**Using Data Analysis Techniques to Evaluate Chinese Air Quality Data**

*Angel Hsu and Elaine Yu*
*Yale School of Forestry and Environmental Studies*
*May 5, 2009*

## Introduction

Beijing's air quality came on center stage when China received the 2008 Olympic bid. Realizing that they had to implement drastic measures to improve the city's air quality, the Chinese government instituted major pollution control measures in Beijing. These included shutting down factories in neighboring Hebei and Shandong provinces, decreasing the number of vehicles on the road by one-half, halting all construction projects, and increasing public transportation in the months leading up the Olympic Games. While the Chinese government touted the marked improvement these measures and other efforts had in cleaning up Beijing's air, there were questions as to the credibility of this claim. One study by Andrews (2008) claimed that the government overestimated the number of "blue-sky days" based on the Air Pollution Index[1] (API) by 22 percent in 2007 and 15 percent in 2008 through statistical manipulation, which suggested that air quality in Beijing was not necessarily improving.

Chinese statistics, particularly economic data (such as GDP growth), have been criticized for being inaccurate and misleading (Sinton, 2001; Economist, 2008). Major problems include lack of availability, interruptions in time series, and inconsistencies between data sources – issues that researchers have recognized with Chinese energy and air quality statistics (Sinton, 2001; Zhang et al., 2007; Akimoto et al., 2006). The Chinese government has acknowledged problems with persistent corruption and fraud in Chinese statistics: the *China Daily* (2003) reported some 10,000 unlawful statistics infractions per year. Although the recent Disclosure of Environmental Information Law enacted by the Ministry of Environmental Protection (MEP) in May 2008 marked a significant advance for transparency of environmental data, it remains to be seen whether the law can be effectively implemented and enforced. Data are distributed amongst various ministries and departments, often with little integration and sharing between agencies. Although the National Bureau of Statistics (NBS) is intended to be a central information body in China, it relies on its provincial counterparts to report much of the data, and is unable to independently verify much of the survey data. In 1993 several provinces failed to report their energy statistics, resulting in a major delay of reports for that year (Sinton, 2001).

Statistical methods and analytical techniques can be applied to evaluate the environmental data themselves. Regression analysis and statistical modeling have shown to be useful in measuring environmental performance. By statistically isolating effects of treatments (e.g. policies) when keeping other variables constant, relationships between outcome and explanatory variables can be estimated (Coglianse and Bennear, 2005). Matching estimator variables based on comparable models or longitudinal time-series studies could provide predictive models and confidence intervals for data. Furthermore, correlation analysis could compare data sets from different sources, helping to spot anomalies or inconsistencies in the data. Applying statistical techniques can help make sense of Chinese environmental data, which often lack coherent calculation methodologies or descriptive source information.

This study applies data analysis techniques to evaluate recent Chinese air quality data from the last five years (2004-2007) on two main fronts: first, official sources of Chinese data such as the National Bureau of Statistics (NBS) and the China National Environmental Monitoring Center (CNEMC) will be compared with independently-derived data sources to attempt to shed light on the data's accuracy and comparability. Second, we will use linear regression models using a variety of explanatory variables, including industrial coal and fuel consumption, population, gross regional product (GRP), and number of vehicles to determine the major contributing factors for a general increase in emissions of sulfur dioxide (SO2) – a major pollutant that contributes to acid rain and regional haze (EPA, 2008).

**Hypothesis**

In the assessment of the quality of official and non-official Chinese air quality datasets, we hypothesize that the underlying distributions and means of the datasets will not be significantly different from each other (H0). The alternative hypothesis (HA) is that the two datasets will be significantly different ($p<0.5$) from each other, suggesting that there may be some actual differences in the data.

In applying linear regression analysis, we predict that industrial coal consumption will be the most significant variable in explaining $SO_2$ emissions. Over 65 percent of $SO_2$ released into the atmosphere is the result of coal combustion, used to generate electricity (EPA, 2008). In China coal comprises 70 percent of electricity generation (WRI, 2009). Other sources of $SO_2$ emissions include industrial sources from the production of cement, iron and steel, and petroleum. Locomotives, large ships, and heavy trucks are also minor sources of $SO_2$ emissions.

**Methods**

Part 1: Comparison of datasets

Three datasets containing data for $SO_2$ and nitrogen dioxide ($NO_2$) emissions from 2006 for Chinese provinces and municipalities were used in this analysis. A total of 23 official provinces, four municipalities (Beijing, Chongqing, Shanghai, and Tianjin), and four autonomous regions (Guanxi, Inner Mongolia, Xinjiang, and Xizang (Tibet) were used in this study. Two of the three data sources were official Chinese government statistics from the CNEMC and the Annual Environmental Statistics Report as collected by the Institute of Public and Environmental Affairs (IPE) (http://air.ipe.org.cn). David Streets and Qiang Zhang of the Argonne National Laboratory from Chinese national energy and fuel data calculated the non-official SO2 and NO2 data from the NASA ARCTAS project. Detailed descriptions of the calculation methodology can be found in Zhang et al. (2009) and Zhao et al. (2009).

The quality of the datasets were first evaluated through simple exploratory data analysis, including pairs plots of the individual datasets against each other to obtain an idea of their correlation. Quantile-quantile (QQNorm) plots were also used to gain a better understanding of the underlying distribution of the data – diagonal QQNorm plots indicate that the distributions are fairly normal. Additional plots, including barcharts, were used to compare separation and proximity of individual data points to get a visual comparison of the data.

$SO_2$ and $NO_2$ concentration data, also collected by IPE, from approximately 115 cities in 2006 were used to gauge the approximate quality of the data from the CNEMC dataset, which is calculated only using the capital city in the related province.

Part 2: Multiple Linear Regression of SO2 emissions

To better understand some of the underlying factors for generally increasing $SO_2$ emissions in China from 2004-2007, we developed step-wise linear regression models to test the significance of selected predictor variables, such as industrial coal and fuel consumption, population, GRP, and number of cars, to determine which factors are significant in explaining the levels of $SO_2$ emissions during this period. These models were prepared by closely analyzing the distribution of the data, performing logistic transformations for data normalization, and removal of outliers to produce the best model. The "best" model was selected based on significance testing and AIC criterion to determine the goodness of fit for the particular model.

Part 3: Cluster Analysis

Based on the results of Part 2, a cluster analysis was also performed to create a visual representation of distance between variables based on several different groupings. To achieve the clustering analysis, we scaled our data and mainly used an hierarchical, agglomerative approach (using the hclust function), grouping provinces into 5 sections.

## Results and Discussion

Part 1: Comparison of datasets

A first look at the datasets for $SO_2$ and $NO_2$ revealed that the information was reported in different units – the CNEMC data were recorded in milligram/cubic meter; the ARCTAS data in gigagrams; and the IPE data in 10,000 tons. Although the CNEMC data could not be converted to comparable units, as they are concentration measures, the ARCTAS data were multiplied by 1,103 tons to make the units consistent.

Pairs plots of the datasets for $SO_2$ and $NO_2$ were then created to create visual comparisons of correlations. From Figures 1 and 2, it appears that the ARCTAS and IPE datasets are highly correlated ($r=0.86$), but the CNEMC are not highly correlated with either of the others ($r=0.26$ and $r=0.38$ for $SO_2$; $r=0.22$ for $NO_2$).

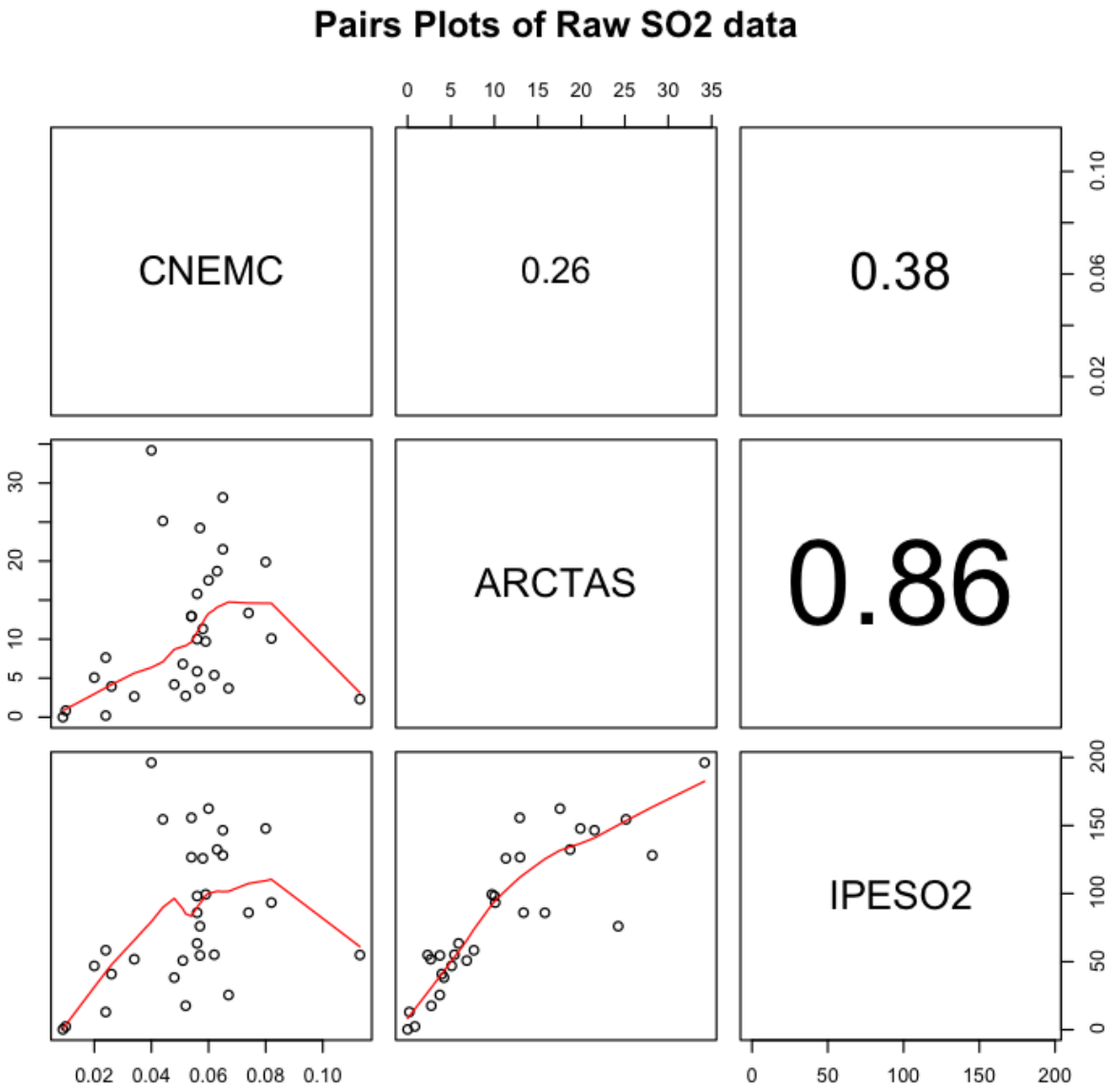


**Figure 1**. Pairs plot of $SO_2$ datasets, correlation (r) values shown in upper right-hand panels. (NB: data for China total omitted).

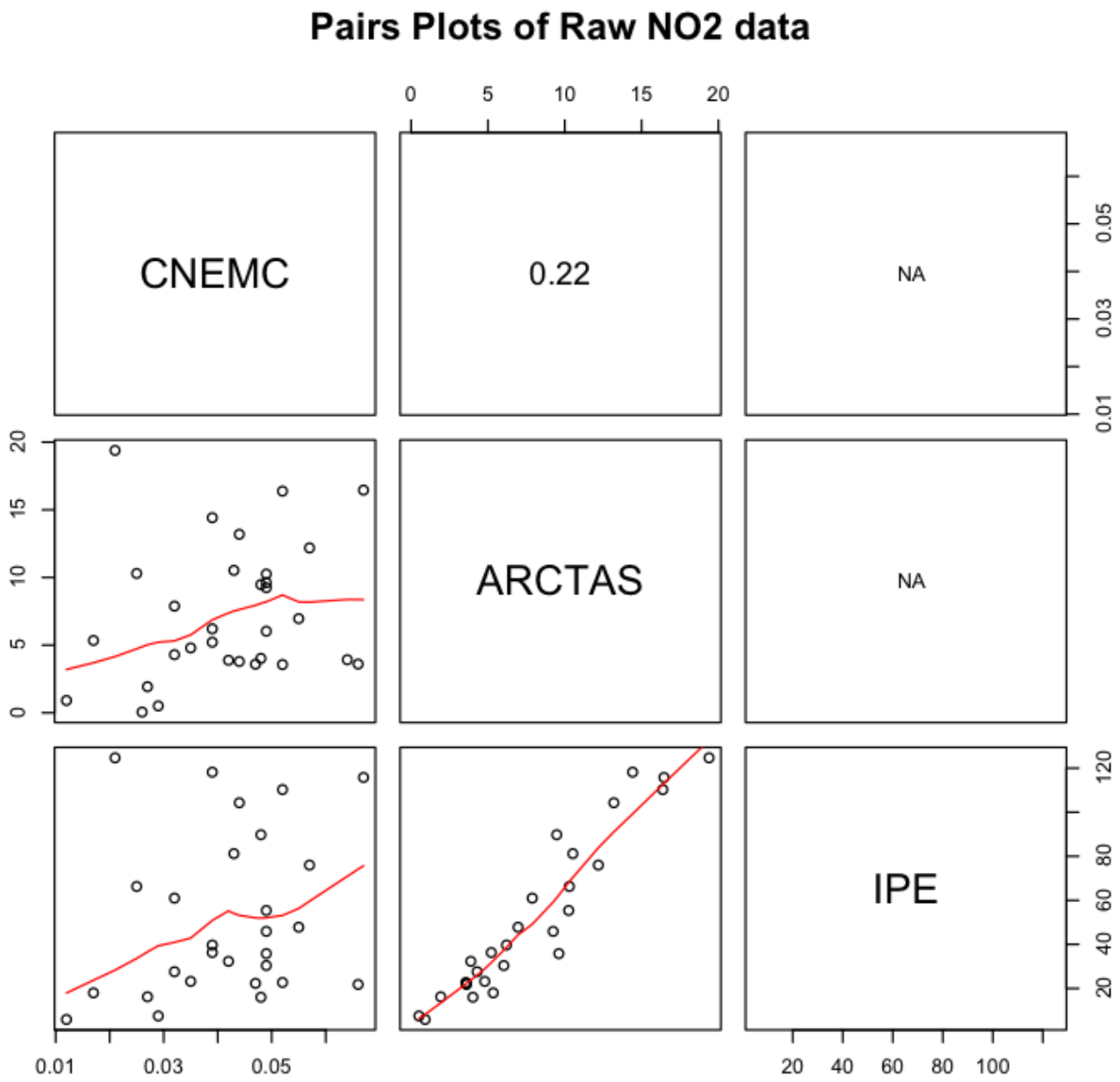


**Figure 2**. Pairs plot of $NO_2$ datasets, correlation (r) values shown in upper right-hand panels. (NB: data for China total omitted).

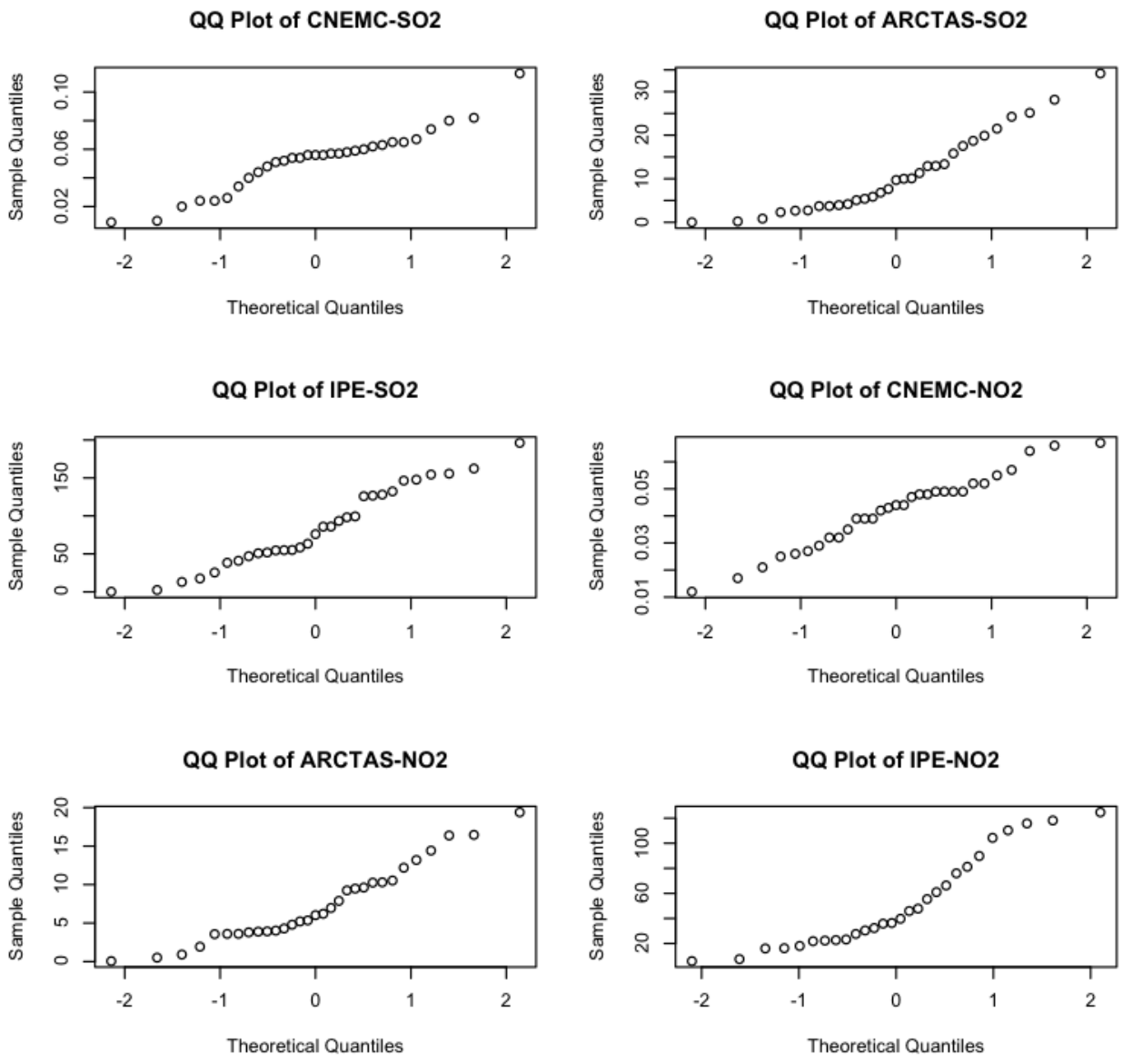


**Figure 3**. Q-Q norm plots for each dataset.

QQ-norm plots (Figure 3) of each of the datasets show that the distributions of the samples appear to be fairly normally distributed.

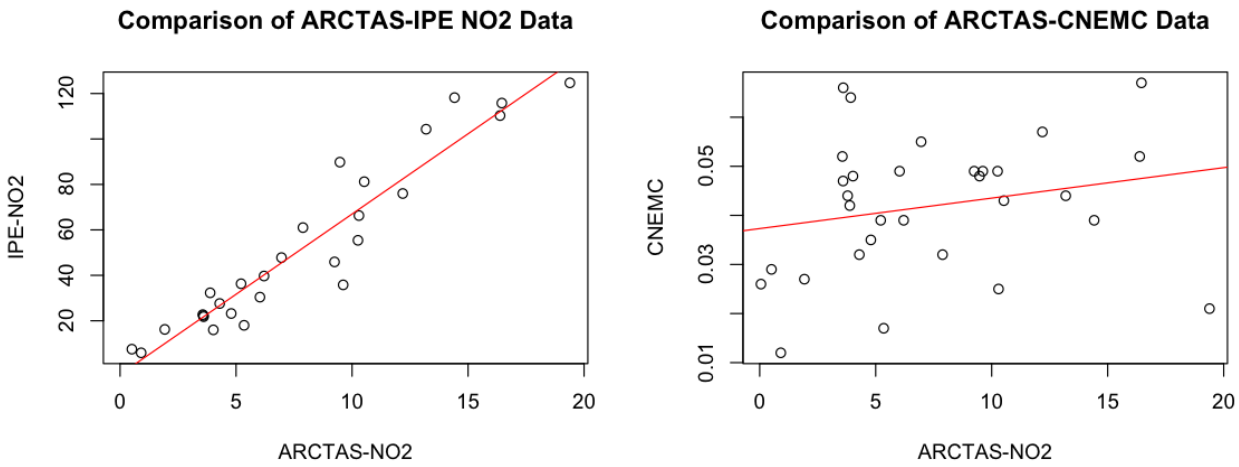


**Figure 4**. Scatterplots of ARCTAS $NO_2$ dataset with CNEMC and IPE datasets, NA values omitted.

Applying linear regression models to the $NO_2$ data, it appears that the ARCTAS dataset is a significant ($p<0.05$) variable in explaining the IPE data for $NO_2$, with an $R^2$ value of .90. The ARCTAS dataset, however, did not have a significant relationship with the CNEMC data ($p=0.3$).

Stripplots of the ARCTAS and IPE data sources for SO2 and NO2 data are in Figures 5 and 6. In most cases, such as Zhejiang (ZJ) province, the numbers for SO2 and NO2 between sources are quite different (15.8 for ARCTAS; 85.9 for IPE for SO2). There appears to be a general trend for the numbers in the IPE dataset to be larger (mean difference=72.5 for the SO2 data). T-tests comparing the two samples from the ARCTAS and IPE datasets for both the $SO_2$ and $NO_2$ data and yielded p-values less than 0, which suggests that the means of the two data sources are significantly different.

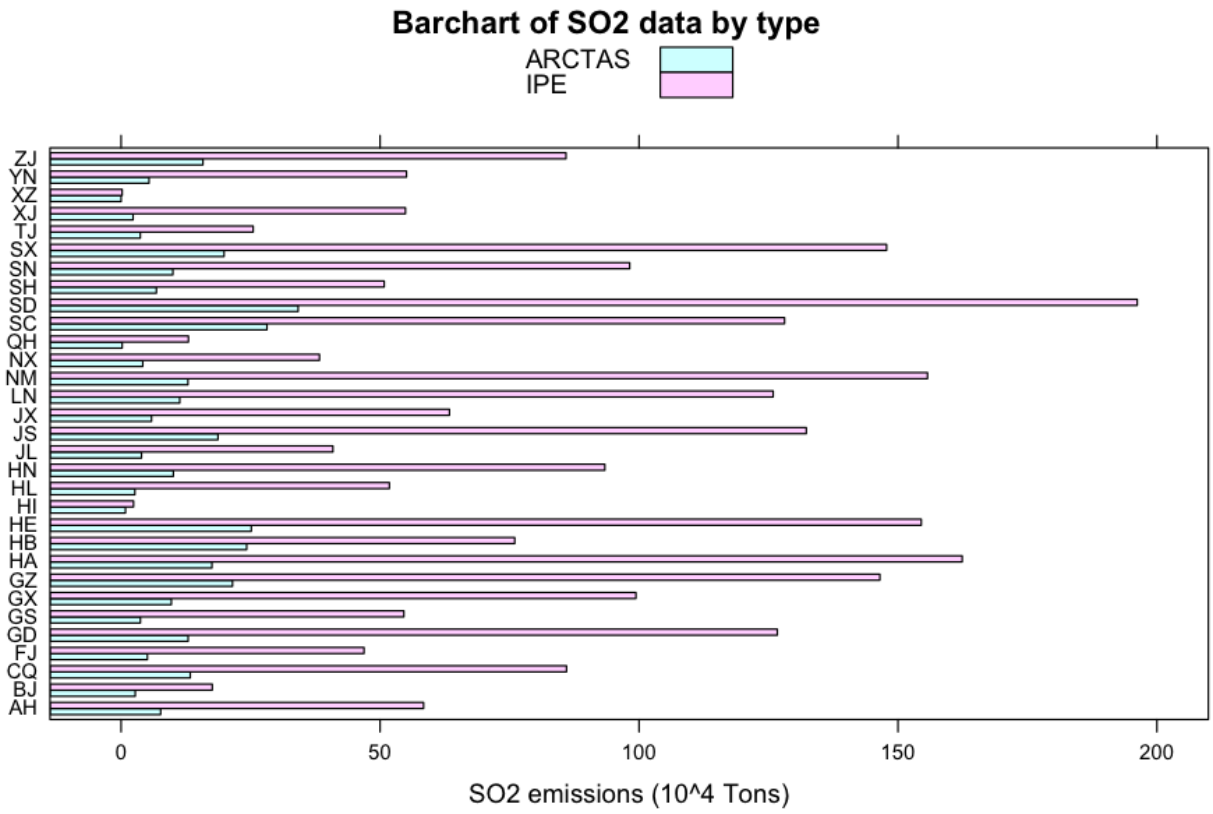


Figure 5. Barchart of $SO_2$ emissions from both the ARCTAS and IPE sources.

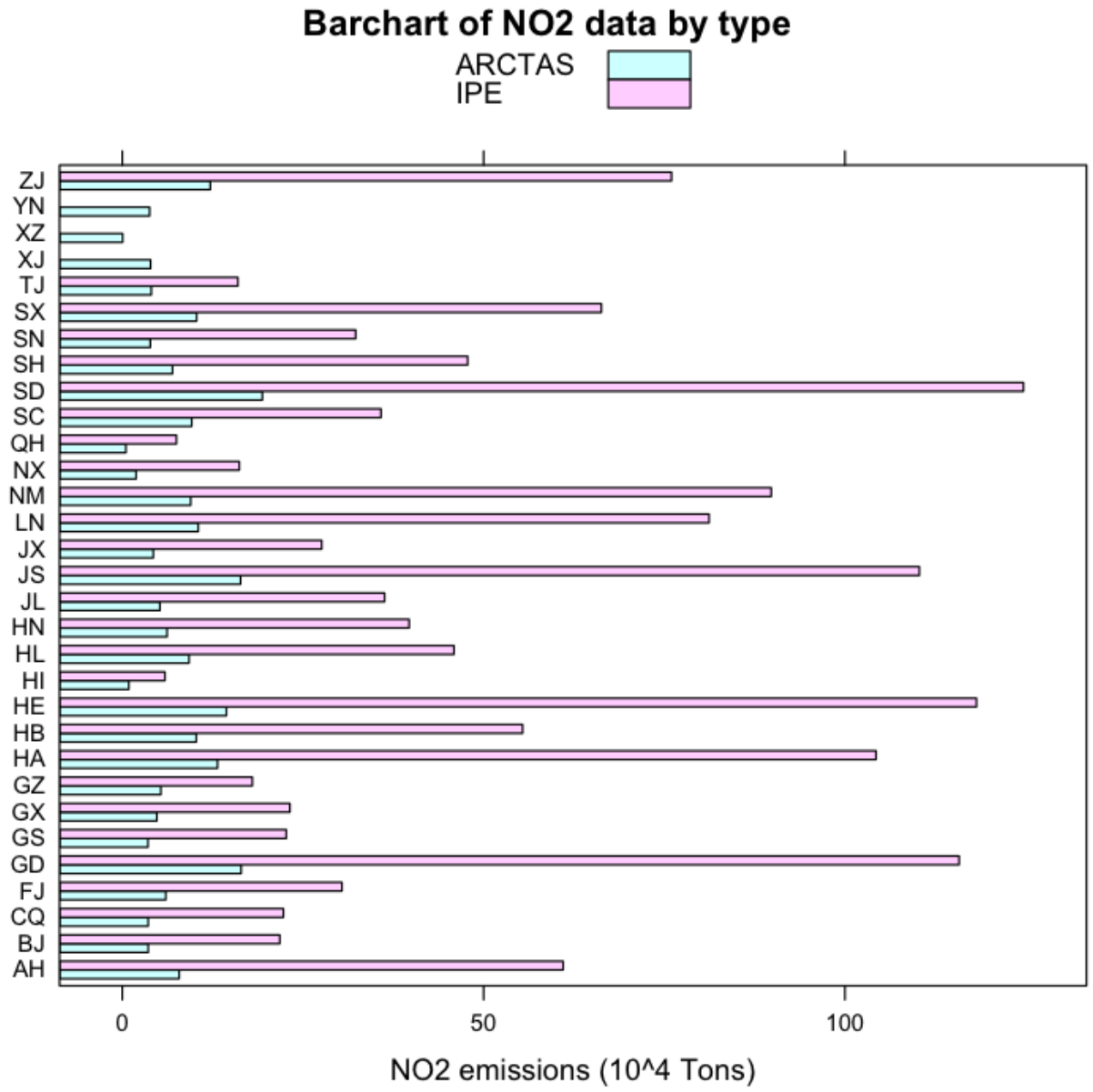


**Figure 6**. Barchart of $NO_2$ emissions from both the ARCTAS and IPE sources.

Because the CNEMC dataset is of $SO_2$ and $NO_2$ concentrations, additional concentration data from cities within the provinces were used to gain a sense of how accurate a reflection these data points might be for air quality in the province as a whole. Figures 7 and 8 show these plots. Normalizing the data per capita may provide an even accurate picture of how the $SO_2$ and $NO_2$ concentrations for capital cities relate to the data from the other cities; however, this population data for the cities was not available at the time of analysis.

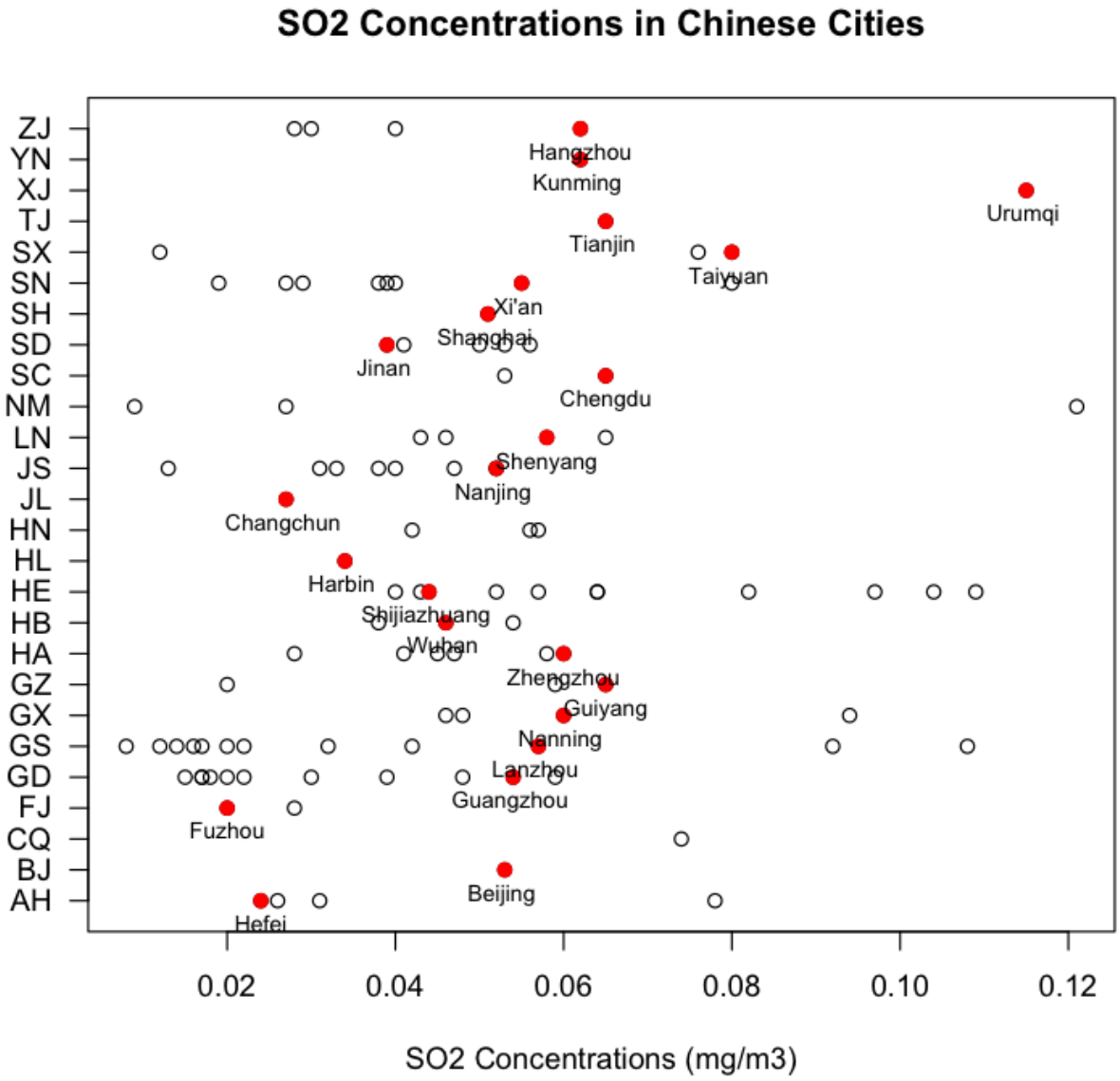


**Figure 7**. Plots $SO_2$ concentration for Chinese cities by province. Red circles correspond to values from the capital cities in each province.

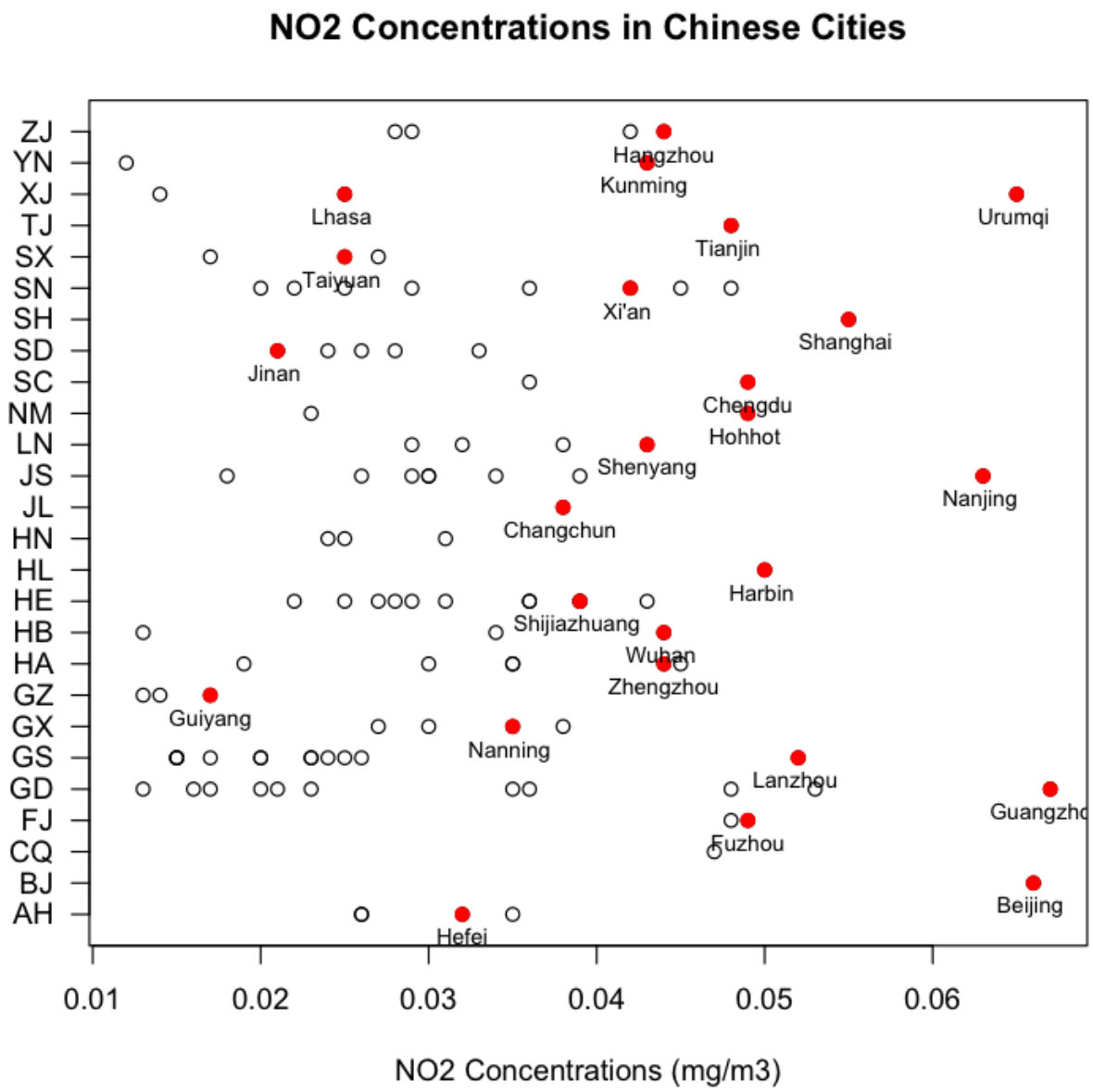


**Figure 8**. Plots $NO_2$ concentration for Chinese cities by province. Red circles correspond to values from the capital cities in each province.

Part 2: Multiple Linear Regression for $SO_2$ Emissions in China

A time-series plot of $SO_2$ emissions from 2004-2007 (www.air.ipe.org) in Figure 9 show general fluctuations in the data across time. $SO_2$ emissions tend to flucuate in a similar direction for each year across provinces. The rather extreme dips in $SO_2$ emissions in Guangdong and Guizhou provinces between 2005 and 2006 may lead to some suspicions in the accuracy of the data reported in those years. Changes in calculation or reporting methodology may be one

reason for this dip; however, the actual reason for these drops are currently unknown. A more probable explanation could be due to improved $SO_2$ policy control. According to Jiming Hao, an air pollution control engineering expert and professor from Tsinghua University, annual average concentrations of $SO_2$ emissions is decreasing overall in China according to data from 1990-2005, but $SO_2$ emissions in China overall are still 4-6 times the $SO_2$ level of developed nations (Hao 2009, manuscript in preparation).

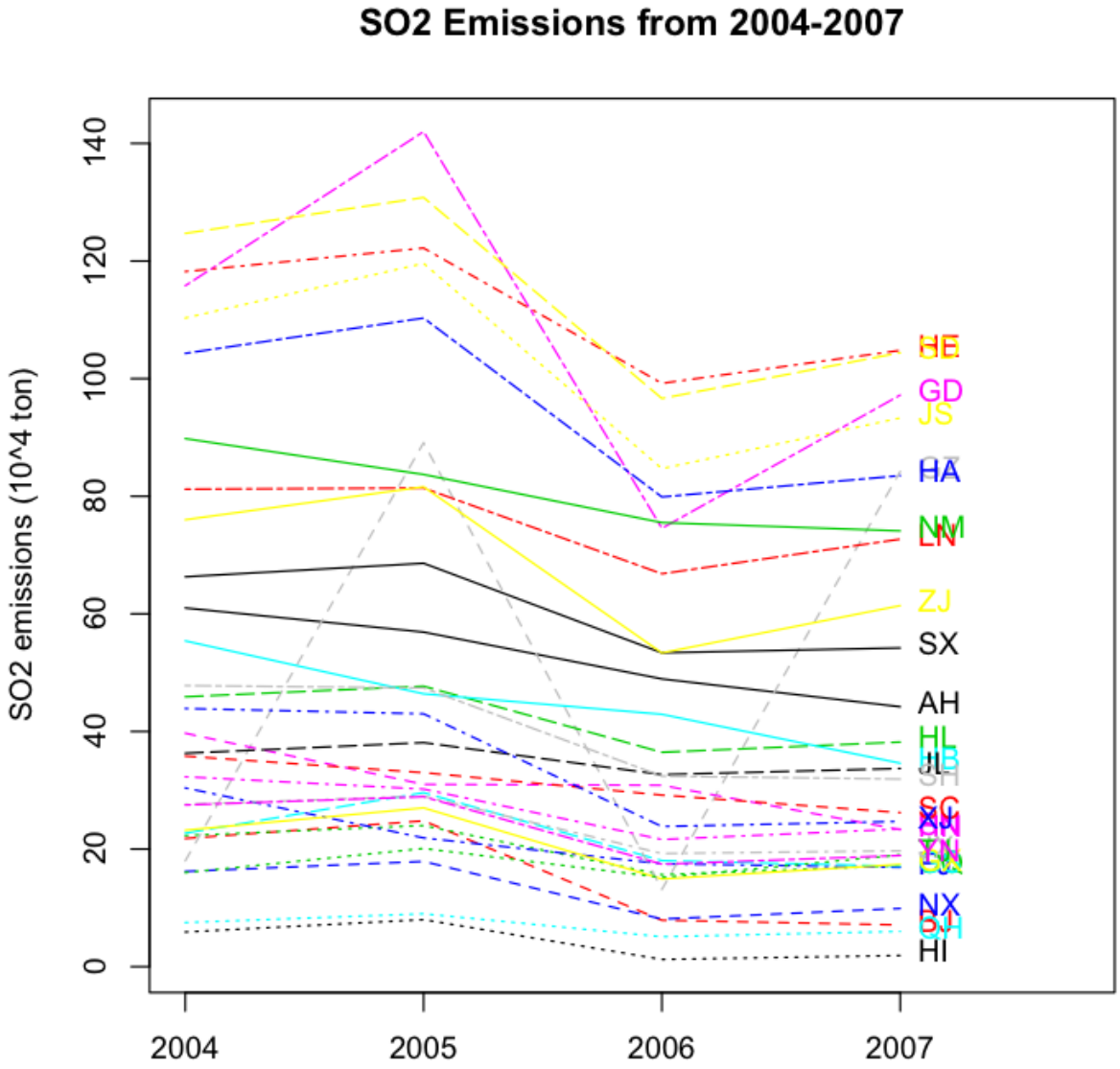


**Figure 9.** $SO_2$ emissions from 2004 - 2007.

The variables used in the linear regression modeling are as follows:
area = square kilometers
pop = year-end population (10,000 persons) for 2004-07
icc = industrial coal consumption (10^4 tons) for 2004-07
ifc = industral fuel consumption (10^4) for 2004-07
grp = gross regional product (100 million yuan) for 2004-07

so2 = sulfur dioxide emissions (10^4 tons) 2004-07
car = number of civil motor vehicles (10,000 units) 2004-07

Figure 10 shows histograms of the variables of interest for the development of a linear regression model. These histograms were analyzed to determine whether the underlying distributions of the data normal. In cases where appropriate, we applied a logistic transformation so that those variables became parametric.

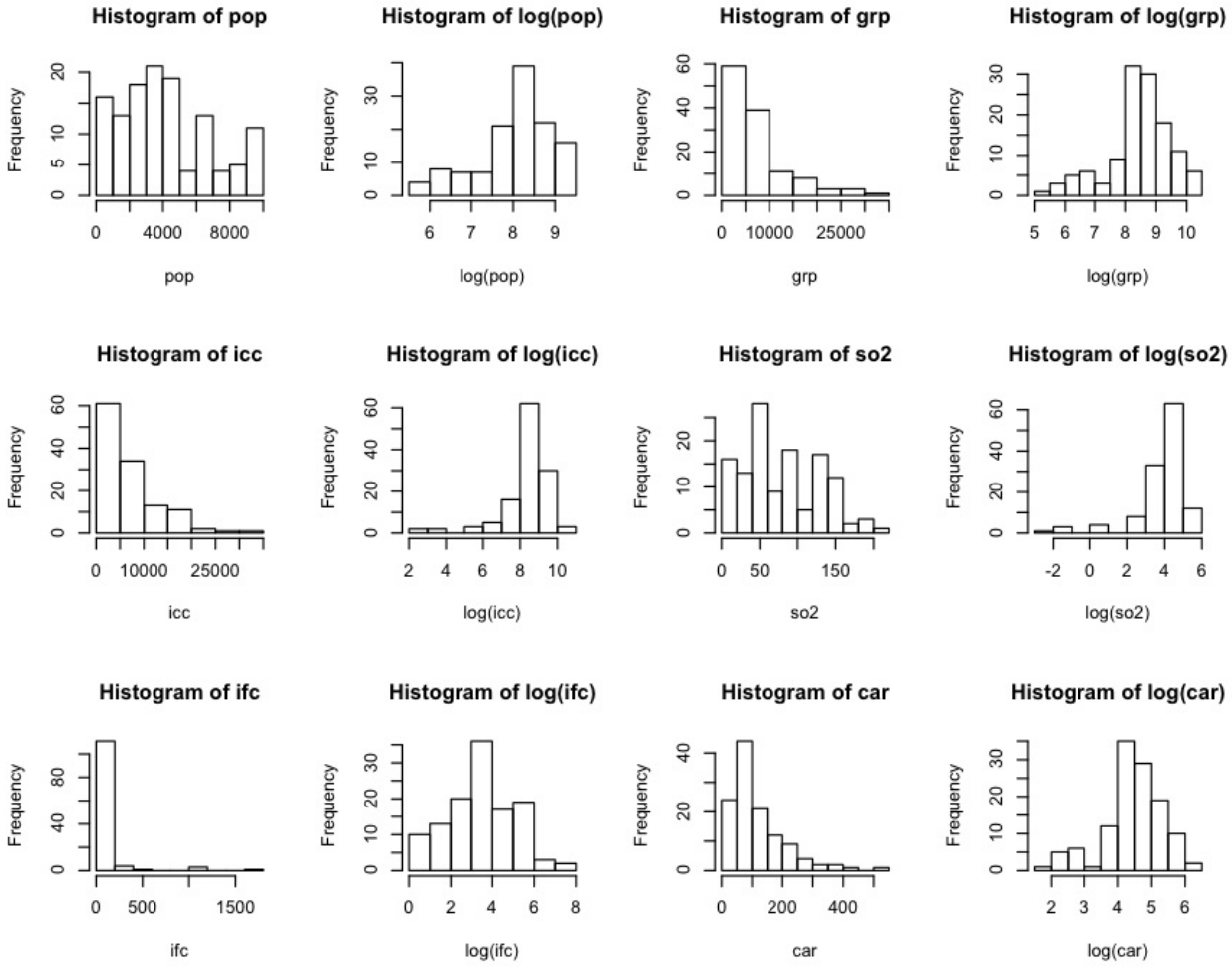


**Figure 10.** Histograms of varaibles of interest.

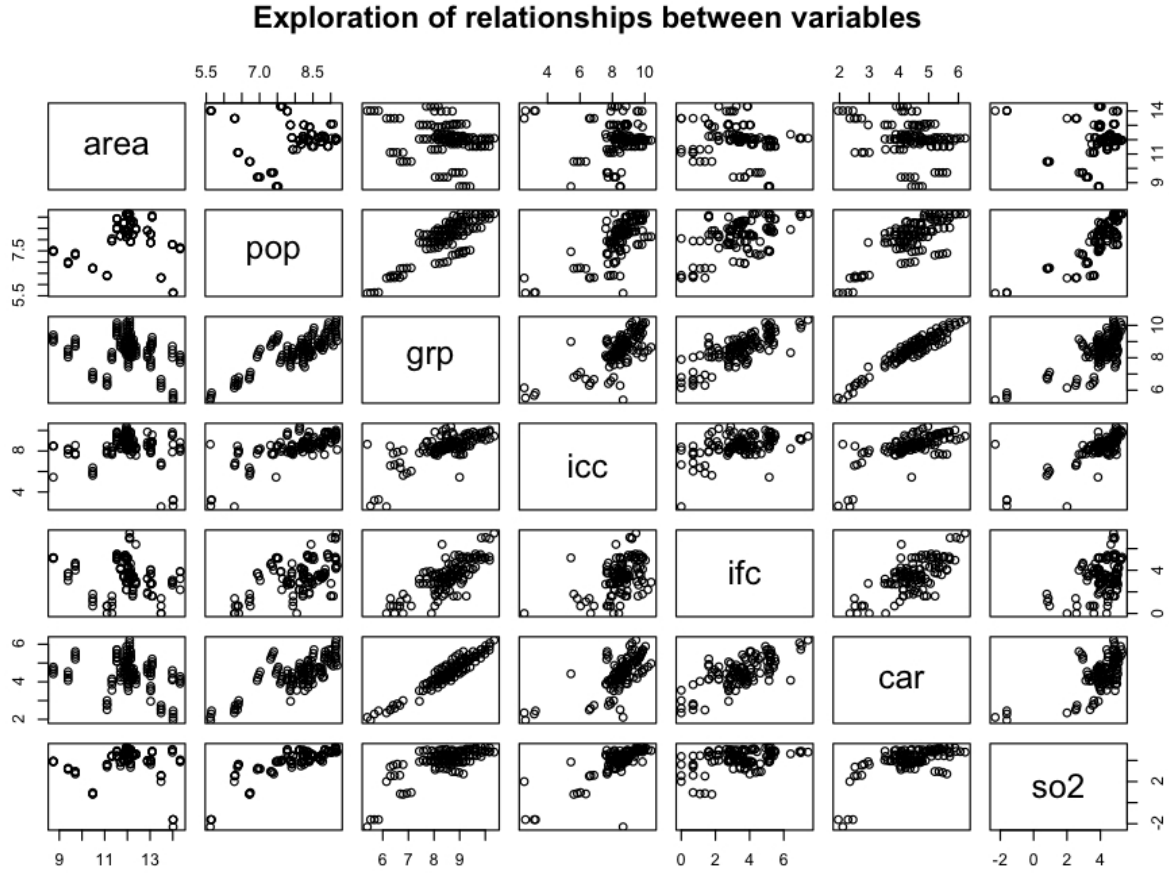


**Figure 11.** Scatterplot matrix of logistically-transformed data.

We then looked at the data more closely in terms of paired relationships (Figure 11). Figure 12 illustrates an example of how we analyzed a pair of variables using industrial fuel consumption and $SO_2$ emissions. We examined individual paired plots that were logistically transformed and then examined potenital outliers using regression diagnostics. We further went back to the dataset to analyze influential outlying data points individually to determine their candidacy for omission and removed them according.

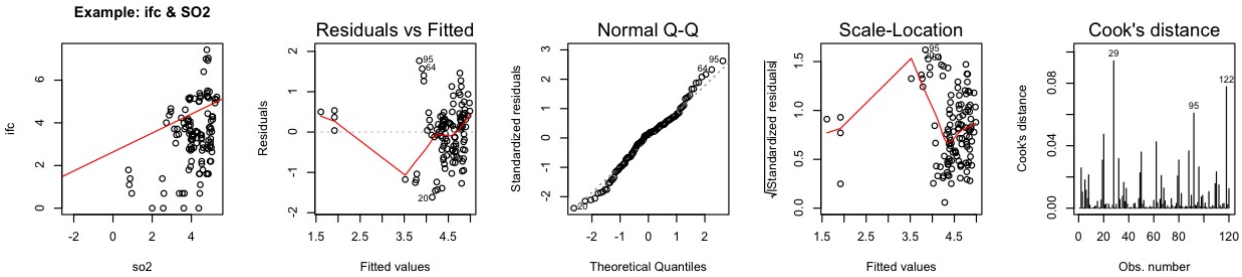


**Figure 12**. Example of analysis for various paired variables. In this example, we show analysis of industrial fuel consumption and $SO_2$ emissions.
1st panel (left-most): plot of logistically transformed $SO_2$ and industrial fuel consumption data [plot(ifc~$SO_2$)]

2nd panel (second left-most): Diagnostic plot, residuals against fitted values.
3rd panel (middle): Diagnostic plot, normal probability plot of residuals.
4th panel (second right most): Diagnostic plot, Scale-Location to determine consistency of variance.
5th panel (right most): Diagnostic plot, Cook's distance.

After outliers were omitted, a scatterplot matrix of the data was created without outliers to serve as a check to determine whether greater normality was achieved. We then created various models to best explain $SO_2$emissions in China [See **Appendix II**]. Because $SO_2$ emissions are spatially distributed, we thought that the emissions data ought to be normalized according to the land area of each province. Using the area data, we created the new variable $SO_2$ per square kilometer ($SO_2$/area). Explanatory variables used in our experimental models included "car", "icc", "ifc", "grp", and interactions between variables. After testing several dozen models using AIC and ANOVA to test each model's predictive power, we determined the following model best fit our $SO_2$ data for 2004-2007:

```
> mod21.1 = lm((so2/area) ~ ifc:grp:pop + icc:grp:pop, dat=lo)
```

Using the population (pop) and gross regional product data (grp), we developed interaction terms for industrial fuel consumption (ifc:grp:pop) and industrial coal consumption (icc:grp:pop).

```
> summary(mod20.1)

Call:
lm(formula = (so2/area) ~ ifc:grp:pop + icc:grp:pop, data = lo)

Residuals:
      Min         1Q    Median         3Q       Max
-0.079814 -0.027382 -0.002850  0.020846  0.094532

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.100607   0.043008   2.339   0.0213 *
ifc:grp:pop -0.003282   0.002283  -1.438   0.1536
grp:pop:icc  0.050743   0.005050  10.047   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03808 on 102 degrees of freedom
Multiple R-squared: 0.5874,    Adjusted R-squared: 0.5793
F-statistic:  72.6 on 2 and 102 DF,  p-value: < 2.2e-16
```

The summary of the model above indicates that the explanatory interaction variable icc:pop:grp (interaction icc:pop:grp = product of the icc, pop, and grp) is highly significant with a positive coefficient of 0.050743 and a p-value less than 0.001. This strongly supports our hypothesis that coal combustion is a leading source of $SO_2$ emissions in China. Futhermore, the interaction of the three variables suggests that icc, pop and grp may be dependent on one another or similary correlated. Increasing population may result in rising demand for energy, which would lead to an increase in industrial coal consumption for power and electricity generation - the primary coal-consuming sector in China (WRI, 2009). Gross regional product may also increase as industries per province grows, indicating that growth in economic output depends on increasing coal consumption to energize production. We will further examine these relationships using clustering analysis in the next section.

The explanatory interaction variable ifc:grp:pop was not statistically significant and carries a negative coefficient of -0.003. The negative coefficient, which suggests that the ifc:grp:pop interaction has a slightly negative impact on $SO_2$ emissions. This could be due to a shift in a province's industrial energy mix from a higher sulfur fuel, such as petroleum, to a fuel such as natural gas, which has lower sulfur content than petroleum. For instance, in the United States, the average sulfur dioxide emissions for natural gas fired-generation is 0.1 lbs/MWh - about 120 times less than oil-fired energy generation (EPA, 2007).

We were also initially surprised that civilian motor vehicle ownership (e.g. the "car" variable) did not bear sufficient significance to be included in the model, given that increasing vehicle populations in China have often been cited as cause for increasing overall air pollution emissions. It is possible that a freight traffic variable including data for the number of heavy trucks would have been a better explanatory variable since it is is likely that heavier vehicles such as cargo trucks emit more $SO_2$ than private vehicles.

## Part 3: Cluster Analysis

Cluster analysis provides the opportunity to assign cities and provinces into groups based on distributional similarity for various variables. We scaled our data and mainly used an hierarchical, agglomerative approach (using the hclust function) to group and cut provinces and cities into five categories (See **APPENDIX III**).


**Figure 13**. Map of Chinese provinces. Source: http://www.maps-of-china.net/images/chinamap.gif/.

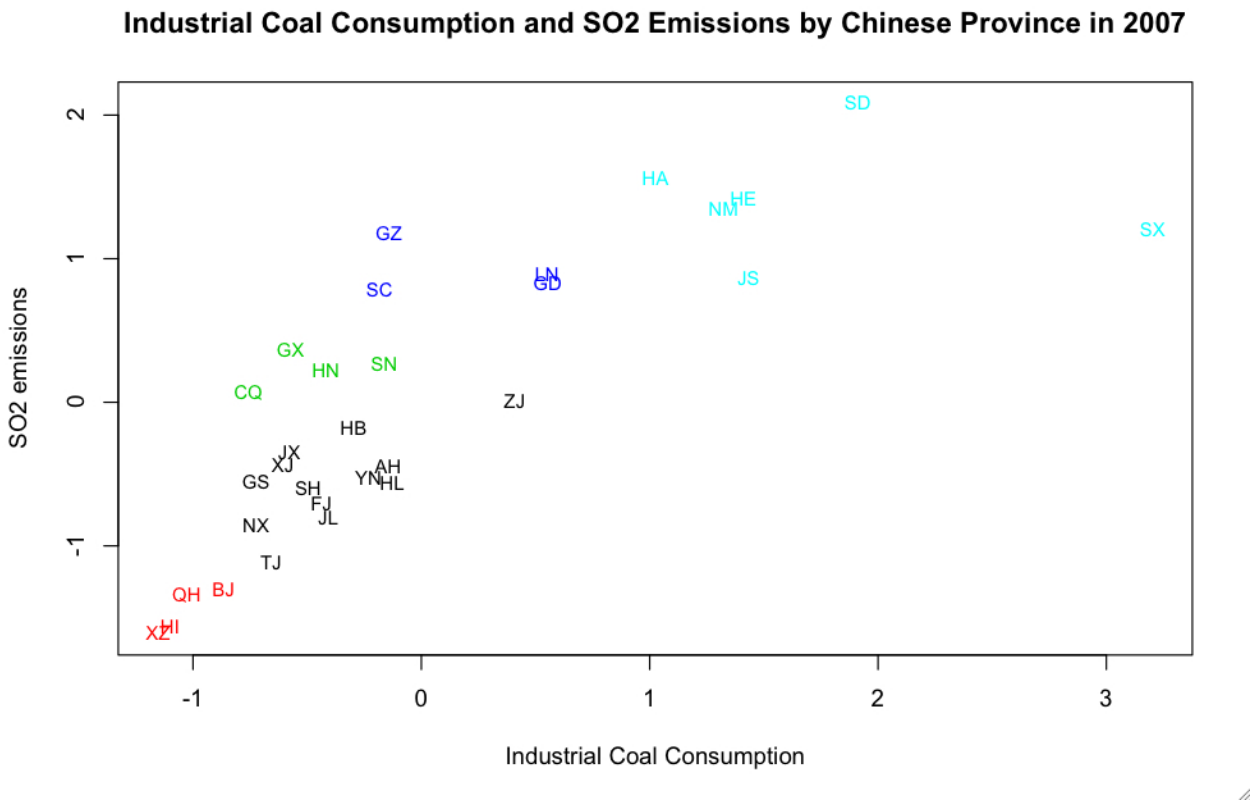


**Figure 14**. Cluster of Chinese provinces by industrial coal consumption and $SO_2$ emissions in 2007.

```
> split(rownames(z), cut5)
$`1`
 [1] "Anhui"        "Fujian"       "Gansu"        "Heilongjiang"
 [5] "Hubei"        "Jiangxi"      "Jilin"        "Ningxia Hui"
 [9] "Shanghai"     "Tianjin"      "Xinjiang"     "Yunnan"
[13] "Zhejiang"

$`2`
[1] "Beijing" "Hainan"  "Qinghai" "Xizang"

$`3`
[1] "Chongqing"      "Guangxi Zhuang" "Hunan"
[4] "Shaanxi"

$`4`
[1] "Guangdong" "Guizhou"   "Liaoning"  "Sichuan"

$`5`
[1] "Hebei"      "Henan"      "Jiangsu"    "Nei Mongol"
```

```
[5] "Shandong"   "Shanxi"
```

Industrial coal consumption and $SO_2$ emissions have a clear positive relationship for 2007 in Figure 14. Based on general knowledge of Chinese cities and regional concentration of industries, this plot makes sense. For instance, as indicated in red, Beijing, Qinghai, Hainan, and Xizang (Tibet) are clustered togther in a group that represents the lowest $SO_2$ emissions and industrial coal consumption. Beijing relocated a significant amount of industry from its municipal boundaries over the past decade to prepare for the Olympic Games and also is more advanced in air pollution control technology versus other cities. Qinghai and Xizang are remote areas with very few industries, while Hainan is also a very small island whose commerce significantly comes from tourism. On the otherside of the spectrum, the cluster analysis clearly groups heavily industrialized regions Hebei, Henan, Jiangsu, Nei Mongol (Inner Mongolia), Shandong, and Shanxi - as indicated in light blue above.

We were also interested in understanding how strong the relationship is between gross regional product and $SO_2$ emissions.

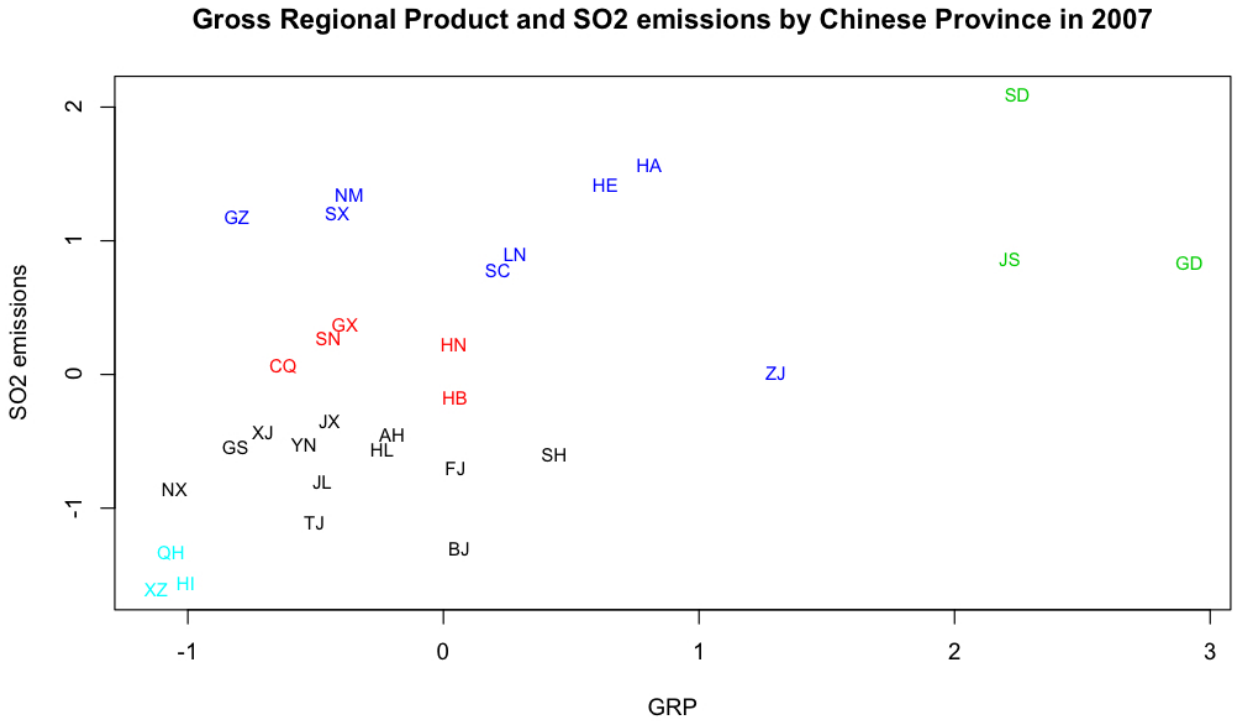


**Figure 15**. Cluster of Chinese provinces by gross regional product and $SO_2$ emissions in 2007.

```
> split(rownames(ooo), cut5)
$`1`
 [1] "Anhui"         "Beijing"      "Fujian"       "Gansu"
 [5] "Heilongjiang" "Jiangxi"      "Jilin"        "Ningxia Hui"
 [9] "Shanghai"      "Tianjin"      "Xinjiang"     "Yunnan"
```

```
$`2`
[1] "Chongqing"      "Guangxi Zhuang" "Hubei"
[4] "Hunan"          "Shaanxi"

$`3`
[1] "Guangdong" "Jiangsu"   "Shandong"

$`4`
[1] "Guizhou"    "Hebei"      "Henan"      "Liaoning"
[5] "Nei Mongol" "Shanxi"     "Sichuan"    "Zhejiang"

$`5`
[1] "Hainan"  "Qinghai" "Xizang"
```

In Figure 15, we can see a less clear relationship between gross regional product and SO2 emissions than the relationship between industrial coal consumption and SO2 emissions (Figure 14), implying that provincial wealth can sometimes be significantly attributed to reasons other than industrial growth. Beijing's current economy, for instance, profits greatly from the service versus industrial sector, thus we see it as rather an outlier in this plot and no longer grouped with Qinghai, Xizang, and Hainan. Additionally, more coastal cities and provinces such as Beijing, Shanghai, Guangdong, Shangdong, and Jiangsu as indicated as right-most outliers in the graph are much more developed than the interior provinces due to their proximity to and location on the Eastern coastline, which lends to more global trade opportunities. These outliers aside, we can generally see a strong relationship between gross regional product and SO2 emissions. There are also some provinces that lean toward being outliers on the left side of the plot, such as Shanxi, Nei Mongol (Inner Mongolia), and Guizhou, whose gross regional product is lower than the national average despite having high coal-mining activity. All three provinces are likely economically limited due to their geographical distance from the Eastern coastline.

**Conclusion**

This study attempted to employ statistical techniques to shed light on air quality data in China. Due to the contested quality of Chinese air quality data, we first examined several datasets for suflur dioxide and nitriogen dioxide emissions in China, from "official" and "non-official" sources, to determine whether simple exploratory data analysis methods and plots would reveal questionable anomalies. We found that the CNEMC dataset, which was comprised of $SO_2$ and $NO_2$ concentration data, was not correlated (r=0.26 and r=0.38 for SO2; and r=0.03 for $NO_2$) with the ARCTAS and IPE datasets. However, we did find a high degree of correlation (r=0.89 for $SO_2$ and r=0.9 for $NO_2$). Second, we sought to comprehend the underlying causes for $SO_2$ emission levels from 2004 to 2007 in Chinese provinces by developing multiple linear regression models. We determined that the strongest variable to explain $SO_2$ emissions per square kilometer is the interaction between industrial coal consumption, population, and gross regional product. Cluster analysis further emphasized these relationships.

**Acknowledgements**

We would like to thank Professor Joseph Chang for his valuable inputs and advice in the development of this project; Teaching Assistants Muting Wan and Arlene Kim for their excellent technical assistance; Professor David Streets and Qiang Zhang at Argonne National Laboratory and Jeremy Schriefels at the U.S. EPA for guidance on Chinese $SO_2$ data; and Doctoral student Jeffrey Chow for his expertise on linear regression modeling.

**References**

_____. 2003. Need for accurate statistics stressed. China Daily. 12 December. http://www.chinadaily.com.cn/en/doc/2003-12/08/content_288371.htm.

_____. 2008. An aberrant abacus. The Economist. http://www.economist.com/finance/displaystory.cfm?story_id=11290833.

Akimoto, H., T. Ohara, J. Kurokawa, N. Horii. 2006. Verification of energy consumption in China during 1996-2003 by using satellite observational data. Atmospheric Environment. 40:7663-7667.

Andrews, S. 2008. Inconsistencies in air quality metrics: 'Blue Sky' days and PM10 concentrations in Beijing. Environmental Research Letters, Vol. 3.

Coglianse, C. and L.D.S. Bennear. 2005. Appendix E: Program evaluation of environmental policies: toward evidence-based decision making in Decision Making for the Environment. G.D. Brewer and P.C. Stern, eds. Washington, D.C.: National Academies Press.

Sinton, J. 2001. Accuracy and reliability of China's energy statistics. China Economic Review. 12:373-383.

U.S. Environmental Protection Agency (EPA). 2007. "Air Emissions." http://www.epa.gov/solar/energy-and-you/affect/air-emissions.html

U.S. Environmental Protection Agency (EPA). 2008. "SO2: What is it? Where does it come from?" http://www.epa.gov/air/urbanair/so2/what1.html.

World Resources Institute. 2009. Climate Analysis Indicator Tool (CAIT). www.cait.wri.org.

Zhang, Q., D.G. Streets, K. He, Y. Wang, A. Richter, J.P. Burrows, I. Uno, C.J. Jang, D. Chen, Z. Yao, and Y. Lei. 2007. NOx emission trends for China, 1995-2004: the view from the ground and the view from space. Journal of Geophysical Research. Vol. H2.

Zhang, Q., et al. (2009), Asian Emissions in 2006 for the NASA INTEX-B Mission, Atmos. Chem. Phys. Diss., 9, 4081-413.

Zhao, Y., et al. (2009), Primary air pollutant emissions of coal-fired power plants in China: Current status and future prediction, Atmos. Environ., 42, 8442-8452.

1 A blue-sky day is defined as a day that has an Air Pollution Index (API) less than 100. For more information on how the API is calculated, see http://fire.biol.wwu.edu/trent/alles/China_API_Rules.pdf.

**Internet Resources**

**China Data Center**
www.chinadataonline.org

**Institute of Public and Environmental Affairs**
www.air.ipe.org.cn

**NASA ARCTAS Emissions Data**
www.cgrer.uiowa.edu/arctas/emission.html

**APPENDIX I - R-SCRIPT FOR PART I: COMPARISON OF DATA SETS**

```
# Exploratory data analysis and plots

# Make sure to setwd
library(YaleToolkit)
dat = read.csv("Air_quality_data.csv", as.is=TRUE, skip=1)

# header info
dath = read.csv("Air_quality_data.csv", header=FALSE, nrow=5, as.is=TRUE)

# get rid of header info
x = read.csv("Air_quality_data.csv", skip=5, as.is=TRUE)

# get rid of column 5, which has a lot of extra info
x = x[-5]

# Exploratory data analysis
dim(x)
whatis(x) # make sure all data are numerical


# Make a data frame for the emissions data for 2006
```

```
e = x[, c(33:34, 53:58, 87:88, 91, 95, 97, 99, 101, 103)]
rownames(e) <- x[,1]
attach(e)

# convert units for Gg (1 Gg = 1,102 tons)?
en = e[,3:8]
en = en*1102/10e4
e[,3:8] = en


# checking sources
sox = e[,c(1,4,9)]
names(sox) = c("CNEMC", "ARCTAS", "IPESO2")
nox = e[,c(2,5,13)]
names(nox) = c("CNEMC", "ARCTAS", "IPE")

panel.cor <- function(x, y, digits=2, prefix="", cex.cor,
cor.method="pearson")
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y, method=cor.method), na.rm=TRUE, use=complete.obs)
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = 1, col='black')

}

pairs(sox[1:31, 1:3], lower.panel=panel.smooth, upper.panel=panel.cor,
main="Pairs Plots of Raw SO2 data")

pairs(sox[1:31, 1:3], upper.panel=panel.cor, lower.panel=function(...)
panel.smooth(..., col.smooth=gray(.5), lty=1), diag.panel=panel.dist_summary,
cex.labels = 2, font.labels=2)

pairs(log(sox[1:31, 1:3]), lower.panel=panel.smooth, upper.panel=panel.cor,
main="Pairs Plots of log(SO2) data")

nox$IPE[29]=""
nox$IPE=as.numeric(IPE)
pairs(nox[1:31, 1:3], lower.panel=panel.smooth, upper.panel=panel.cor,
na.action=na.omit, main="Pairs Plots of Raw NO2 data")
```

```
# omit china total data
sox = sox[-32,]
nox = nox[-32,]

# qqnorm plots for datasets
par(mfrow=c(3,2))
   qqnorm(sox[,1], main="QQ Plot of CNEMC-SO2")
   qqnorm(sox[,2], main="QQ Plot of ARCTAS-SO2")
   qqnorm(sox[,3], main="QQ Plot of IPE-SO2")
   qqnorm(nox[,1], main="QQ Plot of CNEMC-NO2")
   qqnorm(nox[,2], main="QQ Plot of ARCTAS-NO2")
   qqnorm(nox[,3], main="QQ Plot of IPE-NO2")


# closer look at IPE and ARCTAS for NOX
lm1 = lm(nox$IPE ~ nox$ARCTAS)
summary(lm1)

# closer look at CNEMC and IPE data for NOX
lm2 = lm(nox$CNEMC ~ nox$ARCTAS)
summary(lm2)

# plots for both
par(mfrow=c(1,2))
plot(nox$IPE ~ nox$ARCTAS, xlab="ARCTAS-NO2", ylab="IPE-NO2",
main="Comparison of ARCTAS-IPE NO2 Data")
abline(coef(lm1), col="red")
plot(nox$CNEMC ~ nox$ARCTAS, xlab="ARCTAS-NO2", ylab="CNEMC",
main="Comparison of ARCTAS-CNEMC Data")
abline(coef(lm2), col="red")

# hypothesis testing: t-tests
t.test(sox[2:3])
t.test(nox[2:3])


# stripplots and barchart plots

# NO2
temp = stack(nox, select=2:3)
temp2 = rep(x$provcode[1:31], 2)
new = cbind(temp, temp2)
names(new) = c("no2", "type", "code")
stripplot(new$code ~ new$no2, groups = type, auto.key=TRUE, main="Stripplot
```

```
of NO2 data by type", xlab="SO2 emissions (10^4 Tons)")
barchart(new$code ~ new$no2, groups = type, auto.key=TRUE, main="Barchart of
NO2 data by type", xlab="NO2 emissions (10^4 Tons)")

# SO2
temp = stack(sox, select=2:3)
temp2 = rep(x$provcode[1:31], 2)
new = cbind(temp, temp2)
names(new) = c("so2", "type", "code")
stripplot(new$code ~ new$so2, groups = type, auto.key=TRUE, main="Stripplot
of SO2 data by type", xlab="SO2 emissions (10^4 Tons)")
barchart(new$code ~ new$so2, groups = type, auto.key=TRUE, main="Barchart of
SO2 data by type", xlab="SO2 emissions (10^4 Tons)")

# what is the mean difference between the two datasets?
diff = mean(sox[3] - sox[2])
diff2 = mean(nox[3] - nox[2])


### How evaluate CNEMC data? Look at cities
cities = read.csv("Cities.csv", as.is=TRUE, header=FALSE, skip=5)
colnames(cities) = cities[1,]
cities = cities[-1,]
cs = cities[,c(1,3,4,7:9)]
attach(cs)
y =  as.factor(cs$Provcode)

# SO2
plot(cs$so206, y, yaxt='n', ylab='', xlab="SO2 Concentrations (mg/m3)",
main="SO2 Concentrations in Chinese Cities")
axis(2, at=seq_along(levels(y)), labels=levels(y), las=2)
points(cs$so206[cs$Capital=="Y"], y[cs$Capital=="Y"], pch=19, col="red")
text(cs$so206[cs$Capital=="Y"], y[cs$Capital=="Y"],
labels=cs$City[cs$Capital=="Y"], pos=1, offset=0.5, cex=0.75)

# NO2
plot(cs$no206, y, yaxt='n', ylab='', xlab="NO2 Concentrations (mg/m3)",
main="NO2 Concentrations in Chinese Cities")
axis(2, at=seq_along(levels(y)), labels=levels(y), las=2)
points(cs$no206[cs$Capital=="Y"], y[cs$Capital=="Y"], pch=19, col="red")
text(cs$no206[cs$Capital=="Y"], y[cs$Capital=="Y"],
labels=cs$City[cs$Capital=="Y"], pos=1, offset=0.5, cex=0.75)
```

## APPENDIX II - R-SCRIPT FOR PART II: MULTIPLE REGRESSION MODELING SO2

```
# get rid of header info
x = read.csv("/Users/elaineyu/Desktop/Air_quality_data.csv", skip=5,
as.is=TRUE)

# get rid of column 5, which has a lot of extra info
x = x[-5]
x
names(x)
x
#time-series
names(x)
z = cbind(provcode=x[,2], area=x[,4], x[,7:10], x[,12:15], x[,77:88],
x[,105:108])

rownames(z) = z[,1]
z = z[-1]
z= z[-32,]

area1 = x[1:31,4]
area = as.vector(rep(area1, times=4))
area
pop = as.vector(cbind(z[,5], z[,4], z[,3], z[,2]))
grp = as.vector(cbind(z[,9], z[,8], z[,7], z[,6]))
icc = as.vector(cbind(z[,10], z[,11], z[,12], z[,13]))
ifc = as.vector(cbind(z[,14], z[,15], z[,16], z[,17]))
so2 = as.vector(cbind(z[,18], z[,19], z[,20], z[,21]))
car = as.vector(cbind(z[,22], z[,23], z[,24], z[,25]))

#Creating Data Frame
soxy1 = cbind(area, pop, grp, icc, ifc, car, so2)
soxy = as.data.frame(soxy1)
soxy

#Test to see if data is parametric - do we need transformations?

#Individual look:

#Population
par(mfrow=c(3,4))
```

```
hist(pop)
hist(log(pop))
qqnorm(pop)
qqnorm(log(pop)) #Better - more normal

#Gross Regional Product
hist(grp)
hist(log(grp))
qqnorm(grp)
qqnorm(log(grp)) #Better

#Industrial Coal consumption
hist(icc)
hist(log(icc)) #Better
qqnorm(icc)  #Normal
qqnorm(log(icc))

#SoX
hist(so2)
hist(log(so2)) #Better
qqnorm(so2)  #Normal
qqnorm(log(so2))
boxplot(so2) #looks more normal?
boxplot(log(so2))  #Bottom outliers when log is performed
range(so2)

#Vehicles
hist(car)
hist(log(car)) #Better
qqnorm(car)
qqnorm(car)
boxplot(car)

#Industrial Fuel Consumption
hist(ifc)
hist(log(ifc)) #Better
qqnorm(ifc)
qqnorm(log(ifc)) #Better
boxplot(ifc)
boxplot(log(ifc))

#All plots
#Plot all
plot(soxy)
```

```
#log transformation of all data:
lsoxy = log(soxy)
plot(lsoxy, main ="Exploration of relationships between variables")

#Examining Relationships more closely and searching for outliers

par(mfrow=c(1,5))

#SO2 and Industrial Fuel Consumption
plot(ifc~so2, main = "Example: ifc & SO2", dat =lsoxy)
a8 <-lm(car~so2, dat=lsoxy)
abline(a8, col ="red")
plot(a8, which=1:4)
#Outliers: 29, 95, 122

#So2 and Industrial Coal Consumption
plot(icc~so2, main = "Sulfur Dioxide and Industrial Coal Consumption", xlab =
"log(industrial coal consumption)", ylab = "log(So2 Emissions)",dat=lsoxy)
a1 <-lm(icc~so2, dat=lsoxy)
abline(a1, col ="red")
par(mfrow=c(2,2))
plot(a1, which=1:4)
#Outliers: 21, 24, 29

#So2 and Private Vehicles
plot(car~so2, main = "Sulfur Dioxide and Civilian Vehicle
Population",dat=lsoxy)
a <-lm(car~so2, dat=lsoxy)
abline(a, col ="red")
#Identifying outliers
par(mfrow=c(2,2))
plot(a, which=1:4)
#20,64, 95,122

#So2 and Gross Regional Product
plot(grp~so2, dat=lsoxy)
a2 <-lm(grp~so2, dat=lsoxy)
abline(a2, col ="red")
par(mfrow=c(2,2))
plot(a2, which=1:4)
#20,21,29, 51, 82, 122

#Population and SO2
plot(pop~so2, dat=lsoxy)
```

```
a3 <-lm(pop~so2, dat=lsoxy)
abline(a3, col ="red")
par(mfrow=c(2,2))
plot(a3, which=1:4)
#29, 51, 82, 91, 113, 122

#GRP and Industrial Coal Consumption
plot(grp~icc, dat=lsoxy)
a4 <-lm(grp~icc, dat=lsoxy)
abline(a4, col ="red")
par(mfrow=c(2,2))
plot(a4, which=1:4)
#20, 24, 29,

#Car and GRP
plot(grp~car, dat=lsoxy)
a5 <-lm(grp~car, dat=lsoxy)
abline(a5, col ="red")
par(mfrow=c(2,2))
plot(a5, which=1:4)
#33, 64, 91, 95, 122

#ICC and Pop
plot(icc~pop, dat=lsoxy)
a6 <-lm(icc~pop, dat=lsoxy)
abline(a6, col ="red")
par(mfrow=c(2,2))
plot(a6, which=1:4)
#21,29, 60

#Omit Outliers
##20, 21, 24, 29, 33, 51, 60, 64, 82, 91, 95, 113, 122
#See Check also( Removing 3 and 83)
j = rbind(lsoxy[1:2,], lsoxy[4:19,], lsoxy[22:23,], lsoxy[25:28,], lsoxy[30:
32,], lsoxy[34:50,], lsoxy[52:59,], lsoxy[61:63,], lsoxy[65:81,],
lsoxy[84:90,], lsoxy[92:94,], lsoxy[97:112,], lsoxy[114:121,],
lsoxy[120:124,])

#Excluding NAs
o = na.exclude(j)

#Check
lo = log(o)
plot(lo)
```

```
#Modeling
mod = lm(so2 ~ ., dat=lo)
mod
mod0 = step(mod, direction="backward")
mod0

mod1 = lm(so2 ~ icc, dat=lo)
mod2 = lm(so2 ~ icc + ifc, dat=lo)
mod3 = lm(so2 ~ icc + car, dat=lo)
mod4 = lm(so2 ~ icc + ifc + car, dat=lo)
mod5 = lm(so2 ~ icc + ifc + car:grp , dat=lo)
mod6 = lm(so2 ~ icc + ifc + car:pop:grp, dat=lo)

AIC(mod1, mod2, mod3, mod4, mod5)

mod1.1 = lm((so2/area) ~ icc, dat=lo)
mod2.1 = lm((so2/area) ~ icc + ifc, dat=lo)
mod3.1 = lm((so2/area) ~ icc + car, dat=lo)
mod4.1 = lm((so2/area) ~ icc + ifc + car, dat=lo)
mod5.1 = lm((so2/area) ~ icc + ifc + car:grp, dat=lo)

AIC(mod1.1, mod2.1, mod3.1, mod4.1, mod5.1)

mod6.1 = lm((so2/area) ~ icc + grp, dat=lo)
mod7.1 = lm((so2/area) ~ icc + ifc + grp, dat=lo)
mod8.1 = lm((so2/area) ~ icc + car + grp, dat=lo)
mod9.1 = lm((so2/area) ~ icc + ifc + car, dat=lo)
mod10.1 = lm((so2/area) ~ icc + ifc + car:grp, dat=lo)
mod11.1 = lm((so2/area) ~ icc + ifc + car:pop:grp, dat=lo)

mod12.1 = lm((so2/area) ~ icc*grp, dat=lo)
mod13.1 = lm((so2/area) ~ ifc + icc*grp, dat=lo)

#Best ones so far
mod7.1 = lm((so2/area) ~ icc + ifc + grp, dat=lo)
mod13.1 = lm((so2/area) ~ ifc + icc*grp, dat=lo)

#Try again
mod14.1 = lm((so2/area) ~ ifc + icc:grp, dat=lo)
mod15.1 = lm((so2/area) ~ ifc + icc:grp:pop, dat=lo)
mod16.1 = lm((so2/area) ~ ifc + icc:grp:pop + car:grp, dat=lo)
mod17.1 = lm((so2/area) ~ ifc + icc:grp:pop + car:grp:pop, dat=lo)
mod18.1 = lm((so2/area) ~ ifc + icc:grp:pop + car, dat=lo)
```

```
mod19.1 = lm((so2/area) ~ ifc:grp + icc:grp:pop, dat=lo)
mod20.1 = lm((so2/area) ~ ifc:grp:pop + icc:grp:pop, dat=lo)
mod21.1 = lm((so2/area) ~ ifc:grp:pop + icc:grp:pop, dat=lo)
mod22.1 = lm((so2/area) ~ icc + ifc + ifc:grp:pop + icc:grp:pop, dat=lo)
mod23.1 = lm((so2/area) ~ ifc:pop + icc:pop, dat=lo)
mod24.1 = lm((so2/area) ~ grp + ifc:grp:pop + icc:grp:pop, dat=lo)

AIC(mod7.1, mod13.1, mod14.1, mod15.1, mod16.1, mod17.1, mod18.1, mod19.1,
mod20.1, mod21.1, mod22.1, mod23.1, mod24.1, mod25.1)

anova(mod7.1, mod13.1, mod14.1, mod15.1, mod16.1, mod17.1, mod18.1, mod19.1,
mod20.1, mod21.1, mod22.1, mod23.1, mod24.1, mod25.1)
summary(mod20.1)

********
#FOR SO2 TIME SERIES PLOT

z = cbind(provcode=x[,2], x[,7:10], x[,12:15], x[,77:88], x[,97:100])
rownames(z) = z[,1]
z = z[-1]
z = z[-32,]

temp = stack(z,select=17:20)
temp$Year = c(rep(2004,31), rep(2005,31), rep(2006,31), rep(2007,31))
temp = as.data.frame(temp)
names(temp) = c("so2", "id", "Year")

### Time-series plots
# population
temp2 = t(z)
temp2ts = ts(temp2[1:4,], start=2004, frequency=1)

plot(temp2ts, plot.type="single", xlim=c(2004, 2007+.75), xlab="",
ylab="Population", col=1:31)
ylast = window(temp2ts, 2007)
yvals = temp2ts[4,]
text(rep(2007, 31), as.numeric(yvals), colnames(ylast), pos=4, col=1:31)

# perhaps more interesting, SO2
temp3ts = ts(temp2[21:24,], start=2004, frequency=1)
plot(temp3ts, plot.type="single", xlim=c(2004, 2007+.75), xlab="", ylab="SO2
emissions (10^4 ton)", lty=1:31, col=1:31, main="SO2 Emissions from
2004-2007")
ylast = window(temp3ts, 2007)
```

```
yvals = temp3ts[4,]
text(rep(2007, 31), as.numeric(yvals), colnames(ylast), pos=4, col=1:31)
```

## APPENDIX III - R-SCRIPT FOR CLUSTERING

```
#Industrial Coal Consumption and SO2 Emissions by Chinese Province in 2007
x = read.csv("/Users/elaineyu/Desktop/Air_quality_data.csv", skip=5,
as.is=TRUE)

# get rid of column 5, which has a lot of extra info
x = x[-5]
x
names(x)

y <- x[, c("province", "icc07", "so207ipe")]
rownames(y) = 1:nrow(y)
y
yy = y[,-1]
class(yy)
class(y)
yy = y[-1]
rownames(yy) = y$province
yy
zz = yy[-32,]
zz

#To Scale data:
z = scale(zz, center =T, scale =T)
z
plot(hclust(dist(z)), main="Default (complete), scaled")
plot(hclust(dist(z), method="single"), main="Single, scaled")
plot(hclust(dist(z), method="ward"), main="Ward, scaled")

hc1 = hclust(dist(z), method="ward")
plot(hc1)
cut5 = cutree(hc1, k=5)
cut5
apply(z[cut5 == 1,],2,mean)
apply(z[cut5 == 2,],2,mean)
apply(z[cut5 == 3,],2,mean)
apply(z[cut5 == 4,],2,mean)
apply(z[cut5 == 5,],2,mean)

aggregate(z, by=list(cut5), mean)
```

```
split(rownames(z), cut5)

plot((data.frame(z)), main = "Industrial Coal Consumption and SO2 Emissions
by Chinese Province in 2007", xlab= "Industrial Coal Consumption", ylab= "SO2
emissions", type="n")
text(z[,1], z[,2],labels= rownames(z), cex = 0.8, col=cut5)

#GRP AND SO2
y <- x[, c("province", "grp07", "so207ipe")]
rownames(y) = 1:nrow(y)
y
yy = y[,-1]
class(yy)
class(y)
yy = y[-1]
rownames(yy) = y$province
yy
oo = yy[-32,]
oo


To Scale data:
ooo = scale(oo, center =T, scale =T)

plot(hclust(dist(ooo)), main="Default (complete), scaled")
plot(hclust(dist(ooo)), method="single"), main="Single, scaled")
plot(hclust(dist(ooo)), method="ward"), main="Ward, scaled")

hc1 = hclust(dist(ooo), method="ward")
plot(hc1)
cut5 = cutree(hc1, k=5)
cut5
apply(ooo[cut5 == 1,],2,mean)
apply(ooo[cut5 == 2,],2,mean)
apply(ooo[cut5 == 3,],2,mean)
apply(ooo[cut5 == 4,],2,mean)
apply(ooo[cut5 == 5,],2,mean)

aggregate(ooo, by=list(cut5), mean)
split(rownames(ooo), cut5)

plot((data.frame(ooo)), main = "Gross Regional Product and SO2 emissions by
Chinese Province in 2007", xlab= "GRP", ylab = "SO2 emissions", type="n")
text(ooo[,1], ooo[,2],labels= rownames(ooo), cex = 0.8, col=cut5)
```