Urban Air Pollution Structures and Their Social-Environmental Drivers: A Multivariate Analysis Based on PCA and Machine Learning

Introduction

Air pollution is an important public environmental problem faced by major cities around the world (Mayer, 1999). It not only affects the health of residents, but also poses a challenge to the sustainable development of the city (Amegah & Jaakkola, 2016). Dublin, as a city with a dense population, busy traffic and variable meteorological conditions, shows obvious spatiotemporal heterogeneity in the changes of pollutants. In order to understand the underlying patterns behind pollution changes more effectively, this paper attempts to combine multi-source data and explore the pollution structure by using dimensionality reduction and machine learning methods. And evaluate the predictive ability of meteorological and travel factors for the pollution status.

This paper first extracts the potential dimensions of the pollution structure through principal component analysis. Then, the pollution status types are identified through clustering. Finally, the gradient boosting regression model is applied to predict the pollution structure dimension and evaluate the explanatory power of external variables. The research finds that factors such as temperature, wind speed, humidity and traffic volume have certain predictive effects on the pollution structure. This research provides a data basis and methodological support for urban pollution monitoring and intelligent early warning.

Literature Review

Traditional air pollution analysis mostly focuses on a single pollutant. However, in real urban environments, multiple pollutants often co-occur and have complex correlation structures (Pires et al., 2008). Therefore, in recent years, more and more studies have used multivariate methods such as principal component analysis to identify pollution sources and potential structures (Lu et al., 2011). In addition, meteorological conditions such as temperature, humidity, and wind speed play a key role in the accumulation and diffusion of pollution (He et al., 2017). Traffic flow and pedestrian flow are the main driving forces of pollutant emissions (Forehead & Huynh, 2018).

In recent years, machine learning models have been widely used in air quality prediction, especially in the context of integrating multi-source external features (Zimmerman, 2018). However, most studies still model pollutants as independent indicators, and rarely take the potential changes in pollution structure as the research object. Based on existing research, this paper attempts to establish an integrated framework of "dimensionality reduction-clustering-prediction", which not only focuses on the abstract structure of the pollution dimension, but also introduces external variables into the prediction model for explanation, thus expanding the research perspective of pollution modeling.

Dataset Description

This study uses multi-source hourly data for Dublin from 2021 to 2022.

Pollution data include 12 indicators from the Google AirView project (e.g., NO, NO_2, O_3, $PM_{2.5}$), with values showing high variability (e.g., NO: mean = 22.6 μg/m³, max = 468 μg/m³).

Meteorological data from Met Éireann include temperature, wind speed, and humidity, with average values of 13.3 ° C, 9.8 knot, and 74%, respectively.

Traffic data are from SCATS, and pedestrian flow counts are provided by Dublin City Council. Both show strong temporal fluctuation, with foot traffic ranging from 21,000 to over 150,000 per hour.

All data are aggregated by hour and aligned with timestamps. Negative values (0.3% records) were set to 0, assuming sensor noise, and missing values (2%) were estimated by linear interpolation. The final merged dataset contains ~1,900 hourly records.

Table 1

Processed Variable

Type	Time Dimension	Sample size	variable examples
Air pollution	Hour	~1900	NO、NO_2 、PM2.5、CO、O_3
Weather data	Hour	~1900	temperature、wind_speed、humidity
Traffic flow	Hour	~1900	traffic_volume
Pedestrian flow	Hour	~1900	total_foot_traffic

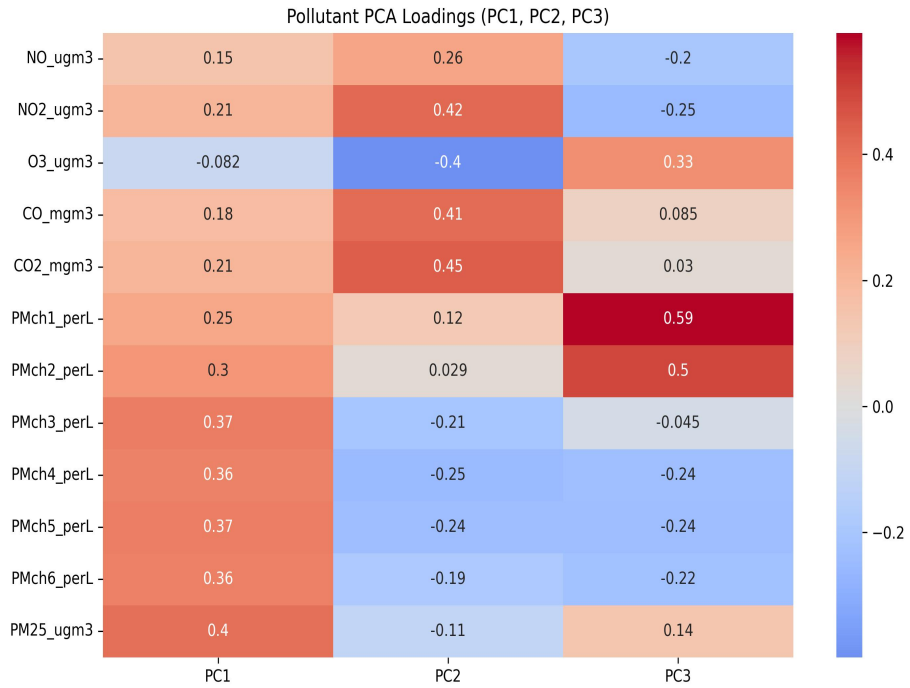
Principal Component Analysis

Correlation Analysis Between Variables And Principal Components

After standardizing the pollutant variables, I used PCA to extract the latent dimensions. The first three principal components explained 47.4%, 19.3%, and 11.9% of the variance, respectively.

Figure 1

Variable Load Heat Map



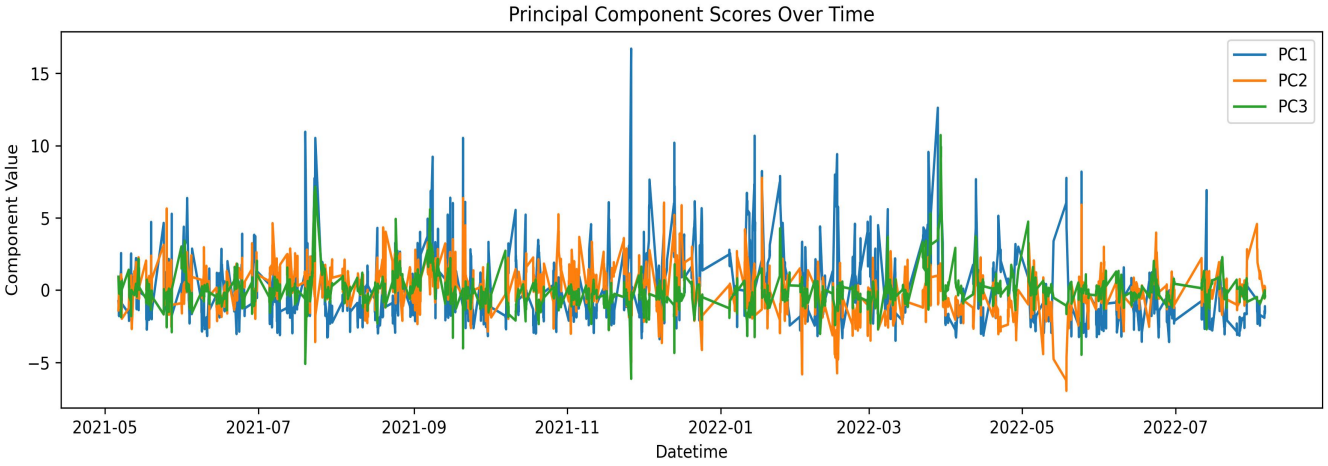
The variable load heat map above shows that PC1 is strongly positively correlated with PMch1 – PMch6 and PM2.5, indicating the intensity of particulate matter pollution. And PC2 mainly reflects the changes in NO_2 , CO_2 , and CO, representing the gas pollution dimension. The PC3 is correlated with O_3 and some light particulate matter, explaining weaker heterogeneity.

Principal Component Time Series Analysis

Furthermore, I analyzed the fluctuation law of the pollution structure over time through the principal component time series diagram. PC1 shows the most intense fluctuation characteristics, indicating that the pollution structure it represents is significantly affected by spatio-temporal factors and external shocks, especially with high frequencies and high values in winter. The fluctuation of PC2 is relatively stable and shows a certain periodicity. Combined with its load structure, it can be regarded as a long-term difference in the structure of pollution sources. The fluctuation range of PC3 is the smallest but contains multiple peaks, and it slightly increases in summer. It may represent a local or short-term pollution incident.

Figure 2

Principal Component Time Series Plot



PCA assumes that the data is linear, which is more suitable for this case after standardization. It is also widely used in pollution structure analysis.

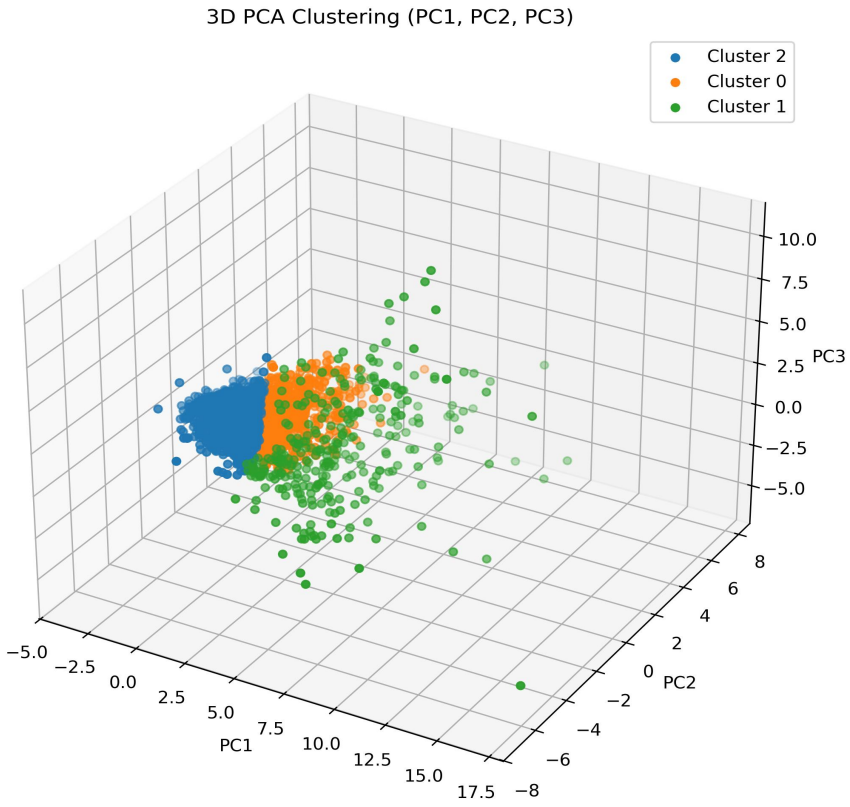
Cluster Analysis

Principal Component Clustering Structure and Pollution Characteristics

In order to further identify the type of pollution structure, this paper constructs a space based on the first three principal components. The KMeans clustering method (k=3) is applied to perform unsupervised classification of pollution samples. Since the principal components have compressed the main information of the original pollution variables, this space can more accurately reflect the differences and internal connections of pollution structures (Sinaga & Yang, 2020 ; Boutsidis et al., 2009). K-means assumes that the clustering results are balanced and spherical, which is very effective for identifying potential pollution types.

Figure 3

Principal Component Clustering Plot



According to the cluster diagram, Cluster 2 is concentrated in the low principal component score range, representing the clean period with low pollutant concentration levels. Cluster 0 is located in the middle area, reflecting the typical structure of daily urban pollution. Cluster 1 is widely distributed and has high principal component scores, indicating that it may correspond to a pollution event or a high pollution structure under high emission conditions.

In addition, I calculated the mean of the pollutants after clustering and drew a bar chart to show the differences in pollutants under each cluster.

Table 2

Average Concentration of Each Pollutant in Different Clusters(Part 1)

Cluster	NO (μg/m³)	NO_2 (μg/m³)	O_3 (μg/m³)	CO (mg/m³)	CO_2 (mg/m³)	PMch1
0	28.66	30.85	37.84	0.46	839.77	26662.47
1	36.30	30.52	45.99	0.47	841.86	54605.66
2	15.04	9.01	55.49	0.38	798.09	12840.80

Table 3

Average Concentration of Each Pollutant in Different Clusters(Part 2)

Cluster	PMch2	PMch3	PMch4	PMch5	PMch6	PM2.5 (μg/m³)
0	3090.07	892.35	399.54	249.92	187.20	8.23
1	7472.58	2037.28	872.75	538.34	396.88	17.76
2	1700.29	613.79	281.82	176.61	127.60	5.20

Figure 4

Pollutant Cluster Mean Histogram

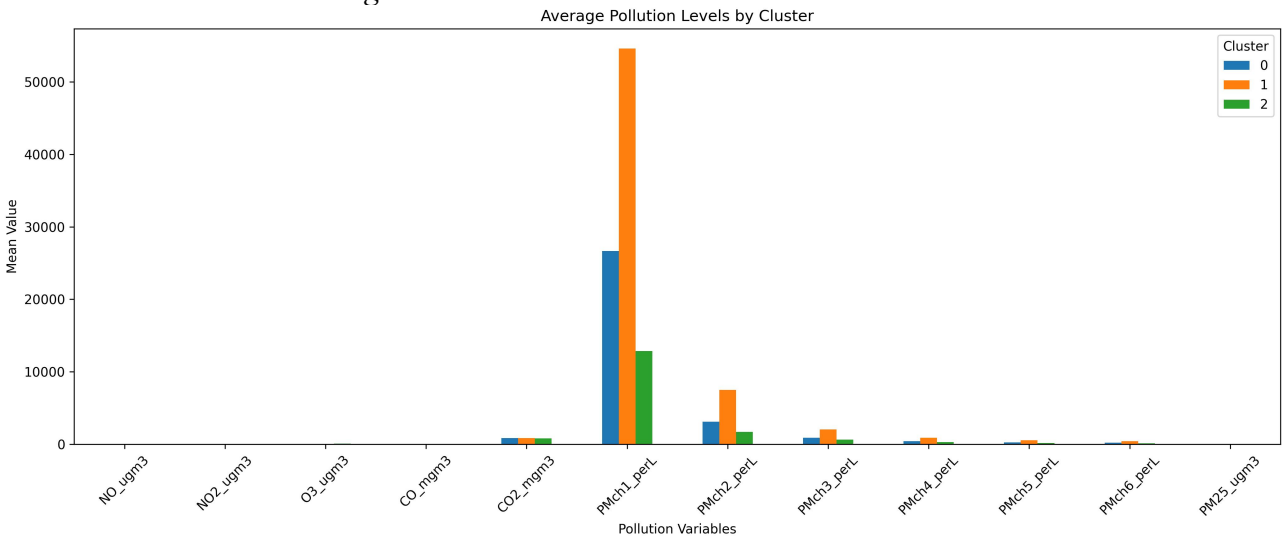


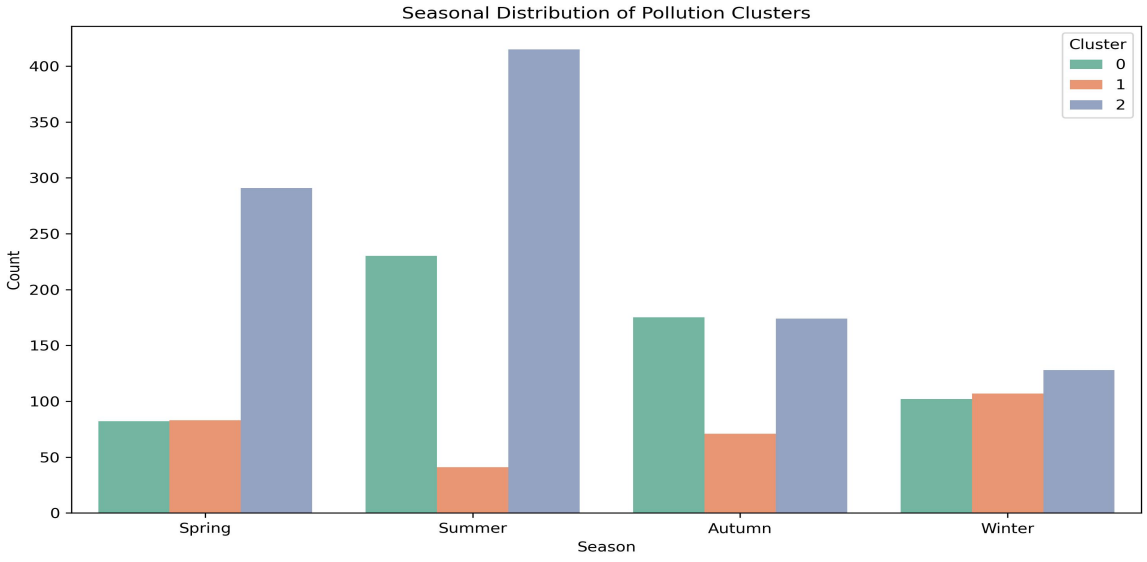
Figure 4 and Tables 2 & 3 show the average levels of the three types of pollution structures in the dimensions of specific pollutants. Cluster 1 corresponds to high concentrations of PM and NO series emissions, representing highly polluting events. Cluster 0 presents as a typical urban background pollution structure, with moderate concentrations of all items. While Cluster 2 shows the lowest concentration of gaseous pollutants and a higher level of ozone. This structural division is helpful for understanding the manifestation patterns of pollution source combinations and pollution intensities in actual time and space.

Seasonality Analysis

This paper also conducts seasonal distribution statistical analysis on the clustering results to further understand the changing trend of pollution structure under different climatic conditions.

Figure 5

Seasonal Distribution of Pollution Clusters



As shown in the figure, the occurrence frequencies of the three types of pollution structures vary significantly throughout the four seasons. Among them, Cluster 2 has the highest occurrence

frequency. Especially in summer and spring, there are obvious peaks, reaching more than 400 times respectively. By contrast, Cluster 1 has the lowest occurrence frequency in summer and slightly increases in winter. The distribution of Cluster 0 is relatively even, but it reaches a relative peak in summer.

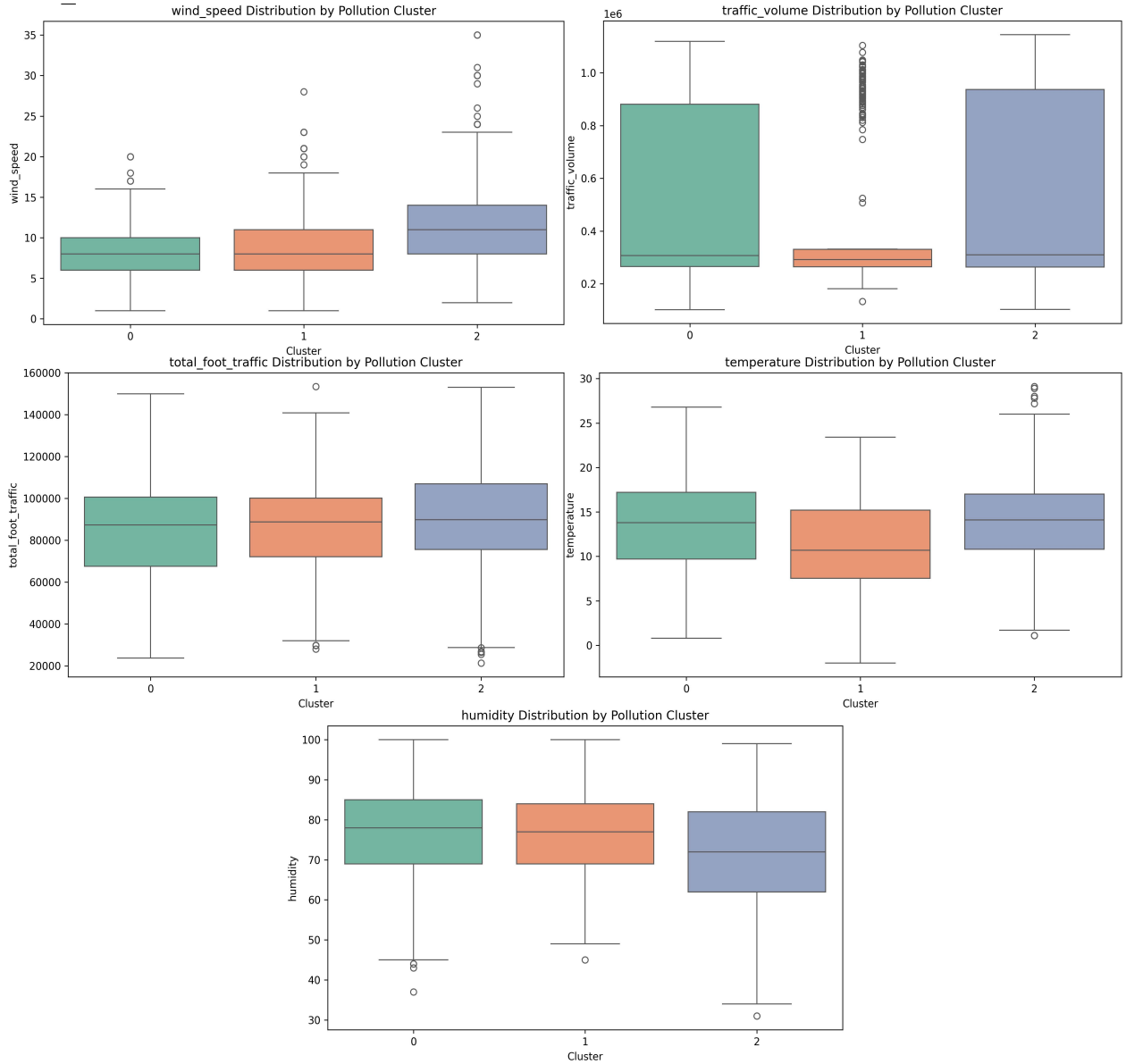
This result further verified the seasonal characteristics of the pollution structure. It also indicates that dimension reduction and cluster analysis can effectively identify pollution patterns driven by climatic conditions.

External variable analysis

In order to further understand the causes of pollution structure and its relationship with meteorological and human activities, I plotted box plots of various structures on five variables: temperature, wind speed, humidity, traffic flow, and pedestrian flow.

Figure 6

Cluster_External Variable Box Plot



The results show that Cluster 2 is generally higher in wind speed and temperature. It indicates that favorable meteorological diffusion conditions are conducive to the formation of structures with lower pollution loads. In contrast, Cluster 1 is lower in both temperature and wind speed. Cluster 0, as a moderately polluted structure, has an external variable distribution that is basically centered. This indicates that the pollution structure is significantly influenced by external meteorological and travel conditions and has obvious temporal characteristics.

OLS Regression Analysis

I continue to build an OLS regression model. The results show that the PC1 can be partially explained, with a model R^2 of 11.1%, which is significantly affected by temperature, wind speed, traffic and pedestrian flow, representing changes in pollution intensity. The PC2 has a strong explanatory power (R^2 = 27.7%), which is dominated by temperature, humidity and wind speed. It reflects the important role of meteorological conditions in the formation of gas pollution structure (such as NO_2 vs O_3). The PC3 represents local pollution events, which are significantly affected by temperature and humidity. But the overall explanatory power is weak (R^2 = 7.5%). This shows that its causes are more complicated, and there are still a large number of unobservable nonlinear or time-lag effects.

Figure 7

PC1 Regression Results

OLS Regression Results

Dep. Variable:	PC1	R-squared:	0.111
Model:	OLS	Adj. R-squared:	0.109
Method:	Least Squares	F-statistic:	47.46
Date:	Sat, 19 Apr 2025	Prob (F-statistic):	2.37e-46
Time:	13:29:27	Log-Likelihood:	-4233.5
No. Observations:	1899	AIC:	8479.
Df Residuals:	1893	BIC:	8512.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.4252	0.536	8.253	0.000	3.374	5.477
temperature	-0.1048	0.013	-8.123	0.000	-0.130	-0.079
wind_speed	-0.1426	0.013	-11.134	0.000	-0.168	-0.117
humidity	-0.0063	0.005	-1.381	0.167	-0.015	0.003
traffic_volume	-7.649e-07	1.86e-07	-4.109	0.000	-1.13e-06	-4e-07
total_foot_traffic	-8.581e-06	2.32e-06	-3.701	0.000	-1.31e-05	-4.03e-06

Omnibus:	621.495	Durbin-Watson:	0.674
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2262.215
Skew:	1.589	Prob(JB):	0.00
Kurtosis:	7.301	Cond. No.	6.59e+06

Figure 8

PC2 Regression Results

OLS Regression Results

Dep. Variable:	PC2	R-squared:	0.277
Model:	OLS	Adj. R-squared:	0.276
Method:	Least Squares	F-statistic:	145.4
Date:	Sat, 19 Apr 2025	Prob (F-statistic):	8.46e-131
Time:	13:29:27	Log-Likelihood:	-3183.7
No. Observations:	1899	AIC:	6379.
Df Residuals:	1893	BIC:	6413.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.8434	0.309	-5.975	0.000	-2.448	-1.238
temperature	0.0470	0.007	6.332	0.000	0.032	0.062
wind_speed	-0.1360	0.007	-18.456	0.000	-0.150	-0.122
humidity	0.0373	0.003	14.226	0.000	0.032	0.042
traffic_volume	-7.644e-08	1.07e-07	-0.714	0.476	-2.86e-07	1.34e-07
total_foot_traffic	-1.943e-06	1.33e-06	-1.456	0.145	-4.56e-06	6.73e-07

Omnibus:	123.312	Durbin-Watson:	0.977
Prob(Omnibus):	0.000	Jarque-Bera (JB):	547.308
Skew:	-0.067	Prob(JB):	1.42e-119
Kurtosis:	5.627	Cond. No.	6.59e+06

Figure 9

PC3 Regression Results

OLS Regression Results

Dep. Variable:	PC3	R-squared:	0.075
Model:	OLS	Adj. R-squared:	0.073
Method:	Least Squares	F-statistic:	30.69
Date:	Sat, 19 Apr 2025	Prob (F-statistic):	4.17e-30
Time:	13:29:27	Log-Likelihood:	-2963.7
No. Observations:	1899	AIC:	5939.
Df Residuals:	1893	BIC:	5973.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.6481	0.275	-5.998	0.000	-2.187	-1.109
temperature	0.0707	0.007	10.694	0.000	0.058	0.084
wind_speed	-0.0046	0.007	-0.694	0.488	-0.017	0.008
humidity	0.0129	0.002	5.505	0.000	0.008	0.017
traffic_volume	2.501e-08	9.54e-08	0.262	0.793	-1.62e-07	2.12e-07
total_foot_traffic	-2.455e-06	1.19e-06	-2.066	0.039	-4.79e-06	-1.25e-07

Omnibus:	1139.063	Durbin-Watson:	0.499
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23236.502
Skew:	2.423	Prob(JB):	0.00
Kurtosis:	19.438	Cond. No.	6.59e+06

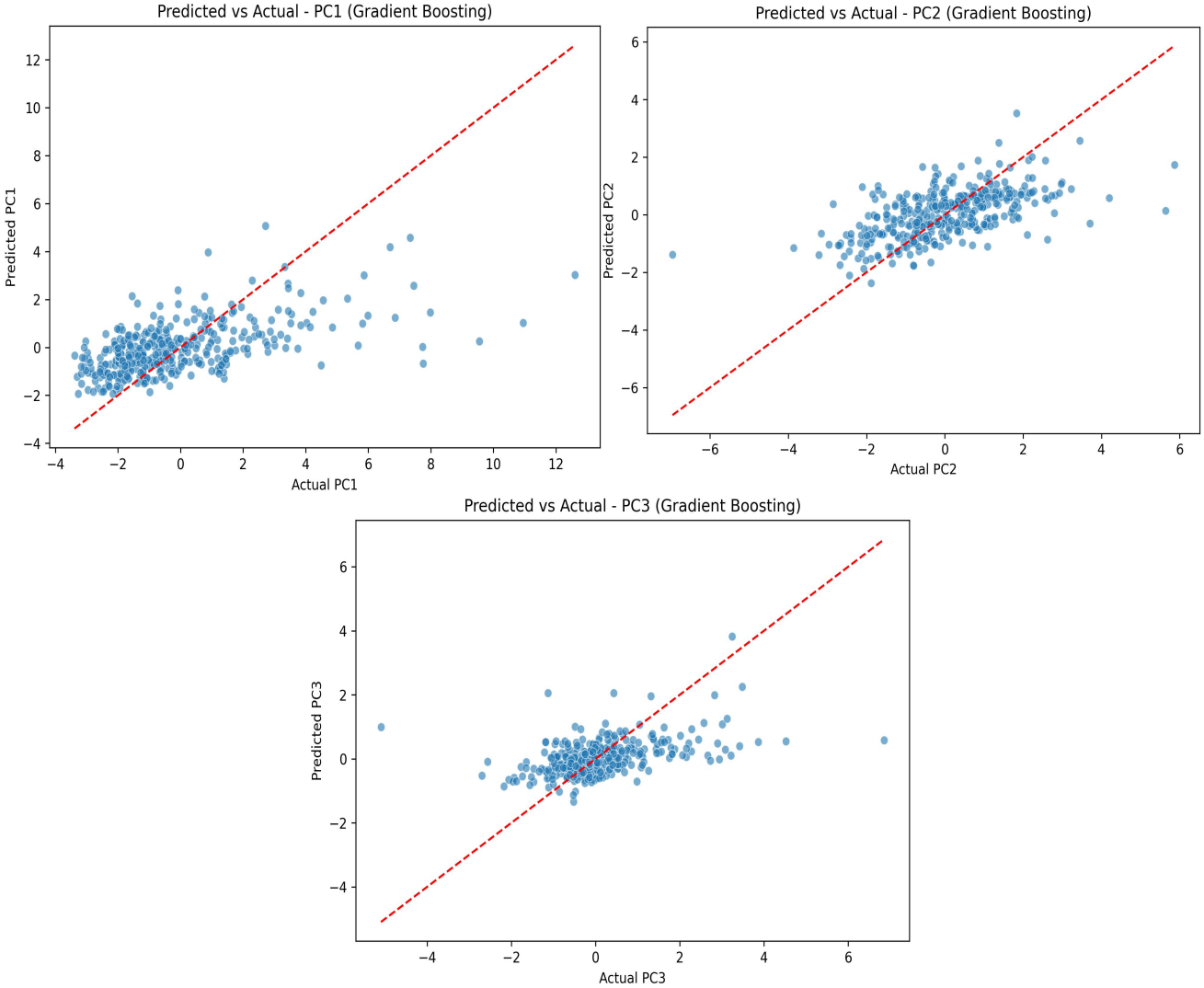
Principal Component Prediction Analysis

OLS provides interpretability, while gradient boosting could captures nonlinear patterns and variable interactions. Both are suitable for datasets with mixed features. I further used Gradient Boosting Regressor to perform predictive modeling on PC1, PC2, and PC3. And optimized the parameters through GridSearchCV. The three Gradient Boosting models all chose the same optimal parameter combination, namely learning_rate = 0.01, max_depth = 5, n_estimators = 300, subsample = 0.8.

This result shows that under the current feature space and sample distribution conditions (Ranjanet al., 2019), this group of parameters has good generalization ability. It is capable of stably capturing the response patterns of external variables to different pollution structure dimensions.

Figure 10

Predicted Values vs Actual Values for PC1, PC2, and PC3



From the perspective of predictive performance, the model of PC2 performs the best (R^2 = 0.3305), indicating that the structure of gaseous pollution is more closely related to meteorological and human flow and traffic variables. The PC1 model came second (R^2 = 0.3095), and the particulate

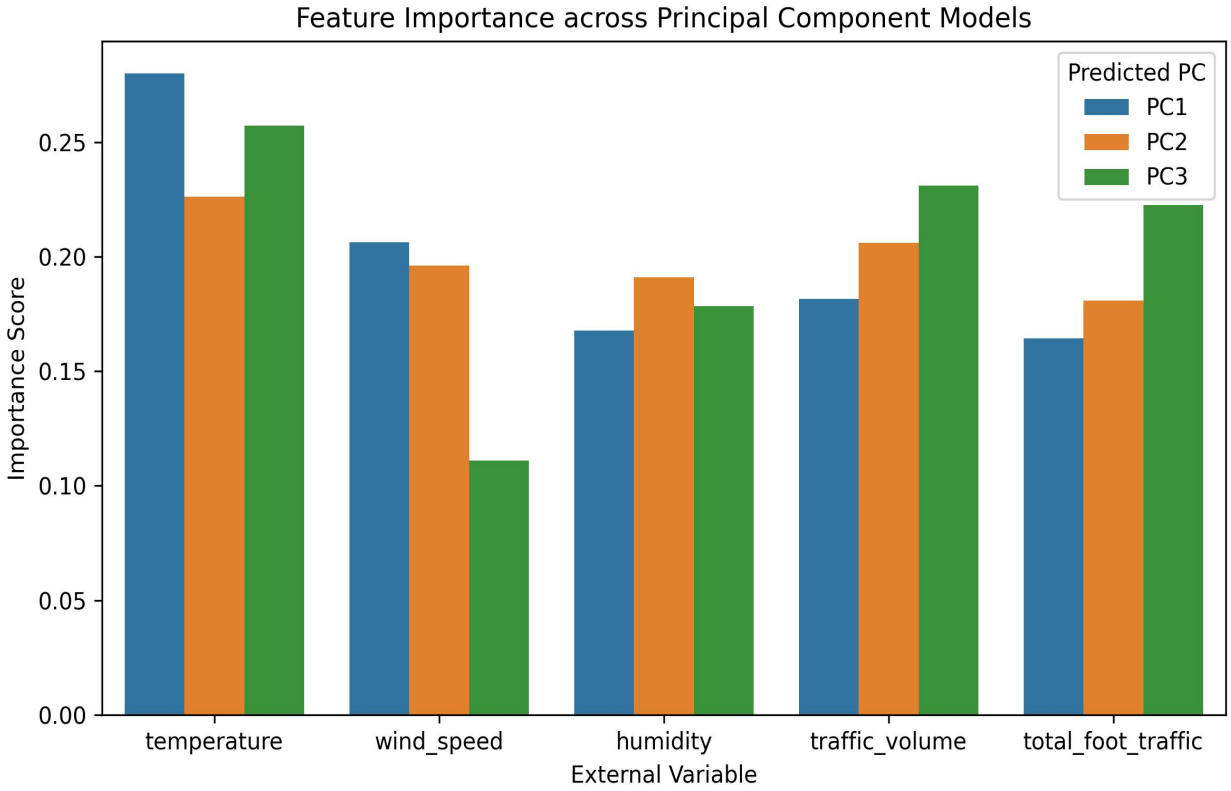
matter concentration was jointly driven by meteorological diffusion and the superposition of emissions. The PC3 model has a relatively weak predictive ability (R^2 = 0.2081), indicating that the changes in the pollution structure it represents are more complex or highly volatile, and are difficult to be fully characterized by macro variables.

Feature Importance Analysis

After completing the prediction model, this paper compares the importance of features in the three principal component prediction models.

Figure 10

Feature Importance Scores for PC1, PC2, and PC3 in the Prediction



The figure shows that temperature is the most explanatory variable in all three models. In particular, the importance score is as high as 0.28 when predicting PC1. The importance of wind speed is similar in PC1 and PC2, about 0.20, but it drops significantly in PC3. In addition, the importance of pedestrian flow in PC3 has increased significantly, reaching 0.23. Traffic flow has a stable contribution in the three principal components, especially in PC2, with a score close to 0.21, which strengthens the close relationship between pollution structure and traffic emissions.

Conclusion

This paper constructs a systematic pollution structure identification and interpretation analysis framework based on weather, traffic and pollution data of Dublin City in 2021-2022. I extracted three types of pollutants using the PCA method and divided their structure using the KMeans clustering method. The analysis shows that different pollution structures show significant differences in pollutant

composition, external meteorological and traffic factors and seasonal distribution. It reflect the dynamic and temporal characteristics of pollution structure. Besides, the machine learning model based on Gradient Boosting is further used to predict the principal component scores respectively. Although the prediction performance is at a medium level (R^2 is between 0.21-0.33), it still reflects the importance of variables such as temperature, wind speed, and traffic volume in explaining the changes in pollution structure. Finally, feature importance analysis reveals that there are differences in the focus of each principal component affected by external variables.

However, this paper still has certain research limitations. The explanatory power of the model is limited, and the current variables cannot fully capture the mechanisms behind the pollution structure. Second, the number of input variables is relatively limited, and key meteorological factors such as wind direction, precipitation, and inversion have not been considered. In addition, pollutant data do not contain spatial location characteristics and lack geographical difference analysis, which limits the further identification and interpretation of spatial pollution structure.

References

Amegah, A. K., & Jaakkola, J. J. (2016). Household air pollution and the sustainable development goals. Bulletin of the World Health Organization, 94(3), 215. 10.2471/BLT.15.155812

Boutsidis, C., Drineas, P., & Mahoney, M. W. (2009). Unsupervised feature selection for the k - means clustering problem. Advances in neural information processing systems, 22. https://proceedings.neurips.cc/paper_files/paper/2009/file/c51ce410c124a10e0db5e4b97fc2af39-Paper.pdf

Forehead, H., & Huynh, N. (2018). Review of modelling air pollution from traffic at street-level-The state of the science. *Environmental Pollution*, *241*, 775-786. https://doi.org/10.1016/j.envpol.2018.06.019

He, J., Gong, S., Yu, Y., Yu, L., Wu, L., Mao, H., ... & Li, R. (2017). Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major Chinese cities. *Environmental pollution*, *223*, 484-496. https://doi.org/10.1016/j.envpol.2017.01.050

Lu, W. Z., He, H. D., & Dong, L. Y. (2011). Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. *Building and Environment*, *46*(3), 577-583. https://doi.org/10.1016/j.buildenv.2010.09.004

Mayer, H. (1999). Air pollution in cities. Atmospheric environment, 33(24-25), 4029-4037. https://doi.org/10.1016/S1352-2310(99)00144-2

Pires, J. C. M., Sousa, S. I. V., Pereira, M. C., Alvim-Ferraz, M. C. M., & Martins, F. G. (2008). Management of air quality monitoring using principal component and cluster analysis—Part I: SO2 and PM10. *Atmospheric Environment*, *42*(6), 1249-1260. https://doi.org/10.1016/j.atmosenv.2007.10.044

Ranjan, G. S. K., Verma, A. K., & Radhika, S. (2019, March). K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In 2019 IEEE 5th international conference for convergence in technology (I2CT) (pp. 1-5). IEEE. 10.1109/I2CT45611.2019.9033691

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727. 10.1109/ACCESS.2020.2988796

Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., & Robinson, A. L. (2018). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, *11*(1), 291-313. https://doi.org/10.5194/amt-11-291-2018