Baruch College

# Study on The Variables Contributing to Air Pollution

STA -9705 Project Report

Divya Krishnan
5-24-2018

Contents

# *Introduction*

## Main Objective

Our main objective is to classify and analyze the variables contributing to air pollution levels in US cities. We used mean hypothesis testing and canonical correlation to test differences due to environmental factors and principal component and factor analysis to interpret weightings of the contributing variables.

## Data Description

Our dataset includes environmental characteristics and census information on a sample of 41 U.S. cities, collected from 1970. It includes one qualitative variable, city name, and seven quantitative variables with differing units. The lack of standardization in the units needs to be accounted for in the factor and principal component analysis. All four of our analysis types assume an independent and randomly sampled dataset. Considering the geographical distance between cities, we believe it is unlikely our data points are dependent, and this is a safe assumption.

We replace the individual city names with a numbered regional categorization for our mean hypothesis analysis. This the regional designation was drawn from recent census data.

| Variable | Summary | Units |
|---|---|---|
| X0 | Region in Mean Hypothesis testing, city for all other analysis types | Region, integer; or city name, string |
| X1 | SO2 Content of air | Micrograms / $m^3$ |
| X2 | Average Annual Temperature | Fahrenheit |
| X3 | Manufacturing enterprises employing 20+ workers | Integer |
| X4 | Population size | Thousands, data from 1970 |
| X5 | Average annual wind spead | mph |
| X6 | Average precipitation | Inches, yearly |
| X7 | Average number of days with precipitation | Days, yearly |

## Methodologies

### Mean Hypothesis Testing

The US is split into four regions: Northeast, Midwest, South, and the West. We used mean hypothesis tests to test if for differences in our input variables across US regions. For mean hypothesis tests we assume that all variables are distributed normally and all samples are independent. Due to the geographic spread of the cities, it is unlikely that any data points (cities) is dependent on another.

### Canonical Correlation

Canonical correlation assesses the linear relationship between a set of input and output variables. We used wanted to establish a linear correlation between SO2 air content and average annual temperature (X1, X2) and environmental factors and basic city characteristics like average precipitation and population size (X3, X4, X5, X6, X7). Canonical correlation is a generalization of multiple correlation and accounts for associations between variables. Canonical correlations can be tested for significance with MANOVA tests.

Canonical correlation is limited by the assumption that the input and output variables (X1-X2 and X3-7) are measured in the same sampling unit.

### Principal Component Analysis

We used Principal Component Analysis (PCA) to identify the principal components (PCs) that contribute most to the variance between variables. Once the principal components are identified we can find the individual contributions of each variable. With PCA, we can identify and interpret the weightings of the air pollution input variables. For PCA, we assume there is no dependence between variables and we make no grouping assumption at the before the analysis. PCA using the covariance matrix has limited interpretability if the variances per variable have a large spread, but this issue can be fixed by substituting the covariance matrix for the correlation matrix.

### Factor Analysis

Similar to PCA, factor analysis is a type of dimensional reduction. In factor analysis, we identify hidden or "latent" variables that contribute to the variance in our given variables. Latent variables are identified by their "loadings" or weighted effects on the given variables. Once these loadings are identified, we can attempt to label or name them if they have obvious common elements, like common environmental traits or city industry.

Factor analysis assumes our observations are independent and randomly sampled. The analysis is limited by the subjectivity of the interpretations. Additionally, it cannot identify causality, it only highlights correlations between similar responses.

## Air Pollution - Analysis

## Mean Hypothesis Testing with MANOVA

### Objective

We wish to see if pollution in a city is independent of where the city is placed in USA geographically.

### Analysis

We do this by classifying of the 50 cities in the dataset into 4 regions- Midwest, Northwest, South and West. Once this is done we can carry out MANOVA test to check if there exist any distinction in the pollution parameters of the regions.

The Hypothesis for our test is:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_a: At\ least\ one\ mean\ differs$

$\mu_i =$ The mean of the region of i th type.

Here $i$ is from 1 to 4 signifying the region

Assuming that the data is multivariate normal and all the observations are independent of each other.

The SAS output gives us the following test statistics:

Wilks = $\prod_{k=1}^{s} \left(\frac{1}{1+\lambda k}\right) = 0.089$

Pillai test = $\sum_{k=1}^{s} \left(\frac{\lambda k}{1+\lambda k}\right) = 1.542$

Lawley Hotelling's test = $\sum_{k=1}^{s} (\lambda k) = 4.280$

Roy's Test = $\left(\frac{\lambda 1}{1+\lambda 1}\right) = 0.729$

We note that P value for all the test is less than 0.001 so we can safely reject the Null Hypothesis implying that pollution is different for each region in USA. Thus `we can say that pollution of the city can be characterized by which region the city falls in.

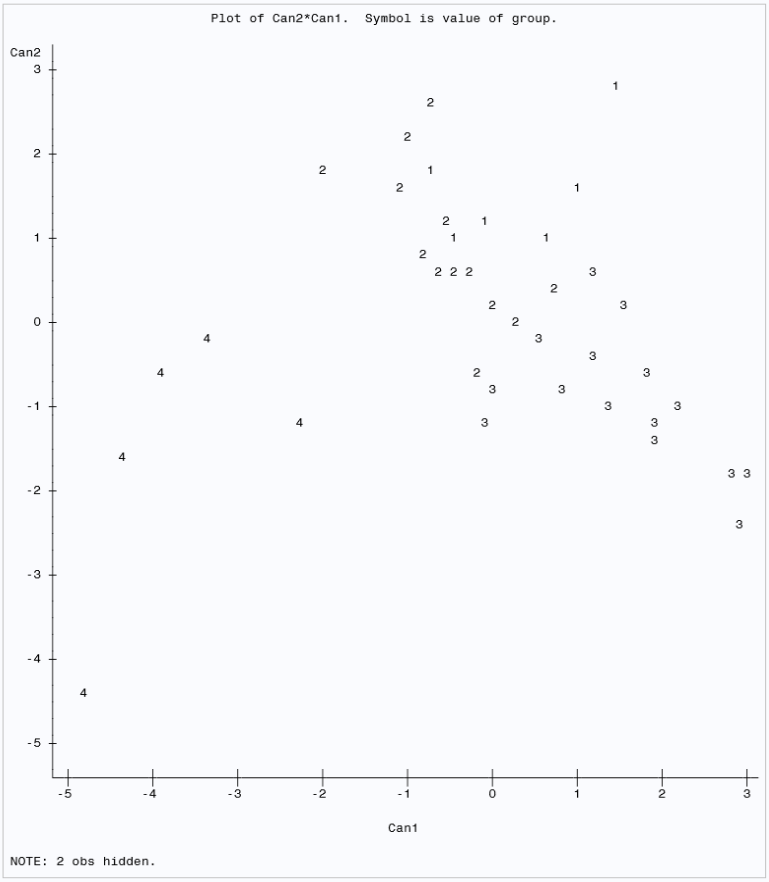
Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for group
E = Error SSCP Matrix

| | | Characteristic Vector V'EV=1 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Characteristic Root | Percent | SO2_content | Avg_temp | No_manufacture | Pop | Avg_wind | avg_precip | Avg_daysprecip |
| 2.69189585 | 62.89 | 0.00142879 | -0.00769857 | -0.00024846 | 0.00019011 | 0.01018659 | 0.03020611 | -0.00293786 |
| 1.23626507 | 28.88 | 0.00328515 | -0.02029305 | 0.00005856 | -0.00006289 | 0.06280229 | 0.00417158 | 0.00015672 |
| 0.35194040 | 8.22 | -0.01152528 | -0.02196740 | 0.00059968 | -0.00021396 | -0.00275688 | 0.01117725 | -0.00360823 |
| 0.00000000 | 0.00 | 0.00327605 | -0.00995763 | -0.00117859 | 0.00111124 | 0.00147043 | 0.00089849 | -0.00044239 |
| 0.00000000 | 0.00 | -0.00202013 | 0.02723271 | 0.00003234 | 0.00002096 | 0.02401568 | -0.01600012 | 0.01052290 |
| 0.00000000 | 0.00 | 0.00283020 | 0.01895110 | -0.00016705 | 0.00000613 | 0.11978336 | -0.00577711 | 0.00000000 |
| 0.00000000 | 0.00 | 0.00329329 | 0.01627798 | 0.00018470 | -0.00000849 | 0.00661571 | -0.00358152 | 0.00000000 |

| Multivariate Statistics and F Approximations | | | | | |
|---|---|---|---|---|---|
| S=3 M=1.5 N=14.5 | | | | | |
| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
| Wilks' Lambda | 0.08959210 | 5.62 | 21 | 89.565 | <.0001 |
| Pillai's Trace | 1.54228470 | 4.99 | 21 | 99 | <.0001 |
| Hotelling-Lawley Trace | 4.28010132 | 6.12 | 21 | 58.474 | <.0001 |
| Roy's Greatest Root | 2.69189585 | 12.69 | 7 | 33 | <.0001 |
| NOTE: F Statistic for Roy's Greatest Root is an upper bound. | | | | | |

In order to study further and have better understanding as to which regions have significantly different pollution parameters we use discriminant analysis.

From the graph it is clear that Region 4-West Coast is significantly most distinct amongst the four regions as it is spatially most separated from the others. Whereas Northeast, Midwest and South fall in the same section of the plot which can imply that they belong to the same group meaning pollution parameters for the cities in these region are somewhat the same and do not depend on their geographical location.



In order to test this further, we will be conducting contrast tests:

CONTRAST 1

$\delta = 3\mu_4 - \mu_1 - \mu_2 - \mu_3$

$H_0: \delta = 0$

$H_a: \delta \neq 0$

| MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Region 4 vs. Region 1,2 and 3 Effect<br>H = Contrast SSCP Matrix for Region 4 vs. Region 1,2 and 3<br>E = Error SSCP Matrix<br><br>S=1 M=2.5 N=14.5 | | | | | |
|---|---|---|---|---|---|
| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
| Wilks' Lambda | 0.31626957 | 9.57 | 7 | 31 | <.0001 |
| Pillai's Trace | 0.68373043 | 9.57 | 7 | 31 | <.0001 |
| Hotelling-Lawley Trace | 2.16185967 | 9.57 | 7 | 31 | <.0001 |
| Roy's Greatest Root | 2.16185967 | 9.57 | 7 | 31 | <.0001 |

Here we have p value less than 0.001 so we reject null hypothesis which confirms our earlier prediction from the scatter plot that Region 4 is different from other regions.

CONTRAST 2

$\delta = 2\mu_1 - \mu_2 - \mu_3$

$H_0: \delta = 0$

$H_a: \delta \neq 0$

Here we have p value as 0.0316 so we can say that group two and three and different from region 1 although there is certain risk in implying that.

| MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall region 2&3 vs. region 1 Effect<br>H = Contrast SSCP Matrix for region 2&3 vs. region 1<br>E = Error SSCP Matrix<br><br>S=1 M=2.5 N=14.5 | | | | | |
|---|---|---|---|---|---|
| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
| Wilks' Lambda | 0.63075171 | 2.59 | 7 | 31 | 0.0316 |
| Pillai's Trace | 0.36924829 | 2.59 | 7 | 31 | 0.0316 |
| Hotelling-Lawley Trace | 0.58540990 | 2.59 | 7 | 31 | 0.0316 |
| Roy's Greatest Root | 0.58540990 | 2.59 | 7 | 31 | 0.0316 |

CONTRAST 3

$\delta = \mu_2 - \mu_3$

$H_0: \delta = 0$

$H_a: \delta \neq 0$

Here we have p value as 0.0001 so we can conclude that pollution in region 2 and region 3 different and we have well defined regions.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall region 3 vs. region 2 Effect
H = Contrast SSCP Matrix for region 3 vs. region 2
E = Error SSCP Matrix

S=1 M=2.5 N=14.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.40750532 | 6.44 | 7 | 31 | 0.0001 |
| Pillai's Trace | 0.59249468 | 6.44 | 7 | 31 | 0.0001 |
| Hotelling-Lawley Trace | 1.45395569 | 6.44 | 7 | 31 | 0.0001 |
| Roy's Greatest Root | 1.45395569 | 6.44 | 7 | 31 | 0.0001 |

CONTRAST 4

$\delta = \mu_1 - \mu_2$

$H_0: \delta = 0$

$H_a: \delta \neq 0$

Here we have p value as 0.0834 so we conclude that pollution parameters are broadly similar for these two regions.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall region 1 vs. region 2 Effect
H = Contrast SSCP Matrix for region 1 vs. region 2
E = Error SSCP Matrix

S=1 M=2.5 N=14.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.68620526 | 2.03 | 7 | 31 | 0.0834 |
| Pillai's Trace | 0.31379474 | 2.03 | 7 | 31 | 0.0834 |
| Hotelling-Lawley Trace | 0.45728990 | 2.03 | 7 | 31 | 0.0834 |
| Roy's Greatest Root | 0.45728990 | 2.03 | 7 | 31 | 0.0834 |

CONTRAST 5

$\delta = \mu_1 - \mu_3$

$H_0: \delta = 0$

$H_a: \delta \neq 0$

Here we have p value as 0.0018 so we conclude that pollution in region 1 and 3 different and we have well defined regions.

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall region 1 vs. region 3 Effect
H = Contrast SSCP Matrix for region 1 vs. region 3
E = Error SSCP Matrix

S=1 M=2.5 N=14.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| Wilks' Lambda | 0.50357504 | 4.37 | 7 | 31 | 0.0018 |
| Pillai's Trace | 0.49642496 | 4.37 | 7 | 31 | 0.0018 |
| Hotelling-Lawley Trace | 0.98580138 | 4.37 | 7 | 31 | 0.0018 |
| Roy's Greatest Root | 0.98580138 | 4.37 | 7 | 31 | 0.0018 |

## Canonical Correlation

### Objective

We wish to test if there exists a linear relationship between SO2 content of air, average temperature recorded in city and manufacturing units with more than 20 employees, population size, wind speed, average number of days with precipitation, average precipitation.

| Variable | Summary | Units |
|---|---|---|
| Y1 | SO2 Content of air | Micrograms / $m^3$ |
| Y2 | Average Annual Temperature | Fahrenheit |
| X1 | Manufacturing enterprises employing 20+ workers | Integer |
| X2 | Population size | Thousands, data from 1970 |
| X3 | Average annual wind spead | mph |
| X4 | Average precipitation | Inches, yearly |
| X5 | Average number of days with precipitation | Days, yearly |

### Analysis

Hypothesis tests:
The Null and alternate hypothesis are:
$H_0$: $\sum yx = 0$

$H_a: \sum yx \neq 0$

$H_0$ which states that there is no linear relation between the variables whereas Ha suggests existence of linear relationship between variables.

| Multivariate Statistics and F Approximations | | | | | |
|---|---|---|---|---|---|
| S=2 M=8.5 N=8.5 | | | | | |
| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
| Wilks' Lambda | 0.00440216 | 13.37 | 40 | 38 | <.0001 |
| Pillai's Trace | 1.84154191 | 11.62 | 40 | 40 | <.0001 |
| Hotelling-Lawley Trace | 33.99556120 | 15.47 | 40 | 30.189 | <.0001 |
| Roy's Greatest Root | 26.83438117 | 26.83 | 20 | 20 | <.0001 |
| NOTE: F Statistic for Roy's Greatest Root is an upper bound. | | | | | |
| NOTE: F Statistic for Wilks' Lambda is exact. | | | | | |

From the SAS output the p-value for Wilks Lambda test(p-value < 0.0001) is much less than $\alpha = 0.05$. Hence we reject the null hypothesis and conclude that there is a linear relationship between $SO_2$ content of air, average temperature recorded in city and manufacturing units with more than 20 employees, population size, wind speed, average number of days with precipitation, average precipitation.

| | Canonical Correlation | Adjusted Canonical Correlation | Approximate Standard Error | Squared Canonical Correlation | Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 0.981872 | 0.972833 | 0.005681 | 0.964073 | 26.8344 | 19.6732 | 0.7893 | 0.7893 |
| 2 | 0.936733 | 0.909927 | 0.019374 | 0.877469 | 7.1612 | | 0.2107 | 1.0000 |

| Test of H0: The canonical correlations in the current row and all that follow are zero | | | | |
|---|---|---|---|---|
| Likelihood Ratio | Approximate F Value | Num DF | Den DF | Pr > F |
| 0.00440216 | 13.37 | 40 | 38 | <.0001 |
| 0.12253130 | 7.54 | 19 | 20 | <.0001 |

From the above output using Wilks lambda test we conclude that r1 = 0.9818 and r2 = 0.9367 are significant since p-values for both r1 and r2 are non-significant. The relative sizes of squared canonical correlations, 0.964 and 0.877 indicates two dimensions of relationship and this is confirmed by Wilks' test.

We obtain the following standardized coefficients for the two canonical variates:

**CANONICAL CORRELATION ANALYSIS**

**The CANCORR Procedure**

**Canonical Correlation Analysis**

| Standardized Canonical Coefficients for the INPUT VARIABLES | | |
|---|---|---|
| | INPUT1 | INPUT2 |
| X1 | 4.5039 | -9.5448 |
| X2 | -5.8428 | 7.7499 |
| X3 | -1.4349 | 1.3068 |
| X4 | -2.5517 | -4.6990 |
| X5 | -2.1804 | 2.4551 |
| X1X2 | 10.2009 | -6.3399 |
| X1X3 | -1.3469 | 15.7422 |
| X1X4 | -3.0335 | 0.3531 |
| X1X5 | -0.4862 | -2.6705 |
| X2X3 | 3.7700 | -14.2802 |
| X2X4 | 2.1547 | 0.1808 |
| X2X5 | 1.4008 | 2.5744 |
| X3X4 | 1.9904 | 4.2126 |
| X3X5 | -1.2491 | -3.0196 |
| X4X5 | 0.9770 | 1.9081 |
| X1SQ | -4.7737 | 0.7635 |
| X2SQ | -6.3444 | 6.0241 |
| X3SQ | 0.8639 | -1.3823 |
| X4SQ | 0.9156 | -0.4933 |
| X5SQ | 1.7453 | -1.0317 |

| Standardized Canonical Coefficients for the YIELD VARIABLES | | |
|---|---|---|
| | YIELD1 | YIELD2 |
| Y1 | -0.0638 | 1.1079 |
| Y2 | 0.9707 | 0.5379 |

## Interpretation

The variables that contribute most to the correlation between u1 and v1 are Y2 and X1,X2,X4,X5,X1X2,X1X4,X2X3,X2X4,X1SQ,X2SQ

The variables that contribute most to the correlation between u2 and v2 are Y1 and X1,X2,X4,X5,X1X2,X1X3,X2X5,X3X4,X3X5,X2X3,X2SQ

We can therefore say that $SO_2$ levels are mostly influenced by population size of the city, and enterprises employing more than 20 workers. Average precipitation and average number of days with precipitation also has some influence on the $SO_2$ level. Average annual wind speed has least contribution.

Similar analysis on average annual temperature variable(Y2) results in manufacturing enterprises employing more than 20 workers as the biggest contributor followed by the population size of the city, average precipitation, and average number of days with precipitation.

## Principal Component Analysis

### Objective

Identify the key contributors to variance in the dataset through identification of the principal components. We use an orthogonal transformation to convert a set of observations into a set of possibly correlated variables into a set of linearly uncorrelated variables. This transformation is defined in such a way that the first principal component has the largest possible variance accounting for as much of the variability of the data as possible.

### Analysis

The covariance matrix shows a large range of variances (min: average wind speed, 2.0410; max: 335,371, population in thousands), indicating one variable will dominate the first component. To remediate this, we substituted the **S** matrix for the **R** matrix to get a more balanced result. This decision even agrees to the fact that our data is not commensurate so in such a situation we must consider correlation matrix as correlation matrix standardizes the data when the scale is different.

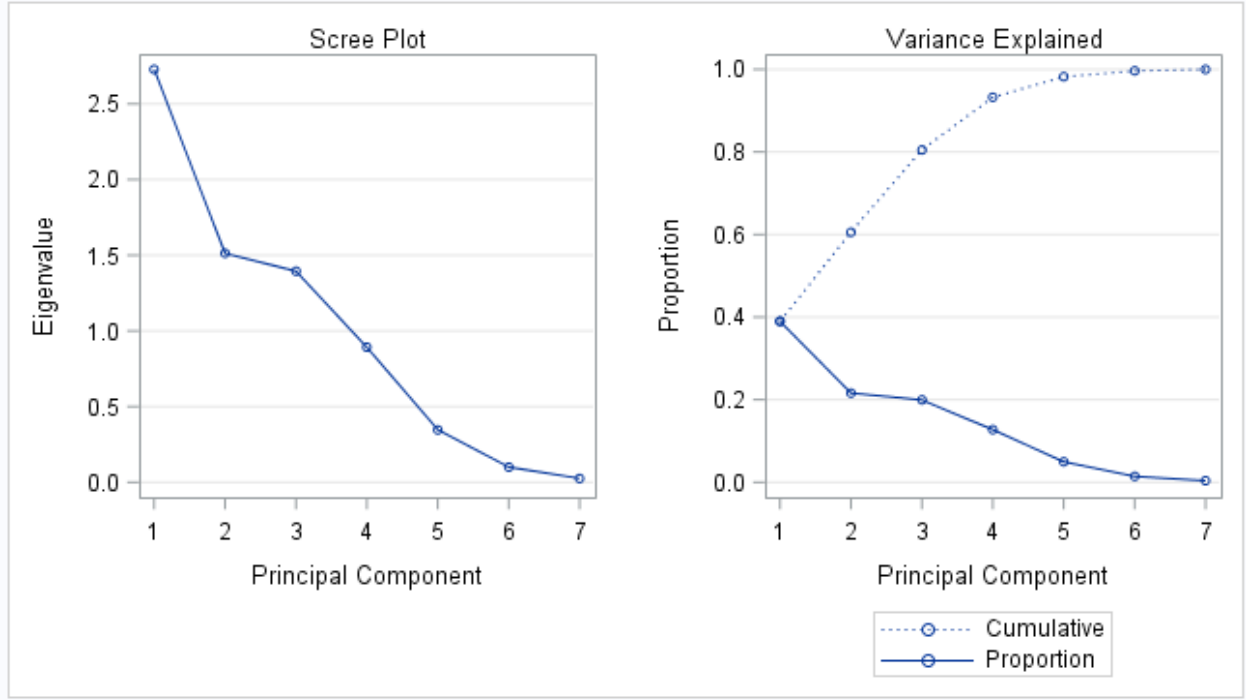
**Covariance Matrix**

| | SO2_content | Avg_temp | No_manufacture | Pop | Avg_wind | avg_precip | Avg_daysprecip |
|---|---|---|---|---|---|---|---|
| SO2_content | 550.9476 | -73.5607 | 8527.7201 | 6711.9945 | 3.1753 | 15.0018 | 229.9299 |
| Avg_temp | -73.5607 | 52.2399 | -773.9713 | -262.3496 | -3.6114 | 32.8630 | -82.4262 |
| No_manufacture | 8527.7201 | -773.9713 | 317502.8902 | 311718.8140 | 191.5481 | -215.0199 | 1968.9598 |
| Pop | 6711.9945 | -262.3496 | 311718.8140 | 335371.8939 | 175.9301 | -178.0529 | 645.9860 |
| Avg_wind | 3.1753 | -3.6114 | 191.5481 | 175.9301 | 2.0410 | -0.2185 | 6.2144 |
| avg_precip | 15.0018 | 32.8630 | -215.0199 | -178.0529 | -0.2185 | 138.5694 | 154.7929 |
| Avg_daysprecip | 229.9299 | -82.4262 | 1968.9598 | 645.9860 | 6.2144 | 154.7929 | 702.5902 |

**Pearson Correlation Coefficients, N = 41**
**Prob > |r| under H0: Rho=0**

| | SO2_content | Avg_temp | No_manufacture | Pop | Avg_wind | avg_precip | Avg_daysprecip |
|---|---|---|---|---|---|---|---|
| SO2_content | 1.00000 | -0.43360<br>0.0046 | 0.64477<br><.0001 | 0.49378<br>0.0010 | 0.09469<br>0.5559 | 0.05429<br>0.7360 | 0.36956<br>0.0174 |
| Avg_temp | -0.43360<br>0.0046 | 1.00000 | -0.19004<br>0.2340 | -0.06268<br>0.6970 | -0.34974<br>0.0250 | 0.38625<br>0.0126 | -0.43024<br>0.0050 |
| No_manufacture | 0.64477<br><.0001 | -0.19004<br>0.2340 | 1.00000 | 0.95527<br><.0001 | 0.23795<br>0.1341 | -0.03242<br>0.8405 | 0.13183<br>0.4113 |
| Pop | 0.49378<br>0.0010 | -0.06268<br>0.6970 | 0.95527<br><.0001 | 1.00000 | 0.21264<br>0.1819 | -0.02612<br>0.8712 | 0.04208<br>0.7939 |
| Avg_wind | 0.09469<br>0.5559 | -0.34974<br>0.0250 | 0.23795<br>0.1341 | 0.21264<br>0.1819 | 1.00000 | -0.01299<br>0.9357 | 0.16411<br>0.3052 |
| avg_precip | 0.05429<br>0.7360 | 0.38625<br>0.0126 | -0.03242<br>0.8405 | -0.02612<br>0.8712 | -0.01299<br>0.9357 | 1.00000 | 0.49610<br>0.0010 |
| Avg_daysprecip | 0.36956<br>0.0174 | -0.43024<br>0.0050 | 0.13183<br>0.4113 | 0.04208<br>0.7939 | 0.16411<br>0.3052 | 0.49610<br>0.0010 | 1.00000 |

The scree plot and the variance explained plots indicate that 80% of the variance can be explained by the first three principal components. So, we decided to keep the first three principal components.




The variance explained by each principal component is the same as the proportion of the associated eigenvalue which are as follows:

z1 = 38.97%
z2 = 21.60%
z3 = 19.93%

**Eigenvalues of the Correlation Matrix**

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 2.72811968 | 1.21578483 | 0.3897 | 0.3897 |
| 2 | 1.51233485 | 0.11736187 | 0.2160 | 0.6058 |
| 3 | 1.39497299 | 0.50298170 | 0.1993 | 0.8051 |
| 4 | 0.89199129 | 0.54521262 | 0.1274 | 0.9325 |
| 5 | 0.34677866 | 0.24649107 | 0.0495 | 0.9820 |
| 6 | 0.10028759 | 0.07477267 | 0.0143 | 0.9964 |
| 7 | 0.02551493 | | 0.0036 | 1.0000 |

The coefficients of the principal components is given by the eigenvectors associated with the eigenvalues of the **R** matrix. Full output for all principal components is below:

| Eigenvectors | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 | Prin6 | Prin7 |
| SO2_content | 0.489699 | 0.084576 | 0.014350 | -.404210 | 0.730394 | 0.183346 | 0.149529 |
| Avg_temp | -.315371 | -.088638 | 0.677136 | 0.185228 | 0.162465 | 0.610661 | -.023664 |
| No_manufacture | 0.541169 | -.225881 | 0.267159 | 0.026272 | -.164101 | -.042734 | -.745181 |
| Pop | 0.487588 | -.282004 | 0.344838 | 0.113404 | -.349105 | -.087863 | 0.649126 |
| Avg_wind | 0.249875 | 0.055471 | -.311265 | 0.861901 | 0.268255 | 0.150054 | 0.015765 |
| avg_precip | 0.000187 | 0.625879 | 0.492036 | 0.183937 | 0.160599 | -.553574 | -.010315 |
| Avg_daysprecip | 0.260179 | 0.677967 | -.109579 | -.109761 | -.439970 | 0.504947 | 0.008217 |

The functions for the three principal components we keep are:
$Z_1 = a_1'y$, $Z_2 = a_2'y$, …..$Z_p = a_p'y$

Such that,

$Var(Z_1)$ is the largest possible variance
$Var(Z_1) > Var(Z_2) > \ldots. > Var(Z_p)$
$Cov(Z_i, Z_j) = 0$ for $i \neq j$

So we get,

$$z1 = 00.489699(X1) - 0.315371(X2) + 0.541169(X3) + 0.487588(X4) + 0.249875(X5) + 0.000187(X6) + 0.260179(X7)$$

$$z2 = 0.084576\,(X1) - 0.088638(X2) - 0.225881(X3) - 0.282004(X4) + 0.055471(X5) + 0.625879(X6) + 0.677967(X7)$$

$$z3 = 0.014350\,(X1) + 0.677136(X2) + 0.267159(X3) + 0.344838(X4) - 0.311265\,(X5) + 0.492036(X6) - 0.109579(X7)$$

## Interpretation of the Principal Components:

The absolute value of a coefficient in a principal component shows the contribution of the corresponding variable.
The relative sizing of the variables that contributed to the first principal component were :
X3 > X1 > X4 > X2 > X7 > X5 > X6. Thus largest contribution to pollution are can be associated with the number of manufacturers in with 20+ workers, SO2 content, and population size.
As together, those three variables contributed 88.62% of the variance to the first principal component. Note that Average precipitation contributed almost nothing to PC1.
The top contributors to PC2 are average precipitation and average days of precipitation. Note that PC 2 is a shape components and represent a contrast with 3*(X3, X4) and X6, X7.

PC3 contributors were distributed between average temperature, average precipitation, population size, and average wind speed.
Thus we can conclude that no principal component is uniquely associated with a single variable and all the variables can be explained by the 3 principal components retained.

## Factor Analysis

### Objective

Factor Analysis is used to describe the variability among correlated variables and reduce the dimensions of a data set by identifying a lesser number of "latent variables" called factors.

$$yi = \mu i + \lambda i1 f1 + \cdots \lambda imfm + \varepsilon i$$

Where f1,…fm are factors and $\lambda$ij are loadings

Var(yi)=$\sum_{j=1}^{m} \lambda ij^2 + \varphi i = hi^2 + \varphi i$

Where is communality $h_i^2$ and $\varphi i$ is specific variance. Communality is the sum of the squared loadings of yi

Since the variables are commensurate, we use R in place of S.

These latent variables are equally as effective as generating the y responses, but cannot be observed or measured. Factor Analysis groups variables by grouping together variables with common response patterns, thus reducing redundancy among variables. The observed variables are modeled into linear combinations of the potential factors, aiming to find independent latent variables. For our purposes, we will see if we can reduce the number of variables that effect city pollution, as to isolate redundancies and understand the contributions of our variables more effectively.

### Analysis

To determine the number of factors we need to explain the data set, we look at the eigenvalues of the correlation matrix. Because the correlation matrix will always yield an average eigenvalue of 1, we expect that only the eigenvalues that are greater than 1 will explain our factors. The eigenvalues of the total variance explained shows that the first 3 eigenvalues are 2.7281, 1.5123, and 1.3950, and these eigenvalues explain 80.51% of the total variance, suggesting that the variables of that contribute to pollution in this data set can be explained by three factors.

| Eigenvalues of the Correlation Matrix: Total = 7 Average = 1 | | | | |
|---|---|---|---|---|
| | Eigenvalue | Difference | Proportion | Cumulative |
| 1 | 2.72811968 | 1.21578483 | 0.3897 | 0.3897 |
| 2 | 1.51233485 | 0.11736187 | 0.2160 | 0.6058 |
| 3 | 1.39497299 | 0.50298170 | 0.1993 | 0.8051 |
| 4 | 0.89199129 | 0.54521262 | 0.1274 | 0.9325 |
| 5 | 0.34677866 | 0.24649107 | 0.0495 | 0.9820 |
| 6 | 0.10028759 | 0.07477267 | 0.0143 | 0.9964 |
| 7 | 0.02551493 | | 0.0036 | 1.0000 |

Variables that were loading at less than .3 were not retained in the initial factor pattern, because we were only interested in seeing which variables are highly loading onto each factor. In the Initial Factor Pattern below, notice that there are some variables loading heavily on multiple factors. For example, Average Temperature and Average Precipitation are both loading onto Factors 2 and 3. Average Temperature is loading onto Factor 3 at .799 and moderately onto Factor 2 at .520, and Average Precipitation appears to be loading onto Factors 2 & 3, at .77 and .581 respectively. When variables load onto multiple factors, it becomes difficult to understand which variables are truly contributing to each of the latent factors.

**7 factors will be retained by the NFACTOR criterion.**

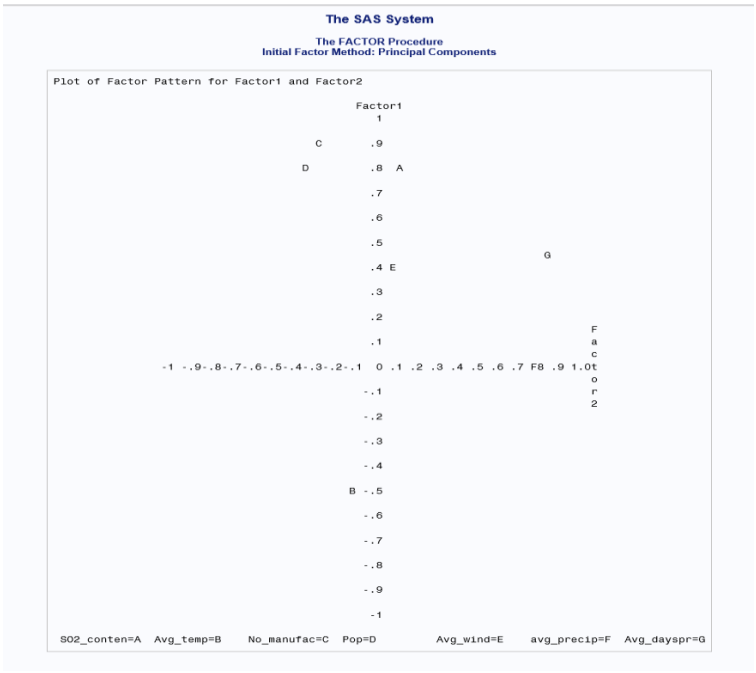
**Factor Pattern**

| | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 | Factor6 | Factor7 |
|---|---|---|---|---|---|---|---|
| SO2_content | 0.80884 | . | . | -0.38176 | 0.43011 | . | . |
| Avg_temp | -0.52090 | . | 0.79976 | . | . | . | . |
| No_manufacture | 0.89385 | . | 0.31554 | . | . | . | . |
| Pop | 0.80535 | -0.34680 | 0.40728 | . | . | . | . |
| Avg_wind | 0.41272 | . | -0.36763 | 0.81403 | . | . | . |
| avg_precip | . | 0.76969 | 0.58114 | . | . | . | . |
| Avg_daysprecip | 0.42974 | 0.83374 | . | . | . | . | . |

**Values less than 0.3 are not printed.**

In the plot of each variable against the two factors, we can see that there is difficulty in discerning the contribution of each variable on the first two factors. We can see that both Average days precipitation and population loads onto both factors 1 & 2, as shown in the factor pattern above.

PreRotated Plot:




To obtain a clearer picture of which variables are loading onto which factors, we employ a varimax rotation of the loadings to better understand which variables are loading onto each factor. A varimax rotation will scale the loadings by dividing them by the corresponding communality, thus maximizing the variance of the squared loading in each column.
For this rotation, we coded priors=smc, which reduces the correlation matrix R by replacing the diagonal of the original observed correlation matrix by the square multiple correlations. All the eigenvalues are greater than zero, R is a full rank matrix hence it is invertible.

So for this data, the SMC gave us the clearest description of which variables are loading onto which Factors. Only two of our variables are loading onto more than one factor, which is a significant improvement from the unrotated data. The rotated factor pattern below shows that there all variables seem to regress towards the axis of either one factor or another, implying that that specific variable is a contributor to that factor.

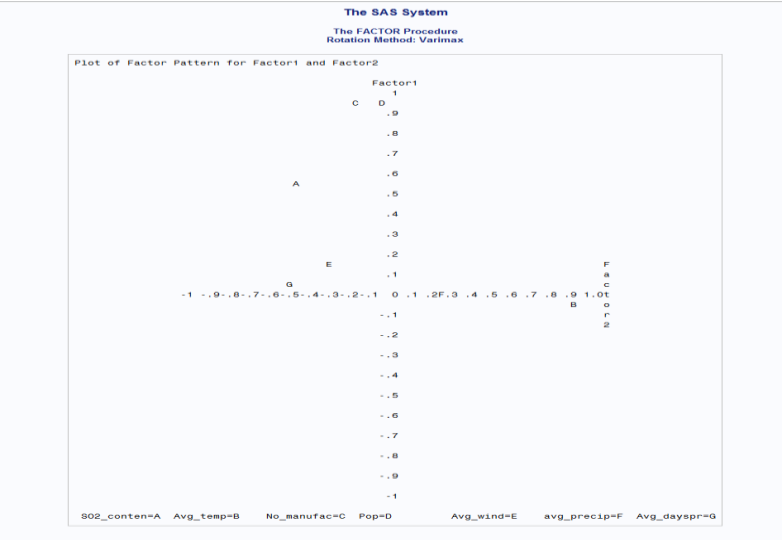
The FACTOR Procedure
Rotation Method: Varimax

| Orthogonal Transformation Matrix | | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 0.86024 | -0.49450 | 0.12429 |
| 2 | -0.38752 | -0.47564 | 0.78968 |
| 3 | 0.33138 | 0.72749 | 0.60079 |

| Rotated Factor Pattern | | | |
|---|---|---|---|
| | Factor1 | Factor2 | Factor3 |
| SO2_content | 0.55077 | -0.47428 | . |
| Avg_temp | . | 0.90284 | . |
| No_manufacture | 0.96998 | . | . |
| Pop | 0.96469 | . | . |
| Avg_wind | . | -0.32818 | . |
| avg_precip | . | . | 0.83935 |
| Avg_daysprecip | . | -0.51460 | 0.69248 |
| Values less than 0.3 are not printed. | | | |

| Variance Explained by Each Factor | | |
|---|---|---|
| Factor1 | Factor2 | Factor3 |
| 2.2026391 | 1.5145770 | 1.2393378 |

Rotated Plot Using Varimax Rotation



Variables with a loading under 0.3 threshold were excluded and are considered to be insignificant, having a complexity of <1. The highest loading for Factor 1 is No_manufacturers, for factor 2 is average temperature, and for factor 3 is average precipitation.

- Factor 1: is most heavily identified with SO2_content, No_Manufacture and Population. There are no commonly known drivers that influence all three of these variables, further investigation may be required to find commonality. Note that factor 1 has two major contributors: Population size and No. of manufacturers.

- Factor 2: Most heavily identified by average temperature, average days precipitation, and So2 content. Again, no commonly known drivers are readily associated with these three variables. Their relationship could be due to variable weather patterns.

- Factor 3: Exclusively explained by average precipitation and average days precipitation, so we can call this factor precipitation.

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0.86024 | -0.49450 | 0.12429 |
| 2 | -0.38752 | -0.47564 | 0.78968 |
| 3 | 0.33138 | 0.72749 | 0.60079 |

Rotated Factor Pattern

| | Factor1 | Factor2 | Factor3 |
|---|---|---|---|
| SO2_content | 0.55077 | -0.47428 | . |
| Avg_temp | . | 0.90284 | . |
| No_manufacture | 0.96998 | . | . |
| Pop | 0.96469 | . | . |
| Avg_wind | . | -0.32818 | . |
| avg_precip | . | . | 0.83935 |
| Avg_daysprecip | . | -0.51460 | 0.69248 |

Values less than 0.3 are not printed.

Variance Explained by Each Factor

| Factor1 | Factor2 | Factor3 |
|---|---|---|
| 2.2026391 | 1.5145770 | 1.2393378 |

Summary Conclusion

- From hypothesis test, we conclude that the pollution parameters are distinct for each of the four regions which means that pollution has geographical dependency
- From Canonical Correlation technique, the $SO_2$ levels and the average temperature have a very high correlation with the population size of the city and the enterprises employing more than 20 workers
- From Principal Component Analysis technique, the largest contributors to the air pollution levels in a city are enterprises employing more than 20 workers, $SO_2$ levels and population size
- From Factor Analysis technique, the main factor is heavily influenced by enterprises employing more than 20 workers, $SO_2$ levels and population size

Therefore, we conclude that the three factors namely, *enterprises employing more than 20 workers, $SO_2$ levels and population size* are the significant contributors the pollution levels in a city

## Bibliography


1. Data Set
   Sokal, R. R., and Rohlf, F. J. (1981), Biometry: The Principles and Practices of Statistics in Biological Research, 2nd ed., San Francisco: W. H. Freeman and Co.
2. Rencher Alvin (2002) Methods of Multivariate Analysis, 2nd ed. United States of America, A John Wiley and Sons Publication Co.

## Appendix – SAS Code

### Hypothesis Testing & Contrast Analysis

```
/* Hypothesis & Contrast Analysis */

TITLE 'Hypothesis & Contrasts';
data pollution;
        infile 'T14_12_POLLUTION.dat' delimiter='09'x;
        input city $ SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;

PROC GLM;
 CLASS group;
 MODEL SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip = group;
 CONTRAST 'Region  4 vs. Region 1,2 and 3'
 group 1 1 1 -3;
MANOVA H=group/PRINTE PRINTH MSTAT = EXACT;
RUN;

PROC GLM;
 CLASS group;
 MODEL SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip = group;
 CONTRAST 'region 2&3 vs. region 1'
 group -2 1 1 0;
MANOVA H=group/PRINTE PRINTH MSTAT = EXACT;
RUN;

PROC GLM;
 CLASS group;
 MODEL SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip = group;
 CONTRAST 'region 3 vs. region 2'
 group 0 -1 1 0;
MANOVA H=group/PRINTE PRINTH MSTAT = EXACT;
RUN;
PROC GLM;
 CLASS group;
 MODEL SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip = group;
 CONTRAST 'region 1 vs. region 2'
 group 1 -1 0 0;
MANOVA H=group/PRINTE PRINTH MSTAT = EXACT;
RUN;
PROC GLM;
 CLASS group;
 MODEL SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip = group;
 CONTRAST 'region 1 vs. region 3'
 group 1 0 -1 0;
MANOVA H=group/PRINTE PRINTH MSTAT = EXACT;
RUN;
```

## Canonical Correlation

```
/* Canonical Correlation */

TITLE 'CANONICAL CORRELATION ANALYSIS';
DATA POLLUTION;
        INFILE 'POLLUTION.dat';
        INPUT  city $  Y1 Y2 X1 X2 X3 X4 X5;
 X1X2 = X1 * X2;    X1SQ = X1 * X1;
 X1X3 = X1 * X3;    X2SQ = X2 * X2;
 X1X4 = X1 * X4;    X3SQ = X3 * X3;
 X1X5 = X1 * X5;    X4SQ = X4 * X4;
 X2X3 = X2 * X3;    X5SQ = X5 * X5;
 X2X4 = X2 * X4;
 X2X5 = X2 * X5;
 X3X4 = X3 * X4;
 X3X5 = X3 * X5;
 X4X5 = X4 * X5;
 PROC CANCORR ALL
 VPREFIX = INPUT VNAME = 'INPUT VARIABLES'
 WPREFIX = YIELD WNAME = 'YIELD VARIABLES';
 WITH Y1 Y2 ;
 VAR X1 X2 X3 X4 X5 X1X2 X1X3 X1X4 X1X5 X2X3 X2X4 X2X5 X3X4 X3X5 X4X5 X1SQ X2SQ
X3SQ X4SQ X5SQ;
RUN;
```

## Principal Component Analysis

```
/* Principal Component Analysis */

Title 'PCA';
data pollution;
        infile 'T14_12_POLLUTION.dat' delimiter='09'x;
        input city $ SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;
proc means data=pollution;
        var SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;
proc corr data=pollution;
        var SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;
proc princomp data=pollution;
        var SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;
proc princomp cov data=pollution;
        var SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;
```

Factor Analysis

```
/* Factor Analysis */

Title 'Factor';
data pollution;
        infile 'T14_12_POLLUTION.dat';
        input city $ SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;
proc factor data=pollution
method= principal
priors=smc
nfactors= 3
rotate= varimax
fuzz=.3 /*deletes factors loadings that are less than .3, so you can see which variables are loading on which
factors, this will yield missing values but paint a clearer picture of which variables are loading*/
outstat=stats
plot nplot=2
out=pollution;
var SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;
proc factor data=pollution
method=principal
priors= 1

rotate= varimax
fuzz= .3;
var SO2_content Avg_temp No_manufacture Pop Avg_wind avg_precip Avg_daysprecip;
run;
proc plot data=pollution;
plot factor1*factor2;
run;
```