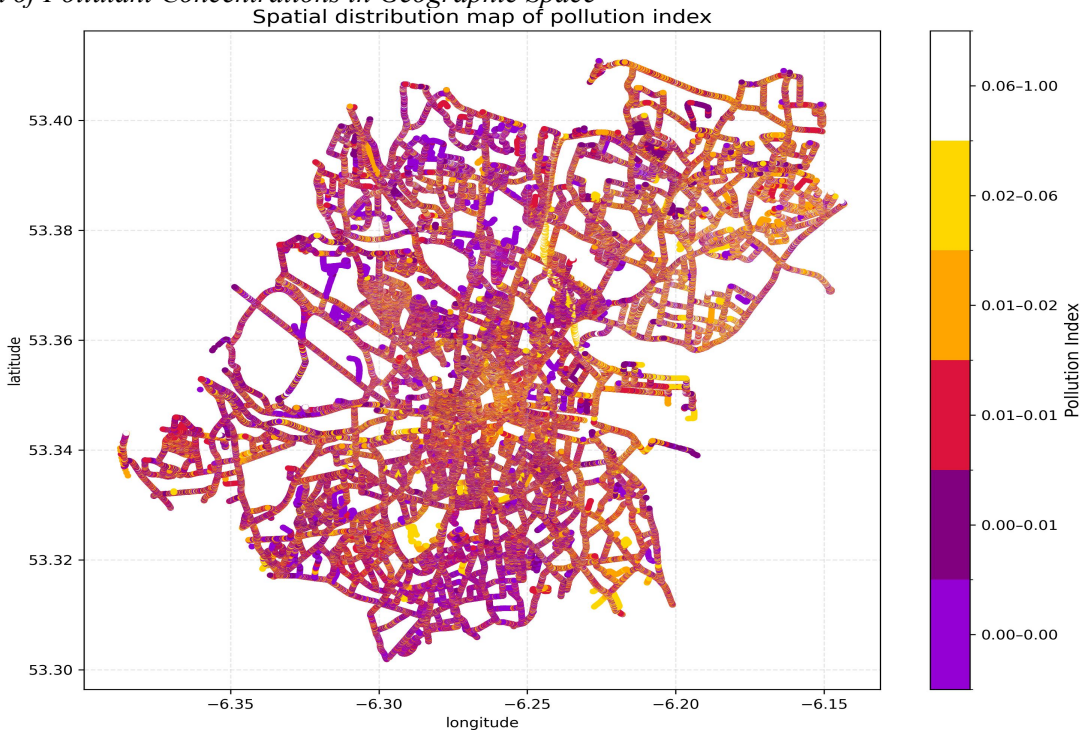


Geographical Spatial Distribution Analysis

Since the original pollutant data was collected from Google mobile monitoring devices, we used the PCA dimensionality reduction method to conduct a comprehensive analysis of all pollutant concentrations and normalized them by coordinate points to reflect the spatial comprehensive distribution characteristics of pollutants in different areas of Dublin.

Figure 5

Distribution of Pollutant Concentrations in Geographic Space



From the figure, it can be observed that the pollution level is higher in the city center and some

traffic-intensive areas (latitude and longitude are approximately 53.34 north latitude and longitude -6.26), while the concentration is lower in the outer areas. This spatial distribution is highly consistent with the actual traffic routes and urban structure.

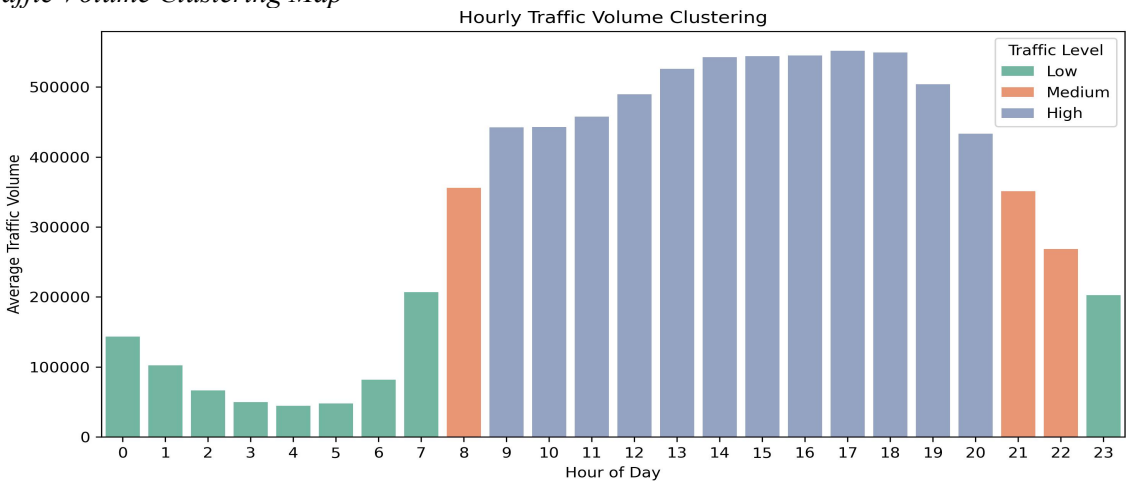
Cluster Analysis and Pollution Differences Based on Traffic Flow

Traffic Volume Cluster Analysis

In order to analyze the impact of traffic intensity on pollutant concentration, we used KMeans to perform clustering based on the hourly average traffic volume, dividing it into three flow levels: peak, medium, and low. The results showed that from 23:00 to the next day on July 7th, the traffic was low-peak, from 9:00 to 20:00, it was peak, and the rest were medium.

Figure 6

Hourly Traffic Volume Clustering Map

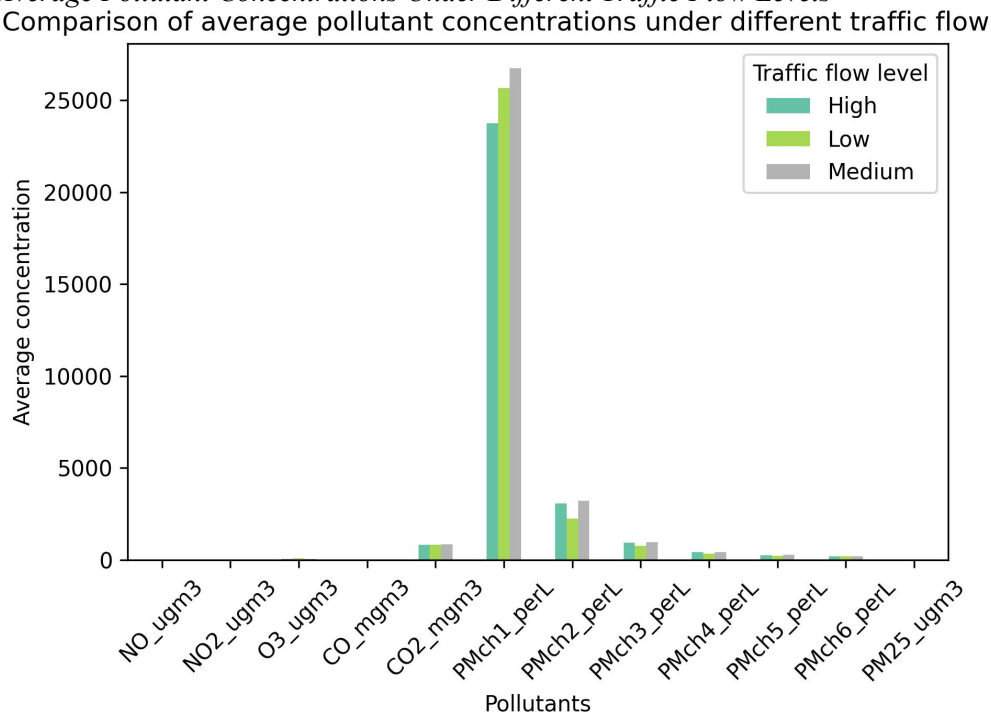


Pollution Concentration Difference Analysis

We further calculated the average pollutant concentrations under different traffic flow levels and found that the pollutant concentrations corresponding to medium flow are generally higher than those corresponding to high and low flow, indicating the existence of a nonlinear pollution mechanism.

Figure 7

Comparison of Average Pollutant Concentrations Under Different Traffic Flow Levels



Conclusion

This project comprehensively applied data preprocessing, visualization analysis, correlation analysis, machine learning modeling and spatial clustering methods to fully reveal the relationship between pollutants and traffic and wind speed in Dublin. In the end, we successfully established a high-precision prediction model for CO_2 concentration. And through spatial and time period analysis, we further assisted in understanding the distribution mechanism of pollutants. Therefore, this study provides a valuable reference for environmental governance and urban planning.