International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Temporal and Cross-Sensor Analysis of Urban Air Pollution Using Multivariate Environmental Data

Ms.Priya Shankar Adhalrao[1], Mrs. Shubhangi Pratik Bomble[2]

[1]Student, MCA, JCEI's Jaihind Institute of Management and Research, Vadgaon Sahani
[2]Assistant Professor, Department of Master of Computer Application, JCEI's Jaihind Institute of Management and Research, Vadgaon Sahani

Abstract

Urban air pollution is not controlled by one variable; pollutant concentrations, sensor responses, temperature and humidity change together over time. This study presents a practical analysis of the UCI Air Quality dataset using a multivariate exploratory approach. The supplied AirQualityUCI.csv file was cleaned by treating the code −200 as missing and by excluding trailing blank records. The analysis used descriptive statistics, monthly aggregation, Pearson correlation, environmental relationships and cross-sensor comparison. The file contained 9,357 populated observations across 15 research variables. Strong relationships were found between benzene and PT08.S2(NMHC) (r = 0.982), CO and benzene (r = 0.931), and CO and PT08.S1(CO) (r = 0.879). NOx and PT08.S3(NOx) showed a moderate negative relationship (r = −0.656), while NO2 and PT08.S4(NO2) showed only a weak relationship (r = 0.157). Monthly results also showed that different pollutants reached their highest mean values in different months. The study highlights useful patterns in the dataset while also showing why missing values, sensor cross-sensitivity and drift must be considered when interpreting air-quality measurements.

Keywords: Air Quality; Air Pollution; Multivariate Analysis; Time-Series Analysis; Chemical Sensors; Correlation Analysis; Environmental Data; Data Visualization

I. INTRODUCTION

Air-quality monitoring produces data from several pollutants and environmental measurements at the same time. Looking at one pollutant in isolation can hide relationships that are useful for understanding urban pollution. Time-based analysis is therefore important because pollutant levels and sensor responses can change across hours, months and environmental conditions.

The UCI Air Quality dataset combines reference measurements for CO, non-methanic hydrocarbons, benzene, NOx and NO2 with responses from five metal-oxide sensors and measurements of temperature, relative humidity and absolute humidity [1]. The original work using this data also examined field calibration of an electronic nose for benzene estimation, showing the importance of relationships between sensor responses and reference measurements [2].

This paper focuses on an interpretable analysis rather than building a website, mobile application or real-time monitoring system. The purpose is to understand what the supplied data itself shows and to present the results in a clear and reproducible form.

II. PROBLEM STATEMENT

Urban air pollution is a multivariate problem. Several pollutants can rise or fall at different rates, while temperature and humidity may also change at the same time. Chemical sensors add another layer of complexity because their responses can be influenced by more than the pollutant they are designed to measure. Missing observations and sensor drift can further affect the reliability of direct comparisons.

There is therefore a need for a simple, transparent analysis that brings temporal behaviour, pollutant relationships, environmental variables and sensor-reference relationships together. A project-level study of this kind can show which patterns are strong enough to be useful and which measurements need more careful interpretation.

III. OBJECTIVES

- To analyze temporal variations in major air pollutants.
- To identify correlations between major pollutant concentrations.
- To study relationships between temperature, relative humidity and absolute humidity and pollution levels.
- To examine relationships between chemical sensor responses and reference pollutant measurements.
- To identify important trends, data-quality issues and limitations in the dataset.
- To present the findings using statistical analysis and clear data visualizations.

IV. RESEARCH QUESTIONS

- What temporal patterns are visible in the major pollutant measurements?
- Which pollutant variables show the strongest relationships with one another?
- How are temperature and humidity associated with pollutant concentrations?
- How closely do chemical sensor responses relate to their corresponding reference measurements?
- What data-quality and sensor-related limitations should be considered when interpreting the results?

V. LITERATURE REVIEW

Earlier work on the underlying monitoring data has mainly focused on sensor calibration, prediction and time-series modelling. De Vito et al. studied field calibration of an electronic nose for benzene estimation in an urban monitoring setting and showed the practical importance of sensor/reference relationships [2]. Kong et al. demonstrated the value of multivariate time-series methods for air-pollution prediction, reinforcing the importance of preserving temporal structure [3]. Ali et al. later found that temporal information can improve machine-learning-based calibration of low-cost air-quality sensors [4].

The present study does not attempt to replace these methods. Instead, it uses a simpler exploratory workflow to make the main relationships in the dataset easier to understand before any predictive or calibration model is considered.

Study / Source	Focus	Key contribution	Relation to this study
De Vito et al. (2008)	Electronic-nose calibration	Benzene estimation and sensor/reference calibration	Supports cross-sensor analysis [2]

UCI Air Quality dataset	Multivariate monitoring data	Reference pollutants, five sensors and environmental variables	Primary data source [1]
Kong et al. (2021)	Multivariate air-pollution forecasting	Uses time-series structure for pollution prediction	Supports temporal analysis [3]
Ali et al. (2024)	Low-cost sensor calibration	Temporal information can improve calibration	Motivates time-aware sensor analysis [4]

VI. RESEARCH GAP

Existing work has examined the dataset and related air-quality datasets for calibration, forecasting and machine-learning applications. These approaches are useful when the goal is prediction or sensor correction, but a project-level exploratory view can still be valuable. The gap addressed here is not the absence of air-quality research; it is the need for a compact and transparent workflow that combines temporal trends, pollutant-to-pollutant correlations, environmental associations and sensor-reference relationships in one analysis.

The study therefore emphasizes interpretation rather than maximum predictive accuracy. This makes the findings easier to reproduce and easier to explain in an academic project setting.

VII. DATASET AND DATA QUALITY

The official UCI record identifies the Air Quality dataset as a multivariate time-series dataset with 9,358 reported instances and 15 features. It contains hourly measurements from five metal-oxide sensors together with reference pollutant concentrations and environmental variables [1]. The supplied CSV used in this study contains 9,357 populated observations after trailing blank records were excluded. All calculations reported here use the supplied file.

Variable group	Variables	Purpose
Reference pollutants	CO(GT), NMHC(GT), C6H6(GT), NOx(GT), NO2(GT)	Reference concentration measurements
Chemical sensors	PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3)	Sensor response measurements
Environmental	T, RH, AH	Temperature and humidity context
Temporal	Date, Time	Time-based analysis

Data-quality treatment

- The semicolon-separated CSV and comma-decimal numeric format were parsed correctly.
- The value −200 was treated as missing/invalid rather than as a real measurement.
- Trailing blank records were excluded before calculating results.
- Date and Time were combined into one datetime field for temporal analysis.

- Reported correlations use pairwise complete observations; no arbitrary value substitution was used.

Variable	Missing / invalid	Percentage
NMHC(GT)	8,443	90.23%
CO(GT)	1,683	17.99%
NO2(GT)	1,642	17.55%
NOx(GT)	1,639	17.52%
PT08.S1(CO), C6H6(GT), PT08.S2, PT08.S3, PT08.S4, PT08.S5, T, RH, AH	366 each	3.91% each

The most important limitation is NMHC(GT), where only 914 observations are non-missing. Conclusions involving NMHC should therefore be treated more cautiously than results based on variables with better coverage.

VIII. PROPOSED METHODOLOGY

The study follows a quantitative exploratory design. The workflow was kept simple so that every reported result can be traced back to the supplied dataset. No application development was required.

Phase	Activity	Output
1	Dataset acquisition and validation	Verified structure and variables
2	Data preprocessing	Analysis-ready data
3	Descriptive analysis	Summary statistics
4	Exploratory analysis	Distributions and trends
5	Correlation analysis	Correlation matrix and key associations
6	Temporal analysis	Monthly pollution patterns
7	Cross-sensor analysis	Sensor/reference relationships
8	Interpretation	Findings, limitations and conclusion

Analysis methods

- Descriptive statistics were used to summarize central tendency and spread.
- Monthly means were calculated to compare pollutant behaviour across the observation period.
- Pearson correlation was used to measure linear association between numeric variables.
- Environmental relationships were examined using correlations between pollutants and T, RH and AH.
- Reference measurements were compared with their nominally related sensor responses.
- Charts were used to make missingness, trends, correlations and sensor relationships easier to interpret

IX. RESULTS AND DISCUSSION

9.1 Descriptive findings

Variable	Mean	Std. Dev.	Minimum	Maximum
CO(GT)	2.153	1.453	0.100	11.900
NMHC(GT)	218.812	204.460	7.000	1189.000
C6H6(GT)	10.083	7.450	0.100	63.700
NOx(GT)	246.897	212.979	2.000	1479.000
NO2(GT)	113.091	48.370	2.000	340.000
Temperature	18.318	8.832	−1.900	44.600
Relative humidity	49.234	17.317	9.200	88.700
Absolute humidity	1.026	0.404	0.185	2.231

The reference pollutant variables show substantial variation rather than a narrow range of values. This supports the use of temporal and correlation analysis instead of relying only on overall averages.

9.2 TEMPORAL ANALYSIS

Monthly aggregation shows that pollutant behaviour is not constant throughout the dataset. The highest monthly mean CO value was observed in December 2004 (2.751). Benzene reached its highest monthly mean in October 2004 (13.525), while NOx reached its highest monthly mean in November 2004 (424.188). NO2 reached its highest monthly mean in February 2005 (160.695). These different peak months show that the pollutants do not follow exactly the same monthly pattern.

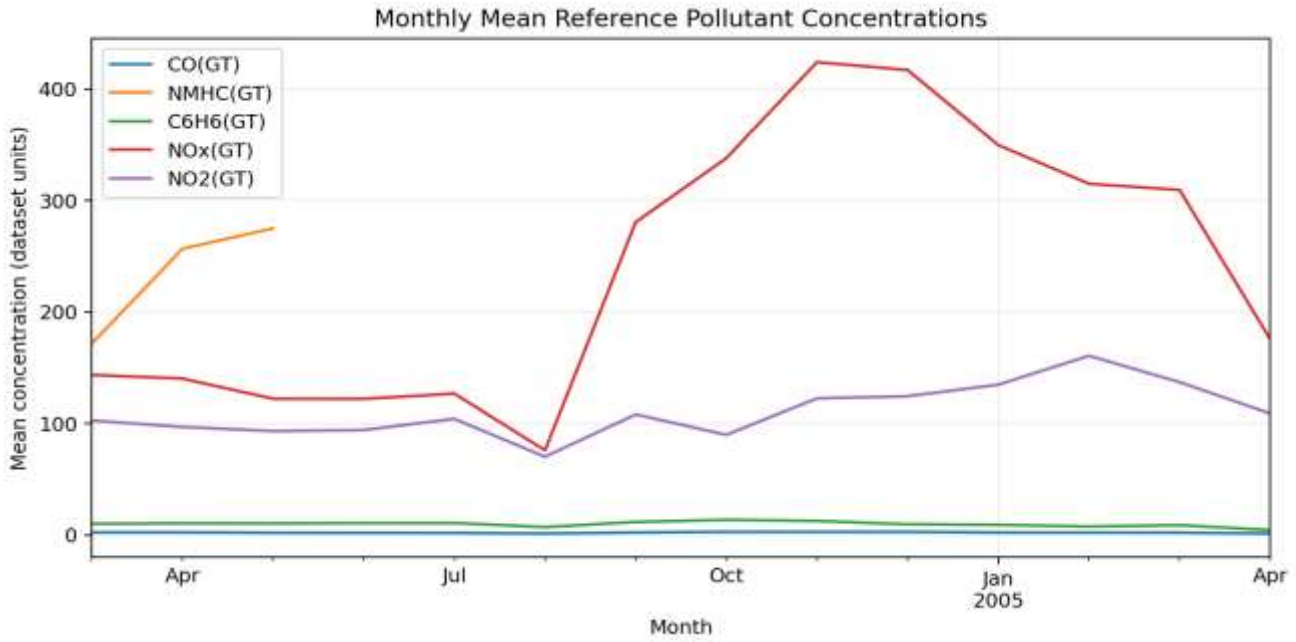


Figure 1. Monthly mean reference pollutant concentrations calculated from the supplied AirQualityUCI.csv file.

9.3 CORRELATION ANALYSIS

The correlation analysis shows several strong positive relationships. The strongest reported relationship was between C6H6(GT) and PT08.S2(NMHC), with r = 0.982. CO and benzene also showed a strong positive relationship (r = 0.931), followed by CO and PT08.S2(NMHC) (r = 0.916). These values indicate that the variables move together strongly in the supplied data, although correlation alone does not prove causation.

Rank	Variable A	Variable B	Pearson r
1	C6H6(GT)	PT08.S2(NMHC)	0.982
2	CO(GT)	C6H6(GT)	0.931
3	CO(GT)	PT08.S2(NMHC)	0.916
4	NMHC(GT)	C6H6(GT)	0.903
5	PT08.S1(CO)	PT08.S5(O3)	0.899
6	PT08.S1(CO)	PT08.S2(NMHC)	0.893
7	CO(GT)	NMHC(GT)	0.890
8	C6H6(GT)	PT08.S1(CO)	0.884
9	PT08.S2(NMHC)	PT08.S5(O3)	0.881
10	CO(GT)	PT08.S1(CO)	0.879

9.4 CORRELATION STRUCTURE

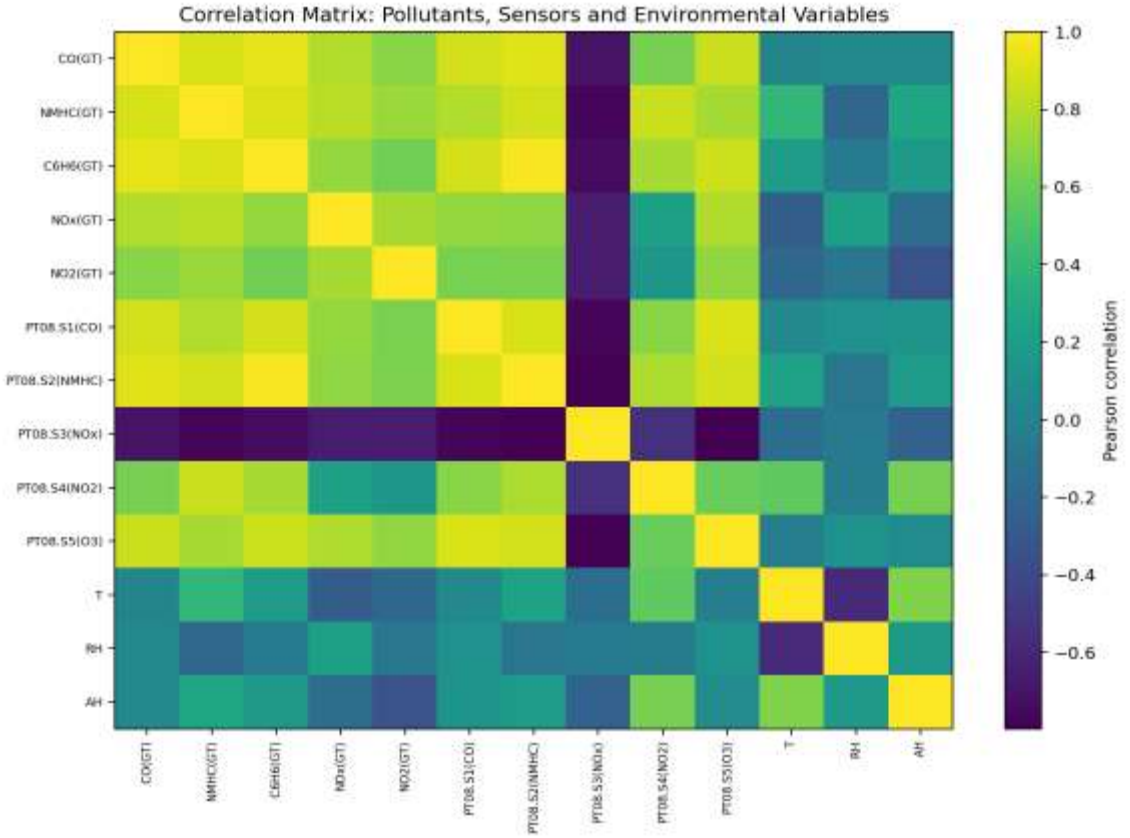


Figure 2. Pearson correlation matrix for pollutants, sensor responses and environmental variables.

The heatmap gives a broader view of the same pattern. Several pollutant and sensor variables are strongly related, while some environmental relationships are much weaker. Strong sensor-to-sensor relationships also suggest that sensor responses should not automatically be interpreted as independent measurements.

9.5 ENVIRONMENTAL RELATIONSHIPS

Pollutant	Temperature r	Relative Humidity r	Absolute Humidity r
CO(GT)	0.022	0.049	0.049
NMHC(GT)	0.392	−0.191	0.270
C6H6(GT)	0.199	−0.062	0.168
NOx(GT)	−0.270	0.221	−0.149
NO2(GT)	−0.187	−0.092	−0.335

Most environmental correlations are weaker than the strongest pollutant and sensor relationships. NMHC(GT) has a moderate positive relationship with temperature ($r = 0.392$), while NO2 has a moderate negative relationship with absolute humidity ($r = -0.335$). These results should be interpreted as associations within this dataset rather than as evidence that one variable directly causes the other.

9.6 CROSS-SENSOR RELATIONSHIPS

Reference variable	Related sensor	Pearson r	Complete pairs
CO(GT)	PT08.S1(CO)	0.879	7,344
C6H6(GT)	PT08.S2(NMHC)	0.982	8,991
NOx(GT)	PT08.S3(NOx)	−0.656	7,396
NO2(GT)	PT08.S4(NO2)	0.157	7,393

X. DISCUSSION

The results show that the dataset contains clear internal relationships among pollutants and chemical sensor responses. The very high benzene–PT08.S2(NMHC) correlation is particularly notable, while the CO–PT08.S1(CO) relationship is also strong. These findings are consistent with the general purpose of paired reference and sensor measurements: sensor signals can contain useful information about the corresponding pollutant environment.

At the same time, the results are not uniformly strong. The NO2–PT08.S4(NO2) relationship is weak ($r = 0.157$), whereas NOx–PT08.S3(NOx) is moderately negative ($r = -0.656$). This difference is a useful reminder that a sensor label alone does not guarantee a simple one-to-one relationship with a reference concentration. Cross-sensitivity, environmental conditions and sensor behaviour can influence the observed association [2], [4].

The temporal results also show that different pollutants reached their highest monthly means in different months. This supports retaining the time dimension during analysis. A single overall mean would hide these changes.

Missingness is another important part of the result rather than a minor preprocessing detail. NMHC(GT) has 8,443 invalid/missing coded observations, leaving only 914 usable values. Analyses involving this variable therefore have a much smaller effective sample than analyses involving variables with better coverage.

XI. RECOMMENDATIONS

- Perform data-quality checks before statistical interpretation, especially for variables with heavy missingness.
- Use temporal information when comparing sensor behaviour because relationships may change across time.
- Use reference measurements when evaluating sensor responses rather than treating sensor outputs as direct ground truth.
- Treat correlation as association, not causation, and avoid causal claims from this exploratory analysis.
- For future work, compare the historical dataset with newer datasets from multiple monitoring locations.
- If prediction or calibration is later required, use chronological validation to avoid leakage from future observations.

XII. LIMITATIONS

- The analysis is based on one historical monitoring dataset rather than current multi-location observations.
- Several variables contain missing/invalid values, with NMHC(GT) affected most severely.
- The study is observational, so correlation does not establish causation.
- Chemical sensors can show cross-sensitivity and drift, which complicate direct interpretation of sensor/reference relationships.
- The supplied file contains 9,357 populated observations, while the official UCI record reports 9,358 instances; the supplied file count was used for calculations.

XIII. CONCLUSION

This study provides a clear exploratory analysis of urban air-quality data by combining temporal analysis, descriptive statistics, pollutant correlations, environmental relationships and cross-sensor comparisons. The strongest relationships were found between benzene and PT08.S2(NMHC), CO and benzene, and CO and PT08.S1(CO). The results also showed that different pollutants reached their highest monthly means at different times, confirming that time is an important part of the dataset.

The analysis also demonstrates why air-quality data should not be interpreted only from overall averages. Missing values, sensor cross-sensitivity, differences between sensor/reference relationships and environmental effects all matter. The study therefore provides a transparent baseline for future work in sensor calibration, forecasting and time-series modelling without requiring a complex software application.

XIV. FUTURE SCOPE

- Compare the historical dataset with recent multi-location air-quality datasets.
- Evaluate robust missing-data reconstruction methods and compare their effect on conclusions.
- Study sensor drift across different time windows.
- Apply explainable forecasting or calibration models with chronological validation.
- Test whether important relationships remain stable across seasons and locations.

REFERENCES

1. S. Vito, "Air Quality," UCI Machine Learning Repository, Dataset 360, 2008, doi: 10.24432/C59K5F.
2. S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," Sensors and Actuators B: Chemical, vol. 129, no. 2, pp. 750–757, 2008, doi: 10.1016/j.snb.2007.09.060.
3. T. Kong, D. Choi, G. Lee, and K. Lee, "Air Pollution Prediction Using an Ensemble of Dynamic Transfer Models for Multivariate Time Series," Sustainability, vol. 13, no. 3, art. 1367, 2021, doi: 10.3390/su13031367.
4. S. Ali, F. Alam, J. Potgieter, and K. M. Arif, "Leveraging Temporal Information to Improve Machine Learning-Based Calibration Techniques for Low-Cost Air Quality Sensors," Sensors, vol. 24, no. 9, art. 2930, 2024, doi: 10.3390/s24092930.
5. F. E. Ahangar, F. R. Freedman, and A. Venkatram, "Using Low-Cost Air Quality Sensor Networks to Improve the Spatial and Temporal Resolution of Concentration Maps," International Journal of Environmental Research and Public Health, vol. 16, no. 7, art. 1252, 2019.
6. "Characterisation and Calibration of Low-Cost PM Sensors at High Temporal Resolution to Reference-Grade Performance," 2023.
7. "An Efficient Spatiotemporal Data Calibration Approach for the Low-Cost PM2.5 Sensing Network: A Case Study in Taiwan," Environment International, 2019.
8. "Spatiotemporal Air Quality Inference of Low-Cost Sensor Data: Evidence from Multiple Sensor Testbeds," Environmental Modelling & Software, vol. 149, art. 105306, 2022.
9. R. Wathore et al., "Exploring Environmental and Temporal Performances of Machine Learning Models for Calibration of a Low-Cost PM2.5 Sensor," Aerosol and Air Quality Research, vol. 26, art. 00094, 2026.
10. "An Integrated Low-Cost Air Quality Sensor and a Multi-Task Calibration Framework for Particulate Matter," Environment International, vol. 207, art. 109981, 2026.