A Particular Look at Urban Particulate: Analyzing Observed Variations in PM2.5 Levels

Dan Quasney

Introduction

Smog in China. Wildfires in California. Haze in New York City. All of these phenomena have garnered considerable attention over the last few years due to the fact that humans are almost always responsible for their creation, and the debilitating impacts each can have on the environment where it occurs. Though each occurs in vastly different geographic areas, they do have something in common: they all generate huge amounts of particulate matter which mixes into the air people breathe. This particulate matter, specifically particles between 0.1 and 2.5 μm called PM2.5, are especially dangerous because their small size allows them to bypass air filters and become even more deeply embedded in pulmonary tissue, perhaps even entering the bloodstream.[1] These particles are generated when exhaust emissions from combustion engines, ambient dust, and industrial exhaust react with atmospheric water and sunlight, and their composition is both seasonal and distinct to different regions in the US. For instance, contaminant levels are at their highest in the eastern United States between July and August, whereas levels peak in the West during the winter months.[2] The health risks that high levels of PM2.5 pose to the public is the primary reason for selecting this topic for the foundations project. This paper attempts to identify the features that contribute to a city's PM2.5 levels and how the relationships between those features influence observable variations in those levels.

Methodology

To identify an adequate municipality size, this study gives careful consideration not only to population, regional location, and proximity to a major metropolitan area but also to the implications of analyzing one municipal level over another. Statewide data tend to be too broad and fail to capture the

---

[1] "Understanding Particle Pollution," Environmental Protection Agency, accessed 11/14/2015 http://www3.epa.gov/airtrends/aqtrnd04/pmreport03/pmunderstand_2405.pdf.

[2] Joseph Pinto, Allen Lefohn, Douglas Shadwick, "Spatial Variability of PM2.5 in Urban Areas in the United States," *Journal of the Air & Waste Management Association (2004),* 440-449.

impacts of smaller stimuli that impart changes on an urban system. This study uses Core-based Statistical Areas (CBSAs) as the primary municipal level to be the focus of the analysis. It also assumes that PM2.5 particulate observed in a given CBSA was also generated in that same CBSA since these areas are tied to surrounding areas socioeconomically by means of transportation and commuting patterns.

Additionally, energy sources on a state level are assumed to be an adequate proxy for energy production plants within a vicinity of CBSAs. Variables within the commuting habits section together represent a valid quality measurement of transportation infrastructure for each metropolitan area. This measurement is crucial for explaining PM2.5 levels since traffic is a major source for urban air pollution. Next, industry characteristics included for each urban agglomeration account for additional pollution sources, specifically manufacturing, transportation/warehousing and utilities due to the high level of combustion required for each. Finally, demographic factors assist in observing differences in CBSAs in terms of their respective size and median income, as this study hypothesizes that denser cities emit less pollution.

Attempting to measure PM2.5 pollution across such a large area is difficult largely because of the long lifecycle of the particles[3] and limited number of measuring stations present in each CBSA. Particulate rates can be heavily impacted by variations in local geography, meaning that mountainous regions or regions with very few measuring stations may only accurately represent hyper-localized particulate levels. Consequently, the analysis incorporated only urban areas with a population size of more than 100,000 residents that are also monitored by more than 3 measurement stations, which reduced the sample size from 352 CBSAs to 98.

Results

Visualizing the data and investigating the normality of the distribution of the dependent variable provided useful insight into the shape of the data before running a regression. The yearly pollution

[3] Jeffrey R. Brook, Richard L. Poirot, Tom F. Dann, Patrick K.H. Lee, Carrie D. Lillyman, Thera Ip, "Assessing Sources of PM2.5 in Cities Influenced by Regional Transport," *Journey of Toxicology and Environmental Health (2007)*, 191-9.

mean appeared to be normally while also having a positive relationship with the CBSA population density (see Figure 1-1 and 1-2 for distributions and K-S test results). In order to select the relevant variables to be included in the model, the correlation matrix served as a critical tool for avoiding multicollinearity (See Table 1-2). As expected, some of the variables are highly correlated, especially the variables within the Industry, Energy Consumption, and Transportation sectors.

The final regression yielded an R-squared value of 0.538, meaning that approximately 54% of the variance can be explained using the chosen variables. An increase in CBSA density increases pollution levels, while higher usage of petroleum and natural gas decreases pollution, which may stem from the correlation of total fossil and coal. While establishing causality is difficult, denser urban areas are also characterized by increased industrial activity, both of which are likely to cause higher consumption of energy and a greater dependency on fuel intensive modes of transportation. The impact of transportation on observed PM2.5 levels is represented solely by the share of commuters using bikes and not by car or public transportation as one might expect (See Figure 1-3). Biking therefore appears to be the differentiator between two cities of same density despite being the lowest percentage of commuter method used on average, as car usage or public transit is not quantitatively different across similarly dense metropolitan areas.

Prevailing indicators of variations in observed levels of PM2.5 are transportation as a function of metropolitan bikeability and energy usage by percentage of petroleum and natural gas used. The increase in the bikeability of an area has been shown to reduce levels of PM2.5 when controlling for other variables. Energy usage was driven primarily by the presence of petroleum and natural gas, as those negatively correlate with coal usage - a considerably worse polluter than the previous two. With greater density in urban areas comes an increase in energy usage and, subsequently, higher PM2.5 levels.

Appendix

List of Variables and Definitions in Preliminary Regression Model			
Segment	**Variable Name**	**Description**	**Source**
Demographics	pollution_mean	Mean of Daily Pollution Levels	EPA AirData 2010
	cbsa_density	CBSA Density	US Census 2010
	population	CBSA Population	US Census 2010
	area	CBSA Area	US Census 2010
	median_income	Median Income	US Census 2010
Energy Consumption	total_fossil	Percentage of Total Fossil Fuel Consumption	US EIA 2013
	nuclear_electric_power	Percentage of Nuclear Power Consumption	US EIA 2013
	renewable_energy	Percentage of Renewable Energy Consumption	US EIA 2013
	coal	Percentage of Coal Consumption	US EIA 2013
	petroleum	Percentage of Petroleum Consumption	US EIA 2013
	natural_gas	Percentage of Natural Gas Consumption	US EIA 2013
Commuting Habits	car	Percentage of Commuters Using Cars	US Census 2010
	walk	Percentage of Commuters Walking	US Census 2010
	bike	Percentage of Commuters Biking	US Census 2010
	other	Percentage of Commuters Using Other Transportation Modes	US Census 2010
	public_trans	Percentage of Commuters Using Public Transit	US Census 2010
Industry	util	Number of Utlities Sector Establishments per CBSA	US Census 2012
	trans_ware	Number of Transportation or Warehouses per CBSA	US Census 2012
	manufact	Number of Manufacturing Facilities per CBSA	US Census 2012

Table 1-1: Variables, definitions, and data sources

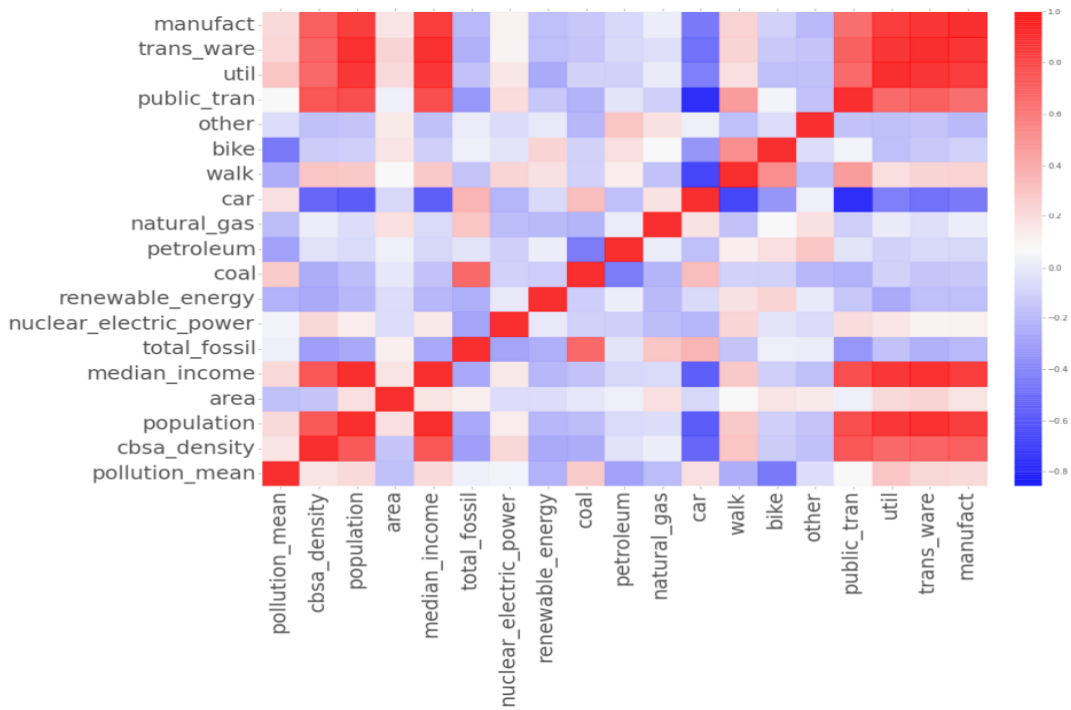

Table 1-2: Correlation matrix visualized as a choropleth

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     np.log(pollution_mean)   R-squared:                       0.593
Model:                                OLS   Adj. R-squared:                  0.513
Method:                     Least Squares   F-statistic:                     7.378
Date:                    Sat, 14 Nov 2015   Prob (F-statistic):           1.01e-09
Time:                            17:54:49   Log-Likelihood:                 32.178
No. Observations:                      92   AIC:                            -32.36
Df Residuals:                          76   BIC:                             7.993
Df Model:                              15
Covariance Type:                nonrobust
==================================================================================================================
                                                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------------------------------------
Intercept                                         -6.0671      4.314     -1.406      0.164       -14.659     2.525
np.log(cbsa_density)                               0.1255      0.040      3.104      0.003         0.045     0.206
np.log(coal + np.exp(-100))                        0.0018      0.002      0.793      0.430        -0.003     0.006
np.log(petroleum + np.exp(-100))                  -0.2767      0.151     -1.838      0.070        -0.576     0.023
np.log(nuclear_electric_power + np.exp(-100))     -0.0002      0.001     -0.362      0.718        -0.001     0.001
np.log(renewable_energy + np.exp(-100))            0.0032      0.043      0.075      0.941        -0.082     0.089
np.log(natural_gas + np.exp(-100))                -0.0898      0.034     -2.660      0.010        -0.157    -0.023
np.log(car + np.exp(-100))                         1.6852      0.899      1.874      0.065        -0.106     3.476
np.log(walk + np.exp(-100))                        0.0944      0.079      1.194      0.236        -0.063     0.252
np.log(bike + np.exp(-100))                       -0.1192      0.039     -3.060      0.003        -0.197    -0.042
np.log(public_tran + np.exp(-100))                 0.0393      0.055      0.712      0.479        -0.071     0.149
np.log(other + np.exp(-100))                       0.0578      0.052      1.120      0.266        -0.045     0.161
np.log(median_income)                             -0.0425      0.123     -0.345      0.731        -0.288     0.203
np.log(util + np.exp(-100))                        0.1065      0.062      1.711      0.091        -0.017     0.230
np.log(trans_ware + np.exp(-100))                  0.0118      0.087      0.136      0.892        -0.161     0.184
np.log(manufact + np.exp(-100))                   -0.0692      0.072     -0.963      0.339        -0.213     0.074
==============================================================================
Omnibus:                        6.129   Durbin-Watson:                   1.697
Prob(Omnibus):                  0.047   Jarque-Bera (JB):                9.545
Skew:                           0.124   Prob(JB):                      0.00846
Kurtosis:                       4.558   Cond. No.                     1.05e+04
==============================================================================
```

Regression Table 1-1: Sample without area and population

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     np.log(pollution_mean)   R-squared:                       0.538
Model:                                OLS   Adj. R-squared:                  0.517
Method:                     Least Squares   F-statistic:                     25.31
Date:                    Sat, 14 Nov 2015   Prob (F-statistic):           6.39e-14
Time:                            17:54:50   Log-Likelihood:                 26.346
No. Observations:                      92   AIC:                            -42.69
Df Residuals:                          87   BIC:                            -30.08
Df Model:                               4
Covariance Type:                nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------------------------
Intercept                          1.2286      0.170      7.233      0.000         0.891     1.566
np.log(cbsa_density)               0.0894      0.018      4.897      0.000         0.053     0.126
np.log(petroleum + np.exp(-100))  -0.2961      0.117     -2.526      0.013        -0.529    -0.063
np.log(natural_gas + np.exp(-100))  -0.0676      0.030     -2.242      0.027        -0.127    -0.008
np.log(bike + np.exp(-100))       -0.1504      0.025     -6.065      0.000        -0.200    -0.101
==============================================================================
Omnibus:                        5.906   Durbin-Watson:                   1.701
Prob(Omnibus):                  0.052   Jarque-Bera (JB):                9.426
Skew:                           0.017   Prob(JB):                      0.00898
Kurtosis:                       4.568   Cond. No.                         62.1
==============================================================================
```

Regression Table 1-2: Regression with only significant explanatory variables

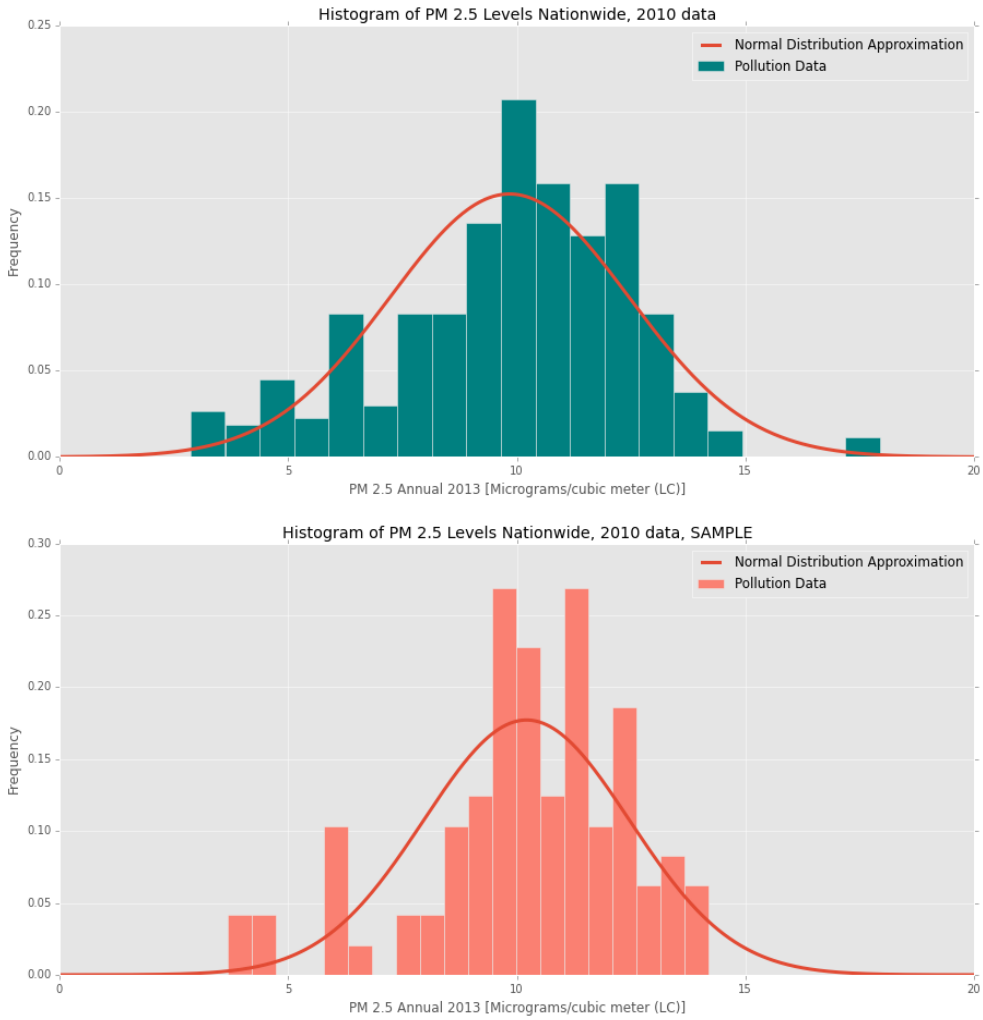

Figure 1-1: Distribution of annual PM2.5 distribution across sampled

For Kolmogorov-Smirnov test for goodness of fit:

H0: The two variables/samples are drawn from the same distribution

H1: The two variables/samples are not drawn from the same distribution

```
ks = scipy.stats.kstest(ofinterest_sample['pollution_mean'], 'norm', args=(m,sd))
print 't score:',ks[0],'p-value:',ks[1]
```

```
t score: 0.112401121469 p-value: 0.181307006788
```

The p-value is larger than α of 5%. Hence, one fails to reject H_0

Figure 1-2: Results of K-S test that compared population means

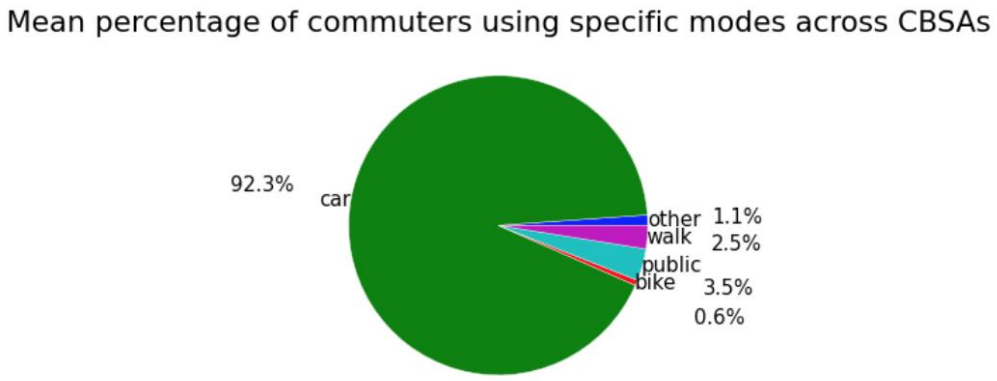

Figure 1-3: Observed commuter modal split