

**Copyright**

**by**

**Xiaoyan Jiang**

**2009**

**Estimating Ground-level PM<sub>2.5</sub> in Texas from Remote Sensing  
Satellite Data with Interpolation and Regression Methods**

by

**Xiaoyan Jiang**

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

August 2009

**The Report committee for Xiaoyan Jiang**

**Certifies that this is the approved version of the following report**

**Estimating Ground-level PM<sub>2.5</sub> in Texas from Remote Sensing  
Satellite Data with Interpolation and Regression Methods**

**APPROVED BY**

**SUPERVISING COMMITTEE:**

**Supervisor: Paul Damien**

**Robert McCulloch**

## **Dedication**

I wish to dedicate this work to my husband, Yihua Cai, whose love and support keep me going each and every day. He has given so much of himself to see that I complete my degree, and it is time that I am able to give something back to him.

## **Acknowledgements**

I would like to begin my acknowledgements expressing my gratitude and appreciation to my supervisor, Dr. Paul Damien. His unwavering support and supervision during the course of my study allowed me to complete this journey. In addition, his input into this report from a scientific standpoint has been substantial. I would also like to thank my committee member, Dr. Robert McCulloch for his suggestions and comments on my report. Special thanks are given to Dr. Daniel Powers who consistently gave me academic advice during the whole period. Appreciation is also given to Dr. Sergey Fomel and Dr. Zong-Liang Yang for their suggestions and support on my thesis report.

In retrospect, there are more people including Ben Garcia and Li Jin that deserve thanks for helping me complete this work. Finally, I need to acknowledge the following institutions that provided the necessary support allowing me to pursue this degree. These include NASA Earth Science Fellowship program, Jackson School of Geoscience, EPA and NOAA.

# **Estimating Ground-level PM<sub>2.5</sub> in Texas from Remote Sensing Satellite Data with Interpolation and Regression Methods**

Xiaoyan Jiang, MSStat

The University of Texas at Austin, 2009

SUPERVISOR: Paul Damien

## **Abstract**

The integration of remote sensing satellite data in air quality monitoring system at a regional scale is an important method to provide high spatial / temporal resolution information. This work focuses on estimating high spatial / temporal resolution ground-level information about particulate matter with aerodynamic diameters less than 2.5  $\mu\text{m}$  (PM<sub>2.5</sub>), with the utilization of MODIS aerosol optical thickness (AOT) data and meteorological data. Several missing data reconstruction techniques including Bayesian inversion, regularization and prediction-error filter are employed to estimate PM<sub>2.5</sub> from satellite data. The results show that several direct missing data interpolation methods have the capability to estimate some distinctive features on the basis of available ground-based measurements, while the PEF method tends to generate more information with the aid of satellite AOT information.

In addition to interpolation methods, general linear regression methods are used to predict ground-level PM<sub>2.5</sub> with the consideration of other factors that have been shown to play an important role in predictions. Ordinary Least Square (OLS) method, when natural log taken on dependent and independent variables, is able to reduce the violation of homoscedasticity. The scatterplot of predicted and measured PM<sub>2.5</sub> shows a strong correlation over the validation region, indicating the ability of the regression model to predict PM<sub>2.5</sub>. Weighted Least Square (WLS) method also has advantage in improving homoscedasticity. The predicted and measured PM<sub>2.5</sub> has a relatively high correlation.

## Table of Contents

Chapter One: Introduction .....	1
Chapter Two: Datasets.....	4
Chapter Three: Estimating PM2.5 using Interpolation Method.....	6
3.1 Satellite AOD Data Reconstruction.....	6
3.2 PM2.5 Data Interpolation.....	7
3.3 Prediction-Error Filtering.....	8
3.4 Discussion.....	9
Chapter Four: Estimating PM2.5 using Linear Regression Method.....	17
4.1 Data Integration and Randomization for Model Development and Testing.....	17
4.2 Regression Methods.....	17
4.3 Regression Model Development.....	18
4.3.1 Descriptive Statistics.....	18
4.3.2 Regression Model Development and Discussion.....	22
Chapter Five: Discussion and Conclusions.....	29
References.....	30
Vita.....	32

### ***List of Tables***

Table 4.1: Summary statistics of PM, AOD, TMP, RH, PBL and UV.

Table 4.2: Regression statistics using the regression equation

$$\log PM = a + b \log AOD + c \log TMP + e \log PBL + f \log UV .$$

Table 4.3: Regression statistics using WLS method.

## ***List of Figures***

- Figure 2.1: Study domain (triangle markers represent ground-based PM2.5 observation stations). Data in the southern part is used to perform regression analysis. Northern part is used to do validation. Interpolation method is only applied to the red highlighted region with relatively dense observations.
- Figure 3.1: Impulse response of the nine-point Laplacian filter (a) gets inverted by recursive filtering (polynomial division) on a helix. (b) Division by  $D(Z)$ . (c) Division by  $D(1/Z)$ . (d) Division by  $D(Z)D(1/Z)$ .
- Figure 3.2: (a) MODIS AOT at 1745 on July06, 2002. (b) Random initial model with covariance specified by the inverse Laplacian filter.
- Figure 3.3: Nine-point filter: Result of missing data interpolation after 2000 iterations using (a) polynomial multiplication and (b) polynomial division on a helix.
- Figure 3.4: Locations of TCEQ PM2.5 stations.
- Figure 3.5: PM2.5 data interpolated using regularization with the Laplacian filter.
- Figure 3.6: PM2.5 data interpolated using shaping regularization with a triangle filter.
- Figure 3.7: Prediction-error filter.
- Figure 3.8: PM2.5 data: binned and mask.
- Figure 3.9: PM2.5 data interpolated using prediction-error filter.
- Figure 3.10: Comparison among different interpolation methods.
- Figure 3.11: Difference between PEF-based simulation and Shaping regularization interpolation.
- Figure 4.1: Histograms of PM, AOD, TMP, RH, PBL and UV for the Southern part.
- Figure 4.2: Histograms of PM, AOD, TMP, RH, PBL and UV for the Northern part.
- Figure 4.3: Histograms of logs of PM, AOD, TMP, RH, PBL and UV for the Southern part.
- Figure 4.4: Histograms of logs of PM, AOD, TMP, RH, PBL and UV for the Northern part.
- Figure 4.5: Residual plots for PM using the regression equation  
$$PM = a + bAOD + cTMP + dRH + ePBL + fUV .$$
- Figure 4.6: Scatterplot of predicted and measured  $PM$ .
- Figure 4.7: Residual plots for  $\log PM$  using the regression equation  
$$\log PM = a + b \log AOD + c \log TMP + e \log PBL + f \log UV .$$
- Figure 4.8: Scatterplot of predicted and measured  $\log PM$ .
- Figure 4.9: Residual plots for  $\log PM$  using WLS method.
- Figure 4.10: Scatterplot of predicted and measured  $Wt \log PM$  .

## Chapter One: Introduction

Aerosols are one of the major air pollutants responsible for human health problems, and are one of the largest uncertainties in climate research. The tiny airborne particulate matter (PM) is a complex mixture of solid and liquid particles that vary in size and composition. PM with aerodynamic diameters less than 2.5  $\mu\text{m}$  is called PM<sub>2.5</sub>, which could cause respiratory and lung diseases [Krewski et al., 2000]. Understanding the impacts of aerosols on Earth's climate system and human health requires long-term monitoring of aerosols or PM<sub>2.5</sub> on a large scale. This task is challenging because operating and maintaining such networks are costly, in particularly for many developing countries.

While long-term monitoring the ground-level aerosols over a large scale does not exist, satellite remote sensing tools and air quality models are being developed to estimate PM<sub>2.5</sub> concentrations. Sophisticated atmosphere chemistry models are developed to estimate detailed PM<sub>2.5</sub> information over specific locations (e.g. CMAQ, CAMX, WRF-CHEM) [Binkowski, 2003; ENVIRON, 1998; Grell, 2005]. However, the predictions of PM<sub>2.5</sub> by these models may be biased for various reasons such as the lack of some chemical reactions and the simple model assumptions, because they also require very detailed emissions inventory to perform simulations. The missing of some emissions inventories could also lead to unrealistic estimations. Monitoring aerosols using either ground-based or remote sensing techniques has attracted a lot of attention during the past few years. The installation of ground-based aerosol measuring stations (e.g., IMPROVE, AERONET and EPA routine sites), and the launch of satellite remote sensing instruments (e.g., Moderate Resolution Imaging Spectroradiometer (MODIS) and Multi-angle Imaging Spectroradiometer (MISR) have improved our view and understanding of aerosols near the surface and in the atmosphere [Kaufman et al., 2002; Di Girolamo et al., 2004].

The MODIS sensors were designed to systematically retrieve aerosol properties over both land and ocean on a daily basis [Kaufman et al. 1997; Tanré et al. 1997]. Some

studies have compared MODIS aerosol data with ground-based measurements, and their results show that it is suitable for monitoring air quality events over local, regional, and global scales [Chu et al. 2003; Wang and Christopher 2003; Hutchison 2008; Engel-Cox et al. 2004]. Although satellite observations, which provide a regional to global coverage, could serve as surrogates for monitoring PM<sub>2.5</sub> air quality to some extent, the clouds and aerosols in the upper atmosphere could contaminate the information retrieved from satellite. For example, over the areas covered by clouds, the satellite retrieved aerosol information is not very promising. However, there are limitations in current ground-based measurements. The ground-based measurements can only cover limited regions, with very sparse observation sites. The alternative way is to combine the two datasets together to provide more robust information about PM<sub>2.5</sub>, so that the improved PM<sub>2.5</sub> information can be further used to guide air quality study and management.

One object of this work is to use satellite data with the aid of interpolation methods to estimate spatially distributed ground-level aerosol concentrations. Some advanced missing data interpolation techniques are implemented to better restore missing satellite information. Several previous studies [e.g., Wang and Christopher, 2003; van Donkelaar et al., 2006; Hutchison et al., 2008] have shown that there is a high correlation between satellite-measured aerosol information and PM<sub>2.5</sub>. Thus, with the improved satellite data, it is possible for us to infer ground-level PM<sub>2.5</sub> information over the regions where there are no observations. The available near-surface PM<sub>2.5</sub> observations are treated as known values in the process of interpolation.

While the direct linkage between the PM<sub>2.5</sub> and satellite measured aerosol optical properties could be used to infer the ground-level PM<sub>2.5</sub>, other meteorological conditions have been shown to play an important role in controlling the level of PM<sub>2.5</sub> near the surface [e.g. Liu et al., 2005]. Consideration of the roles of other meteorological variables is of importance to estimating the PM<sub>2.5</sub> near the surface. In addition to the direct estimation from the satellite data, the second part of this work is to utilize the existing meteorological data along with the satellite measured aerosol products and ground-based measurements to better understand the relationship among them. Linear regression

method is applied to generate best fit linear regression equation among these variables to predict PM2.5 over other regions.

In this study, results are presented aimed at predicting better PM2.5 over the Texas region. We begin in chapter 2 with a brief description of the methods used in this study. In section 3, we apply an interpolation method with the utilization of satellite data to estimate PM2.5 over spatial domain. In chapter 4, linear regression methods used in this work are described and a better regression equation is achieved by doing ordinary least square (OLS) and weighted least square (WLS). The estimated regression equation is then used to predict PM2.5 over the regions where the data are not used in regression process to predict and validate PM2.5.

## Chapter Two: Datasets

The widely measured satellite information about aerosols is aerosol optical thickness (or depth) (AOD), which represents columnar information for ambient conditions and is directly correlated with the aerosol loading in the total atmospheric vertical column [Chu et al., 2002]. It is a dimensionless parameter that quantifies the degree to which aerosols prevent the transmission of light. The AOD data used in this study are measured by MODIS sensors onboard Terra and Aqua satellites, which have 36 spectral channels providing information about atmospheric, land and oceanic conditions. The MODIS provides observations in moderate spatial (from 250m to 1000m) and temporal (1-2 day) resolutions in different spectral regions of the electromagnetic spectrum. The AOD algorithm uses observed radiances in seven wavelengths. In this work, the level 2 AOD data (at 550 nm wavelength) is used, which has a spatial resolution of 10\*10km (MOD04). Studies have shown that PM<sub>2.5</sub> concentrations have a relatively high degree of spatial homogeneity over a 24 h period. Thus, one-year daily MODIS AOD dataset downloaded from the NASA's website (<http://ladsweb.nascom.nasa.gov/data/search.html>) is used in this study. The mean of the AOD measurements from 20\*20km MODIS pixel centered at a given observation site are calculated and matched with the PM<sub>2.5</sub> measurement taken at that site on the same day.

PM<sub>2.5</sub> represents near-surface aerosol concentrations. PM<sub>2.5</sub> mass concentration (ug/m<sup>3</sup>) data are acquired from air quality monitoring stations maintained by the Texas Commission on Environmental Quality (TCEQ), and are downloaded from this website ([http://www.tceq.state.tx.us/compliance/monitoring/air/monops/historical\\_data.html](http://www.tceq.state.tx.us/compliance/monitoring/air/monops/historical_data.html)). The data used in this work are for 2002. Preprocessing is performed to select PM<sub>2.5</sub> data over the regions where the satellite AOD data are available. Thus, the total number of monitoring sites for the PM<sub>2.5</sub> is 40, as shown in Figure 1. AOD data are extracted for the 40 sites. Because in some dates, PM<sub>2.5</sub> measurements are missing, the following analysis only uses available data in these sites.

Meteorological fields are provided by the National Centers for Environmental Prediction (NCEP)'s North American Regional Reanalysis (NARR) data set, which has a domain covering our configured computational area [Mesinger et al., 2006]. The NARR data were generated at a 3-hour interval with the use of the NCEP Eta model, its data assimilation system and a recent version of the Noah LSM at 32 km/45 layer resolution. As the dataset utilized a variety of observations, the generated reanalysis variables are quite close to observations. Therefore, in this study, we use them as observed meteorological variables. Variables of temperature (TMP), relative humidity (RH), planetary boundary layer height (PBL) and wind speeds (UV) play an important role in affecting the ground-level PM<sub>2.5</sub>. Variables near the available PM<sub>2.5</sub> sites are extracted to do regression analysis.

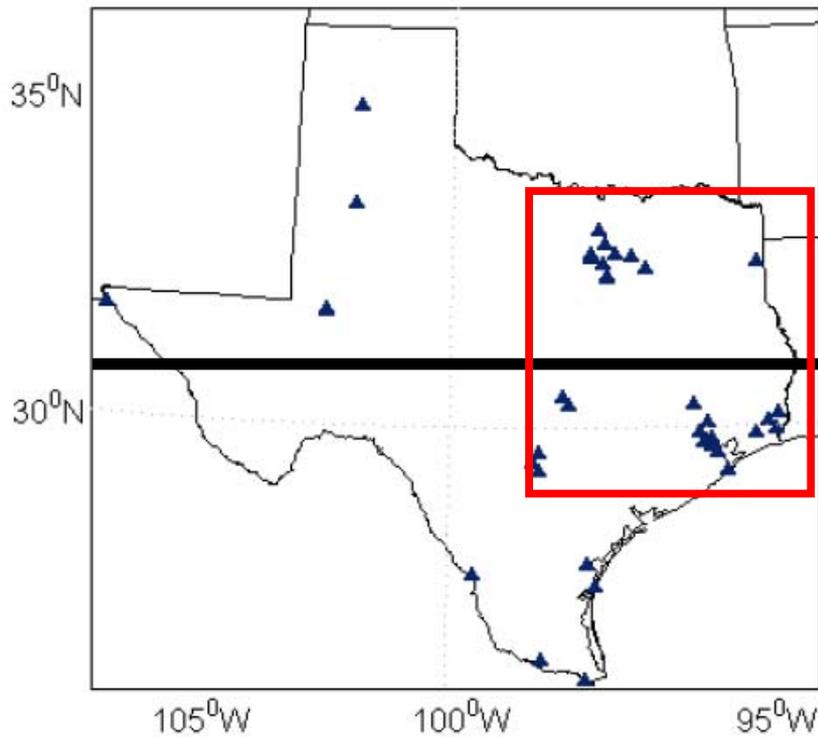


Figure 2.1. Study domain (triangle markers represent ground-based PM<sub>2.5</sub> observation stations). Data in the southern part is used to perform regression analysis. Northern part is used to do validation. Interpolation method is only applied to the red highlighted region with relatively dense observations.

## Chapter Three: Estimating PM2.5 using Interpolation Method

### 3.1 Satellite AOD Data Reconstruction

Due to the contamination of clouds, satellite AOD information is missing over part of the study domain, as shown in Figure 2a. Thus, the first part of this work is to reconstruct missing AOD data using some missing data interpolation methods. Several methods are available, including inverse interpolation, projection onto convex sets (POCS), prediction-error filtering. The philosophy behind the missing data reconstruction is to minimize energy after specified filtering. Using the ideas of inverse interpolation, we can extend the satellite data into the empty part of the domain. Bayesian Inversion suggests the model estimate (given data  $\mathbf{d}$ ) of the form

$$\mathbf{m}_* = \mathbf{m}_0 + \mathbf{C}_m \mathbf{F}^T \left( \mathbf{F} \mathbf{C}_m \mathbf{F}^T + \mathbf{C}_n \right)^{-1} (\mathbf{d} - \mathbf{F} \mathbf{m}_0) \quad (1)$$

We use the above estimate (1) to find missing part of AOD data.  $\mathbf{F}$  is the mask operator which is a diagonal matrix with ones and zeros on the diagonal using ones to mask the known data locations.  $\mathbf{C}_m$  is a stationary filter, and  $\mathbf{C}_n$  is close to zero. The stationary filter used here is the inverse of a nine-point Laplacian filter

$$\begin{aligned} \hat{L}_2(Z_1, Z_2) = 20 & - 4 Z_1 - 4/Z_1 - 4 Z_2 - 4/Z_2 \\ & - Z_1 Z_2 - Z_1/Z_2 - Z_2/Z_1 - 1/(Z_1 Z_2) \end{aligned} \quad (2)$$

To build the inverse, the Laplacian filter is put on a helix using the Wilson-Burg algorithm [Fomel et al., 2003]. The factorization is tested in Figure 3.1, where the impulse response of the Laplacian filter gets inverted by recursive filtering (polynomial division) on a helix.

Over-determined least-squares (polynomial multiplication on a helix) and underdetermined least-squares (polynomial division on a helix) are used as two alternative formulations of Bayesian least-squares inversion.

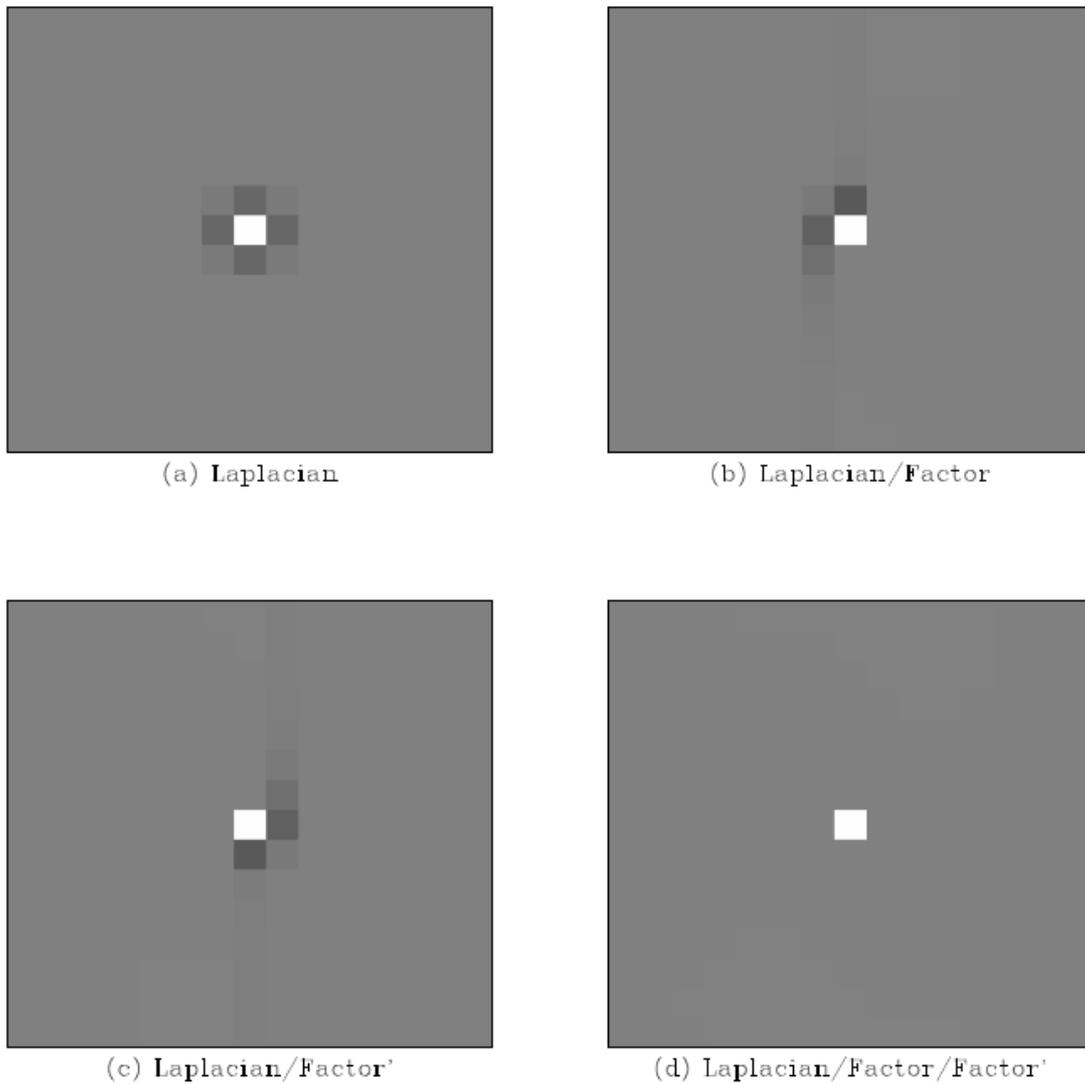


Figure 3.1: Impulse response of the nine-point Laplacian filter (a) gets inverted by recursive filtering (polynomial division) on a helix. (b) Division by  $D(Z)$ . (c) Division by  $D(1/Z)$ . (d) Division by  $D(Z)D(1/Z)$ .

### 3.2 PM2.5 Data Interpolation

One way to infer point-scale PM2.5 information over the regions where there are no observations is to perform missing data interpolation directly with respect to known data. Three methods of missing data interpolation without the utilization of satellite data are applied to the PM2.5 observations in this work.

The first technique we tried is Laplacian regularization, which finds a solution of the regularized least-squares optimization problem

$$\min \left( \|\mathbf{L} \mathbf{m} - \mathbf{d}\|^2 + \epsilon^2 \|\mathbf{R} \mathbf{m}\|^2 \right), \quad (3)$$

where  $\mathbf{d}$  represents irregular data. Here, it refers to PM2.5 observations over different locations.  $\mathbf{m}$  is model estimated on a regular grid.  $\mathbf{L}$  is forward interpolation from the regular grid to irregular locations,  $\epsilon$  is a scaling parameter, and  $\mathbf{R}$  is the regularization operator related to the inverse of the assumed model covariance. We selected  $\mathbf{R}$  as the finite-difference approximation of the Laplacian operator.

The second technique is Shaping regularization, which is an iterative solution of the inverse problem

$$\widehat{\mathbf{m}} = \left( \mathbf{L}^T \mathbf{L} + \mathbf{S}^{-1} - \mathbf{I} \right)^{-1} \mathbf{L}^T \mathbf{d} = \left[ \mathbf{I} + \mathbf{S} \left( \mathbf{L}^T \mathbf{L} - \mathbf{I} \right) \right]^{-1} \mathbf{S} \mathbf{L}^T \mathbf{d}, \quad (4)$$

where  $\mathbf{S}$  is the shaping operator, which is taken as a two-dimensional triangle smoothing here.

The third technique is called inverse distance weighted (IDW) interpolation. It is based on the assumption that the interpolating surface should be influenced most by the nearby points and less by the more distant points (5). The interpolating surface is a weighted average of the points and the weight assigned to each point decreases as the distance from the interpolation point to the observation site increases. The radius parameter can be used to control how many points will be in the region where the interpolation will be performed.

$$\mathbf{x} = \frac{\sum_k w_k x_k}{\sum_k w_k}; \quad (5)$$

### 3.3 Prediction-Error Filtering

In order to utilize available satellite AOD information to estimate spatial distribution of ground-level PM2.5 concentrations, the method of prediction-error filter

(PEF) [Claerbout and Brown, 1999] is implemented to gather statistics from satellite AOD. The PEF plays the role of the so-called "inverse-covariance matrix" in statistical concept. For simplicity, we assume the relationship between AOD and PM2.5 is stationary. This means that their statistical properties do not change. Based on the training data set, which, in this work is reconstructed AOD after using missing data interpolation technique, a PEF is estimated. By deconvolving (polynomial division) random numbers using the estimated PEF, we generated synthesized image that shares the covariance with the training data set. As this synthesized image shares information with the satellite AOD, we then use it as an initial random model to reconstruct PM2.5. Hence, the estimated PM2.5 shares some pattern of satellite AOD data. Since the estimated PM2.5 also includes the information from existing PM2.5 observations, the estimated PM2.5 using PEF is expected to be more informative than the ones using other direct interpolation techniques without obtaining information from other data sources. The main idea of this method is to fill the missing data points with something simulated by other data set.

### **3.4 Discussion**

Figure 3.2 shows the original MODIS AOD data (Figure 3.2a) and a random model (Figure 3.2b) created by dividing random normally distributed noise by  $D(Z)$ , which is derived from the Laplacian filter while converting the filter to a helix. The random model is used as an initial model for the missing data reconstruction. The results of missing data reconstruction from both (overdetermined least-squares and underdetermined least-squares) methods are shown in Figure 3.3 after 2000 conjugate-gradient iterations. Comparing the two methods, we found that the division method performs well in this case. The information around the boundaries is reconstructed.

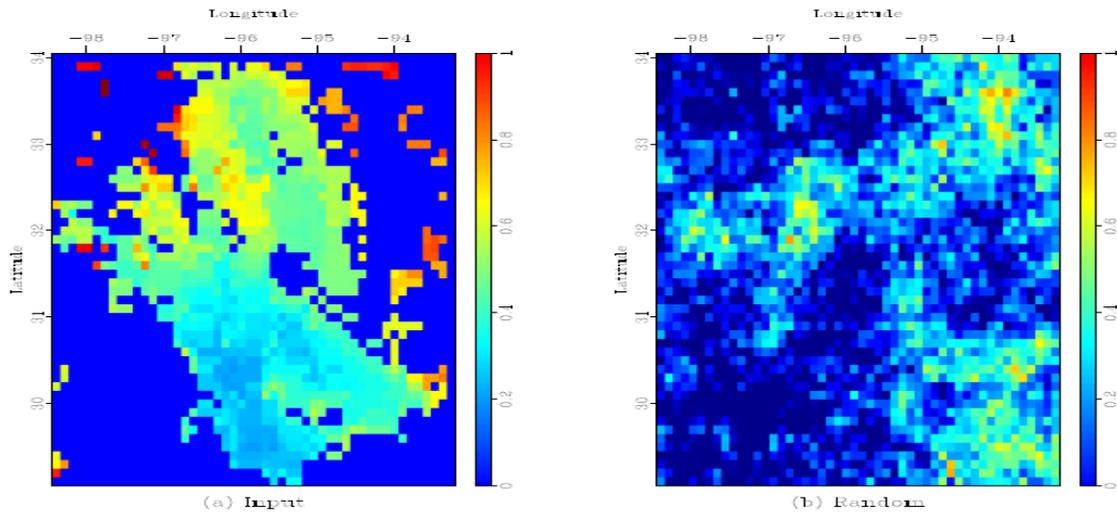


Figure 3.2: (a) MODIS AOT at 1745 on July06, 2002. (b) Random initial model with covariance specified by the inverse Laplacian filter.

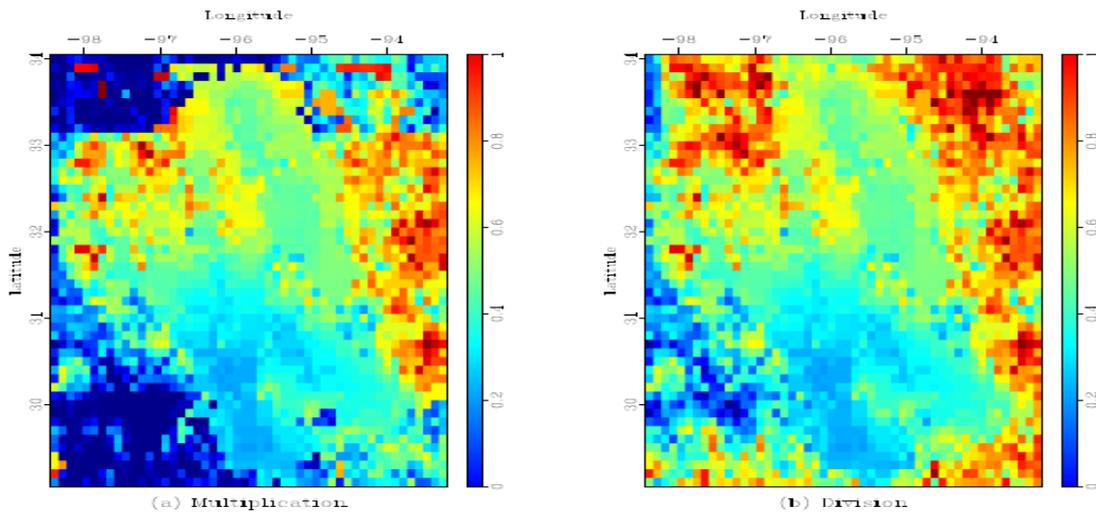


Figure 3.3: Nine-point filter: Result of missing data interpolation after 2000 iterations using (a) polynomial multiplication and (b) polynomial division on a helix.

The major goal of this work is to estimate PM<sub>2.5</sub> information using a PEF estimated from reconstructed satellite AOD. Before doing that, we performed missing data interpolation on irregular point-scale PM<sub>2.5</sub> (Figure 3.4) using three different methods introduced in subsection 3.2. Since there are only a few points over the west part

of the domain, we reselected domain to do interpolation. The reselected domain covers the same area as does the satellite AOD data set.

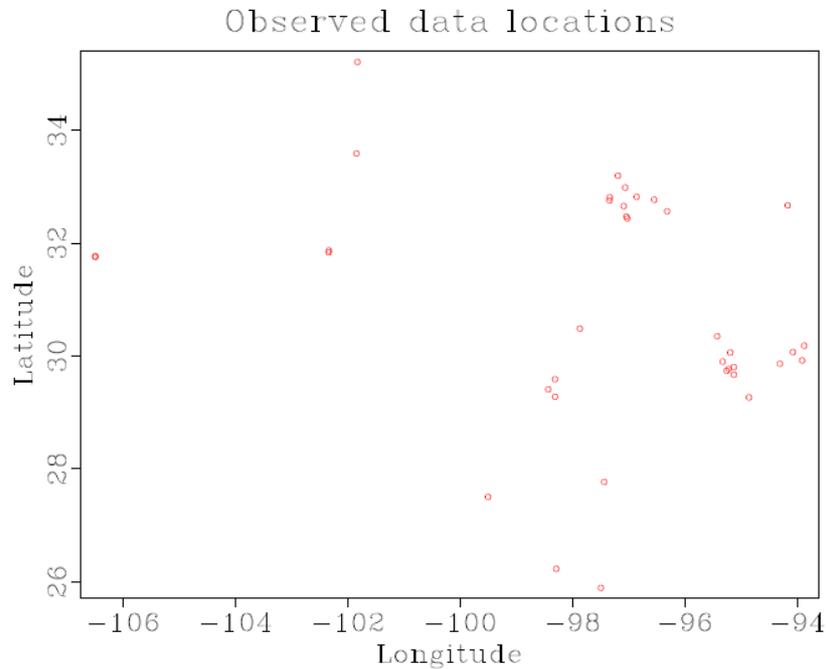


Figure 3.4: Locations of TCEQ PM2.5 stations.

Figure 3.5 shows the interpolation results after 10 and 1000 iterations using the method of Laplacian regularization. Clearly, 10 iterations are not enough for the method to converge. The result after 1000 iterations indicates some kind of pattern about the spatial distribution of PM2.5. The reconstructed high concentrations of PM2.5 over the southeast and northwest corners are also reflected on the AOD figure (Figure 3.2a). We also see low concentrations of PM2.5 over the western region. Using the method of Shaping regularization, the result converges fast (Figure 3.6). The pattern is much closer to that of AOD. The result using the IDW technique (in Figure 3.10) does not show some distinctive features regarding low and high concentrations.

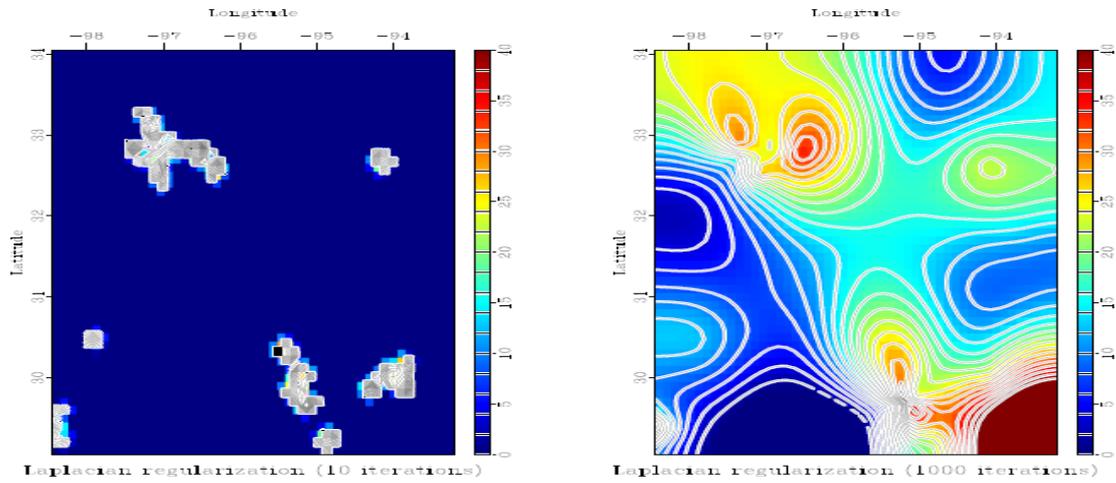


Figure 3.5: PM2.5 data interpolated using regularization with the Laplacian filter.

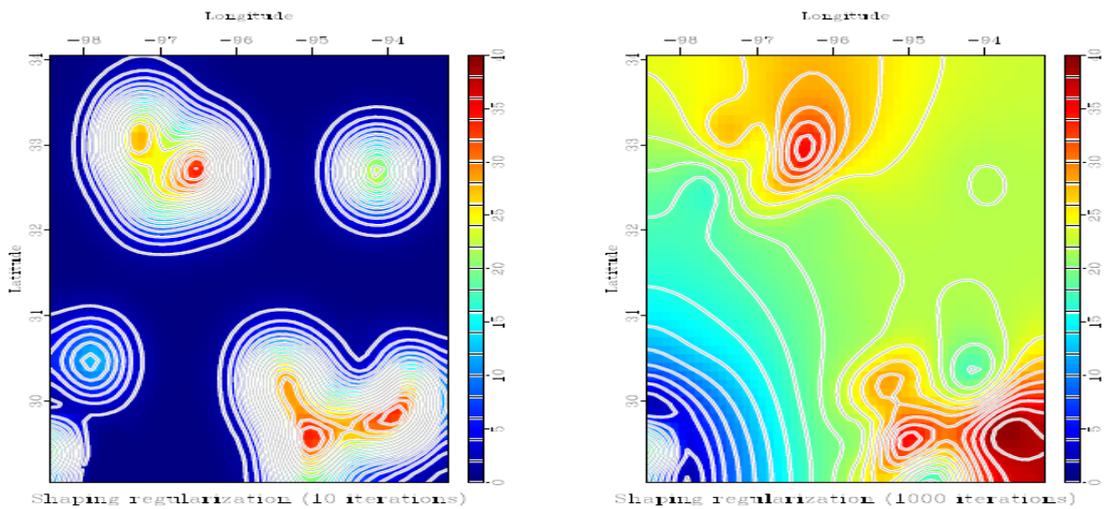


Figure 3.6: PM2.5 data interpolated using shaping regularization with a triangle filter.

To better incorporate the AOD information, the pattern realization using PEF has been performed. Figure 3.7a shows the AOD training image after removing its linear trend. The estimated PEF from Figure 3.7a is then convolved on the training data set (Figure 3.7b). By deconvolving (polynomial division) random numbers using the estimated PEF, we got a synthetic image, shown in Figure 3.7d. The method seems to extract some information with low or high values, as reflected in the simulated data.

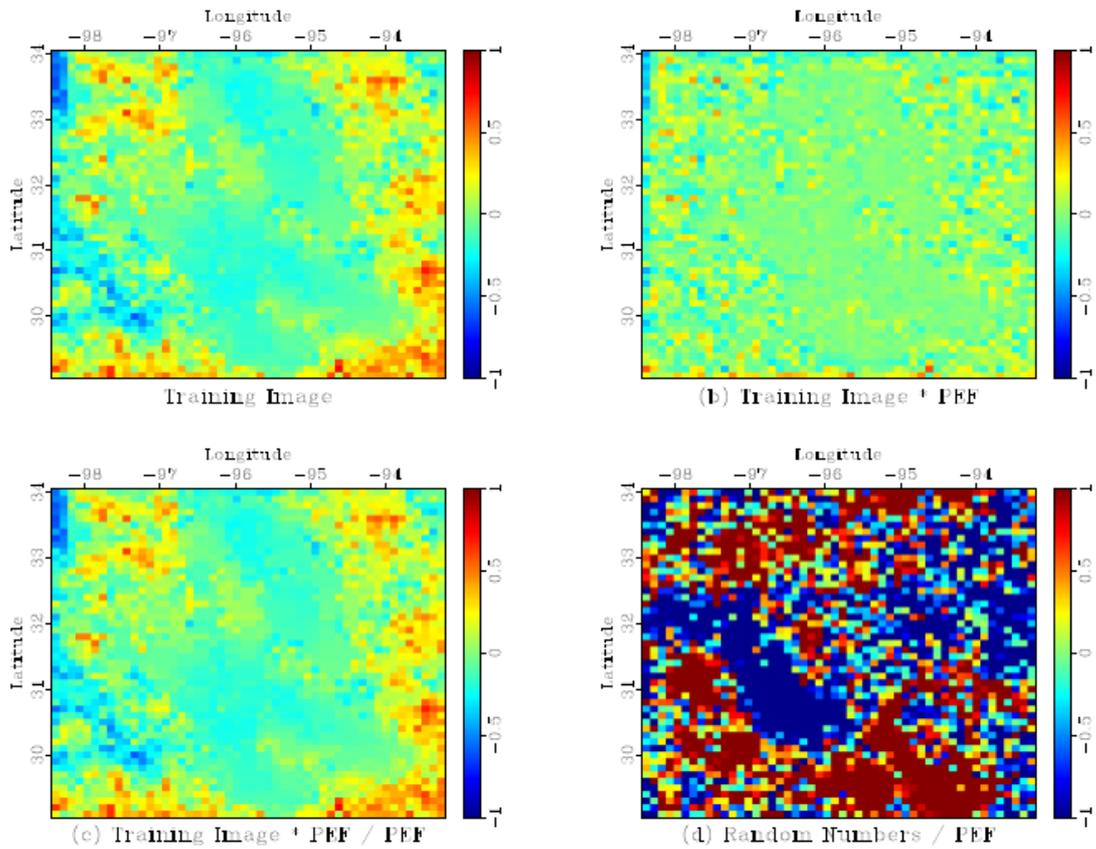


Figure 3.7: Prediction-error filter.

To employ the simulated data to missing data interpolation on irregular ground-based observations, we first set up a Cartesian mesh to show the point-scale data on regular grids (Figure 3.8a). Then, the known data points are masked out (Figure 3.8b). Results using PEF on the point-scale data are shown in Figure 3.9. After 2000 iterations, PEF simulations with and without predictions tend to show similar patterns, which differ from the results using interpolation methods directly. The significant differences lie over the regions where the Laplacian and Shaping regularizations don't result in high concentrations, in particular in the southwest region. As seen from Figure 3b, the reconstructed AOD over this region tends to show somewhat high concentrations information, which is well represented in estimated PM<sub>2.5</sub> information in Figure 3.9. Comparing all of results produced using different techniques; they all show some

distinctive features (Figure 3.10). The calculated difference between the Shaping regularization result and the PEF simulation (Figure 3.11) reflect some features represented in AOD data. Therefore, one can conclude that the integration of AOD to the ground-level PM2.5 estimation could help retrieve information which can't be reconstructed with the existing PM2.5 observations.

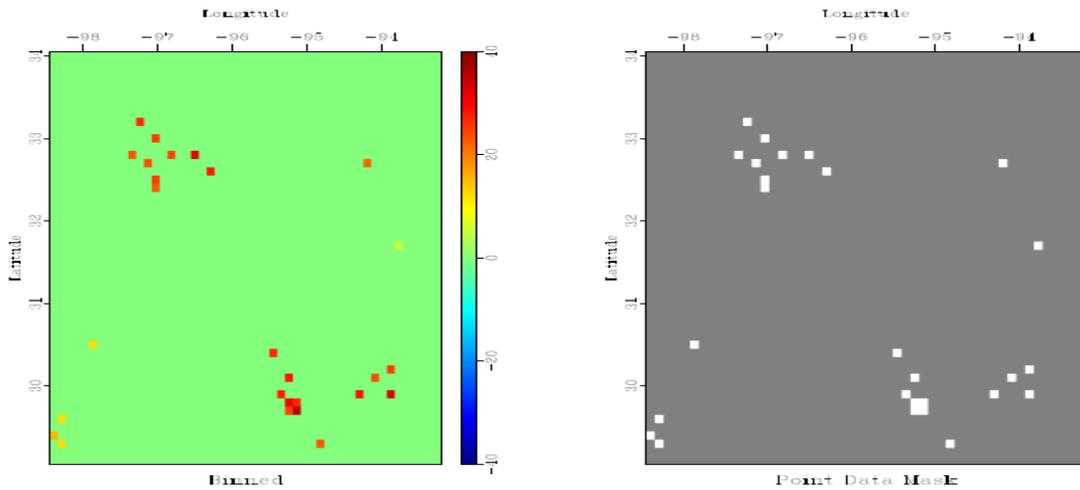


Figure 3.8: PM2.5 data: binned and mask.

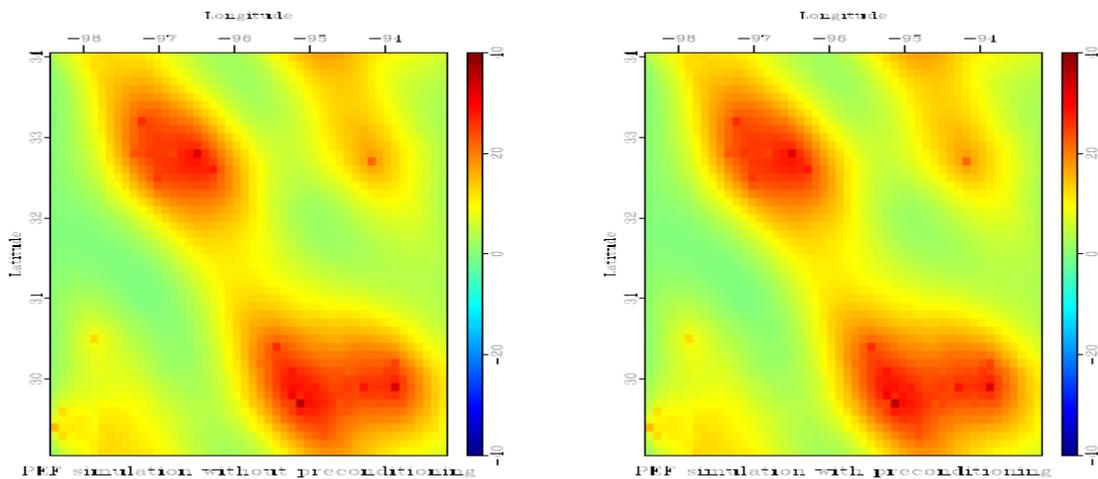


Figure 3.9: PM2.5 data interpolated using prediction-error filter.

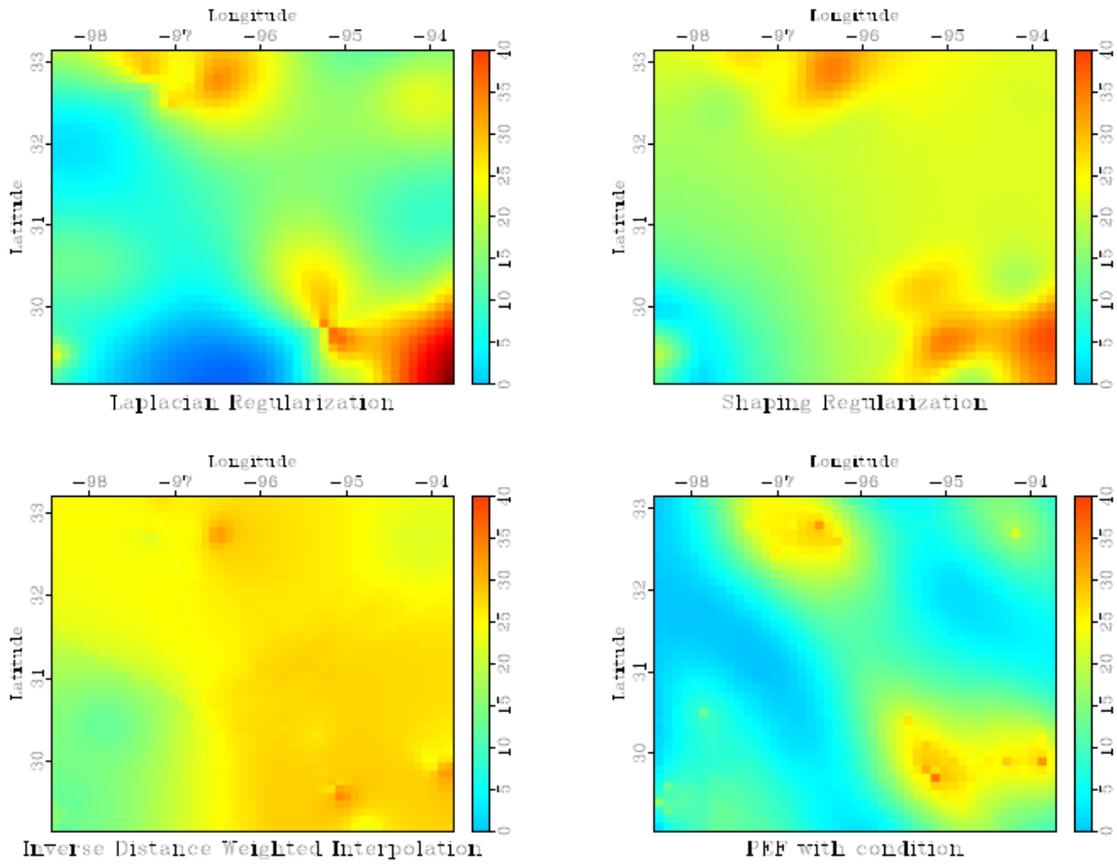


Figure 3.10: Comparison among different interpolation methods.

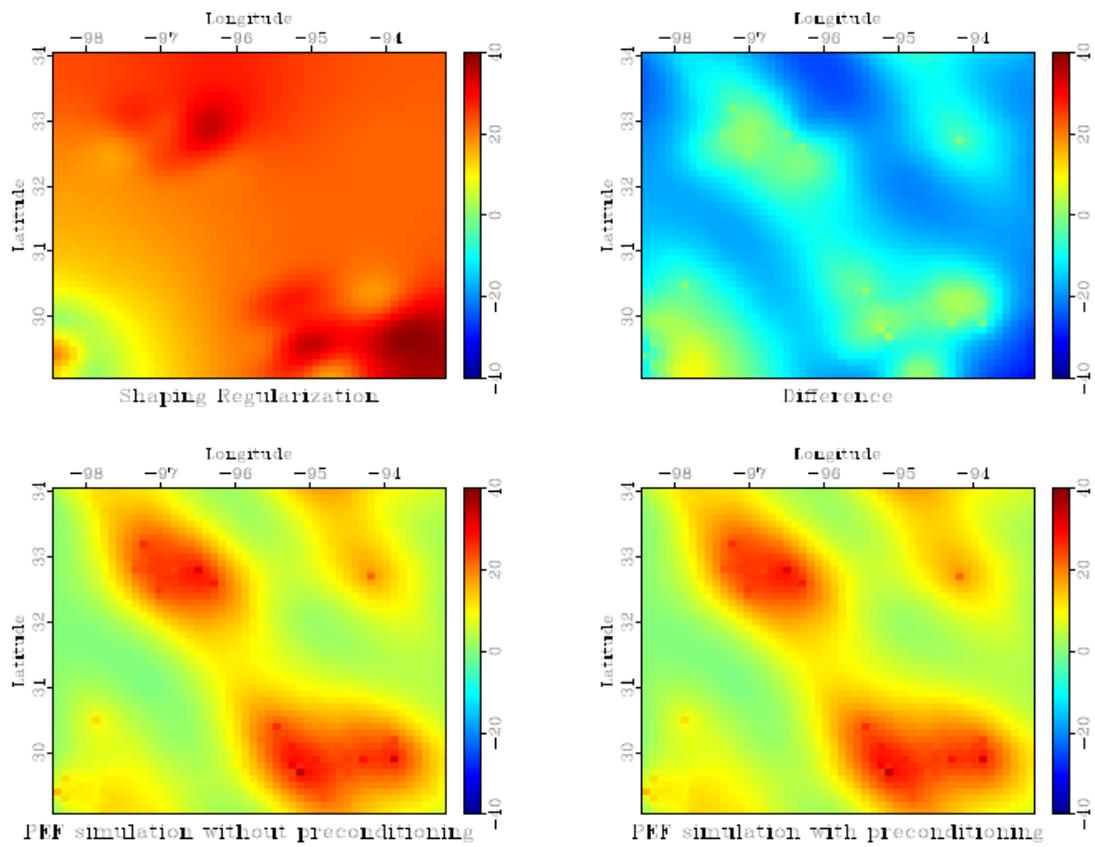


Figure 3.11: Difference between PEF-based simulation and Shaping regularization interpolation.

## **Chapter Four: Estimating PM<sub>2.5</sub> using Linear Regression Method**

### **4.1 Data Integration and Randomization for Model Development and Testing**

A total of 40 sites from TCEQ ground-based monitoring network are used in this study for 2002. Satellite derived AOD and meteorological variables are extracted over these sites in order to do regression analysis. To preserve a subset of the monitoring PM<sub>2.5</sub> data for regression model validation, the dataset was divided into two subsets (southern and northern parts). Therefore, 23 sites in the southern part (Figure 2.1) are used to develop linear regression model, and the remaining sites are used to evaluate the model performance. No auto-correlations were found between different sites; as a result, these sites are treated as independent observations. Thus, we believe that the data points finally collected are randomly divided into the model dataset and validation dataset.

### **4.2 Regression Methods**

In statistics, regression methods are used for analyzing several variables, with the focus on revealing the relationship between a dependent variable and other independent variables. Then the relationship can be used to predict or forecast the dependent variable using other available independent variables [Cook and Weisberg, 1999]. Thus, regression analysis is often used to understand the relationship between the independent variables and dependent variable, and to explore their mathematical relationships for other applications. A lot of effort has been put to develop techniques for regression analysis, such as linear regression and ordinary least squares (OLS) [Neter and Wasserman, 1990; Cook and Weisberg, 1999]. For these types of regression methods, the regression function is defined in terms of a finite number of unknown parameters that can be estimated from the available data. There are other techniques that are nonparametric regression. In this study, only parametric regression is applied to the datasets.

The OLS regression method minimizes the sum of squared distances between the observed responses and the fitted responses from the regression model. This technique provides simple expressions for the estimated parameters. It also calculates other

associated statistical values such as the standard errors of the parameters and standardized residuals. The standardized residuals can provide information regarding homoscedasticity. Under some circumstances (e.g., the data are normally distributed), OLS method usually provides optimal estimates. It should be noted that one of the critical assumptions of OLS regression is homoscedasticity which requires the variance of residual error to be constant for all values of the independent(s). If the independent(s) has/have different error variance at different ranges of their values, then the estimates of the regression coefficients will have unduly large standard errors for some ranges of the dependent and too small for other ranges. In other words, violation of homoscedasticity occurs when error variance is correlated with the magnitude of the dependent, suggesting the magnitude of the dependent is also correlated with the variance of the independents. Plot of standardized residual versus fitted values can be used to easily detect homoscedasticity problem. In addition, normal probability plot of the residuals could be seen skewed at some points if violation of homoscedasticity occurs.

When the homoscedasticity is violated, WLS can be used to compensate for violation of the homoscedasticity assumption by weighting cases differently. The ones with large variances on the independent variables will have small weights and those with small variances will count more in estimating the regression coefficients. That means cases with larger weights contribute more to the fit of the regression line and others with smaller weights contribute less to the fit. Usually, the estimated coefficients are very close to those estimated with OLS, but the standard errors are smaller. In the following analysis, different methods are explored to find the best fit of estimated and observed PM2.5.

## **4.3 Regression Model Development**

### **4.3.1 Descriptive Statistics**

Before developing appropriate regression model, histograms and summary statistics of data are examined. Histograms (Figure 4.1, 4.2, 4.3 and 4.4) of the various parameter distributions showed that, for both the model and the validation datasets, these

variables are unimodal and log-normally distributed. The summary statistics are presented in Table 4.1. The annual mean PM concentration for all sites is 9.253 ug/m<sup>3</sup>. The overall mean AOD is 0.224. The correlation analysis between all variables shows that there is a relatively high correlation between PM and AOD (0.699).

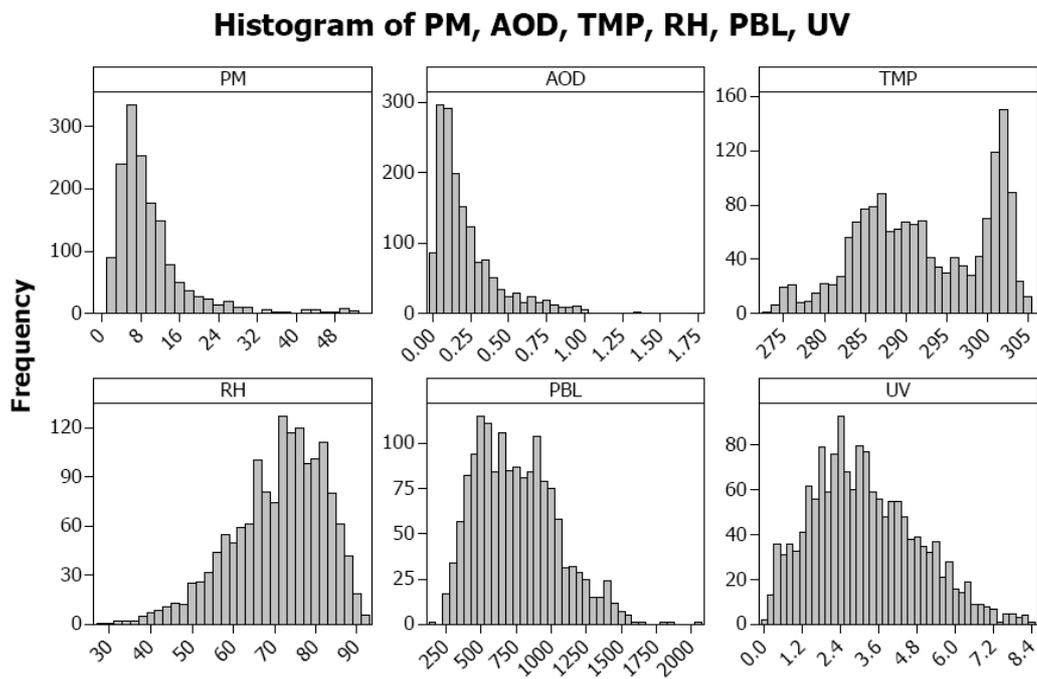


Figure 4.1: Histograms of PM, AOD, TMP, RH, PBL and UV for the Southern part.

### Histogram of PM, AOD, TMP, RH, PBL, UV

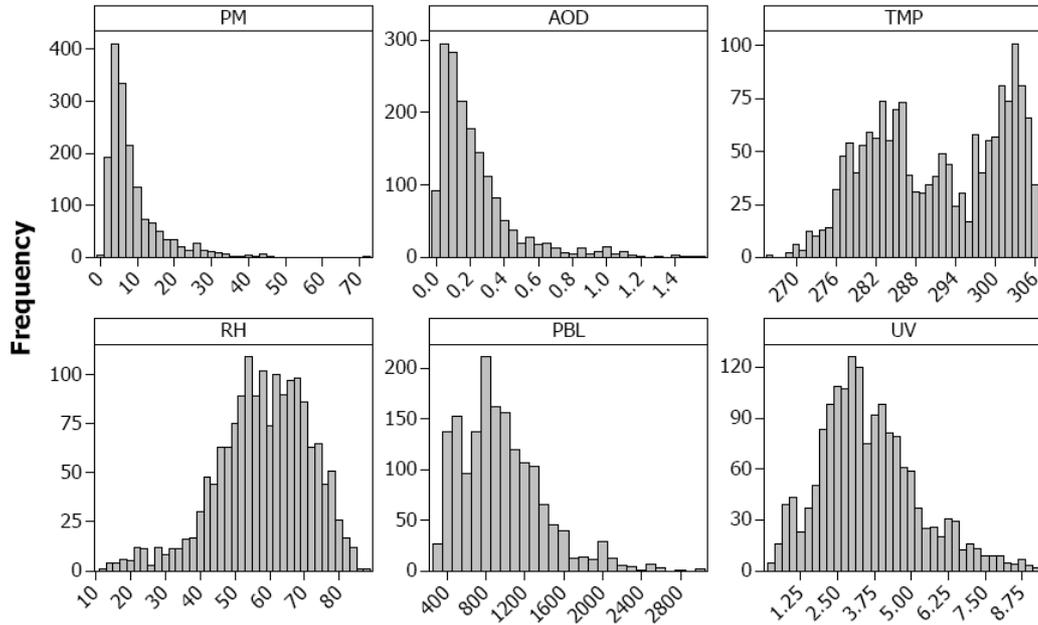


Figure 4.2: Histograms of PM, AOD, TMP, RH, PBL and UV for the Northern part.

### Histogram of LogePM, LogeAOD, LogeTMP, LogeRH, LogePBL, LogeUV

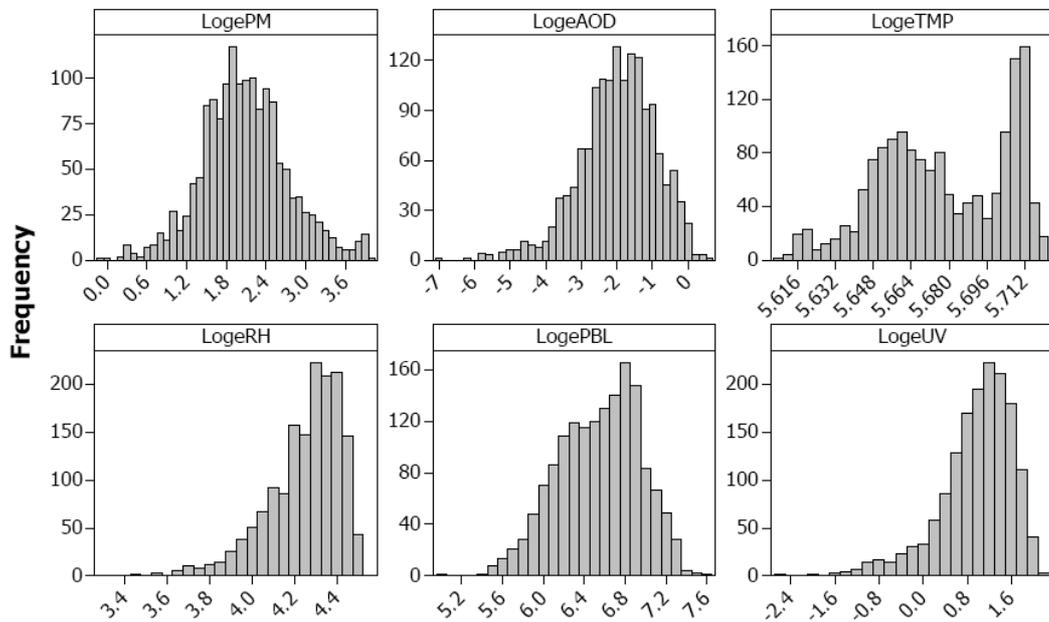


Figure 4.3: Histograms of logs of PM, AOD, TMP, RH, PBL and UV for the Southern part.

## Histogram of LogePM, LogeAOD, LogeTMP, LogeRH, LogePBL, LogeUV

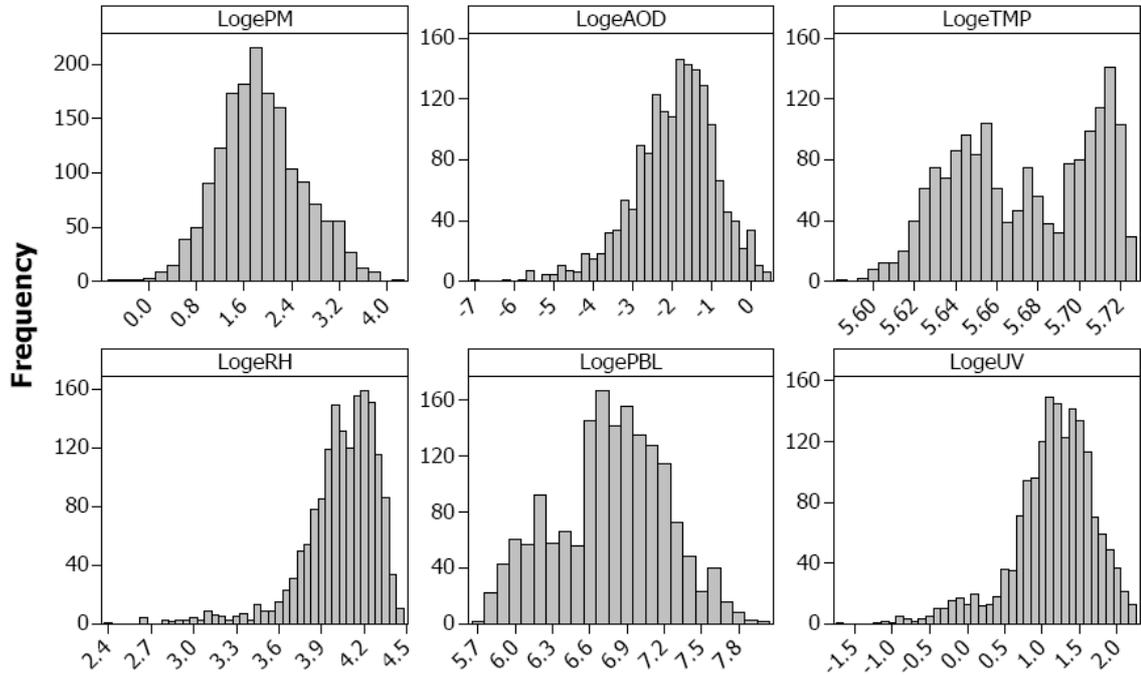


Figure 4.4: Histograms of logs of PM, AOD, TMP, RH, PBL and UV for the Northern part.

Variable	N	Mean	SEMean	StDev	Minimum	Median	Maximum
PM	3212	9.253	0.134	7.576	0.5	7	71.3
AOD	3212	0.22374	0.00392	0.22226	0.001	0.154	1.733
TMP	3212	291.65	0.158	8.98	266.45	291.38	306.37
RH	3212	64.369	0.246	13.956	11.1	65.928	92.371
PBL	3212	856.88	6.71	380.08	148.87	806.95	3026
UV	3212	3.3748	0.0296	1.6802	0.0708	3.1721	9.202

Table 4.1: Summary statistics of PM, AOD, TMP, RH, PBL and UV.

### 4.3.2 Regression Model Development and Discussion

The main purpose of regression analysis is to model the relationship between PM 2.5 and other variables. Hence, in this study, PM is called dependent variable, denoted as  $Y$ ; and others are called independent variables, denoted as  $X_1, X_2, \dots$ . Analysis of

relationships between independent variables shows that correlations among them are quite low, less than 0.5, thus we can say they are independent.

Linear regression is first used to fit the relationship between the dependent (Y) and the independents ( $X_1, X_2, \dots$ ). The linear regression approach assumes that the relationship is linear, thus the model takes the form of

$$Y = a + bX_1 + cX_2 + dX_3 + eX_4 + fX_f,$$

where  $X_1$  is AOD,  $X_2$  is TMP,  $X_3$  is RH,  $X_4$  is PBL and  $X_5$  is UV, and Y is PM.

Therefore,

$$PM = a + bAOD + cTMP + dRH + ePBL + fUV .$$

The dependent variable PM is daily averaged for 2002 with missing dates removed. The independent variables on the right-hand side include AOD, TMP, RH, PBL and UV, which are geographically matched to each PM site. The parameter a is regression constant, b, c, d, e and f are regression coefficients for different independent variables.

The regression model is estimated with the statistical significance of parameter estimate at the 0.05 level. As a first attempt to develop the regression model, regression analysis was performed between PM and other variables. P values for AOD, TMP, PBL and UV are less than 0.001, suggesting the significance between PM and these variables. However, the P value for RH is 0.75, indicating less importance of this variable to PM. Therefore, in the following analysis, RH could be excluded. The calculated R square is about 50.1 %. Plots of residuals for PM as shown in Figure 4.5 reveal non-normality in the probability plot, and violation of homoscedasticity which can be seen from the plot of skewed residuals versus the fitted values. The plot of predicted PM and measured PM (Figure 4.6) also shows strong scattering trend, indicating less accuracy of predicted PM.

### Residual Plots for PM

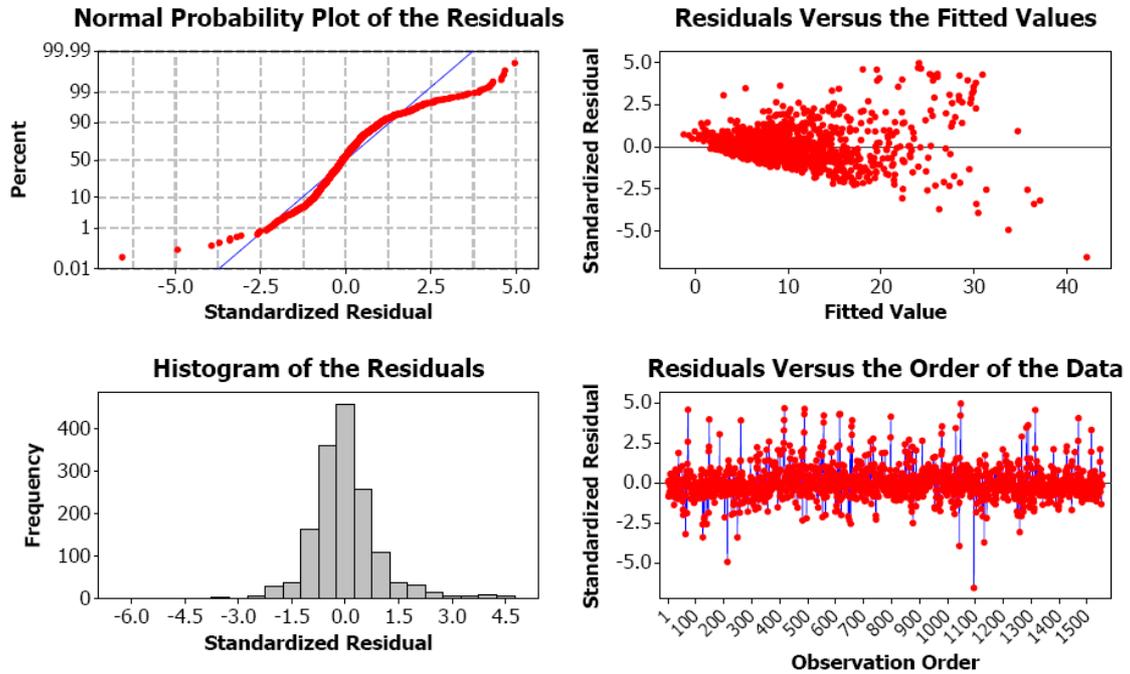


Figure 4.5: Residual plots for PM using the regression equation  $PM = a + bAOD + cTMP + dRH + ePBL + fUV$ .

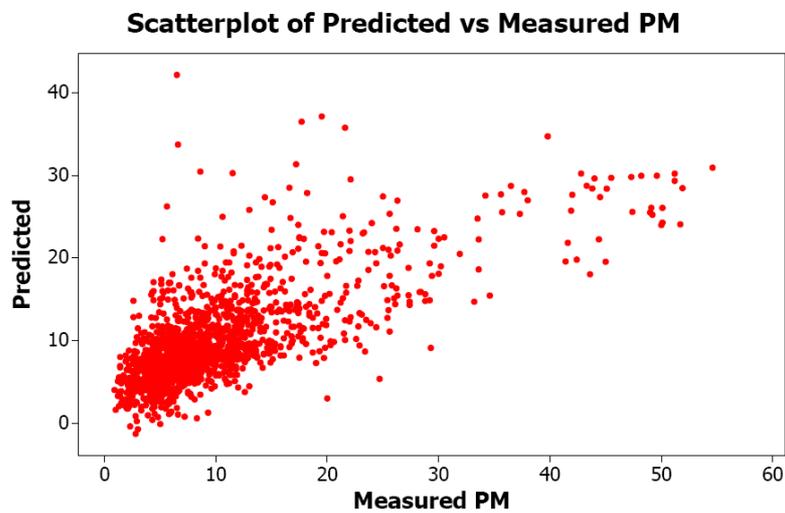


Figure 4.6: Scatterplot of predicted and measured  $PM$ .

The log-normality in Figure 4.3-4.4 suggests that taking natural log of variables might reduce the possibility of violation of homoscedasticity. The second attempt to develop the regression model is to take logs on those variables that show log-normality.

As RH does not highly correlate with PM, it is not included in the analysis. Now the regression equation is modified as follows,

$$\log PM = a + b \log AOD + c \log TMP + e \log PBL + f \log UV$$

The modeled regression equation is

$$\log PM = -30.9 + 0.297 \log AOD + 6.32 \log TMP - 0.326 \log PBL - 0.178 \log UV$$

Overall, the results show there are strong correlations between  $\log PM$  and logs of other variables ( $P < 0.001$ ).

Predictor	Coef	SE Coef	T	P
Constant	-30.896	3.174	-9.73	0
LogAOD	0.2972	0.01398	21.26	0
LogTMP	6.3184	0.5672	11.14	0
LogPBL	-0.32644	0.03713	-8.79	0
LogUV	-0.17753	0.02039	-8.71	0

Table 4.2: Regression statistics using the regression equation  $\log PM = a + b \log AOD + c \log TMP + e \log PBL + f \log UV$ .

When log is taken on variables, the violation of homoscedasticity is reduced as seen in Figure 4.7. Also, the normal probability plot of the residuals shows the strong normality. Thus, one can conclude that the overall results of regression analysis are improved. The scatterplot of predicted and measured  $\log PM$  for the Northern part shows a strong correlation, suggesting a better prediction has been achieved. Overall, the predicted PM values are quite close to measured, so that the estimated regression model can be used to do predictions along with other available variables.

### Residual Plots for LogePM

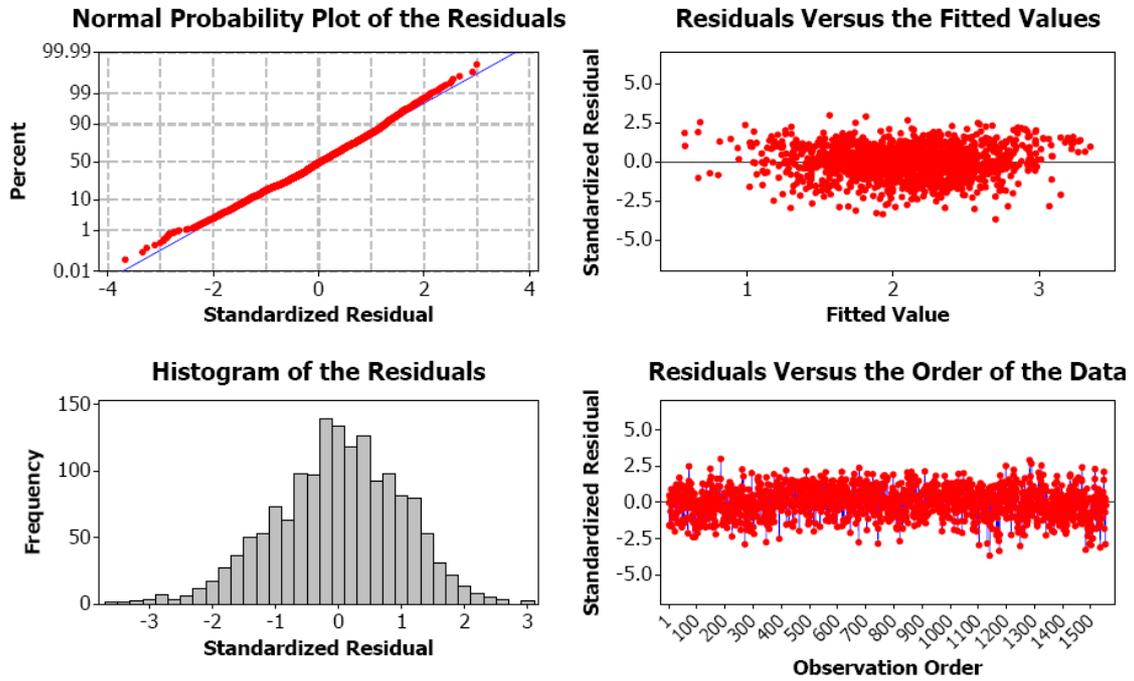


Figure 4.7: Residual plots for  $\log PM$  using the regression equation  $\log PM = a + b \log AOD + c \log TMP + e \log PBL + f \log UV$ .

### Scatterplot of Predicted\_LogePM vs LogePM

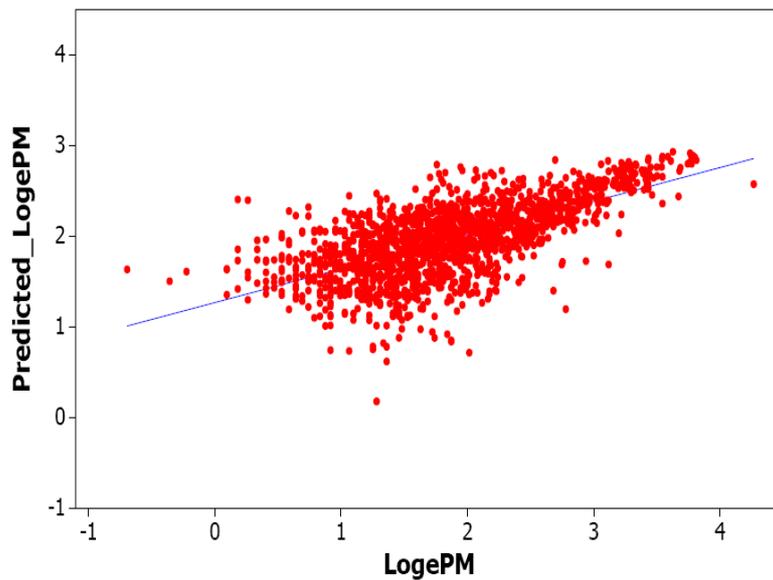


Figure 4.8: Scatterplot of predicted and measured  $\log PM$ .

Another way to correct the problem of heteroskedastic errors or violation of homoscedasticity is to perform WLS regression. As described in subsection 4.2, the logic of WLS regression is to find a weight ( $W_i$ ) that can be used to modify the influence of large errors, and find the ‘best’ fit values of regression constant and coefficients. As in the OLS, the idea is to minimize  $\sum (Y - \hat{Y})^2$ , while in WLS, it is to minimize  $\sum W_i (Y - \hat{Y})^2$ . This process has the effect of minimizing the influence of a case with a large error on the estimation of regression constant and coefficients and maximizing the influence of a case with a small error on the estimation of constant and coefficients.

Technique used in this study to estimate weights is described as follows:

Estimate  $W_i$  by regressing squared residuals ( $e^2$ ) on the offending independent variables, and transform the values of independent variables and dependent variable. This is also called residualizing the independent variables. The regression equation of WLS analysis is presented below:

$$Wt \log PM = 65.3 + 0.250Wt \log AOD - 2.73Wt \log TMP - 2.12Wt \log PBL + 0.857Wt \log UV$$

Predictor	Coef	SE Coef	T	P
Constant	65.254	6.561	9.95	0
WtLogAOD	0.24986	0.01627	15.35	0
WtLogTMP	-2.7298	0.3562	-7.66	0
WtLogPBL	-2.1206	0.1936	-10.95	0
WtLogUV	0.8567	0.1084	7.9	0

Table 4.3: Regression statistics using WLS method.

Compared to the regression results by taking logs on variables, the standard errors of coefficients are quite close, although those from taking logs appear to be smaller. The residuals plots also show the homoscedasticity. Standardized residuals are also normal distributed. Overall performance is acceptable as compared to the above analysis.

In summary, the prediction of ground-level PM<sub>2.5</sub> can be performed by either taking logs on independent variables or using WLS method. The predicted results are very promising. In addition, the analysis suggests that RH is not as important as other variables.

### Residual Plots for WtLogePM

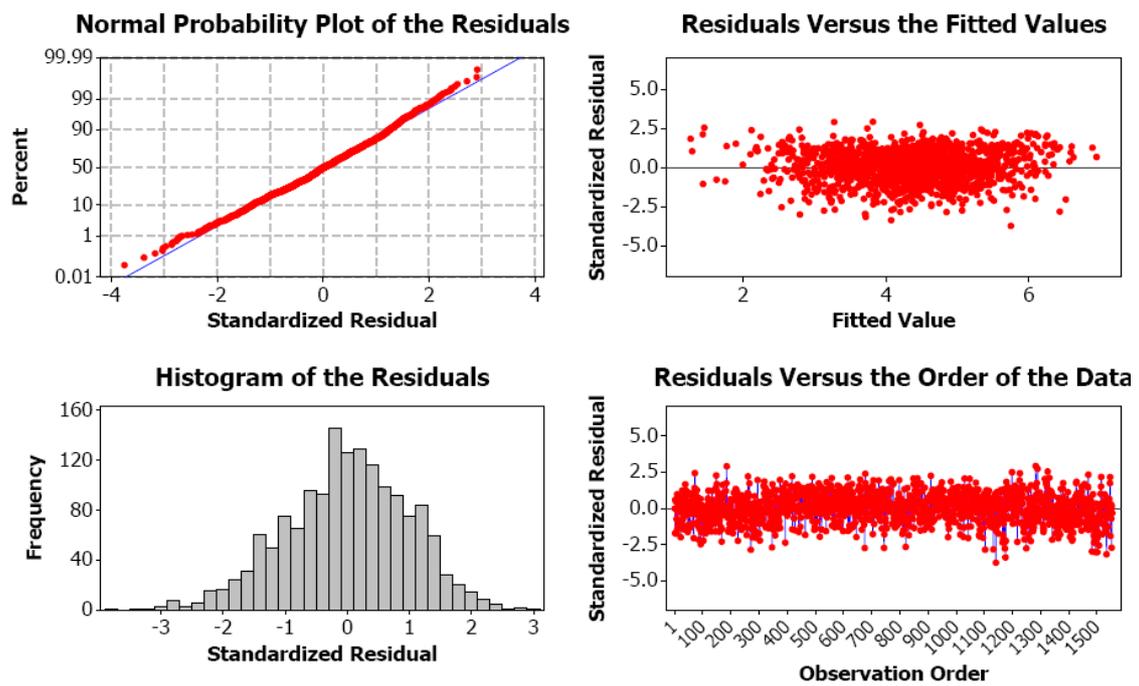


Figure 4.9: Residual plots for log  $PM$  using WLS method.

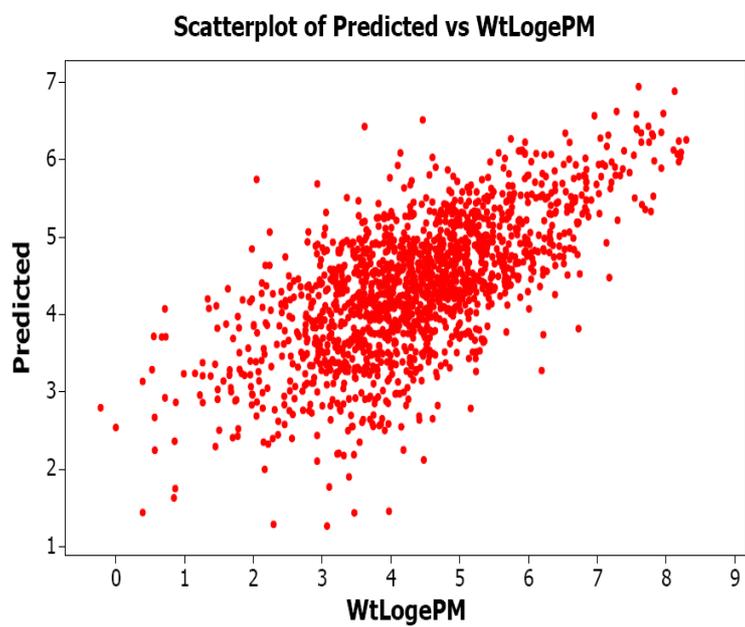


Figure 4.10: Scatterplot of predicted and measured Wtlog *PM* .

## **Chapter Five: Discussion and Conclusions**

This work employed different interpolation techniques and regression methods to reconstruct or estimate ground-level PM<sub>2.5</sub> concentrations. The overall estimations are encouraging. Several direct missing data interpolation methods have the capability to estimate some distinctive features on the basis of available ground-based measurements, while the PEF method tends to generate more information with the aid of satellite AOT information.

General linear regression method is able to predict ground-level PM<sub>2.5</sub> with the consideration of other factors that have been shown to play an important role in predictions. OLS method, when natural log taken on dependent and independent variables, is appropriate for reducing the violation of homoscedasticity. The scatterplot of predicted and measured PM<sub>2.5</sub> shows a strong correlation over the validation region, indicating the ability of the regression model to predict PM<sub>2.5</sub>. WLS method also has advantage in improving homoscedasticity. The predicted and measured PM<sub>2.5</sub> has a relatively high correlation.

This work suggests that some advanced interpolation methods when using remote sensing aerosol product are able to predict ground-level PM<sub>2.5</sub>. The mathematical linear regression methods are applicable in estimating PM<sub>2.5</sub> with the utilization of remote sensing AOD and other meteorological variables. In the analysis, RH does not have strong impacts on PM<sub>2.5</sub> concentration near the surface, while other meteorological variables including TMP, PBL, UV could affect it. This suggests that in the future study, these factors need be considered to infer ground-level PM<sub>2.5</sub> if using regression method. This also implies that in the future work, these factors need to be considered when doing data reconstruction using interpolations. Combination of the two methods might lead to even more promising results. This needs to be further investigated.

## References

- Binkowski, F.S. and S.J. Roselle, 2003, Models-3 Community multiscale air quality (CMAQ) model aerosol component, 1 Model description. *J. Geophys. Res.*, 108(D6) 4183, doi:10.1029/2001JD001409.
- Chu, D. A., Y. J. Kaufman, C. Ichoku, L. A. Remer, D. Tanre', B. N. Holben, 2002, Validation of the MODIS aerosol optical depth retrieval over land. *Geophys. Res. Lett.*, **29**(12):1617.
- Claerbout, J. and M. Brown, 1999, Two-dimensional textures and prediction-error filters, 61st Mtg., Session:1009, *Eur. Assn. Geosci. Eng.*
- Cook, D. and S. Weisberg, 1999, Applied Regression Including Computing and Graphics, *John Wiley & Sons Inc.*
- Di Girolamo, L., T. C. Bond, D. Bramer, D. J. Diner, F. Fettingner, R. A. Kahn, J. V. Martonchik, M. V. Ramana, V. Ramanathan, and P. J. Rasch, 2004, Analysis of Multi-angle Imaging SpectroRadiometer (MISR) aerosol optical depths over greater India during winter 2001–2004, *Geophys. Res. Lett.*, 31, L23115, doi:10.1029/2004GL021273.
- Engel-Cox, J. A., C. H. Holloman, B. W. Coutant, and R. M. Hoff, 2004, Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality, *Atmos. Env.*, 38, 2495-2509.
- ENVIRON, 1998. User's Guide to the Comprehensive Air Quality Model with Extensions (CAMx) Version 2.00. *ENVIRON International Corporation*, 101 Rowland Way, Suite 220, Novato, California 94945-5010.
- Fomel, S., P. Sava, J. Rickett, and J. F. Claerbout, 2003, The Wilson-Burg method of spectral factorization with application to helical filtering, *Geophys. Prospect*, 51(5), 409, doi:10.1046/j.1365-2478.2003.00382.x.
- Grell, G. A., S. E. Peckham, R. Schmitz, S. A. McKeen, G. Frostb, W. C. Skamarock, and B. Eder, 2005, Fully coupled "online" chemistry within the WRF model, *Atmos. Environ.*, 39, 6957–6975.
- Hutchison, K. and et al., 2008, Improving correlations between MODIS aerosol optical thickness and ground-based PM2.5 observations through 3D spatial analyses, *Atmos. Environ.*, 42(3),530-543.
- Neter J. and W. Wasserman, 1990, Applied Linear Statistical Models.
- Kaufman, Y. J., D. Tanré, L. A. Remer, E. Vermote, A. Chu, and B. N. Holben, 1997, Operational remote sensing of tropospheric aerosol over land from EOS moderate resolution imaging spectroradiometer, *J. Geophys. Res.*, 102, 17,051–17,067.
- Kaufman, Y. J., D. Tanre, and O. Boucher, 2002, A satellite view of aerosols in the climate system, *Nature*, 419, 215–223.
- Krewski, D., R. T. Burnett, M. S. Goldberg, K. Hoover, J. Siemiatycki, M. Jerrett, A. Abrahamowicz, W. H. White, 2000, Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality: A Special Report of the Institutes Particle Epidemiology Reanalysis Project, *Health Effects Institute*, Cambridge MA, (97pp).

- Liu, Y., J. Sarnat, V. Kilaru, D. Jacob, P. Koutrakis, 2005, Estimating ground-level PM<sub>2.5</sub> in the eastern United States using satellite remote sensing. *Environ Sci Technol.*, 39(9):3269–3278.
- Mesinger, F., et al., 2006, North American Regional Reanalysis, *Bull. Amer. Meteor. Soc.*, 87, 343 – 360.
- Tanré, D., Y. J. Kaufman, M. Herman, and S. Mattoo, 1997, Remote sensing of aerosol properties over oceans using the MODIS/EOS spectral radiances, *J. Geophys. Res.*, 102, 16,971–16,988.
- van Donkelaar, A., R. V. Martin, and R. J. Park, 2006, Estimating ground-level PM<sub>2.5</sub> using aerosol optical depth determined from satellite remote sensing, *J. Geophys. Res.*, 111, D21201, doi:10.1029/2005JD006996.
- Wang, J., and S. A. Christopher, 2003, Intercomparison between satellite-derived aerosol optical thickness and PM<sub>2.5</sub> mass: Implications for air quality studies, *Geophys. Res. Lett.*, 30(21), 2095, doi:10.1029/2003GL018174.

## **Vita**

Xiaoyan Jiang was born in Changzhou, Jiangsu, China on October 15, 1978, the daughter of Xiuqin Ni and Wangdu Jiang. After completing her undergraduate study at Nanjing University, Nanjing, China, in 2001, she entered the Graduate School at the University of Texas at Austin in 2006.

Permanent Address: No. 12 Jinfeng Cun, Hengshan Qiao Town,  
Changzhou, Jiangsu, China, 213119

This report was typed by author.