DATA319 Spring 2025 Final Project Report

Air Quality Analysis in Urban Italy Using Multivariate Techniques

Introduction

Air pollution is a serious environmental concern posing important risks to public health and urban environments. It is necessary to comprehend the dynamics of air quality in cities to support policy and intervention development to reduce levels of pollution. Our researchers chose the "Air Quality" dataset from the UCI Machine Learning Repository to undertake a close analysis of air pollution trends in Italy's very polluted city. The data set is a full set of hourly series of observations on different pollutants such as carbon monoxide (CO), non-methane hydrocarbons (NMHC), nitrogen dioxide (NO2), and benzene (C6H6), and meteorological parameters such as temperature, relative humidity, and absolute humidity. The whole data set spans one year from March 2004 to February 2005 and provides a good range to verify seasonal and diurnal air quality trends.

The reason why this dataset has been chosen is that it pertains to urban environmental health and can be used in policy-making. One of the leading causes of respiratory and cardiovascular disease is air pollution, and urban areas, because of the large amount of traffic and industrialization, are particularly vulnerable to high levels of pollution. By analyzing this dataset, we aim to uncover correlations between different pollutants, identify time trends in emissions, and ascertain the effect of weather conditions on air quality. To achieve this, we will explore several key research questions: What relationships are there between the different types of pollutants? How do carbon monoxide (CO) levels fluctuate over a year? Which gases correlate with each other and show decreasing trends? Does time of day affect carbon monoxide (CO) emissions? What sort of relationship is there between temperature and gas concentrations? Such insights can guide policymakers in coming up with targeted measures to reduce pollution, such as traffic restrictions during peak hours or weather-based seasonal bans. Additionally, that the dataset includes both sensor response and pollutant concentration allows for a multiperspective examination with the ability to bridge gaps between ground-truth measurements and sensor-based monitoring networks.

It has 15 attributes and holds 9,358 hourly observations, including the values of weather variables and pollutants. Though it is afflicted with such issues as -200 indications for missing data values, this was dealt with well by preprocessing in a way that there was no loss of data values. The data is still valid since it exactly serves the function of responding to our research goals about the relations between pollutants, time trend, and effects of weather factors. For instance, we check if CO levels have daily variations corresponding to peak traffic hours or seasonal variations corresponding to temperature changes. We also seek intercorrelations among pollutants to identify common sources, such as traffic, which have a pattern of simultaneous emission of CO, NMHC, and NO2.

Our analysis applies multivariate techniques like Principal Component Analysis (PCA) and hierarchical clustering to reduce dimensionality and reveal latent patterns in the data. PCA is applied to condense the most contributory variables, and clustering groups similar pollution profiles, e.g., traffic-related emission peaks. Both techniques are extremely valuable to tackle the complexity of the dataset and to extract meaningful information. Furthermore, we examine the fit of the dataset to multivariate normal distributions for assumption checking for our statistical models. By combining these approaches, we aim to present an integrated view of air quality dynamics in the urban environment being studied. The broader implications of the project go beyond the academic. Our findings can be used by city planners and environmental regulators to create more effective air quality monitoring networks, direct pollution reduction strategies, and implement more targeted public health campaigns. For example, if our analysis reveals that CO peaks are found in morning and evening rush hours, city governments might wish to implement congestion pricing or promote public transportation to reduce emissions. Similarly, identifying strong correlations between certain pollutants would make it possible to reduce monitoring campaigns by measuring representative indicator pollutants rather than each contaminant separately. Last but not least, the project contributes to the growing body of research on urban air quality with evidence-based recommendations toward sustainable city planning and healthier environments.

Lastly, the "Air Quality" dataset in the UCI collection offers an actual chance to examine the complex interaction of pollutants with meteorological parameters in the context of a polluted urban area. The present research aims to reveal patterns and trends for informing decision-making as well as enhancing the management of air quality. In trying to answer fundamental research questions using multivariate approaches, we anticipate providing practical contributions that can have implications for public health interventions as well as environmental conservation initiatives. What follows positions our methodology, results, and conclusions in the context of the relevance of the research for both scientific and applied spheres.

Description of Data

The "Air Quality" dataset includes the following variables:

- Date - Date of measurement (Date)
- Time - Time of measurement (Categorical)
- CO(GT) - True hourly averaged concentration of CO in mg/m^3 (Integer)
- PT08.S1(CO) - Hourly averaged sensor response (nominally CO targeted) (Categorical)
- NMHC(GT) - True hourly averaged concentration of Non-Methane Hydrocarbons in µg/m^3 (Integer)
- C6H6(GT) - True hourly averaged concentration of Benzene (C6H6) in µg/m^3 (Continuous)
- PT08.S2(NMHC) - Hourly averaged sensor response (nominally NMHC targeted) (Categorical)
- NOx(GT) - True hourly averaged concentration of NOx in ppb (Integer)
- PT08.S3(NOx) - Hourly averaged sensor response (nominally NOx targeted) (Categorical)
- NO2(GT) - True hourly averaged concentration of NO2 in µg/m^3 (Integer)
- PT08.S4(NO2) - Hourly averaged sensor response (nominally NO2 targeted) (Categorical)
- PT08.S5(O3) - Hourly averaged sensor response (nominally O3 targeted) (Categorical)
- T - Temperature in °C (Continuous)
- RH - Relative Humidity in % (Continuous)
- AH - Absolute Humidity (Continuous)

Methods and Results

The first substantial step was performing extensive data cleaning as well as data preprocessing for the preparation of the dataset for multivariate analysis. Our original dataset had missing data with a placeholder of -200, which we replaced with NaN for efficient missing data handling. Considering that several columns had a very high percentage of missing data, we applied a threshold-based approach where we dropped a column with over 30% missing data for data integrity. This was a very crucial step because the inclusion of columns with a high percentage of missing data can introduce bias or biased results towards a specific analytical interpretation. Having addressed missing data, we merged the separate "Date" as well as "Time" columns into a merged "Date_Time" column, converting the same to a datetime for easy handling of a time factor. We derived some other time-based features like "Hour," "Month," as well as "Weekday" for easy exploration of diurnal, seasonal, as well as weekly trends due to pollutant levels. Our final cleaned dataset contained major pollutants like CO, C6H6, NOx, as well as NO2, with meteorological variables like temperature (T), relative humidity (RH), as well as absolute humidity (AH).

To explore correlations with types of pollution, as with meteorology, we began with exploratory data analysis (EDA) using correlation matrices and plots. Correlation matrices provided estimates of linear correlations of variables with high positive correlations of CO with benzene (C6H6) ($r = 0.93$) and moderate correlations of CO with NOx ($r = 0.79$) and with NO2 ($r = 0.67$). This suggests common pollution sources, possibly automotive activity, releasing all of these contaminants under similar conditions. Temperature was weakly positively correlated with benzene but negatively with NOx and NO2, suggesting seasonal effects. Plots, including heatmaps and time series plots, were employed to investigate these trends. For instance, line plots of levels of CO vs. time suggested winter peaks, possibly due to increased heating needs as well as inversions with temperatures trapping pollution downwind.

As the dataset is multivariate, we applied Principal Component Analysis (PCA) to reduce dimensions and identify the most important factors. PCA transforms the original variables into a new set of linearly uncorrelated principal components (PCs) that account for the maximum variance in the data. We preprocessed the data using the StandardScaler before PCA so that all variables would make proportional contributions to the analysis. The scree plot revealed that the first two principal components explained approximately 46.7% and 27.6% of the variance, cumulatively explaining more than 74% of the variance. The PC1 vs. PC2 biplot revealed CO, C6H6, NOx, and NO2 to be loading heavily on PC1, indicating that these pollutants varied together, while temperature and absolute humidity had larger effects on PC2. This dichotomy emphasized the distinct roles of pollutant concentrations and meteorological conditions in shaping air quality patterns.

To complement PCA, we applied hierarchical clustering to segment periods that share similar pollution profiles. We used hierarchical clustering because it can discover nested structures of data without knowing the number of clusters. Ward's method of linkage, variance-minimizing within groups, and Euclidean distance as the metric were utilized. The resulting dendrogram identified natural clusters for different times of day, i.e., rush hours (morning and evening), when the pollutant levels were highest. This was consistent with our hypothesis that traffic pollution is the reason for short-term air quality fluctuations. Cluster analysis also helped to identify seasonality effects, with the winter months clustering because of high CO and benzene. These findings validated the worth of clustering in identifying patterns over time that would not necessarily be apparent through univariate studies.

Normality test and comparison to multivariate normal distributions were applied to test our statistical assumptions. CO, NO2, and temperature were treated as representative variables for the testing. Shapiro-Wilk tests identified that none of the variables happened to be univariate normal ($p < 0.001$ for all), a common trend in environmental observations under the impact of skewness and outliers. To visually confirm non-normality, we developed quantile-quantile (Q-Q) plots, which confirmed heavy-tailedness of the distributions of the pollutants. Despite this, we proceeded with PCA and clustering since these steps are extremely resilient to small departures from normality. To enable comparison, we generated a multivariate normal set with the same mean and covariance structure as our selected variables. Synthetic data contained more evenly balanced distributions and fewer outlier values, focusing on the complexity of the original data and the effects that outside variables like traffic and weather have.

The application of PCA and hierarchical clustering was driven by their usefulness in addressing our research questions. PCA enabled us to visualize the overall picture of variable correlations, and clustering determined time intervals with analogous pollution profiles. Other methods, such as factor analysis or k-means clustering, were also considered but were less appropriate. Factor analysis is based on the premise of underlying latent variables, which was not our aim in seeking dominant pollutants. K-means clustering would have required us to predetermine the number of clusters, which was not possible with our exploratory orientation. Advantages of our chosen methods were interpretability and insensitivity to correlated variables, whereas drawbacks were sensitivity to outliers and the need for careful preprocessing.

Ethical issues were also a consideration in our approach. The data set, while anonymized, represented actual measurements from a polluted urban area, and so we took pains not to misrepresent or exaggerate results that could potentially have implications for public opinion or policy. Missing data were handled transparently, and results were qualified with caveats about potential sensor drifts or cross-sensitivities described in the original study. By being systematic and open, we aimed to provide believable results that could inform air quality management planning without unnecessarily alarming citizens.

Finally, our methodology incorporated data cleaning, exploratory analysis, dimensionality reduction, clustering, and normality testing in a manner that gave comprehensive treatment to the research questions. Each step was carefully designed to capitalize on the strength of multivariate methods while their limitations were also considered. The combination of the methods allowed us to find sensible trends in air quality data, paving the way for sensible conclusions and directions for future research.

The exploratory analysis revealed some significant trends in the air quality data that respond to our research questions. The correlation matrix revealed strong positive correlations between CO and benzene (C6H6) ($r = 0.93$), between CO and NOx ($r = 0.79$), and between CO and NO2 ($r = 0.67$). These high correlations suggest these pollutants have common sources, most probably vehicular emissions, as a result of the urban location of the measurement. The association between benzene and CO was also high, again suggesting the relationship of traffic and these pollutants since benzene is a known exhaust constituent of vehicles. Meteorological parameters had more varied relationships with pollutants. Temperature showed a weak positive correlation with benzene ($r = 0.19$), but negative ones with NOx ($r = -0.28$) and NO2 ($r = -0.21$), so more favorable temperatures can be associated with less harmful nitrogen oxides, possibly due to greater dispersion of air or to fewer emissions from heating on warmer days.

Principal Component Analysis (PCA) further elucidated the relationships by reducing the dataset dimensionality. The first two principal components (PC1: 46.7%, PC2: 27.6%) accounted for 74.3% of the total variance. The first principal component was dominated strongly by CO, benzene, NOx, and NO2, again highlighting their covariance as well as their common sources. PC2 was dominated by temperature and absolute humidity, however, and highlighted meteorological conditions with a distinct air quality pattern effect. The biplot plot showed that pollutant concentrations were clustered, while temperature and humidity vectors were in opposite directions, highlighting their independent effects. Such separation shows that while pollutants co-vary together due to common emission sources, weather conditions control their concentrations independently. As an example, higher temperatures may scatter pollutants more efficiently, while humidity may affect particulate formation or sensor efficiency.

Hierarchical clustering also identified temporal trends in pollution concentrations, namely diurnal and seasonal trends. The dendrogram indicated distinct clusters for rush hour periods (8-10 AM and 5-8 PM), during which CO concentrations reached median levels of 2.8-3.5 mg/m3. These peaks are as expected from traffic congestion patterns and support the hypothesis that traffic emissions are responsible for short-term pollution peaks. Nighttime hours (12-5 AM) formed a separate cluster with the lowest CO levels (0.6-1.3 mg/m3), reflecting less human activity. Seasonal clustering reflected greater CO and benzene levels during winter months (December-February) due to increased heating requirements and temperature inversions holding pollutants near the surface. Lower levels of NOx and NO2 were seen for summer months (June-August), as well as in line with their negative correlation with temperature.

The time-series analysis of CO concentration generated both long-term and short-term trends. Averages were most elevated in October (2.81 mg/m3) and November (2.76 mg/m3), when temperatures were lower and perhaps more fossil fuel combustion occurred. The lowest averages were in August (1.28 mg/m3), possibly due to favorable dispersion conditions and reduced heating demands. Hourly median CO concentrations also showed a clear diurnal trend, with the highest concentrations at evening rush hours (7-8 PM: 3.5 mg/m3) and lowest at early morning (4-5 AM: 0.6 mg/m3). The trend mirrors urban traffic activity and indicates the effect of human activity on air quality. Of special note, there was a sharp spike of nearly 12 mg/m3 in the late 2004 period that might be linked to an individual pollution event like a traffic jam or factory operation.

The temperature-concentration relationship varied by gas. Linear regression showed no meaningful correlation between temperature and CO (slope = 0.003, p = 0.127), suggesting that CO levels are more strongly related to direct emissions than weather. Benzene did increase with temperature (slope = 0.159, $p < 0.001$), perhaps due to increased evaporation of volatile compounds at higher temperatures. NOx and NO2 had strongly negative correlations with temperature (NOx: slope = -6.51, $p < 0.001$; NO2: slope = -1.149, $p < 0.001$), consistent with better atmospheric dispersion in warmer weather or seasonally lower heating-related emissions. These are consistent with the correlation matrix and PCA findings and further evidence of the effect of temperature on the modulation of some pollutants.

Normality tests confirmed that the data were very far from being multivariate normal. Shapiro-Wilk tests were rejecting normality for CO, NO2, and temperature ($p < 0.001$ each), and Q-Q plots indicated heavy-tailed distributions, particularly for pollutants. This type of skew is common to environmental data, in which extreme values (e.g., peaks in pollution) are infrequent but of ecological significance. Comparison with artificial multivariate normal data highlighted these differences: the real dataset had more outlier values and non-linear relationships, as a result of the complex interactions of emissions, weather, and human activity. These differences apart, PCA and clustering were still effective because these methods do not suffer from mild non-normality. The outcomes of PCA and clustering also provided information about potential sensor cross-sensitivities or drifts. For instance, the consistent covariation of CO and benzene between clusters suggests that their measurements were uniform, but the anti-correlation between NOx and temperature could be either due to actual atmospheric processes or sensor artifacts. The paper initially reported potential sensor drifts, and our observation of seasonal clusters is partly a result of these technical limitations. However, the strong correlation between correlation patterns and known emission sources (e.g., traffic) warrants the quality of the overall data.

As a whole, results indicate that concentrations of pollutants in this urban area are affected by a combination of traffic, season, and weather conditions. CO and benzene increase during rush hours and in cold seasons, and NOx and NO2 reduce with the temperature rise, likely by dispersion. Such trends were properly identified by multivariate methods in which PCA tapped into dominant pollution-meteorology correlations and clustering tapped into actionable temporal trends. These findings provide the foundation for evidence-based targeted interventions in air quality, like rush-hour traffic control or winter emissions regulation. However, the non-normal distributions and potential sensor issues mean that these results must be interpreted with caution, and further validation using other data sources is required.

Discussion

Our analysis provided critical insights into the factors driving air quality patterns in urban Italy. Through the application of multivariate techniques, we found strong evidence that pollutant levels, particularly carbon monoxide (CO), benzene (C6H6), and nitrogen oxides (NOx, NO2), vary together and are strongly influenced by human activities such as traffic and heating. In parallel, meteorological variables like temperature and absolute humidity were found to play independent roles in modulating pollutant concentrations.

The fact that there exists such a strong correlation between CO, nitrogen oxides, and benzene shows that the pollutants have common sources, of which traffic and combustion stand out. The finding concurs with existing literature regarding the state of air quality within cities where congestion on the roads is a proven variable towards having high levels of pollutants. The morning rush hour and evening rush hour peaks of the levels of CO also reflect the relationship between human activity and deterioration of the atmosphere. The results confirm the effects of traffic control measures such as congestion pricing or promotion of mass transport towards capping peak levels of pollutants during times of congestion.

The use of Principal Component Analysis (PCA) successfully reduced dimensionality, allowing us to visualize the strong covariation between pollutants and to distinguish the separate influences of weather conditions. Hierarchical clustering revealed meaningful temporal patterns, with rush hours and winter months clustering together due to elevated pollution levels, consistent with known patterns of human activity and atmospheric behavior.

PCA's differentiation between pollutant levels and meteorology variables provides informative results for environmental authorities and city planners. For instance, the predominance of temperature and humidity in the second principal component suggests the weather's contribution towards pollutant diffusion regardless of emission rates. The differentiation of this type means that measures of limiting pollution should have provision for emission measures as well as for the contribution of weather towards diffusion. The cluster results, once again yielding rush hour and winter month profiles of pollution, provide intelligence for intervention. For instance, warnings of the quality of the air might be scheduled around such times of maximum hazard, and wintertime heating-related emissions might be regulated seasonally.

However, several limitations must be acknowledged. Firstly, the dataset contained notable missing values, although our preprocessing strategies, including threshold-based column removal and interpolation, helped mitigate this issue. Secondly, normality tests and Q-Q plots indicated significant deviations from multivariate normality. While PCA and clustering are robust to these deviations, this could still introduce minor biases into the interpretation of results.

That missing values were present, particularly within columns having more than 30% missing values, had to be treated cautiously so as not to skew the analysis. While our method of eliminating such columns preserved data integrity, the possibility exists that informative variables might have been excluded. It might be worth exploring more advanced imputation techniques in the future in order not to discard these variables. The heavy-tailed Q-Q plots reflective of non-normal distributions of pollutant concentrations are a manifestation of the richness of environmental data, where there exist sporadic but significant extreme values (e.g., spikes of pollutant concentration). While PCA and clustering are resilient towards such deviations, the outcomes have to be interpreted carefully, especially when inferring extremes or rare observations.

Moreover, the scope of the data presents constraints. The dataset covers only one city over a single year, limiting the generalizability of our findings to other urban areas or broader timescales. Seasonal anomalies or unique events during the study period could have disproportionately influenced pollutant levels. Additionally, potential sensor drifts and cross-sensitivities noted in the original dataset documentation could have introduced subtle measurement errors.

The single-city, single-year design restricts extrapolating our findings to other regions with different traffic volumes, industrial output, or atmospheric conditions. For example, tighter emissions regulations or different transit modes in cities may have different pollutant relationships. The duration of the dataset also raises the issue of whether the observed patterns are representative of long-term trends or whether they are subject to anomalous events, e.g., a very cold winter or a one-off surge in traffic. The sensor issues that were noted introduce another level of complexity to interpretation because cross-sensitivities would muddy the distinctions between pollutant measurements. These limitations make it all the more important for validation against other datasets over other locations and longer periods.

Future work could address these limitations by extending the analysis to datasets spanning multiple cities and years to validate the patterns observed. Incorporating additional contextual data, such as traffic volumes, industrial activities, or meteorological forecasts, would enable a more complete modeling of pollution dynamics. Methodologically, exploring alternative clustering techniques such as DBSCAN or advanced dimensionality reduction methods like t-SNE or UMAP could uncover even finer structures within the data.

Extending the analysis to incorporate data from several cities would determine universal patterns versus local trends. Using real-time traffic patterns or industrial emission reports would further improve the accuracy of the model by directly attributing pollution levels to the sources. More sophisticated algorithms, such as DBSCAN, could more aptly detect those occasionally shaped clusters in the data, and t-SNE or UMAP may detect nonlinear patterns that PCA will fail to catch. The inclusion of machine learning techniques, such as random forests or neural networks, would also further increase prediction quality through the modeling of complex variable interactions. All such advances would provide a richer characterization of the dynamics of air quality and improve the accuracy of policy recommendations.

From an ethical standpoint, while the dataset is anonymized and publicly available, careful interpretation is essential. Overstating causal relationships without longitudinal validation could mislead policymakers and the public. Our analysis emphasizes that, while correlations are strong, causation must be approached cautiously, and results should be contextualized appropriately.

The ethical significance of air quality research is that conclusions can influence public health policy and city planning. Mistaken attribution of cause from association can lead to perverse, or even negative, interventions. For example, interpreting temperature as the direct cause of fluctuations in pollutant concentration without controlling for confounding factors like seasonal traffic patterns can lead to spurious policy. Transparency regarding limitations of the dataset, such as sensor drift or missing data, is required if public trust needs to be preserved, as well as in order not to overestimate conclusions. Investigators should only report results, separating reported causal mechanisms from reported associations for evidence-based decision-making.

Overall, this study highlights the effectiveness of multivariate methods in uncovering complex environmental patterns and emphasizes their potential for guiding evidence-based strategies to improve urban air quality. The combined application of PCA and hierarchical clustering resulted in a sound framework for dissecting the multi-faceted interaction between pollution and weather conditions. The strategy revealed actionable insight, such as the need for certain traffic regulation measures during peak hours and periodic adjustments to heating-related pollution. Although with certain limitations, the paper demonstrates the value of multivariate analysis to environmental science and its applicability to real policy concerns. Through further refining of these approaches and expanding the scope of analysis, subsequent research can further enhance our understanding of air quality dynamics and assist in developing healthier cities.

Conclusion

Analysis of the "Air Quality" dataset from the UCI Machine Learning Repository was useful in uncovering patterns of urban air pollution in an Italian city. Correlation of pollutants CO, benzene, and nitrogen oxides was found as strong indicators for a common source of emission, particularly traffic. Diurnal patterns of peak concentrations during rush hours supplemented the evidence for human activities as factors behind the degradation of air quality. Seasonal trends underscored increased pollution in colder months, likely due to higher heating demands and atmospheric inversions. These results underscore the importance of targeted measures, such as traffic restrictions in rush hours and seasonal restrictions on heating emissions, to decrease pollution peaks. The multivariate techniques employed- PCA and hierarchical clustering- were able to disentangle the complex interaction between pollutants and meteorological variables, and present a firm foundation for future air quality studies.

Although the strengths of our approach are persuasive, such limitations as missing data, sensor drifts, and the constrained temporal and spatial nature of the dataset necessitate caution in interpretation. Non-normality of the distributions of pollutant concentrations further complicates the task of statistical modeling, but the robustness of PCA and clustering alleviates some of those challenges. Extension of this study in future studies by incorporating multi-year and multi-city datasets could test our findings and enhance generalizability. More advanced techniques like DBSCAN or t-SNE might identify more subtle trends, and adding additional contextual data (e.g., traffic volume, industrial output) might be able to improve the predictive accuracy. Overcoming these limitations would make the conclusions stronger and more useful for policy-making.

The moral import of this research is high since air quality directly affects public health. Although our analysis revealed actionable patterns, over-emphasizing causality without more validation may be used to craft misguided policies. Honest disclosure of the dataset's limitations-e.g., possible inaccuracies in the sensors- is required to preserve public trust and secure evidence-based policy-making. Policy-makers should use these results as a component of a broader strategy, combining them with local knowledge and other sources of data to support the design of effective air quality interventions.

Finally, the project reiterated the multi-purpose of multivariate analysis in solving complex environmental issues and elucidating successful patterns. From these results came the correlation between human activity, climate, and intensity of air pollutants, upon which successful focus-directed urban air quality management will depend. Enhanced methods, increased scope, and extended futures are ensured when future generation work attempts to reinforce, increase, or build further what is here achieved by demonstrating the applicability of multivariate techniques to shedding light on the air pollution phenomenon. The use of such analytical viewpoints within policy structures can reduce pollution and preserve public health within urban areas across the world.

Dataset:

https://archive.ics.uci.edu/dataset/360/air+quality