**The Report Committee for Kailin Wu**
**Certifies that this is the approved version of the following report:**

# Longitudinal Analysis on AQI in 3 Main Economic Zones of China

**APPROVED BY**
**SUPERVISING COMMITTEE:**

**Supervisor:**

Daniel A.Powers

Matthew A Hersh

# Longitudinal Analysis on AQI in 3 Main Economic Zones of China

**by**

**Kailin Wu, B.S.**

## Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Master of Science in Statistics**

**The University of Texas at Austin**
**May 2014**

# Acknowledgements

I would like to express my special appreciation and thanks to my advisor Professor Daniel Powers, who provided detailed guidance on this report. I have learned a lot about social statistics from Dr. Powers throughout the 2 courses he instructed.

My thanks go to Dr. Matt Hersh. He is a kind person and cares about us students in SSC. Thanks Dr. Matt Hersh for assigning me as teaching assistant in undergraduate class. It’s a valuable experience to work with instructors in SSC.

A special thanks to Vicki L. Keller who gave me much attendance and encouragement on study and living in Austin.

Last, but by no means least, my deepest thanks to my parents. It would not have been possible for me to complete my study in States. Thanks to them for endless support and unconditional love.

# Abstract

## Longitudinal Analysis on AQI in 3 Main Economic Zones of China

Kailin Wu, M.S. Stat

The University of Texas at Austin,2014

Supervisor: Daniel A.Powers

In modern China, air pollution has become an essential environmental problem. Over the last 2 years, the air pollution problem, as measured by PM 2.5(particulate matter) is getting worse. My report aims to carry out a longitudinal data analysis of the air quality index (AQI) in 3 main economic zones in China. Longitudinal data, or repeated measures data, can be viewed as multilevel data with repeated measurements nested within individuals. I arrive at some conclusions about why the 3 areas have different AQI, mainly attributed to factors like population, GDP, temperature, humidity, and other factors like whether the area is inland or by the sea. The residual variance is partitioned into a between-zone component (the variance of the zone-level residuals) and a within-zone component (the variance of the city-level residuals). The zone residuals represent unobserved zone characteristics that affect AQI.

In this report, the model building is mainly according to the sequence described by West et al (2007) with respect to the bottom-up procedures and the reference by Singer, J. D., & Willett, J. B (2003) which includes the non-linear situations. This report also compares the quartic curve model with piecewise growth model with respect to this

data. The final model I reached is a piece wise model with time-level and zone-level predictors and also with temperature by time interactions.

# Table of Contents

# List of Tables

# List of Figures

# INTRODUCTION

China is a fast developing country and its GDP growth is impressive over these years. However, in modern times, the environmental pollution aggravated due to the development of industry. Environmental problems are among the most important social problems in transitional China. In particular, air pollution has become an essential environmental problem. Over the last 2 years, the air pollution problem, as measured by PM 2.5(particulate matter) is getting worse, which harms the economy, society and environment. This phenomenon motivates my research on the correlation between different indexes and air quality to learn what the main effects are and in hence how we may address the corresponding problems.

China has a vast territory and because of the imbalanced economic development in the eastern, central and western areas and also the big difference in energy consumption, the impact of various factors on PM emissions will be different. There are 3 main economic zones in mainland China. From the AQI (air quality index) map below, we can see that PM 2.5 seems clustered in areas. More red and purple dots appear clustered in BER, which means that air pollution is more severe. YRD and PRD have many yellow dots, which means the air pollution in these two zones is moderate.

China's Ministry of Environmental Protection (MEP) is responsible for measuring the level of air pollution in China. As of 1 January 2013, MEP monitors the daily pollution level in 163 of its major cities. The API level is based on the level of six atmospheric pollutants, namely sulfur dioxide (SO2), nitrogen dioxide (NO2), suspended particulates smaller than 10μm in aerodynamic diameter (PM10), suspended particulates smaller than 2.5 μm in aerodynamic diameter (PM2.5), carbon monoxide (CO), and ozone (O3) measured at the monitoring stations throughout each city. In this report, I focus on the analysis on PM 2.5 which is the most severe problem of air pollution in China today. The air quality index I use is based on the China’s air quality standards (GB3095-2012).

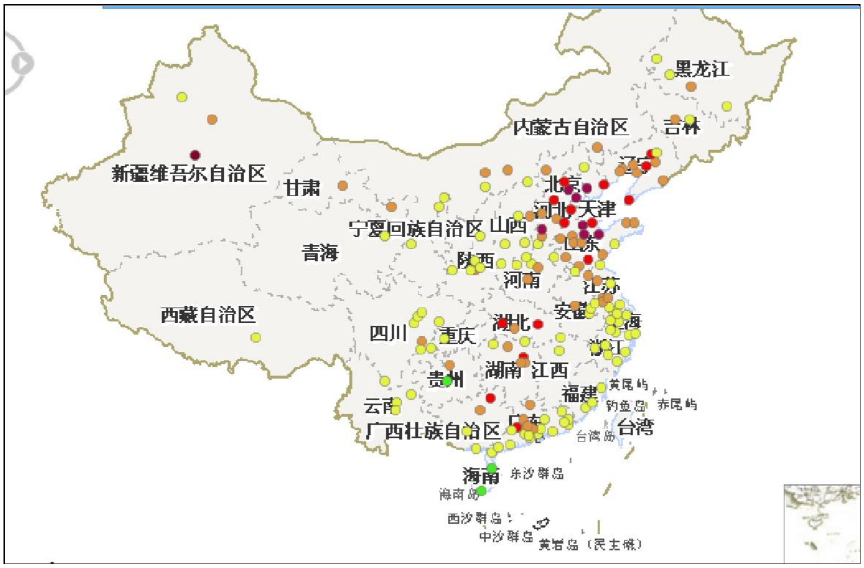


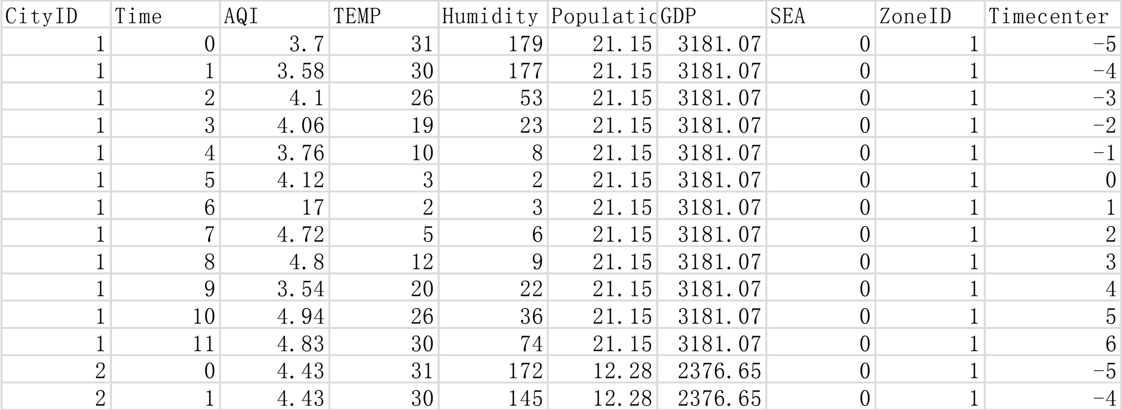
Figure 1: *AQI map of China on 15/07/2013*

The reason why I chose these three areas is because these economic zones are the country’s fastest-growing regions and demonstrate comparatively worse AQI. They share some similar characteristics, such as a developed economy and concentrated population, but also have different characteristics such as geographical location and different energy consumption types. The three economic zones are named BER, YRD and PRD respectively. BER is located in the north part of China. It is the economic hinterland surrounding Beijing and Tianjin. It also includes areas in Hebei, Liaoning and Shandong, which surrounds the Bohai Sea. This economic zone has an importance place because it includes China’s capital Beijing. YRD which refers to Yangtze River Delta or the Golden Triangle of the Yangtze generally comprises the triangle-shaped territory of Shanghai, southern Jiangsu province and northern Zhejiang province of China. The Yangtze River drains into the East China Sea. The urban build-up in the area has given rise what may be the largest concentration of adjacent metropolitan areas in the world. The delta is one of the most densely populated regions on earth, and includes one of the world's largest cities on its banks — Shanghai. PRD (The Pearl River Delta) in Guangdong province, People's Republic of China is the low-lying area surrounding the Pearl River estuary where the Pearl River flows into the South China Sea. It is one of the most densely urbanized regions in the world and one of the main hubs of China's economic growth. It has been the most economically dynamic region of the Chinese Mainland since the launch of China's reform program in 1979.

# DATA STRUCTURE

Longitudinal data, or repeated measures data, can be viewed as multilevel data with repeated measurements nested within individuals. A dataset is longitudinal if it tracks the same type of information on the same subjects at multiple points in time or space. This data is a longitudinal data set from 16 cities chosen from 3 economic zones, 6 for each zone, respectively. The cities' air quality index (AQI) has been recorded for 12 successive months from July, 2012 to July 2013. At the same time, it was recorded city's GDP, population, humidity, and temperature over these 12 months and whether the city is near sea. The city-level variable SEA remains constant for each city across the 12 measurement months.

The data structure for a multilevel analysis of these data is generally different, depending on the specific program that is used. In this report, I use R [CITE AND REFERNCE TO R HERE] to do the data analysis. The regular format in figure 2 is referred to as a 'wide' form data set.



| CityID | Time | AQI | TEMP | Humidity | Populatic | GDP | SEA | ZoneID | Timecenter |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 3.7 | 31 | 179 | 21.15 | 3181.07 | 0 | 1 | -5 |
| 1 | 1 | 3.58 | 30 | 177 | 21.15 | 3181.07 | 0 | 1 | -4 |
| 1 | 2 | 4.1 | 26 | 53 | 21.15 | 3181.07 | 0 | 1 | -3 |
| 1 | 3 | 4.06 | 19 | 23 | 21.15 | 3181.07 | 0 | 1 | -2 |
| 1 | 4 | 3.76 | 10 | 8 | 21.15 | 3181.07 | 0 | 1 | -1 |
| 1 | 5 | 4.12 | 3 | 2 | 21.15 | 3181.07 | 0 | 1 | 0 |
| 1 | 6 | 17 | 2 | 3 | 21.15 | 3181.07 | 0 | 1 | 1 |
| 1 | 7 | 4.72 | 5 | 6 | 21.15 | 3181.07 | 0 | 1 | 2 |
| 1 | 8 | 4.8 | 12 | 9 | 21.15 | 3181.07 | 0 | 1 | 3 |
| 1 | 9 | 3.54 | 20 | 22 | 21.15 | 3181.07 | 0 | 1 | 4 |
| 1 | 10 | 4.94 | 26 | 36 | 21.15 | 3181.07 | 0 | 1 | 5 |
| 1 | 11 | 4.83 | 30 | 74 | 21.15 | 3181.07 | 0 | 1 | 6 |
| 2 | 0 | 4.43 | 31 | 172 | 12.28 | 2376.65 | 0 | 1 | -5 |
| 2 | 1 | 4.43 | 30 | 145 | 12.28 | 2376.65 | 0 | 1 | -4 |

Figure 2: *Wide format data set*

Note that the measurement months are numbered 0,. . ., 11 instead of 1, . . . , 12. This ensures that the intercept represents the mean AQI at the starting point of data collection.

# METHODOLOGY

## MULTILEVEL REGRESSION MODELING

The multilevel regression model has become known in the research literature under a variety of names, such as ‘random coefficient model’ (de Leeuw & Kreft, 1986; Longford, 1993), ‘variance component model’ (Longford, 1987), and ‘hierarchical linear model’ (Raudenbush & Bryk, 1986, 1988). Statistically oriented publications tend to refer to the model as a mixed-effects or mixed model (Littell, Milliken, Stroup, & Wolfinger, 1996). ’ . They all assume that there is a hierarchical data set, with one single outcome or response variable that is measured at the lowest level, and explanatory variables at all existing levels. Conceptually, it is useful to view the multilevel regression model as a hierarchical system of regression equations.

In multilevel research, the data structure in the population is hierarchical, and the sample data are a sample from this hierarchical population. The lowest level (level 1) is usually defined by the individuals. At each level in the hierarchy, we may have several types of variables.

### Multilevel Linear Regression Model

Level-1 is the lowest level of the model which corresponds to a single row in a data set. Level-2 is the clustering level of a model and level-1 units are members of level-2 clusters.

Here is the simplest example of a model in two levels which is known as the unconditional means model. It’s also called intercept-only model. This intercept-only model is useful as a null model that serves as a benchmark with which other models are compared.

Level 1: $Y_{ij} = \beta_{0j} + e_{ij}$ (1)

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$ (2)

Combined model: $Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$

By adding predictors to level 1 or level 2, we can get the linear growth model which is presented below:

Level-1: $Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + e_{ij}$

Level-2: $\beta_{0j} = \gamma_{00} + \beta_{01}X_j + u_{0j}$

$\beta_{1j} = \gamma_{10} + \beta_{11} + u_{1j}$

The assumptions about the level-1 and level-2 residuals are showed below:

$$e_{ij} \sim N(0, \sigma_e^2)$$

$$\begin{bmatrix} u_0 \\ u_1 \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01}^2 \\ \sigma_{01}^2 & \sigma_1^2 \end{bmatrix}\right)$$

**Bottom-up Procedure**

Longitudinal data, or repeated measures data, can be viewed as multilevel data with repeated measurements nested within individuals. To do multilevel regression modeling, we can use an exploratory procedure to select a model. Model building strategies can be either top-down or bottom-up. The top-down approach starts with a model that includes the maximum number of fixed and random effects that are considered for the model. Typically, this is done in two steps. The first step starts with all the fixed effects and possible interactions in the model, followed by removing insignificant effects. The second step starts with a rich random structure, followed by removal of insignificant effects. This procedure is described by West et al (2007). In multilevel modeling, the top-down approach has the disadvantage that it starts with a large and complicated model, which leads to longer computation time and sometimes to convergence problems. In this report, the opposite strategy is used, which is bottom-up: start with a simple model and proceed by adding parameters, which are tested for significance after they have been added. Typically, the procedure starts by building up the fixed part, and follows after with the random part. The advantage of the bottom-up procedure is that it tends to keep the models simple.

First, we start with the simplest possible model, the intercept-only model, which is also called the unconditioned means model and to add the various types of parameters step by step. At each step, we inspect the estimates and standard errors to see which parameters are significant. We start with the fixed regression coefficients, and add variance components at a later stage. The different steps of such a selection procedure are given below:

Step 1: Analyze a model with no explanatory variables. This model is also called intercept-only model. This model is given by the following equation:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

Where $\gamma_{00}$is the regression intercept, and $u_{0j}$ and $e_{ij}$ are the usual residuals at the group and the individual level. The intercept-only model is useful because it gives us an estimate of the intraclass correlation ρ:

$$\sigma_{\mu 0}^2 / \sigma_{\mu 0}^2 + \sigma_e^2$$

Step2: Analyze a model with all lower-level explanatory variables fixed. This means that the corresponding variance components of the slopes are fixed at zero. This model is written as:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + u_{0j} + e_{ij}$$

Where the $X_{pij}$ are the p explanatory variables at the individual level. In this step, we assess the contribution of each individual-level explanatory variable. The significance of each predictor can be tested, and we can assess what changes occur in the first-level and second-level variance terms. We can test the improvement of the final model chosen in this step by computing the difference of the deviance of this model and the previous model (the intercept-only model). This difference approximates a chi-square with degrees of freedom equal to the difference in the number of parameters of both models. If there are 3 levels, this step is repeated on a level-by-level basis.

Step 3: add higher–level explanatory variables. The equation is written below:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{q0}Z_{qj} + u_{0j} + e_{ij}$$

Where the $Z_{qj}$ are the $q$ explanatory variables at the group level. This model allows us to examine whether the group-level explanatory variables explain between-group variation in the dependent variable. Also, if there are 3 levels, this step is repeated on a level-by-level basis.

The models in steps 2 and 3 are often called variance component models, because they decompose the intercept variance into different variance components for each hierarchical level. In a variance component model, the regression intercept is assumed to vary across the groups, but the regression slopes are assumed fixed.

Step 4: Access whether any of the slopes of any of the explanatory variables has a significant variance component between the groups. This is called random coefficient model, testing for random slope variation is best done on a variable by variable basis. The equation is given by:

$$Y_{ij} = \gamma_{00} + \gamma_{p0}X_{pij} + \gamma_{q0}Z_{qj} + u_{pj}X_{pij} + u_{0j} + e_{ij}$$

Where the $u_{pj}$ are the group-level residuals of the slopes of the individual-level explanatory variables $X_{pij}$.

Testing for random slope variation is best done on a variable-by-variable basis. After deciding which of the slopes have a significant variance between groups by using the deviance difference test, we add all these variance components simultaneously in a final model and use chi square tests based on the deviance to test whether the model of Step 4 is fits better that the final model of Step 3. Also, if there are more than two levels, this step is repeated on a level-by-level basis.

The last step is to decide whether to add cross-level interactions between explanatory group-level variables and those individual-level explanatory variables that had significant slope variation found in Step 4. Following the 5 steps explained above, leads to the full model. In this report, the method is mainly based on this bottom-up procedure.

### Multilevel Linear Regression Model

Besides the linear regression model, sometimes, we may come across non-linear regression models. The variable procedure is similar s above, but also cooperated with the model building sequence based on the reference from Singer, J. D., & Willett, J. B (2003).

It also includes the non-linear situation, which is more comprehensive. The sequences are given as:

(a) Examine empirical growth plots

(b) Fit an unconditional means model

(c) Fit an unconditional linear growth model

(d) Fit unconditional non-linear model (e.g., quadratic)

(e) Compare unconditional linear and non-linear models

(f) Add level-1 and level-2 predictors

When it comes to the non-linear regression, we often use polynomial curves to model the pattern of change over time. Polynomial curves are often used for estimating development curves. They are convenient because they can be estimated using standard linear modeling procedures and they are very flexible. However, a general problem with polynomial function is that they often have very high correlations. So sometimes, polynomial functions may cause numerical problems in the estimation.

Another solution to the estimation of non-linear model that often discussed is the use of piecewise linear functions and spline functions (Snijder and Bosker 1999), which are the functions that break up the development curve into different adjacent pieces.

We can use a global chi square test to decide which kind of function to use, polynomial or the piecewise linear function.

# DATA ANALYSIS AND RESULTS

## DATA SOURCE AND VARIABLE IDENTIFICATION

This data is a longitudinal data set from 18 cities chosen from 3 economic zones, 6 for each zone, respectively. The data set contains 3 levels (or hierarchies) which are time, city and zone levels. It has 206 observations with 18 cities. The dependent variable here is the AQI The 6 cities in BER are: Beijing, Tianjin, Qinghuangdao, Shijiazhuang, Tangshan, Langfang; 6 cities in YRD are Shanghai, Hangzhou, Wuxi, Suzhou, Wenzhou, Ningbo and the remaining left cities in PRD are: Guangzhou, Shenzhen, Zhuhai, Dongguan, Xiamen, and Fuzhou. The cities' air quality index (AQI) has been recorded for 12 successive months from July, 2012 to July 2013. At the same time, it was recorded city's GDP, population, humidity, and temperature over these 12 months and whether the city is near sea. The city-level variable SEA remains constant for each city across the 12 measurement months.

The independent variables are given below:

Time: *T* (*t*=1,…,12). There are12 waves, the distance between 2 waves represents 1 month and the history data is from July, 2012 to June, 2013.

Cities: CityID (*j*=1,…, 8)

Zone: ZoneID (*i*=1, 2, 3) And also some potential independent variables:

Temperature: TEMP (in degrees C)

Humidity: HUM (in mm per month)

Population: POP (in million)

GDP: GDP (in US dollars)

Inland or by Sea: SEA (1- near the sea; 0- inland)

I collected these data mainly online. About AQI, I used data collected by PM 2.5 Monitoring System at www.cnpm25.com. Population and GDP of these 18 cities are collected from Wikipedia. Temperature and humidity of these 18 cities during 12months are collected from www.weather.com.cn.

## MODEL BUILDING AND COMPARISONS

### Empirical Growth plots

Most of the following is the result of model building according to the sequence talked in previous part of the report by cooperating the sequence described by Singer, J. D., & Willett, J. B (2003) and by West et al (2007).

First, Examine empirical growth plots. I produced a raw data using the R plotting function `plot(x-time,y-AQI)` and overlay loess model fit by gg plots.

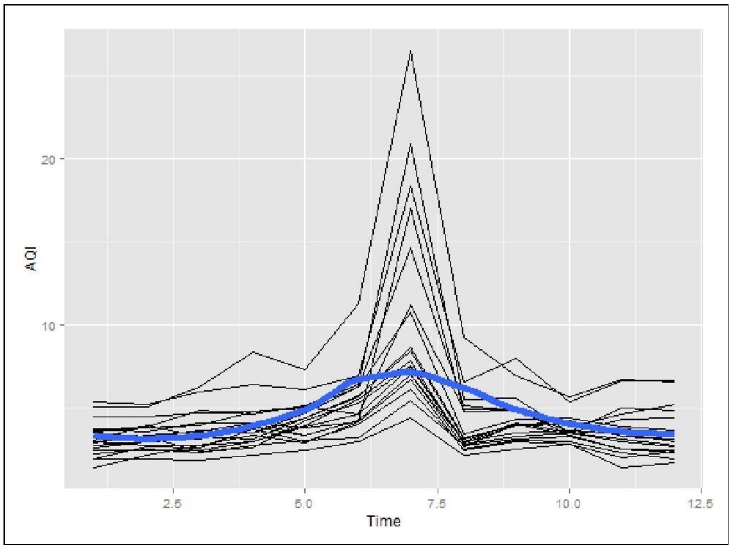


Figure 3: *Raw data plot and overlaid a loess-smoothed line*

From the plots above, we can see that there exists a peak at time 6 (January, 2013). Also, there seems to be a symmetrical pattern centered on January. Apparently, it may not be appropriate to apply a linear growth model to these empirical patterns. Based on the shape of the plots, 2 possible models may be fit to this data set. One is a quadratic mixed model and the other one is a piecewise linear mixed model. Although the linear model seems inappropriate, I still include linear model in the sequence in order to compare the models numerically to attain a complete modeling sequence.

**Unconditional Means Model**

Following the sequence, I fit an unconditional means model according to the specification below.

Level 1: $Y_{tij} = \pi_{0ij} + e_{tij}$ time level

Level 2: $\pi_{0ij} = \beta_{00j} + r_{0ij}$ city level

Level 3: $\beta_{00j} = \gamma_{000} + u_{00j}$ zone level

Here, (*t*=1,…, 12), (*j*=1,…, 8)and (*i*=1, 2, 3).The result of R is presented in the panel below:

```
> summary(model.non)
Linear mixed model fit by maximum likelihood  ['merModLmerTest']
Formula: AQI ~ 1 + (1 | CityID) + (1 | ZoneID)
   Data: report
     AIC      BIC   logLik  deviance   df.resid
  1045.1   1058.6   -518.5   1037.1      212
Random effects:
Groups   Name        Variance  Std.Dev.
CityID   (Intercept)    0.2695    0.5192
ZoneID   (Intercept)   1.6817    1.2968
Residual             6.6488    2.5785
Number of obs: 216, groups: CityID, 18; ZoneID, 3
Fixed effects:
           Estimate   Std. Error     df  t value  Pr(>|t|)
(Intercept)   4.4506     0.7787 2.9994   5.716   0.0106 *
```

~~Table 1: *R output of the unconditional means model fit for the report data*~~

From the output above, we get the AIC value of 1045.1. In this report, I use AIC as the index to select the models using the rationale outlined earlier.

From the random effects part, we can see that the variance in random intercepts between the zones is much larger than the variance between cities which means that the difference between the cities within one cluster is much smaller compared to the difference between the 3 zones. And we can calculate the proportion of variance explained at level 3 (zone level), also called the ICC = 1.68/ (0.27+1.68+6.65) × 100%=19.53%. This means that the zone residuals explain about 20% of the unconditional variation in residuals. This makes sense when looking at the map above Three economic zones share some similar characteristics however, the PM 2.5 indexes are very different from each other. We can see that PM 2.5 is clustered in zones. The characteristics which cause the big difference between these 3 economic zones could be explored by adding predictors to the appropriate levels or hierarchies in our multilevel model.

**Unconditional Growth Piecewise Model**

Next, I analyze an unconditional growth piecewise model with the breakpoint at time6.

Level 1 $AQI_{tij} = \gamma_{0ij} + \gamma_{1ij}T_{1ij} + \gamma_{2ij}T_{2ij} + e_{tij}$

Level 2: $\gamma_{0ij} = \beta_{00j} + r_{0ij}$

$\gamma_{1ij} = \beta_{10j}$

Level 3 $\beta_{00j} = \gamma_{000} + u_{00j}$

$\beta_{10j} = \gamma_{100}$

The result of R is presented in the panel below:

```
Formula: AQI ~ 1 + Time.rate1 + Time.rate2 + (1 | CityID) + (1 | ZoneID)
        Data: report
AIC      BIC   logLik deviance df.resid
975.9    996.1   -481.9    963.9     210
Random effects:
Groups   Name       Variance Std.Dev.
CityID   (Intercept)    0.4408   0.6639
ZoneID   (Intercept)   1.6817   1.2968
Residual           4.5938   2.1433
Number of obs: 216, groups: CityID, 18; ZoneID, 3
Fixed effects:
   Estimate   Std. Error  df       t value   Pr(>|t|)
(Intercept)   6.7920     0.8175   3.6400   8.308   0.0017 **
Time.rate1    0.7425    0.0851 197.9900   8.725 1.11e-15 ***
Time.rate2   -0.8337    0.1041 197.9900  -8.008 9.77e-14 ***
```

Table 2: *R output of the unconditional growth piecewise model*

From the output above, we can see that the intercept, `Time.rate1` and `Time.rate2` are all significant. The intercept means that the population average AQI in the first month (July, 2012) is 6.79. The slope of `time.rate1` is positive and slope of `time.rate2` is negative which means that the air quality became worse from July, 2012 to January 2013 and then AQI became better from January to June. In the first rate, with one month increases, AQI gets higher by 0.7425 and for the second rate AQI gets lower by -0.8337 by each month. Importantly, AIC decreases a lot compared to the first unconditional means model.

### Unconditional Quadratic Growth Curve Model

According to the shape of the plot, a quadratic function is also reasonable so I fit a unconditional quadratic growth curve model below and then compare it to the piecewise above to get a relatively better function.

$$AQI_{tij} \ = \gamma_{0ij} + \gamma_{1ij}T_{1ij} + \gamma_{2ij}T^{2}{}_{2ij} + e_{tij}$$

The R output is shown below:

```
Formula: AQI ~ 1 + Time + Time.2 + (1 | CityID) + (1 | ZoneID)
   Data: report
    AIC      BIC   logLik deviance df.resid
  1009.7   1030.0   -498.9    997.7      210
 Random effects:
 Groups   Name        Variance Std.Dev.
 CityID   (Intercept) 0.3694   0.6078
 ZoneID   (Intercept) 1.6817   1.2968
 Residual             5.4507   2.3347
Number of obs: 216, groups: CityID, 18; ZoneID, 3

Fixed effects:
             Estimate Std. Error       df t value Pr(>|t|)
(Intercept)   2.36246    0.86408   4.54000   2.734   0.0455 *
Time         1.12953    0.17196 197.99000   6.569 4.38e-10 ***
Time.2      -0.09781    0.01506 197.99000  -6.494 6.61e-10 ***
```

Table 3: *R output of the unconditional growth quadratic model*

### Deviance Test between Quadratic and Piecewise Model

We can see that AIC increases and in order to make sure if it's significantly different from the piecewise model, I do the chi square test on the deviance to choose a better one. The chi square test is presented below:

```
> anova(model.qua,model.pwlow)
Data: report
Models:
object: AQI ~ 1 + Time + Time.2 + (1 | CityID) + (1 | ZoneID)
```

```
..1: AQI ~ 1 + Time.rate1 + Time.rate2 + (1 | CityID) + (1 | ZoneID)
       Df     AIC     BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
object  6 1009.72 1029.98 -498.86   997.72
..1     6  975.86  996.11 -481.93   963.86 33.862      0  < 2.2e-16 ***
```

Table 4 : *R output of the deviance test between unconditional growth quadratic model and unconditional piecewise growth model*

The *p*-value is almost 0, which means that the piecewise model is much better than the quadratic model. So I decide to use the piecewise model as the baseline model for the further analysis.

### Piecewise Growth Model Adding Level-1 Predictors

Next, according to the modeling sequence discussed previously, I fit the piecewise model by adding the time-varying covariate temperature and humidity to the model. Plus, this model also adds the time invariant (subject level) predictors SEA and Population, GDP[1].

Level1: $AQI_{tij} \ = \gamma_{0ij} + \gamma_{1ij}T_{1ij} + \gamma_{2ij}T_{2ij} + \gamma_{3ij}GDP_{ij} + \gamma_{4ij}Temp_{tij} + \gamma_{5ij}Humidity_{tij} + \gamma_{6ij}SEA_{ij} + \gamma_{7ij}Population_{ij} + e_{tij}$

Level 2:

$$\gamma_{0ij} = \beta_{00j} + r_{0ij}$$

$$\gamma_{1ij} = \beta_{10j}$$

$$\gamma_{2ij} = \beta_{20j}$$

$$\vdots$$

$$\gamma_{7ij} = \beta_{70j}$$

[1] AS the time period is not too long, here I assume GDP and Population is time-invariant during this period

Level 3 $$\beta_{00j} = \gamma_{000} + u_{00j}$$

$$\beta_{10j} = \gamma_{100}$$

$$\vdots$$

$$\beta_{70j} = \gamma_{700}$$

The R output is shown below:

```
Formula: AQI ~ 1 + Time.rate1 + Time.rate2 + TEMP + SEA + Population +
GDP + Humidity + +(1 | CityID) + (1 | ZoneID)
   Data: report
     AIC      BIC   logLik deviance df.resid
   962.2    999.3   -470.1    940.2      205
Random effects:
 Groups   Name        Variance Std.Dev.
 CityID   (Intercept)    0.05483  0.2342
 ZoneID   (Intercept)   0.05443  0.2333
 Residual            4.46201  2.1123
Number of obs: 216, groups: CityID, 18; ZoneID, 3
Fixed effects:
            Estimate Std. Error        df t value Pr(>|t|)
(Intercept)       7.882442        0.479188    23.820000    16.450   1.67e-14
***Time.rate1   0.364262   0.164552  23.110000   2.214 0.036997 *
Time.rate2   -0.515516   0.172910  32.230000  -2.981 0.005426 **
TEMP     -0.121972   0.027984  21.850000  -4.359 0.000255 ***
SEA       -1.538409   0.398585  20.730000  -3.860 0.000925 ***
Population   0.132198   0.069830  16.450000   1.893 0.076066 .
GDP      -0.056779   0.038534  16.400000  -1.473 0.159565
Humidity     0.005458   0.002761 174.560000   1.977 0.049591 *
```

Table 5: *R output of the piecewise model with level-1 predictors*

From the output above, we can see that the coefficients of `time.tate1`, `time.rate2`, `TEMP`, `SEA` and `Humidity` are all significant. The *p*-value of Population is less than 0.1. The coefficient by sea is negative and the *p*-value is very small. It means that the cities near the sea have lower AQI (better air) than the inland cities. Similarly, when the weather is hotter, the AQI tends to be lower. The coefficient of Humidity here is almost 0 and it means that raining doesn't effect AQI too much. Plus, the coefficient of population here is not statistically significant which means that population has no big effect on AQI.

**Growth Piecewise Model Adding Higher-level Predictors**

The following step is to access whether any of the slopes of any of the explanatory variables has a significant variance component between the groups. This is called random coefficient model, testing for random slope variation is best done on a variable by variable basis. Here, I found that temperature can be added as random coefficient to the zone level (third-level).

Level1: $AQI_{tij} = \gamma_{0ij} + \gamma_{1ij}T_{1ij} + \gamma_{2ij}T_{2ij} + \gamma_{3ij}GDP_{ij} + \gamma_{4ij}Temp_{tij} + \gamma_{5ij}Humidity_{tij} + \gamma_{6ij}SEA_{ij} + \gamma_{7ij}Population_{ij} + e_{tij}$

Level 2:
$$\gamma_{0ij} = \beta_{00j} + r_{0ij}$$
$$\gamma_{1ij} = \beta_{10j}$$
$$\gamma_{2ij} = \beta_{20j}$$
$$.$$
$$.$$
$$.$$
$$\gamma_{7ij} = \beta_{70j}$$

Level 3
$$\beta_{00j} = \gamma_{000} + u_{00j}$$
$$\beta_{10j} = \gamma_{100}$$
$$.$$
$$\beta_{40j} = \gamma_{100} + u_{40j}$$
$$.$$
$$.$$

$$\beta_{70j} = \gamma_{700}$$

```
   Formula: AQI ~ 1 + Time.rate1 + Time.rate2 + TEMP + SEA + Population +
   (1 | CityID) + (1 + TEMP | ZoneID)
    Data: report
      AIC      BIC   logLik deviance df.resid
    959.8    996.9   -468.9    937.8      205

Random effects:
Groups   Name        Variance Std.Dev. Corr
CityID   (Intercept) 0.064506 0.25398
ZoneID   (Intercept) 4.458576 2.11153
         TEMP        0.005485 0.07406  -1.00
Residual             4.284603 2.06993
Number of obs: 216, groups: CityID, 18; ZoneID, 3
Fixed effects:
            Estimate   Std. Error      df t value Pr(>|t|)
(Intercept)     6.531843   1.358571  1.935000   4.808  0.04334 *
Time.rate1     0.638392   0.186102 12.182000   3.430  0.00488 **
Time.rate2    -0.717268   0.180847 16.508000  -3.966  0.00105 **
TEMP       -0.003071   0.060306  3.816000  -0.051  0.96194
SEA        -1.316430   0.363380 19.374000  -3.623  0.00177 **
Population   0.030536   0.028753 15.297000   1.062  0.30470
```

Table 6: *R output of the piecewise model with level-1 predictors*

**Adding Cross-level Interactions**

The last step is to decide whether to add cross-level interactions between explanatory group-level variables and those individual-level explanatory variables that had significant slope variation found in the previous step which is temperature. In this step, by exploring possible cross-level interactions, I decided to add the interaction between the time slopes and temperature.

```
Formula: AQI ~ 1 + Time.rate1 + Time.rate2 + TEMP + SEA + Population +
    TEMP * Time.rate1 + TEMP * Time.rate2 + (1 | CityID) + (1 +      TEMP |
ZoneID)
   Data: report
  AIC      BIC   logLik deviance df.resid
  878.3    922.2   -426.2    852.3      203
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 CityID   (Intercept) 1.6113812 1.26940
 ZoneID   (Intercept) 8.3089212 2.88252
          TEMP        0.0003924 0.01981  -1.00
 Residual             2.4146661 1.55392
Number of obs: 216, groups: CityID, 18; ZoneID, 3

Fixed effects:
                 Estimate Std. Error        df t value Pr(>|t|)
(Intercept)      10.117800   1.873216   3.130000   5.401   0.0111 *
Time.rate1        3.795110   0.347539 201.720000  10.920   <2e-16 ***
Time.rate2       -3.889148   0.327572 206.540000 -11.873   <2e-16 ***
TEMP         0.001909   0.040907  12.520000   0.047   0.9635
SEA          -0.899374   0.771952  21.070000  -1.165   0.2570
Population       -0.084765   0.061303  13.250000  -1.383   0.1896
Time.rate1:TEMP -0.093161   0.008515 186.980000 -10.940   <2e-16 ***
Time.rate2:TEMP 0.106005   0.009648 195.460000  10.987   <2e-16 ***
```

Table 7: *R output of the final piecewise model with interactions*

Similarly, by doing the global chi -square test, I reached the conclusion that the model in the last step with interactions improves the model compare to the previous models. From the output, we found that both the interactions are significant. The interaction between temperature and the first slope of time is negative which means that during the first period, when the temperature gets higher, the AQI tends to get lower. However, when it comes to the second period (January-June), the temperature and time has positive correlation.

By all the steps explained above, it leads to the full model. The full model is a piecewise model with time-level and zone-level predictors and also with temperature by time interactions. The multi-level equations of the final model are presented as below:

Level1: $AQI_{tij} = \gamma_{0ij} + \gamma_{1ij}T_{1ij} + \gamma_{2ij}T_{2ij} + \gamma_{3ij}GDP_{ij} + \gamma_{4ij}Temp_{tij} + \gamma_{5ij}Humidity_{tij} + \gamma_{6ij}SEA_{ij} + \gamma_{7ij}Population_{ij} + e_{tij}$

Level 2:

$$\gamma_{0ij} = \beta_{00j} + r_{0ij}$$

$$\gamma_{1ij} = \beta_{10j}$$

$$\gamma_{2ij} = \beta_{20j}$$

.

.

.

$$\gamma_{7ij} = \beta_{70j}$$

Level 3

$$\beta_{00j} = \gamma_{000} + u_{00j}$$

$$\beta_{10j} = \gamma_{100}$$

.

$$\beta_{40j} = \gamma_{400}+\gamma_{401}T_{1j} + \gamma_{402}T_{2j} + u_{40j}$$

.

.

$$\beta_{70j} = \gamma_{700}$$

# CONCLUSION

In this report, the model building is mainly according to the sequence described by West et al. (2007) with respect to the bottom-up procedures and the reference by Singer, J. D., & Willett, J. B (2003), which includes the non-linear situations. This report also compares the quartic curve model with piecewise growth model with respect to this data. The final model I reached is a piecewise model with time-level and zone-level predictors and also with temperature by time interactions.

From the result, I found that the zone-level explains much more variance than the city-level which explains the clustering pattern of AQI. The differences on AQI in these 3 zones are mainly attributed to the factors like temperature, geographical location and population. Especially, I found that GDP does not significantly affect city's AQI. This conclusion could explain why Shenzhen in PRD and shanghai in YRD, who with their higher GDPs, have much lower AQI than Shijiazhuang in BER.

# DISCUSSION

There are some limitations of this report. The first one is the sample size of my study is not big enough. The data-collection of longitudinal study is time-consuming so my study is lack of observations. With larger sample size at all levels, the estimates and their standard errors would be more accurate. Also, because of the insufficient data, in this report I treat the variable population and GPA during July, 2012 to June, 2013 as time-invariant variables. However, this is not the case. Plus, I could also consider other predictors such as industry emissions and numbers of vehicles in those cities to better predict AQI.

# Appendix

R code

```
install.packages('lme4')
install.packages('ggplot2')
install.packages('lmerTest')
library(lme4)
library(ggplot2)
library(lmerTest)
report <- read.csv("C:/Users/Think/Desktop/reportt.csv")
#change the order form July 2012-2013 June##
p.1 <-ggplot(report, aes(Time, AQI, group = CityID)) + geom_line ()
p.1 + geom_smooth(aes(group = 1), method = 'lm', size = 2, se = F)
p.1+ geom_smooth(aes(group = 1), method = 'loess', size = 2, se = F)
model.non <- lmer(AQI ~ 1 +(1| CityID)+(1| ZoneID), data = report, REML =
F)
model.pwinter<-lmer(AQI~1+Time.rate1+
Time.rate2+TEMP+SEA+Population+TEMP*Time.rate1+TEMP*Time.rate2+(1|
CityID)+(1 +TEMP|ZoneID), data = report, REML = F)
summary(model.pwinter)
summary(model.pwall)
summary(model.pwran2)
summary(model.non)
anova(model.pwall,model.pwran)
report$Time.2 <- report$Time^2
model.pwlow<-  lmer(AQI  ~  1  +Time.rate1  +  Time.rate2+(1   |  CityID)+(1
|ZoneID), data = report, REML = F)
b <-6                                  # break point
prior <- 11 - b                        # time after break
report$Time.rate1 <- ifelse(report$Time <= b, (report$Time - 6), 0)
# period 1 growth rate
report$Time.rate2 <- ifelse(report$Time > b,  (report$Time - 6),0)
# period 2 growth rate
Time<--5:6                                   # full range of time values
Time.rate1<-c(-5:0, rep(0, 6))               # make time values
Time.rate2 <- c(rep(0, 6), 1:6)              # make time values
model.pwall<-lmer(AQI~1+Time.rate1+
Time.rate2+TEMP+SEA+Population+GDP+Humidity++(1  | CityID)+(1 |ZoneID),
data = report, REML = F)
model.pwstep4<- lmer(AQI ~ 1 +Time.rate1 + Time.rate2+TEMP+SEA+(1 |
CityID)+(1 +TEMP+SEA|ZoneID), data = report, REML = F)
model.pwran2<- lmer(AQI ~ 1 +Time.rate1 +
Time.rate2+TEMP+SEA+Population+GDP+(1  | CityID)+(1 +TEMP+SEA|ZoneID),
data = report, REML = F)
```

# References

Hox, Joop J. (2010). *Multilevel Analysis: Techniques and Applications*. 2th ed. New York: Routledge.

Powers, Daniel A. and Yu Xie (2008) *Statistical Methods for Categorical Data Analysis, 2$^{nd}$ Edition*. London: Emerald

Singer, Judith, B. and John B. Willett (2003). *Applied longitudinal data analysis : modeling change and event occurrence*, New York:Oxford University Press

West, B. T., Welch, K. B., and Gatecki, A. T. (2007). *Linear mixed models*. Boca Raton, FL: Chapman &Hall.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.