Air Pollution in the People's Republic of China

TF: Kela Roberts, HUID: 20192773, Anonymous Author: JM[1]

[1] *Harvard Extension, Statistics E-100*

This project measures air pollution data in five cities in the People's Republic of China: Beijing, Shanghai, Chengdu, Shenyang and Guangzhou, with data provided by the US Department of State. ANOVA tests were conducted to evaluate differences in pollution between cities. This was found to be significant at $p < 0.001$. Differences between particular cities were explored using pairwise t-tests with Bonferroni correction and TukeyHSD tests. All cities except (Shanghai, Guangzhou) were found to be significantly different at $p < 0.001$. With outliers removed, all cities including (Shanghai, Guangzhou) were found to be significantly different, though at $p < 0.01$. Quadratic and sinusoidal functions were used to explore the relationship between mean pollution values and hour of the day for each of the five cities. The association was statistically significant in all cities at $p < 0.001$, with all the five different models R squared at least > 0.53.

I. INTRODUCTION

Air and other forms of pollution are serious concerns for the inhabitants of many nations, but the severity and cost to the environment and public health is on a different scale for citizens of China. Here, air, water and soil pollution at very high intensities have been linked to increased rates of pulmonary and heart disease, cancer and premature death[1]. In the nation's capital of Beijing, the average air pollution is twice the value recommended by the World Health Organization[2] for safe levels of airborne pollutants, and on many days, visibility is severely reduced due to the density of airborne particulates.

During important national events, such as the 2008 Olympics in Beijing, or the recent APEC conference in Beijing in November 2014, the government has taken measures to reduce pollution to improve the image of the city, such as by shutting down factories, reducing the number of cars on the road, and seeding the clouds to cause rain and wind, leading to only a temporary fix.

Many covariates are believed to cause such high levels of airborne pollutants, including manufacturing and car emissions, meteorological events such as wind conditions, and regional geography. But because of the complexity of the data, as well as surrounding political controversy over the reliability of readings, even basic information, such as the difference in average pollution levels across cities, and the least and worst polluted times of day, represents information that is not clearly disseminated to the people. As a result, many conflicting beliefs about pollution exist. Extracting and revealing such information can guide citizens of countries like China on where to live and spend their time, and when is the healthiest time of day to be outside exercising.

The primary goal of this analysis was to evaluate differences in air pollution across geography and time, and to see if any trends could be teased out relating to hour of the day and average pollution level in a given city.

II. METHODS

The data was gathered from the U.S. Department of State, and is available publicly at stateair.net. It is collects every hour in five Chinese cities, where the US has an embassy or consulate. The data for Beijing, Shanghai and Chengdu includes readings from 2012 - 2014; the data for Guangzhou includes readings from 2011 to 2014; the data for Shenyang includes the two years from 2013 to 2014. The hourly readings (the column 'Value' in the dataset) produce a statistic known as the Air Quality Index (AQI), created by the US Environmental Protection Agency, which is a measure of the amounts of ground-level ozone, particle pollution, carbon monoxide, sulfur dioxide, and nitrogen dioxide in the air. The AQI index ranges from 0 to 500, with values above 50 considered increasingly unhealthy. It is not uncommon for cities in China to register values beyond 500. You can see the index in Table I.

TABLE I: EPA Air Quality Index

AQI Range	Air Condition
0-50	Good
51-100	Moderate
101-150	Unhealthy for sensitive groups
151-200	Unhealthy
201-300	Very Unhealthy
301-500	Hazardous

Pollution levels in China can exceed 500 AQI. Different countries and organizations maintain their own indices of pollution levels and its health effects.

In comparing hourly pollution values, averages of hourly times were found, i.e. the average pollution level for a given city at 9 am, the average pollution level for a given city at 10am, etc... This was computed by gathering all of the pollution values read at a given hour for a given city, then dividing by the total number of hours. In almost all cases the p values used were 0.001, except for the pairwise t-tests without outliers for mean differences between cities, in which case an association was found only at $p < 0.01$. The pollution value was mea-

sured against hour, because of the researcher's interest in investigating pollution levels across the course of a day where the conventional wisdom for best and worst times of day for airborne pollution is conflicting. Some people, for example, recommend that outdoor exercise be done in the early morning, before rush hour, while others believe that the pollution is best at night. Outliers were defined as those points beyond 1.5∗IQR + Q3.

III. RESULTS

A. Differences across cities

In evaluating the difference between pollution levels across cities, an ANOVA test found there to be significance at $p < .001$ with outliers included and excluded. In assessing the individual differences between pollution values across cities while keeping outliers, pairwise t-tests with Bonferroni corrections found that almost all cities, except the pair of (Shanghai, Guangzhou) were found to be significantly different, as shown in Table II.

TABLE II: Significance in pollution levels across cities, with outliers

	Beijing	**Shanghai**	**Chengdu**	**Guangzhou**
Shanghai	p = 0			
Chengdu	p = 0	p = 0		
Guangzhou	p = 0	p = 1	p = 0	
Shenyang	p = 0	p = 0	p = 0	p = 0

Shanghai and Guangzhou are not found to be significantly different.

Removing the most extreme outliers ($Values > 1.5(IQR) + Q3$), we get the following boxplot:

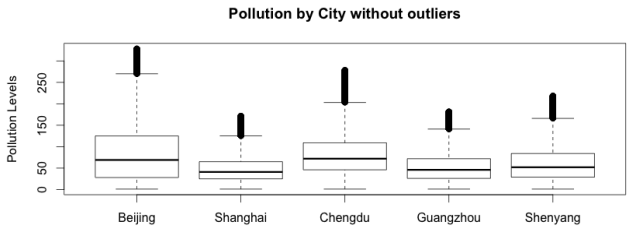

FIG. 1: Boxplot of pollution values by city, extreme outliers removed

After outliers were removed, all cities were found to be significantly different, though the p value for (Beijing, Chengdu) was only significant at $p < 0.01$, as see in Table III.

TABLE III: Significance in pollution levels across cities, without outliers

	Beijing	**Shanghai**	**Chengdu**	**Guangzhou**
Shanghai	p = 0			
Chengdu	p = 0.01	p = 0		
Guangzhou	p = 0	p = 0	p = 0	
Shenyang	p = 0	p = 0	p = 0	p = 0

Beijing and Chengdu are found to be significant at $p < 0.01$.

TABLE IV: Output of TukeyHSD with conf level 0.05, outliers included

Cities	Difference	Lower	Upper	P
Shanghai-Beijing	-36.8	-38.28	-35.32	0
Chengdu-Beijing	-1.66	-3.05	-0.27	0.01
Guangzhou-Beijing	-33.66	-34.98	-32.33	0
Shenyang-Beijing	-24.22	-25.83	-22.62	0
Chengdu-Shanghai	35.14	33.59	36.7	0
Guangzhou-Shanghai	3.15	1.65	4.64	0
Shenyang-Shanghai	12.58	10.83	14.33	0
Guangzhou-Chengdu	-31.99	-33.41	-30.58	0
Shenyang-Chengdu	-22.56	-24.24	-20.88	0
Shenyang-Guangzhou	9.43	7.81	11.05	0

B. Differences across time

In evaluating the effect of time on pollution levels, we tested the null hypotheses that there is no association between hour of the day and pollution levels. For this hypothesis, a linear model explaining pollution value by hour of the day was significant at $p < 0.001$, however the adjusted R-squared was near zero. Decomposing the hourly value by mean (i.e. the mean for each hour of the day) revealed a more interpretable account of the variation in hourly differences, which shows a small dip in the early morning hours, followed by a large one starting in the late morning and ending in the late afternoon, before rising back up again as night falls, as seen in the figure below:

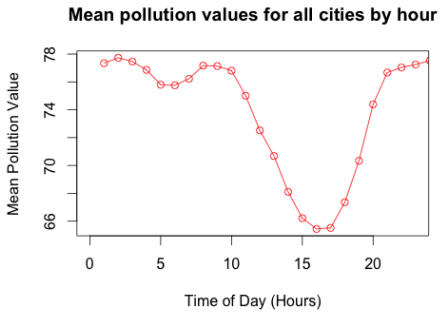

FIG. 2: Average air pollution for all cities by the hour;

However, this averaging of hours as averages of cities doesn't provide much in the way of useful information. Better is a further decomposition of the relationship between hour and value, separated out into cities, where

we observe statistically significant relationships between mean pollution value and time, and a variety of quadratic an trigonometric functions that can capture a high degree of variation in the data $Rsquared > .53$, at least, for all the models of the cities. Similar in all cities was a mid to late afternoon drop in pollution levels:

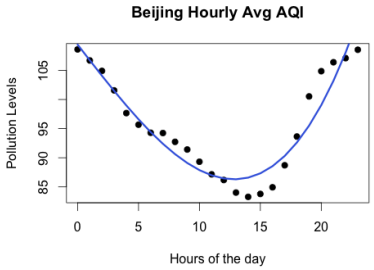

FIG. 3: Average air pollution in Beijing peaks in the middle of the night, and falls to its lowest point near 3pm. $p < 0.001$, adjusted R-squared value of 0.9. Data points in black, a curve of fit in blue.

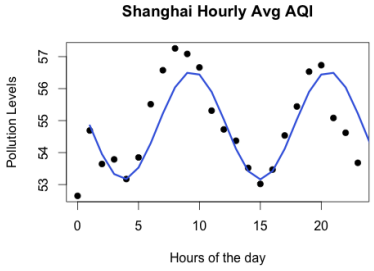

FIG. 4: Averages in Shanghai follow a sinusoidal curve, with peaks at 10 am and 9pm. Notice a similar trough to Beijing at 3 pm. $p < 0.001$, adj R-squared of 0.71.

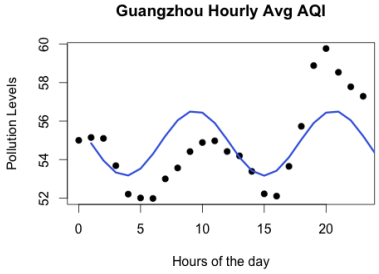

FIG. 5: Avg air pollution in Guangzhou also follows a sinusoidal curve, with troughs at 5 am and 3 pm. $p < 0.001$, adjusted R-squared of 0.71.

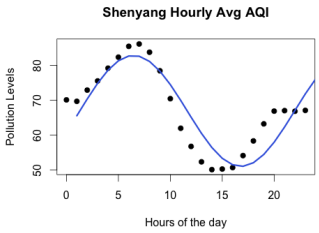

FIG. 6: Avg air pollution in Shenyang also follows a sinusoidal curve, with a min at 3pm. $p < 0.001$, adjusted R-squared value of 0.88.

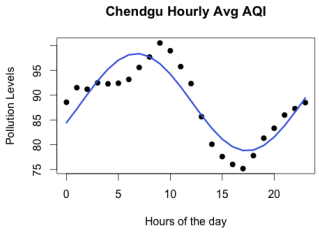

FIG. 7: Avg in Chengdu, with a max at 9 am and a min at 5pm. $p < 0.001$, adjusted R-squared of 0.80.

C. A full model

A full model, explaining pollution value against City, Year, Month, Day and Hour, was created to describe the data, with outliers included and again when excluded. With outliers included, all covariates were found to be significant ($p = 0$), however the model does not explain much, with the Rsquared value at 0.093. When outliers were removed and the model was run again, one dummy variable representing the city of Chengdu was found to be not significant at $p = 0.44$; however, it's stepwise removal from the model led to a decrease in the adjusted Rsquared value (from 0.103 to 0.096), so, the original model, outliers excluded, seen below, was considered best:

TABLE V: Full model with outliers excluded. P value of Chendgu is not significant

Variable	Estimate	Std. Error	t value	p-val
(Intercept)	1.14E+04	5.03E+02	22.551	0
Shanghai	-3.44E+01	5.53E-01	-62.325	0
Chengdu	3.93E-01	5.12E-01	0.767	0.443
Guangzhou	-3.29E+01	4.84E-01	-67.996	0
Shenyang	-2.01E+01	6.07E-01	-33.154	0
Year	-5.59E+00	2.50E-01	-22.364	0
Month	-1.74E+00	5.67E-02	-30.76	0
Day	1.24E-01	1.94E-02	6.387	0
Hour	-1.63E-01	2.47E-02	-6.601	0

Residual standard error: 50.79 on 88049 degrees of freedom
Multiple R-squared: 0.103, Adjusted R-squared: 0.1029
F-statistic: 1264 on 8 and 88049 DF, p-value: <2.2e-16

IV. CONCLUSIONS AND DISCUSSION

This study has shown that there are statistically significant differences in pollution rates across cities in China. Within the group of five cities tested, Beijing is certainly the worst polluted, followed by Chengdu, Shenyang, Guangzhou and finally Shanghai. Statistical significance was not the case for a difference in pollution values between Guangzhou - Shanghai when outliers were included during the pairwise t-tests. However, even without outliers excluded, a 95% confidence interval $(1.65, 4.64)$ shows only a slight difference. Even less was the difference for Chengdu - Beijing at 95%, with the confidence interval being $(-1.66, -3.05)$.

There are many potential reasons for these differences across cities, though such covariates as would explain them were not available to be analyzed. Some of the posited reasons are traffic, industrial production, geography, climate, population density and even architectural density. Despite creating a full model of pollution based on City, Year, Month, Day and Hour, with a $p < 0.001$, the adjusted R-squared value was only 0.103. A full regression model that includes not only City, Year, Month, Day and Hour would likely do a much better job of capturing more of the variation in the data set.

What is more surprising, and accessible given the current data, is the movement of pollution over time in these cities. Despite very high variation at given hours for a given city, an average pollution value taken to describe those hours shows a predictable fluctuation in pollution levels over the course of a day.

In the case of Beijing, the conventional wisdom that supposes lower pollution levels during the night, because of the absence of most traffic, can be rejected. Pollution across a given day in Beijing peaks in the middle of the night, and then slopes down along a predictable U-shaped curve, dropping to a minimum in the middle of the afternoon. In fact, despite differences in models for the five cities, all of them show a reliable 3 - 5 pm minimum in average pollution, the result of which is unknown but could be the subject of further study.

The other cities' time distributions were modeled with sinusoidal functions, with the functions modeling Shanghai and Guangzhou having very similar characteristics. They are both coastal cities, so perhaps the climates could be the cause of this. Chengdu and Shenyang were also modeled using sine curves.

Considering the implications of these findings, for average citizens, who can make real-time decisions concerning outdoor activity, hourly published pollution readings would be a better statistic when deciding about a time for outdoor recreation. However, these are not always available, and readings may vary in specific parts of a city. For organizations such as schools trying to schedule outdoor sporting events in advance, on the other hand, such information would be very helpful in scheduling a healthier time for outdoor activity.

A. Assumptions of the data

Also to note, despite clearing the most extreme outliers, many assumptions of normality were violated in producing these models, including a distribution of residuals that was fairly right skewed (FIG 8) as well as a plot of the residuals against predictions that contained a pattern not captured by the model (FIG 9).

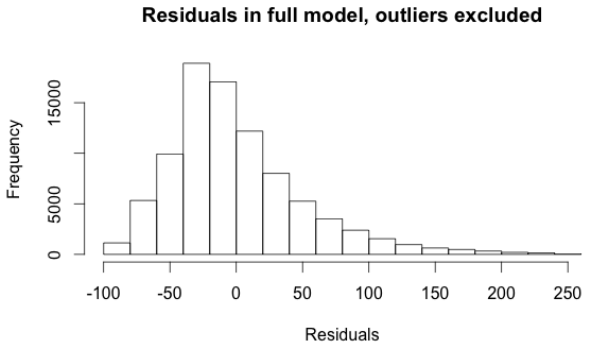

FIG. 8: Histogram of residuals shows a right skew.

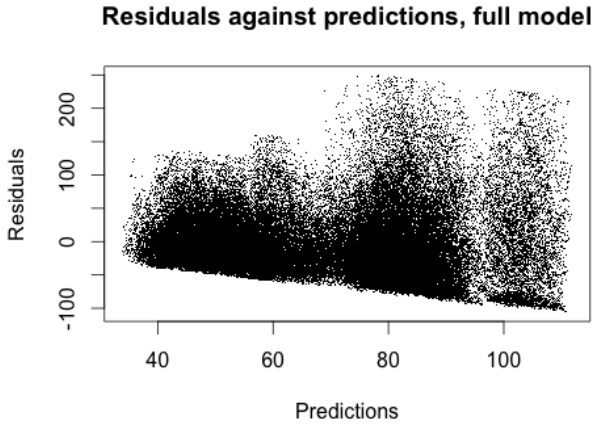

FIG. 9: Plot of residuals against predictions reveals a pattern in the data not captured by the model.

It is also worth discussing how inferential statistics bears on this data set. After all, it features hourly collections for specific cities over a 2-3 year period, depending on the city. This might seem like a population, rather than a sample. However, the data set is a sample, when we consider that the population values can vary both across time (in the past and the future), as well as across space. Because of China's constantly changing economic and industrial landscape, however, it would be difficult to credibly extrapolate beyond a few years in either direction. The argument for space is a bit more persuasive. The readings taken by the US Department of State represent single points in what are very often quite large

cities, where pollution values can vary in different parts of the same place. These differences may not always be extreme, but occasionally they are. Thus, we are making inferences about the city as a whole. In fact, one of the chief complaints by the Chinese government of the US reported statistics is this fact that they only measure at a single point in the city, whereas the Chinese have multiple stations scattered throughout. Future studies on the subject include questions such as 1) can indoor air pollution be predicted by outdoor air pollution, 2) how do US and Chinese pollution readings vary over time, and 3) what are people's beliefs about air pollution? how do they coincide with reality? Also, of course, as mentioned before, an attempt should be made to make a more complete model, including the various covariates that might influence the level of air pollution.

ACKNOWLEDGMENTS

I would like to thank the Department of State for access to their data, R and RStudio, LaTex and TexWorks, as well as Thomas Talhelm, co-founder of SmartAir Beijing for suggesting these research questions and pointing the author in the direction of the data.

[1]Xu, Xiping, et al. "Air pollution and daily mortality in residential areas of Beijing, China." Archives of Environmental Health: An International Journal 49.4 (1994): 216-222.

[2] "WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide", Global update 2005.

V. APPENDIX

There were two different kinds of outliers in this data set. The first were actual errors, such as when the machine reporting the readings was unable to capture a pollution reading. These were coded as "-999" for the Value column of the data set, and often included a "Missing" field in the Valid column. These were removed – negative pollution values are not possible.

Another kind of outlier that affected the data set were not errors, but what may be considered severe pollution events. These data points reflect extreme pollution values, far beyond the index created by the EPA. One such event occurred in January of 2012, when readings of 994 AQI were registered, which is more than 10 standard deviations from the mean. Sometimes these events coincide with the Chinese new year, when fireworks are being set off. Large outliers can be seen in the plot of Beijing data in Figure 8:

Removing these outliers, or any of the 4.4% of the Beijing data that could be considered an outlier according to the $1.5*(IQR)+Q3$ rule of thumb was tricky, because they are not errors in the data. However, leaving these values in sometimes led to violations in assumptions for tests like ANOVA. With outliers left in, the standard deviation of Beijing was more than twice the standard deviation of Guangzhou, as can be seen in Table V:

For the sake of the reliability of our tests, all hypothesis testing was conducted with outliers in the data and outliers taken out. Even if outliers removed, however, our data was not normally distributed, though the sample sizes were very high. Altogether, there were 88058 data points for all cities collected over three years.

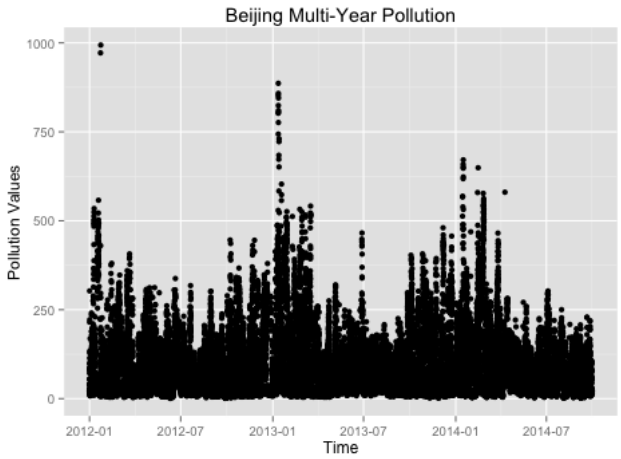

FIG. 10: Outliers visible in Beijing pollution values from 2012 to 2014. All values > 286.5, about 4.4% of the data, could be considered outliers.

City	Mean	Standard Deviation
Beijing	95.69114	89.57082
Shanghai	54.82972	46.24321
Chengdu	88.33361	60.75883
Guangzhou	54.75179	39.53022
Shenyang	67.47949	55.04043

TABLE VI: The standard deviation of Beijing is more than twice that of Chengdu

When attempting to fit a linear model to describe the seasonality in the movement of pollution over time for each city, a simple line failed to explain much of the variation of the data. Instead, quadratic and sinusoidal functions were used to approximate the change in pollution values over time. These equations can be seen below:

TABLE VII: Quadratic and sinusoidal functions to model pollution over time. All p values of 0.

City	**Equation**	**Adj.R**
Beijing	$\hat{y} = 109.42 + -2.7(t) + .005(t^3)$	0.90
Shanghai	$\hat{y} = 54.85 - 1.69 * sin(t)$	0.71
Guangzhou	$\hat{y} = 54.54 + 2.35 * cos(t)$	0.54
Shenyang	$\hat{y} = 66.90 + 15.85 * sin(t) + -1.38 * cos(t)$	0.88
Chengdu	$\hat{y} = 88.58 + 8.90 * sin(t) - 4.13 * cos(t)$	0.80