Leading Causes of US Air Pollution

Adam Hetherwick and Jose Arreola-Muñoz

June 11, 2024

1 Introduction

A common predicament of large metropolitan cities is that with such a high density of vehicles and factories comes a high proportion of smog and pollution. Pollution is a major concern because it can cause chronic diseases and physical ailments such as asthma, heart disease, lung cancer, etc. Such issues require preventative measures, like mapping and analyzing the air sulpher dioxide concentration for different United States cities. Our data is made available to us as an R site package taken from Hand et al. (1994). We aim to analyze which climatic and human ecological variables contribute to a cities' air pollution the most.

This analysis will provide us with key insights as to which variables are the highest contributors to sulpher dioxide concentration. We aim to group cities by similarities to predict whether or not future cities of similar status will have high or low pollution levels. We will also perform statistical computations and hypothesis testing to ensure normality of the predictors and responses, understand relationships between variables, group cities based on similar characteristics, and predict pollution level. The following section will introduce the dataset in higher detail.

2 Data Description and Visualization

The data for our analysis involves 41 United States cities with 6 predicting variables and 1 response, sulpher dioxide concentration. The sulpher dioxide concentration (SO_2) is a good predictor for a cities pollution, as it is a primary pollutant from fossil fuel combustion at power plants (EPA 2024). In order to group cities by similarities, we will examine the average annual temperature in degrees F, the number of manufacturers employing more than 20 workers, population size in thousands, average annual wind speed in miles per hour, average annual rainfall in inches and days. Examining the distribution of the sulpher dioxide content will give us more information on the behavior of the response.

In order to properly analyze this dataset, we must convert it to a form compatible with multivariate analysis. The data was originally given as each row denoting a cities' SO_2 and predictors. To convert it to a multivariate sense, we can think of it as two samples of cities, one with low polluted cities, and one with high polluted cities, distinguished by their SO_2 in comparison to the median SO_2. This will convert the dataset to a multivariate case with $p = 6$, one for each predictor, and $n_1 = 22$ for low pollution, and $n_2 = 19$ for high pollution.

A critical aspect of multivariate analysis is the assumption of normality of the response and the predictors. In order to ensure this assumption we can conduct the Shapiro-Wilks test which has the following W test statistic and hypothesis:

$$\mathbf{W} = \frac{(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{a_i x_{(i)}})^{\mathbf{2}}}{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \bar{\mathbf{x}})^{\mathbf{2}}}$$

H_0: The data is not significantly different than a Gaussian population.

H_A: The data is significantly different than a Gaussian population.

	SO2	temp	manu	popul	wind	precip	predays
p_value	9.723376e-06	0.02214972	2.781101e-09	3.622798e-08	0.697258	0.03725311	0.2419457

Figure 1: Shapiro-Wilks normality test p-values.

From Figure 1, we can see that the response, along with 4 of the predictors fail the normality test and reject the null hypothesis. In order to address this, we can calculate the log transformation of each of these variables then conduct the p-values once more.

	SO2	temp	manu	popul	wind	precip	predays
p_value	0.0967	0.1938	0.6744	0.3676	0.6973	0.0373	0.2419

Figure 2: Shapiro-Wilks normality test p-values with log-transformed variables.

After the log transformation of the non-normal variables and response, all but 1 of the variables are now normal as seen in Figure 2. The only non-normal variable is the average annual rainfall in inches. It is worth noting that various transformations of this variable resulted in lower p-values, such as the log-transformation ($p = 1.005e^{-06}$), square root transformation ($p = 0.0002481$), and the cubic root transformation ($p = 3.833e^{-05}$). We will proceed with our analysis with caution knowing that the original p value is nearly > 0.05, though not exactly. We can now continue with our data description by observing the behavior of the transformed response.

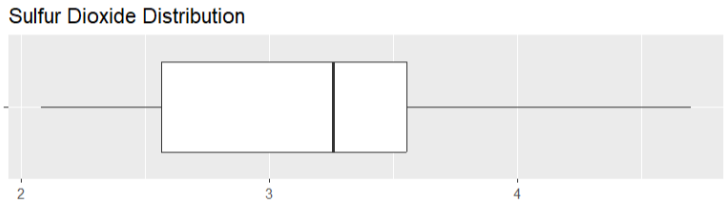

Figure 3: Sulpher Dioxide Concentration Distribution.

Figure 3 above illustrates this nature, which consists of a slight skew towards cities with higher pollution. It is also worth noting that there are no outliers. Next, we can examine the relationship between the transformed predictors and the transformed response, grouped by high and low pollution.

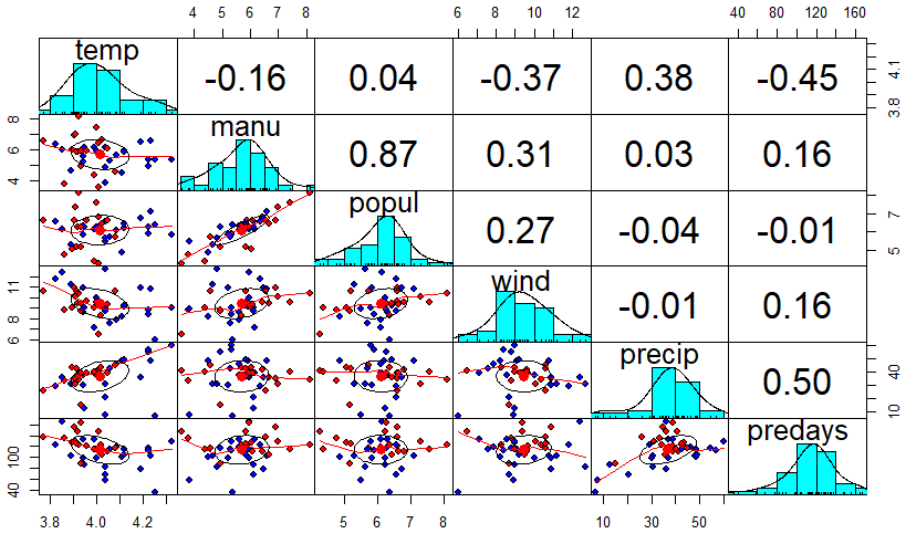

Figure 4: Multiple Predictor Dependency Plot.

Figure 4 includes the relationship between each predictor, along with the R^2 correlation coefficient. Here we can see that the only predictors with a strong correlation are the number of manufacturers employing over 20 workers and the population size in thousands $R^2 = 0.87$. This begs the ever so infamous question of whether or not it is causation or correlation. Average rainfall in inches and days per year along with rainfall in days and average temperature also have higher correlations (0.50 and -0.45 respectively). Another descriptive method to examine predictor dependence and pivotal aspect of multivariate analysis is the covariance matrix.

	temp	manu	popul	wind	precip	predays
temp	0.02	-0.01	0.01	-0.10	0.85	-1.12
manu	-0.01	0.48	0.29	0.16	1.17	4.89
popul	0.01	0.29	0.25	0.07	-0.17	-0.15
wind	-0.10	0.16	0.07	2.98	0.01	11.29
precip	0.85	1.17	-0.17	0.01	221.46	271.13
predays	-1.12	4.89	-0.15	11.29	271.13	801.12

Figure 5: Low Pollution Covariance Matrix

	temp	manu	popul	wind	precip	predays
temp	0.01	-0.02	0.00	-0.02	0.27	-0.54
manu	-0.02	1.43	1.15	0.76	-0.61	-0.77
popul	0.00	1.15	1.15	0.60	-0.58	-0.72
wind	-0.02	0.76	0.60	1.05	-0.50	-0.07
precip	0.27	-0.61	-0.58	-0.50	49.55	26.51
predays	-0.54	-0.77	-0.72	-0.07	26.51	338.78

Figure 6: High Pollution Covariance Matrix

Figures 5 and 6 display the covariance matrices for the high and low pollution subsets. We see that for both the low polluted and high polluted cities, a lot of the values off the main diagonal are near 0 which implies no covariance. For low polluted cities, the average number of high precipitation days per year has strong covariance with average rainfall in inches (same for high polluted cities) and average annual wind speed. It is also worth noting that both these covariance matrices are positive semi-definite, as seen by the eigenvalues being greater than or equal to 0.

```
High_Pollution_Covariance_Matrix_Eigenvalues 341.1943  47.1603 2.9217 0.5612 0.1269 0.0055
Low_Pollution_Covariance_Matrix_Eigenvalues  908.3188 114.5814 2.7108 0.6511 0.0439 0.0040
```

Figure 7: Eigenvalues of High and Low Pollution Covariance Matrices.

Despite some eigenvalues being close to 0, Figure 7 shows all of them are above 0, signifying that the covariance matrices are indeed positive semi-definite. This is important to ensure for further analysis. We can now examine the long anticipated relationship between the transformed predictors and the response.

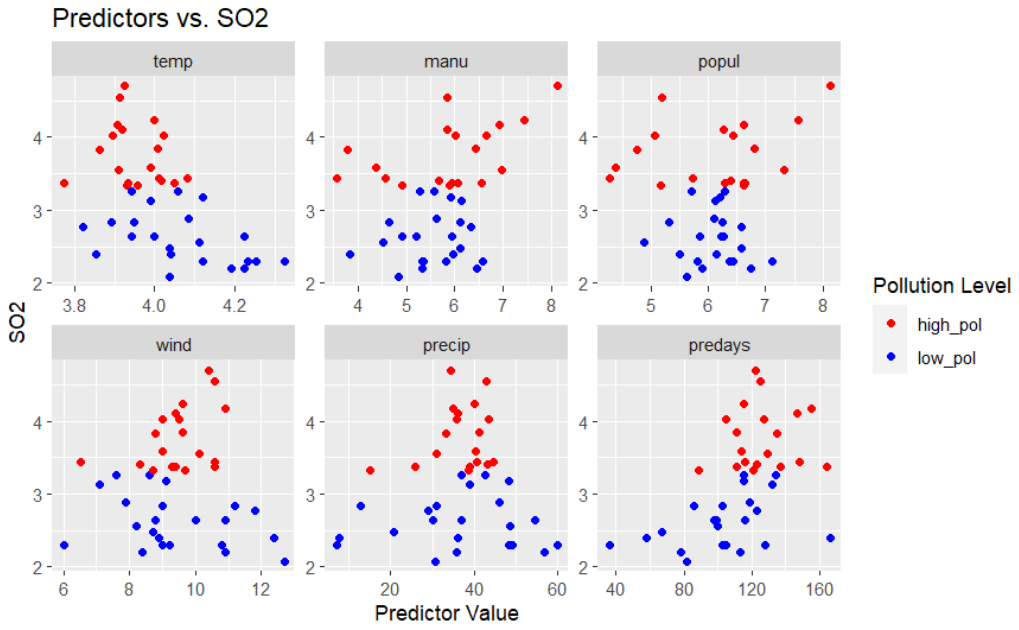

Figure 8: Relationship between Predictors and Response.

Figure 8 shows that no predictor has a distinct linear relationship with the response. A linear decision boundary may not work well. To ensure this, we can use Box's M test to test if the covariance matrices are statistically equivalent. The following section will explain in further detail the tests required for analysis and prediction.

3 Methods

In this section we will perform various statistical methods and testing to predict the pollution level of a city given their human and environmental variables. We will begin our analysis by performing the Box's M test to examine which type of decision boundary we should choose.

3.1 Box's M Test

Box's M test tests whether or not the covariance matrices for the two groups are statistically equivalent. The formal hypothesis and test statistic are presented here.

$$\mathbf{u_1} = \mathbf{-2(1-c_1)ln(M)}$$

$$\nu_1 = \mathbf{n_1 - 1}, \nu_2 = \mathbf{n_2 - 1}$$

$$\mathbf{c_1} = \frac{(\frac{1}{\nu_1} + \frac{1}{\nu_2} - \frac{1}{\nu_1+\nu_2})(\mathbf{2p^2 + 3p - 1})}{\mathbf{6(p+1)}}$$

$$\mathbf{M} = \frac{|\mathbf{S_1}|^{\frac{\nu_1}{2}}|\mathbf{S_2}|^{\frac{\nu_2}{2}}}{|\mathbf{S_{pl}}|^{\frac{\nu_1}{2}+\frac{\nu_2}{2}}}$$

$$\mathbf{S_{pl}} = \frac{\nu_1\mathbf{S_1} + \nu_2\mathbf{S_2}}{\nu_1 + \nu_2}$$

$$\mathbf{S_1} = \frac{\mathbf{1}}{\mathbf{n_1 - 1}}\sum_{\mathbf{i=1}}^{\mathbf{n_1}}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})'$$

$$\mathbf{S_2} = \frac{\mathbf{1}}{\mathbf{n_2 - 1}}\sum_{\mathbf{i=1}}^{\mathbf{n_2}}(\mathbf{y_i} - \bar{\mathbf{y}})(\mathbf{y_i} - \bar{\mathbf{y}})'$$

H_0: The data is not significantly different than a Gaussian population.

H_A: The data is significantly different than a Gaussian population.

Given our u_1 test statistic for the Box's M test is larger than our critical value $\chi^2_{0.95,21}$ ($39.66 > 32.67$), we reject the null hypothesis and conclude the covariance matrices are not statistically equal. This supports separating the cities' pollution levels by a quadratic function.

3.2 Quadratic Discriminant Analysis

We can use the following formula to calculate the quadratic decision boundary in order to predict pollution class. This in turn will separate our classes by a quadratic function.

$$\delta_k(x) = -\tfrac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log \pi_k$$

$$\max_{1\le k\le K} \delta_k(x)$$

```
[1] 0.8780488
Confusion Matrix and Statistics

          Reference
Prediction high_pol low_pol
  high_pol       17       3
  low_pol         2      19
```

Figure 9: QDA Accuracy and Confusion Matrix.

As seen in Figure 9, the classification rate is quite high for QDA. We anticipate that the this model will yield a higher classification rate compared to LDA.

3.3 Linear Discriminant Analysis

To compare our classification results to the Quadratic Discriminant Analysis, we can maximize the linear discrimination function, t calculate the accuracy rate and confusion matrix:

$$\delta_{\mathbf{k}}(\mathbf{x}) = \mathbf{x^T}\sum\nolimits^{-\mathbf{1}}\mu_{\mathbf{k}} - \mathbf{0.5}\mu_{\mathbf{k}}^{\mathbf{T}}\sum\nolimits^{-\mathbf{1}}\mu_{\mathbf{k}} + \mathbf{log}\pi_{\mathbf{k}}$$

```
[1] 0.7317073
Confusion Matrix and Statistics

          Reference
Prediction high_pol low_pol
  high_pol       13       5
  low_pol         6      17
```

Figure 10: LDA Accuracy and Confusion Matrix.

As anticipated, our LDA accuracy rate is lower than that of the QDA. The LDA decision boundary incorrectly predicts high pollution and low pollution. This is due to the fact that $\sum_{\mathbf{1}} \neq \sum_{\mathbf{2}}$. We can now also try to perform dimension reduction via Principal Component Analysis then run LDA and QDA on the reduced data subset. We hope this will have higher accuracy than the previous two discriminant methods.

3.4 Principal Component Analysis

Principal Component Analysis (PCA) takes the first k eigenvectors associated with the k largest eigenvalues of the covariance matrix, then projects the data onto those eigenvectors. This reduces the most variance in k dimensions while retaining important qualities in the original data. We will use k=2 here because as seen in Figure 7, the eigenvalues for both the high and low polluted cities drop off immensely after k=2, thus k=2 will explain the most variance. To interpret the principal components, they are the vectors of the covariance matrix that explain the most variance. The PCA variance explained plot for some k, along with the reduced dimension PCA plot can be found below:

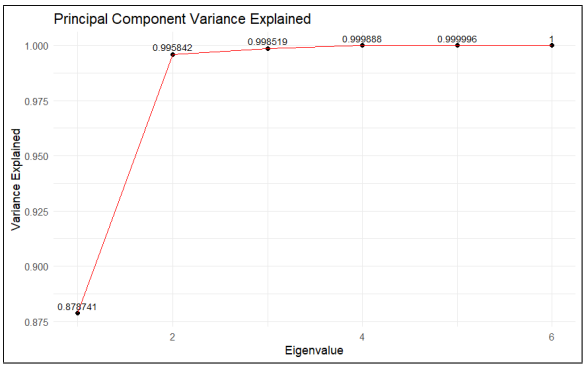

Figure 11: PCA Variance Explained Plot

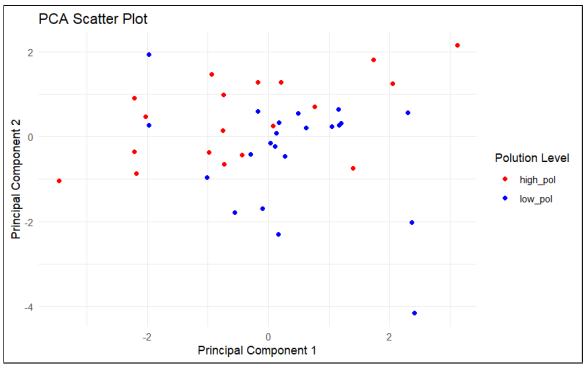

Figure 12: Reduced Dimension by PCA Plot.

As expected, most of the variance (99.58%) is explained with just the first 2 components, thus k=3 would not help much. The PCA scatter plot seems to work fairly well though there is still some overlap within the two groups. This overlap will likely make QDA and LDA work not as well. For sanity, we can check this with the accuracy and confusion matrix below:

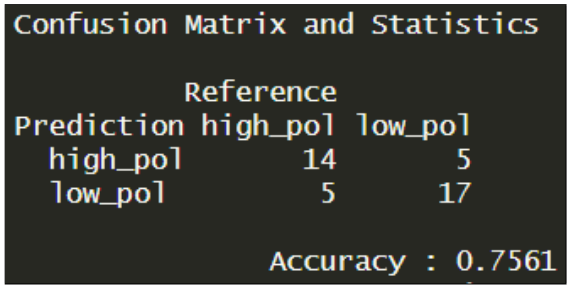

Figure 13: QDA Accuracy and Confusion Matrix (PCA)

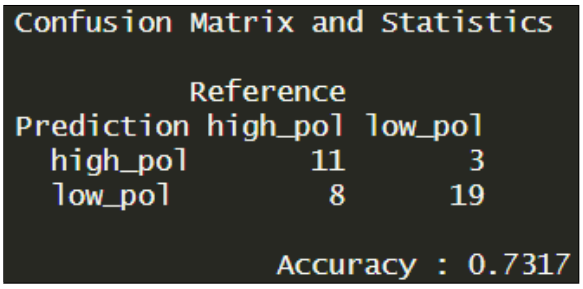

Figure 14: LDA Accuracy and Confusion Matrix (PCA)

Here we see that accuracy decreases for both high and low polluted cities, and the confusion matrix shows such errors. This is expected as due to the PCA scatter plot not separating the groups well (Figure 12). In order to predict the pollution rate with higher accuracy, we may opt for various machine learning models like the Support Vector Machine and Random Forest.

3.5 Support Vector Machine

Support Vector Machine transforms data to a feature space by including the original dimensions and introducing a new one via a kernel function, then finds the optimal decision plane to classify points based on their label. This is thus a supervised learning method that we think will work well for our data. We will opt for the 3 arguably most popular kernel functions in the linear, radial basis function, and non-linear. The accuracy rates by kernel method can be found in Figure 15.

Figure 15 shows that the RBF kernel has the largest accuracy rate with 90.24%, while the other two methods are lower. It is worth noting that the linear kernel function has an identical accuracy rate as

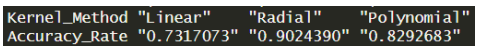

Figure 15: Support Vector Machine Accuracy by Kernel Function.

LDA on the original data. This verifies our curiosity that the data can be predicted in a more efficient manner. We can now also compare our methods to the Random Forest model.

3.6 Random Forest

The Random Forest model is a decision trees based model that predicts a class k by conditioning over the predictors as above or below a certain threshold. For example, in Figure 8, this could look something like: if temp is greater than 4 and manu is greater than 7, then it is high_pol. We centered and scaled our data to have mean 0 and variance 1. Below is the prediction rate and confusion matrix for the random forest model:

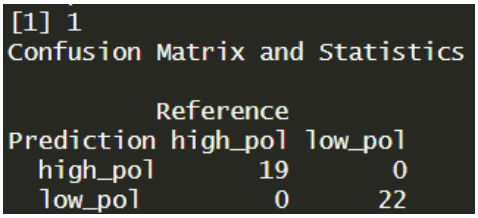

Figure 16: Random Forest Accuracy and Confusion Matrix

The Random Forest method produced an accuracy rate of 100%, denoting perfect prediction. This could be due to the decision trees presence being really beneficial.

The LDA and QDA methods worked fairly well though the random forest method worked the best. These methods give us ample information on which is best for prediction pollution rates in a city given their human and environmental variables. The following sections will discuss conclusions and discussion outlooks.

4 Discussion

4.1 Box's M Test

Box's M test indicated that the covariance matrices of the pollution data from different cities are not statistically equal. This result was expected given the bunched up behavior seen in Figure 8. Consequently, a quadratic discriminant analysis (QDA) approach was pursued to account for this variance difference.

4.2 Classification Performance

The classification performance of QDA and linear discriminant analysis (LDA) was evaluated, with QDA achieving a higher classification accuracy (87.80%) compared to LDA (73.70%). This suggests that the quadratic separation better captures the underlying data structure, likely due to the non-equal covariance matrices identified earlier.

4.3 Principal Component Analysis (PCA)

PCA was applied to reduce the dimensionality of the data. By selecting k=2, based on the steep drop in eigenvalues and the variance explained plot, the first two principal components were chosen. The PCA scatter plot (Figure 12) showed some overlap between the high and low pollution groups, reflecting the inherent complexity of the data. This overlap indicates potential challenges for both QDA and LDA, as evidenced by decreased classification accuracies 75.61 (QDA) & 73.17 (LDA) (Figures 13 and 14), when using PCA-reduced data.

4.4 Support Vector Machine (SVM)

SVM was employed with three kernel functions: linear, radial basis function (RBF), and non-linear. The RBF kernel achieved the highest accuracy rate of 90.24%, outperforming the linear and non-linear kernels. The linear kernel's accuracy mirrored that of LDA, confirming that the original data's linear separability can be efficiently utilized by simpler models.

4.5 Random Forest

The Random Forest model outperformed all other methods, achieving a perfect accuracy rate of 100%. Air pollution data likely involves non-linear relationships and interactions between various factors such as temperature, industrial activity, traffic density, and meteorological conditions. Random Forest can model these intricate dependencies more effectively than linear methods.

5 Conclusion

The analysis demonstrated various statistical and machine learning techniques in classifying city pollution levels. The Box's M test revealed a significant difference in covariance matrices, indicating the preferred use of QDA. This indication was later showed to hold true when comparing the classification rate between QDA and LDA. Moreover, PCA provided a useful dimensionality reduction but indicated overlapping group structures that posed challenges for linear methods. The data did not benefit from using linear dimension reduction as the classification rates were both lower when using QDA and LDA.

SVM with an RBF kernel showed promise, achieving a high accuracy rate, while the Random Forest model delivered perfect classification. These results suggest that more sophisticated models like Random Forest can handle the intricacies of pollution data more effectively than traditional linear methods.

6 Group member contribution

Adam Hetherwick did the introduction, and exploratory data analysis section, and we both worked together on the methods. Jose did the discussion and conclusion section, thus making the contribution equal.

7 References

1. Hand et al. (1994) data set 26, USAIR.DAT, originally from Sokal and Rohlf (1981). https://search.r-project.org/CRAN/refmans/gamlss.data/html/usair.html
2. "Sulfur Dioxide Basics." EPA, Environmental Protection Agency, 31 Jan. 2024, www.epa.gov/so2-pollution/sulfur-dioxide-basics

8 Code Appendix

Code for Data Description and Visualization along with the Methods can be found at the GitHub:

1. Adam Hetherwick's GitHub: https://github.com/AJHetherwick/Statistics_Projects/blob/main/STA_135_Final_Project.Rmd